Scaling Limits of Algorithms on Large Matrices

Raghav Somani

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:
Sewoong Oh, Chair
Soumik Pal
Zaid Harchaoui

Program Authorized to Offer Degree:
Paul G. Allen School of Computer Science and Engineering

University of Washington

Abstract

Scaling Limits of Algorithms on Large Matrices

Raghav Somani

Chair of the Supervisory Committee:
Sewoong Oh
Paul G. Allen School of Computer Science and Engineering

Advances in large-scale machine learning have been driven by simple and lightweight iterative algorithms. This dissertation explores a broad class of such algorithms that operate on large random matrices, where matrix coordinate processes interact through *mean-field* dynamics. Examples of such algorithms include stochastic gradient-based methods for optimizing the weight matrices of deep neural networks (DNNs), Monte Carlo Markov Chain (MCMC) algorithms for sampling from random matrix models, and the forward pass algorithm in DNNs with weight matrices at each layer.

We demonstrate that, under mild assumptions, iterative algorithms and dynamics on large finite-dimensional matrices exhibit well-defined analytical scaling limits as the algorithm step-sizes approach zero and the dimensionality of the matrix-valued iterates grows to infinity. These scaling limits can be described as processes on infinite exchangeable arrays (IEAs) and analytically characterized as smooth curves on the metric space of graphons and measure-valued graphons (MVGs). The scaling limit of the process can also be described via McKean-Vlasov type stochastic differential equations (SDEs), similar to those studied in the theory of interacting particle systems.

In deriving these findings, we develop a theory of gradient flows on graphons. We introduce new metrics on the space of MVGs that provide a natural notion of convergence for our limit theorems, equivalent to the convergence of IEAs. The analysis reveals that the

scaling limits of popular algorithms like stochastic gradient descent (SGD) and an MCMC sampling algorithm coincide and are *gradient flows* on the space of graphons, uncovering an interesting connection between sampling and optimization. The analysis also demonstrates the *propagation of chaos* phenomenon in large-scale systems, indicating that as the system size grows, the coordinate evolutions become statistically independent.

Finally, we apply these analytical tools to analyze the feedforward dynamics in a linear residual neural network as its depth and width grow to infinity. We again find the propagation of chaos phenomenon at play, demonstrating that as the network size grows, the evolution of any finite set of independently chosen neurons, from the input layer to the output layer, for any fixed input, becomes independent. Moreover, this neuron evolution can be described as a Gaussian process, with drift and diffusion components fully determined by the weights of the limiting network. This allows us to provide an optimal control framework for the risk minimization problem in such infinitely deep and wide networks.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

As I embarked on this Ph.D. journey, I could not have imagined the profound change it would bring me. These years have been a time of significant personal growth. In recognition of the many individuals who have contributed to this transformative journey, I divide my acknowledgments into eight parts.

Firstly, I am deeply grateful to my advisor, Professor Sewoong Oh, whose steadfast support has been fundamental throughout my Ph.D. journey. I am especially thankful for Sewoong's willingness to invest in my potential from the very beginning, right from when we first met in person during NeurIPS 2018, and later during the graduate school visit in 2019. I knew that he would be the perfect advisor who would respect my research interests and provide me with the freedom to explore them even if they are a little different from his own. Along with this, Sewoong's enthusiasm, cheerfulness, calmness, and constant guidance are some of the ideal qualities I could ever ask for in an academic advisor. I thank Sewoong for allowing me to pause and rethink the directions we wanted to pursue, humbly holding back in areas where I had less inclination and conviction, and pivoting to an area that I could call my own research, which has manifested well in the form of this dissertation.

Secondly, I extend my heartfelt thanks to my dissertation reading committee members. I am grateful to Soumik for mentoring me in refining my intuitions into rigorous research ideas and questions that led to the success of this dissertation. His dedication to clarity and his unwavering scholarly urge to push the boundaries of understanding have consistently awed and inspired me. I am especially thankful for his invaluable support and suggestions during times when I oscillated between various career paths. I am very grateful to Professor Zaid Harchaoui for his insightful feedback on my research directions. His deep knowledge

of existing literature and his ability to connect various research topics in novel ways have greatly helped me in refining and better communicating my research to diverse audiences.

Coming from a background in mathematics, computing, machine learning (and optimization) prior to starting my Ph.D., I have always seen reflections of my academic self in the three committee members spanning three departments at the University of Washington (UW). Their combined support has been instrumental in shaping my graduate school journey, providing a solid foundation for my academic pursuits, and guiding me through all phases – both challenging and rewarding.

Thirdly, I am thankful to Raghavendra Tripathi for his collaborative spirit and meticulous attention to technical detail, which greatly enhanced our understanding of complex topics that we encountered in our research together. I am thankful for the invaluable research and career advice from Professor Siva Athreya, which has significantly shaped my outlook on various aspects of life. I am deeply grateful to my close collaborators for their dedication to a first-principles approach in our research, providing enough room for creative and free thinking in the midst of rigorous research exercises that were usually unburdened by academic deadlines. I believe that this process has sharpened both my intuition and critical thinking abilities across various facets of life.

Fourthly, this work was greatly supported by the foundational contributions and collaborations of the individuals who prepared and motivated me to pursue a Ph.D. I am particularly grateful to my pre-doctoral mentor, Dr. Praneeth Netrapalli. Praneeth introduced me to foundational machine learning and optimization with the help of interesting research problems, and later helped me refine some of the research questions that I eventually asked in this Ph.D., significantly influencing my research path. I am also thankful to Praneeth and Dr. Prateek Jain for jointly recognizing my research acumen early on during my undergraduate final year and giving me the opportunity to work with them as a research fellow at Microsoft Research India (MSRI) for two wonderful years before I joined as a Ph.D. student at UW.

Fifthly, I value the instruction and guidance from Professors Shayan Oveis Gharan, Anna Karlin, Yin Tat Lee, and Dmitriy Drusvyatskiy, whose course teachings were pivotal in the early stages of my research. I appreciate the Artificial Intelligence & Machine Learning Lab in the Gates Center for Computer Science and Engineering (CSE) for providing a supportive and engaging research environment and workspace. I thank Elise Dorough and Joe Eckert for their consistent administrative support throughout my Ph.D. I also want to give special acknowledgment to the 'Kantorovich Initiative' and the 'Institute for Foundations of Machine Learning' for providing a stimulating audience and valuable feedback throughout my research journey. I thank the 'National Science Foundation' for facilitating the funding for the majority of my research with my research collaborators. I thank The D. E. Shaw Group for the internship opportunity, which allowed me to demonstrate my research skills and learn about the vast potential of my research from my colleagues.

Sixth, I am immensely grateful to my housemates – Aditya Kusupati, Sanjana Sharma, and Tapan Chugh – who have made our residence and life in Seattle a true haven. I fondly remember when Aditya, Tapan, and I, while working at MSRI together, decided to join UW, planning to stay together and work on machine learning problems. As aspiring young researchers, we agreed that Aditya would focus on ML applications, I would provide the necessary theoretical explanations, and Tapan would improve the systems behind them. We are glad that some of our early research ideas successfully worked out, and we constantly learned so much from each other during our stay together. I also thank Aditya, Tapan, and Sanjana for tagging me along on random drives, various hikes, and exploring places, given my reluctance to learn driving. Special thanks to Sanjana, who joined our house in the midst of our Ph.D. and made it a home. Watching TV was so much fun when Sanjana commanded the remote. As I have always expressed to her, thanks for raising the entropy of our social lives.

I also cherish the fun times hanging out with Sahil Verma, Sudheesh Singanamalla,

Priyal Suneja, Venkatesh Potluri, and Pratyush Patel. I especially thank the 'desi-grads' friend community in CSE UW, who provided me with a social environment reminiscent of my home country, making me feel closer to home under the gloomy skies of Seattle. I hope the group stays active and grows as newer graduate students join and stay in touch with the older ones.

Seventh, my deepest gratitude is reserved for my family and friends. To my parents, for providing me with the best education they could. They have played a pivotal role in helping me make some of the best decisions of my life. My father, coming from a commerce and finance background, initially wanted me to pursue a similar path, given that this was the path taken by the vast majority of people across our entire ancestral family tree. However, he trusted my academic and career decisions throughout this journey and provided the best mentorship by learning about my ever-evolving interests and the places I attended. My mother complemented my father by ensuring I was surrounded by a good social and friend circle, trusting my decisions as long as she knew I would be happy pursuing them. As their only child, I was not burdened with the socioeconomic responsibilities that might be expected. I thank them for teaching me important life lessons during difficult times, which helped me thrive, learn, and persist both in my academic as well as personal life.

I thank my grandmother, uncle, and elder sister's family for their boundless love and encouragement. I thank my late grandfather for the love he showered on me when I was a young schoolboy. I thank my family-in-law for their unwavering love, trust, and support for my constraints and decisions.

I warmly thank my friends, from my enduring 'B3 family' to my childhood school friends, who have supported and cheered for me and always had my back. Cheers to the crazy memories we've had together, and may we create many more.

Finally, to my wonderful wife, Surabhi, who has been a cornerstone in my life since my early undergraduate days. Her love, support, strength, and companionship have been my

anchor during turbulent times. She was the first person with whom I shared my desire to pursue a Ph.D. in the USA, which could take another five or more years of staying apart. It was a surprise to her, but she stood by me and supported me throughout this journey. Despite being poles apart (literally), she ensured she was always emotionally available. Her trust and love know no borders. Thank you for being my partner in every decision we have made together. Thank you for listening to all my dilemmas and helping me make important choices along the way. I am glad that we have stood strong across years and oceans. This Ph.D., and indeed our entire academic journey, is our collective achievement – a testament to our shared sacrifices, support, and love. Thank you for believing in me more than I believed in myself. I am happy that you are proud of how far we have come. With this chapter also coming to an end, I hope we will be able to spend more time close to each other. Thank you for everything, Surabhi!

I thank and acknowledge everyone who has been part of the journey, directly or indirectly, consciously or unconsciously. I end this section thanking life for the freedom and opportunities it provides.

DEDICATION

To the memory of my beloved grandfather,

Late Shri Jugal Kishore Somani.

Chapter 1

INTRODUCTION

Iterative algorithms are fundamental to machine learning, large-scale optimization, and statistical estimation, forming the basis for many advanced computational methods. These algorithms often operate as discrete-time stochastic processes over high-dimensional spaces, such as real-valued vectors and matrices. Popular examples of such algorithms and dynamics include Stochastic Gradient Descent (SGD) [RM51, Bub15, BCN18, Net19], Metropolis-Hastings sampling algorithm [MRR$^+$53], Oja's algorithm [Oja82], and the forward pass computation of a Deep Neural Network (DNN) with a set of trainable weight matrices. With growing data and computation, recent advancements in machine learning have been made possible by using large, internet-scale data sets to train increasingly larger parametric machine learning models for extended periods on ever-more efficient computational infrastructures.

Since the ambient space comprises a large set of parameters that update with each iteration, this results in complex interactions between the evolutions of every parameter. As the dimension of the underlying parametric space grows, it becomes imperative to study how the many possibly dependent (possibly stochastic) processes corresponding to each parameter interact and evolve. These iterative algorithms exhibit two pivotal properties:

1. Updates in these algorithms depend solely on the configuration of previous iterates rather than on any specific labeling or ordering of the parameters. This observation ensures that the algorithm's behavior remains invariant under any permutation of parameter indices, reflecting an invariance in the representation of the parameter space.

2. The evolution of iterates is governed by an aggregate effect analogous to mean-field

theory, where the influence of any single parameter on the update of another is relatively insignificant. However, the collective influence of all parameters plays a significant role in determining the trajectory of each iterate, embodying a global rather than a local dynamical perspective.

The various symmetries in the interaction of these processes allow us to derive simple and beautiful descriptions of their joint evolution and characterize various phenomena that arise at scale. We will develop a framework of analysis that provides a common ground for a family of iterative algorithms exhibiting such properties, enabling us to understand various large-scale behaviors that simplify our understanding of these algorithms.

We commonly observe mean-field interactions in various physical systems. One such example in the natural sciences involves the behavior of gas molecules under classical physics laws. In this system, each molecule is influenced by the average effect of all other molecules rather than detailed pairwise interactions. To describe this macroscopically, one considers the statistical distribution of molecules' positions and velocities – commonly the Maxwell-Boltzmann distribution. As the number of molecules tends toward infinity (the thermodynamic limit), the system's macroscopic state increasingly represents individual molecules' states. Here, the mean-field theory simplifies the description of molecular dynamics by focusing on "average properties" rather than "complex many-body interactions". Developments in the theory of large interacting particle systems and optimal transport have provided a geometry and an analytical framework to discuss the behaviors of such large mean-field interacting particle systems that enjoy complete permutation symmetry [Szn91, Vil03, Vil12, San15].

The training dynamics of two-layer neural networks exhibit mean-field interactions similar to those in systems of interacting gas molecules. Drawing a direct analogy from interacting particles, we can see how the weights of neurons (analogous to the positions and velocities, i.e., the states of particles) in a two-layer neural network change under the potential field of the loss function, which Gradient Descent (GD) follows over discrete time steps. Recent analyses of Stochastic Gradient Descent (SGD) over two-layer neural networks have sought to

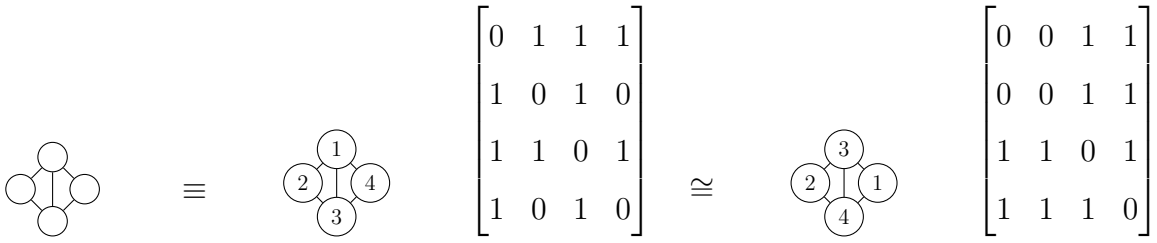


Figure 1.1: Symmetry in unlabeled graphs.

understand these interactions using the theory of Wasserstein gradient flows and McKean-Vlasov equations, which arise from such mean-field interactions [SMN18, CB18, RVE18, MMM19, CCP19, AOY19, NP20, SS20a, SS20b, TR20, BC21]. These studies have provided insights into phenomena like the propagation of chaos, which essentially means that as the width of the hidden layer (i.e., the number of hidden neurons) grows towards infinity, the evolution of the weights of neurons becomes statistically independent.

In many interesting applications, however, the "particles" are edge weights in a graph (or its adjacency matrix) whose vertex labels are exchangeable, but the edges themselves are not. This arrangement does not afford complete permutation symmetry since permuting the edges changes the connectivity of the underlying graph. See Figure 1.1 for an illustration. The adjacency matrices of such unlabeled graphs are classified as exchangeable matrices, where the distribution of any submatrix remains invariant under permutations of row and column labels. In this context, a mean-field interaction suggests that while the impact of a single edge on the evolution of another is negligible, the collective influence of the entire unlabeled graph significantly shapes the trajectory of every edge. Motivated by the dynamics of various classes of algorithms over graphs, this thesis aims to extend Wasserstein calculus to higher-order exchangeable structures like infinite two-dimensional exchangeable arrays and to investigate the characterization of scaling limits for various iterative stochastic algorithms over such large matrix-valued processes that exhibit mean-field interactions.

To further elucidate the concept of symmetry in neural networks, consider the setup of

a deep neural network. We note some crucial observations: First, relabeling neurons within any given layer does not affect the system's overall behavior. The network's output remains unchanged under the permutation of neurons in any layer as long as the connections in the computational graph are maintained. Second, this symmetry is not a complete permutation symmetry since the permutation symmetry between any two adjacent layers is shared by a common set of computational nodes called hidden neurons. See Figure 1.3 for a diagram. These observations prevent one from using the theory of interacting particle systems to derive scaling limits as the number of hidden neurons goes to infinity.

With these observations, we will consider three classes of iterative algorithms where such a symmetry exists. We first briefly describe each of the classes to set the context of the rest of the thesis:

1. Optimization: Optimization algorithms, such as those based on stochastic gradients, play a crucial role in large-scale optimization and machine learning [BB07]. Each step of the algorithm involves evaluating an estimate of the negative gradient of the objective risk function and moving in that direction. Popular examples of such dynamics include the training dynamics of large deep-learning models using SGD-based algorithms. See Section 2.2.1 for details of such algorithms.

2. Sampling: When finding the minimizers of an objective function or computing its gradient is computationally expensive, one often considers sampling from its corresponding Gibbs measure with some temperature parameter. In many circumstances of interest, it is well-known that as the temperature approaches zero, the Gibbs measure concentrates around the minimizers of the objective Hamiltonian function [Hwa80]. Thus, one may replace the problem of optimization of the Hamiltonian function with a problem of sampling from the Gibbs measure at a low temperature. One can sample from the Gibbs measure by running suitable Markov chains with an invariant distribution given by the Gibbs measure [vLA87]. Due to this relation between sampling and optimization, these algorithms are also called zeroth-order optimization algorithms. See

Section 2.2.2, where we discuss a relaxed Metropolis chain algorithm operating over stochastic block models.

3. Multiplicative updates: Another class of algorithms relies on the iterative application of linear transformations rather than additive changes as seen in optimization and sampling algorithms. These matrices are random-scaled perturbations of identity. This class includes algorithms such as Oja's algorithm, gossip algorithms, and the forward pass in deep residual neural networks. See Section 2.2.3, where we describe the setup in greater detail.

The analysis of algorithms categorized under cases 1 and 2 becomes complex at large scales due to the mean-field interactions among the coordinates of the iterates that are exchangeable matrices. As previously discussed, the restricted permutation symmetry limits our ability to apply established theories of interacting particle systems and McKean-Vlasov theory to characterize the scaling limits of dynamics over such spaces. Additionally, for dynamics categorized under case 3, analyzing the dependencies and correlations arising from the progressively multiplicative updates, given their matrix-multiplicative nature, is essential to characterize their scaling limits effectively.

Despite these challenges, the possibility of identifying limiting objects to characterize the scaling limits is promising. This optimism stems from the observation that the core principle of the algorithm remains consistent, regardless of the ambient space's dimensionality. We see this in classical interacting particle systems, which enjoy complete permutation symmetry. This symmetry allows us to describe the system as a function of the empirical measure of a set of interacting particles. The evolution of their empirical measure can equivalently describe the evolution of the set of particles. The space of probability measures, known for its elegant geometry derived from the theory of optimal transport [Vil03, San15], enables one to embed all empirical measures obtained from various particle sets into the single common space of probability measures. Since this space is complete, the limit as the number of particles grows can be described by the limit of the empirical measures in this space. Since this space is

also a geodesic metric space, this framework allows us to connect particle dynamics with the corresponding dynamics on the space of measures, providing a calculus facilitating the desired scaling limits. This thesis aims to extend this framework to dynamics over higher-order exchangeable structures, such as two-dimensional arrays from popular matrix-valued algorithms and graph dynamics.

Exchangeable matrices can be considered adjacency matrices of large, dense, edge-weighted, unlabeled graphs. The concept of limits for such large unlabeled, edge-weighted graphs, known as graphons, has been extensively studied [Lov12]. Central to our analysis is the novel application of gradient flows [AGS08] and dynamics over infinite exchangeable arrays. We explore each class of algorithms, initially studying their continuous-time limits at fixed dimensions, developing an analytical framework to embed iterates in a common, suitable limiting space, and ultimately deriving their scaling limit descriptions as dimensions grow to infinity.

The research and study have been progressively developed.

- In case 1, the existing theory of graph limits, called *graphons*[Lov12], combined with the general theory of gradient flow on metric spaces[AGS08], allows us to develop a theory of gradient flow on graphons. We show that standard Euclidean gradient flow on matrices can be embedded into the space of graphons, over which we can consistently study their limits. We demonstrate that these limits naturally turn out to be well-defined gradient flows on graphons. The task then reduces to obtaining continuous-time limits of iterative algorithms over finite dimensions, which we can embed into the space of graphons to further take their scaling limits under a suitable topology. This framework helps us understand macroscopic evolution and demonstrates phenomena such as the propagation of chaos that arises from scaling the dimensions to infinity.

- We extend the framework discussed in the above item to incorporate the microscopic statistical details of the evolution of every coordinate in the exchangeable array system by characterizing the scaling limit as smooth deterministic curves in the space of

measure-valued graphons (MVG). We provide amenable metrics on the space of MVGs that stay consistent with the topology and geometry of graphons. We show a correspondence between two-dimensional real-valued infinite exchangeable arrays and the metric space of MVGs. The scaling limit characterizations of the matrix-valued processes adopt the form of McKean-Vlasov equations, which are traditionally linked to the study of interacting particle systems and optimal transport.

- Beyond first-order stochastic optimization algorithms, as discussed in case 2, our analysis extends to demonstrate that the scaling limit of a simple Monte Carlo Markov Chain (MCMC) sampling algorithm over a stochastic block model (SBM), also leads us to a McKean-Vlasov description over MVGs. In the asymptotically zero-noise case, this curve, when projected down to graphons, coincides with the gradient flow of the Hamiltonian function on the space of graphons. This highlights a beautiful connection between sampling and optimization, illustrating that an algorithm that does not use any gradients can still mimic a gradient flow under the scaling of time and dimensions.

- Moving to case 3, we consider the scaling limits of iterative products of random and biased perturbations of the identity matrix. Even when these product matrices contain Gaussian noise, we show that the Gaussian chaos arising from successive products of Gaussian random variables converges to Gaussian processes due to the Central Limit Theorem (CLT) as the matrix size increases, simplifying the distributional description of the infinite product. This analysis allows us to characterize Deep Linear Residual Networks as Gaussian processes beyond their random initialization and once again unveil the phenomenon of propagation of chaos. We discuss this as an application of our study in Chapter 7.

In the following sections of the chapter, we will give an introduction and overview of the algorithm classes we consider, discuss the permutation symmetries appearing in Deep Neural Networks (DNN) and use them to parameterize wide and Deep Linear Residual Networks

and discuss the contributions and the structure of this dissertation.

1.1 Finite dimensional iterative algorithms

Iterative stochastic optimization algorithms serve as the workhorses of machine learning [Cau47, Bub15, BCN18]. The specifics of these common Markov chains will be discussed in Section 2.2.1. For a more detailed overview, readers are referred to the following monographs [Ben99, KY03, Bor09, MB11, KC12]. Analyzing such discrete stochastic processes can be challenging in general. A common approach, inspired by stochastic approximation theory, involves examining their continuous-time counterparts. The insights gained from continuous-time dynamics can be very valuable, as phenomena demonstrated to occur in continuous time approximately apply even to discrete-time algorithms when the step-sizes are small enough.

As discussed in case 1 and 2, iterative algorithms can be used to optimize (or sample from the Gibbs measure of) real-valued functions that are *permutation invariant*, that is, the output of these functions does not change if the rows and columns of the adjacency matrix are permuted using a common permutation. This property is crucial in giving rise to the mean-field interaction between the entries of the iterates.

Definition 1.1. For any $n \in \mathbb{N}$, we say that a function R_n on symmetric $n \times n$ matrices is *permutation invariant* if

$$R_n(X_n) = R_n\Big(\big(X_{n,(\pi(i),\pi(j))}\big)_{(i,j)\in[n]^2}\Big),$$

for all $n \times n$ symmetric matrices X_n and all permutations π of $[n] \coloneqq \{1, 2, \dots n\}$.

We will make some observations through an example.

Example 1.1. Consider the function R_n on symmetric $n \times n$ real-valued matrices for any $n \in \mathbb{N}$ defined as $R_n \colon X_n \mapsto n^{-3} \operatorname{tr}(X_n^3)$. If X_n is the adjacency matrix of a graph, then R_n computes the triangle density of the graph. We now make a few observations here:

1. In this example, the gradient is the map $\nabla R_n \colon X_n \mapsto 3n^{-3}X_n^2$. Observe that if we look at the update for any index $e \in [n]^2$ of X_n, we see that the e-th coordinate of the gradient evaluated at X_n not only depends on the element $X_{n,e}$, but also on the entire matrix X_n. Such interactions are called *mean-field interactions* as we will study in the later chapters.

2. The update rule is consistent if we relabel the rows and columns by some common permutation of the index set $[n]$. Moreover, $(R_n, \nabla R_n)$ is permutation invariant.

3. When n is large, the evaluation of the function R_n does not change by much as n continues to increase, given the input to R_n approximates as graphs in a certain sense. We will return to this notion of convergence later in Chapter 2.

More generally, in the case of Deep Neural Networks (DNNs), the output of the DNN does not change if the neurons in any of the hidden layers are permuted. This permutation-invariant property in DNNs is discussed in detail in Section 1.4. Since we consider unlabeled graphs, well-defined functions over unlabeled graphs always satisfy this property.

Gradient descent, when used with small step sizes, approximates the Euclidean gradient flow obtained as a solution to Cauchy's problem [Cau47]:

$$\frac{\mathrm{d}}{\mathrm{d}t}X_{n,e}(t) = -\,\partial_e R_n(X_n(t)), \qquad e \in [n]^2, \qquad t \in \mathbb{R}_+,$$

with a given initialization $X_n(0) = X_{n,0}$. Here, $\mathbb{R}_+$ denotes the set of all non-negative real numbers, which is used to index time, and ∂_e refers to the partial derivative with respect to the e-th matrix entry. The solution of this Cauchy's problem, which exists and is unique when ∇R_n is Lipschitz [Lin94], is often called the *gradient flow* of R_n on $\mathbb{R}^n$. These gradient flow curves possess a nice desirable property: under the time limit (i.e., $t \to \infty$) they converge to the set of stationary points of a wide class of optimization objectives. A natural question that arises in several applications is whether the solution to the above Cauchy problem has a *scaling limit* as $n \to \infty$. That is, is there a simple description of the limit of these solutions

as $n \to \infty$? One can also ask the reverse question that what class of discrete time algorithms can approximate the gradient flow in finite dimensions. If such a class can be identified, then the conclusions obtained by taking the scaling limit of the gradient flow as the dimension grows can equally apply to such other optimization algorithms. We will introduce one such wide class of algorithms in Section 2.2.1.

Consider the problem of 'soft' optimization of a Hamiltonian function H_n defined over $\mathcal{M}_{n,+}$, the set of symmetric matrices taking values in $[0, 1]$. Instead of finding the actual minimizers, which can be computationally expensive, one often considers a Gibbs measure corresponding to H_n, whose density with respect to the Lebesgue measure on $\mathcal{M}_{n,+}$, is proportional to $\mathrm{e}^{-\beta H_n}$, for some inverse temperature parameter $\beta > 0$. It is well known that, in many circumstances of interest, as $\beta \to \infty$, the Gibbs measure concentrates around the minimizers of H_n [Hwa80]. Thus, one may replace the problem of optimization of H_n by a problem of sampling from the Gibbs measure for a large β. This is achieved by running suitable stochastic processes with an invariant distribution given by the Gibbs measure [vLA87]. A large class of commonly used models, including those in statistical physics and exponential random graph models (ERMs), falls under this umbrella [Che16]. One might wish to sample from such a Gibbs measure, whether trying to find graphs that approximately minimize the Hamiltonian (i.e., serving as an approximate non-parametric maximum likelihood estimator, such as MCMLE [Che16, Chapter 3.3]) or sampling from a Bayesian posterior distribution. Although Metropolis or Gibbs sampling algorithms are popular choices for running MCMC algorithms, their mixing times are generally not known. In Section 2.2.2, we will introduce a relaxed Metropolis chain algorithm that can be used to sample from the Gibbs measure corresponding to H_n. In Chapter 6 provide non-asymptotic error bounds for this algorithm.

In Chapter 3, we will see that such a class of iterative stochastic optimization algorithms, as well as the relaxed Metropolis chain algorithm, are closely related to stochastic differential equations (SDEs) of the following form:

$$\mathrm{d}X_{n,e}(t) = b_{n,e}(X_{n,e}(t), X_n(t))\,\mathrm{d}t + \Sigma_{n,e}(X_{n,e}(t), X_n(t))\,\mathrm{d}B_{n,e}(t), \qquad t \in \mathbb{R}_+, \tag{1.1}$$

for $e \in [n]^2$, and with initialization $X_n(0) = X_{n,0}$. Here, b_n and Σ_n are some matrix-valued drift and diffusion coefficients functions, respectively, that are coordinate-wise functions of the corresponding coordinates of the matrix, as well as of the entire matrix. As discussed above, the way in which these functions depend on the entire matrix is permutation invariant, which results in the mean-field interaction between the elements of the matrix. Here, B_n is an $n \times n$ symmetric matrix-valued process with coordinatewise independent standard Brownian motions. These SDEs are of the McKean-Vlasov type exhibiting the mean-field evolution of the matrices. When the domain of the algorithm has a boundary, the SDE that we will obtain will contain reflection terms necessary to accommodate the boundary conditions. Consequently, one might again ask whether such general classes of iterative optimization and iterative sampling algorithms have a scaling limit as the step-size goes to zero and the size of the matrices increases to infinity.

Moving to algorithms of the multiplicative form, we also consider the class of dynamics that generate a Markov process via repeated linear transformations. Here, each linear transformation is a random and possibly biased perturbation of the identity. The resulting linear transformation can be considered as an iterated product of small transformations, which, when applied to an initial state, allows us to understand the trajectory of the state as a function of the number of iterations. It is natural to ask how this trajectory behaves as we scale down the perturbation (with respect to identity) for each update and, at the same time, increase the number of such updates. In order to arrive at a non-trivial continuous-time limit of such discrete iterations, it is important to scale the perturbations as a function of the number of iterations. Specifically, if $m \in \mathbb{N}$ is the total number of iterations, then any drift component of the perturbation needs to scale as m^{-1}, whereas the i.i.d. random noise in each perturbation needs to scale as $m^{-1/2}$ for the limit to be non-trivial. Let us look at an example where $n = 1$.

Example 1.2. For any time $t \in \mathbb{R}_+$ and $m \in \mathbb{N}$, consider the following product of $m \in \mathbb{N}$

many perturbations of 1; i.e., for every m, consider:

$$P_m(t) \coloneqq \prod_{i=1}^{\lfloor mt \rfloor} \left(1 + \frac{\mu}{m} + \frac{\sigma}{\sqrt{m}} X_i\right), \qquad t \in \mathbb{R}_+,$$

where $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$ are the drift and diffusion coefficients, and $(X_i)_{i\in[m]}$ are i.i.d. standard Gaussian random variables. It follows from [DM83] that as $m \to \infty$, the càdlàg process P_m weakly converges to the stochastic process $P\colon t \mapsto \mathrm{e}^{\mu t + \sigma B_t - \sigma^2 t/2}$ as $m \to \infty$, where B is a standard Brownian motion (BM). Notice that the process P is nothing but the stochastic exponential of the process $Y\colon t \mapsto \mu t + \sigma B_t$.

In Section 2.2.3, we will introduce such iterative products in higher dimensions, where the terms of the products may not necessarily commute as matrices and may even change smoothly with each iteration. It will be crucial to scale the drift and noise terms appropriately based on the dimension to arrive at non-trivial limits. In Section 1.4.2, we will see that such iterated matrix products arise in the forward pass computation in large Deep Residual Networks with linear activation functions. Consequently, it is relevant to revisit the question whether this class of dynamics over matrices has a scaling limit as the number of iterations and the dimensionality of the matrices increase. In the context of Deep Residual Networks, the number of iterations corresponds to the network's depth, and the dimensions relate to the number of neurons in each network layer. This will allow us to also uncover the scaling limits of the evolution of neurons along the depth of a network for a given set of random weight matrices. We will discuss this application in Chapter 7, and find a similar propagation of chaos behavior in the evolution of neurons along the depth.

In the following section, we will provide an overview of the classical theory of interacting particle systems, where such scaling limits have been studied. A similar question is encountered in this theory; however, instead of symmetric two-dimensional arrays that correspond to weighted edges, the theory examines a one-dimensional array of particles with complete permutation symmetry.

1.2 Classical Interacting Particle Systems (IPS)

The theory of interacting particle systems (IPS) study evolutions of one dimensional arrays that correspond to the positions of the interacting particles [Gär88]. The system consists of a large number of interacting particles, where the effect of each individual particle on the whole system is negligible, but the cumulative effect of all the particles is significant. This is the mean-field model of interaction. In this context, one encounters a similar system of SDEs where multidimensional diffusions of particle positions X_n interact through their empirical distribution as:

$$\mathrm{d}X_{n,i}(t) = b_n(X_{n,i}(t), \hat{\mu}_n(t))\,\mathrm{d}t + \sigma_n(X_{n,i}(t), \hat{\mu}_n(t))\,\mathrm{d}B_{n,i}(t), \qquad t \in \mathbb{R}_+, \tag{1.2}$$

for $i \in [n]$ and with initialization $X_n(0)$. Here, b_n and σ_n are some drift and diffusion coefficient functions respectively. Here, $X_{n,i}(t)$ is the position of the i-th particle at time $t \in \mathbb{R}_+$, $(B_{n,i})_{i\in[n]}$ is a vector-valued process with coordinatewise independent Brownian motions, and $\hat{\mu}_n(t) \coloneqq n^{-1}\sum_{i\in[n]} \delta_{X_{n,i}(t)}$ is the empirical probability measure of all the n particles at time $t \in \mathbb{R}_+$. Any drift and diffusion that is symmetric in the coordinates (i.e., "mean-field interactions") can be represented through some functions b and σ via an SDE of the above form. The study of such systems originated from the probabilistic study of the Boltzmann and Vlasov equations [Kac56, McK75, Dob79, Tan79]. For modern surveys, readers are referred to [Szn91, Vil12, CD22, Jab14].

In this mean-field model, each particle's influence is approximated by the average influence of all the other particles, making the limiting characterization simpler to work with. Under suitable assumptions, it is known that the process of empirical distribution of the particle system, $t \mapsto \hat{\mu}_n(t)$, converges to the solution of well-known partial differential equations (PDEs) as $n \to \infty$. The convergence is often obtained via *propagation of chaos*, where in the limit as $n \to \infty$, a finite collection of randomly chosen particles evolve independently

and are identically distributed according to the McKean-Vlasov SDE:

$$\begin{aligned} \mathrm{d}X(t) &= b(X(t), \mu(t))\,\mathrm{d}t + \sigma(X(t), \mu(t))\,\mathrm{d}B(t), \\ \mu(t) &= \mathrm{Law}(X(t)), \end{aligned} \qquad t \in \mathbb{R}_+,$$

where $\mu(t)$ is the weak limit of the measures $(\hat{\mu}_n(t))$ as $n \to \infty$ at time $t \in \mathbb{R}_+$. This essentially indicates that the dynamics of a randomly chosen particle can be described by a deterministic drift and diffusion, that depends both on the position of the particle and the distribution of the entire ensemble of particles.

There has been a recent surge of interest in the application of the above convergence in the context of neural networks, as seen in [SMN18, CB18, RVE18, MMM19, CCP19, AOY19, NP20, SS20a, SS20b, TR20, BC21, CCFRF22, SABP22]. For instance, consider a single hidden layer neural network as shown in in Figure 1.2 with n hidden neurons. Given an input $x_0 \in \mathbb{R}^d$, and the network weights represented as each row of a matrix $A \in \mathbb{R}^{n\times d}$, the output $\hat{y}(x_0)$ of the network can be computed as $\hat{y}(x_0) = \frac{1}{n}\sum_{i=1}^{n} \sigma(A_{i,*}x_0)$ for some activation function $\sigma\colon \mathbb{R} \to \mathbb{R}$. The goal of risk minimization is to minimize the expected risk function $R_n(A) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(Y, \hat{y}(X))]$ for some data distribution μ supported on $\mathbb{R}^d \times \mathbb{R}$ and some loss function $\ell\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$. The permutation symmetry in the network can be observed by noticing that if we permute the neurons using a permutation $\pi \in S_n$, the output of the network stays the same as seen from the identity $\hat{y}(x_0) = \frac{1}{n}\sum_{i=1}^{n} \sigma\big(A_{\pi(i),*}x_0\big)$. Therefore we have $R_n(A) = R_n(A^\pi)$ where A^π is the matrix A with rows permuted by π.

In this setup, the particles are the neurons of the network, and the evolution of these particles along the iterative optimization process is governed by a mean-field interaction coming from algorithms like stochastic gradient descent (SGD). It has been shown in works like [SMN18, CB18] that as $n \to \infty$ and as the step-size of the algorithm goes to zero, the dynamics of the SGD algorithm appropriately converge to a Wasserstein gradient flow of a risk function R defined over the law of the weights of the neurons of the network (i.e., $\mu = \lim_{n\to\infty} \hat{\mu}_n = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \delta_{A_{i,*}}$). Further, under some conditions on $\mathcal{D}$, σ, and ℓ, the Wasserstein gradient flow converges to the global minima of the risk function. On the

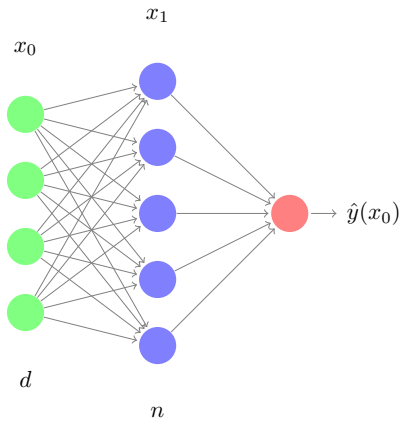


Figure 1.2: Single hidden layer Neural Network with n neurons

contrary, if one sticks to the Euclidean setup corresponding to the domain of R_n, such insights are not immediate majorly due to the Euclidean non-convexity of the function R_n for every $n \in \mathbb{N}$.

1.3 Interacting Graph Systems: Beyond IPS

The difference between the classical interacting particle system and the system of interacting graphs that we consider, lies in the fact that the objective function (which determines the drift) need not be symmetric in all the n^2 (up to matrix symmetry) many coordinates of the argument matrix. Therefore, such a function need not be expressed as a function of the empirical distribution of the matrix entries. Since these interactions are obtained via a different symmetry group, these interactions may not be of the usual mean-field type. Due to this fundamental difference, even if SDE (1.2) appears to be of the same type as SDE (1.1), the classical McKean-Vlasov theory does not apply immediately. This leads us to ask the same question again if does the system of interacting graphs under the kind of mean-field interaction discussed, exhibit properties like propagation of chaos when the size of the matrices go to infinity.

The permutation symmetry in matrices as we described earlier, can be captured by

graphons (see Section 2.5 for a preliminary introduction) and measure-valued graphons (MVG) (see Chapter 5). In other words, such functions can be considered as functions on the space of graphons instead of probability measures. This abstraction, which involves considering the appropriate symmetry group and characterizing the scaling limit, proves advantageous in several aspects. We will see that here too, the entries of the matrix exhibit the propagation of chaos phenomenon, i.e., if we sample a finite submatrix out of the system, then, conditioned on the labels of the submatrix, the entries of the submatrix evolve independently as the size of the system grows to infinity. We will also find in Chapter 4 that this symmetry prescribes much weaker conditions for obtaining exponential rates of convergence to the minimizers.

In the next few sections, we will introduce the wide class of dynamics that we will study.

1.4 Permutation symmetries in Deep Neural Networks (DNNs)

In this section, we show that the Deep Neural Networks (DNN), in this illustration, a feed forward DNN is permutation invariant in a certain sense. See Figure 1.3, where the NN consists of $m \in \mathbb{N}$ hidden layers, with widths $n = (n_0, n_1, \dots n_m) \in \mathbb{Z}^{m+1}$ an initial input $H_0 \in \mathbb{R}^{n_0}$, and a terminal output $\hat{y}(H_0) \in \mathbb{R}$ (say), and an intermediate sequence of transformations

$$\mathbb{R}^{n_0} \ni H_{n,0} \mapsto H_{n,1} \in \mathbb{R}^{n_1} \mapsto H_{n,2} \in \mathbb{R}^{n_2} \mapsto \cdots \mapsto H_{n,m} \in \mathbb{R}^{n_m} \mapsto \hat{y} \in \mathbb{R}.$$

Each transformation involves a matrix $\theta^{(m)}_{n,k+1} \in \mathbb{R}^{n_{k+1} \times n_k}$, a vector $b^{(m)}_{n,k+1} \in \mathbb{R}^{n_{k+1}}$, and the transformation is defined as

$$H_{n,k} = \sigma\Big(\theta^{(m)}_{n,k} H_{n,k-1} + b^{(m)}_{n,k}\Big), \qquad k \in [m], \tag{1.3}$$

where the function $\sigma \colon \mathbb{R} \to \mathbb{R}$ acts coordinate-wise. Finally, take $\hat{y}$ to be just the average of elements in $H_{n,m}$.

Given some probability distribution μ on $\mathbb{R}^{n_0} \times \mathbb{R}$, the goal is to minimize the risk function R_n, given by quadratic loss here for specificity,

$$R_n\left(\left(\theta^{(m)}_{n,k}, b^{(m)}_{n,k}\right)_{k\in[m]}\right) \coloneqq \mathbb{E}_{(X,Y)\sim\mu}\big[(Y - \hat{y}(X))^2\big], \tag{1.4}$$

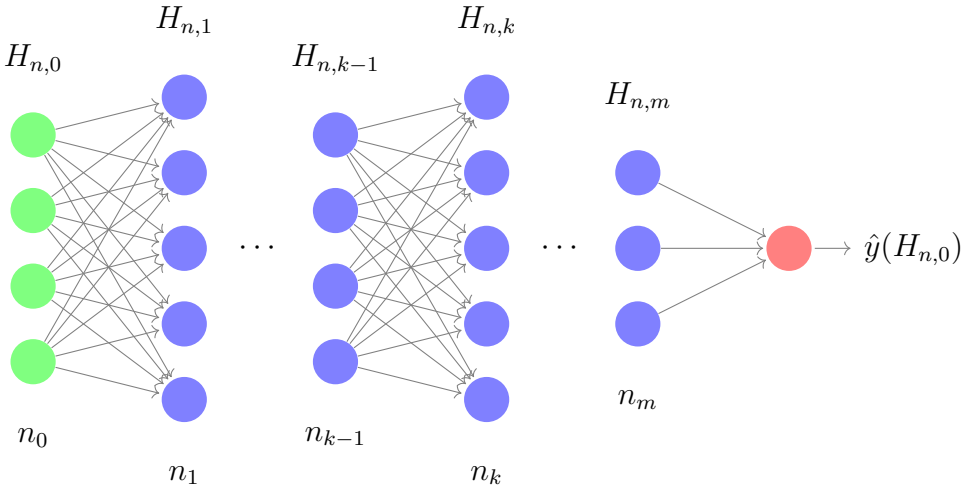


Figure 1.3: Finite width Neural Network with m hidden layers.

where the minimization is over all choices of the sequence of matrices $\theta_{n,k}^{(m)}$ and vectors $b_{n,k}^{(m)}$, for $k \in [m]$. The entries of the matrix $\theta_{n,k}^{(m)}$ can be thought of as associated with the edges of the bipartite graph connecting the nodes in layers $k-1$ and k. The output R_n in equation (1.4) does not depend on the labeling of the nodes in either layer, in the sense that if we relabel the nodes and correspondingly permute the rows and columns of $\theta_{n,k}^{(m)}$ and the vector $b_{n,k}^{(m)}$, the output R_n remains the same. Therefore the risk function R_n can be thought of as a function of edge weights of a sequence of bipartite graphs that is invariant under vertex relabeling.

One can ask the question: as the number of vertices in each layer goes to infinity, is there a scaling limit for the gradient flow of R_n? This is a multivariate generalization of the set-up described in Section 2.2.1, where instead of a single graph we have a sequence of m graphs, all bipartite, and successive graphs share vertices.

In Section 1.4.1, we will see how the permutation symmetry in DNNs can be used to take appropriate averages of two DNNs sharing the same architecture. In Section 1.4.2, we will take up a simple setup of a linear residual neural network, were we will fix a model for the random weight matrices, and ask if there can be a scaling limit description of the network as the depth and the width goes to infinity.

1.4.1 SGD and permutation symmetries in Deep Neural Networks (DNNs)

DNNs typically consist of a sequence of matrices that share row/column labels with their adjacent ones. Most modern DNNs possess permutation symmetries in their parametric representations. That is, their output is invariant under permutations applied to the rows/columns of the matrices appearing in DNN representation. The goal is to obtain the sequence of matrices that minimizes the risk function R_n for $n \in \mathbb{N}$. Authors in [AHS23] empirically study the effectiveness of SGD in optimizing the non-convex DNN risk functions R_n for large $n \in \mathbb{N}$. For simplicity, consider the special case when the DNN is parameterized through a single finite symmetric matrix and therefore does not involve shared labels. Let $(U_{n,k})_{k\in\mathbb{Z}_+}$ and $(V_{n,k})_{k\in\mathbb{Z}_+}$ be the SGD iterations, starting at two independent initializations, say, $U_{n,0} \neq V_{n,0}$. Authors in [AHS23] observe that $(U_{n,k})_{k\in\mathbb{Z}_+}$ and $(V_{n,k})_{k\in\mathbb{Z}_+}$ can be "aligned" by optimizing over the set of all permutations. That is, for every $k \in \mathbb{Z}_+$, they solve for

$$\pi_k^* \in \underset{\pi_k \in S_n}{\arg\min} \left\| U_{n,k} - V_{n,k}^{\pi_k} \right\|_{\mathrm{F}}^2,$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm, S_n is the set of all permutations of $[n]$, and $V_{n,k}^{\pi_k}$ is the matrix $V_{n,k}$ with rows and columns relabeled by the permutation $\pi_k \in S_n$. The authors observe an emergent property of SGD called "linear mode connectivity" (LMC) [FDRC20]. This property essentially says that R_n does not fluctuate a lot on $W_{n,k}(\lambda)$ for large $k \in \mathbb{Z}_+$, where

$$W_{n,k}(\lambda) = (1-\lambda)U_{n,k} + \lambda V_{n,k}^{\pi_k^*}, \qquad \lambda \in [0,1].$$

Further, they observe that $R_n(W_{n,k}(\lambda))$ approaches a constant uniformly on $\lambda \in [0,1]$ as n goes to infinity. Authors in [BSM+22] observe through experiments that for a fixed and large enough $k \in \mathbb{Z}_+ \setminus \{0\}$, the permutation π_k^*, has negative convexity gap

$$R_n\Big((1-\lambda)U_{n,0} + \lambda V_{n,0}^{\pi_k^*}\Big) - \Big[(1-\lambda)R_n(U_{n,0}) + \lambda R_n\Big(V_{n,0}^{\pi_k^*}\Big)\Big].$$

Following these empirical observations and the hypothesis made by the authors in [ESSN22], it makes sense to consider DNNs up to its permutation symmetries, and as

a consequence study limiting behaviors of stochastic optimization algorithms over the space of graphons.

1.4.2 Forward pass in Deep Linear Residual Networks

Consider the simple setup of a linear residual neural network. The feedforward computation of the network initialized under the Depth-μP scaling [YYZH24, BNL$^+$24] determines the scale of the network weight matrices, the scaling $n^{-1/2}m^{-1/2}$ at every layer, and the output scaling of n^{-1}, where $m \in \mathbb{N}$ is the number of hidden layers (depth) of the network, and $n \in \mathbb{N}$ is the number of neurons in each layer (width). Let $d \in \mathbb{N}$ be the input dimension. The network computes its output as:

$$H_{n,0} = Jx, \quad H_{n,k} = H_{n,k-1} + \frac{1}{\sqrt{nm}}\theta_{n,k}^{(m)} H_{n,k-1}, \quad k \in [m], \quad \hat{y}(x) = \frac{1}{n}\sum_{i=1}^{n} H_{n,m,i}. \tag{1.5}$$

Here, the matrix $J \in \mathbb{R}^{n\times d}$ is a fixed sampling matrix with (possibly random) entries of the order $\Theta(1)$.

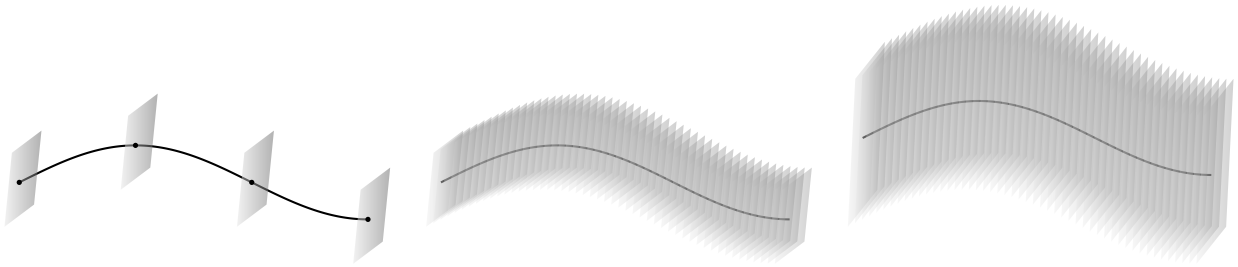
Since this is a residual networks, the identification of a neuron stays the same across all the layers. The trainable matrices $\Theta_n^{(m)} = \left(\theta_{n,k}^{(m)}\right)_{k\in[m]}$ for every $m \in \mathbb{N}$ are the $n \times n$ dimensional weight matrices corresponding to every layer. At the time of initialization, these matrices are set to have i.i.d. $N(0,1)$ entries, that is $\theta_{n,k}^{(m)} = G_{n,k}^{(m)}$, where the elements in $G_{n,k}^{(m)}$ are all i.i.d. standard Gaussian for every $k \in [m]$ and every $m \in \mathbb{N}$.

As the network is trained, the matrices $\Theta_n^{(m)}$ change to have a non-zero mean, which allows the network to learn via an empirical risk minimization process. A reasonable way through which these weight matrices $\Theta_n^{(m)}$ can be modeled is

$$\theta_{n,k}^{(m)} = \frac{1}{\sqrt{nm}} M_{n,k}^{(m)} + G_{n,k}^{(m)}, \qquad k \in [m],\ m \in \mathbb{N}, \tag{1.6}$$

where $\left(M_{n,k}^{(m)}\right)_{k\in[m]}$ are random bias matrices with means $\left(A_{n,k}^{(m)}\right)_{k\in[m]}$ respectively.

In Chapter 7, we will take up this model of deep networks, and provide a scaling limit description of the network's feedforward computation as the depth as well as the width goes to infinity (see Figure 1.4) for an illustration.



(a) Fixed depth, fixed width. (b) Infinite depth, fixed width. (c) Infinite depth, width $\to \infty$.

Figure 1.4: Each sub-figure shows the relative position of weight matrices along the depth interval $[0, 1]$. Matrix size corresponds to network width, and the number of matrices to network depth. The black curve indicates the underlying drift curve to which these sequences converge. See Chapter 7 for more details.

We will pose this problem as characterizing the limits of iterated product of matrices, where each matrix is a random and possibly biased perturbation of identity (see Section 6.3). This will allow us to characterize the evolution dynamics of the preactivations (for a given input x), as it gets transformed from the input to the output layer. We will show that as the depth and width of the network goes to infinity, any set of finitely chosen neuron, evolve as independent Gaussian processes along the network's depth, scaled appropriately.

1.5 Contributions and Dissertation Structure

This thesis addresses several classes of algorithms and dynamics over matrices that are commonly used and observed in machine learning and large scale optimization. It contains six chapters and a discussion.

Chapter 2 This chapter introduces the reader to the preliminaries necessary for understanding the subsequent chapters of the thesis. In Chapter 2.2, we describe in detail the classes of finite-dimensional matrix/graph-valued iterative algorithms introduced in Section 1.1. From Chapter 4 onwards, the iterates of these algorithms, as well as their scaling limits,

will be suitably embedded in various topological spaces. These spaces will be studied and used throughout the thesis to describe the scaling limits as curves. To this end, Chapter 2.3 discusses the preliminaries of the general theory of curves on metric spaces, adapted from [AGS08]. To incorporate the boundaries of the domain of the algorithms, we introduce a variation of SDE in equation (1.1) that includes boundary conditions using local time processes. In Chapter 2.5, we introduce the first topological space over which we will take our scaling limits with respect to the dimension of the iterate matrices (the second is introduced and discussed in detail in Chapter 5). Finally, in Chapter 2.6, we present the concept of exchangeability and exchangeable arrays, which lie at the heart of this dissertation and fundamentally allow the existence of such scaling limits.

Chapter 3 In this chapter, we consider each of the algorithm classes described in Chapter 2.2 and derive their continuous-time limits while keeping the dimension of the matrices fixed. These continuous-time limits are described via weak solutions of matrix-valued SDEs of the form given in equation (1.1). The SDEs contain both a drift and a diffusion term, which depend element-wise on the coordinates of the matrix as well as on the entire matrix. This interaction gives rise to the mean-field interaction between the coordinate processes of the matrix. In Chapter 6, we will extend the discussion to the dimension limits of the ambient space of these curves, which will be appropriately described as smooth curves over the metric spaces of graphons and measure-valued graphons.

Chapter 4 In Chapter 4, we begin by exploring how to take limits with respect to the dimension of the matrices. Since most optimization algorithms utilize the first-order information of the objective in some way, a naive guess would be to continue to expect some form of 'gradient' even when the dimension is taken to be infinity. Drawing motivation from the theory of Wasserstein calculus [Vil03, San15], we develop a theory of gradient flows on graphons, where we embed the matrix-valued iterates of the algorithms in the space of graphons. In this chapter, we demonstrate that Euclidean gradient flows on

finite-dimensional matrices suitably converge to gradient flows on the metric space of graphons as the dimension of the matrices tends to infinity.

Chapter 5 Despite already having a space over which we can take our infinite dimension limits, there is detail that is lost in the process due to the very nature of the space of graphons, which are only perfectly suited for adjacency matrices of simple unweighted dense graphs. In this chapter, we illustrate why the topological space of graphons is too coarse for capturing the microscopic description of matrix-valued processes when the coordinates of the matrices are real-valued. As a remedy, using the theory of exchangeability and infinite exchangeable arrays (IEAs) discussed in Chapter 2.6, we show that measure-valued graphons (MVGs) are the most suitable analytical objects to fully describe the scaling limit of our finite-dimensional exchangeable matrix-valued processes as the dimension goes to infinity. We provide two metrics on the topological space of MVGs, which help us quantify the convergence to the scaling limit in Chapter 6.

Chapter 6 In this chapter, we consider the finite-dimensional SDEs obtained in Chapter 3 as the continuous-time scaling limits of the iterative algorithms described in Chapter 2.2. We use the topology of graphons and measure-valued graphons (MVGs), or equivalently infinite exchangeable arrays (IEAs), to provide their respective scaling limits as the dimension of the ambient space of the processes goes to infinity. These scaling limits are characterized as McKean-Vlasov type SDEs over graphons and MVGs.

In Chapter 6.2.2, we demonstrate a connection between sampling and optimization by showing that a zeroth-order Monte Carlo-based sampling algorithm, despite not using any gradient information of any function, converges to the gradient flow of the objective Hamiltonian function on the space of graphons. This surprising and anticipated link between sampling and optimization is further analyzed, and we derive the non-asymptotic rates of this convergence.

In Chapter 6.3, we show that an appropriately scaled product of iterated matrices,

which are small perturbations of the identity map, converges to a process over IEAs as the dimension of the matrices tends to infinity. Such characterizations are useful in qualitatively analyzing various algorithms in the continuous-time setting. We apply these findings to deep learning (also see Chapter 7), deriving interesting characterizations and phenomena that emerge due to the scale of the problem.

Chapter 7 This chapter discusses an application where the developed machinery provides tangible conclusions for neural networks. In Chapter 7, we study the evolution of an input data point as it is transformed across the layers of a linear residual network using random weight matrices. The linearity of the layers reduces the asymptotic analysis to understanding the iterated products of random matrices discussed in Chapter 1.4.2.

We start by describing a scaling limit for this iterated product under the depth maximal update parameterization. For a fixed width (dimension of the weight matrices), using the results from Chapter 3.3, we show that this scaling limit is characterized by a non-commutative exponential of a matrix-valued stochastic process as the depth (number of time-scaled iterations) approaches infinity. We then extend our analysis to the limit where the width of the network (dimensionality of the matrices) also approaches infinity.

In this regime, any finitely chosen set of neuron evolutions along the network depth becomes independent. Passing to the infinite depth limit simplifies neuron evolution, leading to the phenomenon of propagation of chaos, where neurons behave independently as the network's width and depth grow infinitely large. Furthermore, the evolution of any randomly chosen neuron can be described by a Gaussian process, with the mean and variance explicitly related to the weights of the limiting network. This description allows us to formulate an explicit optimal control problem for risk minimization.

We conclude with a discussion of future research directions that build upon the contributions of this dissertation.

1.5.1 Authorship and Publication details

The body of the thesis is derived from a sequence of progressive research works, some of which are published and others under review. The work from [OPST23] forms Chapter 4 of this manuscript. The work from [HOP+22] forms Chapters 3.1 and Chapter 6.1 of this manuscript. The work from [APST23] forms Chapter 5, Chapter 3.2 and Chapter 6.2 of this manuscript. The work from [STH+24] forms Chapter 3.3, Chapter 6.3 and Chapter 7 of this manuscript.

These works were jointly co-authored with Raghavendra Tripathi, with shared responsibility for the problem formulation and theoretical contributions, under the guidance of the reading committee of this dissertation and external collaborators.

Chapter 2

BACKGROUND AND PRELIMINARIES

In this chapter, we will introduce the necessary background common to the following chapters. We start by introducing some notation in Section 2.1 that we will use throughout all chapters in this thesis. In Section 2.2, we elaborate on the different finite-dimensional algorithms introduced in Section 1.1. In Section 2.3, we will cover some definitions to set the stage for defining several kinds of curves on metric spaces that result from interpolating and taking the limits of the discrete-time iterative algorithms. In Section 2.4, we will introduce a variant of SDE (1.1) that accommodates boundary conditions whenever our algorithms are over closed cubic finite-dimensional domains. In Section 2.5, we will introduce a theory of limits of graphs/matrices and familiarize the reader with the space of *graphons*, which will be one of the limiting topological spaces that we will use to embed the iterates across varying dimensions. In Section 2.6, we will introduce the notion of exchangeability and relate it to graphons. In most of this dissertation, we will be working with symmetric matrices, and we will establish some notation to make this convenient.

2.1 Notation

This chapter establishes the notation used throughout this thesis for various mathematical constructs and concepts.

- For any index set X, the Cartesian product $X \times X$ is denoted by X^2. The set $X^{(2)}$ represents $X^2/\sim$, where the equivalence relation $\sim$ is defined through the relation $(a, b) \sim (b, a)$ for all $a, b \in X$. This notation is particularly useful in the context of domains of symmetric functions like two dimensional arrays.

- The symmetric group on the finite set $[n] \coloneqq \{1, 2, \ldots, n\}$ is denoted by S_n.
- The symbol ∇_e and ∂_e refers to the partial derivative with respect to the e-th variable.
- The set of all non-negative real numbers and non-negative integers is denoted by $\mathbb{R}_+$ and $\mathbb{Z}_+$ respectively.
- For any metric space (Ω, d), its its Borel sigma algebra is denoted by $\mathcal{B}(\Omega)$, and the set of all Borel probability measures is denoted by $\mathcal{P}(\Omega)$.
- For any $n \in \mathbb{N}$, the set of $n \times n$ real-valued symmetric matrices is denoted by $\mathcal{M}_n(\mathbb{R})$, the set of $[-1, 1]$-valued symmetric matrices is denoted by $\mathcal{M}_n$, and the set of $[0, 1]$-valued symmetric matrices is denoted by $\mathcal{M}_{n,+}$.
- The symbol $\circ$ is used both for the matrix Hadamard (elementwise) product as well as function composition with the meaning being clear from the context.

These notations and definitions will be consistently used throughout the thesis to ensure clarity and precision in the presentation of the mathematical concepts and results.

2.2 Finite dimensional iterative algorithms

In this section, we expand on each class of algorithms over finite dimensions introduced in Section 1.1. In Section 2.2.1, we discuss the projected (noisy) stochastic gradient descent algorithm. In Section 2.2.2, we discuss a relaxed Metropolis chain algorithm. In Section 2.2.3, we discuss the setup of iterated products of matrices that appear in various multiplicative algorithms.

2.2.1 First Order Stochastic Optimization Algorithms

In this section, we will discuss the class of stochastic optimization algorithms over finite-dimensional Euclidean spaces. To do this, we will first establish some notation. We will fix

$n \in \mathbb{N}$ throughout this entire introduction. However, since we will be scaling these dimensions later, starting from Chapter 4, we will maintain the subscripts and scaling factors of n throughout this introduction. The specific choices of scalings that will appear will become apparent when we discuss the limit spaces in which these finite-dimensional spaces will be embedded.

Our algorithms will operate on the set $\mathcal{M}_n$ of symmetric matrices with entries restricted in $[-1, 1]$. This is an arbitrary choice of a compact set in $\mathbb{R}$, and can be chosen to suit the application. We will be interested in optimizing a permutation invariant function $R_n \colon \mathcal{M}_n \to \mathbb{R}$ that satisfies certain differentiability properties. Let $\boldsymbol{\tau}_n \coloneqq (\tau_{n,k})_{k \in \mathbb{Z}_+}$ be a sequence of positive step-sizes (also known as learning rate). Here, $\mathbb{Z}_+$ denotes the set of all non-negative integers. Given the step-size sequence $\boldsymbol{\tau}_n$, we can define a monotonically increasing sequence of times $(t_{n,k})_{k \in \mathbb{Z}_+}$, defined as a cumulative sum of $\boldsymbol{\tau}_n$, i.e., $t_{n,0} = 0$ and $t_{n,k} \coloneqq \sum_{j=0}^{k-1} \tau_{n,j}$ for any $k \in \mathbb{N}$. We will only consider step size sequences $\boldsymbol{\tau}_n$ that have a diverging sum, i.e., $\lim_{k \to \infty} t_{n,k} = \infty$. We define the norm of the step size sequence $\boldsymbol{\tau}_n$ as $|\boldsymbol{\tau}_n| \coloneqq \sup_{k \in \mathbb{Z}_+} \tau_{n,k}$, which we assumed to be finite.

Since the updates of general iterative algorithms can move the iterates out of $\mathcal{M}_n$, we use a projection operator $P \colon \mathbb{R} \to [-1, 1]$ to project back the entries of the matrix iterates in $[-1, 1]$. This can be done by defining the map P as $P(x) = \arg\min_{z \in [-1,1]} |x - z|^2$ for $x \in \mathbb{R}$. The map P has an explicit form for $x \in \mathbb{R}$, as

$$P(x) \coloneqq \begin{cases} -1, & \text{if } x \in (-\infty, -1), \\ x, & \text{if } x \in [-1, +1], \\ +1, & \text{if } x \in (+1, +\infty). \end{cases}$$

The operator P can be used coordinatewise over matrices. We now describe some iterative optimization algorithms.

Definition 2.1 (Projected GD)**.** Let $n \in \mathbb{N}$ and let $R_n \colon \mathcal{M}_n \to \mathbb{R}$ be a differentiable function. The projected GD iterates of R_n starting at $X_{n,0} \in \mathcal{M}_n$ are defined to be a

sequence of symmetric matrices $(X_{n,k})_{k\in\mathbb{Z}_+}$ as

$$X_{n,k+1} = P\big(X_{n,k} - n^2\tau_{n,k}\nabla R_n(X_{n,k})\big), \qquad k \in \mathbb{Z}_+. \tag{PGD}$$

The reader shall note that there is a multiplicative factor of n^2 further associated with the step-size in all the first order algorithms. This important choice will ensure that the curves across every dimension $n \in \mathbb{N}$ can be embedded appropriately on the limiting space of graphons. We will discuss this in Section 2.5.2 and later in Chapter 4.

In most practical machine learning applications, computing gradients of risk functions can be computationally intensive. Hence, in practice, stochastic approximation algorithms based on projected Stochastic Gradient Descent (SGD) are instead used to minimize such functions since they are often faster to simulate [RM51, KW52]. The details of this common Markov chain are described later in the section, and the reader can refer to the monographs [Ben99, KY03, Bor09, MB11, KC12] for a detailed overview. Roughly, if the current state is a symmetric matrix X_n, one jumps to a new state by taking a small step along the negative Euclidean gradient $-\nabla R_n(X_n)$, and potentially adding independent, centered, and variance-bounded noise to each matrix entry (up to matrix symmetry). Each matrix entry is then projected on to the interval $[-1, 1]$ to satisfy the entrywise constraint. These stochastic algorithms are usually variants of deterministic algorithms, in which the iteration updates are replaced by their stochastic estimates (with randomness generally derived from the training data). More formally, let $(\xi_{k+1})_{k\in\mathbb{Z}_+}$ be an i.i.d. sequence of random variables with some distribution $\mathcal{D}$ over some arbitrary measurable space $(\Omega, \mathcal{A})$.

Let $g_n \colon \mathcal{M}_n \times \Omega$ be a function such that

$$\nabla R_n = \mathbb{E}_{\xi\sim\mathcal{D}}[g_n(\,\cdot\,;\xi)].$$

We now introduce the stochastic analogues of the Projected GD iteration scheme (definition 2.1). Further, we will consider two ways to introduce noise at each iteration.

1. **Small noise**: We can replace the Euclidean derivative ∇R_n in equation (PGD) by its unbiased stochastic proxy random variable $g_n(\,\cdot\,;\xi_{k+1})$.

2. **Large noise**: We can add an additive scaled noise to the iterates in equation (PGD) before applying the projection operation, as we describe in Definition 2.2 below.

The small noise appears in algorithms like Stochastic Gradient Descent (SGD) and its variants, and the large noise appears when there is additional noise added on top of these algorithms for reasons that include escaping saddle points that is crucial when the objective function is non-convex. We will use the operator $\circ$ over symmetric matrices to denote the Hadamard (elementwise) product.

Definition 2.2 (Projected SGD with and without noise)**.** Let Σ_n be a map from $\mathcal{M}_n$ to $n \times n$ symmetric matrices with non-negative entries. The projected noisy SGD iterates starting from $X_{n,0} \in \mathcal{M}_n$ is defined to be a sequence of symmetric matrices $(X_{n,k})_{k\in\mathbb{Z}_+}$ given iteratively as

$$X_{n,k+1} = P\Big(X_{n,k} - n^2\tau_{n,k}g_n(X_{n,k};\xi_{k+1}) + \tau_{n,k}^{1/2}G_{n,k}\Big), \qquad k \in \mathbb{Z}_+, \tag{PNSGD}$$

where $(G_{n,k})_{k\in\mathbb{Z}_+}$ is an $n \times n$ symmetric matrix-valued martingale difference sequence independent of $(\xi_{k+1})_{k\in\mathbb{Z}_+}$. We also consider the noise $G_{n,k}$ for $k \in \mathbb{Z}_+$, of the form $G_{n,k} = \Sigma_n(X_{n,k}) \circ Z_{n,k}$ where $(Z_{n,k})_{k\in\mathbb{Z}_+}$ is a sequence of independent $n \times n$ symmetric random matrices with standard normal entries (up to matrix symmetry).

In practice it makes sense to generalize the large noise structure to consider non-diagonal diffusion coefficient functions as an approximation to SGD [LTE19]. Notice that if we were to use a non-diagonal diffusion coefficient, we would need to scale down matrix-matrix product by some factor of n and account for the non-trivial correlations arising due to it. Such correlations, as we will see, have been studied for the class of dynamics falling in case 3 discussed earlier in this chapter. For the purposes of optimization algorithms, we will stick to diagonal diffusion coefficient functions.

In Chapter 3, we will show that as we take the step-size to zero, the process of matrices thus obtained converges to the solution of the SDE in equation (1.1). Later, in Chapter 6, we will demonstrate that these continuous-time, finite-dimensional, matrix-valued processes

converge to a well-defined curve on the space of measure-valued graphons, which can be described using a McKean-Vlasov SDE. Furthermore, upon projection to the space of graphons, this curve, obtained for SGD, converges to a gradient flow on graphons (developed in Chapter 4).

2.2.2 Zeroth Order Optimization and Sampling

Let $H_n \colon \mathbb{R}^n \to \mathbb{R}$ be a measurable function that is differentiable. A natural stochastic process is the Langevin diffusion:

$$\mathrm{d}X_n(t) = -\nabla H_n(X_n(t))\,\mathrm{d}t + \sqrt{\frac{2}{\beta}}\,\mathrm{d}B_n(t),$$

where B_n is standard n-dimensional Brownian motion and $\beta > 0$ is commonly called the inverse temperature parameter. In practice, SGD based algorithms are used to mimic the paths of the Langevin diffusion in discrete time. As $\beta \to \infty$, the paths of the Langevin diffusion converge to that of the gradient flow of H_n, namely $\dot{X}_n(t) = -\nabla H_n(X_n(t))$, which in a sense gives the fastest decay of the Hamiltonian. On the other hand, on discrete spaces or when the gradient of the Hamiltonian is not well defined, one employs a MCMC algorithm [Dia09], such as the celebrated Metropolis algorithm [Ric10, Section 2.4], to sample from the Gibbs distribution.

In this section, we will introduce a Monte Carlo Markov Chain (MCMC) algorithm called *relaxed Metropolis chain* algorithm on a stochastic block model (SBM). SBMs are a widely used family of models of random graphs (see [HLL83, Ver18]). The base Markov chain runs on a SBM with n communities, with r members in each community, with an acceptance-rejection step specified by the permutation invariant Hamiltonian function $H_n \colon \mathcal{M}_{n,+} \to \mathbb{R}$, and the inverse temperature parameter β. Our algorithm includes a novel relaxation procedure after each accept-reject step which introduces a further positive parameter σ. Let us first define the Empirical Stochastic Block Model (ESBM).

Definition 2.3 (Empirical Stochastic Block Model (ESBM))**.** For $n, r \in \mathbb{N}$ let $q \equiv$

$(q_{i,j})_{(i,j)\in[n]^2} \in \mathcal{M}_{n,+}$, and let $N = nr$. A random simple graph with N vertices is called ESBM$[n, r, q]$ if

- for $i \in [n]$, there are r many vertices having color i,
- for $i, j \in [n]$, $i \neq j$, $r^2 q_{i,j}$ many edges (unordered pairs of vertices $\{u, v\}$) are drawn by randomly sampling without replacement where one vertex has color i and the other has color j,
- for $i \in [n]$, $\binom{r}{2} q_{i,i}$ many edges are drawn by randomly sampling without replacement unordered pairs of vertices of color i, and
- the samplings in the last two items are done independently for all pairs $(i, j) \in [n]^{(2)}$.

To construct the Gibbs probability measure on $\mathcal{M}_{n,+}$, we will be interested in ESBM$[n, r, q]$ random graphs where the entries of q are also random. For each $r \in \mathbb{N}$, consider the uniform distribution μ_r on the discrete set $\{i/r^2 \mid i \in \{0\} \cup [r^2]\}$ and ν_r the uniform distribution on the discrete set $\left\{i/\binom{r}{2} \;\middle|\; i \in \{0\} \cup \left[\binom{r}{2}\right]\right\}$. Define $U_{r,n}$ to be the probability measure on $\mathcal{M}_{n,+}$ where each entry above the diagonal is independently distributed as μ_r and the diagonal entries are independently distributed as ν_r. Thus $U_{r,n}$ can be viewed as a discrete uniform distribution on the set of possible edge-densities.

Recall that H_n is the Hamiltonian function on the space of $n \times n$ symmetric matrices. Fix a positive sequence $(\gamma_r)_{r\in\mathbb{N}}$ such that

$$\lim_{r\to\infty} \gamma_r \log^2 r = 0, \quad \text{and} \quad \lim_{r\to\infty} \frac{\gamma_r r^2}{\log r} = \infty. \tag{2.1}$$

The importance of these conditions will become clear when we derive the algorithm's scaling limit in Section 3.2.

Fix $\beta > 0$ and let $\beta_{r,n} \coloneqq \beta n^{-2}/\gamma_r$. Consider a family of Gibbs probability measures on $\mathcal{M}_{n,+}$ given by

$$\widehat{Q}_{r,n,\beta}(\mathrm{d}q) = \frac{1}{Z_{r,n,\beta}} \mathrm{e}^{-\beta_{r,n} H_n(q)} U_{r,n}(\mathrm{d}q) = \frac{1}{Z_{r,n,\beta}} \mathrm{e}^{-\beta\gamma_r^{-1} n^{-2} H_n(q)} U_{r,n}(\mathrm{d}q), \tag{2.2}$$

where $Z_{r,n,\beta}$ is the normalizing constant. As each $q \in \mathcal{M}_{n,+}$ corresponds to a simple random graph in ESBM$[n, r, q]$, $\widehat{Q}_{r,n,\beta}$ can be thought of as a random probability distribution on simple graphs over nr vertices. We will denote the model specified by $\widehat{Q}_{r,n,\beta}$, as ESBM$[n, r, \beta, \mathcal{H}]$. It should be emphasized that the measure $\widehat{Q}_{r,n,\beta}$ depends on the choice of the parameter γ_r. Note that the above model closely resembles commonly used framework in exponential random graphs (see [Cha17] and references therein).

The following Metropolis chain algorithm (see [LP17, Section 3.2]) can be used to sample from ESBM$[n, r, \beta]$.

- *Base Markov Chain*: The state space of the chain is the set $\mathcal{S}_{r,n}$ of all simple graphs on nr vertices with n colors assigned to equal number of vertices. The base chain starts at an arbitrary graph $G(0) = G$ in the state space. Suppose, for $\ell \geq 0$, the Markov chain has completed ℓ steps, $\{G(p)\}_{p=0}^{\ell}$, and is at graph $G(\ell)$. For $(i, j) \in [n]^{(2)}$, let $m_{i,j}(\ell)$ denote the number of edges between vertices of color $i \in [n]$ and color $j \in [n]$ in $G(\ell)$. The next step in the Markov chain is generated as follows.

  - For every $(i, j) \in [n]^{(2)}$, $i \neq j$, if $m_{i,j}(\ell) \notin \{0, r^2\}$, then, toss a fair coin. If the coin comes up heads, then delete an edge between color i and color j, chosen at random, and if the coin turns up tails, place an additional edge between color i and color j at random. Replace r^2 by $\binom{r}{2}$ if $i = j$.
  - For every $(i, j) \in [n]^{(2)}$, if $m_{i,j}(\ell) = 0$, then toss a fair coin. If the coin comes up heads, then add an additional edge, chosen at random, and if the coin turns up tails, do nothing. Similarly, if $m_{i,j}(\ell) = r^2$, $i \neq j$ (or $\binom{r}{2}$, if $i = j$), then toss a fair coin. If the coin turns up heads then delete an existing edge, chosen at random, otherwise do nothing.
  - Do these independently for every pair $(i, j) \in [n]^{(2)}$.

  The resulting graph is $G(\ell + 1)$ and $q(\ell + 1) = (q_{i,j}(\ell + 1))_{(i,j)\in[n]^2}$ be its edge density matrix. It is not hard to see that the base chain viewed as a process on edge densities

is also a Markov chain that is reversible with respect to the uniform distribution $U_{r,n}$.

- *Metropolis Chain*: We run the base chain for $s_r \approx \gamma_r^2 r^4$ many steps followed by an accept-reject step. Suppose we started the base chain at graph G and edge density matrix q. After running the base chain for s_r many steps we arrive at a graph G' and a corresponding edge density matrix q'.

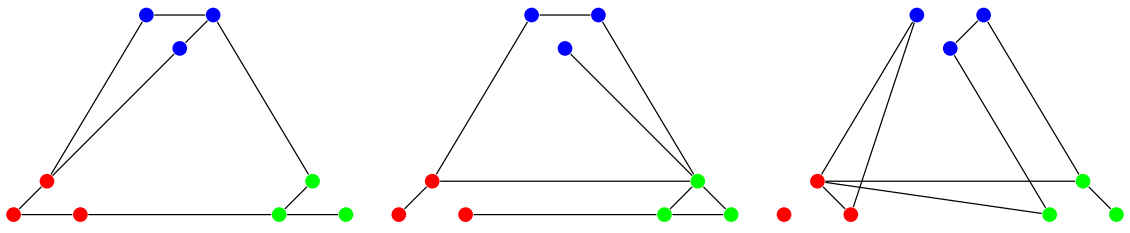
  - Accept-reject step: Accept G' as the next state of the Metropolis chain with probability $\exp\left(-\beta_{r,n}\left(H_n(q') - H_n(q)\right)^+\right)$, otherwise, remain at G. Here $x^+ \coloneqq \max\{x, 0\}$.

It is standard to see the unique invariant distribution of this Metropolis Markov chain is the Gibbs measure $\widehat{Q}_{r,n,\beta}$. We will explore scaling limits of the chain as $r, n \to \infty$, $\gamma_r, \beta_{r,n}$ as specified above and when $s_r = O(\gamma_r^2 r^4)$. But, first, we introduce an additional relaxation step.

- *Relaxed Metropolis Chain*: After every Metropolis accept-reject step, we run the base chain for an additional $\ell_{r,n}(\sigma) = O(\sigma^2 n^{-4} \gamma_r r^4)$ many steps, for some $\sigma > 0$, and always accept the last state.

We illustrate the relaxed Metropolis chain algorithm in Figure 2.1. Starting with a community density matrix q corresponding to the SBM shown in Figure 2.1a, s_r edges are independently flipped across each pair of communities to obtain q' (for illustration, say $s_r = 1$). The resulting graph is accepted with a probability depending on $H_n(q') - H_n(q)$, yielding an SBM as shown in Figure 2.1b. Finally, in the relaxation step, we independently flip $\ell_{r,n}(\sigma)$ edges across each pair of communities to obtain the SBM shown in Figure 2.1c (for illustration, say $\ell_{r,n}(\sigma) = 2$).

Thus, our final Markov chain repeatedly runs the base chain for s_r many steps, performs an accept-reject step and then runs another $\ell_{r,n}(\sigma)$ many steps of the base chain. We call this the *relaxed Metropolis chain*. Since $\ell_{r,n}(0) = 0$, when $\sigma = 0$, we recover the true Metropolis



(a) (b) (c)

Figure 2.1: Relaxed Metropolis chain iteration for $r = 3, n = 3$.

chain. However, note that the relaxed chain has a different invariant distribution for any positive σ.

Remark 2.4. The reader can make a small observation that the base chains when run for s_r and $\ell_{r,n}(\sigma)$ many steps, as we shall see later, shall correspond to contributing the small and large noise as also present in the (PNSGD) algorithm (see Definition 2.2 in Section 2.2.1).

We will analyze this algorithm in detail in Section 3.2 and show that as $r \to \infty$, the matrix iterates q, suitably time scaled, converge to an SDE of the form described in equation (1.1). In Chapter 6 we will show that as $n \to \infty$, the matrix-valued process described by the SDE, converges to a curve on measure-valued graphons described by a McKean-Vlasov equation. Moreover, we will see that upon projection to the space of graphons, this curve is nothing but a gradient flow on the space of graphons, showing a remarkable connection between sampling (i.e., gradient-free) and optimization.

2.2.3 Iterative product of matrices

For any fixed $n \in \mathbb{N}$, let $H_{n,0} \in \mathbb{R}^n$ be an initial state vector of a system with n interacting particles. Here, we will look at a special kind of mean-field interaction where the updates are of the multiplicative form. Iterative algorithms belonging to the class of multiplicative updates transform the initial state $H_{n,0}$ to obtain the final state $H_{n,m}$ of the system using

$m \in \mathbb{N}$ successive linear transformations. Each linear transformation is a random biased perturbation of the identity, scaled suitably as a function of m.

Given any state vector $H_{n,k}$ for $k \in [m-1] \cup \{0\}$, the next state $H_{n,k+1}$ of the system is obtained by applying the linear transformation corresponding to the matrix $I_n + X^{(m)}_{n,k+1}$. Here, $X^{(m)}_{n,k+1} \in \mathbb{R}^{n \times n}$ is the perturbation matrix that can possibly be random with a non-zero mean. As $m \to \infty$, the matrices $\left(X^{(m)}_{n,k}\right)_{k \in [m]}$ individually shrink down to zero. However, one can expect the net effect of all the linear transformations to be independent of m as it grows large. This phenomenon is attributed to the cumulative effect of the successive linear transformations, which compensates for the diminishing perturbations, thereby maintaining the overall transformation's magnitude irrespective of the increasing iterations.

Similar to the noise model in Section 2.2.1, we will consider two kinds of noise to model the perturbation—small noise and large noise. Let

$$X^{(m)}_{n,k+1} \coloneqq \frac{\mu_n}{m} M^{(m)}_{n,k+1} + \frac{\sigma_n}{\sqrt{m}} G^{(m)}_{n,k+1}, \qquad k \in [m], \tag{2.3}$$

such that $\mathbb{E}\left[M^{(m)}_{n,k+1}\right] = A^{(m)}_{n,k+1} \in \mathbb{R}^{[n]^2}$ contributes a drift, and $G^{(m)}_{n,k+1}$ is an independent matrix with each entry i.i.d. as $N(0,1)$, which contributes to a diffusion. The scales m^{-1} and $m^{-1/2}$, of the matrices $M^{(m)}_{n,k+1}$ and $G^{(m)}_{n,k+1}$ ensure that as we take $m \to \infty$, they contribute to the small and large noise effects in the update of the state of the system. Since n is fixed, we can consider the parameters μ_n and σ_n to be any real constants. Later in Chapter 6.3 (see Section 6.3.1), we will scale μ_n and σ_n with respect to n as we consider the limit as $n \to \infty$.

With this setup, the effective linear transformation acting upon $H_{n,0}$ at iteration $k \in [m]$ is

$$P^{(m)}_n(k) \coloneqq \left(I_n + X^{(m)}_{n,k}\right) \ldots \left(I_n + X^{(m)}_{n,2}\right)\left(I_n + X^{(m)}_{n,1}\right),$$

with $P^{(m)}_n(0) \coloneqq I_n$. Since we apply m updates, we can rescale the iterates to consider piecewise constant interpolations $P_{n,m}$ of $\left(P^{(m)}_n(k)\right)_{k=0}^{m}$ defined as $P_{n,m}(t) \coloneqq P^{(m)}_n(\lfloor mt \rfloor)$ for $t \in [0,1]$. In Chapter 3.3, we will show that as $m \to \infty$, the curve of linear operators $P_{n,m}$ has an appropriate limit.

We remark that in special cases, the matrices $\left(X_{n,k}^{(m)}\right)_{k\in[m]}$ for every $m \in \mathbb{N}$ can be considered as the prefix subsequence of a common infinite sequence of matrices. However, we adhere to the triangular array setup for reasons of generality and to capture various applications of interest (see Chapter 3.3 and Chapter 7 for a discussion).

With the above-discussed setup, we find that in order to understand the evolution of an initial state as both $m, n \to \infty$, it is crucial to understand the scaling limit of the product of random matrices of the form described. Beginning with the seminal work of Bellman [Bel54], Furstenberg and Kesten [FK60], and Berger [Ber84], there has been significant work on the law of large numbers and CLT-type results for (the entries of) the product of random matrices and operators [Fur63, Tut65, Wat84, Joh94, TN13, DMN$^+$21]. Concentration inequalities for the product of random matrices have also been extensively studied [EH18, Bag20, KMS20, HW20, CHKT21]. Iterated products of matrices have been studied in the context of random walks on groups [Led01, Fur02, BQ16]. In Chapter 3.3, we will study the continuous-time scaling limit of the matrix product by taking $m \to \infty$ and obtain a continuous-time process on $n \times n$ matrices as they converge to infinite exchangeable arrays (IEAs) (see Section 2.6). In Chapter 6.3, we will study the evolution of the coordinates of the matrix-valued process as $n \to \infty$. As a consequence, we will explore a beautiful application of this analysis in the context of deep learning in Chapter 7.

2.3 Curves on metric spaces

To obtain continuous time scaling limits of the iterative schemes as for the ones defined in Definition 2.1 and Definition 2.2, we will use piecewise constant interpolations of the iterates.

Definition 2.5 (Piecewise constant interpolation)**.** Given a sequence $(a_k)_{k\in\mathbb{Z}_+}$ over any domain, and a sequence of positive step sizes $\boldsymbol{\tau} = (\tau_k)_{k\in\mathbb{Z}_+}$, the *piecewise constant interpolation* of $(a_k)_{k\in\mathbb{Z}_+}$ as a right-continuous curve $a\colon \mathbb{R}_+ \to \{a_k\}_{k\in\mathbb{Z}_+}$ defined as

$$a(t) \coloneqq a_k, \qquad \text{if} \quad t \in [t_k, t_{k+1}),$$

for some $k \in \mathbb{Z}_+$, where $t_0 = 0$ and $t_k \coloneqq \sum_{j=0}^{k-1} \tau_j$ for any $k \in \mathbb{N}$.

Throughout this dissertation, we will encounter curves that will be absolutely continuous with respect to a certain metric. We state the definition for the sake of completeness.

Definition 2.6. Let (X, d) be a metric space, then a curve $\omega\colon \mathbb{R}_+ \to (X, d)$ is *absolutely continuous* with respect to d if there exists $m \in L^1(\mathbb{R}_+)$ such that for all $0 \le r < s < \infty$,

$$d(\omega(r), \omega(s)) \le \int_r^s m(t)\,\mathrm{d}t.$$

We denote the set of all absolutely continuous curves on (X, d) by $\mathrm{AC}(X, d)$.

2.4 System of reflected diffusions

To cater to the continuous time limits of the algorithms with projections as described in Section 2.2.1, we will need to introduce the notion of local time processes.

For $n \in \mathbb{N}$, consider the domain $\mathcal{M}_n = [-1, 1]^{[n]^2}$ (and $\mathcal{M}_{n,+} = [0, 1]^{[n]^2}$). Notice that $\mathcal{M}_n$ (and $\mathcal{M}_{n,+}$) is a cube, and is closed with respect to the usual topology. Consider the SDE:

$$\begin{aligned}\mathrm{d}X_{n,e}(t) = b_{n,e}(X_{n,e}(t), X_n(t))\,\mathrm{d}t + \Sigma_{n,e}(X_{n,e}(t), X_n(t)) \circ \mathrm{d}B_{n,e}(t) \\ + \mathrm{d}L^-_{n,e}(t) - \mathrm{d}L^+_{n,e}(t),\end{aligned} \tag{2.4}$$

for $t \in [0, T]$ for some fixed $T \in \mathbb{R}_+$ and starting at $X_n(0) = X_{n,0} \in \mathcal{M}_n$ (or $\mathcal{M}_{n,+}$). Here Σ_n is a map from $[-1, 1] \times \mathcal{M}_n$ (or from $[0, 1] \times \mathcal{M}_{n,+}$) to the set of $n \times n$ symmetric matrices with non-negative entries, B_n is a $n \times n$ symmetric matrix valued process containing a set of standard Brownian motions $(B_{n,e})_{e\in[n]^2}$, and the processes L^-_n and L^+_n are local times at the boundary of $\mathcal{M}_n$. More precisely, they satisfying the following conditions:

1. The processes X_n, L^+_n and L^-_n are adapted processes.

2. The process L^-_n and L^+_n are coordinatewise non decreasing processes a.e.

3. For every $e \in [n]^2$,

$$\begin{aligned}\int_0^\infty \mathbb{1}\{X_{n,e}(t) > -1\}\,\mathrm{d}L^-_{n,e}(t) = 0, \qquad \text{and} \\ \int_0^\infty \mathbb{1}\{X_{n,e}(t) < +1\}\,\mathrm{d}L^+_{n,e}(t) = 0.\end{aligned} \tag{2.5}$$

The first condition in equation (2.5) changes to the condition $\int_0^\infty \mathbb{1}\{X_{n,e}(t) > 0\}\, \mathrm{d}L^-_{n,e}(t) = 0$ if we consider the set $\mathcal{M}_{n,+}$. We say that (X_n, L^+_n, L^-_n) solves the Skorokhod problem with respect to the set $\mathcal{M}_n$. Following [KLRS07, Definition 1.2], the strong solution (X_n, L^+_n, L^-_n) of the Skorokhod problem exists and is unique if b_n and Σ_n are Lipschitz with respect to $\|\cdot\|_{\mathrm{F}}$. We will denote the Skorokhod map as $\mathrm{Sko}(\,\cdot\,)$.

The Lipschitz property of the Skorokhod map

Let Y_1 and Y_2 be two real valued stochastic processes. Let $\Lambda_{[-1,1]}$ denote the Skorokhod map that maps the set of càdlàg functions on $[0, T]$ to itself. If $(X_1 \coloneqq \Lambda_{[-1,1]}(Y_1), L^+_1, L^-_1)$ and $(X_2 \coloneqq \Lambda_{[-1,1]}(Y_2), L^+_2, L^-_2)$ solve the Skorokhod problem with respect to the set $[-1, 1]$, then the Skorokhod map $\Lambda_{[-1,1]}$ is 4-Lipschitz under the uniform metric [KLRS07, Corollary 1.6], i.e.,

$$\sup_{t\in[0,T]} |X_1(t) - X_2(t)| \leq 4 \sup_{t\in[0,T]} |Y_1(t) - Y_2(t)|, \qquad \forall\ T \in \mathbb{R}_+. \tag{2.6}$$

2.5 Limits of large graphs and Graphons

In this dissertation, we will study various limits of finite dimensional matrices. Since we will be working with permutation invariant functions of matrices, we can as well consider these function to be functions of 'unlabeled' matrices, that is, one would identify two matrices to be the same if their rows and columns are permuted by a common permutation. This identification allows us to think of such unlabeled matrices as adjacency matrix of unlabeled graphs. Because we will be working with matrices of different dimensionalities, it is important to make sense of how can two matrices of possibly different dimensions be compared. Due to this relation between matrices and graphs, the question rather boils down to understanding the notion of convergence of unlabeled graphs [LS06, BCL$^+$08, BCL$^+$12]. For simplicity, we will first work with unweighted simple graphs (i.e., graphs with edge weights in $\{0, 1\}$). A simple graph G is a graph without loops or multiple edges. We denote the vertex and edge set of G by $V(G)$ and $E(G)$, respectively.

Definition 2.7 (Convergence of unweighted simple graphs)**.** Let $(G_n)_{n\in\mathbb{N}}$ be a sequence of unweighted graphs such that $\lim_{n\to\infty}|V(G_n)| = \infty$. For $n \in \mathbb{N}$ and $k \in [|V(G_n)|]$, let $G_n[k]$ denote the random induced subgraph of G_n with k vertices with vertex set $[k]$. The sequence $(G_n)_{n\in\mathbb{N}}$ converges if for every $k \in \mathbb{N}$ and every labeled graph G with k vertices with vertex set $[k]$, the limit $\lim_{n\to\infty}\mathbb{P}\{G_n[k] = G\}$ exists.

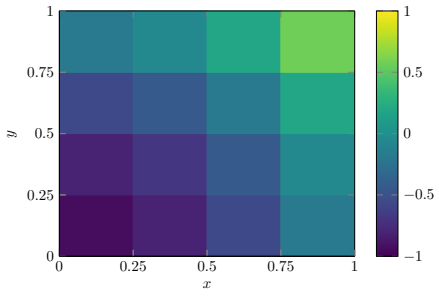
Definition 2.7 describes when a sequence of unlabeled unweighted graphs converges, but it does not characterize the limit objects of such graphs. In order to complete the space of unweighted graphs, we will need to define what is *kernel* is. A kernel is a Borel measurable function $w\colon [0,1]^{(2)} \to [-1,1]$ that is symmetric, i.e., $w(x,y) = w(y,x)$ for a.e. $(x,y) \in [0,1]^{(2)}$. Two kernels are identified if they are equal Lebesgue a.e. We will denote the set of all kernels by $\mathcal{W}$. More generally, for a bounded interval $I \subset \mathbb{R}$, let $\mathcal{W}_I$ be the set of all functions $w \in \mathcal{W}$ with $w(x,y) \in I$. Given a function $w \in \mathcal{W}$, we can think of the interval $[0,1]$ as the set of nodes, and of the value $w(x,y)$ as the weight of the edge $(x,y) \in [0,1]^{(2)}$.

Definition 2.8 (Kernels and finite dimensional matrices)**.** Any finite dimensional matrix can be embedded in $\mathcal{W}$. Let $Q_n \coloneqq \{Q_{n,i}\}_{i\in[n]}$ be defined as $Q_{n,1} = \left[0, \frac{1}{n}\right]$, $Q_{n,2} = \left(\frac{1}{n}, \frac{2}{n}\right]$, $\ldots$, and $Q_{n,n} = \left(\frac{n-1}{n}, 1\right]$. Given any matrix $A_n \in \mathcal{M}_n$, set $w_{A_n}(x,y) = A_n(v(x), v(y))$, where $v(x) = i$ whenever $x \in Q_{n,i}$ for some $i \in [n]$. Informally, we replace each entry (i,j) by a square of size $\frac{1}{n} \times \frac{1}{n}$ with the constant function $A_{n,(i,j)}$ on this square. It follows from our discussion that, for any $n \in \mathbb{N}$, the set $\mathcal{M}_n$ can be naturally identified with a subset of finite dimensional kernels, $\mathcal{W}_n \subset \mathcal{W}$. This identification/embedding will be denoted by K (as in $K(A_n)$ is the kernel corresponding to the matrix A_n) and its inverse will be denoted by M_n (as in matrix). See Figure 2.2 for an example.

If two graphs (or matrices) are isomorphic, the two may be identified by permuting the labels on the vertices. This may be done for kernels by the following equivalence relation. A map $\varphi\colon [0,1] \to [0,1]$ is *Lebesgue measure-preserving*, if it is measurable and push-forwards the Lebesgue measure to itself, i.e. $|\varphi^{-1}(A)| = |A|$ for all Borel measurable sets $A \subseteq [0,1]$. Throughout this dissertation, we will refer to a Lebesgue measure-preserving map simply as

$$\frac{1}{16}\begin{bmatrix} -16 & -15 & -12 & -7 \\ -15 & -14 & -11 & 1 \\ -12 & -11 & -6 & 4 \\ -7 & 1 & 4 & 9 \end{bmatrix}$$

(a) Symmetric matrix A_4.



(b) Kernel representation of A_4.

Figure 2.2: Kernels and finite dimensional matrices.

measure-preserving map and denote the set of all such maps by $\mathcal{T}$. We will denote the set of all invertible measure-preserving maps by $\mathcal{I}$. For $w \in \mathcal{W}$ and $\varphi\colon\ [0,1] \to [0,1]$, we define w^{φ} by $w^{\varphi}(x,y) = w(\varphi(x),\varphi(y))$ when $(x,y) \in [0,1]^{(2)}$. So one defines an equivalence relation $\cong$ on $\mathcal{W}$ such that $w_1 \cong w_2$ if there exist measure preserving transformations $\varphi_1, \varphi_2\colon\ [0,1] \to [0,1]$ and $w \in \mathcal{W}$ such that $w_1 = w^{\varphi_1}$, and $w_2 = w^{\varphi_2}$. We will call $\widehat{\mathcal{W}} := \mathcal{W}/\cong$ as the space of graphons and we will refer to any element as a graphon. Wherever it is clear from the context, for any kernel $w \in \mathcal{W}$, we will use an abuse of notation and use the same symbol w to denote the equivalence class, or the graphon, corresponding to the kernel. Wherever it shall be important to consider the equivalence set, we will denote by $[w] \in \widehat{\mathcal{W}}$, the equivalence set of the kernel w.

Given a graphon $[w] \in \widehat{\mathcal{W}}$ and $k \in \mathbb{N}$, we can obtain a $\{0,1\}$-valued $k \times k$ exchangeable symmetric array as we will describe in Definition 2.9.

Definition 2.9 (Sampling random unweighted simple graphs from graphons)**.** Given a graphon $[w] \in \widehat{\mathcal{W}}$, one can sample a random graph $G_k[w]$ with $k \in \mathbb{N}$ vertices by sampling k i.i.d. Uni$[0,1]$ random variables $(U_i)_{i\in[k]}$, and assigning edge weights $w(U_i, U_j)$ for all $(i,j) \in [k]^{(2)}$. It is important to note that the graph $G_k[w]$ thus obtained, has been provided a vertex labeling that depends on the representative kernel in $w^{\varphi} \in [w]$ for some $\varphi \in \mathcal{T}$.

The convergence of unweighted graph sequences can also be characterized in terms of homomorphism densities with respect to simple finite graphs. Let H and G be two simple graphs with edge-weights 1. We define $\hom_H(G)$ as the number of homomorphisms from H to G, i.e., the number of adjacency preserving maps $V(H) \to V(G)$, and the homomorphism density of H in G as

$$T_H(G) = \frac{1}{|V(G)|^{|V(H)|}} \hom_H(G).$$

Following Definition 2.7 we see that a sequence of unweighted finite simple graphs $(G_n)_{n\in\mathbb{N}}$ converge if $(T_H(G_n))_{n\in\mathbb{N}}$ has a limit in $[0,1]$ for every finite simple graph H.

The homomorphism density of a simple graph $H = ([k], E)$ for $k \in \mathbb{N}$, in a kernel $w \in \mathcal{W}$ is defined as

$$T_H(w) \coloneqq \int_{[0,1]^k} \prod_{\{i,j\}\in E} w(x_i, x_j) \prod_{v\in[k]} \mathrm{d}x_v. \tag{2.7}$$

For instance, if H is the triangle graph, then $T_\triangle(w) = \int_{[0,1]^3} w(x_1,x_2)w(x_2,x_3)w(x_3,x_1)\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_2$. The sequence of unweighted simple graphs $(G_n)_{n\in\mathbb{N}}$ with adjacency matrices $(X_n)_{n\in\mathbb{N}}$ converges to the graph limit w if $\lim_{n\to\infty} T_H(K(X_n)) = T_H(w)$ for every finite simple graph H. We will abuse notation and use $T_H(G_n)$ with $T_H(X_n)$ interchangeably wherever convenient. From equation (2.7) it follows that for any $\varphi \in \mathcal{T}$, $T_H(w) = T_H(w^\varphi) =: T_H([w])$ for every $w \in \mathcal{W}$ and finite simple graph H. Since the limit is independent of the vertex labeling of the graph sequence, the limiting graphons are therefore well-defined limits of unlabeled graphs. See Figure 2.3 for an example.

Following the counting lemma and the inverse counting lemma [Lov12, Section 7.2, Lemma 10.23, Lemma 10.32], the above discussed notion of convergence of graphs/graphons coincides with convergence with respect to a metric. To this fact, we introduce the *cut metric* on $\widehat{\mathcal{W}}$. Before this, we will first define what is called the *cut norm* on the space of bounded symmetric Borel measurable functions.

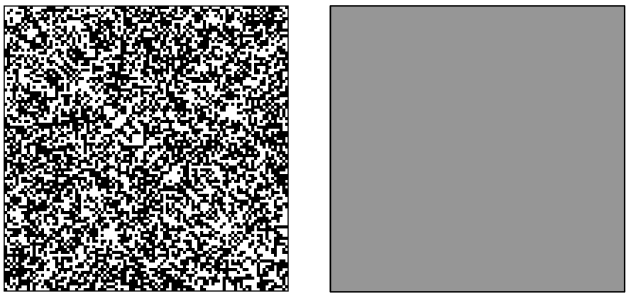


Figure 2.3: Graphon limit of an Erdős–Rényi graph $G(n, 0.5) \to w \equiv 0.5$ [Lov12, Figure 1.5].

Definition 2.10 (Cut norm)**.** The cut norm $\|\cdot\|_{\square}\colon \mathcal{W} \to \mathbb{R}_+$ is defined as

$$\|w\|_{\square} \coloneqq \sup_{S,T\subseteq[0,1]} \left| \int_{S\times T} w(x,y)\,\mathrm{d}x\,\mathrm{d}y \right|, \tag{2.8}$$

for all $w \in \mathcal{W}$ where S and T are Borel measurable subsets of $[0,1]$. The definition extends to all bounded symmetric Borel measurable functions.

The cut norm was first defined for matrices in [FK99] and was later adapted to kernels in [BCL$^+$08]. The cut norm is used to define the cut metric over the set of graphons.

Definition 2.11 (Cut metric [BCL$^+$08])**.** The cut metric $\delta_{\square}\colon \widehat{\mathcal{W}} \times \widehat{\mathcal{W}} \to \mathbb{R}_+$ is defined as

$$\delta_{\square}([w_0],[w_1]) \coloneqq \inf_{\varphi_0,\varphi_1\in\mathcal{T}} \|w_0^{\varphi_0} - w_1^{\varphi_1}\|_{\square} = \inf_{\varphi\in\mathcal{I}} \|w_0 - w_1^{\varphi}\|_{\square}, \qquad [w_0],[w_1] \in \widehat{\mathcal{W}}. \tag{2.9}$$

It is worth pointing that $\delta_{\square}$ naturally extends to kernels, but it only defines a pseudometric on $\mathcal{W}$. In fact, $\delta_{\square}(w_1, w_2) = 0$ if and only if there exists $u \in \mathcal{W}$ and Lebesgue measure preserving transforms $\varphi_1, \varphi_2\colon [0,1] \to [0,1]$ such that $w_i = u^{\varphi_i}$, for $i \in [2]$. The kernels w_1, w_2 in this case are said to be *weakly isomorphic*. In other words, graphons can be defined as the class of kernels identified up to weak isomorphism.

An important fact about the topology generated by $\delta_{\square}$ on $\widehat{\mathcal{W}}$ is that it is compact [LS06], [Lov12, Section 9.3]. Following the Weierstrass Theorem [San15, Box 1.1], it therefore follows that minimizers of lower semicontinuous functions on $\big(\widehat{\mathcal{W}}, \delta_{\square}\big)$ necessarily exist. This also allows for the existence of the iteration updates of the implicit Euler scheme, which serve as a building block of gradient flows on graphons, as we will discuss in Chapter 4.

2.5.1 Going beyond unweighted graphs

An *edge-weighted graph* G is a graph with a weight $\beta_{i,j} = \beta_{i,j}(G) \in \mathbb{R}$ associated with each edge $\{i,j\} \in E(G)$. Set $\beta_{i,j} = 0$ if $\{i,j\} \notin E(G)$. Let $(G_n)_{n\in\mathbb{N}}$ be a sequence of simple edge-weighted graphs with vertex set $[n] \coloneqq \{1,\dots,n\}$. We call $(G_n)_{n\in\mathbb{N}}$ a *dense graph sequence* if the number of edges $E(G_n)$ in G_n is $\Theta(n^2)$.

Following from the definition of convergence of unlabeled graphs via homomorphism density functions, and the fact that any real-valued matrix can be embedded in kernels, we can extend the domain of homomorphism functions to also consider the convergence of edge-weighted graphs. This allows us to consider any function over $\mathcal{M}_n$ for any $n \in \mathbb{N}$ to be appropriately defined over $\mathcal{W}_n$, enabling us to adopt the topology of the cut metric over the range of such functions. In particular, for any function $R\colon \mathcal{W} \to \mathbb{R}$ and any $n \in \mathbb{N}$, we can interpret a real-valued permutation invariant function R_n over $\mathcal{M}_n$ to be the restriction of R via the map K, i.e., $R_n = (R \circ K)\,|_{\mathcal{M}_n}$. The same can be done for the Hamiltonian H_n through the function $\mathcal{H}\colon \mathcal{W} \to \mathbb{R}$. Since we only consider R_n to be permutation invariant functions on $\mathcal{M}_n$, this corresponds to R being a well-defined function over $\widehat{\mathcal{W}}$. On the other hand, if $(K(X_n))_{n\in\mathbb{N}}$ converges to a kernel $w \in \mathcal{W}$, then if $R\colon \left(\widehat{\mathcal{W}}, \delta_\square\right) \to \mathbb{R}$ is a continuous function, then $\lim_{n\to\infty}(R \circ K)(X_n) = \lim_{n\to\infty} R_n(X_n) = R([w])$.

More generally, given any norm on the space of bounded Borel measurable symmetric function, we can define an induced metric. We will use another complete, separable metric which has properties that become useful when we define a 'gradient flow' over $\widehat{\mathcal{W}}$ in Chapter 4.

Definition 2.12 (Invariant L^2 metric [BCCH18, Jan16])**.** The invariant L^2 metric $\delta_2\colon \widehat{\mathcal{W}} \times \widehat{\mathcal{W}} \to \mathbb{R}_+$ is defined as

$$\delta_2([w_0],[w_1]) \coloneqq \inf_{\varphi_0,\varphi_1\in\mathcal{T}} \|w_0^{\varphi_0} - w_1^{\varphi_1}\|_2 = \inf_{\varphi\in\mathcal{I}} \|w_0 - w_1^{\varphi}\|_2, \qquad [w_0],[w_1] \in \widehat{\mathcal{W}}, \tag{2.10}$$

where $\|\cdot\|_2\colon L^2\big([0,1]^{(2)}\big) \to \mathbb{R}_+$ is the usual L^2-norm.

Let us mention in passing that the invariant L^2 metric is closely related to the popular Gromov-Wasserstein metric [Mém11] used to compare two metric measure spaces or their

sample equivalents [DSS$^+$20]. This can be seen by considering $[0,1]$ as a metric measure space where the measure is the Lebesgue measure and, for a given bounded metric $\mathbf{d}$, one defines a graphon as $w(x,y) = \mathbf{d}(x,y)$ for $x, y \in [0,1]$. Then the Gromov-Wasserstein distance (for $p = 2$) between $([0,1], \lambda_{[0,1]}, \mathbf{d})$ and $([0,1], \lambda_{[0,1]}, \mathbf{d}')$, for two distances $\mathbf{d}$ and $\mathbf{d}'$, is the same as computing the invariant L^2 distance between the corresponding graphons.

Unlike $\left(\widehat{\mathcal{W}}, \delta_\square\right)$, the metric space $\left(\widehat{\mathcal{W}}, \delta_2\right)$ does not enjoy compactness. Following the triangle inequality, it follows that convergence in δ_2 implies convergence in $\delta_\square$, i.e., the topology generated by δ_2 is stronger than the one generated by $\delta_\square$. As an example, consider the Erdős-Rényi graph sequence $\left(G(n,p)\right)_{n\in\mathbb{N}}$ for some $p \in (0,1)$. This sequence has a limit under $\delta_\square$ which is the all p graphon, but does not have any δ_2 limit point.

The δ_2 metric however is useful for its geometric properties. We will see in Chapter 4, that $\left(\widehat{\mathcal{W}}, \delta_2\right)$ is a geodesic metric space. This will allow us to make sense of a notion of gradient on the space of graphons, that will be connected to the standard Euclidean gradient over matrices.

We will see in Chapter 5 that the above discussed notion of convergence of unlabeled weighted graphs (or equivalently unlabeled matrices) is rather weak. As an example, consider two sequences of graphs, $(G_{n,1/2})_{n\in\mathbb{N}}$ and $(K_{n,1/2})_{n\in\mathbb{N}}$, where for every $n \in \mathbb{N}$, $G_{n,1/2}$ is the Erdős-Rényi unweighted graph with the probability of any edge present is $1/2$, and $K_{n,1/2}$ is the complete weighted graph where every graph has weight $1/2$. Both the sequences converge to the graphon $[w_{1/2}] \equiv 1/2$. We will address this issue in Chapter 5 by introducing measure-valued graphons (MVGs) and strengthening the topology, with respect to which we will take all our scaling limits.

2.5.2 Spatial and temporal scaling of curves

In this section, we will make a crucial observation on the scaling involved when we transition between curves (or iterates) in $\mathcal{W}_n$ and curves (or iterates) in $\mathcal{M}_n$ for any $n \in \mathbb{N}$.

Notice that the set of kernels $\mathcal{W}$ is a subset in the space $L^2([0,1]^{(2)})$ of square integrable functions on $[0,1]^{(2)}$. This naturally allows us to use the usual $L^2([0,1]^{(2)})$-norm $\|\cdot\|_2$ on

$\mathcal{W}$. With the natural embedding that we defined above, we immediately obtain the identity $\|X_n\|_{\mathrm{F}}^2 = n^2\|w_n\|_2^2$ where $w_n = K(X_n)$ for any $X_n \in \mathcal{M}_n$, and $\|\cdot\|_{\mathrm{F}}$ is the usual Frobenius norm on matrices.

For any function $R_n\colon \mathcal{M}_n \to \mathbb{R}$ consider the implicit Euler scheme described as follows. Starting from $X_{n,\tau} \in \mathcal{M}_n$ with a step size $\tau > 0$, the next iterate, say $X_{n,\tau,+} \in \mathcal{M}_n$ is obtained as

$$X_{n,\tau,+} \in \operatorname*{arg\,min}_{Z_n \in \mathcal{M}_n} \left[R_n(Z_n) + \frac{1}{2\tau}\|Z_n - X_{n,\tau}\|_{\mathrm{F}}^2 \right].$$

If we set $w_{n,\tau} = K(X_{n,\tau})$, then notice that the above inclusion relation can be written as

$$w_{n,\tau,+} \in \operatorname*{arg\,min}_{u_n \in \mathcal{W}_n} \left[(R_n \circ M_n)(u_n) + \frac{n^2}{2\tau}\|u_n - w_{n,\tau}\|_2^2 \right],$$

under the natural identification map between $\mathcal{M}_n$ and $\mathcal{W}_n$, where $u_n = K(Z_n)$ and $w_{n,\tau,+} = K(X_{n,\tau,+})$. The above update on $\mathcal{W}_n$ is precisely the implicit Euler update rule on $\mathcal{W}_n \subset L^2([0,1]^{(2)})$ with the step size τ/n^2. Thus we see that the 'spatial scaling' between $\mathcal{M}_n$ and $\mathcal{W}_n$ translates into a 'temporal scaling' between the implicit Euler scheme sequences on $\mathcal{M}_n$ and $\mathcal{W}_n$.

The above analysis extends even to the explicit Euler scheme that require us to have the existence of a gradient map of the function R_n. In order to relate the two iterations, i.e., one on $\mathcal{M}_n$ and the other on $\mathcal{W}_n$ obtained by applying K, we need to first define a notion of a gradient on the space of graphons. There is a natural notion of gradient of functions defined on $\widehat{\mathcal{W}}$ that we call Fréchet-like derivative, and is related to the Euclidean gradients in finite dimensions by a scaling of n^2. We first define the Fréchet-like derivative on kernels.

Definition 2.13 (Fréchet-like derivative on $\mathcal{W}$)**.** The Fréchet-like derivative of $R\colon \mathcal{W} \to \mathbb{R}$ at $v \in \mathcal{W}$ is given by $DR(v) \in L^\infty\big([0,1]^{(2)}\big)$ that satisfies the following condition,

$$\lim_{\substack{w \in \mathcal{W}, \\ \|w-v\|_2 \to 0}} \frac{R(w) - R(v) - (\langle DR(v), w\rangle - \langle DR(v), v\rangle)}{\|w - v\|_2} = 0, \tag{2.11}$$

where $\langle\cdot,\cdot\rangle$ is the usual inner product on $L^2\big([0,1]^{(2)}\big)$. If R admits a Fréchet-like derivative at every $V \in \mathcal{W}$, we say that R is Fréchet differentiable.

We make another observation: for any $v \in \mathcal{W}$, the Fréchet-like derivative of R is "coupled" with v. That is, for any $\varphi \in \mathcal{T}$, $DR(v^\varphi) = (DR(v))^\varphi$.

Lemma 2.14. *Let $n \in \mathbb{N}$. Let $R\colon \mathcal{W} \to \mathbb{R}$ be an invariant function that is Fréchet differentiable according to Definition 2.13, that is, $DR(v_n)$ exists for every $v_n \in \mathcal{W}_n$. If $R_n \coloneqq R \circ K$ is differentiable up to the boundary of $\mathcal{M}_n$, then*

$$n^2(\nabla R_n \circ M_n)(v_n) = (M_n \circ DR)(v_n), \qquad v_n \in \mathcal{W}_n.$$

We will provide a proof of the above lemma in Chapter 4, where we will discuss the Fréchet-like derivative in greater detail. Simply put, n^2 times the Euclidean gradient of R_n at a matrix argument X_n can be identified as the Fréchet-like derivative DR of R at the kernel $K(X_n)$. The time in the Euclidean gradient in Definition 2.1 is therefore scaled by n^2 following Lemma 2.14.

Given the relation obtained in Lemma 2.14, we can map the iterates of any stochastic gradient based optimization algorithm on $\mathcal{W}_n$. Since the Fréchet-like derivative is coupled with the kernel at which it is evaluated, we can consider the equivalent sets and directly map the iterates of such algorithms in $\mathcal{M}_n$ to iterates in $\widehat{\mathcal{W}}_n$. In particular, we can consider piecewise constant interpolations (see definition 2.5) of the $\widehat{\mathcal{W}}_n$-valued iterates obtained from any iterative algorithm on $\mathcal{M}_n$. Since any algorithm of interest can now be embedded on the space of graphons (and subsequently also on measure-valued graphons), we are interested in characterizing the scaling limit of these curves under a suitable topology.

2.6 Exchangeability

Working with unlabeled matrices, if looked upon from a probabilist's lens is the same as working with exchangeable arrays. Exchangeability essentially means that if we were to pick a random finite sub-array out of an array of random variables, the distribution of the finite sub-array does not depend on the labeling of the original array. More formally put, we can consider an array of real-valued random variables $(X_e)_{e\in[n]^{(2)}}$, and say it is exchangeable if $(X_e)_{e\in[n]^{(2)}}$ has the same distribution as $(X_{\pi(e)})_{e\in[n]^{(2)}}$, where if $e = (i,j) \in [n]^{(2)}$, then

$\pi(e) \coloneqq (\pi(i), \pi(j))$. Similarly, infinite exchangeable arrays (IEAs) are defined via the same definition except that we only consider finite permutations of $\mathbb{N}$.

To understand the relation between IEAs and graphons, it is useful to consider an analogy from the classical de Finetti's Theorem [Kal05, Theorem 1.1]. An infinite exchangeable sequence of random variables $(X_i)_{i\in\mathbb{N}}$ corresponds to a random probability distribution in the sense that the sequence of finite empirical distributions $\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ converges weakly, in probability, to this (random) probability distribution, as $n \to \infty$. Moreover, given an instance of this random distribution, the conditional distribution of the exchangeable sequence is i.i.d. with the same law. Similar correspondence between IEAs and (random) graphons is achieved by a powerful generalization of de Finetti's theorem due to Aldous and Hoover [Ald81, Hoo82, Ald82, Kal89]. It states that for every infinite symmetric exchangeable array $X = (X_{i,j})_{(i,j)\in\mathbb{N}^{(2)}}$ there exists a Borel measurable function $f\colon [0,1]\times[0,1]^{(2)}\times[0,1] \to \mathbb{R}$ satisfying $f(\,\cdot\,, x, y, \,\cdot\,) = f(\,\cdot\,, y, x, \,\cdot\,)$ for a.e. $(x,y) \in [0,1]^{(2)}$, and a collection of i.i.d. $\mathrm{Uni}([0,1])$ random variables $U, \{U_i\}_{i\in\mathbb{N}}, \{U_{i,j} = U_{\{i,j\}}\}_{i,j\in\mathbb{N}}$ on some probability space such that the IEA Y defined as $Y_{i,j} = f(U, U_i, U_j, U_{\{i,j\}})$ for all $(i,j) \in \mathbb{N}^{(2)}$, has the same distribution as X.

Thus, via an Aldous-Hoover representation, an IEA X induces a random graphon $w^{(U)}$, defined as

$$w^{(u)}(x,y) \coloneqq \mathbb{E}[f(U, U_1, U_2, U_{1,2}) \mid U = u, (U_1, U_2) = (x,y)], \tag{2.12}$$

for $(x,y) \in [0,1]^{(2)}$ and $u \in [0,1]$.

In Chapter 5, we will see that this correspondence between an IEA to a graphon is not one to one. As a remedy, we will introduce *measure-valued graphons* (MVGs) that will serve the complete correspondence. In Chapter 6 we will show that the scaling limits of SDEs obtained in Chapter 3 as limits of various algorithms and dynamics described in Chapter 1, will correspond to a curve on the space of exchangeable arrays, and equivalently, on the space of MVGs. This limiting curve can be projected down to a curve on the space of graphons with a loss of microscopic information on edge-distributions.

2.7 Conclusion

The concepts and tools introduced in this chapter lay the foundational groundwork for several key arguments and constructs that will be extensively developed in the subsequent chapters of this thesis. In Chapter 3, we will revisit each algorithm outlined in Section 2.2 and derive their continuous-time limits within finite dimensions, utilizing the definitions provided in Section 2.3. These continuous-time curves will then undergo further analysis in Chapter 6 to ascertain their scaling limits as the dimensionality of the ambient space approaches infinity. Given that these algorithms function within finite-dimensional bounded Euclidean spaces, it is essential to employ the stochastic process approach described in Section 2.4 to accurately determine their continuous-time limits. Additionally, Sections 2.5 and 2.6 offer both analytical and probabilistic frameworks essential for describing the spaces on which these scaling limits will be defined.

With the background and setup established, we are now prepared to construct arguments supported by examples and proofs that will substantiate the thesis of this dissertation.

Chapter 3

CONVERGENCE OF ALGORITHMS TO FINITE DIMENSIONAL SDES

In this chapter, we will fix the dimension of all iterative algorithms to $n \in \mathbb{N}$, and take their continuous time limits to obtain SDEs like equation (2.4) as discussed in Chapter 1. To do this for each class of iterative dynamics, we will need some mild assumptions that will ensure the consistency of these algorithms as we take the continuous time limit.

3.1 Noisy Stochastic Gradient Descent Algorithm

In this section, we will consider the Noisy Stochastic Gradient Descent algorithm defined in Definition 2.2 and show that as the step-size goes to zero, the sequence of iterations of the algorithm converge to the solution of an SDE:

$$\mathrm{d}X_n(t) = -n^2 \nabla R_n(X_n(t)) + \Sigma_n(X_n(t)) \circ \mathrm{d}B_n(t) - \mathrm{d}L_n^+(t) + \mathrm{d}L_n^-(t), \qquad t \in \mathbb{R}_+, \ \text{(RSDE)}$$

where $X_n(0) = X_{n,0}$ is the initialization of the algorithm. Here B_n is an $n \times n$ symmetric matrix valued process whose entries are independent Brownian motions up to matrix symmetry, and the tuple (X_n, L_n^+, L_n^-) solves the Skorokhod problem with respect to the set $\mathcal{M}_n$ (see Section 2.4).

To show this convergence, we will need some mild assumptions that we state next.

Assumption 3.1. We make the following assumptions on R and $\phi = DR$:

1. For every $n \in \mathbb{N}$, the function R_n is in $C^1(\mathcal{M}_n)$ up to the boundary of $\mathcal{M}_n$.

2. The map $\phi := DR$ is κ_2-Lipschitz with respect to $\|\cdot\|_2$, for some constant $\kappa_2 \in \mathbb{R}_+$, That is,

$$\|\phi(w_1) - \phi(w_2)\|_2 \leq \kappa_2 \|w_1 - w_2\|_2, \qquad \forall \ \ w_1, w_2 \in \mathcal{W}.$$

We now put some mild assumptions on the "small noise" and "large noise" that we discussed in Section 2.2.1. Let $g\colon \mathcal{W} \times \Omega \to L^\infty\big([0,1]^{(2)}\big)$ be a function, such that, if we define $g_n(X_n;\xi) = g(K(X_n);\xi)$ for a.e. $\xi \in \Omega$, for every $n \in \mathbb{N}$, and for $X_n \in \mathcal{M}_n$, then

$$\nabla R_n = \mathbb{E}_{\xi\sim\mathcal{D}}[g_n(\,\cdot\,;\xi)].$$

For R to be a well-defined function on graphons, we will need to assume that the law of the random variable $g(w;\xi)$ for $\xi \sim \mathcal{D}$ is invariant under measure-preserving transformations for all $w \in \mathcal{W}$, i.e., $\mathrm{Law}(g(w;\xi)) = \mathrm{Law}(g(w^\varphi);\xi)$ for all $\varphi \in \mathcal{T}$.

Assumption 3.2. We assume the following about the "small noise":

1. Law of the random variable $g(w;\xi)$ for $\xi \sim \mathcal{D}$ is invariant under measure preserving transformations for all $w \in \mathcal{W}$, i.e., $\mathrm{Law}(g(w;\xi)) = \mathrm{Law}(g(w^\varphi;\xi))$ for all $\varphi \in \mathcal{T}$.

2. The random variable $g(\,\cdot\,;\xi)$ for $\xi \sim \mathcal{D}$ has uniformly bounded variance over all finite dimensional kernels. That is, there exists $\sigma \geq 0$ such that for all $v \in \cup_{n\in\mathbb{N}}\mathcal{W}_n$,

$$\mathbb{E}_{\xi\sim\mathcal{D}}\big[\|g(v;\xi) - \phi(v)\|_2^2\big] \leq \sigma^2.$$

3. For every $n \in \mathbb{N}$, the function $g_n(\,\cdot\,;\xi) = g(\,\cdot\,;\xi) \circ K$ is in $C^0(\mathcal{M}_n)$ up to the boundary of $\mathcal{M}_n$ for all $\xi \in \Omega$.

Assumption 3.3. We assume the following about the "large noise":

1. There exists a function $\Sigma\colon \mathcal{W} \to L^\infty([0,1]^{(2)})$ such that the diffusion coefficient functions $(\Sigma_n)_{n\in\mathbb{N}}$ are restrictions of Σ, i.e., for every $n \in \mathbb{N}$, $\Sigma_n = M_n \circ \Sigma \circ K$ on $\mathcal{M}_n$.

2. The map $\Sigma\colon \mathcal{W} \to L^\infty([0,1]^{(2)})$ is κ_2-Lipschitz in $\|\cdot\|_2$ and uniformly bounded in $\|\cdot\|_\infty$ by some constant $M_\infty \in \mathbb{R}_+$, i.e., for all $u, v \in \mathcal{W}$,

$$\|\Sigma(u) - \Sigma(v)\|_2 \leq \kappa_2\|u - v\|_2, \qquad \text{and} \qquad \|\Sigma(u)\|_\infty \leq M_\infty.$$

We are now ready to state the theorem.

Theorem 3.1 (SDE limit of algorithms [HOP+22]). *Let $n \in \mathbb{N}$ be fixed, and suppose Assumptions 3.1, 3.2 and 3.3 hold. Let $W_n \colon \mathbb{R}_+ \to \mathcal{M}_n$ be the piecewise constant interpolation (see Definition 2.5) of the iterates $(X_{n,k})_{k\in\mathbb{Z}_+}$ of the projected noisy SGD algorithm (Definition 2.2). Then W_n converges weakly in the space of càdlàg processes to a process X_n as $|\boldsymbol{\tau}_n| \to 0$ that satisfies the SDE:*

$$\mathrm{d}X_n(t) = -n^2 \nabla R_n(X_n(t))\,\mathrm{d}t + \Sigma_n(X_n(t)) \circ \mathrm{d}B_n(t) + \mathrm{d}L_n^-(t) - \mathrm{d}L_n^+(t), \quad t \in \mathbb{R}_+, \tag{3.1}$$

staring at $X_n(0) = X_{n,0}$. Here B_n is a $n \times n$ symmetric matrix-valued process with coordinates independent standard Brownian motions up to matrix symmetry, and (X_n, L_n^+, L_n^-) solves the Skorokhod problem with respect to $\mathcal{M}_n$.

The proof of the above theorem is provided in Appendix A.1

Practitioners also use variants of SGD under the "small noise" setup where instead of having a single unbiased stochastic proxy of the gradient, an average over independent batches of stochastic gradients is used at every step. Authors in [MLPA22] derive weak SDE approximations of various popularly used stochastic optimization algorithms that use batches. However this existing literature does not cover SDEs with boundary terms.

3.1.1 Stochastic Gradient Descent (SGD) Algorithm

When $\Sigma_n \equiv 0$, equation (3.1) reduces to

$$\mathrm{d}X_n(t) = -n^2 \nabla R_n(X_n(t))\,\mathrm{d}t + \mathrm{d}L_n^-(t) - \mathrm{d}L_n^+(t), \quad t \in \mathbb{R}_+, \quad X_n(0) = W_{n,0}, \tag{3.2}$$

such that (X_n, L_n^+, L_n^-) solves the Skorokhod problem on $\mathcal{M}_n$ (see Section 2.4 for details). Moreover, it is shown in Appendix A.1.1 that the solution of equation (3.2) is the same as the solution of (3.3) given below. Furthermore, we will show in Chapter 4 that if the solution $X_n \colon \mathbb{R}_+ \to \mathcal{M}_n$ of

$$\mathrm{d}X_n(t) = -n^2 \nabla R_n(X_n(t)) \circ \mathbb{1}\{G_n(X_n(t))\}\,\mathrm{d}t, \qquad t \in \mathbb{R}_+, \tag{3.3}$$

exists, where $G_n(A)$ is the subset of $[n]^2$ defined as

$$\begin{aligned} G_n(A) \coloneqq & \big\{(i,j) \in [n]^2 \;\big|\; |A(i,j)| < 1\big\} \\ & \cup \big\{(i,j) \in [n]^2 \;\big|\; A(i,j) = 1, \partial_{i,j} R_n(A) > 0\big\} \\ & \cup \big\{(i,j) \in [n]^2 \;\big|\; A(i,j) = -1, \partial_{i,j} R_n(A) < 0\big\}, \end{aligned} \tag{3.4}$$

for all $A \in \mathcal{M}_n$, then X_n is a gradient flow on $\mathcal{M}_n$ in a suitable sense.

3.2 Relaxed Metropolis-Hastings Algorithm

In this Section, we will consider the Relaxed Metropolis chain algorithm as described in Section 2.2.2. Notice that for any $n \in \mathbb{N}$, we can interpret the real-valued permutation invariant function H_n over $\mathcal{M}_n$ to be the restriction of $\mathcal{H}$ via the map K, i.e., $H_n = (\mathcal{H} \circ K)\,|_{\mathcal{M}_{n,+}}$. We will call the function $\mathcal{H}$, the Hamiltonian function. In this Section, we will show that the relaxed Metropolis chain algorithm on a the state space $\mathcal{S}_{r,n}$, converges to an SDE on $\mathcal{M}_{n,+}$ as $r \to \infty$. Following equation 2.2, we will denote the model specified by $\widehat{Q}_{r,n,\beta}$, as ESBM$[n, r, \beta, \mathcal{H}]$, where $\mathcal{H}$ is the Hamiltonian defined on $\widehat{\mathcal{W}}$ such that $\mathcal{H} \circ K$ restricted on $\mathcal{M}_{n,+}$ is the function H_n.

We will set some assumptions that will be necessary to derive the continuous time limit of the algorithm.

Assumption 3.4. Let $\mathcal{H}\colon \widehat{\mathcal{W}} \to \mathbb{R}$ be bounded below, Fréchet-like differentiable with $D\mathcal{H}$ denoting its Fréchet-like derivative (see Definition 2.13) and satisfy

$$\frac{\lambda}{2}\|u - v\|_2^2 \le \mathcal{H}(v) - \mathcal{H}(u) - \langle D\mathcal{H}(u), v - u\rangle \le \frac{L}{2}\|u - v\|_2^2, \tag{3.5}$$

for every $u, v \in \mathcal{W}_{[0,1]}$, for some constants $\lambda \in \mathbb{R}$ and $L > 0$. Further, assume that $D\mathcal{H}$ is $(\|\cdot\|_\square \to \|\cdot\|_\infty)$-Lipschitz.

Definition 3.2. Let $\mathcal{H}$ satisfy Assumptions 3.4. Let $\beta > 0$. Define $b_0\colon \mathcal{W}_{[0,1]} \to L^\infty\big([0,1]^{(2)}\big)$ as

$$b_0(w) \coloneqq -2\beta D\mathcal{H}(w) \exp\big(\beta^2 n^{-2}\|D\mathcal{H}(w)\|_2^2\big)\overline{\Phi}\Big(\sqrt{2}\beta n^{-1}\|D\mathcal{H}(w)\|_2\Big), \qquad w \in \mathcal{W}_{[0,1]},$$

where $\overline{\Phi}$ is the right tail of standard Gaussian, i.e., $\overline{\Phi}(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-y^2/2)\,\mathrm{d}y$, for $x \in \mathbb{R}_+$. For any $n \in \mathbb{N}$, we will denote the restriction of b_0 to $\mathcal{M}_{n,+}$ by b_n. That is, $b_n = M_n \circ b_0 \circ K$.

In the context of the Metropolis chain algorithm, with an abuse of notation, we will use the notation b and b_0 interchangeably. By Lemma A.3 in Appendix A.2.1, $n^{-4} b_n(w) = \mathbb{E}\big[Z \exp\big(-\beta_{r,n}\gamma_r \langle \nabla H_n, Z\rangle_{\mathrm{F}}^+\big)\big]$ where Z is $n \times n$ symmetric matrix with i.i.d. Gaussian entries, and $\beta_{r,n} \coloneqq \beta n^{-2}/\gamma_r$ as defined in Section 2.2.2. Thus $\|b_n\|_\infty < \infty$.

Remark 3.3. It follows from Assumption 3.4 that $\|D\mathcal{H}\|_\infty \le C$ for some $C \ge 0$ and therefore $\|D\mathcal{H}(v)\|_2 \le C$ for every $v \in \mathcal{W}$. Since $\mathrm{e}^x \to 1$ as $x \to 0$ and $\overline{\Phi}(x) \to \frac{1}{2}$ as $x \to 0$, it follows that $\|b(v) + \beta D\mathcal{H}(v)\|_\infty \to 0$ as $n \to \infty$.

We now recall that the relaxed Metropolis algorithm samples from the ESBM$[r, n, \beta, \mathcal{H}]$ model. Given $G(k) \in \mathcal{S}_{r,r}$ and the matrix $q_{n,k}^{(r)} \in \mathcal{M}_{n,+}$ of edge-densities for any $k \in \mathbb{Z}_+$, we run the relaxed Metropolis chain consisting of the following steps.

1. Run the base chain for $s_r \coloneqq \lceil \gamma_r^2 r^4 \rceil$ many steps. Let $\widetilde{G}(k+1) \in \mathcal{S}_{r,n}$ be the graph obtained after such s_r many steps. Let $\widetilde{q}_{n,k+1}^{(r)}$ denote the matrix of edge densities of $\widetilde{G}(k+1)$.

2. Given $G(k), \widetilde{G}(k+1)$ for any $k \in \mathbb{Z}_+$, define

$$Y(k+1) = \begin{cases} \widetilde{G}(k+1), & \text{w.p.} \quad \exp\left(-\beta_{r,n}\Big[H_n\Big(\widetilde{q}_{n,k+1}^{(r)}\Big) - H_n\Big(q_{n,k}^{(r)}\Big)\Big]^+\right), \\ G(k), & \text{otherwise,} \end{cases}$$

   where $a^+ = \max\{0, a\}$ for $a \in \mathbb{R}$ and $\beta_{r,n} \coloneqq \beta n^{-2}/\gamma_r$ as defined in Section 2.2.2. Let $p_{n,k+1}^{(r)}$ be the matrix of edge-densities for $Y(k+1)$. Observe that

$$p_{n,k+1}^{(r)} = \begin{cases} \widetilde{q}_{n,k+1}^{(r)}, & \text{if} \quad Y(k+1) = \widetilde{G}(k+1), \\ q_{n,k}^{(r)}, & \text{if} \quad Y(k+1) = G(k). \end{cases}$$

3. After the accept-reject step, we again run the base chain starting from $Y(k+1) \in \mathcal{S}_{r,n}$ for $\ell_{r,n} \coloneqq \lceil n^{-4}\sigma^2\gamma_r r^4 \rceil$ many steps for some $\sigma > 0$. Let the graph obtained thereafter be $G(k+1)$, and let $q^{(r)}_{n,k+1}$ be the edge density matrix of $G(k+1)$.

This procedure gives a Markov chain $(G(k))_{k\in\mathbb{N}}$ on the state space $\mathcal{S}_{r,n}$ with corresponding process of edge-density matrix $\left(q^{(r)}_{n,k}\right)_{k\in\mathbb{N}}$. In the following we show that, as $r \to \infty$ the latter process converges to as SDE with drift b_n as defined in Definition 3.2. For fixed $n \in \mathbb{N}$, the adjacency matrix of $G(k)$ converges to the corresponding edge-density matrix $q^{(r)}_{n,k}$ in the cut metric as $r \to \infty$ uniformly. In Section 6.2.2 we will show that we can interpret the limiting deterministic curve on the space of graphons as the cut limit of the process of adjacency matrices of $(G(k))_{k\in\mathbb{N}}$ as $r \to \infty$ followed by $n \to \infty$.

3.2.1 Heuristic analysis of the graph edge-density process

Before we state our result we analyze heuristically the process of edge-density matrix $\left(q^{(r)}_{n,k}\right)_{k\in\mathbb{Z}_+}$ as defined above. To this end, for any $k \in \mathbb{Z}_+$, define $\Delta q^{(r)}_{n,k} \coloneqq q^{(r)}_{n,k+1} - q^{(r)}_{n,k}$ and let $\mathcal{F}_k$ be the sigma algebra generated by $\left\{q^{(r)}_{n,i} \,\middle|\, i = 0, \ldots, k\right\}$. Given $k \in \mathbb{Z}_+$, let us analyze $\mathbb{E}\left[\Delta q^{(r)}_{n,k} \,\middle|\, \mathcal{F}_k\right]$. Notice that

$$\mathbb{E}\left[\Delta q^{(r)}_{n,k} \,\middle|\, \mathcal{F}_k\right] = \mathbb{E}\left[\widetilde{\Delta} q^{(r)}_{n,k} \,\middle|\, \mathcal{F}_k\right] + \mathbb{E}\left[q^{(r)}_{n,k+1} - p^{(r)}_{n,k+1} \,\middle|\, \mathcal{F}_k\right],$$

where $\widetilde{\Delta} q^{(r)}_{n,k} \coloneqq p^{(r)}_{n,k+1} - q^{(r)}_{n,k}$. Notice that given $p^{(r)}_{n,k+1}$, the increment of the (i,j)-th coordinate, $q^{(r)}_{n,k+1,(i,j)} - p^{(r)}_{n,k+1,(i,j)}$, has the same distribution as the reflected random walk of step-size $\frac{1}{r^2}$ run for $\ell_{r,n}$ steps. Assuming that the random walk does not hit the boundary during relaxation (hitting the boundary is rare), this is very small. It follows that $\mathbb{E}\left[\Delta q^{(r)}_{n,k} \,\middle|\, \mathcal{F}_k\right] \approx \mathbb{E}\left[\widetilde{\Delta} q^{(r)}_{n,k} \,\middle|\, \mathcal{F}_k\right]$

$$= \mathbb{E}\left[\left(\widetilde{q}^{(r)}_{n,k+1} - q^{(r)}_{n,k}\right) \exp\left(-\beta_{r,n}\left[\mathcal{H}\left(K\left(\widetilde{q}^{(r)}_{n,k+1}\right)\right) - \mathcal{H}\left(K\left(q^{(r)}_{n,k}\right)\right)\right]^+\right)\right]. \tag{3.6}$$

By Assumption 3.4,

$$\mathcal{H}\left(K\left(\widetilde{q}^{(r)}_{n,k+1}\right)\right) - \mathcal{H}\left(K\left(q^{(r)}_{n,k}\right)\right) - \left\langle D\mathcal{H}\left(K\left(q^{(r)}_{n,k}\right)\right), K\left(\widetilde{q}^{(r)}_{n,k+1}\right) - K\left(q^{(r)}_{n,k}\right)\right\rangle$$

$$\approx \left\| K\left(\widetilde{q}^{(r)}_{n,k+1}\right) - K\left(q^{(r)}_{n,k}\right)\right\|_2^2.$$

On the other hand, given $q^{(r)}_{n,k}$, the increment of each coordinate $\widetilde{q}^{(r)}_{n,k+1,(i,j)} - q^{(r)}_{n,k,(i,j)}$ for every $(i,j) \in [n]^{(2)}$ has the same distribution as a symmetric random walk (with reflections at the boundary) with step-size $1/r^2$ run for $s_r = \gamma_r^2 r^4$ steps. In particular, $\mathbb{E}\left[\left\| K\left(\widetilde{q}^{(r)}_{n,k+1}\right) - K\left(q^{(r)}_{n,k}\right)\right\|_2^2\right] = \gamma_r^2$. Two important and non-trivial consequences of this heuristic are the following:

1. Due to a concentration of measure argument, $\left\| K\left(\widetilde{q}^{(r)}_{n,k+1}\right) - K\left(q^{(r)}_{n,k}\right)\right\|_2^2 \leq C\gamma_r^2 \log r$ for some constant $C > 0$ with high probability.

2. $\widetilde{q}^{(r)}_{n,k+1} - q^{(r)}_{n,k}$ has approximately the same distribution as $\gamma_r Y_n$ where Y_n is an $n \times n$ symmetric matrix of independent standard Gaussians. Notice that if $q^{(r)}_{n,k,(i,j)} \in \{0,1\}$ for any $(i,j) \in [n]^{(2)}$, this is not true. This approximation is valid only when all the coordinates of $q^{(r)}_{n,k}$ are sufficiently away from $\{0,1\}$. With a careful analysis, one can show that this is indeed the case except for a negligible fraction of time.

Assuming the above heuristics and using equation (3.6) we obtain that with high probability, $\mathbb{E}\left[\Delta q^{(r)}_{n,k} \,\middle|\, \mathcal{F}_k\right]$

$$\begin{aligned}
&\approx \mathbb{E}\left[\left(\widetilde{q}^{(r)}_{n,k+1} - q^{(r)}_{n,k}\right) \exp\left(-\beta_{r,n}\left\langle D\mathcal{H}\left(K\left(q^{(r)}_{n,k}\right)\right), K\left(\widetilde{q}^{(r)}_{n,k+1}\right) - K\left(q^{(r)}_{n,k}\right)\right\rangle^{+}\right)\right] \\
&= \gamma_r\, \mathbb{E}\left[Y_n \exp\left(-\beta_{r,n}\gamma_r\left\langle \nabla H_n\left(q^{(r)}_{n,k}\right), Y_n\right\rangle_{\mathrm{F}}^{+}\right)\right], \qquad (3.7)
\end{aligned}$$

where we used the fact that for any two $n \times n$ symmetric matrices $A_n, B_n \in \mathcal{M}_{n,+}$ we have $\langle A_n, B_n\rangle_{\mathrm{F}} = n^2\langle K(A_n), K(B_n)\rangle$ and that the fact that $D\mathcal{H} = n^{-2}\nabla H_n$ (see Lemma 2.14). The expectation in the last expression above is very amenable to analysis. It follows from Lemma A.3 that $\mathbb{E}\left[Y_n \exp\left(-\beta_{r,n}\gamma_r\left\langle \nabla H_n\left(q^{(r)}_{n,k}\right), Y_n\right\rangle_{\mathrm{F}}^{+}\right)\right] = n^{-4} b_n\left(q^{(r)}_{n,k}\right)$ where b_n is defined in Definition 3.2.

The above heuristic can now be summarized as follows. With high probability

$$\mathbb{E}\left[\Delta q^{(r)}_{n,k} \,\middle|\, \mathcal{F}_k\right] \approx \gamma_r n^{-4} b_r\left(q^{(r)}_{n,k}\right), \qquad (3.8)$$

provided that the all coordinates of $q_{n,k}^{(r)}$ are away from $\{0,1\}$. Now let us analyze the conditional covariance of $\Delta q_{n,k}^{(r)}$. Recall that $\mathbb{E}\Big[\Big\|K\Big(\widetilde{q}_{n,k+1}^{(r)}\Big)-K\Big(q_{n,k}^{(r)}\Big)\Big\|_2^2\,\Big|\,\mathcal{F}_k\Big]\le\gamma_r^2$. On the other hand, given $p_{n,k+1}^{(r)}$, the increment $q_{n,k+1,(i,j)}^{(r)}-p_{n,k+1,(i,j)}^{(r)}$ of coordinate $(i,j)\in[n]^{(2)}$ has the same distribution as the symmetric random walk with step-size $\frac{1}{r^2}$ (reflected at the boundary $\{0,1\}$) running for $\ell_{r,n}\approx n^{-4}\gamma_r\sigma^2r^4$ steps. In particular, each coordinate has variance $\approx n^{-4}\gamma_r\sigma^2$. Also note that given $p_{n,k}^{(r)}$, the coordinates of $q_{n,k+1}^{(r)}-p_{n,k+1}^{(r)}$ are independent. In particular,

$$\mathrm{Cov}\Big(\Delta q_{n,k}^{(r)}\,\Big|\,\mathcal{F}_k\Big)=n^{-4}\gamma_r\sigma^2I_n+O(\gamma_r^2),\tag{3.9}$$

where $O(\gamma_r^2)$ means that each coordinate of $\mathrm{Cov}\Big(\Delta q_{n,k}^{(r)}\,\Big|\,\mathcal{F}_k\Big)$ differs from $n^{-4}\gamma_r\sigma^2I_n$ at most by a constant factor of γ_r^2.

For a fix $t>0$, we will define $t_{r,n}\coloneqq\lfloor tn^4/\gamma_r\rfloor$. Also define $q_n^{(r)}\colon\mathbb{R}_+\to\mathcal{M}_{n,+}$ to be a piecewise constant interpolation of $\Big(q_{n,k}^{(r)}\Big)_{k\in\mathbb{Z}_+}$ given by

$$q_n^{(r)}(t)\coloneqq q_{n,t_{r,n}}^{(r)},\qquad t\in\mathbb{R}_+,\tag{3.10}$$

In particular, we obtain

$$q_n^{(r)}(t)-q_n^{(r)}(0)=\sum_{k=0}^{t_{r,n}-1}\mathbb{E}\Big[\Delta q_{n,k}^{(r)}\,\Big|\,\mathcal{F}_k\Big]+\sum_{k=0}^{t_{r,n}-1}\Big(\Delta q_{n,k}^{(r)}-\mathbb{E}\Big[\Delta q_{n,k}^{(r)}\,\Big|\,\mathcal{F}_k\Big]\Big),\qquad t\in\mathbb{R}_+.$$

Using the heuristic derived in (3.8) and (3.9), one expects that

$$q_n^{(r)}(t)-q_n^{(r)}(0)\approx\sum_{k=0}^{t_{r,n}-1}\gamma_rn^{-4}b_n\Big(q_{n,k}^{(r)}\Big)+\sum_{k=0}^{t_{r,n}-1}\Delta M_{n,k}^{(r)},\qquad t\in\mathbb{R}_+,\tag{3.11}$$

where $\Big(\Delta M_{n,k}^{(r)}\Big)_{k\in\mathbb{Z}_+}$ is a $\mathcal{M}_n$-valued martingale difference sequence with uniform coordinate-wise variance $\gamma_rn^{-4}\sigma^2$. We must caution that the approximation in (3.11) is not valid if $q_{n,k}^{(r)}$ is close to boundary $\{0,1\}$. The heuristic calculations have been derived under the assumption that all coordinates of $q_{n,k}^{(r)}$ are away from $\{0,1\}$. Ignoring this boundary contribution, it is reasonable to conclude that

$$q_n^{(r)}(t)-q_n^{(r)}(0)\approx\int_0^t b_n\big(q_n^{(r)}(s)\big)\,\mathrm{d}s+\sigma B_n(t),\qquad t\in\mathbb{R}_+,$$

where B_n is an $n \times n$ matrix with i.i.d. Brownian motions (up to matrix symmetry), in the interior of the state space. In view of this, it is reasonable to expect to that if the process $\left(q_{n,k}^{(r)}\right)_{k\in\mathbb{Z}_+}$ spends negligible proportion of time at the boundary, then

$$q_n^{(r)}(t) - q_n^{(r)}(0) \approx \int_0^t b_n\big(q_n^{(r)}(s)\big)\,\mathrm{d}s + \sigma B_n(t) + L_n^{(0)}(t) - L_n^{(1)}(t), \qquad t \in \mathbb{R}_+,$$

where $\left(q_n^{(r)}, L_n^{(0)}, L_n^{(1)}\right)$ solves the Skorokhod problem on the cube $\mathcal{M}_{n,+}$. That is, each coordinate process of $q_n^{(r)}$ satisfies the above SDE with reflection at the boundary $\{0, 1\}$. This heuristic argument can be made precise (see Theorem 3.4) and it is one of the main takeaways of this section. Before we state the main theorem, we make a brief digression to the Skorokhod problem and the Skorokhod map which will play a crucial role in Theorem 3.4 and its proof.

Theorem 3.4 ([APST23])**.** *Let $\mathcal{H}$ satisfy Assumption 3.4 and let $(\gamma_r)_{r\in\mathbb{N}}$ satisfy condition* (2.1)*. Let $D([0,\infty), \mathcal{M}_{n,+})$ be the space of right continuous functions with left limits equipped with the topology of uniform convergence over compact subsets. Let $q_n^{(r)} \colon \mathbb{R}_+ \to \mathcal{M}_{n,+}$ be a piecewise interpolation of $\left(q_{n,k}^{(r)}\right)_{k\in\mathbb{Z}_+}$ (see equation* (3.10)*). Then, $q_n^{(r)}$ converges weakly in $D([0,\infty), \mathcal{M}_{n,+})$ to a process X_n over compact time intervals, with continuous path that satisfies the SDE*

$$\mathrm{d}X_n(t) = b_n(X_n(t)))\,\mathrm{d}t + \sigma\,\mathrm{d}B_n(t) + \mathrm{d}L_n^{(0)}(t) - \mathrm{d}L_n^{(1)}(t), \qquad t \in \mathbb{R}_+, \tag{3.12}$$

with initial condition $X_n(0) = q_{n,0}^{(r)}$, where B_n is a symmetric $n \times n$ matrix with whose coordinates are i.i.d. Brownian motions (up to matrix symmetry) and $\left(X_n, L_n^{(0)}, L_n^{(1)}\right)$ solves the Skorokhod problem w.r.t. the finite dimensional cube $\mathcal{M}_{n,+}$ (see Section 2.4).

We provide the proof of Theorem 3.4 in Appendix A.2

3.3 Iterated product of matrices

In Section 2.2.3, we introduced the general setup of algorithms that are of the multiplicative form and rely on products of iterated matrices. In this section, we will take the scaling limit of

the iterated product of such random matrices to obtain their continuous-time counterpart. To recall, the k-th random matrix is a perturbation of the identity and takes the form $I_n + X_{n,k}^{(m)}$, where

$$X_{n,k}^{(m)} = \frac{\mu_n}{m} M_{n,k}^{(m)} + \frac{\sigma_n}{\sqrt{m}} G_{n,k}^{(m)}, \tag{3.13}$$

where $\mathbb{E}\left[M_{n,k}^{(m)}\right] = A_{n,k}^{(m)} \in \mathbb{R}^{[n]^2}$ and $M_{n,k}^{(m)}$ is an independent matrix with entries i.i.d. as $N(0,1)$.

Let's separate the two components of the perturbation and see what we can expect.

1. Consider a triangular array of $n \times n$ deterministic matrices $\left(\left(A_{n,k}^{(m)}\right)_{k\in[m]}\right)_{m\in\mathbb{N}}$ and define the following iterated product of matrices

$$P_n^{(m)}(k) \coloneqq \left(I_n + \frac{\mu_n}{m} A_{n,k}^{(m)}\right) \dots \left(I_n + \frac{\mu_n}{m} A_{n,2}^{(m)}\right) \left(I_n + \frac{\mu_n}{m} A_{n,1}^{(m)}\right), \qquad k \in [m], \tag{3.14}$$

where μ_n is a dimension dependent scaling factor and $P_n^{(m)}(0) = I_n$. Note that $P_n^{(m)}$ satisfies following difference equation

$$P_n^{(m)}(k+1) - P_n^{(m)}(k) = \frac{\mu_n}{m} A_{n,k+1}^{(m)} P_n^{(m)}(k), \qquad k \in [m-1].$$

It is reasonable to expect that $P_n^{(m)}$ admits a scaling limit as $m \to \infty$ for every $n \in \mathbb{N}$. That is, under appropriate conditions on the curve A_n defined as $A_n(t) \coloneqq \lim_{m\to\infty} A_{n,\lfloor mt \rfloor}^{(m)}$ for $t \in [0,1]$, we should expect that the curve $P_{n,m}$ defined as $P_{n,m}(t) \coloneqq P_n^{(m)}(\lfloor mt \rfloor)$ for $t \in [0,1]$, converges to an absolutely continuous curve, say P_n, satisfying $\frac{\mathrm{d}}{\mathrm{d}t} P_n(t) = \mu_n A_n(t) P_n(t)$ as $m \to \infty$.

If $A_n(t) \equiv A_n$ is a constant curve, then the solution to this differential equation is $P_n(t) = \mathrm{e}^{t\mu_n A_n}$. For more general curve A_n, one may guess the solution of the above differential equation to be $P_n(t) = \mathrm{e}^{\int_0^t \mu_n A_n(s)\,\mathrm{d}s}$. However, this is incorrect – unless $A_n(s)$ and $A_n(s')$ commute for all $s, s' \in [0,t]$. However, the correct solution can be defined using a non-commutative analogue of exponential that we define later in this section.

2. Consider the case where instead of $\left(A_{n,k}^{(m)}\right)_{k\in[m]}$, we have i.i.d. matrices $\left(G_{n,k}^{(m)}\right)_{k\in[m]}$ with i.i.d. standard Gaussian coordinates and instead of equation (3.14), we consider the iterated product of matrices such that

$$P_n^{(m)}(k+1) - P_n^{(m)}(k) = \frac{\sigma_n}{\sqrt{m}} G_{n,k+1}^{(m)} P_n^{(m)}(k).$$

Following a similar heuristic as above, one may expect that $P_{n,m}$ converges to a matrix valued SDE satisfying $\mathrm{d}P_n(t) = \sigma_n \,\mathrm{d}B_n(t)P_n(t)$ for $t \in [0,1]$ as $m \to \infty$, where B_n is an $n \times n$ matrix whose coordinates are i.i.d. Brownian motions.

In Theorem 3.8 we show that, under appropriate assumptions and suitable time scaling, the iterated matrix product has a scaling limit that is given as the unique solution Y_n of an SDE. Moreover, we give the solution explicitly as a non-commutative analogue of exponential of Y_n, denoted $\mathrm{Texp}[Y_n]$, that we define later (see Definition 3.5).

Various authors have studied similar problems – in fixed dimension $n \in \mathbb{N}$. For instance, it is shown in [EH18] that if

$$Q_{n,k}^{(m)} = \left(I_n + \frac{1}{m}A_{n,k}\right) \ldots \left(I_n + \frac{1}{m}A_{n,2}\right)\left(I_n + \frac{1}{m}A_{n,1}\right), \quad k \in [m], \tag{3.15}$$

then $Q_{n,m}^{(m)}$ converges to e^{A_n} where $A_n \coloneqq \lim_{m\to\infty} \frac{1}{m}\sum_{k=1}^m A_{n,k}$ (assuming this limit exists) as $m \to \infty$. A particularly important case that this result covers is the case when $\{A_{n,k}\}_{k\in\mathbb{N}}$ are i.i.d. and have the expectation $\mathbb{E}[A_{n,k}] = A_n$. In this particular case, the rate of convergence was investigated in [HW20, KMS20]. It follows easily from the above result that if $(A_{n,k})_{k\in\mathbb{N}}$ is a sequence of matrices such that $\lim_{m\to\infty} \frac{1}{m}\sum_{k=1}^m A_{n,k} = A_n$ then $Q_{n,\lfloor mt\rfloor}^{(m)} \to \mathrm{e}^{tA_n}$ as $m \to \infty$. However, this theorem does not apply to the triangular array of matrices. If we define the triangular array $A_{n,k}^{(m)} \coloneqq A_{n,k}$ for all $k \in [m]$ and all $m \in \mathbb{N}$, then our result (Theorem 3.8) recovers this result (see Example 3.3). It should be noted that if $\left(\left(A_{n,k}^{(m)}\right)_{k\in[m]}\right)_{m\in\mathbb{N}}$ is a triangular array of matrices such that $A_n \coloneqq \lim_{m\to\infty} \frac{1}{m}\sum_{k=1}^m A_{n,k}^{(m)}$ exists and also the limit $P_n(t) \coloneqq \lim_{m\to\infty} P_n^{(m)}(\lfloor mt\rfloor)$ exists, it is not necessarily true that $P_n(t) = \mathrm{e}^{tA_n}$.

We now ready to explore the scaling limit of iterated product of matrices as defined in (3.14) as $m \to \infty$ while the dimension n is kept fixed. We begin with some definitions and notations.

Definition 3.5 (Time ordered exponential)**.** Let $t \mapsto Y_n(t)$ be an $n \times n$ matrix valued càdlàg semimartingale. For any $k \in \mathbb{N}$ and $t \in \mathbb{R}_+$, let $\Delta_k(t) \coloneqq \big\{(s_1, \ldots, s_k) \in [0,t]^k \,\big|\, t \geq s_k \geq s_{k-1} \geq \ldots s_1 \geq 0\big\}$ and define

$$J_k(Y_n)(t) \coloneqq \int_{\Delta_k(t)} \mathrm{d}Y_n(s_k) \ldots \mathrm{d}Y_n(s_1), \qquad k \in \mathbb{N}, \qquad J_0(Y_n) \equiv I_n.$$

We define the non-commutative exponential $\mathrm{Texp}[\,\cdot\,]$ of Y_n as

$$\mathrm{Texp}[Y_n](t) \coloneqq \sum_{k=0}^{\infty} J_k(Y_n)(t), \qquad \text{and} \qquad \Gamma(Y_n)(t) \coloneqq \mathrm{Texp}[Y_n](t) - I_n, \qquad t \in \mathbb{R}_+. \tag{3.16}$$

If $t \mapsto Y_n(t) = tA_n$ for some fixed matrix A_n, then $\mathrm{Texp}[Y_n](t) = \mathrm{e}^{tA_n}$ for every t. Even for a general (deterministic) absolutely continuous curve $t \mapsto Y_n(t)$, the non-commutative exponential $\mathrm{Texp}[Y_n](t)$ admits a beautiful interpretation that we explain in Example 3.1.

Definition 3.6 (Poisson point process)**.** Let N be a unit intensity Poisson point process on $\mathbb{R}_+$. For every $t \in \mathbb{R}_+$, define N_t as the set of atoms from N occurring up to time t, such that $N_t(\omega) = N(\omega) \cap [0,t]$ for every realization ω. The set N_t is ordered in a non-decreasing manner, reflecting the chronological order of atoms up to time t. Recall that conditioned on $|N_t| = k$, the distribution of ordered tuple of points $0 \leq s_1 \leq s_2 \leq \ldots \leq s_k \leq t$ in N_t has uniform distribution on $\Delta_k(t)$. We refer the reader to [Bré20] for more detail on Poisson point processes.

Example 3.1. Suppose Y_n is a deterministic and absolutely continuous curve. Let $Y_n(t) = \int_0^t A_n(s)\,\mathrm{d}s$ for $t \in \mathbb{R}_+$, where the integral is applied coordinatewise. For $k \geq 1$, it follows that

$$J_k(Y_n)(t) = \int_{\Delta_k(t)} A_n(s_k)A_n(s_{k-1}) \ldots A_n(s_1) \prod_{j=1}^{k} \mathrm{d}s_j$$

$$= |\Delta_k(t)| \int_{\Delta_k(t)} A_n(s_k) A_n(s_{k-1}) \dots A_n(s_1) \, \mathrm{d}\sigma_{k,t}(s_k, \dots, s_1)$$
$$= \mathrm{e}^t \, \mathbb{E}\left[\prod_{\alpha \in N_t} A_n(\alpha) \,\middle|\, |N_t| = k \right] \mathbb{P}\{|N_t| = k\},$$

where $|\Delta_k(t)|$ is the volume (that is k-dimensional Lebesgue measure) of the simplex $\Delta_k(t)$, $\sigma_{k,t}$ is the uniform measure on $\Delta_k(t)$, and the last line follows by observing that $\mathbb{P}\{|N_t| = k\} = \mathrm{e}^{-t} \frac{t^k}{k!} = \mathrm{e}^{-t} |\Delta_k(t)|$ for every $t \in \mathbb{R}_+$. We define an empty product of matrices to be I_n, and always interpret $\prod$ of a finite collection of matrices indexed by time as denoting ordered multiplication going from left to right with increasing time indices. With this notation,

$$\mathrm{Texp}[Y_n](t) = \mathrm{e}^t \, \mathbb{E}\left[\prod_{\alpha \in N_t} A_n(\alpha) \right], \qquad t \in \mathbb{R}_+.$$

The following proposition gives a characterization of $\mathrm{Texp}[\,\cdot\,]$ of a semimartingale that justifies the name non-commutative exponential.

Proposition 3.7. *Let Y_n be a continuous $\mathcal{M}_n$-valued semimartingale. Then, there exists a pathwise unique $\mathcal{M}_n$-valued process Z_n satisfying*

$$Z_n(t) = I_n + \int_0^t \mathrm{d}Y_n(t) \cdot Z_n(t), \qquad t \in \mathbb{R}_+.$$

Moreover, $Z_n(t) = \mathrm{Texp}[Y_n](t)$ for all $t \in \mathbb{R}_+$.

The proof of Proposition 3.7 is provided in Appendix A.3.

We consider the product defined in equation (3.13), where $\left(G_{n,k}^{(m)}\right)_{k \in [m]}$ is a sequence of i.i.d. matrices with zero mean, and consider the following product

$$P_{n,m}(t) \coloneqq \prod_{k=1}^{\lfloor mt \rfloor} \left(I_n + X_{n,k}^{(m)} \right), \qquad m \in \mathbb{N}, \quad t \in [0,1]. \tag{3.17}$$

With Assumption 3.5 stated below, we are ready to state the main theorem of this section.

Assumption 3.5. There exists $C, D \geq 0$ such that

1. For every $m \in \mathbb{N}$, and $k \in [m]$, $\left\| A_{n,k}^{(m)} \right\|_{\max} \leq C$, and $\mathrm{Cov}\left(M_{n,k}^{(m)}, \preceq \right) nDI_n$.

2. For every $m \in \mathbb{N}$, $\left(G_{n,k}^{(m)}\right)_{k\in[m]}$ is an i.i.d. sequence of matrices with i.i.d. $N(0,1)$ entries.

3. The piecewise constant interpolations $A_{n,m}$ of $\left(A_{n,k}^{(m)}\right)_{k\in[m]}$, defined as $A_{n,m}(t) \coloneqq A_{n,\lfloor mt\rfloor}^{(m)}$ for $t \in [0,1]$, uniformly converge to an absolutely continuous curve A_n as $m \to \infty$.

Theorem 3.8 (Convergence to SDE for fixed dimension [STH+24])**.** *Let* $\left(\left(X_{n,k}^{(m)}\right)_{k\in[m]}\right)_{m\in\mathbb{N}}$ *be the triangular array defined in equation* (3.13)*. Under Assumption 3.5, the curve* $P_{n,m}$ *(as defined in equation* (3.17)*) uniformly converges to* $\mathrm{Texp}[Y_n]$ *as* $m \to \infty$*, where*

$$Y_n(t) \coloneqq \mu_n \int_0^t A_n(s)\,\mathrm{d}s + \sigma_n B_n(t), \qquad t \in [0,1],$$

and B_n *is a* $n \times n$ *matrix with i.i.d. BM coordinates.*

The proof of Theorem 3.8 is provided in Appendix A.3.

Example 3.2. Consider a simple example in the case when $n = 1$. Let $B(t)$ be the standard one dimensional Brownian motion. Then,

$$J_k(B)(t) = \frac{1}{k!}H_k(B_t),$$

where H_k is the k-th Hermite polynomial. In particular, we get that $\mathrm{Texp}[B](t) = \mathrm{e}^{B_t - t/2}$. In other words, $\mathrm{Texp}[\,\cdot\,]$ agrees with the so-called stochastic exponential for Brownian motion.

Now consider the product

$$P_m(t) \coloneqq \prod_{i=1}^{\lfloor mt\rfloor}\left(1 + \frac{X_i}{\sqrt{m}}\right),$$

where X_i are i.i.d. Gaussian random variables. It follows from Theorem 3.8 that $P_m(t)$ converges to $\mathrm{e}^{B_t - t/2}$ where B_t is a standard BM (compare with [DM83]).

Example 3.3. Let $(A_{n,k})_{k\in\mathbb{N}}$ be a sequence of $n \times n$ matrices and assume that $\lim_{m\to\infty} \frac{1}{m} \sum_{k=1}^{m} A_{n,k} = A_n$. Define a triangular array of $n \times n$ matrices $A_{n,k}^{(m)} = A_{n,k}$ for $k \in [m]$ and $m \in \mathbb{N}$. It is easily checked that

$$\frac{1}{m} \sum_{k=1}^{\lfloor mt \rfloor} A_{n,k}^{(m)} \to tA_n, \qquad t \in [0,1].$$

It follows from Theorem 3.8 that

$$P_{n,m}(1) = \left(I_n + \frac{1}{m} A_{n,m}\right) \ldots \left(I_n + \frac{1}{m} A_{n,2}\right) \left(I_n + \frac{1}{m} A_{n,1}\right)$$

converges to e^{A_n} as $m \to \infty$. This recovers the main result in [EH18].

In Chapter 6.3, we will emphasize on the role of μ_n, σ_n as we will take a suitable limit of the matrix-valued process $\mathrm{Texp}[Y_n]$ as we will take $n \to \infty$ and interpret the limit as a process on infinite exchangeable arrays (IEAs). In Chapter 7, we will consider an application in deep learning to make more sense of how does the application of $\mathrm{Texp}[Y_n]$ on an initial state, say $H_{n,0}$ behave when we take $n \to \infty$. We will end this section by describing some examples of popular iterative falling under this class of algorithms and provide argue about their continuous time behavior.

3.3.1 Examples

In the following subsections, we derive the continuous-time limit of some of the popular algorithms and derive immediate conclusions about their continuous time convergence to stationarity.

Oja's Algorithm

In this section, we analyze the Oja's algorithm [Oja82] which is perhaps, the most popular algorithm for Streaming Principle Component Analysis. It is well known that under very mild conditions, given i.i.d. sampled from a distribution, Oja's algorithm asymptotically converges to the top eigenvector of the second moment matrix of the distribution [LWLZ18, HNWW21].

Consider $m \in \mathbb{N}$ i.i.d. samples $\{x_{n,i}\}_{i\in[m]}$ from a distribution over $\mathbb{R}^n$ with second moment $\Sigma_n \succeq 0$. The Oja's algorithm starts with an initialization non-zero vector $p_{n,0} \in \mathbb{R}^d$ computes $q_{n,0} = p_{n,0}/\|p_{n,0}\|_2$ and at every step $k \in [m]$ sets

$$p_{n,k} = \left(I_n + \frac{1}{mn} \cdot n x_{n,k} x_{n,k}^\top\right) q_{n,k-1}, \qquad q_{n,k} = p_{n,k}/\|p_{n,k}\|_2.$$

Now notice that for any $t \in [0,1]$, the $\lfloor mt \rfloor$-th iterate

$$q_{n,\lfloor mt \rfloor} = P_{n,m}(t) \cdot \prod_{k=1}^{\lfloor mt \rfloor} \frac{1}{\|p_{n,k}\|_2} \cdot q_{n,0},$$

where $P_{n,m}(t) \coloneqq \prod_{k=1}^{\lfloor mt \rfloor} \big(I_n + \frac{1}{mn} \cdot n x_k x_k^\top\big)$. By assumption, we have $\mathbb{E}\big[x_{n,k} x_{n,k}^\top\big] = \Sigma_n$ for every $k \in [m]$. Since we are always sampling i.i.d. samples for every $m \in \mathbb{N}$, it also holds that $\lim_{m\to\infty} \Sigma_n = \Sigma_n$ for every $t \in [0,1]$. Applying Theorem 3.8, we find that $(P_{n,m})_{m\in\mathbb{N}}$ uniformly converges to $t \mapsto \mathrm{Texp}\big[\int_0^{\cdot} \Sigma_n\big](t) = \mathrm{e}^{t\Sigma_n}$.

Since the Oja's algorithm re-scales the iterates at every iteration, in the limit as $m \to \infty$, the vector $q_{n,\lfloor mt \rfloor}$ therefore converges to the unit vector corresponding to $\mathrm{e}^{t\Sigma_n} q_0$. Let us call it $q_n(t)$. If Σ_n has an eigen-decomposition $V_n \Lambda_n V_n^\top$ for $\Lambda_n = \mathrm{diag}\big((\lambda_{n,i})_{i=1}^n\big)$ having its diagonals arranged in descending order, and $z_n(t) \coloneqq V_n^\top q_n(t)$, and $\Delta_n \coloneqq \lambda_{n,1} - \lambda_{n,2}$, then

$$\|q_n(t) - v_{n,1}\|_2^2 = \|z_n(t) - e_{n,1}\|_2^2, \tag{3.18}$$

where $e_{n,1}$ is the first element of the standard canonical basis of $\mathbb{R}^n$. Notice that

$$\begin{aligned} z_{n,1}^2(t) &= \frac{z_{n,1}^2(0)}{z_{n,1}^2(0) + \sum_{j=2}^n \mathrm{e}^{-2t(\lambda_{n,1}-\lambda_{n,j})} z_{n,j}^2(0)}, \\ z_{n,i}^2(t) &= \frac{\mathrm{e}^{-t(\lambda_{n,1}-\lambda_{n,i})} z_{n,i}^2(0)}{z_{n,1}^2(0) + \sum_{j=2}^n \mathrm{e}^{-2t(\lambda_{n,1}-\lambda_{n,j})} z_{n,j}^2(0)}, \qquad i \in [n] \setminus \{1\}. \end{aligned} \tag{3.19}$$

Using the fact that $\Big((1+x)^{1/2} - 1\Big)^2 \le x$, for all $x \ge 0$, we find that

$$\frac{1}{n}\|q_n(t) - v_{n,1}\|_2^2 \le 2\frac{\|z_n(0)\|_\infty^2}{z_{n,1}^2(0)} \cdot \mathrm{e}^{-2t\Delta_n}.$$

We find the the mean squared error of the estimation of the top eigenvector goes down exponentially with time and depends on the Δ_n.

Gossip Algorithms

Gossip algorithms are distributed algorithms used to average values over the nodes of a graph. Simple applications arise when certain sensors capture values over a small region or space. To combat minor fluctuations in their readings, the sensors need to average their readings in a distributed manner. Distributed averaging also arises in many applications such as coordination of autonomous agents, estimation and distributed data fusion on ad-hoc networks, and decentralized algorithms.

Let $G_n = ([n], E)$ be an undirected graph with adjacency matrix $A_n \in \{0,1\}^{[n]^{(2)}}$, and let $x_n(0) \in \mathbb{R}^n$ represent the initial values stored on the nodes of the graph. The goal of the gossip algorithm is for each node to estimate $\frac{1}{n}\sum_{i=1}^{n} x_{n,i}(0)$ in a distributed and asynchronous manner. That is, the target is to approximate $\frac{1}{n}1_n 1_n^\top x_n(0)$ over the set of nodes.

Each node has an independent clock that ticks at the times of a rate 1 Poisson process. This corresponds to a single clock ticking at rate n Poisson process at times $\{Z_k\}_{k\in\mathbb{N}}$. Let I_k denote the $[n]$-valued random variable denoting the node whose clock ticked at time Z_k, for every $k \in \mathbb{N}$. Let us denote $[Z_k, Z_{k+1})$ as the k-th time slot for every $k \in \mathbb{N}$. Given A_n, we can compute a matrix P_n such that for every $i \in [n]$, $P_{i,j}$ denotes the probability that node i is connected to node j. That is, if D_n is the diagonal matrix storing the degree of every node on its diagonal, then $P_n = D_n^{-1}A_n$. Let us assume that P_n is doubly stochastic.

The algorithm runs as follows. Let $m \in \mathbb{N}$ be the total number of steps that the algorithm runs in a single round. Let $I_k = i \in [n]$ at the k-th time slot. Then, node i chooses a neighbor $j \in [n]$ with probability $P_{i,j} > 0$ and node i updates its value as

$$x_{n,i}(k+1) = (1-\alpha_m)x_{n,i}(k) + \alpha_m \cdot \frac{1}{d_i}\sum_{j\in N_i} x_{n,j}(k), \qquad k \in \mathbb{Z}_+,$$

where $N_i \subseteq [n]$ denotes the neighbor set of node i, and $|N_i| = d_i$. Let $B_n = I_n + \alpha_m(P_n - I_n)$. Observe that this operation can be written as a linear transformation of $x_n(k)$. That is, $x_n(k+1) = Z_{n,k}x_n(k)$, where $Z_{n,k}$ with probability $1/n$ (for the event when the clock of node i ticks in the k-th time slot) is the matrix that is all zero rows, except the i-th row being the i-th row of B_n. Hence, $\mathbb{E}[Z_{n,k}] = I_n + \frac{\alpha_m}{n}(P_n - I_n)$ for every $k \in [m]$.

Therefore, after time slot $\lfloor nmt \rfloor$, the values stored on the nodes of the algorithm is

$$x_n(\lfloor nmt \rfloor) = \prod_{i=1}^{\lfloor nmt \rfloor} Z_{n,k} \cdot x_n(0), \qquad t \in \mathbb{R}_+.$$

If $\alpha_m = m^{-1}$, then following Theorem 3.8,

$$\prod_{i=1}^{\lfloor nmt \rfloor} Z_{n,k} \to \mathrm{Texp}\left[\int_0^{\cdot} (P_n - I_n)\,\mathrm{d}s\right](t) = \exp(t(P_n - I_n)), \qquad t \in \mathbb{R}_+.$$

Therefore, the vector $x_n(\lfloor nmt \rfloor)$ converges to $\exp(t(P_n - I_n))x_n(0)$ as $m \to \infty$.

Let $(\lambda_{n,i})_{i\in[n]}$ denote the eigenvalues of P_n in descending order. Since P_n is a stochastic matrix, all its eigenvalues are non-negative, moreover $\lambda_{n,1} = 1$ and the corresponding eigenvector is 1_n. We can now compute the limit of the error as $m \to \infty$ as

$$\begin{aligned}\lim_{m\to\infty} \left\| x_n(\lfloor nmt \rfloor) - \frac{1}{n} 1_n 1_n^\top x_n(0) \right\|_2 &= \left\| \exp(t(P_n - I_n))x_n(0) - \frac{1}{n} 1_n 1_n^\top x_n(0) \right\|_2 \\ &\leq \left\| \frac{1}{n} 1_n 1_n^\top - \exp(-t(I_n - P_n)) \right\|_2 \|x_n(0)\|_2.\end{aligned}$$

Since the only non-zero eigenvalue of $\frac{1}{n} 1_n 1_n^\top$ is 1, the operator norm of the matrix $\frac{1}{n} 1_n 1_n^\top - \exp(-t(I_n - P_n))$ is $\exp(t(1 - \lambda_{n,2}))$. Therefore we get that the relative error in the limit $m \to \infty$ is bounded by $\mathrm{e}^{-t(1-\lambda_{n,2})}$.

We find that relative error depends on how small $\lambda_{n,2}$ be. As an example, if G were a random d-regular graph, then it is well known that $\lambda_{n,2}$ converges to $\Theta(d^{-1/2})$ with high probability [FKS89], yielding that as $n \to \infty$, the rate of convergence is $\mathrm{e}^{-T(1-\Theta(d^{-1/2}))}$ with probability 1.

Convergence of Stochastic Gradient Descent (SGD)

Let $(a_i, b_i)_{i\in[m]} \subset \mathbb{R}^n \times \mathbb{R}$ be a set of $m \in \mathbb{N}$ input output pairs of data points. The output if modeled as a linear function of the input, gives us the standard linear regression model, where the objective is to find a suitable linear predictor. If one assumes that $b_i = n^{-1}\langle a_i, x^* \rangle + \epsilon_i$ for $i \in [m]$ for some $x^* \in \mathbb{R}^n$ such that ϵ_is are all i.i.d. centered random variables independent

of $(a_i)_{i\in[m]}$, then the objective is to solve the minimization problem

$$\arg\min_{x\in\mathbb{R}^n} \frac{1}{2m}\sum_{i=1}^{m}\big(n^{-1}\langle a_i, x\rangle - b_i\big)^2.$$

Let us consider the particle SGD algorithm [Chi22] that starts at $x_1 \in \mathbb{R}^n$ and at iteration $k \in \mathbb{N}$, computes x_{k+1} as

$$x_{k+1} = x_k - n\eta_{n,m,k}a_k\big(n^{-1}\langle a_k, x_k\rangle - b_k\big) = \Big(I_n - \frac{\eta_m}{n}\cdot na_ka_k^\top\Big)x_k + n\eta_m a_k b_k.$$

Unrolling the expression, following the popular choice [Net19], if we set $\eta_{n,m,k} = \frac{1}{\lambda_{k,n}m}$, where $\lambda_{k,n} > 0$ is the least eigenvalue of $\mathbb{E}\big[a_ka_k^\top\big]$ for every k, we get that at iteration $\lfloor mt\rfloor$ for any $t \in [0,1]$,

$$x_{\lfloor mt\rfloor} = \prod_{k=1}^{\lfloor mt\rfloor}\left(I_n - \frac{na_ka_k^\top}{\lambda_{k,n}mn}\right)\cdot x_1 + \frac{n}{m}\sum_{k=1}^{\lfloor mt\rfloor}\prod_{j=k+1}^{\lfloor mt\rfloor}\left(I_n - \frac{na_ja_j^\top}{\lambda_{j,n}mn}\right)\cdot\frac{a_k}{\lambda_{k,n}}(n^{-1}a_k^\top x^* + \epsilon_k). \tag{3.20}$$

Let us define the piecewise constant interpolation of $\big(\mathbb{E}\big[a_ia_i^\top\big]\big)_{i\in[m]}$ as $s \mapsto \Sigma_{n,m}(s)$ and assume that Σ_m uniformly converges to a curve Σ_n as $m \to \infty$. Similarly define a, ϵ, λ_n to the limit of the piecewise continuous interpolation of $(a_i)_{i\in[m]}$, $(\epsilon_i)_{i\in[m]}$ and $(\lambda_{i,n})_{i\in[m]}$ respectively as $m \to \infty$.

Then, following Theorem 3.8, if $(t \mapsto x_{\lfloor mt\rfloor}) \to x$ as $m \to \infty$, then

$$\begin{aligned} x(t) = \operatorname{Texp}\left[-\int_0^{\cdot}\left(\frac{\Sigma_n}{\lambda_n}\right)(s)\,\mathrm{d}s\right](t)x(0) \\ + \int_0^t \operatorname{Texp}\left[-\int_0^{\cdot}\tau_s\left(\frac{\Sigma_n}{\lambda_n}\right)(r)\,\mathrm{d}r\right](s)\frac{a(s)a(s)^\top}{\lambda_n(s)}x^*\,\mathrm{d}s \\ + n\int_0^t \operatorname{Texp}\left[-\int_0^{\cdot}\tau_s\left(\frac{\Sigma_n}{\lambda_n}\right)(r)\,\mathrm{d}r\right](s)\frac{a(s)}{\lambda_n(s)}\epsilon(s)\,\mathrm{d}s. \end{aligned}$$

Taking expectation over a and ϵ, we get

$$\begin{aligned} \mathbb{E}[x(t)] = \operatorname{Texp}\left[-\int_0^{\cdot}\left(\frac{\Sigma_n}{\lambda_n}\right)(s)\,\mathrm{d}s\right](t)x(0) \\ + \int_0^t \operatorname{Texp}\left[-\int_0^{\cdot}\tau_s\left(\frac{\Sigma_n}{\lambda_n}\right)(r)\,\mathrm{d}r\right](s)\left(\frac{\Sigma_n}{\lambda_n}\right)(s)x^*\,\mathrm{d}s. \end{aligned}$$

In the simple case when Σ and λ_n are the constant curve of a positive definite matrix and its smallest eigenvalue respectively, the above equation reduces to

$$\begin{aligned}\mathbb{E}[x(t)] &= \operatorname{Texp}\left[-\int_0^{\cdot} \frac{\Sigma_n}{\lambda_n}\,\mathrm{d}s\right](t)x(0) + \int_0^t \operatorname{Texp}\left[-\int_0^{\cdot} \frac{\Sigma_n}{\lambda_n}\,\mathrm{d}r\right](s)\frac{\Sigma_n}{\lambda_n}\,\mathrm{d}s \cdot x^* \\ &= \mathrm{e}^{-t\Sigma_n/\lambda_n}x(0) + \int_0^t \mathrm{e}^{-s\Sigma_n/\lambda_n}x^*\,\mathrm{d}s = \mathrm{e}^{-t\Sigma_n/\lambda_n}x(0) + (I_n - \mathrm{e}^{-t\Sigma_n/\lambda_n})x^*,\end{aligned}$$

which implies

$$\mathbb{E}[x(t)] - x^* = \mathrm{e}^{-t\Sigma_n/\lambda_n}(x(0) - x^*). \tag{3.21}$$

In other words, $\mathbb{E}[x(t)]$ converges exponentially fast to the minimizer x^* of the optimization problem. This rate, as one might expect depends on the condition number of the second moment matrix. Note that equation (3.21) captures the general case when the sequence of input output pairs have time-varying distribution such that the second moment of the distribution of the input converges to an L^2-absolutely continuous curve under the chosen scaling.

Applications in financial modeling

The iterated product of matrices obtained from a triangular array is particularly important when there is a fixed notion of continuous time, and measurements are made at various levels of discretization. In finance modeling, m may denote the number of intervals within a year, ranging from trading days to microseconds of annual trading activity. Here, n represents the count of (dependent) financial instruments, often numbering in tens of thousands. To examine changes in an instrument's price from time-step $k-1$ to k, consider the price of the i-th instrument, $H_{n,k,i}$, written in the form:

$$H_{n,k-1,i} + \frac{1}{m}\cdot\left(\frac{1}{n}\sum_{j=1}^{n} M^{(m)}_{n,k,(i,j)}H_{n,k-1,j}\right) + \frac{1}{\sqrt{m}}\cdot\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n} G^{(m)}_{n,k,(i,j)}H_{n,k-1,j}\right). \tag{3.22}$$

The above expression indicates that the price $H_{n,k,i}$ of instrument i at every time step $k \in [m]$, increases at a rate influenced by a linear combination of the (noisy) growth rates $M^{(m)}_{n,k,(i,j)}$ of

all instruments $j \in [n]$. Additional noise is contingent upon the price of all other instruments. Since there is an absolute notion of time $t \in [0, 1]$ where $t = 0$ and $t = 1$ indicate the start and end of a financial year, our analysis establishes a uniform limiting framework applicable across all trading frequencies. Essentially, the evolution of the price of n (or possibly infinite) financial instruments is dictated by the curve $t \mapsto A_n(t)$, representing the continuous time-varying return. Examples of such models include the dynamics of monetary reserves of n banks interacting via lending mechanisms [RJPLH15].

It is easy to see that when $n = 1$, and A_1 is a constant curve, we recover the classical geometric Brownian motion.

3.4 Conclusion

In this chapter, we derive the fixed-dimension continuous-time limits of the three iterative algorithms described in Chapter 2.2. To take the limit as the dimensions increase, we must first better understand how stochastic processes on a large number of coordinates can be described compactly and analytically. We then move to Chapter 4, where we develop a calculus on graphons (see Chapter 2.5) and argue that curves such as Euclidean gradient flows can be suitably interpreted within the space of graphons, and moreover, they converge to well-defined gradient flows on graphons. With further development of the analytical space and the topology over these exchangeable arrays, we will be prepared in Chapter 6 to discuss the scaling limits of these finite-dimensional processes as the dimensionality increases to infinity.

Chapter 4

GRADIENT FLOW ON GRAPHONS

In this chapter, we will develop a theory of gradient flows on the space of limits of dense edge-weighted, unlabeled graphs, known as graphons. We will show that gradient flows on finite-dimensional matrices of permutation-invariant functions can be suitably embedded in the space of graphons, where these curves converge to gradient flows as the dimensions of the matrices increase to infinity. This analysis acts as a building block for us to later characterize scaling limits of iterative processes obtained via algorithms discussed in Section 1.1.

We will start in Section 4.1 by introducing the analogous notion of gradient flows developed for classical particle systems to motivate our developments. In Section 4.2 we will introduce some background necessary for the developments later in this chapter. With the help of the general theory of gradient flows on metric spaces [AGS08], in Section 4.3, we will discuss the construction and argue for the existence of such flows on the metric space of graphons. In Section 4.3.3, we will introduce a notion of derivative on graphons and relate it to the Euclidean gradient. In Section 4.4, we will then show that gradient flows of permutation-invariant functions on finite-dimensional matrices converge to a gradient flow on graphons under certain consistency conditions. In Section 4.5 we will show that just like Wasserstein gradient flows can be described with the help of continuity equations [San15, Chapter 5], we will demonstrate that gradient flows on graphons can be described by a consistent family of continuity equations. We will conclude in Section 4.6 by discussing some examples and deriving implications on how a gradient flow description allows us to determine the time rate of convergence to first order stationary points of the objective function.

4.1 Introduction

Let $x_1, x_2, \ldots, x_n$ be n vectors in $\mathbb{R}^d$. Let $f_n \colon \left(\mathbb{R}^d\right)^n \to \mathbb{R}$ be a permutation invariant function of those n variables. Here *permutation invariant* means $f_n(x_1, \ldots, x_n) = f_n(x_{\pi_1}, \ldots, x_{\pi_n})$ where π is any permutation of the set $[n] \coloneqq \{1, 2, \ldots, n\}$. Consider the Cauchy problem

$$\frac{\mathrm{d}}{\mathrm{d}t} x_i(t) = -\,\partial_i f_n(x_1(t), \ldots, x_n(t)), \quad i \in [n], \qquad t \in \mathbb{R}_+, \tag{4.1}$$

with given initial conditions $(x_i(0))_{i\in[n]}$. The solution to this problem (which exists and is unique when, say, ∇f_n is Lipschitz [Lin94]) is often called the gradient flow of f_n. A natural question that appears in several applications is whether the solution to the above Cauchy problem has a *scaling limit* as n goes to infinity.

In order for such a limit to exist, it is imperative that there is some consistency over the dimension parameter n. The permutation invariance of f_n offers a resolution. In fact, if $(x_i)_{i\in[n]}$ is thought of as positions of particles in space, f_n can be thought of as a function acting on the empirical distribution $\mu_n \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Here μ_n is a discrete probability measure that puts mass $1/n$ on the positions of each of the n particles, represented by the delta mass $\delta_\cdot$.

Consider $\mathbb{R}^d$ with the usual Euclidean metric and let $F \colon \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ be a suitable function. The function F induces a sequence of permutation invariant functions $(f_n)_{n\in\mathbb{N}}$ as above by the definition

$$f_n\left(x_1, \ldots, x_n\right) \coloneqq F\left(\mu_n\right), \quad n \in \mathbb{N}.$$

For such an f_n for any $n \in \mathbb{N}$, the evolution (4.1) can be thought of as an evolution on the space of probability measures by defining

$$\mu_n(t) \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}, \quad t \in \mathbb{R}_+.$$

Now the following question makes sense. Suppose that the sequence of initial measures $(\mu_n(0))_{n\in\mathbb{N}}$ *converges* to a limiting probability measure $\mu(0)$ where the convergence is typically in the sense of weak convergence of probability measures. Does the sequence of curves $\left((\mu_n(t))_{t\in\mathbb{R}_+}\right)_{n\in\mathbb{N}}$ converge to some limiting curve on $\mathcal{P}(\mathbb{R}^d)$ possibly after rescaling time?

The answer to the above, under suitable assumptions on F, is the so-called Wasserstein gradient flow [Vil03, San15] of F on the metric space $\mathcal{P}\left(\mathbb{R}^d\right)$ equipped with the Wasserstein-2 metric, $\mathbb{W}_2$. There is now a general theory of curves of maximal slopes (also known as gradient flows) developed for functions on metric spaces which may lack a differentiable structure [AGS08]. The Wasserstein space is a prominent example that has been thoroughly studied [AGS08, San17]. As discussed in Chapter 1, there has been a recent surge in interest in the application of the above convergence of gradient flows in the context of single hidden layer neural networks, see [SMN18, CB18, RVE18, MMM19, CCP19, AOY19, NP20, SS20a, SS20b, TR20, BC21].

However, we are interested in optimization problems where the arguments can be thought of as weights attached to the edges of a large dense graph. Let $G = ([n], E)$ be a graph. For $e = \{i, j\} \in E$ one has an associated variable $X_{n,(i,j)} = X_{n,(j,i)}$ that we take it to be real-valued in this article. For all the applications we consider, we can take $X_e = 0$ if $e \notin E$. Thus our variables can be arranged in an $n \times n$ symmetric matrix X_n. Let $R_n \colon \mathcal{M}_n(\mathbb{R}) \to \mathbb{R} \cup \{\infty\}$ be a function of such matrices. The crucial difference from the previous set-up is that we want R_n to satisfy a permutation invariance property with respect to relabeling the vertices of G as described in Chapter 1 in Definition 1.1. In other words, such functions are invariant under graph isomorphisms of G. One can now ask the same question as before. Consider the gradient flow Cauchy problem

$$\frac{\mathrm{d}}{\mathrm{d}t} X_{n,e}(t) = -\,\partial_e R_n\left(X_n(t)\right), \quad e \in [n]^2, \tag{4.2}$$

with given $X_n(0)$. Is there a suitable scaling limit as n goes to infinity? This paper answers this question in affirmative under reasonable conditions on R_n.

We restrict ourselves to the case where the edge weights X_n all lie in the bounded interval $[-1, 1]$. Without loss of generality, we can take our graph to be the complete graph with its weighted adjacency matrix X_n. Just like empirical distributions of particle systems converge to probability measures, these graph adjacency matrices with bounded edge weights, identified up to graph isomorphisms, converge to a limiting object called

graphons [LS06, BCL$^+$08, BCL$^+$12]. This is intimately connected with the theory of exchangeable arrays in probability theory [Ald81, Ald82, Hoo82, Kal89]. For a definitive modern account of exchangeable arrays and their connections with the limits of large graphs, the reader is referred to [DJ08, Aus08, Aus12, Aus15]. The theory of graph limits and graphons has found applications in many fields including extremal graph theory, combinatorics, data analysis, biology. We refer the reader to [DGKR15, LZ17, BG20, BEFLY21] and references therein for further details.

Our gradient flow of R on graphons will be with respect to the δ_2 metric defined in Definition 2.12. The δ_2 metric shall play a similar role as the $\mathbb{W}_2$ metric does on probability measures. The metric space $(\widehat{\mathcal{W}}, \delta_2)$ is a geodesic space (Proposition 4.13). Our gradient flows will be with respect to this metric. The other, called the cut metric, $\delta_\square$, is more traditional [BCL$^+$08, Lov12] and is important here for the topology that it generates and will play the role of the metric of weak convergence of probability measures. Thus, although our gradient flows will be defined with respect to the invariant L^2 metric δ_2, the various convergence statements will be with respect to the cut metric $\delta_\square$.

We need a set-up that is similar to particle systems and their limiting probability measures. Every symmetric matrix in $\mathcal{M}_n$ identified up to the same permutation on rows and columns can be embedded in the space of block graphons $\widehat{\mathcal{W}}_k \subseteq \widehat{\mathcal{W}}$ (see Section 2.5 for details). Thus, any function $R\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ induces a sequence of functions $\big(R\mid_{\widehat{\mathcal{W}}_n}\colon \widehat{\mathcal{W}}_n \to \mathbb{R} \cup \{\infty\}\big)_{n\in\mathbb{N}}$, by restriction, and a sequence of invariant functions $(R_n\colon \mathcal{M}_n \to \mathbb{R} \cup \{\infty\})_{n\in\mathbb{N}}$.

4.2 Background and Preliminaries

In this Section, we will complement the background already provided in Section 2.5 that will be essential for the completeness of this chapter.

Let $R\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ be a function. We define the effective domain eff-Dom(R) of $\mathbb{R}$ as the set $\{[u] \in \widehat{\mathcal{W}} \mid R([u]) < \infty\}$.

For every permutation $\pi \in S_n$ we can define an invertible Lebesgue measure preserving

map $\tilde{\pi}\colon [0,1] \to [0,1]$ such that $\tilde{\pi}$ is an increasing affine homeomorphism from $Q_{n,i}$ to $Q_{n,\pi(i)}$ for each $i \in [n]$. We denote the set of all such maps by the set $\mathcal{I}_n$. Following this, Definition 2.11 and [BCL$^+$08, Lemma 3.5], we have another equivalent definition which will be useful for us, i.e.,

$$\delta_\square([w_0],[w_1]) = \lim_{n\to\infty} \min_{\tilde{\pi}\in\mathcal{I}_n} \left\| w_0 - w_1^{\tilde{\pi}} \right\|_\square, \qquad [w_0],[w_1] \in \widehat{\mathcal{W}}. \tag{4.3}$$

We denote the metrics induced by the cut norm and the L^2-norm as $d_\square$ and d_2 respectively. The space $(\widehat{\mathcal{W}}, \delta_\square)$ is a compact metric space [LS07], [Lov12, Section 9.3] while the metric space $(\widehat{\mathcal{W}}, \delta_2)$ is complete and separable, but not compact. It is clear that convergence in δ_2 implies the convergence in $\delta_\square$, that is, the topology generated by $\delta_\square$ is weaker than the topology generated by δ_2. The following Lemma says that the metric δ_2 is lower semicontinuous with respect to $\delta_\square$.

Lemma 4.1. *[Lov12, Lemma 14.16] The metric δ_2 is sequentially $\delta_\square$-lower semicontinuous, i.e., if sequences $([u_n])_{n\in\mathbb{N}}, ([v_n])_{n\in\mathbb{N}} \subset \widehat{\mathcal{W}}$, and $[u],[v] \in \widehat{\mathcal{W}}$ such that $([u_n])_{n\in\mathbb{N}} \xrightarrow{\delta_\square} [u]$ and $([v_n])_{n\in\mathbb{N}} \xrightarrow{\delta_\square} [v]$, then*

$$\liminf_{n\to\infty} \delta_2([u_n],[v_n]) \geq \delta_2([u],[v]).$$

Definition 4.2 (Extensions to L^p kernels for $p \in [1,\infty]$)**.** Sometimes in our text, we will consider kernels and matrices whose entries are not necessarily in $[-1,1]$, but are rather elements in $L^2\big([0,1]^{(2)}\big)$ or $L^\infty\big([0,1]^{(2)}\big)$. For any $n \in \mathbb{N}$, just like we defined $\mathcal{W}_n$, we can restrict our attention to the subset of functions $L^p_n\big([0,1]^{(2)}\big) \subset L^p\big([0,1]^{(2)}\big)$ for every $p \in [1,\infty]$ which contain symmetric measurable step functions over $Q_n \times Q_n$. Using the equivalence relation $\cong$, just like we defined $\widehat{\mathcal{W}}$ and $\widehat{\mathcal{W}}_n$, we can similarly define $\widehat{L}^p\big([0,1]^{(2)}\big) \coloneqq L^p\big([0,1]^{(2)}\big)/\cong$ and $\widehat{L}^p_n\big([0,1]^{(2)}\big) \coloneqq L^p_n\big([0,1]^{(2)}\big)/\cong$ for any $p \in [1,\infty]$. When it is clear from the context, we will also call the elements in $\widehat{L}^\infty$ graphons. For simplicity, we use K and M_n for $n \in \mathbb{N}$ even when the kernels are in $L^p\big([0,1]^{(2)}\big)$ for $p \in [1,\infty]$.

4.2.1 Gradient flows on metric spaces

The theory of gradient flow on a general metric space is well-developed by now and can be found in [AGS08]. Since our goal is to define gradient flows on $(\widehat{\mathcal{W}}, \delta_2)$, the definitions below are sometimes not the most general versions as given in [AGS08] but adapted to our particular setting.

Definition 4.3 (Metric derivative)**.** For a metric space (X, d), and any $T \in \mathbb{R}_+$, the *metric derivative* $|\omega'|(t)$ of a curve $\omega = (\omega_t)_{t\in[0,T]}$ in X at $t \in (0, T)$ is defined as

$$|\omega'|(t) \coloneqq \lim_{s\to t} \frac{d(\omega_s, \omega_t)}{|s - t|}, \tag{4.4}$$

provided this limit exists.

If $\omega \in \mathrm{AC}(X, d)$, then the limit in equation (4.4) exists for a.e. $t \in (0, T)$ and $|\omega'| \in L^1([0, T])$ [AGS08, Theorem 1.1.2]. In other words, every absolutely continuous curve in a metric space has metric derivative defined almost everywhere. And conversely, if the metric derivative $|\omega'|(t)$ exists for a.e. $t \in (0, T)$ and $|\omega'| \in L^1([0, T])$, then ω is absolutely continuous.

We now need to define some notion for the derivative of a function $F\colon X \to \mathbb{R} \cup \{\infty\}$. On a metric space the usual notion of derivative can not be defined. However, the following [AGS08, Definition 1.2.4] acts as a substitute in many situations of interest.

Definition 4.4 (Local slope)**.** The *local slope* $|\partial R|(v)$ of $R\colon X \to \mathbb{R} \cup \{+\infty\}$ on a metric space (X, d), at $v \in \text{eff-Dom}(R)$ is defined as

$$|\partial R|(v) \coloneqq \limsup_{w\in X,\, d(v,w)\to 0} \frac{(R(v) - R(w))^+}{d(v, w)}. \tag{4.5}$$

The definition below is narrower than the one in [AGS08, Definition 1.3.2] since we restrict our choice of *upper gradient* in that definition to the local slope [AGS08, Theorem 1.2.5].

Definition 4.5 (Curves of maximal slope)**.** On a metric space (X, d), any locally absolutely continuous curve $\omega = (\omega_t)_{t\in[0,T]}$ in X on a finite time horizon $T > 0$ is a *curve of maximal*

slope for the function $R\colon X \to \mathbb{R} \cup \{+\infty\}$ with respect to its local slope, if $R \circ \omega = G$ a.e. for some non-increasing map G on $(0,T)$, and

$$G'(t) \leq -\frac{1}{2}|\omega'|^2(t) - \frac{1}{2}|\partial R|^2(\omega_t), \qquad \text{a.e.} \quad t \in (0,T). \tag{4.6}$$

On a general metric space, a curve of maximal slope can be referred to as a gradient flow although the concept of gradient itself is absent. See [AGS08, Section 1.3] for the intuition.

4.2.2 Geodesic metric structure on Graphons

In this section we will show that $\left(\widehat{\mathcal{W}}, \delta_2\right)$ is a geodesic metric space. Before stating this formally, we first introduce some definitions that make this concept morr formal.

Definition 4.6 (Length)**.** Given the metric space (X,d), and a curve $\omega = (\omega_t)_{t\in[0,T]}$ in X, the *length* of ω is defined as

$$\ell(\omega) \coloneqq \sup\left\{ \sum_{k=0}^{n-1} d\big(\omega_{t_k}, \omega_{t_{k+1}}\big) \;\middle|\; n \in \mathbb{N}, 0 = t_0 < t_1 < \cdots < t_n = T \right\}.$$

It is clear from Definition 4.6 that for any absolutely continuous curve $\omega = (\omega_t)_{t\in[0,T]}$ in X and $x, y \in X$ such that $\omega_0 = x$, $\omega_T = y$, we have $\ell(\omega) \geq d(x,y)$. Given $x, y \in X$ it is natural to ask if there is an absolutely continuous curve ω from x to y that achieves the length $\ell(\omega) = d(x,y)$. Such a curve is called a *geodesic* between x and y. If there exists a geodesic ω between any two points $x, y \in X$, we say that (X,d) is a geodesic metric space. In a geodesic metric space, notions like convexity and semiconvexity make sense. We make those precise in the following definitions.

Definition 4.7 (Geodesic metric space)**.** A metric space (X,d) is called a *geodesic metric space* if for all $x, y \in X$

$$d(x,y) = \min\{\ell(\omega) \mid \omega \in \mathrm{AC}(X,d), \omega_0 = x, \omega_1 = y\} \ .$$

Definition 4.8 (Constant speed geodesics)**.** On a metric space (X,d), a curve $\omega = (\omega_t)_{t\in[0,1]}$ in X is a *constant speed geodesic* if for all $0 \leq r \leq s \leq 1$,

$$d(\omega_r, \omega_s) = d(\omega_0, \omega_1)(s - r). \tag{4.7}$$

Note that if a curve ω satisfies equation (4.7), then ω is clearly Lipschitz and hence absolutely continuous. It is easy to see that such a curve ω is indeed a geodesic and the metric derivative $|\omega'|(t) = d(\omega_0, \omega_1)$ for a.e. $t \in [0, 1]$. This justifies the name 'constant speed geodesic'.

Remark 4.9. It is also worth pointing that not only every geodesic, but every absolutely continuous curve can be reparametrized so that it becomes Lipschitz [San15, Box 5.1] under the new parametrization.

We now make precise the notion of convexity in metric spaces. On a metric space, we first define convexity (and semiconvexity) along curves. If a function is convex (or semiconvex) along every constant speed geodesic, then we call it convex with respect to the metric.

Definition 4.10 (λ-semiconvexity along curves w.r.t. a metric)**.** On a metric space (X, d), a function $R\colon X \to \mathbb{R} \cup \{\infty\}$ is said to be λ-*semiconvex* with respect to the metric d along a curve $\omega = (\omega_t)_{t\in[0,1]}$ in X for some $\lambda \in \mathbb{R}$, if

$$R(\omega_t) \leq (1-t)R(\omega_0) + tR(\omega_1) - \frac{1}{2}\lambda t(1-t)d^2(\omega_0, \omega_1), \tag{4.8}$$

for all $t \in [0, 1]$. Particularly, if the above inequality holds for $\lambda = 0$, then we say that R is convex with respect to the metric d along the curve ω.

Definition 4.11 (λ-geodesic semiconvexity w.r.t. a metric)**.** On a metric space (X, d), a function $R\colon X \to \mathbb{R} \cup \{\infty\}$ is λ-*geodesically semiconvex* with respect to the metric d, if for any $v_0, v_1 \in \text{eff-Dom}(R)$ there exists a constant speed geodesic $\omega = (\omega_t)_{t\in[0,T]}$ on (X, d) (Definition 4.8) with $\omega_0 = v_0$ and $\omega_1 = v_1$ such that R is λ-semiconvex on ω with respect to the metric d for some $\lambda \in \mathbb{R}$ (Definition 4.10).

Now we are ready to provide some statements that are used in the proof of our main results but are also of independent interest. The two key results in this section are Lemma 4.12 and Proposition 4.13. Proposition 4.13 states that $(\widehat{\mathcal{W}}, \delta_2)$ is a geodesic metric space.

Lemma 4.12. *The invariant L^2 metric between two graphons $[u], [v] \in \widehat{\mathcal{W}}$ satisfies*

$$\delta_2([u],[v]) = \min \int_r^s \|w_t'\|_2 \, \mathrm{d}t, \tag{4.9}$$

for any $0 \le r < s \le 1$, where the minimum is taken over $(w_t)_{t\in[r,s]} \in \mathrm{AC}(\mathcal{W}, d_2)$ with domain $[r, s]$ such that $w_r \in [u]$ and $w_s \in [v]$.

As a consequence of above statements, we obtain that $(\widehat{\mathcal{W}}, \delta_2)$ is a geodesic space. To the best of our knowledge it has not been recorded in the earlier literature.

Proposition 4.13 ([OPST23])**.** *The space $(\widehat{\mathcal{W}}, \delta_2)$ is a geodesic metric space.*

The proofs of Lemma 4.12 and Proposition 4.13 have been provided in Appendix B.1.

Since $(\widehat{\mathcal{W}}, \delta_2)$ is a geodesic metric space, the usual notions of geodesic convexity and semiconvexity makes sense in $(\widehat{\mathcal{W}}, \delta_2)$. In the subsequent sections, we will need a notion of generalized geodesics (defined below) and show that generalized geodesics exist.

Definition 4.14 (Generalized geodesics on $(\widehat{\mathcal{W}}, \delta_2)$)**.** Let $[w_0], [w_1] \in \widehat{\mathcal{W}}$. For every $[w] \in \widehat{\mathcal{W}}$, one can construct an absolutely continuous curve ϑ (depending on $[w]$) as follows. From Lemma B.1, we obtain $\varphi, \varphi_0, \varphi_1 \in \mathcal{T}$ such that

$$\delta_2([w],[w_0]) = \|w^{\varphi} - w_0^{\varphi_0}\|_2, \quad \text{and} \quad \delta_2([w],[w_1]) = \|w^{\varphi} - w_1^{\varphi_1}\|_2. \tag{4.10}$$

Define the curve $\vartheta := ([w_t])_{t\in[0,1]}$, where $w_t := (1-t)w_0^{\varphi_0} + tw_1^{\varphi_1}$ for every $t \in [0, 1]$. This curve ϑ is called a *generalized geodesic* (*with base* $[w]$) between the graphons $[w_0]$ and $[w_1]$ with respect to δ_2. Often, when the base is clear from the context, we simply refer it as a generalized geodesic. From the construction, we can see that any geodesic between $[w_0], [w_1] \in \widehat{\mathcal{W}}$ is also a generalized geodesic (with suitably chosen base) between them.

Finally, we show that $\delta([w], \cdot\,)$ is generalized geodesically semiconvex for every $[w] \in \widehat{\mathcal{W}}$.

Lemma 4.15. *If $[w], [w_0], [w_1] \in \widehat{\mathcal{W}}$, then there exists $\vartheta = (\vartheta_t)_{t\in[0,1]} \in \mathrm{AC}(\widehat{\mathcal{W}}, \delta_2)$ such that $\vartheta_0 = [w_0]$, $\vartheta_1 = [w_1]$, and $\delta_2^2([w], \cdot\,)/2$ is 1-semiconvex over ϑ w.r.t. δ_2.*

The proof of Lemma 4.15 is provided in Appendix B.1.

4.3 Construction and Existence of Gradient Flows

As discussed in Section 4.2.1, we can define what are called *curves of maximal slope*, that are also known as *gradient flows* of a function under a certain set of conditions [AGS08]. In this section, we begin by defining their construction in Section 4.3.1, followed by providing sets of conditions under which they exist in Section 4.3.2. Finally in Section 4.4 we provide conditions for when finite dimensional gradient flows, each defined on $\widehat{\mathcal{W}}_n$, $n \in \mathbb{N}$, converge to a gradient flow on $\widehat{\mathcal{W}}$.

4.3.1 Construction of Gradient Flows

In order to obtain a curve of maximal slope of a function $R\colon \widehat{\mathcal{W}} \to \mathbb{R}$ starting from a graphon $[u_0] \in \widehat{\mathcal{W}}$, we use the iterative implicit Euler scheme. Given a step size $\tau > 0$, consider the potential function Φ_R of R, defined as

$$\Phi_R(\tau, [u]; \cdot\,) \coloneqq R + \frac{1}{2\tau}\delta_2^2(\cdot\,, [u]), \tag{4.11}$$

and the set-valued resolvent operator J_τ defined as $J_\tau([u]) \coloneqq \arg\min_{\widehat{\mathcal{W}}} \Phi_R(\tau, [u]; \cdot\,)$ for all $[u] \in \widehat{\mathcal{W}}$. For a sequence $\boldsymbol{\tau} \coloneqq (\tau_k)_{n\in\mathbb{N}}$ of positive time steps with $|\boldsymbol{\tau}| \coloneqq \sup_{n\in\mathbb{N}} \tau_k < \infty$, we can associate a partition of the time interval $(0, \infty)$ as

$$P_{\boldsymbol{\tau}} \coloneqq \left\{ I_{\boldsymbol{\tau}}^k \coloneqq (t_{\boldsymbol{\tau}}^{k-1}, t_{\boldsymbol{\tau}}^k] \right\}_{n\in\mathbb{N}}, \qquad \tau_k = t_{\boldsymbol{\tau}}^k - t_{\boldsymbol{\tau}}^{k-1},$$

if $t_{\boldsymbol{\tau}}^0 = 0$ and $\lim_{k\to\infty} t_{\boldsymbol{\tau}}^k = \infty$. Given such a sequence $\boldsymbol{\tau}$ and starting from $[u_{\boldsymbol{\tau},0}] \in \widehat{\mathcal{W}}$, one can now obtain a sequence of graphons $([u_{\boldsymbol{\tau},n}])_{n\in\mathbb{N}}$ by an iterative minimization, i.e., by setting $[u_{\boldsymbol{\tau},n}] \in J_{\tau_n}([u_{\boldsymbol{\tau},n-1}])$, provided $J_{\tau_n}([u_{\boldsymbol{\tau},n-1}]) \neq \emptyset$ for all $n \in \mathbb{N}$.

Given such a sequence of iterates $([u_{\boldsymbol{\tau},n}])_{n\in\mathbb{N}}$, we can consider its piecewise constant interpolation to obtain a curve $\overline{[u_{\boldsymbol{\tau}}]}\colon \mathbb{R}_+ \to \widehat{\mathcal{W}}$ called the *discrete solution*. Given a sequence of step sizes $(\boldsymbol{\tau}_k)_{n\in\mathbb{N}}$ with $\lim_{k\to\infty}|\boldsymbol{\tau}_k| = 0$, the discrete solution gives a reasonable candidate for a curve of maximal slope of the function R on $(\widehat{\mathcal{W}}, \delta_2)$.

4.3.2 Existence of Gradient Flows

In this section, we will consider the suggested construction of gradient flows from Section 4.3.1 and provide conditions when this actually can be done. We first note some definitions using which we will present our first existence result.

Definition 4.16 (Generalized minimizing movements)**.** For a function R, its corresponding functional Φ_R as defined in equation (4.11), and an initial datum $[u_0] \in \widehat{\mathcal{W}}$, we say that a curve $\omega = (\omega_t)_{t\in\mathbb{R}_+}$ in $\widehat{\mathcal{W}}$ is a *generalized minimizing movement* (GMM) for Φ_R starting from $[u_0] \in \widehat{\mathcal{W}}$ if there exists a sequence of sequences $(\boldsymbol{\tau}_k)_{k\in\mathbb{N}}$ with $\lim_{k\to\infty}|\boldsymbol{\tau}_k| = 0$ and a corresponding sequence of discrete solutions $(\overline{[u_{\boldsymbol{\tau}_k}]})_{n\in\mathbb{N}}$ such that for all $t \in \mathbb{R}_+$,

$$\begin{gathered}\lim_{k\to\infty} R([u_{\boldsymbol{\tau}_k,0}]) = R([u_0]), \quad \limsup_{k\to\infty} \delta_2([u_{\boldsymbol{\tau}_k,0}], [u_0]) < \infty, \\ \delta_{\square}\text{-}\lim_{k\to\infty} \overline{[u_{\boldsymbol{\tau}_k}]}(t) = \omega_t.\end{gathered} \tag{4.12}$$

There is a related definition of *minimizing movement* (MM) curves that can be found in [AGS08, Definition 2.0.6] where the conditions in equation (4.12) need to hold for all sequences of partitions with vanishing norm. The set of all minimizing movements and generalized minimizing movements on the metric space $(\widehat{\mathcal{W}}, \delta_2)$ with respect to the metric $\delta_{\square}$ starting from $[u_0] \in \text{eff-Dom}(F)$ are denoted by $\text{MM}_{\delta_2,\delta_{\square}}(\Phi_R, [u_0])$ and $\text{GMM}_{\delta_2,\delta_{\square}}(\Phi_R, [u_0])$ respectively. From their definitions it can be verified that the set of minimizing movements is contained in the set of generalized minimizing movements. See [AGS08, Definition 2.0.6] for the precise difference between them. Since $(\widehat{\mathcal{W}}, \delta_2)$ is a bounded metric space, the second conditions in equation (4.12) and [AGS08, equation 2.0.10] are trivially satisfied.

The existence of the curve of maximal slope is dealt in detail in [AGS08, Chapter 2]. Under certain topological compatibility conditions of the metric δ_2, we can distill out our first theorem for existence of gradient flows on $\widehat{\mathcal{W}}$.

Theorem 4.17 (Existence of curves of maximal slope-I [OPST23])**.** *Suppose* $R\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ *satisfies the following conditions.*

1. *R is $\delta_{\square}$-lower semicontinuous on* eff-Dom(R).

2. *Its local slope* $|\partial R|$ *is* $\delta_\square$*-lower semicontinuous in* eff-Dom(R).

3. R *is* $\delta_\square$*-continuous on the sublevel sets of* $|\partial R|$.

Then every curve $\omega \in \mathrm{GMM}_{\delta_2,\delta_\square}(\Phi_R, [u_0])$ *for* $[u_0] \in$ eff-Dom(R) *is a curve of maximal slope.*

In practice, it is difficult to compute $|\partial R|$ or to ascertain its $\delta_\square$-lower semicontinuity. This makes it difficult to apply Theorem 4.17 on natural examples. Later in Theorem 4.23 we show that, when $R\colon \mathcal{W} \to \mathbb{R} \cup \{\infty\}$ admits a Fréchet-like derivative that is λ-semiconvex on $(\mathcal{W}, d_2)$ for some $\lambda \in \mathbb{R}$, the existence of a curve of maximal slope follows without requiring $\delta_\square$-lower semicontinuity of $|\partial R|$.

4.3.3 Fréchet-like derivative and Existence of Gradient Flows

Recall that a function $R\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ also defines an invariant function defined on $\mathcal{W}$ such that it agrees with R. Using an abuse of notation, we will use the same symbol, R, to denote it. We have discussed Fréchet-like derivative in Definition 2.13, but we will state it again here for completeness of this section.

Definition 4.18 (Fréchet-like derivative on $\mathcal{W}$)**.** Suppose $R\colon \mathcal{W} \to \mathbb{R} \cup \{\infty\}$ is an invariant function. Let $v \in$ eff-Dom(R). The *Fréchet-like derivative* at v is given by any $\phi \in L^\infty\big([0,1]^{(2)}\big)$ that satisfies the following condition,

$$\lim_{w \in \mathcal{W},\, \|w-v\|_2 \to 0} \frac{R(w) - R(v) - (\langle \phi, w\rangle - \langle \phi, v\rangle)}{\|w - v\|_2} = 0, \tag{4.13}$$

where $\langle \cdot, \cdot \rangle$ is the usual inner product on $L^2\big([0,1]^{(2)}\big)$. If R admits a Fréchet-like derivative at every $v \in$ eff-Dom(R), we denote the map that takes v to the corresponding ϕ by $D_{\mathcal{W}}R$. In that case we say that R is Fréchet differentiable.

In [DGKR15], the authors consider Gâteuax and Fréchet derivatives of functions on graphons with respect to the cut metric. However, as they remark [DGKR15, Remark 2.18, page 195], such a notion of Fréchet derivative is too weak to cover natural functions such as homomorphism densities.

Let $R\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$. Lemma B.5 (see Appendix B.2.2) justifies saying R has Fréchet-like derivative on $\widehat{\mathcal{W}}$ if R has a Fréchet-like derivative on $\mathcal{W}$. Note that the Lemma B.5 says that not only can the Fréchet-like derivative be thought of as a graphon, but also the two graphons $[D_{\mathcal{W}}R(v)]$ and $[v]$ are 'coupled' in the sense that they are two sets of "edge weights" associated with the edges of the same exchangeable continuum "graph". We make a formal definition to capture this relationship.

Definition 4.19 (Coupled graphons)**.** For any $r \in \mathbb{N}$, we define the set $[w_1] \odot [w_2] \odot \cdots \odot [w_r] \subseteq \widehat{\mathcal{W}}^r$ with initial labeling $(w_1, w_2, \ldots, w_r) \in \mathcal{W}^r$ as

$$\bigodot_{i=1}^{r} [w_i] \coloneqq \{(w_i^{\varphi})_{i=1}^{r} \mid \varphi \in \mathcal{T}\}. \tag{4.14}$$

Without loss of generality, we will always refer to the elements in $\odot_{i=1}^{r}[w_i]$ with the initial labeling $(w_i)_{i=1}^{r}$ unless specified. Since we can also relabel elements in $L^{\infty}\big([0,1]^{(2)}\big)$ (i.e., apply the map $v \mapsto v^{\varphi}$, for $v \in L^{\infty}\big([0,1]^{(2)}\big)$ and $\varphi \in \mathcal{T}$), we can generalize Definition 4.19 to tuples with elements in $L^{\infty}\big([0,1]^{(2)}\big) \supset \mathcal{W}$. That is, we can consider sets of the form

$$\bigodot_{i=1}^{r} [v_i] \coloneqq \{(v_i^{\varphi})_{i=1}^{r} \mid \varphi \in \mathcal{T}\}, \tag{4.15}$$

with initial labeling $\big(v_i \in L^{\infty}\big([0,1]^{(2)}\big)\big)_{i=1}^{r}$. Therefore, from Lemma B.5, if $v \in [v] \in \widehat{\mathcal{W}}$, and $\phi = D_{\mathcal{W}}R(v)$, then $(v, \phi) \in [v] \odot [\phi]$.

For $(v, \phi) \in [v] \odot [\phi]$, we define the set $G_v \subseteq [0,1]^2$ as

$$G_v \coloneqq \{|v| < 1\} \cup \{v = 1, \phi > 0\} \cup \{v = -1, \phi < 0\}. \tag{4.16}$$

Since $(v, \phi) \in [v] \odot [\phi]$, the set G_v is well defined on $\widehat{\mathcal{W}}$. For any $\varphi \in \mathcal{T}$, $G_{v^{\varphi}} = (G_v)^{\varphi} \coloneqq \{(\varphi(x), \varphi(y)) \in [0,1]^2 \mid (x, y) \in G_v\}$.

The next lemma gives an expression for the local slope of R in terms of its Fréchet-like derivative.

Lemma 4.20. *Let $R\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ be a function and $R\colon \mathcal{W} \to \mathbb{R} \cup \{\infty\}$ its invariant extension. Assume that for each $[v] \in \widehat{\mathcal{W}}$ the Fréchet-like derivative $D_{\mathcal{W}}R(v)$ exists for all*

$v \in [v]$, then the local slope (Definition 4.4) of R at $[v]$ satisfies

$$|\partial R|([v]) = \eta_R([v]) \coloneqq \sup_{w \in \mathcal{W}} \frac{(\langle \phi, v\rangle - \langle \phi, w\rangle)^+}{\|v - w\|_2} = \|\phi \mathbb{1}_{G_v}\{\,\cdot\,\}\|_2, \tag{4.17}$$

where $v \in [v]$, and $\phi = D_{\mathcal{W}} R(v)$. In particular, $|\partial R|([v]) = \|\phi\|_2$ if $v \in \{u \in \mathcal{W} \mid |u| < 1 \text{ a.e.}\} \cap \text{eff-Dom}(R)$.

The proof of Lemma 4.20 is provided in Appendix B.2.2.

Remark 4.21. We can define a similar expression for the set valued function G when eff-Dom(R) is a cubic domain. As an example, when eff-Dom$(R) = \{w \in \mathcal{W} \mid a \le w \le b \text{ a.e.}\}$ for some $-1 \le a \le b \le 1$ (see Section 4.6.1 for a discussed example), we can define $G_v \subseteq [0,1]^{(2)}$ for any $v \in$ eff-Dom(R) as

$$G_v = \{a < v < b\} \cup \{v = b, \phi > 0\} \cup \{v = a, \phi < 0\}, \tag{4.18}$$

for $(v, \phi \coloneqq D_{\mathcal{W}} R(v)) \in [v] \odot [\phi]$. Lemma 4.20 continues to hold when $v \in$ eff-Dom$(R) \subset \mathcal{W}$ whenever eff-Dom(R) is a cubic domain. In this case, the set valued function G is defined as described above and the proof of Lemma 4.20 can be modified accordingly.

Remark 4.22. Lemma 4.20 has an important consequence that will be used later. As the metric derivative of a gradient flow is given by its local slope at each point, Lemma 4.20 says that if ω is a gradient flow of R, then its local slope is given by the L^2-norm of its Fréchet-like derivative, i.e., $|\partial R|(\omega_t) = \big\|\phi(\omega_t)\mathbb{1}_{G_{\omega_t}}\{\,\cdot\,\}\big\|_2 = \big\|D_{\widehat{\mathcal{W}}} R(\omega_t)\mathbb{1}_{G_{\omega_t}}\{\,\cdot\,\}\big\|_2$ for all $t > 0$. Here for any $t > 0$,

$$D_{\widehat{\mathcal{W}}} R(\omega_t)\mathbb{1}_{G_{\omega_t}}\{\,\cdot\,\} \coloneqq \Big\{ \big(D_{\mathcal{W}} R(u_t)\mathbb{1}_{G_{u_t}}\{\,\cdot\,\}\big)^{\varphi} \in L^\infty\big([0,1]^{(2)}\big) \;\Big|\; \varphi \in \mathcal{T} \Big\},$$

for $u_t \in \omega_t$. Since the L^2-norm is invariant under measure preserving transformations [Jan13, Lemma 5.5], the L^2-norms of graphons are well-defined. In fact, if one defines a kernel valued curve $(w_t)_{t\in[0,T]}$ by setting $w_t' = -D_{\mathcal{W}} R(w_t)\mathbb{1}_{G_{w_t}}\{\,\cdot\,\}$ pointwise, then the curve $t \mapsto \omega_t = [w_t]$ is a gradient flow (a.k.a. curve of maximal slope). This is shown in Lemma B.7 which in turn shows the existence of a gradient flow under suitable assumption (See Theorem 4.23).

We now prove the existence of a curve of maximal slope if R satisfies reasonable assumptions. Moreover, as mentioned in the introduction, we show that the curve of maximal slope is the natural image of an absolutely continuous curve in $(\mathcal{W}, d_2)$.

Theorem 4.23 (Existence of curve of maximal slope-II [OPST23])**.** *Let $R\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ be a real valued function such that the Fréchet-like derivative $D_{\widehat{\mathcal{W}}}R([w])$ exists for all $[w] \in$ eff-Dom(R). For $w_0 \in [w_0] \in$ eff-Dom(R) and $t \geq 0$ define*

$$w_t \coloneqq w_0 - \int_0^t \phi(w_s)\mathbb{1}_{G_{w_s}}\{\,\cdot\,\}\,\mathrm{d}s, \qquad t \in \mathbb{R}_+,$$

where the above integral is pointwise. If R is λ-semiconvex w.r.t. d_2, then the curve $t \mapsto \omega_t = [w_t]$ is a curve maximal slope for R starting at $[w_0] \in$ eff-Dom(R).

The proof of Theorem 4.23 is provided in Appendix B.2.2.

Remark 4.24. An important consequence of the above Theorem is that if ω is a gradient flow of R then there exists an absolutely continuous curve $(w_t)_{t\in[0,T]} \in \mathrm{AC}(\mathcal{W}, d_2)$ such that $[w_t] = \omega_t$, $|\omega'|(t) = \left\|D_{\mathcal{W}}R(w_t)\mathbb{1}_{G_{w_t}}\{\,\cdot\,\}\right\|_2$ and $w_t' = -D_{\mathcal{W}}R(w_t)\mathbb{1}_{G_{w_t}}\{\}$, for each $t \in (0, T]$.

Remark 4.25. If R is $\delta_\square$-lower semicontinuous, λ-geodesically semiconvex for $\lambda \in \mathbb{R}_+$, and bounded from below, then one can say more about the convergence rate of a gradient flow to a minimizer of R. When $\lambda > 0$, let ω^* be the unique minimizer of R. Then following [AGS08, Remark 4.0.5, part (d)], a gradient flow ω of R on $\widehat{\mathcal{W}}$ starting at $\omega_0 \in \widehat{\mathcal{W}}$ satisfies

$$\delta_2(\omega_t, \omega^*) \leq \mathrm{e}^{-\lambda t}\delta_2(\omega_0, \omega^*), \qquad t \in \mathbb{R}_+.$$

In the limiting case when $\lambda = 0$ the exponential decay does not occur, in general, but some weaker results on the asymptotic behavior of ω hold. Following [AGS08, Corollary 4.0.6], ω satisfies

$$R(\omega_t) - R(\omega_\infty) \leq \frac{\delta_2^2(\omega_0, \omega_\infty)}{2t}, \qquad t \in \mathbb{R}_+,$$

for some minimum point ω_∞ of R, such that the map $t \mapsto \delta_2(\omega_t, \omega_\infty)$ is non-increasing. Moreover, $\lim_{t\to\infty} \delta_2(\omega_t, \omega_\infty) = 0$.

Finite dimensional Fréchet-like derivative

Recall the partition $Q_n \coloneqq \{Q_{n,i}\}_{i\in[n]}$ defined for any $n \in \mathbb{N}$ in Section 2.5. Given an invariant function $R\colon \mathcal{W} \to \mathbb{R} \cup \{\infty\}$, we can restrict its domain to kernels in $\mathcal{W}_n$ and still consider the Fréchet-like derivative $D_{\mathcal{W}_k}R\colon \mathcal{W}_n \cap \text{eff-Dom}(R) \to L_n^\infty\big([0,1]^{(2)}\big)$.

There are two equivalent ways of doing this. First, suppose the Fréchet-like derivative of R at v is given by $D_{\mathcal{W}}R(v) = \phi$. Then define $D_{\mathcal{W}_n}R(v) = \phi_n$ by conditional expectations as $\phi_n \coloneqq \mathbb{E}[\phi \mid \mathcal{F}_n]$, where $\mathcal{F}_n \coloneqq \sigma(Q_n \times Q_n)$. The object ϕ_n is referred to as a 'quotient' obtained by a 'stepping' of ϕ in [BCL$^+$08, Section 3.3] and [Lov12, Section 9.2.1] respectively. Since $\phi = D_{\mathcal{W}}R(v)$, by the *Tower Property* of conditional expectations we obtain that when $w, v \in \mathcal{W}_n$,

$$\begin{aligned} &\langle \phi_n, w\rangle - \langle \phi_n, v\rangle = \langle \phi, w\rangle - \langle \phi, v\rangle, \\ \implies &\lim_{\substack{w\in\mathcal{W}_n,\\ \|w-v\|_2\to 0}} \frac{R(w) - R(v) - (\langle \phi_n, w\rangle - \langle \phi_n, v\rangle)}{\|w - v\|_2} = 0. \end{aligned} \tag{4.19}$$

The second method of defining ϕ_n is to relate it to the Euclidean gradient over $n\times n$ symmetric matrices. This is done in Lemma 2.14 below.

In any case, we can define $\eta_{R,n}\colon \widehat{\mathcal{W}}_n \cap \text{eff-Dom}(R) \to \mathbb{R}_+$ for the function $R\colon \widehat{\mathcal{W}} \to \mathbb{R}\cup\{\infty\}$ as follows. If $v \in [v] \in \widehat{\mathcal{W}}_n \cap \text{eff-Dom}(R)$, then

$$\eta_{R,n}([v]) \coloneqq \sup_{w\in\mathcal{W}_n} \frac{(\langle \phi_n, v\rangle - \langle \phi_n, w\rangle)^+}{\|v - w\|_2}. \tag{4.20}$$

We can also define the local slope $|\partial_n R|$ restricted to $\widehat{\mathcal{W}}_n$ as

$$|\partial_n R|([v]) \coloneqq \limsup_{[w]\in\widehat{\mathcal{W}}_n,\, \delta_2([w],[v])\to 0} \frac{(R([v]) - R([w]))^+}{\delta_2([w],[v])}, \tag{4.21}$$

for $[v] \in \widehat{\mathcal{W}}_n \cap \text{eff-Dom}(R)$. Then, by a similar argument as shown in the proof of Lemma 4.20, we have the following corollary.

Corollary 4.26. *Let $R\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ be a function. Assume that for $[v] \in \widehat{\mathcal{W}}_n \cap$ eff-Dom(R) the Fréchet-like derivative $D_{\mathcal{W}}R(v)$ exists for all $v \in [v]$. Then the local slope (Definition 4.4) of R at $[v]$ satisfies $|\partial_n R|([v]) = \eta_{R,n}([v])$.*

4.4 Scaling limits of finite dimensional Gradient Flows

Recall that the goal of this section is to show that the Euclidean gradient flows on matrices converge in suitable sense to gradient flow on graphons. In the previous section, we establish that the gradient flows in very general settings can be obtained as the limits of discrete solutions. In this section, we show that iterates of $J_\tau|_{\widehat{\mathcal{W}}_n}$ converge in suitable sense to the iterates of $J_\tau|_{\widehat{\mathcal{W}}}$ as $n \to \infty$.

More formally, for any $n \in \mathbb{N}$, define $J_\tau^{(n)}$ to be the resolvent operator on $\widehat{\mathcal{W}}_n$ as

$$J_\tau^{(n)}([u]) \coloneqq \operatorname*{arg\,min}_{\widehat{\mathcal{W}}_n} \Phi_R(\tau, [u]; \cdot) = \operatorname*{arg\,min}_{\widehat{\mathcal{W}}_n} \left\{ R + \frac{1}{2\tau}\delta_2^2([u], \cdot) \right\}, \tag{4.22}$$

for any $\tau > 0$, $[u] \in \widehat{\mathcal{W}}_k$.

The following Lemma essentially shows Γ-convergence of the penalized functionals Φ_R, restricted to $\widehat{\mathcal{W}}_n$, as $n \to \infty$.

Proposition 4.27. *Fix some $\delta_\square$-continuous function $R\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ and some step size $\tau > 0$. Consider a sequence $([u_n] \in \widehat{\mathcal{W}}_n)_{n\in\mathbb{N}}$ such that $([u_n])_{n\in\mathbb{N}} \xrightarrow{\delta_\square} [u]$ as $n \to \infty$ for some $[u] \in \widehat{\mathcal{W}}$. For each $n \in \mathbb{N}$, let $\left[u_{n,\tau}^+\right] \in \arg\min_{\widehat{\mathcal{W}}_n} \Phi_R(\tau, [u_k]; \cdot)$. Suppose $\left[u_{\infty,\tau}^+\right]$ is any $\delta_\square$-limit point of the sequence $\left(\left[u_{n,\tau}^+\right]\right)_{n\in\mathbb{N}}$. Then $\left[u_{\infty,\tau}^+\right] \in \arg\min_{\widehat{\mathcal{W}}} \Phi_R(\tau, [u]; \cdot)$.*

The proof of Proposition 4.27 is provided in Appendix B.3.

In Section 2.5.2 we discussed that implicit Euler iteration on $\widehat{\mathcal{W}}_n$ for any $n \in \mathbb{N}$ can be viewed as the time scaling of the implicit Euler method on the Euclidean space of $n \times n$ symmetric matrices. The following lemma complements that discussion by saying that the gradient flow on $\mathcal{W}_n$ can be obtained from the Euclidean gradient flow on the space of $n \times n$ symmetric matrices.

Let $n \in \mathbb{N}$, and $(w_{n,t} = w_k(t) \in \mathcal{W}_n \cap \text{eff-Dom}(R))_{t\in\mathbb{R}_+}$ be the Euclidean coordinate gradient flow of R. This may be obtained by suitably scaling the solution of the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} M_n(w_n(t)) = -(\nabla R_n \circ M_n)(w_n(t)), \tag{4.23}$$

with initial condition $w_n(0) = w_{n,0} \in \mathcal{W}_n \cap \text{eff-Dom}(R)$, until the process hits the boundary when one or more entries is ± 1. At the boundary, however, the gradient might push the process outside $\mathcal{M}_n$ and it needs some care to have a proper definition. Instead, we consider the Euclidean gradient flow as the limit of implicit Euler iterations as the step size tends to zero. This definition is valid everywhere and is equivalent to the previous one on Euclidean spaces.

As a consequence of Lemma 2.14, we obtain that the Euclidean coordinate-wise gradient flow on $\mathcal{W}_n$ is the gradient flow on $\mathcal{W}_n$. We are now ready to prove Theorem 4.28. For completeness, we reproduce the theorem statement below.

Theorem 4.28 (Convergence of Gradient Flows [OPST23])**.** *Suppose* $R\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ *satisfies the following conditions:*

1. *R is continuous in $\delta_\square$,*

2. *R is λ-semiconvex (Definition 4.10) along generalized geodesics on $(\widehat{\mathcal{W}}, \delta_2)$ (Definition 4.14), for some $\lambda \in \mathbb{R}$.*

Consider the gradient flow $\omega^{(n)} = \left(\omega_t^{(n)}\right)_{t\in\mathbb{R}_+} \subset \widehat{\mathcal{W}}_n$ of R on each $\widehat{\mathcal{W}}_n$, starting at some $\omega_0^{(n)} = [u_{n,0}]$ for $n \in \mathbb{N}$. Assume that the sequence $([u_{n,0}])_{n\in\mathbb{N}} \xrightarrow{\delta_\square} [u_0]$, and $|\partial R|([u_0]) < \infty$ and $\limsup_{n\to\infty} |\partial R|([u_{n,0}]) \le G < \infty$, for some $G \ge 0$. Then,

$$\limsup_{n\to\infty} \sup_{t\in[0,T]} \delta_\square\left(\omega_t^{(n)}, \omega_t\right) = 0, \tag{4.24}$$

for any $T \in \mathbb{R}_+$, where $\omega = (\omega_t)_{t\in\mathbb{R}_+}$ is the unique minimizing movement curve [AGS08, Definition 2.0.6, page 42] on $\widehat{\mathcal{W}}$ for the function R starting at $\omega_0 = [u_0]$ [AGS08, Theorem 4.0.4]. In addition, if the conditions for the existence of curves of maximal slope (Theorem 4.17 or Theorem 4.23) hold, then ω is also a curve of maximal slope.

The proof of Theorem 4.28 is provided in Appendix B.3.

Remark 4.29. Condition 2 in Theorem 4.28 is satisfied if the invariant extension $R\colon \mathcal{W} \to \mathbb{R} \cup \{\infty\}$ of $R\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ is λ-semiconvex on $(\mathcal{W}, d_2)$ (see Section 4.2, and Definition 4.11).

4.5 Continuity Equations

It is well-known that any absolutely continuous curve in the Wasserstein space can be represented as the solution of a continuity equation [San15, Section 5.3]. Something analogous is partially true for graphons as well. However, the presence of the boundary in $\widehat{\mathcal{W}}$ makes the situation more delicate and we can only characterize AC curves via the continuity equation until it hits the boundary.

Before we state the main result, we introduce some notations. Let $v \in L^1([0,1]^{(2)})$ and let $[w] \in \widehat{\mathcal{W}}$. Let $w \in [w]$ be a representative of $[w]$. For any $n \in \mathbb{N}$, we can define $X_n\colon [0,1]^n \to \mathcal{M}_n$ and $v_n\colon \mathcal{M}_n \to \mathbb{R}^{[n]^{(2)}}$ as

$$\begin{aligned} X_n((u_\ell)_{\ell=1}^n) &\coloneqq (w(u_i,u_j))_{(i,j)\in[n]^{(2)}}, \\ v_n(z)(i,j) &= \mathbb{E}[v(U_i,U_j) \mid X_n((U_\ell)_{\ell=1}^n) = z], \end{aligned} \tag{4.25}$$

where $\{U_i\}_{i\in\mathbb{N}}$ are i.i.d. as Uni$[0,1]$. Intuitively, formula (4.25) means that we average the edge weights from v over all embedding of the vertex labeled weighted graph z in the graphon w. Since X_{n-1} is a leading principle submatrix of X_n, i.e., $X_{n-1} = (X_n(i,j))_{(i,j)\in[n-1]^{(2)}}$, we have that $\sigma(X_{n-1}) \subseteq \sigma(X_n)$, which defines a filtration $\mathcal{F} = (\mathcal{F}_n \coloneqq \sigma(X_n))_{n\in\mathbb{N}}$. It is clear that $(v_n)_{n\in\mathbb{N}}$ is a martingale with respect to the filtration $\mathcal{F}$. We also note that that v_n is a function of the graphon and not its kernel representative. We record both these observations as lemma below.

Lemma 4.30. *For every $i,j \in \mathbb{N}$, the process $(v_n(X_n)(i,j))_{n=\max\{i,j\}}^\infty$ is a martingale with respect to the filtration $\mathcal{F}$.*

Lemma 4.31. *For any $\varphi \in \mathcal{T}$, we have $v_n^\varphi(X_n^\varphi) = v_n(X_n)$, for all $n \in \mathbb{N}$, where $v^\varphi(x,y) \coloneqq v(\varphi(x),\varphi(y))$ for all $(x,y) \in [0,1]^{(2)}$.*

The proof of Lemma 4.31 is provided in Appendix B.4.

Suppose we are given some $\omega = (\omega_t)_{t\in[0,1]} \in \mathrm{AC}(\widehat{\mathcal{W}}, \delta_2)$. From Lemma B.2, we obtain $(w_t)_{t\in[0,1]} \in \mathrm{AC}(\mathcal{W}, d_2)$ such that $\omega_t = [w_t]$ for all $t \in [0,1]$. It follows from the Radon-Nikodým property [Huf77, page 30, Theorem 5] that there exists $v_t \coloneqq w_t' \in L^2([0,1]^{(2)})$, for a.e. $t \in [0,1]$, such that $w_t - w_0 = \int_0^t v_s \,\mathrm{d}s$, where the integral is pointwise.

For any $t \in [0,1]$, let $v_{n,t}$ and $X_{n,t}$ be defined as in equation (4.25) with w_t replacing the role of w and v_t replacing v. The following Proposition 4.32 shows that the $\rho_{n,t} = \mathrm{Law}(X_{n,t}((U_i)_{i=1}^n))$ satisfies continuity equation with the velocity $v_{n,t}$ defined as

$$v_{n,t}(z) \coloneqq \mathbb{E}\Big[(w_t'(U_i,U_j))_{(i,j)\in[n]^{(2)}} \;\Big|\; (w_t(U_i,U_j))_{(i,j)\in[n]^{(2)}} = z\Big], \qquad z \in \mathcal{M}_n.$$

Proposition 4.32 ([OPST23])**.** *Let $n \in \mathbb{N}$, and $\rho_{n,t} = \mathrm{Law}(X_{n,t}((U_i)_{i=1}^n))$. Then, for a.e. $t \in [0,1]$, the continuity equation $\partial_t \rho_{n,t} + \mathrm{div}(v_{n,t}\rho_{n,t}) = 0$ holds weakly with Dirichlet boundary conditions. That is, for any continuously differentiable test function R_n on $\mathcal{M}_n$, vanishing at the boundary,*

$$\partial_t\left(\int R_n(z)\,\mathrm{d}\rho_{n,t}(z)\right) - \int \nabla R_n(z) v_{n,t}(z)\,\mathrm{d}\rho_{n,t}(z) = 0, \qquad \text{a.e.} \quad t \in [0,1].$$

The proof of Proposition 4.32 is provided in Appendix B.4.

Proposition 4.32 shows that an absolutely continuous curve in $(\widehat{\mathcal{W}}, \delta_2)$ determines a family of continuity equations. In [HOP$^+$22], the authors take this analogy further and show that in the presence of noise the limiting curve can be described by a certain McKean-Vlasov type equation.

4.6 Examples and Simulations

In this section, we find some natural classes of examples of functions on graphons with Fréchet-like derivatives. For any graphon $[w] \in \widehat{\mathcal{W}}$ and any $n \in \mathbb{N}$, sample $\{Z_i\}_{i\in[n]}$ i.i.d. from Uni$[0,1]$. Let $G_n[w] = (W(Z_i,Z_j))_{(i,j)\in[n]^{(2)}}$. Let $\rho_n([w]) = \mathrm{Law}(G_n[w])$ denote its law. Consider the functions $R\colon \widehat{\mathcal{W}} \to \mathbb{R}$ defined through the function composition $R = H_n \circ \rho_n$, for different choices of $H_n\colon \mathcal{P}(\mathcal{M}_n) \to \mathbb{R}$. Since ρ_n produce distributions over exchangeable $n \times n$ arrays, the function R is well-defined over $\widehat{\mathcal{W}}$.

4.6.1 Linear functions

Let $H_n \colon \mathcal{P}(\mathcal{M}_n) \to \mathbb{R}$ be defined as a linear function such that R takes the form

$$R(w) = \langle R_n, \rho_n([w]) \rangle \coloneqq \int_{\mathcal{M}_n} R_n(z) \rho_n([w])(\mathrm{d}z), \qquad w \in \mathcal{W}. \tag{4.26}$$

We next show that if R satisfies mild conditions, the Fréchet-like derivative of R has a closed form expression.

Lemma 4.33. *If for all $[w] \in \widehat{\mathcal{W}}$,*

1. *$R_n \in L^1(\rho_n([w]))$, and $\nabla R_n \in L^\infty(\rho_n([w]))$,*

2. *R_n is continuously differentiable on an open set containing $\operatorname{supp}(\rho_n([w]))$, and*

3. *R_n satisfies either one of the following conditions:*

   (a) *For any $\varepsilon > 0$ there is some $\delta_\varepsilon > 0$ such that for all $X, X_0 \in \mathcal{M}_n$ satisfying $\|X - X_0\|_2 \leq \delta_\varepsilon$,*

   $$|R_n(X) - R_n(X_0) - \langle \nabla R_n(X_0), X - X_0 \rangle| \leq \varepsilon \|X - X_0\|_2,$$

   where $\langle \cdot, \cdot \rangle \colon \mathcal{M}_n(\mathbb{R}) \times \mathcal{M}_n(\mathbb{R}) \to \mathbb{R}$ and $\| \cdot \|_2$ are the standard Frobenius inner product and Frobenius norm over $n \times n$ matrices respectively.

   (b) *There is a constant $C_0 \in \mathbb{R}_+$ such that*

   $$\sup_{X, X_0 \in \mathcal{M}_n} |R_n(X) - R_n(X_0) - \langle \nabla R_n(X_0), X - X_0 \rangle| \leq C_0,$$

then R admits a Fréchet-like derivative satisfying

$$(D_{\mathcal{W}} R)(w)(x, y) = \sum_{i,j=1}^{n} \mathbb{E}\left[\left(\nabla R_n \Big((w(Z_p, Z_q))_{(p,q) \in [n]^{(2)}} \Big) \right)_{i,j} \,\middle|\, (Z_i, Z_j) = (x, y) \right],$$

for $(x, y) \in [0, 1]^{(2)}$, and $w \in \mathcal{W}$.

The proof of Lemma 4.33 is provided in Appendix B.5.

We show that geodesic semiconvexity of R on $(\widehat{\mathcal{W}}, \delta_2)$, is implied by the semiconvexity of R_n on $\mathcal{M}_n$.

Lemma 4.34. *If R_n is λ-semiconvex on the convex set $\mathcal{M}_n$, then R is generalized geodesically $n(n-1)\lambda$-semiconvex on $\widehat{\mathcal{W}}$. In particular, it is also geodesically $n(n-1)\lambda$-semiconvex on $\widehat{\mathcal{W}}$.*

The proof of Lemma 4.34 can be found in Appendix B.5.

In the following subsections, we examine special cases of linear functions.

Scalar Entropy function

The scalar entropy function $\mathcal{E}\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ is defined as

$$\mathcal{E}([w]) \coloneqq \int_{[0,1]^2} h(w(x,y))\,\mathrm{d}x\,\mathrm{d}y, \qquad [w] \in \widehat{\mathcal{W}},$$

where $h\colon \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is the convex entropy function $h(p) \coloneqq p\log p + (1-p)\log(1-p)$, if $p \in (0,1)$, $h(0) = h(1) = 0$, and $h(p) = \infty$, otherwise. Observe that it can also be thought of as a linear function with $n = 2$. If

$$R_2\Big(\big(x_{(i,j)}\big)_{(i,j)\in[2]^{(2)}}\Big) \coloneqq h(x_{(1,2)}), \qquad x \in \mathcal{M}_2, \tag{4.27}$$

then from equations (4.26) and (B.51),

$$\begin{aligned} R([w]) = \langle R_2, \rho_2([w])\rangle &= \mathbb{E}_{\{Z_i\sim\mathrm{Uni}[0,1]\}_{i=1}^2}\Big[R_2\Big(\big(w(Z_i,Z_j)\big)_{(i,j)\in[2]^{(2)}}\Big)\Big] \\ &= \mathbb{E}_{\{Z_i\sim\mathrm{Uni}[0,1]\}_{i=1}^2}[h(w(Z_1,Z_2))] = \mathcal{E}([w]), \end{aligned} \tag{4.28}$$

for $[w] \in \widehat{\mathcal{W}}$. Since, $h(p)/p$ is bounded when $\epsilon \le p \le 1-\epsilon$ for some $\epsilon \in (0, 1/2)$, it follows that $\mathcal{E}$ restricted to $\widehat{\mathcal{W}}_\epsilon \coloneqq \{[w] \in \widehat{\mathcal{W}} \mid \epsilon \le w \le 1-\epsilon \text{ a.e.}\}$, is continuous with respect to the weak-$*$ topology on $L^2([0,1]^2)$. Since this is a weaker topology than the one generated by $\delta_\square$, by [Lov12, Lemma 8.22], it follows that $\mathcal{E}$ restricted to $\widehat{\mathcal{W}}_\epsilon$ is $\delta_\square$-continuous.

Since $h''(p) = 1/(p(1-p)) \geq 4$ for $p \in \text{eff-Dom}(h)$, it follows that h is 4-semiconvex on $\mathbb{R}$. Let $E\colon \mathcal{W} \to \mathbb{R} \cup \{\infty\}$ be the invariant extension of $\mathcal{E}$ such that $E(w) \coloneqq \mathcal{E}([w])$ for all $w \in \mathcal{W}$. Then for any $w_0, w_1 \in \mathcal{W}$, defining $w_t \coloneqq (1-t)w_0 + tw_1$ for $t \in [0,1]$, we have

$$\begin{aligned} E(w_t) &= \int_0^1 \int_0^1 h(w_t(x,y))\,\mathrm{d}x\,\mathrm{d}y \\ &\leq \int_0^1 \int_0^1 [(1-t)h(w_0(x,y)) + th(w_1(x,y))]\,\mathrm{d}x\,\mathrm{d}y \\ &\qquad - \int_0^1 \int_0^1 \frac{4}{2} t(t-1)(w_0 - w_1)^2(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= (1-t)E(w_0) + tE(w_1) - \frac{1}{2} 4t(1-t)\|w_0 - w_1\|_2^2, \end{aligned} \tag{4.29}$$

which implies that E is 4-semiconvex on $(\mathcal{W}, d_2)$ (see Definition 4.11). Following Remark 4.29, $\mathcal{E}$ is 4-semiconvex along generalized geodesics on $(\widehat{\mathcal{W}}, \delta_2)$.

Suppose w is valued in $(0,1)$ a.e.. Then,

$$\mathbb{E}[\langle \nabla R_2 \circ w_2, G_2[u] \rangle] = \int_0^1 \int_0^1 \log\left(\frac{w(z_1,z_2)}{1-w(z_1,z_2)}\right) u(z_1,z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2, \tag{4.30}$$

for all $u \in \mathcal{W}$. By the characterization of the Fréchet-like derivative in equation (B.57), we get that $\phi_{\mathcal{E}} = D_{\mathcal{W}} R_2(w)$ is given by

$$\phi_{\mathcal{E}}(x,y) = \log\left(\frac{w(x,y)}{1-w(x,y)}\right), \qquad \text{a.e.} \quad (x,y) \in [0,1]^{(2)}. \tag{4.31}$$

If $[w] \in \widehat{\mathcal{W}}_\epsilon$ for some $\epsilon \in (0, 1/2)$, then by Lemma 4.20 the local slope of entropy $|\partial \mathcal{E}|([w])$ is given by $|\partial \mathcal{E}|([w]) = \|\phi_{\mathcal{E}}\|_2$.

Note that $\widehat{\mathcal{W}}_\epsilon$ is closed in $\widehat{\mathcal{W}}$ and since $w \mapsto \|w\|_2$ is $\delta_\square$-lower semicontinuous [Lov12, Lemma 14.15], it follows that the local slope of entropy, $|\partial \mathcal{E}|$, is also $\delta_\square$-lower semicontinuous on $\widehat{\mathcal{W}}_\epsilon$. Following Remark 4.24, we conclude that starting from $[w_0] \in \widehat{\mathcal{W}}_\epsilon$ the gradient flow of $\mathcal{E}$ flows with the velocity $-D_{\widehat{\mathcal{W}}} \mathcal{E}([w_0])$ at $[w_0]$ if $\epsilon \leq w_0 \leq 1 - \epsilon$ a.e. Consider the flow $w_t(x,y)$ obtained by solving

$$w_t'(x,y) = -\log\left(\frac{w_t(x,y)}{1-w_t(x,y)}\right), \qquad \text{a.e.} \quad (x,y) \in [0,1]^{(2)}, \tag{4.32}$$

with initial condition $[w_0] \in \widehat{\mathcal{W}}_\epsilon$. Then it follows that $(\omega_t \coloneqq [w_t])_{t\in\mathbb{R}_+}$ is a gradient flow of $\mathcal{E}$ starting from $\omega_0 = [w_0]$.

Following Section 4.5, from equation (4.31) it follows that the velocity of gradient flow for scalar entropy function satisfies

$$v([w])(x,y) = -\log\left(\frac{w(x,y)}{1-w(x,y)}\right), \qquad (x,y)\in[0,1]^{(2)}, \tag{4.33}$$

if $0 < w < 1$ a.e. Let $\{U_i\}_{i=1}^\infty$ be i.i.d. Uni$[0,1]$, then using equation (4.25) we can compute the velocities $(v_n)_{n\in\mathbb{N}}$ appearing in the continuity equation as

$$\begin{aligned} v_n(z)(i,j) &= \mathbb{E}[v(U_i,U_j) \mid X_n((U_\ell)_{\ell=1}^n) = z] \\ &= -\mathbb{E}\left[\log\left(\frac{w(U_i,U_j)}{1-w(U_i,U_j)}\right) \,\middle|\, X_n((U_\ell)_{\ell=1}^n) = z\right] \\ &= -\log\left(\frac{z(i,j)}{1-z(i,j)}\right), \qquad z\in\mathcal{M}_n, \quad i,j\in[n]. \end{aligned}$$

Takeaway 4.1. It is worth emphasizing that, following equation (4.32), for any $\epsilon \in (0,1/2)$, the stationary point for the gradient flow of $\mathcal{E}$ is the constant half graphon, which is also the minimizer of $\mathcal{E}$ on $\widehat{\mathcal{W}}$. We also observe that the curve $(\omega_t)_{t\in\mathbb{R}_+}$ always stays inside $\widehat{\mathcal{W}}_\epsilon$ if $\omega_0 \in \widehat{\mathcal{W}}_\epsilon$. Since $\mathcal{E}$ is 4-semiconvex, it follows from Remark 4.25 that we are guaranteed an exponential rate of convergence of the gradient flow to the minimizer.

Homomorphism functions

For $n \in \mathbb{N}\setminus\{1\}$ we can consider interactions such as the homomorphism functions that are continuous in the cut-metric:

$$T_H([w]) \coloneqq \mathbb{E}\left[\prod_{\{i,j\}\in E(H)} w(Z_i,Z_j)\right] = \int_{\mathcal{M}_n} R_n(z)\rho_n(\mathrm{d}z), \tag{4.34}$$

where H is a simple graph with $|V(H)| = n$, $E(H) = \{e_i\}_{i=1}^m$, $\{Z_i\}_{i\in[n]}$ are i.i.d. uniformly in $[0,1]$, and $R_n\big(\big(x_{(i,j)}\big)_{(i,j)\in[n]^{(2)}}\big) = \prod_{\{i,j\}\in E(H)} x_{(i,j)}$. In particular, the function R_n is a monomial for every simple graph H. Since

$$(\nabla R_n(X))_{(p,q)} = \mathbb{1}_{E(H)}\{p,q\} \cdot \prod_{\{i,j\}\in E(H)\setminus\{\{p,q\}\}} X_{(i,j)}, \tag{4.35}$$

for $p, q \in [k]$, the action of the Fréchet-like derivative $\phi_{T_H} = D_{\mathcal{W}} T_H(w)$ on $u \in \mathcal{W}$ according to equation (B.57) is given by

$$\sum_{\ell=1}^{m} \mathbb{E}\left[\prod_{r=1}^{\ell-1} w(Z_{e_r}) \cdot u(Z_{e_\ell}) \cdot \prod_{r=\ell+1}^{m} w(Z_{e_r})\right], \tag{4.36}$$

where $Z_e \coloneqq \big(Z_{e(1)}, Z_{e(2)}\big)$ for all $e \in E(F)$. Following a similar approach to equation (B.55), we obtain that ϕ_{T_H} satisfies

$$\begin{aligned} \phi_{T_H}(x,y) &= \sum_{\ell=1}^{m} \mathbb{E}\left[\prod_{r=1, r\neq\ell}^{m} w(Z_{e_r}) \,\middle|\, Z_{e_\ell} = (x,y)\right] \\ &= \sum_{\ell=1}^{m} \mathbf{t}_{x,y}(H_{e_\ell}, w), \quad x, y \in [0,1], \end{aligned} \tag{4.37}$$

where H_e is the graph obtained by removing edge e from H and $\mathbf{t}_{x,y}(H_e, w)$ is the *homomorphism density of a partially labeled graph* [Lov12, Section 7.4] defined

$$\mathbf{t}_{x,y}(H_e, w) \coloneqq \mathbb{E}\left[\prod_{f \in E(H_e)} w(Z_f) \,\middle|\, Z_e = (x,y)\right].$$

Note that for fixed w and H_e, we can think of $(x,y) \mapsto \mathbf{t}_{x,y}(H_e, w)$ as a bounded measurable function on $[0,1]^2$. We now show that the kernel valued map $w \mapsto \mathbf{t}_{(\cdot,\cdot)}(H_e, w)$ is continuous with respect to $\|\cdot\|_\square$. To this end, let $w, w' \in \mathcal{W}$ and consider any product function $(x,y) \mapsto f(x)g(y)$ where $\|f\|_\infty, \|g\|_\infty \leq 1$. Note that

$$\mathbf{t}_{x,y}(H_e, w) = \int_{[0,1]^{k-2}} \prod_{\{i,j\} \in E(H_e)} w(x_i, x_j) \prod_{\ell \in V(H_e)\setminus e} \mathrm{d}x_\ell.$$

For each edge $\{p,q\} \in E(H_e)$, consider the integral

$$I_{p,q} \coloneqq \int_{[0,1]^n} (w(x_p, x_q) - W'(x_p, x_q)) \prod_{\{i,j\} \in E(H_e)\setminus\{p,q\}} w(x_i, x_j) \cdot \prod_{\ell \in V(H_e)\setminus e} \mathrm{d}x_\ell \; f(x)g(y)\,\mathrm{d}x\,\mathrm{d}y.$$

It follows from [Lov12, Lemma 8.10] (or see the second last display in the proof of [Lov12, Lemma 10.24]) that $|I_{p,q}| \leq \|w - w'\|_\square$. From the same references above, it follows that

$$\left|\int_{[0,1]^2} (\mathbf{t}_{x,y}(H_e, w) - \mathbf{t}_{x,y}(H_e, w'))f(x)g(y)\,\mathrm{d}x\,\mathrm{d}y\right| \leq \sum_{\{p,q\} \in E(H_e)} |I_{p,q}| \leq |E(H_e)|\|w - w'\|_\square.$$

Taking supremum over Borel measurable functions f, g such that $\|f\|_\infty, \|g\|_\infty \leq 1$, we get that $w \mapsto \mathbf{t}_{(\cdot,\cdot)}(H_e, w)$ is Lipschitz continuous with respect to $\|\cdot\|_\square$. Since $|\mathbf{t}_{x,y}(H_e, w)| \leq 1$ for $w \in \mathcal{W}$, it follows that $\|\phi_{T_H}(w)\|_\infty \leq |E(H)|$, that is, ϕ_{T_H} is uniformly bounded on $\mathcal{W}$.

For example, when H is a triangle, the velocity ϕ_{T_H} follows from equation (4.37) as $\phi_{T_H}(x,y) = 3\int_0^1 w(x,z)w(z,y)\,\mathrm{d}z$, for a.e. $x, y \in [0,1]$, i.e., thrice the 'operator product' of w with itself [Lov12, Section 7.4]. If H is a path on 3 vertices and 2 edges, then $\phi_{T_H}(x,y) = \int_0^1 w(x,z)\,\mathrm{d}z + \int_0^1 w(y,z)\,\mathrm{d}z = \deg(x) + \deg(y)$, where $\deg(x) \coloneqq \int_0^1 w(x,z)\,\mathrm{d}z$ for a.e. $x, y \in [0,1]$.

Obtaining the expression for the Fréchet-like derivative in equation (4.37), given a gradient flow ω of T_H, we can compute the velocity of the gradient flow as $D_{\widehat{\mathcal{W}}}T_H(\omega_t)\mathbb{1}\{G_{\omega_t}\}$ for $t > 0$.

The Hessian of the function R_n can be easily computed as

$$\partial_{x_{(i,j)}x_{(p,q)}} R_n = \prod_{\{a,b\} \in E(H) \setminus \{\{i,j\},\{p,q\}\}} x_{(a,b)},$$

if $\{i,j\} \neq \{p,q\}$ are both edges in $E(H)$, and zero otherwise.

Since every $x_{(i,j)} \in [-1,1]$ for all $(i,j) \in [n]^{(2)}$, every entry of the Hessian is uniformly bounded in $[-1,1]$ and hence $\|\mathrm{Hess}(R_n)\|_{\mathrm{op}} \leq n(n-1)/2$. Therefore, R_n is $(-n(n-1)/2)$-semiconvex w.r.t. d_2. It follows from Lemma 4.34 that homomorphism function T_H is $(-n^2(n-1)^2/2)$-semiconvex w.r.t. δ_2. In fact, since $\mathrm{Hess}(R_n)(\{i,j\},\{p,q\})$ is non-zero only when $\{i,j\},\{p,q\} \in E(H)$, it follows that $\|\mathrm{Hess}(R_n)\|_{\mathrm{op}} \leq \sqrt{m(m-1)} \leq m$. This would yield that T_H is $(-mn(n-1))$-semiconvex w.r.t. δ_2. This is useful when H is sparse.

Takeaway 4.2. Note that in this example, it is not clear if the minimizer is a constant graphon or not. If H has odd number of edges however constant graphon $w \equiv -1$ is trivially a minimizer. In the case of graphs H with even number of edges, explicitly determining the minimizer is trickier. As we discussed above, it is also not clear if the homomorphism density function is convex, therefore we cannot guarantee an exponential rate of convergence of the gradient flow to a minimizer following Remark 4.25. To alleviate this, one can regularize the

objective function by adding the scalar entropy function. We discuss this in the next example in Section 4.6.1. However, in the simulation discussed in Section 4.6.4, it does appear that the gradient flow converges to the minimizer of the function of interest at an exponential rate.

Linear combination of Scalar Entropy and Homomorphism function

Let $\beta \in \mathbb{R}$ and let H be a finite simple graph with $n \in \mathbb{N}$ vertices and $m \in \mathbb{N}$ edges. Define the function $G = G_{\beta,H} \coloneqq \mathcal{E} + \beta T_H$ on the set $\widehat{\mathcal{W}}_\epsilon$ for $\epsilon \in (0, 1/2)$. The function G is of particular interest in the theory of exponential random graph models. The function G is $\delta_\square$-continuous on $\widehat{\mathcal{W}}_\epsilon$ and since $\mathcal{E}$ is 4-semiconvex and T_H is λ-semiconvex w.r.t. δ_2, it follows that G is also $(4 + \beta\lambda)$-semiconvex w.r.t. δ_2 for some $\lambda \in \mathbb{R}$ that we estimate in Section 4.6.1. The gradient flow of G, therefore, exists and the Fréchet-like derivative $\phi_G = D_{\mathcal{W}} G = \phi_{\mathcal{E}} + \beta\phi_{T_H}$. Since $\mathcal{E}$ is 4-semiconvex and T_H is $(-n^2(n-1)^2/2)$-semiconvex w.r.t. δ_2, it follows that $G_{\beta,H}$ is $(4 - \beta n^2(n-1)^2/2)$-semiconvex w.r.t. δ_2. Therefore, $G_{\beta,H}$ is at least 0-semiconvex (i.e., convex) w.r.t. δ_2 when $\beta \leq 8/n^2(n-1)^2$.

Takeaway 4.3. Note that $\beta < 8/n^2(n-1)^2$ guarantees exponential rates of convergence of the gradient flow curve. See Remark 4.25.

4.6.2 Interaction energy

In the optimal transport literature, linear functionals of probability measures are often called *potential energy* [San15, page 249]. Inspired by similar definitions, one can define *interaction energy*. Let $R_n \colon \mathcal{M}_n \times \mathcal{M}_n \to \mathbb{R}$ be a function defined on pairs of $n \times n$ symmetric matrices with entries in $[-1, 1]$. Given a graphon $[w] \in \widehat{\mathcal{W}}$, as before let $\rho_n = \mathrm{Law}(G_n[w])$. This defines a function $\mathbb{F}_n \colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ given by

$$\mathbb{F}_n([w]) \coloneqq \int_{\mathcal{M}_n} \int_{\mathcal{M}_n} R_n(z, z') \rho_n(\mathrm{d}z) \rho_n(\mathrm{d}z').$$

Although it looks different than before, this is also a particular case of equation (B.57) as shown below. Define two independent sequences of i.i.d. Uni$[0, 1]$ random variables:

$(Z_1, \dots, Z_n)$ and $(Z'_1, \dots, Z'_n)$, and consider the corresponding matrices

$$X_n \coloneqq (w(Z_i, Z_j))_{(i,j)\in[n]^{(2)}}, \quad \text{and} \quad X'_n \coloneqq (w(Z'_i, Z'_j))_{(i,j)\in[n]^{(2)}}.$$

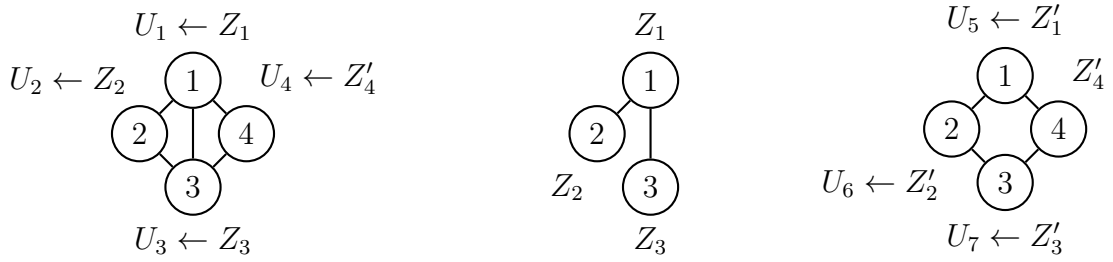
Then X_n and X'_n are i.i.d. samples from ρ_n and $\mathbb{F}_n([w]) = \mathbb{E}[R_n(X_n, X'_n)]$. On the other hand, one can concatenate $(Z_i)_{i=1}^n$ and $(Z'_i)_{i=1}^n$ to construct a single vector $(Z_1, \dots, Z_n, Z'_1, \dots, Z'_n)$ of dimension $2n$ and consider the corresponding $2n \times 2n$ symmetric matrix X_{2n} whose block diagonal components are X_n and X'_n. By defining $\bar{f}_n(X_{2n}) \coloneqq R_n(X_n, X'_n)$, we represent $\mathbb{F}_n([w]) = \mathbb{E}\big[\bar{f}_{2n}(X_{2n})\big] = \int \bar{f}_{2n}(w)\rho_{2n}(\mathrm{d}w)$ and equation (B.57) continues to hold.

An example of interaction energy is given by "variance of homomorphism functions". As before, let H be a simple graph and W_n, W'_n be i.i.d. sampled $n \times n$ symmetric matrices from a graphon $[w]$. Consider the function

$$\mathbb{F}_n([w]) \coloneqq \frac{1}{2}\,\mathbb{E}\left[\left(\prod_{e\in E(H)} w(Z_e) - \prod_{e\in E(H)} w(Z'_e)\right)^2\right] = \mathrm{Var}\left[\prod_{e\in E(H)} w(Z_e)\right], \tag{4.38}$$

by symmetry, where $Z_e \coloneqq \big(Z_{e(1)}, Z_{e(2)}\big)$ and $Z'_e \coloneqq (Z'_{e(1)}, Z'_{e(2)})$ for all $e \in E(H)$. In fact, the above identity holds for whenever we take $R_n(z, z') = (g_n(z) - g_n(z'))^2$, for any function g_n on $\mathcal{M}_n$ that is square-integrable w.r.t. ρ_n. Unfortunately this particular function $\mathbb{F}_n$ in (4.38) is continuous in δ_2 but not in $\delta_\square$. See [Jan13, Theorem 10.15]. Hence, although the curve of maximal slope exists due to the existence of Fréchet-like derivatives, this particularly natural example does not satisfy the assumptions of our convergence theorem.

A similar but slightly different example of interaction which does satisfy our conditions can be constructed as follows. For a simple graph H with n vertices, consider its simple subgraphs H_1 and H_2 with n_1 and n_2 vertices respectively, such that every vertex and edge in H is contained in at least one of the subgraphs. Pool all the uniform random variables $\{Z_i\}_{i=1}^{n_1} \cup \{Z'_i\}_{i=1}^{n_2}$ to get a single set of $n_1 + n_2 \geq n$ i.i.d. Uni$[0,1]$ random variables $\{U_i\}_{i=1}^{n_1+n_2}$ such that $\{U_i\}_{i=1}^{n_1} = \{Z_i\}_{i=1}^{n_1}$ and $\{U_i\}_{i=n_1+1}^{n} = \big\{Z'_j\big\}_{j\in V(H)\setminus V(H_1)}$. Refer to Figure 4.1 for an illustration. We can then define $I_{H_1,H_2}(\,\cdot\,; H)\colon \widehat{\mathcal{W}}_\epsilon \to \mathbb{R} \cup \{\infty\}$ as



(a) H. (b) H_1. (c) H_2.

Figure 4.1: Illustration for the assignment of random variables $\{U_i\}_{i=1}^{n_1+n_2}$.

$$I_{H_1,H_2}([w];H) \coloneqq \log\left(\mathbb{E}\left[\prod_{\{i,j\}\in E(H_1)} w(Z_i,Z_j)\right]\right) + \log\left(\mathbb{E}\left[\prod_{\{i,j\}\in E(H_2)} w(Z'_i,Z'_j)\right]\right) - 2\log\left(\mathbb{E}\left[\prod_{\{i,j\}\in E(H)} w(U_i,U_j)\right]\right), \tag{4.39}$$

for some $\epsilon \in (0,1)$. Each of the terms in the expression in equation (4.39) is the logarithm of the homomorphism densities of a simple graph. Logarithms of homomorphisms are well studied in graph theory and, in particular, related to the max-cut problem (see [Lov12, Remark 5.4, Example 5.18]).

We can construct a graph H_1H_2 by forming the disjoint union of H_1 and H_2, identifying the vertices of the same label, adding all the edges between vertices with the same label according to [Lov12, Section 4.2]. H_1H_2 can have multiple edges. Then, using [Lov12, Proposition 7.1] for the homomorphism density as a simple graph parameter, we get that the determinant of the connection matrix of the homomorphism density

$$T_{H_1H_1}([W])T_{H_2H_2}([W]) - T^2_{H_1H_2}([W]) \geq 0, \quad \text{i.e.,} \quad \frac{T_{H_1H_1}([W])T_{H_2H_2}([W])}{T^2_{H_1H_2}([W])} \geq 1. \tag{4.40}$$

By the assumption that each vertex and edge of H is contained in at least one of the subgraphs, if we identify the multiple edges between the same pair of vertices in H_1H_2 we get back H (see Figure 4.1). Since T is a simple graph parameter, we have $T_{H_1H_2} = T_H$,

$T_{H_1H_1} = T_{H_1}$ and $T_{H_2H_2} = T_{H_2}$, by definition. Thus, taking logarithms in the final expression of (4.40), we get $I_{H_1,H_2}(\,\cdot\,; H) \geq 0$. It is exactly zero if H_1 and H_2 are vertex disjoint. Thus, one may think that $I_{H_1,H_2}(\,\cdot\,; H)$ measures the dependence of the two subgraphs on the homomorphism density.

We can similarly construct higher order interactions by considering multiple subgraphs instead of just two. In that case, one may consider the logarithm of the determinant of the connection matrix of the homomorphism density and define I suitably.

The argument in Section 4.6.1 shows that this function satisfy all our conditions. The computation for its Fréchet-like derivative also follows from Section 4.6.1 followed by the application of the chain rule for derivatives. Logarithms of determinants of matrices play an important role in displacement convexity of Wasserstein optimal transport (see [McC97, proof of Theorem 2.2]). It is an interesting coincidence that they also appear in this context.

4.6.3 Internal energy

Similar to potential and interaction energies, one can define non-linear functions of ρ_n corresponding to what are called 'internal energies' in the optimal transport literature. Let u be a real-valued function on $\mathbb{R}_+$ such that $u(0) = 0$. For a probability measure ρ on an Euclidean space, define the function

$$U(\rho) \coloneqq \begin{cases} \int_{\mathcal{M}_n} u(\rho(z))\,\mathrm{d}z, & \text{if } \rho \text{ is absolutely continuous,} \\ \infty, & \text{otherwise.} \end{cases}$$

Here, we have used the standard abuse of notation in the optimal transport literature of denoting an absolutely continuous measure and its density by the same notation. This defines a nonlinear function on graphons in the following manner. For $1 \leq l \leq n$, consider a function $G_{n,l} \colon \mathcal{M}_n \to [-1,1]^l$ that selects a particular subset of length l from the upper-diagonal elements of a $n \times n$ matrix. Consider the pushforward $\rho_{n,l}([w]) \coloneqq (G_{n,l})_\sharp\,(\rho_n([w]))$. Since ρ_n is generated by n i.i.d. Uni$[0,1]$ random variables, and $l \leq n$, it is easy to come up with examples where $\rho_{n,l}$ has a density. For example, take $w(x,y) = \sin(x+y)$, $n = 3, l = 2$, and

$G_{n,l}(A) = (A_{1,2}, A_{1,3})$. Thus $(G_{n,l} \circ G_n)([w]) = (\sin(Z_1 + Z_2), \sin(Z_1 + Z_3))$. This random vector has a probability density with respect to the Lebesgue measure. Thus, the function $U(\rho_{n,l}([w]))$ for $[w] \in \widehat{\mathcal{W}}$ has a non-empty domain. A prominent example of such functions is the differential entropy for which $u\colon x \mapsto x \log x$ where one computes the differential entropy of $\rho_{n,l}([w])$.

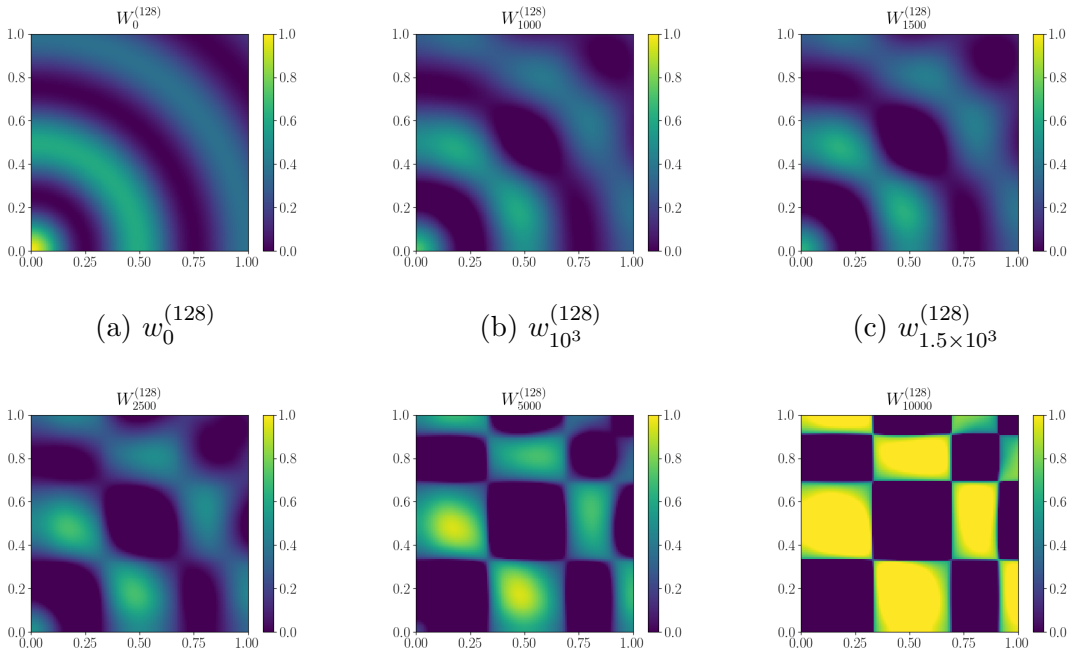
Such functions cannot have Fréchet-like derivatives since a necessary condition for its existence is that the function must be continuous in L^2. On the other hand, one cannot expect a discrete to continuum convergence as considered here. Even in the case of Wasserstein gradient flows, gradient flow of entropy is obtained by adding Brownian noise to particles and not by taking limits of Euclidean gradient flows [JKO98].

4.6.4 Computational example from Extremal Graph Theory

We give an interesting example in this section that suggests how the tools developed in this paper can be used to provide heuristics or search for counterexamples in many problems that are of interest in extremal graph theory.

Mantel's theorem [Man07] (a special case of Turán's theorem) states that the maximum number of edges in an n-vertex triangle-free graph is $n^2/4$. Further, any Hamiltonian graph with at least $n^2/4$ edges must either be the complete bipartite graph $K_{n/2,n/2}$ or it must be pancyclic [Bon71].

This suggests that if one maximizes the edge density subject to the condition that triangle density is 0, then the maximizer should correspond to a complete bipartite graph. Our current theory does not allow for such constrained optimization. However, one can attempt to computationally "verify" this result by simulating gradient flows of linear combinations of homomorphism functions, $T_\triangle - \alpha T_-$, over $(\widehat{\mathcal{W}}, \delta_2)$ for sufficiently small weight $\alpha > 0$. We make an arbitrary choice of weight, say $\alpha = 1/10$ for numerical simulation and consider minimizing $T_\triangle - T_-/10$ over $[w] \in \widehat{\mathcal{W}}$ such that $0 \leq w \leq 1$ a.e., where $\triangle$ and $-$ are the triangle and the edge graphs respectively. This is akin to minimizing triangle density while also maximizing the edge density as much as possible. We see that graphons with small



(a) $w_0^{(128)}$ (b) $w_{10^3}^{(128)}$ (c) $w_{1.5\times10^3}^{(128)}$

(d) $w_{2.5\times10^3}^{(128)}$ (e) $w_{5\times10^3}^{(128)}$ (f) $w_{10^4}^{(128)}$

Figure 4.2: A gradient descent simulation over $T_{\triangle} - T_{-}/10$

function values have small triangle density and large edge density. We set $n = 128$, step size $\tau = 10^{-3}$ and use the forward Euler method (i.e., Algorithm 2.1) starting from an initial graphon $\left[w_0^{(n)}\right] \in \widehat{\mathcal{W}}_n$ as shown in Figure 4.2a. Figure 4.2 shows instances of the iterative process after 10^3, 1.5×10^3, 2.5×10^3, 5×10^3 and 10^4 many steps. We see in Figure 4.2f that after 10^4 iteration, the kernel $w_{10^4}^{(n)}$ is close to the one corresponding to a complete bipartite graph as one would expect from Mantel's theorem. Our Theorem 4.28 implies that one should expect a similar evolution for all large values of n.

Linear combinations of the homomorphism density functions are non necessarily strictly convex (see Section 4.6) and hence the theory does not guarantee exponential rates of convergence. However, in the simulation discussed above, we statistically observe an exponential rate of convergence.

Such a computational and optimization driven approach can be used to study conjectures in extremal graph theory, especially for producing counterexamples. For instance, studying the non-negativity of a linear combination of homomorphism density is an important problem in extremal graph theory and it is known to be undecidable [Lov12, Section 16.6.1]. Although our techniques would not yield a proof but it can be a useful tool for the search of counterexamples. We leave this as an interesting direction for further investigation.

Chapter 5

MEASURE-VALUED GRAPHONS (MVG) AND INFINITE EXCHANGEABLE ARRAYS (IEA)

We have studied in Section 2.6 about infinite exchangeable arrays (IEAs). For two-dimensional IEAs, the Aldous and Hoover [Ald81, Hoo82, Ald82, Kal89] states that for every infinite symmetric exchangeable array $\mathbf{X} = (X_{i,j})_{(i,j)\in\mathbb{N}^{(2)}}$ there exists a Borel measurable function $f\colon [0,1]^4 \to \mathbb{R}$ satisfying $f(\,\cdot\,, x, y, \,\cdot\,) = f(\,\cdot\,, y, x, \,\cdot\,)$ for a.e. $(x,y) \in [0,1]^{(2)}$, and a collection of i.i.d. Uni$[0,1]$ random variables $U, \{U_i\}_{i\in\mathbb{N}}, \{U_{i,j} = U_{\{i,j\}}\}_{i,j\in\mathbb{N}}$ on some probability space such that the IEA $\mathbf{Y}$ defined as $Y_{i,j} = f(U, U_i, U_j, U_{\{i,j\}})$ for all $(i,j) \in \mathbb{N}^{(2)}$, has the same distribution as $\mathbf{X}$. Therefore, the Aldous-Hoover representation of an IEA $\mathbf{X}$ induces a random graphon $w^{(U)}$, defined as

$$w^{(u)}(x,y) \coloneqq \mathbb{E}[f(U, U_1, U_2, U_{1,2}) \mid U = u, (U_1, U_2) = (x,y)], \tag{5.1}$$

for $(x,y) \in [0,1]^{(2)}$ and $u \in [0,1]$.

In [DJ08, Section 5] it is shown that, when the entries in IEA take values in $\{0,1\}$, there is a one-to-one correspondence between IEAs and probability distributions on the space of graphons. However, when the IEA takes more nontrivial values, this is no longer the case. We illustrate this via an example.

Example 5.1. Let $\mathbf{G} = (G_{i,j})_{(i,j)\in\mathbb{N}^{(2)}}$ be an IEA such that every $G_{i,j}$ is an i.i.d. Ber$(1/2)$ random variables. Let $\mathbf{K} = (K_{i,j})_{(i,j)\in\mathbb{N}^{(2)}}$ be a constant IEA such that $K_{i,j} \equiv 1/2$ for all $i,j \in \mathbb{N}$. Both the IEAs $\mathbf{G}$ and $\mathbf{K}$ correspond to a constant graphon $w_p \equiv 1/2$.

In this chapter, we propose a remedy for this by expanding the definition of a graphon to take values in $\mathcal{P}([-1,1])$, the set of Borel probability measures on $[-1,1]$. We call them measure-valued graphons (MVGs).

Although MVGs have already been introduced in the literature by [LS10], our contribution here are twofold.

1. First, we define two cut-like metrics, $\Delta_{\blacksquare}$ and $\mathbb{W}_{\blacksquare}$ (see Definitions 5.10 and 5.11), that capture the topology of convergence as in [LS10] under which the space of MVGs is a compact topological space. See Theorem 5.15 and Examples 5.2-5.5 where we illustrate the strengths and nuances of this convergence.

2. Second, we relate this convergence to the convergence of infinite exchangeable arrays (IEAs). An (IEA) $\mathbf{X}$, see [Kal05, Chapter 7], is a doubly-indexed sequence of random variables $(X_{i,j})_{(i,j)\in\mathbb{N}^{(2)}}$ defined on a single probability space whose joint distribution is invariant under finite permutations of its rows and columns. That is, if ς is any finite permutation on $\mathbb{N}$, then the joint distribution of $\left(X_{\varsigma i,\varsigma j}\right)_{(i,j)\in\mathbb{N}^{(2)}}$ is the same as that of the original array. It follows from the Aldous-Hoover representation theorem [Kal89], that the IEAs are in one-to-one correspondence with random measure-valued graphons (MVG). In Theorem 5.21 we prove a homeomorphism between the set of probability measures on the space of MVGs and the space of laws of IEA, each equipped with the corresponding weak topology. This is a generalization of [DJ08, Theorem 5.3] suited to our applications below.

We note that after the contributions made in [APST23], the authors in [ADW23] also developed a similar metric on MVGs (which they call probability graphons) that captures the same topology. The authors in [ADW23] consider the convergence of weighted graphs which can take edge-weights in a general Polish space and show that the topology induced on probability graphons is independent of the metric on the Polish space.

5.1 Measure-valued graphons

Just like we define kernels in 2.5, we can extend it to define measure-valued kernels.

Definition 5.1 (Measure-valued kernel)**.** A *measure-valued kernel* is a measurable function $W\colon [0,1]^{(2)} \to \mathcal{P}([-1,1])$ such that $W(x,y) = W(y,x)$ for a.e. $(x,y) \in [0,1]^{(2)}$. Here $\mathcal{P}([-1,1])$ is the space of probability measures on the interval $[-1,1]$ equipped with the Borel sigma-algebra generated by the topology of weak convergence. We will denote the set of all measure-valued kernels by $\mathfrak{W}$.

Measure-valued graphons are defined similarly later in Definition 5.4 where we denote by $\widehat{\mathfrak{W}}$, the set of all graphons. More details can be found in [LS10, KKLS14] where a notion of convergence based on *decorated* homomorphism functions is also discussed. In this paper we show that this convergence is metrizable via a metric similar to the Wasserstein metric for probability measures.

Definition 5.2 (Natural projection from $\widehat{\mathfrak{W}}$ to $\widehat{\mathcal{W}}$)**.** Given a measure-valued kernel $W \in \mathfrak{W}$, we can define a corresponding kernel $w \in \mathcal{W}$ defined as $w(x,y) \coloneqq \int_{[-1,1]} \zeta \, W(x,y)(\mathrm{d}\zeta)$ for a.e. $(x,y) \in [0,1]^{(2)}$. This naturally defines a projection from $\widehat{\mathfrak{W}}$ to $\widehat{\mathcal{W}}$. We will often refer to this projection as *natural projection* and denote $w = \mathbb{E}[W]$. This map from $(\widehat{\mathfrak{W}}, \Delta_{\blacksquare})$ to $(\widehat{\mathcal{W}}, \delta_{\square})$ is 1-Lipschitz as seen from Definition 5.10.

Suppose an IEA $\mathbf{X}$ has an Aldous-Hoover representation given by a function f. Analogous to equation (5.1), one can define a random measure-valued kernel $W \in \mathfrak{W}$ (and hence an MVG) as follows. For $(x,y) \in [0,1]^{(2)}$, set

$$W^{(u)}(x,y) \coloneqq \mathrm{Law}(X_{i,j} \,|\, U = u, (U_i, U_j) = (x,y)), \quad u \in [0,1]. \tag{5.2}$$

Conversely, an MVG W generates an IEA in the following manner. Let $(U_i)_{i\in\mathbb{N}}$ denote an i.i.d. sequence of Uni$[0,1]$ random variables. Define an IEA $\mathbf{X}$, where, given $(U_i = u_i)_{i\in\mathbb{N}}$, $X_{i,j}$ for $(i,j) \in \mathbb{N}^{(2)}$, is independently sampled from the probability measure $W(U_i, U_j) \in \mathcal{P}([-1,1])$. This indeed remedies the issue raised in Example 5.1 with IEAs $\mathbf{G}$ and $\mathbf{K}$. The MVG corresponding to $\mathbf{G}$ is $W_{\mathbf{G}}(x,y) \equiv \frac{1}{2}(\delta_0 + \delta_1)$ while the MVG corresponding to K is $W_{\mathbf{K}}(x,y) \equiv \delta_{1/2}$, where $\delta_\cdot$ refers to the delta mass. Note that the random graphon w corresponding to IEA $\mathbf{X}$ in (5.1) can be obtained from the random MVG W as defined

in (5.2) via the natural projection of W. If we take $n \times n$ blocks of the IEA $\mathbf{X}$, one can show that, as $n \to \infty$, this sequences of random exchangeable matrices converge to deterministic limits in the spaces of both graphons and measure-valued graphons. The two limits are related by the natural projection.

This correspondence between IEAs and random MVGs recovers the results in [DJ08] for general exchangeable arrays. In Section 5.1, we define the space of MVGs and notion of convergence that defines the topology on the space of MVGs. We introduce two new metrics on MVGs analogous to the usual cut metric and the Wasserstein metric respectively. In Theorem 5.15 we show the equivalence of all these. We, then, prove a correspondence theorem between the compact metric space of probability measures on MVGs and the space of IEA equipped with the weak topology in Theorem 5.21. This allows us to consider processes on exchangeable matrices/graphs and take their limits either as IEAs or as MVGs. This is what we will do this in Chapter 6 to obtain scaling limits of iterative algorithms as one of the major contributions of this thesis.

As mentioned already in Section 2.5, the convergence of the homomorphism density functions T_F, for simple graphs F, can be used to define a notion of convergence for weighted graphs as well. However, a better approach to the convergence of weighted graphs is given by the convergence of *decorated homomorphism density functions* that we describe below (see [LS10]). In the following, we will use I to denote the compact interval $[-1, 1]$ and $\mathcal{C} \equiv C(I)$, to denote the space of continuous functions on I.

Definition 5.3 (Decorated graph [LS10, Section 2.1])**.** Let $m \geq 1$ and $\mathcal{D} \subseteq \mathcal{C}$. Let $F = ([m], E)$ be a simple graph. The pair (F, f) is called a $\mathcal{D}$-*decorated graph* where $f \colon E(F) \to \mathcal{D}$ is a function from the edges $E(F)$ of F to $\mathcal{D}$. We will refer to F as the *skeleton* and f as the *decoration* of the decorated graph (F, f). If there is no confusion, the decoration of a graph will be implicitly assumed without mention and we will denote $f(\{i, j\})$ by $F_{i,j}$ for $\{i, j\} \in E(F)$.

Throughout this chapter, a decorated graph will mean a $\mathcal{C}$-decorated graph unless stated

otherwise. Let $W \in \mathfrak{W}$ be a measure-valued kernel (see Definition 5.1) and F a decorated graph. Following [LS10, Section 2.5] one can define the (decorated) homomorphism density $T_{\mathrm{d}}(F, W)$ of F in W as

$$T_{\mathrm{d}}(F, W) \coloneqq \int_{[0,1]^m} \left(\prod_{\{j,k\} \in E(F)} \int_{[-1,1]} F_{j,k}(\zeta) W(x_j, x_k)(\mathrm{d}\zeta) \right) \prod_{i=1}^{m} \mathrm{d}x_i. \tag{5.3}$$

We are now ready to define measure-valued graphons.

Definition 5.4 (Measure-valued graphon [LS10, Definition 2.4])**.** Define an equivalence relation $\sim$ on $\mathfrak{W}$ such that $W \sim U$ if $T_{\mathrm{d}}(F, W) = T_{\mathrm{d}}(F, U)$ for every decorated graph F. Let $\widehat{\mathfrak{W}} \coloneqq \mathfrak{W}/\!\sim$ be equipped with the weakest topology that makes $W \mapsto T_{\mathrm{d}}(F, W)$ continuous for every decorated graph F. We will call $\widehat{\mathfrak{W}}$ (equipped with this topology), the space of measure-valued graphons. A measure-valued graphon (MVG) is an element in $\widehat{\mathfrak{W}}$. Naturally, $W_n \to W$ in $\widehat{\mathfrak{W}}$ if $T_{\mathrm{d}}(F, W_n) \to T_{\mathrm{d}}(F, W)$ for every $\mathcal{C}$-decorated graph F. We refer to this topology as the *usual topology* on MVG throughout this paper.

Analogous to the space of kernels $\mathcal{W}$, one defines an equivalence relation $\cong$ on $\mathfrak{W}$ such that $W_1 \cong W_2$ if there exist measure preserving transformations $\varphi_1, \varphi_2 \colon [0,1] \to [0,1]$ and $W \in \mathfrak{W}$ such that $W_1 = W^{\varphi_1}$, and $W_2 = W^{\varphi_2}$. It follows from [KK19, Theorem 11(ii)] that $W_1 \cong W_2$ if and only if $T_{\mathrm{d}}(F, W_1) = T_{\mathrm{d}}(F, W_2)$ for every decorated graph F. In particular, the space of MVGs can be equivalently defined as $\widehat{\mathfrak{W}} \coloneqq \mathfrak{W}/\!\cong$. Unlike as it were necessary in Chapter 4 to distinguish, wherever it is clear from the context, for any measure-valued kernel $W \in \mathfrak{W}$, we will use an abuse of notation and use the same symbol W to denote the equivalence class, or the measure-valued graphon, corresponding to the measure-valued kernel.

5.1.1 Embedding matrices and graphons into MVG

Recall that a weighted graph or (equivalently a symmetric matrix) can be identified with a kernel (and hence a graphon). Similarly, a weighted graph or a graphon can be identified

with a measure-valued kernel (and hence a measure-valued graphon). Let M be an $n \times n$ symmetric matrix with entries in $I = [-1, 1]$. Let $\mathcal{M}_n$ denote the set of all such matrices. Let F be a $\mathcal{C}$-decorated graph on $[m]$ vertices. One can define the homomorphism density of F in M, denoted $T_{\mathrm{d}}(F, W)$, as

$$T_{\mathrm{d}}(F, M) \coloneqq \frac{1}{n^m} \sum_{i_1,\ldots,i_m} \prod_{\{j,k\} \in E(F)} F_{j,k}(M_{i_j,i_k}), \tag{5.4}$$

where the summation runs over all indices $i_1, i_2, \ldots, i_m$ taking values in $[n]$. We make some simple observations. Observe that $T_{\mathrm{d}}(F, M) = T_{\mathrm{d}}(F, M^\sigma)$ where σ is any permutation of $[n]$ and $M^\sigma_{i,j} = M_{\sigma(i),\sigma(j)}$ for all $(i, j) \in [n]^{(2)}$. Also note that one can naturally associate a measure-valued kernel, say $W_M \in \mathfrak{W}$, with a symmetric matrix $M \in \mathcal{M}_n$ as follows. For $n \in \mathbb{N}$, let $Q_n \coloneqq \{Q_{n,i}\}_{i\in[n]}$ be a partition of the interval $[0, 1]$ into contiguous intervals of equal length as defined in Section 2.5. Set $W_M(x, y) = \delta_{M_{i,j}}$ whenever $(x, y) \in Q_{n,i} \times Q_{n,j}$ for some $(i, j) \in [n]^{(2)}$. For any decorated graph F we have $T_{\mathrm{d}}(F, W_M) = T_{\mathrm{d}}(F, M)$. Therefore, when there is no scope of ambiguity we make no distinction between a symmetric matrix M and the corresponding MVG W_M. Similarly, if $w \in \mathcal{W}$ is a graphon, we can define a corresponding MVG, say W by setting $W(x, y) = \delta_{w(x,y)}$ for a.e. $(x, y) \in [0, 1]^{(2)}$ and the notion of homomorphism density extends naturally. We denote this map taking a matrix/graphon to the corresponding MVG by $\mathcal{K}$.

Let $(M_n \in \mathcal{M}_n)_{n\in\mathbb{N}}$ be a sequence of matrices with growing dimension and let $W \in \widehat{\mathfrak{W}}$. We say that $(M_n)_{n\in\mathbb{N}} \to W$ as $n \to \infty$ if $T_{\mathrm{d}}(F, M_n) \to T_{\mathrm{d}}(F, W)$ as $n \to \infty$ for every decorated graph F. In particular, we will often say $(M_n)_{n\in\mathbb{N}} \to W$ as $n \to \infty$ where $(M_n)_{n\in\mathbb{N}}$ is a sequence of matrices, or $(w_n)_{n\in\mathbb{N}} \to W$ as $n \to \infty$ where $(w_n)_{n\in\mathbb{N}}$ is a sequence of graphons and W is an MVG. It is to be understood that this convergence is with respect to decorated graphs, or equivalently these statements mean that MVG corresponding to M_n (or w_n) converge to W in $\widehat{\mathfrak{W}}$ as $n \to \infty$. For an $n \times n$ finite exchangeable random matrix X, we can define a measure valued kernel W_X as $W_X(x, y) = \mathrm{Law}(X_{i,j})$ whenever $(x, y) \in Q_{n,i} \times Q_{n,j}$ for some $(i, j) \in [n]^{(2)}$. We will denote this map by $\mathcal{K}$, i.e., $\mathcal{K}(X) = W_X$. Note that the measure valued kernel cannot recover the joint distribution among the entries

of X, unless they are mutually independent.

Remark 5.5. Recall that $W \mapsto w \coloneqq \mathbb{E}[W]$ is Lipschitz (see Definition 5.2). It follows that if $(M_n)_{n\in\mathbb{N}}$ is a sequence of matrices such that $(M_n)_{n\in\mathbb{N}} \to W$ as $n \to \infty$ for some $W \in \mathfrak{W}$ in the MVG sense, then $(M_n)_{n\in\mathbb{N}} \to w$ as $n \to \infty$ in cut-metric as well. This illustrates that convergence in MVG sense is stronger than cut convergence.

5.1.2 Topology on MVGs

In this section we introduce an alternate notion of convergence for MVGs and two metrics on $\widehat{\mathfrak{W}}$. We then show that this new notion of convergence and the metrics introduced in this section give the same topology on $\widehat{\mathfrak{W}}$ as defined in Definition 5.4.

Definition 5.6 (Homomorphism density)**.** Let F be a finite simple connected graph and let $W \in \mathfrak{W}$. The *homomorphism density* of F into W, denoted $T_F(W)$, is a probability measure on $I_F \coloneqq [-1,1]^{E(F)}$ is defined as a mixture of probability measures as

$$T_F(W) \coloneqq \int_{[0,1]^{V(F)}} \bigotimes_{\{i,j\}\in E(F)} W(x_i, x_j) \prod_{v\in V(F)} \mathrm{d}x_v. \tag{5.5}$$

The measure in equation (5.5) is to be interpreted as the unique measure on I_F such that for any bounded measurable function $\varphi\colon I_F \to \mathbb{R}$, we have

$$\int_{I_F} \varphi(\zeta) T_F(W)(\mathrm{d}\zeta) = \int_{[0,1]^{V(F)}} \left(\int_{I_F} \varphi(\zeta) \bigotimes_{\{i,j\}\in E(F)} W(x_i, x_j)(\mathrm{d}\zeta) \right) \prod_{k\in V(F)} \mathrm{d}x_k. \tag{5.6}$$

Let $\{U_i\}_{i\in V(F)}$ be a collection of i.i.d. Uni$[0,1]$ random variables, then

$$\int_{I_F} \varphi(\zeta) T_F(W)(\mathrm{d}\zeta) = \mathbb{E}\left[\left\langle \varphi, \bigotimes_{\{i,j\}\in E(F)} W(U_i, U_j) \right\rangle \right], \tag{5.7}$$

where $\langle \cdot, \cdot \rangle$ is the usual duality between continuous functions on I_F and probability measures on I_F and expectation is taken with respect to the random variables $\{U_i\}_{i\in\mathbb{N}}$. It is important to note that the homomorphism density of a simple graph F into a graphon w (see Section 2.5), $T_F(w)$ is a real number in $[0,1]$ whereas the homomorphism density of a

simple graph F into an MVG W, $T_F(W)$, is a (mixture) of probability measures. Secondly, in the context of MVGs, $T_F(W)$ is defined for a simple graph F and it is a (mixture of) probability measure on I_F, on the other hand $T_{\mathrm{d}}(F, W)$ is defined for a decorated graph F and it is a real number.

Definition 5.7 (Convergence of MVGs)**.** A sequence of MVGs $(W_n)_{n\in\mathbb{N}}$ converge to a MVG W *in hom-density sense* if $\lim_{n\to\infty} T_F(W_n) = T_F(W)$ weakly for every finite simple graph F.

The above definition naturally extends to any measure-valued symmetric matrix M. And, using the embedding defined in Section 5.1.1, the definition can be naturally extended to symmetric matrices and graphons. We skip the details to avoid repetitions.

5.1.3 Two metrics on MVGs

We now introduce the metrics on MVGs. Let $\mathcal{L}$ be the set of all Lipschitz functions $\psi\colon [-1,1] \to \mathbb{R}$ with bounded Lipschitz norm, i.e., $\|\psi\|_{\mathrm{BL}} = \max\{\|\psi\|_\infty, \|\psi\|_{\mathrm{Lip}}\} \le 1$. Define an operator $\Gamma\colon \mathcal{L} \times \mathfrak{W} \to \mathcal{W}$ defined as

$$\Gamma(\psi, W)(x, y) \coloneqq \int_{-1}^{1} \psi(\zeta) W(x, y)(\mathrm{d}\zeta). \tag{5.8}$$

Definition 5.8 (Generalized Cut norm on $\mathfrak{W}$)**.** For any $W \in \mathfrak{W}$, define $\|\cdot\|_{\blacksquare}\colon \mathfrak{W} \to \mathbb{R}_+$ as

$$\|W\|_{\blacksquare} \coloneqq \sup_{\psi\in\mathcal{L}} \|\Gamma(\psi, W)\|_{\square},$$

where Γ is as defined in (5.8).

Remark 5.9. Recall from Section 5.1.1 that both a kernel and a finite matrix can be associated with a corresponding MVG. With this association, we can reference $\|w\|_{\blacksquare}$ or $\|A\|_{\blacksquare}$ for $w \in \mathcal{W}$ or $A \in \cup_{r\in\mathbb{N}} \mathcal{M}_r$. That is, the definition of *generalized cut norm* extends to both kernels and matrices. We'll adopt this notation moving forward.

Recall that $\mathcal{T}$ is the set of all Lebesgue measure preserving maps $\varphi\colon [0,1] \to [0,1]$ and for any $W \in \mathfrak{W}$ and $\varphi \in \mathcal{T}$, we define $W^\varphi \in \mathfrak{W}$ as $W^\varphi(x, y) \coloneqq W(\varphi(x), \varphi(y))$ for a.e. $(x, y) \in [0,1]^{(2)}$.

Definition 5.10 (Generalized Cut metric on $\widehat{\mathfrak{W}}$)**.** Define $\Delta_{\blacksquare}\colon \widehat{\mathfrak{W}} \times \widehat{\mathfrak{W}} \to \mathbb{R}_+$ as

$$\Delta_{\blacksquare}(W_1, W_2) \coloneqq \inf_{\varphi_1,\varphi_2\in\mathcal{T}} \|W_1^{\varphi_1} - W_2^{\varphi_2}\|_{\blacksquare}, \qquad W_1, W_2 \in \widehat{\mathfrak{W}}.$$

Let μ_1 and μ_2 be two finite measures with the same total mass m. The extension of the Wasserstein distance between μ_1 and μ_2 is defined as $\mathbb{W}_1(\mu_1, \mu_2) = \sup_{\psi\in\mathcal{L}} \int \psi \,\mathrm{d}(\mu_1 - \mu_2)$, where $\mathcal{L}$ is the set of all bounded Lipschitz functions with bounded Lipschitz norm at most 1. Since we are working with a bounded metric space, this definition is equivalent to the standard definition (see [Vil03, Section 1.2.1, Corollary 1.16]) which considers all Lipschitz functions.

Definition 5.11 (Wasserstein Cut metric on $\widehat{\mathfrak{W}}$)**.** Define $\mathbb{W}_{\blacksquare}\colon \widehat{\mathfrak{W}} \times \widehat{\mathfrak{W}} \to \mathbb{R}_+$ as

$$\mathbb{W}_{\blacksquare}(W_1, W_2) \coloneqq \inf_{\varphi_1,\varphi_2\in\mathcal{T}} \sup_{S,T\subseteq[0,1]} \mathbb{W}_1\left(\int_{S\times T} W_1^{\varphi_1}(x,y)\,\mathrm{d}x\,\mathrm{d}y, \int_{S\times T} W_2^{\varphi_2}(x,y)\,\mathrm{d}x\,\mathrm{d}y\right).$$

We also make the following definition. For $W_1, W_2 \in \mathfrak{W}$, define the Wasserstein-2 metric D_2 on $\mathfrak{W}$ as

$$D_2^2(W_1, W_2) \coloneqq \int_{[0,1]^2} \mathbb{W}_2^2(W_1(x,y), W_2(x,y))\,\mathrm{d}x\,\mathrm{d}y, \tag{5.9}$$

where $\mathbb{W}_2$ is the Wasserstein-2 metric on $\mathcal{P}([-1,1])$.

Definition 5.12 (Invariant Wasserstein-2 metric on $\widehat{\mathfrak{W}}$)**.** Define $\Delta_2\colon \widehat{\mathfrak{W}} \times \widehat{\mathfrak{W}} \to \mathbb{R}_+$ as

$$\Delta_2^2(W_1, W_2) \coloneqq \inf_{\varphi_1,\varphi_2\in\mathcal{T}} D_2^2(W_1^{\varphi_1}, W_2^{\varphi_2}), \qquad W_1, W_2 \in \widehat{\mathfrak{W}}.$$

Proposition 5.13. *Let $\mathbb{W}_{\blacksquare}$ and $\Delta_{\blacksquare}$ be as defined in Definitions 5.10 and 5.11. Then, $\mathbb{W}_{\blacksquare}$ and $\Delta_{\blacksquare}$ are metrics on $\widehat{\mathfrak{W}}$. Furthermore, $\mathbb{W}_{\blacksquare} = \Delta_{\blacksquare}$.*

The proof of Proposition 5.13 is provided in Appendix C.1.

Lemma 5.14 shows that $\Delta_{\blacksquare} \le \Delta_2$.

Lemma 5.14. $\Delta_{\blacksquare} \le \Delta_2$.

The proof of Lemma 5.14 is provided in Appendix C.1.

We can now state our first main result.

Theorem 5.15 ([APST23])**.** *Let $W, (W_n)_{n\in\mathbb{N}} \subset \widehat{\mathfrak{W}}$. Then, the following limits are equivalent, as $n \to \infty$.*

1. *$W_n \to W$ in $\widehat{\mathfrak{W}}$, that is, $T_{\mathrm{d}}(F, W_n) \to T_{\mathrm{d}}(F, W)$ for every decorated graph F.*
2. *$W_n \to W$ in homomorphism density sense, i.e., $T_F(W_n) \to T_F(W)$ weakly for every finite simple graph F.*
3. *$\Delta_{\blacksquare}(W_n, W) \to 0$.*
4. *$\mathbb{W}_{\blacksquare}(W_n, W) \to 0$.*

The proof of Theorem 5.15 is provided in Appendix C.1.

Remark 5.16. It follows from [Lov12, Theorem 17.9] that $\widehat{\mathfrak{W}}$ is compact (with respect to the usual topology). In order to apply that theorem, notice that our MVG is a K-graphon in the terminology of Lovász for $K = [-1, 1]$. The set $\mathcal{B}$ can be taken to the countable set of polynomials. By Theorem 5.15, $\left(\widehat{\mathfrak{W}}, \mathbb{W}_{\blacksquare}\right)$ (or $(\widehat{\mathfrak{W}}, \Delta_{\blacksquare})$) is a compact metric space.

It is clear from Definition 5.10 and Theorem 5.15 that if $(W_n)_{n\in\mathbb{N}} \to W$ in $\widehat{\mathfrak{W}}$ then $(\Gamma(\psi, W_n))_{n\in\mathbb{N}} \to \Gamma(\psi, W)$ in $\delta_{\square}$ for every bounded continuous function ψ defined on $[-1, 1]$. However, the convergence $W_n \to W$ is stronger since it implies *simultaneous convergence* of all kernels $\Gamma(\psi, W)$. We now give some examples to illustrate the difference between the convergence of graphons and the convergence of MVGs.

Example 5.2. For $k \in \mathbb{Z}_+$, let $\psi_k \colon [-1, 1] \to \mathbb{R}$ be the map given by $\zeta \mapsto \zeta^k$. Let $W \in \mathfrak{W}$. We will call $\Gamma(\psi_k, W)$ the *moment graphon* of W (if we need to emphasize k, we will call it *k-th moment graphon*). For simplicity, we will also denote $\Gamma(\psi_k, W)$ by $m_k(W)$. It is easy to see that $(W_n)_{n\in\mathbb{N}} \to W$ in $\widehat{\mathfrak{W}}$ as $n \to \infty$ implies $(m_k(W_n))_{n\in\mathbb{N}} \to m_k(W)$ in $\delta_{\square}$

as $n \to \infty$, for all $k \in \mathbb{Z}_+$. Since the convergence in $\delta_\square$ metric implies that for each k, there is a sequence of Lebesgue measure preserving transforms $\varphi_{n,k}\colon [0,1] \to [0,1]$ such that $\left\|m_k(W_n^{\varphi_{n,k}}) - m_k(W)\right\|_\square \to 0$ as $n \to \infty$. However, $W_n \to W$ in $\widehat{\mathfrak{W}}$ as $n \to \infty$ implies that $\varphi_{n,k}$ could be chosen to be independent of k. I.e., there exists a sequence of common 'labellings' (φ_n) such that $\|m_k(W_n^{\varphi_n}) - m_k(W)\|_\square \to 0$ as $n \to \infty$. This is what we mean by simultaneous convergence.

Example 5.3. Consider a sequence of kernels $(a_n)_{n\in\mathbb{N}\cup\{\infty\}}$, i.e. $a_n\colon [0,1]^{(2)} \to [-1,1]$, for $n \in \mathbb{N}\cup\{\infty\}$. For every $n \in \mathbb{N}$, define a measure-valued kernel $W_n \in \widehat{\mathcal{W}}$ by setting $W_n(x,y) = \delta_{a_n(x,y)}$, $(x,y) \in [0,1]^2$. Let $\psi \in C(I)$ be a continuous test function such that $\|\psi\|_\infty \le 1$. Then $\Gamma(\psi, W_n)(x,y) = \psi(a_n(x,y))$, $(x,y) \in [0,1]^{(2)}$. Suppose $(W_n)_{n\in\mathbb{N}} \to W_\infty$ in $\widehat{\mathfrak{W}}$, then $(\Gamma(\psi, W_n))_{n\in\mathbb{N}} \to \Gamma(\psi, W_\infty)$ in $\delta_\square$. In particular, taking $\psi(z) = z^k$, for every $k \in \mathbb{N}$, it follows that simultaneously $\left(a_n^k\right)_{n\in\mathbb{N}} \to a_\infty^k$ in $\delta_\square$. It is well-known that $\delta_\square(a_n, a) \to 0$ does not imply $\delta_\square(a_n^2, a^2) \to 0$ in general. This illustrates that convergence in the MVG sense is a stronger notion than the cut convergence.

Example 5.4. Let $a\colon [0,1]^{(2)} \to [0,1]$ be a kernel. Define a measure-valued kernel W_a as $W_a(x,y) \coloneqq (1-a(x,y))\delta_0 + a(x,y)\delta_1$ for $(x,y) \in [0,1]^{(2)}$. That is, $W(x,y)$ is $\mathrm{Ber}(a(x,y))$ for $(x,y) \in [0,1]^2$. Let ψ be any bounded measurable function on $[0,1]$. Then, $\Gamma(\psi, W_a)(x,y) = (1-a(x,y))\psi(0) + a(x,y)\psi(1)$. If $(a_n)_{n\in\mathbb{N}}$ is a sequence of graphons such that $(W_{a_n})_{n\in\mathbb{N}} \to W_a$ then $(a_n)_{n\in\mathbb{N}} \to a$ in $\delta_\square$. Conversely, in this example, it is easy to verify that if $(a_n)_{n\in\mathbb{N}} \to a$ in $\delta_\square$ then $\sup_\psi \|\Gamma(\psi, a_n) - \Gamma(\psi, a)\|_\square \to 0$ as $n \to \infty$ where the supremum is taken over all continuous and bounded functions $\psi \in C([0,1])$. In particular, we conclude that $(W_{a_n})_{n\in\mathbb{N}} \to W_a$ in $\widehat{\mathfrak{W}}$ if and only if $(a_n)_{n\in\mathbb{N}} \to a$ in $\delta_\square$.

Example 5.5. Let $a, b \in \mathcal{W}$ such that $a(x,y) \ge 0, b(x,y) \ge 0$ and $a(x,y) + b(x,y) \le 1$. Define a "ternoulli" MVG as $W_{a,b}(x,y) = a(x,y)\delta_{-1} + (1 - a(x,y) - b(x,y))\delta_0 + b(x,y)\delta_{-1}$. If $(W_n)_{n\in\mathbb{N}} \to W_{a,b}$ as $n \to \infty$ in $\widehat{\mathfrak{W}}$ then $(a_n)_{n\in\mathbb{N}} \to a, (b_n)_{n\in\mathbb{N}} \to b$ as $n \to \infty$. Conversely, suppose that (a_n, b_n) are "coupled graphons" and $(a_n)_{n\in\mathbb{N}} \to a$ and $(b_n)_{n\in\mathbb{N}} \to b$ under a common labeling as $n \to \infty$. I.e., there exists a sequence $\varphi_n \in \mathcal{T}$ such that $\|a_n^{\varphi_n} - a\|_\square +$

$\|b_n^{\varphi_n} - b\|_\square \to 0$ as $n \to \infty$. Note that there exists a common sequence of measure-preserving transforms for both a_n and b_n. Then, $(W_{a_n,b_n})_{n\in\mathbb{N}} \to W_{a,b}$ as $n \to \infty$.

Sampling from MVGs

Recall that one can generate an IEA from an MVG. We now describe a similar procedure to generate a measure-valued random matrix from an MVG. Let $(U_i)_{i\in\mathbb{N}}$ be an i.i.d. sequence of Uni$[0,1]$ random variables defined on a common probability space, say $(\Omega, \mathcal{F}, \mathbb{P})$. Let $W \in \mathfrak{W}$. For any $n \in \mathbb{N}$ we define the *sampled n-MVG*, denoted $\mu(n, W)$, as

$$\mu(n,W)(i,j) \coloneqq W(U_i, U_j), \qquad (i,j) \in [n]^{(2)}. \tag{5.10}$$

Note that we can identify $\mu(n, W)$ with a random MVG. In the next lemma we show that the random MVG $\mu(n, W)$ converges to W as $n \to \infty$, $\mathbb{P}$-almost surely.

Lemma 5.17. *Let $W \in \widehat{\mathfrak{W}}$. For $n \in \mathbb{N}$, let $\mu(n, W)$ be defined as in* (5.10)*. Then $\mu(n, W) \to W$ in $\widehat{\mathfrak{W}}$ as $n \to \infty$, $\mathbb{P}$-almost surely. That is, $\mathbb{P}$-almost surely, for any finite simple graph F we have $T_F(\mu(n, W)) \to T_F(W)$, weakly as $n \to \infty$.*

Lemma 5.17 follows directly from [KKLS14, Theorem 3.8] by taking $\mathcal{B} = C[-1, 1]$ and $\mathcal{Z} = M([-1, 1])$ the space of finite Radon measures on $[-1, 1]$. We, therefore, skip the proof of Lemma 5.17.

We now describe a sampling procedure to generate weighted graphs from an MVG. Let $n \in \mathbb{N}$ and $W \in \mathfrak{W}$. For every $n \in \mathbb{N}$, define $\mathbb{G}(n, W)$ to be a random (weighted) graph on $[n]$ with edge-weights $\mathbb{G}(n, W)(i, j) \sim W(U_i, U_j)$ and are conditionally independent given (U_i, U_j) for every $(i, j) \in [n]^{(2)}$. Note that the adjacency matrix of $\mathbb{G}(n, W)$ is a random $n \times n$ symmetric matrix with entries in $I = [-1, 1]$ and we will not make any distinction between the adjacency matrix and the graph. Lemma 5.18 shows that almost surely $\mathbb{G}(n, W) \to W$ as $n \to \infty$ (see Section 5.1.1).

Lemma 5.18. *Let $W \in \widehat{\mathfrak{W}}$ and let $\mathbb{G}(n, W)$ be defined for every $n \in \mathbb{N}$ as above. Then, $\mathbb{P}$-almost surely, $\mathbb{G}(n, W) \to W$ as $n \to \infty$. That is, $\mathbb{P}$-almost surely, for every decorated*

graph F,

$$T_{\mathrm{d}}(F, \mathbb{G}(n, W)) \to T_{\mathrm{d}}(F, W), \qquad \text{as } n \to \infty.$$

The proof of Lemma 5.18 is provided in Appendix C.1.

An immediate consequence of Lemma 5.18 is that every MVG can be obtained as the limit of finite weighted graphs. In fact, MVGs were introduced as the limits of finite weighted in [LS10, Section 2.5].

5.2 Infinite Exchangeable Arrays

Recall the correspondence between IEAs and random MVGs described in the Introduction. In this section, we formalize that correspondence.

Let $\mathcal{S}$ be the set of all symmetric infinite arrays with their elements taking values in $[-1, 1]$ with 0 diagonal. That is, let

$$\mathcal{S} \coloneqq \Big\{ \mathbf{x} \in \mathbb{R}^{\mathbb{N}^2} \;\Big|\; x_{i,j} = x_{j,i} \in [-1, 1], \;\; x_{i,i} = 0 \quad \forall \;\; i, j \in \mathbb{N} \Big\}.$$

Equip $\mathcal{S}$ with the product topology under which it is compact. Equip $\mathcal{S}$ with the corresponding Borel sigma-algebra. Let Π_∞ be the set of all finite permutations of $\mathbb{N}$. Observe that Π_∞ has a natural action on $\mathcal{S}$ given by $\mathbf{x}^\sigma(i, j) \coloneqq \mathbf{x}(\sigma(i), \sigma(j))$ for all $(i, j) \in \mathbb{N}^2$. Observe that an IEA is an $\mathcal{S}$-valued random variable $\mathbf{X}$ whose law is invariant under the action of Π_∞. Let $\mathcal{P}(\mathcal{S})$ be the space of Borel probability measures on $\mathcal{S}$. Let $\mathcal{P}_{\mathrm{e}}(\mathcal{S}) \subseteq \mathcal{P}(\mathcal{S})$ be the set of *exchangeable probability measures* on $\mathcal{S}$, that is, $\mathcal{P}_{\mathrm{e}}(\mathcal{S}) \coloneqq \{\rho \in \mathcal{P}(\mathcal{S}) \mid \rho = \mathrm{Law}(\mathbf{X}), \mathbf{X} \text{ is an IEA}\}$. Throughout our discussion we will assume that $\mathcal{P}_{\mathrm{e}}(\mathcal{S})$ inherits the subspace topology from $\mathcal{P}(\mathcal{S})$, that is, it is equipped with the topology of weak convergence unless stated otherwise.

Recall the correspondence between IEAs and (random) MVGs defined in (5.2). Theorem 5.21 generalizes [DJ08, Theorem 5.3] and makes formal the idea that IEA are in one-to-one correspondence with random measure-valued graphons. Moreover, the convergence of IEAs is equivalent to convergence of the corresponding MVGs. We first extend the definition of decorated homomorphism densities to an IEA.

Definition 5.19 (Homomorphism density of IEAs w.r.t. decorated graphs)**.** Let $\mathbf{X}$ be an IEA. For every decorated graph F, define $T_{\mathrm{d}}(F, \mathbf{X}) \coloneqq \mathbb{E}\Big[\prod_{\{i,j\}\in E(F)} F_{i,j}(X_{i,j})\Big]$.

The following assertion is immediate from the definition and Theorem 5.21. For the importance of it, we record it as a Proposition. We skip the proof.

Proposition 5.20. *Let $\mathbf{Y}$ be an IEA and let $W_{\mathbf{Y}}$ be the corresponding (random) measure-valued graphon as described above. Then, for any decorated graph F we have $T_{\mathrm{d}}(F, \mathbf{Y}) = \mathbb{E}[T_{\mathrm{d}}(F, W_{\mathbf{Y}})]$. In particular, if $\mathbf{X}, (\mathbf{X}_n)_{n\in\mathbb{N}}$ are infinite exchangeable arrays then $(\mathbf{X}_n)_{n\in\mathbb{N}} \to \mathbf{X}$ weakly as $n \to \infty$ (with respect to the product topology) if and only if $\lim_{n\to\infty} T_{\mathrm{d}}(F, \mathbf{X}_n) = T_{\mathrm{d}}(F, \mathbf{X})$ for every decorated graph F.*

We now state and prove the main result of this section.

Theorem 5.21 (Homeomorphism Theorem [APST23])**.** *Let $\widehat{\mathfrak{W}}$ be the compact space of MVG equipped with its usual topology. Let $\mathcal{P}(\widehat{\mathfrak{W}})$ be the space of Borel probability measures on $\widehat{\mathfrak{W}}$ equipped with the weak topology. Then, $\mathcal{P}(\widehat{\mathfrak{W}})$ is homeomorphic to $\mathcal{P}_{\mathrm{e}}(\mathcal{S})$.*

The proof of Theorem 5.21 is provided in Appendix C.2.

5.2.1 Examples

Recall that by the Aldous-Hoover representation theorem, we know that every exchangeable array $\mathbf{X}$ can be written as $X_{i,j} = f(U, U_i, U_j, U_{i,j})$ for some Borel measurable function f. Throughout our discussion, we always assume that $U, \{U_i\}_{i\in\mathbb{N}}, \{U_{i,j} = U_{\{i,j\}}\}_{i,j\in\mathbb{N}}$ is a collection of i.i.d. Uni$[0,1]$ random variables on some probability space. We now give examples of IEAs and their Aldous-Hoover representation which in turn yields the corresponding (random) MVG. These examples emphasize that the graphons do not capture general IEAs while MVGs do. We first begin with a definition.

Definition 5.22 (Vertex, extrinsic, and edge dependence)**.** Let $\mathbf{X}$ be an IEA and let $f\colon [0,1]^4 \to \mathbb{R}$ be a corresponding Aldous-Hoover function. We say that f has *vertex* dependence if f depends on the second and third argument. Similarly, we will say that f

has *extrinsic* (respectively, *edge*) dependence if f depends on the first (respectively, fourth) argument. An IEA that doesn't have extrinsic dependence is called *pure* and corresponds to a deterministic MVG.

We must emphasize that Aldous-Hoover function for an IEA is not unique and is often not known explicitly. However, the above definition does not depend on the choice of Aldous-Hoover function (see [Kal89]).

Example 5.6 (Edge dependence - Mixture of two Dirac masses)**.** Let $\mathbf{X}$ be an infinite exchangeable array such that $X_{i,j}$s are all i.i.d. Bernoulli random variables, Ber(1/2). Let $f\colon [0,1]^4 \to \mathbb{R}$ be defined as $f(u,x,y,z) = \mathbb{1}\{z \leq 1/2\}$ for $(u,x,y,z) \in [0,1]^4$. We see that $\mathbf{X}$ is directed by f. On the other hand, let $\widetilde{\mathbf{X}}$ be an IEA such that $\widetilde{X}_{i,j}$s are all i.i.d. (up to matrix symmetry) and $\widetilde{X}_{i,j} \sim \frac{1}{2}\delta_{-1/2} + \frac{1}{2}\delta_{3/2}$. Then, $\widetilde{\mathbf{X}}$ is directed by an Aldous-Hoover function g where $g\colon [0,1]^4 \to \mathbb{R}$ is defined as $g(u,x,y,z) = \frac{1}{2} - \mathbb{1}\{z \leq 1/2\} + \mathbb{1}\{z > 1/2\}$. Note that the graphons and MVGs corresponding to $\mathbf{X}$ and $\widetilde{\mathbf{X}}$ (see equations (5.1) and (5.2)) are given by $w_{\mathbf{X}} \equiv \frac{1}{2} \equiv w_{\widetilde{\mathbf{X}}}$ while $W_{\mathbf{X}} \equiv \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ and $W_{\widetilde{\mathbf{X}}} \equiv \frac{1}{2}\delta_{-1/2} + \frac{1}{2}\delta_{3/2}$. Note that the graphons and the measure-valued graphons corresponding to $\mathbf{X}$ and $\widetilde{\mathbf{X}}$ are deterministic. This is reflected by the Aldous-Hoover representations f and g which are both independent of their first coordinates.

Example 5.7 (Extrinsic and edge dependence - correlated Gaussians)**.** Consider an infinite exchangeable array $\mathbf{X}$ such that $\{X_{i,j}\}_{(i,j)\in\mathbb{N}^{(2)}}$ are standard Gaussian random variables such that $\mathrm{Cov}(X_{i,j}, X_{l,m}) = 1/2$ whenever $\{i,j\} \neq \{l,m\}$. Let $\Phi\colon [0,1] \to \mathbb{R}$ be a function that pushes forward the Lebesgue measure on $[0,1]$ to the standard Gaussian measure on $\mathbb{R}$. And, let $f\colon [0,1]^4 \to \mathbb{R}$ be defined by $f(u,x,y,z) = \frac{1}{\sqrt{2}}\Phi(x) + \frac{1}{\sqrt{2}}\Phi(z)$ for all $(u,x,y,z) \in [0,1]^4$. It is easy to verify that $\mathbf{X}$ is directed by f. We, therefore, obtain for a.e. $(x,y) \in [0,1]^{(2)}$,

$$w_{\mathbf{X}}(x,y) \equiv \frac{\Phi(U)}{\sqrt{2}}, \quad W_{\mathbf{X}}(x,y) \equiv \mathrm{Law}\left(\mathcal{N}\left(\frac{\Phi(U)}{\sqrt{2}}, \frac{1}{2}\right)\right).$$

Note that $\Phi(U)$ is a standard normal random variable. Also note that $w_{\mathbf{X}}$ is a random kernel and $W_{\mathbf{X}}$ is a random MVG.

Following the same approach as above, one can more generally take $f(u,x,y,z) = \alpha\Phi(u) + \beta(\Phi(x)+\Phi(y)) + \gamma\Phi(z)$, say, where $\alpha^2+2\beta^2+\gamma^2 = 1$. And, define $X_{i,j} = f(U, U_i, U_j, U_{i,j})$ to obtain Gaussian exchangeable arrays with various correlation structures. This would yield

$$w_{\mathbf{X}}^{(u)}(x,y) = \alpha\Phi(u) + \beta(\Phi(x)+\Phi(y)), \qquad W_{\mathbf{X}}^{(u)}(x,y) = \mathrm{Law}\Big(\Big(\mathcal{N}\Big(w_{\mathbf{X}}^{(u)}(x,y), \gamma^2\Big)\Big)\Big),$$

for $u,x,y \in [0,1]$. Because of the extrinsic dependence, this IEA is not pure. Note that in this case the graphons $w_{\mathbf{X}}$ and the measure-valued graphon $W_{\mathbf{X}}$ are indeed random.

Example 5.8 (Vertex and edge dependence - Stochastic Block Model (SBM))**.** We now describe an exchangeable array that can be seen as limits of a certain sequence of SBM. Fix $p \in [0,1]$. For every $n \in \mathbb{N}$, color every vertex $i \in [n]$ blue with probability p and red with probability $(1-p)$ independently of each other. More formally, this is associating an independent $p\delta_1 + (1-p)\delta_{-1}$ distributed random variable $C(i)$ with $i \in [n]$, where 1 represents color 'blue' and -1 represents the color 'red'. Fix $p_{bb}, p_{rr}, p_{rb} \in [0,1]$. For each $\{i,j\} \subseteq [n]$, create an edge with probability p_{bb} if both i and j are colored blue, with probability p_{rr} if both are colored red and with probability p_{rb} otherwise. This defines an SBM with two communities 'blue' and 'red'. Let A_n denote the adjacency matrix of this SBM on vertex set $[n]$. It is easy to see that A_n is an exchangeable matrix, that is, $\mathrm{Law}(A_n) = \mathrm{Law}(A_n^\sigma)$. It is natural to ask, if A_n converges to some infinite exchangeable array as $n \to \infty$. This is indeed the case. Here we describe the infinite exchangeable array $\mathbf{X}$ that arises as the limit of $(A_n)_{n\in\mathbb{N}}$.

In order to define the Aldous-Hoover function for infinite exchangeable arrays, we first define some sets for notational simplicity. Fix $p \in [0,1]$. Define $B = [0,p]^2$, $R = [p,1]^2$ and $D = [0,p]\times[p,1] \cup [p,1]\times[0,p]$. Let $f\colon [0,1]^4 \to \{0,1\}$ be defined as

$$f(u,x,y,z) = \mathbb{1}_B\{(x,y)\}\mathbb{1}_{[0,p_{bb}]}\{z\} + \mathbb{1}_R\{(x,y)\}\mathbb{1}_{[0,p_{rr}]}\{z\} + \mathbb{1}_D\{(x,y)\}\mathbb{1}_{[0,p_{rb}]}\{z\},$$

for $u,x,y,z \in [0,1]$. The infinite exchangeable array $\mathbf{X}$ can be defined as $X_{i,j} := f(U, U_i, U_j, U_{i,j})$ for $i,j \in \mathbb{N}$. The corresponding graphon and measure-valued graphon are

$w^{(u)}$ and $W^{(u)}$ defined as

$$w^{(u)}(x,y) = p_{bb}\mathbb{1}_B\{(x,y)\} + p_{rr}\mathbb{1}_R\{(x,y)\} + p_{rb}\mathbb{1}_D\{(x,y)\},$$
$$W^{(u)}(x,y) = \mathrm{Ber}(p_{bb})\mathbb{1}_B\{(x,y)\} + \mathrm{Ber}(p_{rr})\mathbb{1}_R\{(x,y)\} + \mathrm{Ber}(p_{rb})\mathbb{1}_D\{(x,y)\},$$

for a.e. $(x,y) \in [0,1]^2$. This example can be generalized to distributions other than Bernoulli in a straightforward manner.

5.2.2 Finite exchangeable arrays to IEAs

The previous section establishes that the weak convergence of a sequence of IEAs is equivalent to the weak convergence of corresponding (random) MVGs (see Theorem 5.21). In practice, we are often interested in taking limits of finite exchangeable matrices. For instance, we would like to say that $G(n,1/2)$ converges to the IEA $\mathbf{G}$. One way to do this is to identify $G(n,1/2)$ with the corresponding (random) MVG, say $W_{\mathbf{G}_n}$ and show that $W_{\mathbf{G}_n} \to W_{\mathbf{G}}$ where $W_{\mathbf{G}}$ is the MVG corresponding to the IEA $\mathbf{G}$ (see Section 5.1.1). However, it is more natural to show the convergence of a sequence of finite exchangeable matrices to an IEA and deduce the convergence to an MVG from there. This is what we do in this section.

A (random) symmetric matrix $A \in \mathcal{M}_n$ is called (finite) *exchangeable* if $\mathrm{Law}(A) = \mathrm{Law}(A^\sigma)$ for every permutation σ of $[n]$. Given an $n \times n$ exchangeable matrix A, we can construct an IEA follows. Let $\{U_i\}_{i\in\mathbb{N}}$ be a family of i.i.d. Uni$[0,1]$ random variables independent of A. Define an IEA $\mathbf{X}$ such that $X_{i,j} \coloneqq A_{\lceil nU_i\rceil,\lceil nU_j\rceil}$. In plain words, each coordinate (up to matrix symmetry) of $\mathbf{X}$ is chosen independently and uniformly at random from the coordinates of A. With this correspondence, for every decorated graph F, define $T^{(0)}_{\text{finite}}(F,\,\cdot\,)$ over exchangeable matrices as $T^{(0)}_{\text{finite}}(F,A) \coloneqq T_F(\mathbf{X})$. On the other hand, analogous to Definition 5.19 we have following definition for homomorphism density into exchangeable matrices.

Definition 5.23 (Homomorphism density for exchangeable matrices)**.** Let A be an $n \times n$ exchangeable matrix. Let F be a decorated graph such that $|V(F)| < n$. Define $T^{(1)}_{\text{finite}}(F,A) \coloneqq \mathbb{E}\Big[\prod_{\{i,j\}\in E(F)} F_{i,j}(A_{i,j})\Big]$.

Remark 5.24. It is easy to see that $\left|T^{(1)}_{\text{finite}}(F,A) - T^{(0)}_{\text{finite}}(F,A)\right| \leq C(F)n^{-1}$ where $C(F)$ is a constant depending only on F. Therefore, both $T^{(0)}_{\text{finite}}$ and $T^{(1)}_{\text{finite}}$ determine the same limit as $n \to \infty$. Also note that using the embedding described in Section 5.1.1, we can define $T_{\text{d}}(F,A)$ as in equation (5.4). Notice that $T^{(0)}_{\text{finite}}(F,A) = \mathbb{E}[T_{\text{d}}(F,A)]$.

This motivates the following definition for the convergence of finite exchangeable matrices to an IEA.

Definition 5.25 (Convergence of exchangeable matrices)**.** Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of $n \times n$ symmetric exchangeable matrices. We say that $(A_n)_{n\in\mathbb{N}} \to \mathbf{X}$ as $n \to \infty$ if for every decorated graph F we have $T^{(1)}_{\text{finite}}(F,A_n) = \mathbb{E}[T_{\text{d}}(F,A_n)] \to T_{\text{d}}(F,\mathbf{X})$ as $n \to \infty$.

We end this section with some examples of finite exchangeable matrices converging to an IEA.

Example 5.9. Let $V\colon \mathbb{R}^{(2)} \to [-1,1]$ be a C^2 function such that $V(x,y) = V(y,x)$ and $\|\nabla^2 V\|_\infty \leq 1$, where $\nabla^2 V$ is the Hessian of V. Define $\mathcal{V}\colon \mathcal{P}(\mathbb{R}) \to \mathbb{R}$ as $\mathcal{V}(\mu) \coloneqq \frac{1}{2}\iint_{\mathbb{R}^2} V(x,y)\mu(\mathrm{d}x)\mu(\mathrm{d}y)$. Define $\mathcal{V}_n\colon \mathbb{R}^n \to \mathbb{R}$ as $\mathcal{V}_n(x_1,\ldots,x_n) = \mathcal{V}(\mu_n)$ where $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ for every $\{x_i\}_{i\in[n]} \subset \mathbb{R}$. In particular, $\mathcal{V}_n(x_1,\ldots,x_n) \coloneqq \frac{1}{2n^2}\sum_{i,j=1}^n V(x_i,x_j)$. Let $H_n(\mathbf{x}) \in \mathcal{M}_n$ be the Hessian matrix of $\mathcal{V}_n$ at $\mathbf{x} \in \mathbb{R}^n$. Then, $n^2 H_n(\mathbf{x})_{(i,j)} = \partial_{1,2}V(x_i,x_j)$ if $i \neq j$, and $n^2 H_n(\mathbf{x})_{(i,i)} = \partial_{1,1}V(x_i,x_i)$ for $(i,j) \in [n]^{(2)}$. Now, let $\{X_i\}_{i\in\mathbb{N}}$ be i.i.d. random variables distributed according to some probability measure $\mu \in \mathcal{P}([-1,1])$ and let $\mathbf{X}_n = (X_1,\ldots,X_n)$. Then, $\mathcal{H}^{(n)} = n^2 H_n(\mathbf{X}_n)$ is an exchangeable matrix and $\mathcal{H}^{(n)} \to \mathcal{H}^{(\infty)}$, where $\mathcal{H}^{(\infty)}$ is an exchangeable array defined as

$$\mathcal{H}^{(\infty)}_{(i,j)} = \begin{cases} \partial_{1,2}V(X_i,X_j), & \text{if } i \neq j, \\ \partial_{1,1}V(X_i,X_i), & \text{if } i = j, \end{cases} \qquad (i,j) \in \mathbb{N}^{(2)}.$$

For concreteness, assume that $V(x,y) = c(x-y)$ for $(x,y) \in \mathbb{R}^{(2)}$ for some even C^2 function $c\colon \mathbb{R} \to [-1,1]$. In that case, notice that $\mathcal{H}^{(\infty)}_{(i,j)} = -c''(X_i - X_j)$ for $(i,j) \in \mathbb{N}^{(2)}$. Also assume, for simplicity, that $\{X_i\}_{i\in\mathbb{N}}$ are i.i.d. Uni$[0,1]$. Then, c'' is the Aldous-Hoover

representation function and the MVG corresponding to $\mathcal{H}^{(\infty)}$ is nothing but $W^\infty \in \widehat{\mathfrak{W}}$ defined as $W^\infty(x,y) \coloneqq \delta_{-c''(x-y)}$ for a.e. $(x,y) \in \mathbb{R}^{(2)}$.

Example 5.10. One can consider higher order polynomials of measures. That is, for any $k \in \mathbb{N} \setminus \{1\}$, define $\mathcal{V}\colon \mathcal{P}(\mathbb{R}) \to \mathbb{R}$ as $\mathcal{V}(\mu) \coloneqq \int_{\mathbb{R}^k} V(x_1,\dots,x_k)\mu(\mathrm{d}x_1)\dots\mu(\mathrm{d}x_k)$. Define $V_n\colon \mathbb{R}^n \to \mathbb{R}$ as $x \mapsto V_n(x) = \mathcal{V}(\mu_n)$ where $\mu_n \coloneqq \frac{1}{n}\sum_{j=1}^n \delta_{x_i}$. This amounts to evaluating the expectation of V when its arguments are sampled uniformly with replacement from the entries of $x \in \mathbb{R}^n$. Let $H_n(x)$ be the Hessian matrix of V_n at $x \in \mathbb{R}^n$. Let us define $G\colon \mathbb{R}^2 \to \mathbb{R}$ as $G(a,b) \coloneqq$

$$\sum_{i,j\in[k],i\neq j} \int_{\mathbb{R}^{k-2}} \partial_{i,j} V(x_1,\dots,x_{i-1},a,x_{i+1},\dots,x_{j-1},b,x_{j+1},\dots,x_k) \prod_{m\in[k]\setminus\{i,j\}} \mu(\mathrm{d}x_m).$$

Let $(X_i)_{i\in\mathbb{N}}$ be i.i.d. random variables with distribution $\mu \in \mathcal{P}(\mathbb{R})$. Then, $n^k H_n(X_1,\dots,X_n) \to \mathbf{H}$, as $n \to \infty$, where $\mathbf{H}_{i,j} = G(X_i,X_j)$ for $(i,j) \in \mathbb{N}^{(2)}$.

Example 5.11. Consider a Markov chain $(X_n)_{n\in\mathbb{N}}$ on $[0,1]$ with a unique stationary distribution $\pi \in \mathcal{P}([0,1])$. Let $W\colon [0,1]^2 \to [0,1]$ be a kernel such that W is continuous $\pi\times\pi$ a.e. For each $n \in \mathbb{N}$, let $(Y_1,\dots,Y_n)$ be a uniform permutation of $(X_1,\dots,X_n)$ and let $\mathcal{H}^{(n)}$ be an exchangeable matrix defined by $\mathcal{H}^{(n)}_{i,j} = W(Y_i,Y_j)$, $i,j \in [n]$. Let $\{V_i\}_{i\in\mathbb{N}}$ be a collection of i.i.d. random variables with distribution π and let $\mathcal{H}^{(\infty)}$ be an exchangeable array such that $\mathcal{H}^{(\infty)}_{i,j} = W(V_i,V_j)$. Then, $\mathcal{H}^{(n)} \to \mathcal{H}^{(\infty)}$ as $n \to \infty$.

5.3 Conclusion

In this chapter, we introduced the richer analytical structure of measure-valued graphons, which build upon the geometry and topology of graphons to establish a direct correspondence with two-dimensional real-valued infinite exchangeable arrays. Leveraging this correspondence and the metrics developed herein, Chapter 6.2 will explore the SDEs derived in Chapter 3, deriving their scaling limits as absolutely continuous curves on $\widehat{\mathfrak{W}}$, characterized by MVG McKean-Vlasov equations, and examining the non-asymptotic convergence in specific cases.

Chapter 6

SCALING LIMITS OF SDES WITH INCREASING DIMENSIONS

In Chapter 3, we derived the continuous-time limits for the class of discrete-time algorithms discussed in Section 1.1 for finite dimensions. That is, under a suitable choice of time scaling, for every finite-dimensional matrix-valued process obtained from an iterative algorithm, we can weakly describe the evolution of the matrix via an SDE of the form:

$$\begin{aligned} \mathrm{d}X_{n,e}(t) = b_{n,e}(X_{n,e}(t), X_n(t))\,\mathrm{d}t + \Sigma_{n,e}(X_{n,e}(t), X_n(t)) \circ \mathrm{d}B_{n,e}(t) \\ + \mathrm{d}L^-_{n,e}(t) - \mathrm{d}L^+_{n,e}(t), \end{aligned} \tag{6.1}$$

For $X_n(0) = X_{n,0} \in \mathcal{M}_n$, where the tuple (X_n, L^+_n, L^-_n) solves the Skorokhod problem with respect to the set $\mathcal{M}_n$ (see Section 2.4), we remark that the cubic domain $\mathcal{M}_n$ can be replaced by any other cubic domain (which may not necessarily satisfy matrix symmetry), including $\mathbb{R}^{[n]^2}$. The SDE (6.1) can then be appropriately modified without altering its essence.

An interesting and important observation we make is that the system described by (6.1), due to the equivariance of the functions $b_n(z,\cdot)$ and $\Sigma_n(z,\cdot)$ with respect to the group S_n for any z within its domain, exhibits a mean-field interaction. That is, for any $e \in [n]^2$, the evolution $t \mapsto X_{n,e}(t)$ of $X_{n,e}$ at any instant $t \in \mathbb{R}_+$, as determined by the drift and diffusion coefficients, depends on all the coordinate evolutions $X_n(t)$ at time $t \in \mathbb{R}_+$ in a symmetric manner. This symmetric dependency is evident when permuting the rows and columns of the matrix by any permutation $\pi \in S_n$. For simplicity, let us consider $\mathbb{R}^{[n]^{(2)}}$ as the domain and observe that

$$\begin{aligned} \mathrm{d}Y_{n,e}(t) = \mathrm{d}X_{n,\pi(e)}(t) &= b_{n,\pi(e)}(X_{n,\pi(e)}(t), X_n(t))\,\mathrm{d}t + \Sigma_{n,\pi(e)}(X_{n,\pi(e)}(t), X_n(t)) \circ \mathrm{d}B_{n,\pi(e)}(t) \\ &= b_{n,e}(X_{n,\pi(e)}(t), X^\pi_n(t))\,\mathrm{d}t + \Sigma_{n,e}(X_{n,\pi(e)}(t), X^\pi_n(t)) \circ \mathrm{d}B_{n,\pi(e)}(t) \end{aligned}$$

$$= b_{n,e}(Y_{n,e}(t), Y_n(t))\,\mathrm{d}t + \Sigma_{n,e}(Y_{n,e}(t), Y_n(t)) \circ \mathrm{d}B_{n,\pi(e)}(t),$$

where $\pi(e) = (\pi(e_1), \pi(e_2))$, and $Y_n = X_n^\pi = \left(X_{n,\pi(e)}\right)_{e\in[n]^{(2)}}$ for any X_n. Since B_n and B_n^π have the same law, the SDEs above are equivalent for any $\pi \in S_n$. This brief analysis also tells us that the process X_n is exchangeable (see Section 2.6).

The goal of this chapter is to understand how the sequence of processes $(X_n)_{n\in\mathbb{N}}$, defined on bounded cubic domains like $\mathcal{M}_n$ and $\mathcal{M}_{n,+}$, converges to a suitable well-defined limit. To make sense of this convergence, we will utilize the compact topologies discussed in Chapters 4 and 5. Analytically, this limit can be interpreted as a curve on graphons and measure-valued graphons (MVGs) with interesting properties. The scaling limit shall be described by McKean-Vlasov type SDEs on graphons and MVGs. Drawing an analogy with classical interacting particle systems (see Sections 1.2 and 1.3 for an introductory discussion), the limiting curve can be regarded as the 'scaling limit' of the class of iterative algorithms discussed, which substantiates the central hypothesis of this thesis. This completes the cycle of ideas, allowing us to extend the Wasserstein calculus to higher-order exchangeable structures.

In Section 6.1, we will use the graphon topology (induced by the cut metric) to derive a scaling limit of $(X_n)_{n\in\mathbb{N}}$. In Chapter 6.2, we will use the MVG topology (induced by the Wasserstein cut metric) to derive the scaling limit. Both limits will demonstrate the propagation of chaos phenomenon, which essentially states that as the dimensionality increases, any finitely chosen coordinates of the system tend to evolve independently. In Chapter 6.3, we will see that products of iterated matrices, as discussed in Section 2.2.3, converge appropriately to IEAs. We will work out certain examples of algorithms to deduce interesting properties from the analysis. This analysis will later allow us to show a beautiful implication in the context of deep learning that we will discuss at length in Chapter 7.

6.1 Scaling limits as curves on Graphons via McKean-Vlasov equations

The study of particle systems under mean-field interaction is a classical topic in probability theory [Gär88]. It involves multidimensional diffusions that interact through their empirical distributions of the type

$$\mathrm{d}X_i(t) = b\left(X_i(t), \hat{\mu}^{(n)}(t)\right)\mathrm{d}t + \mathrm{d}B_i(t), \quad i \in [n], \quad t \in \mathbb{R}_+, \tag{6.2}$$

where $n \in \mathbb{N}$, $X_i(t) \in \mathbb{R}^d$ for all $i \in [N]$ and for some $d \in \mathbb{N}$, and $\hat{\mu}^{(n)}(t) \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{X_i(t)}$, is the empirical distribution of the vector $(X_i(t))_{i\in[n]}$ at time $t \in \mathbb{R}_+$, and $(B_i)_{i\in[N]}$ is a vector of i.i.d. standard d-dimensional Brownian motions. Any drift that is symmetric in the coordinates ("mean-field interactions") can be represented as (6.2) for some suitable function b. Often, the SDE (6.2) includes a reflection term to constrain the coordinate process to a subset of the Euclidean space [Szn84]. The study of such systems originated from the probabilistic study of the Boltzmann and Vlasov equations due to Kac [Kac56], McKean [McK75], Dobrushin [Dob79], Tanaka [Tan79] and many others. For modern surveys, see Sznitman [Szn91], Villani [Vil12], Chaintron and Diez [CD22] and Jabin [Jab14].

Under suitable assumptions, as the number of particles go to infinity, it is known that the process of empirical distributions of the particle system converges to the solutions of families of well-known PDEs. The convergence is often obtained via *propagation of chaos* where, in the large particle limit, a finite collection of randomly chosen particles evolve independently and identically distributed according to the McKean-Vlasov SDE [Gär88]:

$$\mathrm{d}X(t) = b\left(X(t), \mu(t)\right)\mathrm{d}t + \mathrm{d}B(t),$$

where $\mu(t)$ is the law of $X(t)$ for $t \in \mathbb{R}_+$.

In this chapter we study an analogous evolution of symmetric matrices where the coordinates interact via a suitably symmetric function. As an example, consider the function R_n defined on $\mathcal{M}_{n,+}$, given by

$$R_n(A) \coloneqq \frac{1}{2}\left(n^{-2}\sum_{i,j=1}^n A(i,j) - e\right)^2 + \frac{1}{2}\left(n^{-3}\operatorname{tr}\left[A^3\right] - \tau\right)^2 + \mathcal{E}_n(A), \tag{6.3}$$

where $e, \tau \in [0,1]$ are fixed and $\mathcal{E}_n(A) = n^{-2}\sum_{i,j=1}^n h(A(i,j))$ where $h\colon [0,1] \to \mathbb{R}$ is the convex entropy function defined as $h(p) \coloneqq p\log p + (1-p)\log(1-p)$, if $p \in (0,1)$, and $h(0) = h(1) = 0$. The importance of this particular example will soon become apparent.

The function R_n satisfies a *permutation invariance* property, that is, its value does not change if we permute the rows and columns of the matrix A by the same permutation over $[n] \coloneqq \{1, 2, \ldots, n\}$. Consider the following diffusion on symmetric $n \times n$ matrices

$$\mathrm{d}X_n(t) = -n^2 \nabla R_n(X_n(t))\,\mathrm{d}t + \beta\,\mathrm{d}B_n(t) + \mathrm{d}L_n(t), \qquad t \in \mathbb{R}_+, \tag{6.4}$$

where B_n is a system of $n \times n$ symmetric matrix-valued process of coordinatewise independent Brownian motions and L_n is the coordinatewise bounded variation local time process that constrains each coordinate process to stay in the interval $[0,1]$ (see Section 2.4 for details). One may ask what is an appropriate notion of limit of such a process as $n \to \infty$? Do weak solutions of SDE (6.4) exhibit propagation of chaos?

Note that the function R_n in equation (6.4) is not covered by the classical McKean-Vlasov theory since $R_n(A)$ is not symmetric in the n^2 (up to symmetry) many entries of a matrix A. Therefore, R_n cannot be expressed as a function of the empirical distribution of the entries of the argument matrix. The same is true for any arbitrary differentiable function over $n \times n$ symmetric matrices that is invariant under permuting the rows and the columns using the same permutation. Spectral functions, for example, satisfy such an invariance, as do functions on edge-weighted graphs (represented by their adjacency matrices) that are invariant under vertex relabeling. This particular class of symmetry is captured, not by empirical measures but, by *graphons*. In other words, such functions can be thought of as functions on the space of graphons instead of measures.

Analogous to the classical McKean-Vlasov theory, we show in this section that, under suitable assumptions, (6.4) exhibits a propagation of chaos. Furthermore, in $n \to \infty$ limit, the coordinates of X_n become conditionally independent and the evolution of a randomly chosen coordinate can be described by a novel graphon valued McKean-Vlasov equation.

Recently various authors [DGL16, BBW19, Cop22, DM22] have investigated McKean-

Vlasov limits for interacting particles systems on dense graphs. In these work, the McKean-Vlasov system describes the evolution of random particle from an infinite ensemble where the underlying interaction is determined by a graph or graphon. Extensions to the sparse regime can be found in [LRW19, OR19, BCN20, BCW20, ORS20]. We note that our McKean-Vlasov limit describes the evolution of graphon itself, and not the distribution of any particle system. We borrow the name McKean-Vlasov to stress that each edge-weight evolves by an *ensemble* effect of all the other edge weights but that ensemble is a graphon and not the empirical distribution of any particle system as done in the papers cited above.

Consider the function R_n defined in (6.3). Notice that if we regard $A \in \mathcal{M}_{n,+}$ as the adjacency matrix of a weighted graph, then $n^{-2}\sum_{i,j=1}^{n} A_{i,j}$ can be thought of as the edge-density of the graph while $n^{-3}\operatorname{tr}[A^3]$ can be regarded as the density of triangles in the graph. Consider the problem of minimizing $\mathcal{E}_n(A)$ over all $A \in \mathcal{M}_{n,+}$ subject to the condition that the edge-density and the triangle-density are e and τ, respectively. The non-convexity of this problem makes it very hard and indeed the minimizers in general are not unique [NRS23b, KRRS17]. For certain feasible regime of (e, τ), this minimization problem has been studied in [NRS23b]. In some regimes where the minimizer is known to be unique, the minimizer is characterized. Similar results were obtained in [KRRS17] when $\tau = e^3$. In general, however, determining the structure of minimizers is extremely hard and even reasonable guesses are not available in most cases. This problem has rich connections with extremal combinatorics [Raz08, PR17] and exponential random graph models [BCM21, CV11]. While minimizing $\mathcal{E}_n$ with such 'hard constraints' is difficult, notice that minimizing $R_n(A) := \left(n^{-2}\sum_{i,j=1}^{n} A(i,j) - e\right)^2/2 + \left(n^{-3}\operatorname{tr}[A^3] - \tau\right)^2/2 + \mathcal{E}_n(A)$ can be considered as a relaxation of this problem. Our method provides a numerical scheme like algorithms discussed in Section 2.2.1 to obtain minimizers of such problems. Following Section 3.1, SDE (6.4) arises as the continuous-time limit of the projected stochastic gradient descent algorithm which is used in practice to optimize R_n. As mentioned above, we establish that the curves described by (6.4) converges to a deterministic curve on the space of graphons. Under appropriate assumptions on R_n (see Section 6.1.3) and in zero-noise limit, the (deterministic) limiting

curve on the space of graphons is a gradient flow and hence converges to the minimizer exponentially fast.

6.1.1 Deriving the graphon McKean-Vlasov scaling limit

In Section 6.1, we assume for simplicity that the diffusion coefficient function at coordinate e is not an explicit function of the coordinate $X_{n,e}$ of the process X_n. The result can be extended to take this dependence into account. The general case where this dependence is explicitly considered has been analyzed in Section 6.2.

Let $\mathcal{E}$ be a standard Borel space. The sets $[n]^{(2)}$ and $\mathbb{N}^{(2)}$ will refer to the set of natural number pairs (i,j) in $\mathbb{N}^2$ and $[n]^2$ respectively, such that $i<j$. Recall that an $\mathcal{E}$-valued exchangeable (symmetric) array refers to a doubly indexed collection of random elements $\left(\zeta_{i,j} \coloneqq \zeta_{\{i,j\}} \in \mathcal{E}\right)_{(i,j)\in\mathbb{N}^{(2)}} =: \zeta$ that remain invariant in law under finite permutations of natural numbers $\mathbb{N}$. Two special cases of $\mathcal{E}$ that are important to us are $\mathcal{E}=[-1,1]$ and $\mathcal{E}=C[0,\infty)$ with the usual Borel topology. The Aldous-Hoover representation theorem [Ald85, Hoo79, Hoo82] says that given any exchangeable array as above, there exists a measurable function $f\colon [0,1]\times[0,1]^{(2)}\times[0,1]\to\mathcal{E}$ such that $\zeta_{i,j}=f\left(U,U_i,U_j,U_{i,j}\right)=f\left(U,U_j,U_i,U_{i,j}\right)$ for $(i,j)\in\mathbb{N}^{(2)}$, where U, $(U_i)_{i\in\mathbb{N}}$, $\left(U_{i,j}=U_{\{i,j\}}\right)_{(i,j)\in\mathbb{N}^{(2)}}$ are i.i.d. Uni$[0,1]$ random variables. The function f is typically not unique. Following [Aus08], we say that ζ is directed by f.

The relationship between exchangeable arrays and graphons follows from the Aldous-Hoover representation [DJ08]. Assume that $\zeta_{i,j}$s are real valued and take values in the closed interval $[-1,1]$. An infinite exchangeable array gives rise to a *random* graphon reminiscent of the de Finetti representation theorem for exchangeable sequences of random variables. Although we believe that the following result is well-known, we could not find a statement to this effect in the literature. However, it inspires our later constructions.

Lemma 6.1. *Let $\zeta\in[-1,1]^{\mathbb{N}^{(2)}}$ be an infinite exchangeable array directed by f. Consider the family of symmetric kernels $(w_u)_{u\in[0,1]}$ defined by*

$$w_u(x,y)\coloneqq\mathbb{E}[f(u,x,y,V)],\qquad u\in[0,1],\quad (x,y)\in[0,1]^{(2)}, \tag{6.5}$$

where the above expectation is with respect to a Uni$[0,1]$ *random variable* V*. Then, for* $u \in [0,1]$*, given* $\{U = u\}$*,*

$$\lim_{n\to\infty} \delta_{\square}\left(K\Big((\zeta_{i,j} = f(u, U_i, U_j, U_{i,j}))_{(i,j)\in[n]^{(2)}}\Big), [w_u]\right) = 0, \qquad a.s. \tag{6.6}$$

The proof of Lemma 6.1 is provided in Appendix D.1.

Remark 6.2. As a corollary of the previous result, although the function f is not unique in the Aldous-Hoover representation, the law of the random graphon $[w_U]$ is indeed unique.

Consider $(C[0,\infty))^{\mathbb{N}^{(2)}}$ with the natural filtration generated by the coordinate process. Enlarge the filtration by expanding the probability space to accommodate the countably many i.i.d. Uni$[0,1]$ random variables $(U_i)_{i\in\mathbb{N}}$ and including the sigma algebra generated by them in the sigma algebra at time zero. Endow this filtered probability space with a probability measure P^∞ that denote the joint law of $(U_i)_{i\in\mathbb{N}}$ and that of an independent array of countably many independent Brownian motions (BMs) $\big\{B_{i,j} = B_{\{i,j\}}\big\}_{(i,j)\in\mathbb{N}^{(2)}}$. Finally we turn the natural filtration to one that is right-continuous and complete, thereby satisfying the so-called usual conditions and denote it by $\mathcal{F} = (\mathcal{F}_t)_{t\in\mathbb{R}_+}$. All our processes will be adapted to this filtration associated with this set-up. Note that all uniform random variables $(U_i)_{i\in\mathbb{N}}$ are measurable with respect to $\mathcal{F}_0$.

We will consider the functions $\{b_n\}_{n\in\mathbb{N}}$ to be restrictions of some function $b\colon [-1,1]\times\mathcal{W} \to L^\infty\big([0,1]^{(2)}\big)$, i.e., $b_n(z,\,\cdot\,) = M_n \circ b(z,\,\cdot\,) \circ K$ on $\mathcal{M}_n$ for all $z \in [-1,1]$ and all $n \in \mathbb{N}$. With this generalization, we override Assumption 3.1(2) with the following assumption on the drift function b.

Assumption 6.1. For a.e. $(x,y) \in [0,1]^{(2)}$, $w_1, w_2 \in \mathcal{W}$ and $z_1, z_2 \in [-1,1]$, the drift function $b\colon [-1,1]\times\mathcal{W} \to L^\infty([0,1]^{(2)})$ satisfies

1. There exists $L \in \mathbb{R}_+$ such that

$$\sup_{w\in\mathcal{W}} |b(z_1,w)(x,y) - b(z_2,w)(x,y)| \le L|z_1 - z_2|.$$

2. There exists $\kappa \in \mathbb{R}_+$ such that

$$\sup_{z\in[-1,1]} \|b(z,w_1) - b(z,w_2)\|_2 \leq \kappa \|w_1 - w_2\|_2.$$

Observe that Assumption 6.1 implies Assumption 3.1(2) for $\kappa_2^2 = 2(L^2 + \kappa^2)$ and that $\|b(z,w)\|_\infty \leq C$ uniformly over all $z \in [-1,1]$ and $w \in \mathcal{W}$.

We construct a diffusion as follows. Let $b\colon [-1,1] \times \mathcal{W} \to L^\infty\big([0,1]^{(2)}\big)$ be satisfy Assumption 6.1. Given $w_0 \in \mathcal{W}$, let $X \coloneqq \big(X_{i,j} \coloneqq X_{\{i,j\}}\big)_{(i,j)\in\mathbb{N}^{(2)}}$, be the solution of the following system of SDE taking values in $[-1,1]^{\mathbb{N}^{(2)}}$ with the initial condition $(X_{i,j}(0) = w_0(U_i, U_j))_{(i,j)\in\mathbb{N}^{(2)}}$, and satisfying

$$\begin{aligned} \mathrm{d}X_e(t) &= b(X_e(t), \gamma(t))(U_e)\,\mathrm{d}t + \Sigma\left(\gamma(t)\right)(U_e)\,\mathrm{d}B_e(t) \\ &\qquad + \mathrm{d}L_e^-(t) - \mathrm{d}L_e^+(t), \qquad \text{(Graphon-MKV)} \\ \gamma(t)(x,y) &\coloneqq \mathbb{E}[X_{1,2}(t) \mid U_1 = x, U_2 = y], \end{aligned}$$

for $e \in \mathbb{N}^{(2)}$, $(x,y) \in [0,1]^{(2)}$ and $t \in \mathbb{R}_+$. Here $U_e \coloneqq (U_{e(1)}, U_{e(2)})$. The processes L_e^- and L_e^+ are such that (X_e, L_e^+, L_e^-) solves the Skorokhod problem with respect to $[-1,1]$ (see Section 2.4). The kernel valued process $\gamma\colon \mathbb{R}_+ \to \mathcal{W}$ is adapted to the sigma algebra generated by the uniform random variables $(U_i)_{i\in\mathbb{N}}$, and the independent BMs $(B_e)_{e\in\mathbb{N}^{(2)}}$. Note that if the solution X of the system of SDEs (Graphon-MKV) exists, then conditioned over the sigma algebra $\mathcal{F}_0$, the coordinate processes of X are all independent but not necessarily identically distributed.

It is not obvious if an infinite-dimensional stochastic process satisfying (Graphon-MKV) exists, although it is obvious that such a process, if it exists, will be an infinite exchangeable array taking values in $\mathcal{E} = C[0,\infty)$. Under Assumption 6.1 we show that the process (X, γ) is indeed well-defined. As will be made clear in Proposition 6.3, the limiting object γ is the counterpart to the measure-valued solution of the McKean-Vlasov equation, while every $X_{i,j}$ for $(i,j) \in \mathbb{N}^{(2)}$ is the counterpart to the non-linear evolution of a randomly chosen particle evolving in the McKean-Vlasov interacting system. It should be noted that the

particles in this McKean-Vlasov interaction correspond to the edges of the graphs not the vertices. The McKean-Vlasov equation here describes how the graphon itself evolves in time and it is different from the McKean-Vlasov system described in the introduction where the McKean-Vlasov equation describes the evolution of particles which may possibly depend on some underlying graphon.

Next, we state Proposition 6.3 that shows the infinite dimensional stochastic process satisfying equations (Graphon-MKV) indeed exists.

Theorem 6.3 (Existence of the McKean-Vlasov SDE [HOP+22])**.** *Assume that the drift functions* $b\colon [-1,1]\times\mathcal{W} \to L^\infty\big([0,1]^{(2)}\big)$ *satisfies Assumption 6.1, and the diffusion coefficient function* $\Sigma\colon \mathcal{W} \to L^\infty\big([0,1]^{(2)}\big)$ *is bounded and* κ_2*-Lipschitz in* $\|\cdot\|_2$ *(Assumption 3.3). Then, given any kernel* $w_0 \in \mathcal{W}$*, there exists a pathwise unique strong solution to the coupled system* (Graphon-MKV) *in the following sense. In any probability space supporting countably many i.i.d.* Uni$[0,1]$ *random variables* $(U_i)_{i\in\mathbb{N}}$ *and an independent infinite (symmetric) array of i.i.d. standard Brownian motions* $(B_{i,j})_{(i,j)\in\mathbb{N}^{(2)}}$*, one can construct an infinite exchangeable array of reflected diffusions* $(X_{i,j})_{(i,j)\in\mathbb{N}^{(2)}}$ *that satisfy* (Graphon-MKV) *and every* $X_{i,j}$ *is pathwise unique.*

Moreover, for every $t \in \mathbb{R}_+$*,* $[\gamma(t)]$ *can be recovered as the* $\delta_\square$ *limit of the sequence of graphons* $\big(\big[K\big((X_{i,j}(t))_{(i,j)\in[n]^2}\big)\big]\big)_{n\in\mathbb{N}}$ *locally uniformly in time. That is, for any* $t \in \mathbb{R}_+$*,*

$$\lim_{n\to\infty} \sup_{s\in[0,t]} \delta_\square\left(\left[K\Big((X_e(s))_{e\in[n]^{(2)}}\Big)\right], [\gamma(s)]\right) = 0, \qquad a.s. \tag{6.7}$$

The proof of Theorem 6.3 is provided in Appendix D.1.1.

Next, in Proposition 6.4 we provide the expression for the velocity of the curve γ in the special case for the SDE obtained for the projected noisy stochastic gradient descent algorithm (see Definition 2.2).

Proposition 6.4. *Suppose that* $\Sigma \equiv \beta > 0$ *and* $b(z,w) = -\phi(w)$ *where* $\phi = DR$ *for the risk function* R *(see Chapter 4). Then, the limiting curve* γ *in Proposition 6.3 has a velocity*

$$\dot\gamma(t) = -\phi(\gamma(t)) - \Big[p^{(+1)}_{\beta^2 t}(w_0, \phi\circ\gamma, \beta) - p^{(-1)}_{\beta^2 t}(w_0, \phi\circ\gamma, \beta)\Big], \tag{6.8}$$

where $p_s^{(\pm 1)}(w_0, \phi \circ \gamma, \beta)(x, y)$ *is the density of the real-valued reflected Brownian motion* Z *at* ± 1*, at time* $s \in \mathbb{R}_+$*, starting at* $Z(0) = w_0(x, y)$*, satisfying*

$$\mathrm{d}Z(s) = -\frac{1}{\beta^2}\phi(\gamma(s/\beta^2))(x, y)\,\mathrm{d}s + \mathrm{d}B(s) + \mathrm{d}L^-(s) - \mathrm{d}L^+(s), \qquad s \in \mathbb{R}_+,$$

where (Z, L^+, L^-) *solves the Skorokhod problem with respect to the set* $[-1, 1]$ *(see Section 2.4).*

The proof of Proposition 6.4 is provided in Appendix D.1.

Remark 6.5. Note that the (pointwise) velocity of the curve γ at time $t \in \mathbb{R}_+$ is not $-(\phi \circ \gamma)(t)$ when $\beta > 0$. That is, γ is not a gradient flow of the function R when $\beta > 0$, and the effect of the boundary $\{-1, 1\}$, as seen in (6.8), is qualitatively different from that when $\beta = 0$.

To show that the finite dimensional processes converge as $n \to \infty$, we will need to put further assumptions on the drift and diffusion functions.

Assumption 6.2. There exists a constant $\kappa_\square \in \mathbb{R}_+$ such that for all $w_1, w_2 \in \mathcal{W}$, the drift function $b\colon [-1, 1] \times \mathcal{W} \to L^\infty\big([0, 1]^{(2)}\big)$ and the diffusion coefficient function $\Sigma\colon \mathcal{W} \to L^\infty\big([0, 1]^{(2)}\big)$ satisfy

$$\sup_{(x,y)\in[0,1]^2} \sup_{z\in[-1,1]} |b(z, w_1)(x, y) - b(z, w_2)(x, y)| \le \kappa_\square \|w_1 - w_2\|_\square, \qquad \text{and}$$

$$\sup_{(x,y)\in[0,1]^2} |\Sigma(w_1)(x, y) - \Sigma(w_2)(x, y)| \le \kappa_\square \|w_1 - w_2\|_\square.$$

Theorem 6.6 (Convergence [HOP+22])**.** *Assume that the drift functions* $b\colon [-1, 1] \times \mathcal{W} \to L^\infty\big([0, 1]^{(2)}\big)$ *satisfies Assumption 6.1 and Assumption 6.2, and the diffusion coefficient function* $\Sigma\colon \mathcal{W} \to L^\infty\big([0, 1]^{(2)}\big)$ *is bounded and* κ_2*-Lipschitz in* $\|\cdot\|_2$ *(Assumption 3.3). Then, for any sequence of initial kernels* $\big(w_0^{(n)} \in \mathcal{W}_n\big)_{n\in\mathbb{N}}$ *that converges to* $w_0 \in \mathcal{W}$ *in the* $L^2\big([0, 1]^{(2)}\big)$ *norm* $\|\cdot\|_2$*, i.e., whenever*

$$\lim_{n\to\infty} \left\| w_0^{(n)} - w_0 \right\|_2 = 0, \tag{6.9}$$

the process of random kernels $(K(X_n(t)))_{t\in\mathbb{R}_+}$ *obtained from solutions of the SDEs* (6.1)*, converges locally uniformly in the cut norm as* $n \to \infty$*, in probability, to the limiting process* $\gamma\colon \mathbb{R}_+ \to \mathcal{W}$*, with* $\gamma(0) = w_0$*, established in Theorem 6.3.*

The proof of the Theorem 6.6 is provided in Appendix D.1. The proof also includes a non-asymptotic rate of convergence as Remark D.4. We will strengthen the rate later in Section 6.2.

Remark 6.7. The assumption $\left\|w_0^{(n)} - w_0\right\|_2 \to 0$ can not be weakened to $\left\|w_0^{(n)} - w_0\right\|_\square \to 0$ as $n \to \infty$. To see this, take $\nabla R_n \equiv 0$ and $\Sigma \equiv 1$ and let $w_0 \equiv 0$. It is clear that $\gamma(t) \equiv 0$ for all $t \geq 0$.

On the other hand, let ξ be a random variable taking values $-1/2$ and $+1$ with probability $2/3$ and $1/3$ respectively. And, let $w_0^{(n)}$ be the step-kernel corresponding to $n \times n$ symmetric random matrix whose entries (on and above the diagonal) are i.i.d. and has the same distribution as ξ. Then, $\left\|w_0^{(n)} - w_0\right\|_\square \to 0$ almost surely. However, in this case, the coordinates of X_n are i.i.d. (up to the matrix symmetry) and have the same distribution as an RBM (reflected at ± 1) with initial distribution ξ. In particular, $K(X_n(t))$ converges to $w(t) \equiv \mathbb{E}[X_{n,1,2}(t)]$. It is therefore sufficient to show that $\mathbb{E}[X_{n,1,2}(t)]$ is not identically 0 for a.e. $t \in \mathbb{R}_+$.

To see this, we argue by contradiction. If $\mathbb{E}[X_{n,1,2}(t)] = 0$ for all $t \geq 0$ then $\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}[X_{n,1,2}(t)] = 0$. Using [RY04, Exercise 1.12, pg-407], we obtain that $\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}[X_{n,1,2}(t)] = \frac{2}{3}(p_t(-\frac{1}{2}) - p_t(\frac{3}{2})) + \frac{1}{3}(p_t(2) - 1) \neq 0$, where p_t is the standard heat kernel at time t. This yields a contradiction.

Remark 6.8. We should also remark that arranging for $w_0^{(n)}$ such that $\left\|w_0^{(n)} - w_0\right\|_2 \to 0$ as $n \to \infty$ is not difficult. For any w_0 and $n \in \mathbb{N}$, let $w_0^{(n)}$ be the $L^2\big([0,1]^{(2)}\big)$ projection of w_0 on $\mathcal{W}_n$. Then $w_0^{(n)}$ satisfies this condition.

It is worth noting that presence of noise and the boundary $\{-1, 1\}$ in our problem makes it non-trivial. To see this, consider (3.1) for a constant function R_n (i.e., $\nabla R_n \equiv 0$) and

without the local times, say starting at $w_{n,0} \in \mathcal{M}_n$. The solution is a symmetric matrix of independent Brownian motions. It can be easily checked that, if $\lim_{n\to\infty}\|w_{n,0} - w_0\|_\square = 0$, then $\lim_{n\to\infty} \sup_{t\in[0,T]}\big\|X^{(n)}(t) - w_0\big\|_\square = 0$ for any finite $T > 0$. However, if we consider (3.1) again with $\nabla R_n \equiv 0$ but with reflection at the boundary, the coordinate processes are independent reflected Brownian motions. In this case the cut limit of $X^{(n)}(t)$ is also the cut limit of the kernel $\mathbb{E}\big[X^{(n)}(t)\big]$. But reflecting Brownian motions do not have constant expectations in time due to boundary effect. Hence, the limit of $X^{(n)}(t)$ is not constant in t. But, if this limit were a gradient flow, it would be a constant.

6.1.2 Scaling limit without added noise

Recall from Section 3.1.1 that when $\Sigma_n \equiv 0$, equation (3.1) reduces to

$$\mathrm{d}X_n(t) = -n^2\nabla R_n(X_n(t))\,\mathrm{d}t + \mathrm{d}L_n^-(t) - \mathrm{d}L_n^+(t), \quad t \in \mathbb{R}_+, \quad X_n(0) = w_{n,0}, \tag{6.10}$$

such that (X_n, L_n^+, L_n^-) solves the Skorokhod problem on $\mathcal{M}_n$ (see Section 2.4 for details). Moreover, it is shown in Section 3.1.1 that the solution of (6.10) is the same as the solution of (6.11) given below. Furthermore, it is shown in Theorem 4.23 and Theorem 4.28 that if the solution $X_n\colon \mathbb{R}_+ \to \mathcal{M}_n$ of

$$\mathrm{d}X_n(t) = -n^2\nabla R_n(X_n(t)) \circ \mathbb{1}_{G_n(X_n(t))}\{\,\cdot\,\}\,\mathrm{d}t, \qquad t \in \mathbb{R}_+, \tag{6.11}$$

exists, where $G_n(A)$ is the subset of $[n]^2$ (defined in equation (A.22)), then X_n is a gradient flow on $\mathcal{M}_n$ in a suitable sense. Further, it is shown in 4.28 that under reasonable assumptions on R, the sequence of solutions $(X_n)_{n\in\mathbb{N}}$ of equation (6.11) obtained for all natural numbers $n \in \mathbb{N}$, converge to an absolutely continuous curve $w\colon \mathbb{R}_+ \to \mathcal{W}$ (appropriately in the cut metric (see Definition 2.11)), which is a curve of maximal slope [AGS08] (a.k.a. gradient flow) of R, as $n \to \infty$. This yields the following.

Theorem 6.9. *Suppose Assumptions 3.1 and 3.2 hold. Let R be continuous in the cut norm, and λ-semiconvex with respect to $\|\cdot\|_2$ for some $\lambda \in \mathbb{R}$ (see Section 2.3 for definitions). For*

every $n \in \mathbb{N}$, let $X_n \colon \mathbb{R}_+ \to \mathcal{M}_n$ be a gradient flow of R_n staring at $X_n(0) = w_{n,0} = M_n\big(w_0^{(n)}\big) \in \mathcal{W}_n$, and satisfying equation (6.10). If $\big(w_0^{(n)}\big)_{n\in\mathbb{N}}$ converges to $w_0 \in \mathcal{W}$ in the cut norm, then,

$$\lim_{n\to\infty} \sup_{s\in[0,T]} \|K(X_n(s)) - w(s)\|_\square = 0,$$

for any $T > 0$, where w defined as

$$w(t) \coloneqq w_0 - \int_0^t \phi(w(s))\mathbb{1}_{G_{w(s)}}\{\,\cdot\,\},$$

for $t \in \mathbb{R}_+$, is the gradient flow for R.

The above theorem shows that the projected stochastic gradient descent algorithm, that is an explicit Euler scheme, described in Definition 2.2 without the large noise, converges to a gradient flow on graphons that we developed in Chapter 4 using implicit Euler scheme.

We should mention that our method allows us to also obtain a non-asymptotic rate of convergence. We refer the reader to Remark D.4 for details. We will strengthen the rate later in Proposition 6.17.

6.1.3 Examples

In this section we will verify our assumptions for a class of functions introduced as linear functions in Section 4.6.1. Let $\{Z_i\}_{i\in[n]}$ be i.i.d. Uni$[0,1]$. For any kernel $w \in \mathcal{W}$ and any $n \in \mathbb{N}$, sample a random matrix $G_n[w]$ as $G_n[w] \coloneqq (w(Z_i, Z_j))_{(i,j)\in[n]^{(2)}} \in \mathcal{M}_n$. Let $\rho_n([w])$ denote its law, i.e., $\mathrm{Law}(G_n[w]) = \rho_n([w])$. Now let $R \colon \mathcal{W} \to \mathbb{R}$ be defined as a linear function, i.e.,

$$R(w) \coloneqq \int_{\mathcal{M}_n} R_n(z)\rho_n([w])(\mathrm{d}z), \qquad \forall\ w \in \mathcal{W},$$

Let $(\Omega, \mathcal{A})$ be the standard measurable space on $[0,1]^n$. Let $\ell \colon \mathcal{W}\times\Omega$ be the function defined as

$$\ell(w, Z) \coloneqq R_n\Big((w(Z_i, Z_j))_{(i,j)\in[n]^{(2)}}\Big).$$

Let R_n satisfy Assumption 3.1(1) and let R admit a Fréchet-like derivative evaluation map $\phi\colon \mathcal{W} \to L^\infty\big([0,1]^{(2)}\big)$ (see Section 4.6 for conditions). The map ϕ then satisfies

$$\phi(w)(x,y) = \sum\nolimits_{(i,j)\in[n]^2} \mathbb{E}\Big[\nabla R_n\Big((w(Z_p,Z_q))_{(p,q)\in[n]^{(2)}}\Big) \,\Big|\, (Z_i,Z_j) = (x,y)\Big], \tag{6.12}$$

and $D_{\mathcal{W}}\ell(\,\cdot\,;Z)$ for $Z \in [0,1]^n$ satisfies

$$(D_{\mathcal{W}}\ell(\,\cdot\,;Z))(w)(x,y) = \sum\nolimits_{(i,j)\in[n]^2} \nabla R_n\Big((w(Z_p,Z_q))_{(p,q)\in[n]^{(2)}}\big|_{(Z_i,Z_j)=(x,y)}\Big), \tag{6.13}$$

for $w \in \mathcal{W}$ and $(x,y) \in [0,1]^{(2)}$.

Scalar Entropy and Homomorphism density

Examples like the scalar entropy and the homomorphism density functions considered in Section 4.6, all satisfy Assumption 3.1 for some $\kappa_2 \in \mathbb{R}_+$ since $\|\mathrm{Hess}(R_n)\|_{\mathrm{op}}$ exists and is bounded uniformly in the domain. Specifically, for homomorphism density function $R = T_F$ for a simple graph F with n vertices and m edges $\{e_l\}_{l=1}^m$, the constants $\kappa_2 = mn(n-1)$, and for scalar entropy $R = \mathcal{E}$, the constant $\kappa_2 = 2\epsilon^{-1}(1-\epsilon)^{-1}$ on its domain $\mathcal{W}_\epsilon \coloneqq \{W \in \mathcal{W} \mid \epsilon \le W \le 1-\epsilon\}$ where $\epsilon \in (0,1/2)$. Since this implies that there exists $M_\infty \in \mathbb{R}_+$ such that $\|\phi(w)\|_\infty \le M_\infty$ for all w in the domain, these example also satisfy Assumption 3.2 for $\sigma = M_\infty$.

In the following, we define $b\colon [-1,1] \times \mathcal{W} \to L^\infty\big([0,1]^{(2)}\big)$ as $b(w(x,y),w)(x,y) = -\phi(w)(x,y)$ for all $w \in \mathcal{W}$ and a.e. $(x,y) \in [0,1]^{(2)}$. We will now verify Assumption 6.2 when R is the sum of scalar entropy and some homomorphism density T_F for a simple graph F with n vertices and m edges. Note that for this example, we have

$$b(z,w)(x,y) = \log\frac{z}{1-z} + \phi_{T_F}(w)(x,y), \qquad z = w(x,y) \in [\epsilon, 1-\epsilon], \tag{6.14}$$

for a.e. $(x,y) \in [0,1]^{(2)}$ where from equation 4.37,

$$\begin{aligned}\phi_{T_F}(w)(x,y) &= \sum_{l=1}^m \mathbb{E}\left[\prod_{r=1,r\neq l}^m w(Z_{e_r}) \,\middle|\, Z_{e_l} = (x,y)\right] \\ &=: \sum_{l=1}^m \mathbf{t}_{x,y}(F_{e_l}, w), \quad (x,y) \in [0,1],\end{aligned}$$

$Z_e = (Z_{e(1)}, Z_{e(2)})$ and F_{e_l} is the simple graph obtained from F by removing the edge e_l. It is shown in Section 4.6.1 that the map $w \mapsto \mathbf{t}_{(\cdot,\cdot)}(F_e, w)$ continuous as a map from $(\mathcal{W}, d_\square)$ to $\big(L^\infty\big([0,1]^{(2)}\big), d_\square\big)$. To show that $\phi_{T_F}(\,\cdot\,)(x,y)$ is Lipschitz in the cut norm for every $(x,y) \in [0,1]^{(2)}$, it is sufficient to show that $\mathbf{t}_{x,y}(F_e, \,\cdot\,)$ is Lipschitz in the cut norm for $e \in \{e_l\}_{l=1}^m$. For $w_1, w_2 \in \mathcal{W}$, note that

$$\mathbf{t}_{x,y}(F_e, w_1) - \mathbf{t}_{x,y}(F_e, w_2) = \sum_{\{p,q\} \in E(F_e)} I_{p,q},$$

where for any $\{p,q\} \in E(F_e)$,

$$I_{p,q} \coloneqq \int_{[0,1]^{n-2}} (w_1(x_p, x_q) - w_2(x_p, x_q)) \prod_{(i,j) \in E(F_e)\setminus\{p,q\}} w_1(x_i, x_j) \prod_{v \in V(F_e)\setminus e} \mathrm{d}x_v. \tag{6.15}$$

Following the proof in [Lov12, Lemma 10.24], we get $|I_{p,q}| \le \|w_1 - w_2\|_\square$, which yields

$$|\mathbf{t}_{x,y}(F_e, w_1) - \mathbf{t}_{x,y}(F_e, w_2)| \le (m-1)\|w_1 - w_2\|_\square, \tag{6.16}$$

i.e., the Lipschitz constant of $\mathbf{t}_{x,y}(F_e, \,\cdot\,)$ for every $e \in E(F)$ is $m-1$. This implies that the Lipschitz constant of $\phi(\,\cdot\,)(x,y)$ with respect to $\|\,\cdot\,\|_\square$ is $m(m-1)$. Therefore, for b as in equation (6.14), we have

$$\begin{aligned} |b(z, w_1)(x,y) - b(z, w_2)(x,y)| &= |\phi_{T_F}(w_1)(x,y) - \phi_{T_F}(w_1)(x,y)| \\ &\le m(m-1)\|w_1 - w_2\|_\square. \end{aligned} \tag{6.17}$$

Therefore b (as in equation (6.14)) satisfies Assumption 6.2 with $\kappa_\square = m(m-1)$.

Quadratic functions of homomorphism density

More generally, let $k \in \mathbb{N}$ and let $\{F^1, \dots, F^k\}$ be a family of finite simple graphs. Let $c_1, \dots, c_k \in [0,1]$ be fixed constants. Define a function $R\colon \mathcal{W} \to \mathbb{R}$ as

$$R(w) \coloneqq \frac{1}{2} \sum_{\alpha=1}^{k} (T_{F^\alpha}(w) - c_\alpha)^2.$$

Note that a lower bound on R is achieved if $T_{F^\alpha} \equiv c_\alpha$ for all $\alpha \in [k]$. We note that R being a sum of squares of k many functions satisfies Assumption 3.1(2).

Moreover, let $\phi\colon \mathcal{W} \to L^\infty\big([0,1]^{(2)}\big)$ denote the Fréchet-like derivative evaluation map of R. It follows from chain-rule that

$$\phi(w)(x,y) = \sum_{\alpha=1}^{k} (T_{F^\alpha}(w) - c_\alpha)\phi_{T_{F^\alpha}}(w)(x,y).$$

Note that $w \mapsto \phi_{T_{F^\alpha}}(w)$ satisfies Assumption 3.1(2) with $\kappa_{2,\alpha} = m_\alpha(m_\alpha - 1)$ where m_α is the number of edges in F^α. Further note that for any finite graph F and $U, V \in \mathcal{W}$ we have $|T_F(U) - T_F(V)| \le |E(F)|\|U - V\|_\square \le |E(F)|\|U - V\|_2$. A simple calculation using the fact that $|(T_{F^\alpha}(w) - c_\alpha)| \le 1$ for all w and that $\|\phi_{T_F}(w)\|_2 \le |E(F)|$, we obtain that ϕ satisfies Assumption 3.1(2) with

$$\kappa_2 \le \sum_{\alpha=1}^{k} (m_\alpha^2 + \kappa_{2,\alpha}) \le km^2,$$

where $m = \max_{\alpha\in[n]} m_\alpha$.

Similarly, for any edge e in a finite simple graph F, note $w \mapsto \mathbf{t}_{x,y}(F_e, w)$ is $(m-1)$-Lipschitz in cut norm for every $(x,y) \in [0,1]^{(2)}$ and $w \mapsto T_F(w)$ is m-Lipschitz in cut norm where m is the number of edges in F. Using the fact that $\|\phi_{T_F}(w)\|_\infty \le m$ and $T_F(w) \in [0,1]$ for every $w \in \mathcal{W}_0$, we conclude that $\phi(\,\cdot\,)(x,y)$ is km^2-Lipschitz with respect to $\|\cdot\|_\square$ for a.e. $(x,y) \in [0,1]^{(2)}$ and hence ϕ satisfies Assumption 6.2.

Entropy minimization with edge-triangle constraints

We conclude with the discussion of the example mentioned in Section 6.1. Recall the problem of minimizing the scalar entropy $\mathcal{E}$ over $\widehat{\mathcal{W}_0}$ with prescribed edge density $T_-(\,\cdot\,) = e \in [0,1]$ and triangle density $T_\triangle(\,\cdot\,) = \tau \in [0,1]$ (see Section 4.6). As mentioned in [NRS23b], in general this problem does not admit unique minimizer.

Let us consider a relaxation of this problem. Let $\psi\colon \mathbb{R} \to \mathbb{R}$ be a non-decreasing convex function such that $\psi'(-\log(2)) =: A > 1$. Consider minimizing the function

$$w \mapsto R(w) \coloneqq \frac{1}{2}\left((T_-(w) - e)^2 + (T_\triangle(w) - \tau)^2\right) + \psi(\mathcal{E}(w)).$$

Since ψ is non-decreasing, minimizing $\mathcal{E}$ is equivalent to minimizing $\psi \circ \mathcal{E}$. On the other hand, the term $\frac{1}{2}\left((T_{-}(w)-e)^2+(T_{\triangle}(w)-\tau)^2\right)$ penalizes any deviation from the marginal constraint on the edge and triangle densities.

It follows from the previous discussion that $w \mapsto \frac{1}{2}(T_{-}(w)-e)^2+\frac{1}{2}(T_{\triangle}(w)-\tau)^2$ is λ-semiconvex with $\lambda=-8$. On the other hand, $\mathcal{E}$ is 4-semiconvex and therefore $\psi \circ \mathcal{E}$ is $4A$-semiconvex. In particular, if $A>2$ then R is strongly convex and hence admits a unique minimizer and the gradient flow converges exponentially fast to the minimizer of R. In this case, the gradient flow of R converges exponentially fast to the minimizer.

For instance, take $\psi = 4 \cdot \mathrm{id}$ and consider the optimization algorithm described in Definition 2.2. For every $n \in \mathbb{N}$, $X_n \in \mathcal{M}_n$, and $(i,j) \in [n]^{(2)}$, we can evaluate $g_{n,(i,j)}(X_n;\xi)$ as

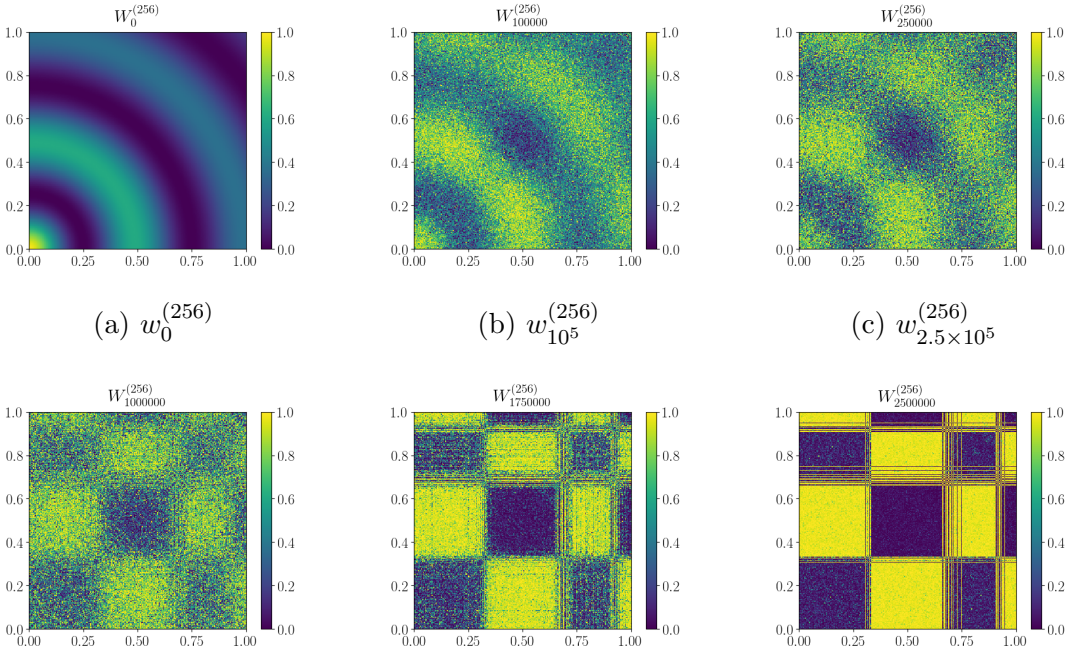
$$\begin{aligned} g_{n,(i,j)}(X_n;\xi) \coloneqq 4\log\left(\frac{X_n(i,j)}{1-X_n(i,j)}\right) &+ (X_n(i_1,i_2)-e) \\ &+ (X_n(i_3,i_4)X_n(i_4,i_5)X_n(i_5,i_3)-\tau)X_n(i,i_6)X_n(i_6,j), \end{aligned}$$

where $\xi=(i_z)_{z\in[6]} \overset{\text{i.i.d.}}{\sim} \mathrm{Uni}([n])^6$. Notice that $\mathbb{E}_\xi[g_n(X_n;\xi)] = \nabla R_n(X_n)$, and Assumption 3.2 is satisfied. Theorem 3.1 and Theorem 6.9 tell us that the (PNSGD) algorithm in the absence of large noise, converges to the minimizer of R as the step size of the algorithm goes to zero, and $n \to \infty$.

If one takes $\psi = \mathrm{id}$ then the function R is not guaranteed to be convex. Therefore, there may be multiple minimizers of R as mentioned in [NRS23b]. Since R is not strictly convex, the gradient flow may not converge to the minimizer, however, it does converge to a stationary point with polynomial rate.

6.1.4 Computational example from Extremal Graph Theory

Following a similar setup as in Section 4.6.4, we simulate the projected noisy SGD algorithm (see Definition 2.2) with $\alpha=1$, $n=256$, step size $\tau=10^{-3}$, $\Sigma_n \equiv 1/3$ from an initial graphon $\left[w_0^{(n)}\right] \in \widehat{\mathcal{W}}_n$ as shown in Figure 6.1a. Figure 6.1 shows instances of the iterative process after 10^5, 2.5×10^5, 10^6, 1.75×10^6 and 2.5×10^6 many steps.



(a) $w_0^{(256)}$ (b) $w_{10^5}^{(256)}$ (c) $w_{2.5\times10^5}^{(256)}$

(d) $w_{10^6}^{(256)}$ (e) $w_{1.75\times10^6}^{(256)}$ (f) $w_{2.5\times10^6}^{(256)}$

Figure 6.1: A noisy stochastic gradient descent simulation over $T_{\triangle} - T_{-}$

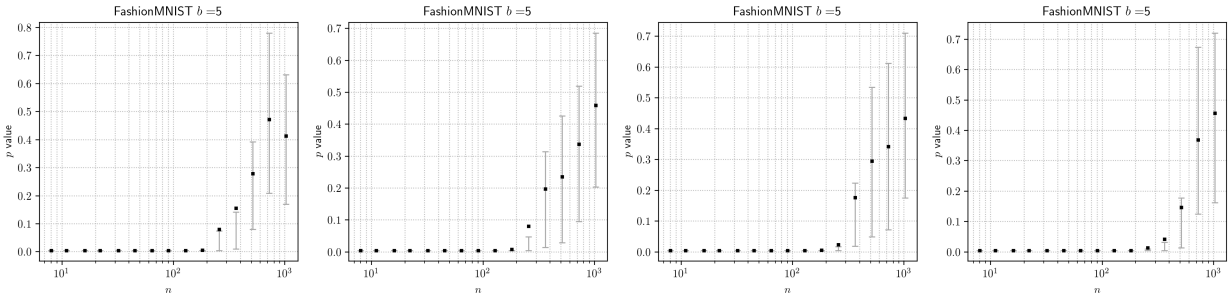
We see in both the Figures 4.2f and 6.1f that after sufficiently many iteration, the approximate solution is close to the one corresponding to a complete bipartite graph as one would expect from Mantel's theorem. Theorem 4.28 and Theorem 6.6 implies that one should expect similar evolutions for all large values of n.

6.1.5 Evidence of propagation of chaos in DNNs

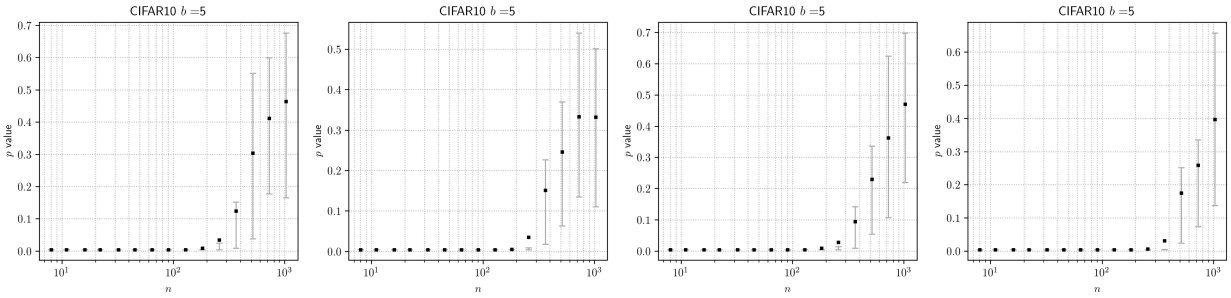
In this section, we aim to empirically verify the propagation of chaos phenomenon along the training dynamics of deep neural networks (DNNs). This phenomenon implies that if we pick fixed and finitely many edges from the graph, then the edge weights tend to become statistically independent along the process as the size of the graphs increases (see Theorem 6.6). For this purpose, we consider feed-forward DNNs of the form described in Section 1.4, shown in Figure 1.3, with $b = 5$ layers. For generality, we use affine maps instead of just linear ones. We consider the popularly used ReLU (Rectified Linear Unit) activation function at each layer, i.e., $x \mapsto \sigma(x) = \max\{0, x\}$ (see equation (1.3)). The loss function ℓ is the squared error, i.e., $\ell(\hat{y}, y) = (\hat{y} - y)^2$ for all $\hat{y}, y \in \mathbb{R}$, such that the risk function on the coupled sequence of parameter matrices A becomes $R_n(A) \coloneqq \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\ell(\hat{y}(X), Y)\right]$ where $\mathcal{D}$ is the data distribution on $\mathbb{R}^d \times \{0, 1\}$. For simplicity, we consider the uniform width scenario where $n_k = n \in \mathbb{N}$ for all $k \in [5]$, i.e., each layer has the same number of neurons.

We consider two popular datasets: CIFAR-10 [KH$^+$09, Chapter 3] and FashionMNIST [XRV17]. Both datasets are used to benchmark machine learning algorithms under various applications. Each dataset contains $N = 6 \times 10^4$ images under 10 categories. For our simulations, each image is vectorized to obtain a $d = 3072$ (for CIFAR-10) and a $d = 784$ (for FashionMNIST) dimensional real-valued vector $x \in \mathbb{R}^d$, respectively. Images in the first five categories are labeled $y = 0$, and the images in the remaining categories are labeled $y = 1$. This gives us 6×10^4 instances of $(X, Y) \sim \mathcal{D}$, using which we train the DNN.

We train our DNN using the batched SGD algorithm with a batch size $B = 10$ (i.e., taking an average of B unbiased estimates of the gradient of R_n) and a step size sequence $\boldsymbol{\tau}$ of the form $(\tau_k = 10^{-2}K^{-1})_{k\in\mathbb{Z}_+}$, where K is the number of epochs (number of passes over



(a) Dataset: FashionMNIST. x-axis: n, y-axis: p-value with interquartile range.



(b) Dataset: CIFAR10. x-axis: n, y-axis: p-value with interquartile range.

Figure 6.2: Evidence of propagation of chaos in DNNs

the training data) and satisfies $K = \lceil k/\lceil N/B \rceil \rceil$. We run the training process for 40 epochs.

To statistically verify the propagation of chaos phenomenon along the entire training process, we collect observations after every epoch. A 5-layer DNN contains $L - 1$ many $n \times n$ matrices that are not directly connected to the input or the output of the DNN (see Figure 1.3). We independently sample 10^5 many $m \times m$ sub-matrices from each of the 4 hidden layers for $m = 2$, and statistically test the hypothesis that the m^2 random variables in the sub-matrices are statistically independent. To do this, we use a kernel-based joint independence test [PBSP18]. We further use a bootstrapping of 200 samples to improve the estimates used in the hypothesis test.

In the sub-figures in Figure 6.2, we visualize our observations as follows. The plots in

any sub-figure correspond to the 4 hidden layers of the DNN. In each plot, we vary n, the width of each layer of the DNN, and plot the statistics of the p-values of the statistical test. From the setup described above, for each n, we have 40 p-values obtained across the entire training dynamics, one for each epoch. We plot the average (using a black square marker) and the interquartile range of the p-values.

For both datasets, we observe that the p-values exceed 0.05 as soon as n becomes moderately large. This suggests that as n grows, we fail to reject the null hypothesis of the test, which posits that the m^2 random variables are statistically independent. Thus, this empirically validates the phenomenon of propagation of chaos in the parameter matrices of a feedforward DNN.

6.2 Scaling limits as curve on MVGs via McKean-Vlasov equations

In this section, we will use the topology of MVG (see Chapter 5) to derive the scaling limit of the processes $(X_n)_{n\in\mathbb{N}}$ defined via solutions of SDEs (6.1). We introduce the following general class of deterministic curves in the space of MVGs described by a stochastic differential equation (SDE). Suppose we are given a pair of functions

$$b\colon [-1,1]\times\mathfrak{W}\to L^\infty\big([0,1]^{(2)}\big),\ \text{and}\quad \Sigma\colon [-1,1]\times\mathfrak{W}\to L^\infty([0,1]^{(2)}), \tag{6.18}$$

where $L^\infty([0,1]^{(2)})$ is the set of all functions $f\colon [0,1]^{(2)}\to\mathbb{R}$ such that $\|f\|_\infty<\infty$.

Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space that supports a standard Brownian motion B and a pair of independent Uniform$[0,1]$ random variables U,V. Given $W_0\in\mathfrak{W}$, consider the following coupled system (X,W,U,V) of one-dimensional reflected diffusion X, a curve W on $\mathfrak{W}$ and the pair of uniform random variables, such that given $(U,V)=(u,v)$, the process $X(\cdot)$ satisfies the initial condition $X(0)\sim W_0(u,v)$ and the SDE

$$\begin{aligned}\mathrm{d}X(t) &= b\left(X(t),\Gamma(t)\right)(u,v)\,\mathrm{d}t+\Sigma(X(t),\Gamma(t))(u,v)\,\mathrm{d}B(t)\\ &\qquad+\mathrm{d}L^-(t)-\mathrm{d}L^+(t),\qquad\qquad\text{(MVG-MKV)}\\ \Gamma(t)(x,y)&\coloneqq\mathrm{Law}(X(t)\mid(U,V)=(x,y)),\quad\ \forall\ (x,y)\in[0,1]^{(2)},\end{aligned}$$

where (X, L^+, L^-) solves the Skorokhod problem [KLRS07] with respect to $[-1, 1]$ (see Section 2.4 for details). The system described by equation (MVG-MKV) will be referred to as the MVG *McKean-Vlasov SDE*. Under appropriate assumptions on b and Σ, Proposition 6.10 shows that the MVSDE admits a pathwise unique solution. Notice that Γ is a deterministic curve on measure-valued kernels, and thus, on measure-valued graphons. The similar McKean-Vlasov SDE introduced in Section 6.1.3 in equation Graphon-MKV obtains convergence only in the sense of graphons and, hence, cannot capture the convergence of general exchangeable arrays. However, there is a corresponding deterministic curve on graphons $\gamma(t) = \mathbb{E}[\Gamma(t)]$ given by the natural projection map. It is useful to think of γ as capturing the evolution of macroscopic properties while Γ describing the microscopic properties.

Where do such processes appear? In equation (6.1) we consider a general class of diffusion on symmetric $n \times n$ matrices whose coordinates are exchangeable and are evolving under a suitable mean-field interaction. In Theorem 6.11 we prove that processes in this general class have corresponding deterministic limits that are examples of (MVG-MKV). This is natural since one can spot that (MVG-MKV) is equivalently characterized by an IEA of independent diffusions satisfying the McKean-Vlasov SDE generated by an i.i.d. sequence of Uniform$[0, 1]$ random variables. Such diffusions naturally arise in the context of SGD of functions defined on the space of graphs. For another example, consider the problem of "soft" optimization described in Section 2.2.2 where, for $\beta > 0$, one may wish to consider a Gibbs measure on $\widehat{\mathcal{W}}$ with a "density" proportional to $\exp(-\beta\mathcal{H})$ for some Hamiltonian $\mathcal{H}$. However, as there is no canonical measure on the space of graphons, this does not seem feasible. On the other hand, consider $\mathcal{H}$ restricted to the space of $n \times n$ graphons $\widehat{\mathcal{W}}_n$. By pulling back the natural map from kernels to graphons and identifying $n \times n$ kernels with $n \times n$ symmetric $[0, 1]$-valued matrices, one can think of $\mathcal{H}$ as a function H_n on symmetric matrices, i.e., $H_n = \mathcal{H} \circ K$ on $\mathcal{M}_{n,+}$. One can define a natural Gibbs measure on $\mathcal{M}_{n,+}$ corresponding to H_n. A large class of commonly used models fall in this umbrella. See the thesis [Che16] for a historical development and some beautiful real-world applications. In particular, it appears in statistical physics models such as the Curie-Weiss models [Che16, Chapter 4], the

exponential random graph models (ERM) [Che16, Chapter 5]. We may wish to sample from such a Gibbs measure whether we are trying to find graphs that approximately minimize the Hamiltonian (i.e., an approximate non-parametric maximum likelihood estimator such as MCMLE [Che16, Chapter 3.3]) or we are sampling from a Bayesian posterior distribution. Although Metropolis or the Gibbs sampling algorithms are popular choices to run MCMC algorithms, their mixing times are generally not known. Another example comes from a series of works of Radin, Sadun and others [RRS14, NRS20, NRS23a, RS23] on the so-called edge-triangle model. Their focus is on a typical graphs with a given number of edges and triangles and to show that they exhibit phase transitions. In [NRS23a, Section 3.1] the authors construct an MCMC scheme to sample from an edge-triangle model (see Section 6.1.3). They justify convergence, not theoretically, but empirically. Given a target edge density e and a triangle density τ, one may easily construct a Hamiltonian that gets minimized when the edge-density and the triangle densities are e and τ, respectively. Then, sampling from this Gibbs measure will approximately sample from an edge-triangle model.

In Section 6.2.2 we will show that, for suitable $\mathcal{H}$, satisfying a semiconvexity condition (Assumption 3.4), the edge density matrix obtained from the Metropolis chain admits limits that are particular cases of (MVG-MKV). With stricter convexity assumptions we will also be able to say something about the exponential rate of convergence. In particular, this is true for all linear combinations of homomorphism functions (see Section 4.6.1). Notice that given $W \in \mathfrak{W}$, consider the corresponding kernel $w = \mathbb{E}[W] \in \mathcal{W}_{[0,1]}$ via the natural projection. Therefore, any function $b_0 \colon [-1,1] \times \mathcal{W} \to\!\!\to L^\infty\big([0,1]^{(2)}\big)$ naturally gives a function $b \colon [-1,1] \times \mathfrak{W} \to L^\infty\big([0,1]^{(2)}\big)$ via the pullback $b(z,W) = b_0(z,w)$. Let $\mathcal{H}$ be a Hamiltonian on $\mathfrak{W}$ that admits a Fréchet-like derivative $D\mathcal{H}$ (see Definition 4.18), and fix a parameter $\beta > 0$. The solution to the McKean-Vlasov SDE (MVG-MKV) with the drift function induced by $b_0(w) = -\beta D\mathcal{H}(w)$ and constant $\Sigma \equiv \sigma$ is analogous to the Langevin diffusion on Euclidean spaces. This family of processes arises as the limit of both stochastic gradient descent on symmetric matrices as well as the following Metropolis chain on the popular stochastic block models (SBM) [EG18].

Note that the graphon convergence of $(X_n)_{n\in\mathbb{N}}$ as defined in SDE (6.1), comes with a loss of microscopic information as discussed earlier in Section 6.1. It may be reasonable to expect that the matrix valued process X_n converges to an IEA as $n \to \infty$ and one should rather consider the convergence of X_n to a deterministic curve on MVG space. In this section we accomplish this convergence and do so for a larger class of functions drift and diffusion coefficient functions than those allowed in Theorem 6.3. This shall allow us to cover a larger class of functions (risk R or Hamiltonian $\mathcal{H}$). We begin with an illustrative example.

Example 6.1. Let F be the triangle graph where two of its edges are decorated by $x \mapsto x^2$, and the third edge is decorated by $x \mapsto x$. Consider the function $R \coloneqq T_{\mathrm{d}}(F, \cdot\,)$. Note that R restricts naturally to a function $R_n \colon \mathcal{M}_n \to \mathbb{R}$ as defined in equation (5.4). One can easily see that $\partial_{(i,j)} T_{\mathrm{d}}(F, \cdot\,)(X_n) = \frac{4}{n^3} \sum_{k=1}^n X^2_{n,(i,k)} X_{n,(k,j)}$ for every $(i,j) \in [n]^{(2)}$. Let $\{B_{n,e}\}_{e\in[n]^{(2)}}$ be a collection of i.i.d. Brownian motions. Consider the following SDE on $\mathcal{M}_n$:

$$\mathrm{d}X_{n,(i,j)}(t) = -\frac{4}{n} \sum_{k=1}^n X^2_{n,(i,k)}(t) X_{n,(k,j)}(t)\,\mathrm{d}t + \mathrm{d}B_{n,(i,j)}(t) + \mathrm{d}L^-_{n,(i,j)}(t) - \mathrm{d}L^+_{n,(i,j)}(t),$$

where $(i,j) \in [n]^{(2)}$ and (X, L^+, L^-) solves the Skorokhod problem on $\mathcal{M}_n$. The above SDE can be recovered as a continuous time limit of the PNSGD algorithm defined in Definition 2.2 when we consider $T_{\mathrm{d}}(F, \cdot\,)$ to be the optimization objective. We remark that the function R does not satisfy the assumptions of Theorem 6.3 as it is not continuous in the cut-metric (see [Jan13, Example C.3]). However, the function R does satisfy Assumption 6.3 for b defined as $b(z, W)(x, y) \coloneqq$

$$\begin{aligned}
&\sum_{\ell=1}^m \mathbb{E}\left[\prod_{s=1}^{\ell-1} \Gamma(F_{e_s}, W)(Z_{e_s}) \cdot \Gamma(F'_{e_\ell}, W)(Z_{e_\ell}) \cdot \prod_{s=\ell+1}^{m} \Gamma(F_{e_s}, W)(Z_{e_s}) \,\middle|\, Z_{e_\ell} = (x,y)\right], \\
&=: \sum_{\ell=1}^m \mathbf{t}_{x,y}(\partial_{e_\ell} F, W), \qquad z \in [-1,1], W \in \mathfrak{W}, (x,y) \in [0,1]^{(2)}, \qquad (6.19)
\end{aligned}$$

where $\{e_s\}_{s=1}^m$ is the set of edges of the skeleton of the triangle graph with $m = 3$ edges, $Z_e \coloneqq (Z_{e(1)}, Z_{e(2)})$ for an edge $e \in E(F)$ and $\partial_e F$ denotes the graph obtained by replacing the decoration at edge $e \in E(F)$ with its derivative. This can be seen by following a very similar argument in Section 6.1.3 and Lemma C.3.

As a consequence of our main result (Theorem 6.11), we will see that the solution of the SDE on $\mathcal{M}_n$ as defined above, converges to the solution X to a MVG McKean-Vlasov SDE, which in this example, takes the form:

$$\begin{aligned} \mathrm{d}X(t) &= -4m_2(W)(u,v)m_1(W)(u,v)\,\mathrm{d}t + \mathrm{d}B(t) + \mathrm{d}L^-(t) - \mathrm{d}L^+(t), \\ W(t)(x,y) &= \mathrm{Law}(X(t) \mid (U,V) = (x,y)), \qquad (x,y) \in [0,1]^{(2)}, \qquad t \in \mathbb{R}_+, \end{aligned}$$

given $(U,V) = (u,v)$. Here m_1 and m_2 evaluate the first and second moment graphons as defined in Example 5.2. This example, naturally extends to all decorated homomorphism density functions, thereby expanding the scope of Theorem 6.3.

More generally we consider the following family of diffusions on symmetric matrices. For $n \in \mathbb{N}$, let $\Sigma_n \colon [-1,1] \times \mathcal{M}_n$ be the restrictions of Σ, i.e., $\Sigma_r(z,X) = \Sigma(z, \mathcal{K}(X))$ and similarly define $b_n \colon [-1,1] \times \mathcal{M}_n$, where $\mathcal{K}$ is defined in Section 5.1.1. In Section 6.1.1, we will discuss the existence of the solution to SDE (MVG-MKV) and argue the weak convergence of the sequence of processes $(X_n)_{n\in\mathbb{N}}$ to the solution of SDE (MVG-MKV).

6.2.1 Deriving the MVG McKean-Vlasov scaling limit

Recall (MVG-MKV) described in Section 6.2. Following a standard Picard's iteration argument, as done in Theorem D.3, it can be shown that MVG McKean-Vlasov SDE (MVG-MKV) admits a pathwise unique solution under appropriate assumptions on b and Σ. For completeness, we record this as Proposition 6.10 but skip the proof.

Assumption 6.3. Recall the definition of the generalized cut norm, $\|\cdot\|_\blacksquare$, from Definition 5.8. Let b, Σ be as in (6.18) and satisfy global Lipschitz conditions, that is, there exists $L, \kappa_\blacksquare \in \mathbb{R}_+$ such that

$$\begin{aligned} \sup_{W\in\mathfrak{W}} \|b(z_1,W) - b(z_2,W)\|_\infty, \sup_{W\in\mathfrak{W}} \|\Sigma(z_1,W) - \Sigma(z_2,W)\|_\infty &\le L|z_1 - z_2|, \\ \sup_{z\in[-1,1]} \|b(z,W_1) - b(z,W_2)\|_\infty, \sup_{z\in[-1,1]} \|\Sigma(z,W_1) - \Sigma(z,W_2)\|_\infty &\le \kappa_\blacksquare \|W_1 - W_2\|_\blacksquare, \end{aligned}$$

for all $W_1, W_2 \in \mathfrak{W}$ and $z_1, z_2 \in [-1,1]$.

Proposition 6.10. *Let b and Σ be as above. Let $\left(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{R}_+}, \mathbb{P}\right)$ be a filtered probability space satisfying the usual conditions that supports a pair of independent* Uniform$[0,1]$ *random variables U, V (measurable with respect to $\mathcal{F}_0$) and a Brownian motion B (adapted to the filtration $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$). Then, for any $W_0 \in \mathfrak{W}$, there exists a pathwise unique strong solution (X, Γ) to the MVG McKean-Vlasov SDE* (MVG-MKV).

In Theorem 6.11 we show that under appropriate assumption on $(X_n(0))_{n\in\mathbb{N}}$, the process X_n converges, uniformly on compact intervals of time, to a deterministic curve Γ on MVGs as $n \to \infty$. And, this curve Γ is described by the McKean-Vlasov system (MVG-MKV).

Theorem 6.11 ([APST23])**.** *Suppose Assumption 6.3 holds. Let $W_0 \in \mathfrak{W}$ and let W be described by the MVG McKean-Vlasov SDE* (MVG-MKV) *with initial condition $W(0) = W_0$. Let X_n be the solution of equation* (6.1) *with initial conditions $X_n(0) \in \mathcal{M}_n$. Suppose that*

$$\lim_{n\to\infty} D_2(X_n(0), W_0) = 0.$$

Then, for any finite time horizon $T > 0$, almost surely,

$$\lim_{n\to\infty} \sup_{t\in[0,T]} \Delta_{\blacksquare}(\mathcal{K}(X_n(t)), \Gamma(t)) = 0. \tag{6.20}$$

The proof of Theorem 6.11 closely parallels the proof of Theorem 6.6. Therefore, we only give a sketch of the proof highlighting only the crucial differences. We provide the proof in Appendix D.2.

Remark 6.12. Note that the proof is stable with respect to small perturbations of drift. More precisely, suppose b_n in (6.1) is replaced by $\widetilde{b}_n$ such that $\left\|\widetilde{b}_n - b_n\right\|_\infty \leq \alpha_n$ for some $\alpha_n \to 0$ as $n \to \infty$. Then, the proof continues to hold and we still obtain the limiting McKean-Vlasov SDE with the same drift b as in Theorem 6.11.

Remark 6.13. In practice, we are often interested in the functions b and Σ in (6.1) that are obtained as the pull back of some functions b_0, Σ_0 defined on $[-1,1] \times \mathcal{W}$ as defined in Section 6.2. Throughout this remark, we always assume this. It follows from Remark 5.5

(i.e. by Lipschitzness of $\Gamma \mapsto \gamma = \mathbb{E}[\Gamma]$) that under appropriate conditions on the $X_n(0)$, solution of SDE (6.1) converges, in probability, to a deterministic curve γ on the space of graphons uniformly on compact time intervals.

Furthermore, note that the McKean-Vlasov equation (MVG-MKV) depends only on $\gamma(t) = \mathbb{E}[\Gamma(t)]$. One can, therefore, say that $\gamma(t)$ satisfies a graphon McKean-Vlasov SDE that satisfies on $(U, V) = (u, v)$

$$\begin{aligned} \mathrm{d}X(t) &= b_0(X(t), \gamma(t)) + \Sigma_0(X(t), \gamma(t)) + \mathrm{d}L^-(t) - \mathrm{d}L^+(t), \\ \gamma(t)(x, y) &= \mathbb{E}[X(t) \mid (U, V) = (x, y)], \end{aligned} \qquad t \in \mathbb{R}_+. \tag{6.21}$$

Thus we recover Theorem 6.6.

We should emphasize the point made in Remark 5.5 once again here. The same IEA gives rise to both McKean-Vlasov SDEs (MVG-MKV) and (6.21) (i.e., (Graphon-MKV)). The crucial difference is X_n converging to this IEA in cut-metric is equivalent to only checking the convergence of homomorphism densities with respect to simple graphs, while X_n converging to this IEA in MVG is equivalent to the convergence of homomorphism densities with respect to decorated simple graphs which is a bigger class of test functions.

6.2.2 Convergence of Metropolis chain algorithm to a Gradient flow on Graphons

In this section we will show that the relaxed Metropolis algorithm discussed in Section 2.2.2, after taking $r \to \infty$ and obtaining the continuous-lime SDE in Theorem 3.4, converges a McKean-Vlasov SDE description of the form of SDE (MVG-MKV) following Theorem 6.11. Further, we will show that as σ is taken to 0, the curve γ obtained after projecting Γ on graphons, is nothing but the gradient flow of the Hamiltonian on the space of graphons.

In Proposition 6.14 we show that, as $n \to \infty$, the paths of this process X_n converge to a deterministic curve on MVG, describe by a McKean-Vlasov SDE (MVG-MKV) with drift b given by $-\beta D\mathcal{H}$ and $\Sigma \equiv \sigma$. By taking the natural projection from MVG to graphons, this implies that the paths of q also converge (see Remark 6.15) to a deterministic curve on the space of graphons in the same scaling limit. Therefore, this deterministic curve on graphons

can be interpreted as the limiting evolution of the adjacency matrices of the sequence of graphs $G(\cdot)$ and is parameterized by $\beta > 0$ and $\sigma > 0$.

Let b_0 be as in Definition 3.2. Recall from Remark 3.3 that $\|b_0 + \beta D\mathcal{H}\|_\infty \to 0$ as $n \to \infty$. As an immediate consequence of Theorem 6.11 (see Remark 6.12) we obtain that the process X_n converges to the McKean-Vlasov SDE defined in (MVG-MKV) with drift given by $-\beta D\mathcal{H}$ as $n \to \infty$. That is, given a pair of Uni($[0,1]$) i.i.d. random variables (U, V) and a standard Brownian motion B on some probability space $(\Omega, \mathcal{G}, \mathbb{P})$, consider the following SDE conditioned on $\{(U,V) = (u,v)\}$,

$$\begin{aligned} \mathrm{d}X(t) &= -\beta D\mathcal{H}(\mathbb{E}[\Gamma^\sigma(t)])(u,v)\,\mathrm{d}t + \sigma\,\mathrm{d}B(t) + \mathrm{d}L^{(0)}(t) - \mathrm{d}L^{(1)}(t), \\ \Gamma^\sigma(t)(x,y) &= \mathrm{Law}(X(t) \,|\, (U,V) = (x,y)), \qquad (x,y) \in [0,1]^{(2)}, \end{aligned} \tag{6.22}$$

for $t \in \mathbb{R}_+$, where $\left(X, L^{(0)}, L^{(1)}\right)$ solves the Skorokhod problem with respect to $[0,1]$ (see Section 2.4).

Proposition 6.14. *Let X_n be a solution of (3.12) with initial condition $X_n(0) \in \mathcal{M}_{n,+}$. If $\lim_{n\to\infty} \|\mathcal{K}(X_n(0)) - W_0\|_\blacksquare = 0$, then X_n converges in MVG sense, in probability, uniformly over compact time intervals, to a deterministic curve Γ^σ in the space of MVGs as $n \to \infty$. Moreover, Γ^σ is described by (6.22) with initial condition W_0.*

Remark 6.15. Consider the curve γ^σ in the space of graphons defined as $\gamma^\sigma(t) \coloneqq \mathbb{E}[\Gamma^\sigma(t)]$ for all $t \in \mathbb{R}_+$. It follows from Remark 6.13 that the random curves $(X_n)_{n\in\mathbb{N}}$ converge in cut-metric, uniformly on compact intervals of time, to γ^σ in probability. And, γ^σ can be recovered as a solution of a MKV SDE satisfying on $\{(U,V) = (u,v)\}$

$$\begin{aligned} \mathrm{d}X(t) &= -\beta D\mathcal{H}(\gamma^\sigma)(u,v)\,\mathrm{d}t + \sigma\,\mathrm{d}B(t) + \mathrm{d}L^{(0)}(t) - \mathrm{d}L^{(1)}(t), \\ \gamma^\sigma(t)(x,y) &= \mathbb{E}[X(t) \,|\, (U,V) = (x,y)], \qquad (x,y) \in [0,1]^{(2)}, \quad t \in \mathbb{R}_+, \end{aligned} \tag{6.23}$$

where $\left(X, L^{(0)}, L^{(1)}\right)$ solves the Skorokhod problem with respect to $[0,1]$ (see Section 2.4).

When $\sigma = 0$, the graphon McKean-Vlasov (6.23) reduces to a deterministic evolution γ of kernels given by

$$\gamma(t)(x,y) = \gamma(0)(x,y) - \beta \int_0^t D\mathcal{H}(\gamma(s))(x,y)\mathbb{1}_{G_{\gamma(s)}}\{\,\cdot\,\}\,\mathrm{d}s, \tag{6.24}$$

for $(x,y) \in [0,1]^{(2)}$, $t \in \mathbb{R}_+$ and initialization $\gamma(0) \in \mathcal{W}_{[0,1]}$, where $G_u \subseteq [0,1]^{(2)}$ for any $u \in \mathcal{W}_{[0,1]}$ is defined as

$$\begin{aligned} G_u \coloneqq &\big\{(x,y) \in [0,1]^{(2)} \,\big|\, u(x,y) = 1, D\mathcal{H}(u)(x,y) < 0\big\} \\ &\cup \big\{(x,y) \in [0,1]^{(2)} \,\big|\, 0 < u(x,y) < 1\big\} \\ &\cup \big\{(x,y) \in [0,1]^{(2)} \,\big|\, u(x,y) = 0, D\mathcal{H}(u)(x,y) > 0\big\}. \end{aligned} \tag{6.25}$$

This can be seen by defining $L^{(0)}$ and $L^{(1)}$ as

$$\begin{aligned} L^{(1)}(t) &\coloneqq + \int_0^t b(\gamma(s))(u,v) \mathbb{1}\{\gamma(s)(u,v) = 1, b(\gamma(s)) > 0\} \,\mathrm{d}s, \\ L^{(0)}(t) &\coloneqq - \int_0^t b(\gamma(s))(u,v) \mathbb{1}\{\gamma(s)(u,v) = 0, b(\gamma(s)) < 0\} \,\mathrm{d}s, \end{aligned}$$

on $\{(U,V) = (u,v)\}$, and observing that the process $(X, L^{(0)}, L^{(1)})$ solves the Skorokhod problem w.r.t. $[0,1]^{(2)}$ (see Section 2.4). It is clear that γ is a constant factor reparametrization of gradient flow of $\mathcal{H}$ on the space of graphons. We now show that this is indeed the case, that is, as $\sigma \to 0$ the curve γ^σ converges to γ on the space of graphons under the cut metric, uniformly over compact intervals of time. To this end, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space equipped with a family of i.i.d. uniform random variables $\{U_i\}_{i\in\mathbb{N}}$ and a collection of independent linear BM $\big\{B_{(i,j)}\big\}_{(i,j)\in\mathbb{N}^{(2)}}$. We can therefore define an IEA X^σ on this probability space such that

$$\begin{aligned} \mathrm{d}X^\sigma_{(i,j)}(t) &= -\beta D\mathcal{H}(\mathbb{E}[\Gamma^\sigma(t)])(U_i,U_j) \,\mathrm{d}t + \sigma \,\mathrm{d}B_{(i,j)}(t) + \mathrm{d}L^{(0)}_{(i,j)}(t) - \mathrm{d}L^{(1)}_{(i,j)}(t), \\ \Gamma^\sigma(t)(x,y) &= \mathrm{Law}(X^\sigma_{(i,j)}(t) \,|\, (U_i,U_j) = (x,y)), \qquad (x,y) \in [0,1]^{(2)}, \end{aligned}$$

for $t \in \mathbb{R}_+$, where $\big(X^\sigma, L^{(0)}, L^{(1)}\big)$ solves the Skorokhod problem with respect to $[0,1]$ (see Section 2.4).

Let γ be as defined in (6.24). Recall that $\gamma(t)$ can be naturally identified with an MVG $\Gamma(t)$ defined as $\Gamma(t)(x,y) \coloneqq \delta_{\gamma(t)(x,y)}$ for $(x,y) \in [0,1]^{(2)}$. On the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ we define another IEA given by $X_{(i,j)}(t) = w(t)(U_i,U_j)$ for $(i,j) \in \mathbb{N}^{(2)}$. Notice that the IEA X satisfies the McKean-Vlasov SDE given by

$$\begin{aligned} \mathrm{d}X_{(i,j)}(t) &= -\beta D\mathcal{H}(\mathbb{E}[\Gamma(t)])(U_i,U_j) \,\mathrm{d}t + \mathrm{d}L^{(0)}_{(i,j)}(t) - \mathrm{d}L^{(1)}_{i,j}(t), \\ \Gamma(t)(x,y) &= \mathrm{Law}(X_{(i,j)}(t) \,|\, (U_i,U_j) = (x,y)), \qquad (x,y) \in [0,1]^{(2)}, \end{aligned} \qquad t \in \mathbb{R}_+, \tag{6.26}$$

where $\left(X, L^{(0)}, L^{(1)}\right)$ solves the Skorokhod problem with respect to $[0,1]$ (see Section 2.4). Note that given $\{U_i\}_{i\in\mathbb{N}}$ the IEA X is deterministic. In particular, $\Gamma(t)(x,y) = \delta_{\gamma(t)(x,y)}$ for a.e. $(x,y) \in [0,1]^{(2)}$.

Proposition 6.16. *Let $\gamma_0 \in \mathcal{W}_{[0,1]}$ be a kernel. Let γ^σ and γ be defined in equation* (6.23) *and* (6.24)*. Then, for every finite $t > 0$, $\sup_{s\in[0,t]} \|\gamma^\sigma(s) - \gamma(s)\|_\square \leq 2C\sigma^2 t \exp(Ct^2)$ for some universal constant $C > 0$.*

The proof of Proposition 6.16 is provided in Appendix D.2.1.

Notice that the drift is a constant multiple of $-D\mathcal{H}$, the direction of steepest descent of $\mathcal{H}$. When $\sigma = 0$, it is clear that the limiting curve is a time-reparametrization of the gradient flow of $\mathcal{H}$ on $\mathcal{W}$. Proposition 6.16 shows that, as $\sigma \to 0$, the family of limiting curves on graphons converges to a time-changed gradient flow of $\mathcal{H}$. Finally, Proposition 6.18 establishes the exponential convergence rate of this flow under appropriate convexity conditions on the Hamiltonian.

6.2.3 Rate of convergence to McKean-Vlasov SDE and to equilibrium

Proposition 6.17 (Non-asymptotic high probability error bound)**.** *Let X_n^σ be a solution to equation* (6.1) *for $\Sigma_n \equiv \sigma$. Let the assumptions of Theorem 6.11 and Proposition 6.16 hold. If the initial condition is i.i.d., then*

$$\sup_{s\in[0,t]} \|\mathcal{K}(X_n^\sigma(s)) - \Gamma(s)\|_\blacksquare^2 \leq C_t n^{-1/56} \log^{3/2} n + (64n^{-1/14} + 2C\sigma^2 t)\mathrm{e}^{C\beta^2\kappa_\blacksquare^2 t^2}, \tag{6.27}$$

with probability at least $1 - 5n^{-3/7} - tn^{-\frac{2}{7\kappa^2 t}}$. Here $\Gamma(s)(x,y) = \delta_{\gamma(s)(x,y)}$ for a.e. $(x,y) \in [0,1]^{(2)}$, and $s \in \mathbb{R}_+$ following equation (6.23)*; and $\kappa = 32\sqrt{6}(L^2 + 2\kappa_\blacksquare^2)^{1/2}$.*

The proof of Proposition 6.17 is provided in Appendix D.2.1.

Proposition 6.18 (Convergence McKean-Vlasov (6.24) to equilibrium when $\sigma = 0$)**.** *Let $\mathcal{H}$ be $\delta_\square$-lower semicontinuous and satisfy Assumption 3.4 with $\lambda \geq 0$ and $L \in [\lambda, \infty) \cup \{\infty\}$.*

Let γ be the graphon valued curve as defined in (6.24)*. Let $\gamma_* \in \widehat{\mathcal{W}}_{[0,1]}$ be a minimizer of $\mathcal{H}$, then*

$$\mathcal{H}(\gamma(t)) - \mathcal{H}(\gamma_*) \leq \frac{\delta_2^2(\gamma(0), \gamma_*)}{2\beta t}, \qquad t \in \mathbb{R}_+. \tag{6.28}$$

Moreover, if $\lambda > 0$ and $L < \infty$ and $\gamma_* \in \widehat{\mathcal{W}}_{[0,1]}$ is the unique minimizer of the strongly convex function $\mathcal{H}$, then for $t \in \mathbb{R}_+$,

$$\delta_2(\gamma(t), \gamma_*) \leq \mathrm{e}^{-\beta\lambda t}\delta_2(\gamma(0), \gamma_*), \quad \mathcal{H}(\gamma(t)) - \mathcal{H}(\gamma_*) \leq \frac{L}{2}\mathrm{e}^{-2\beta\lambda t}\delta_2^2(\gamma(0), \gamma_*). \tag{6.29}$$

The proof of Proposition 6.18 is provided in Appendix D.2.1.

When $\mathcal{H}$ is a linear combinations of homomorphism densities, it is only semiconvex. However, one may regularize $\mathcal{H}$ by adding a large enough multiple of scalar entropy to make it strictly convex (see Section 4.6) and satisfy the conditions for exponential convergence in Proposition 6.18.

6.2.4 Computational example from Extremal Graph Theory

To illustrate our results in this chapter, we give a concrete example with numerical simulations. To motivate our example again from Section 4.6.4 and Section 6.1.4, we first recall the celebrated Mantel's theorem [Man07] from extremal graph theory. It states that the maximum number of edges in an n-vertex triangle-free graph is $n^2/4$. Further, any Hamiltonian graph with at least $n^2/4$ edges must either be the complete bipartite graph $K_{n/2,n/2}$ or it must be pancyclic [Bon71]. One may attempt to computationally verify this theorem by considering a "softer" version of the problem. That is, consider the Hamiltonian $\mathcal{H}(\,\cdot\,) \coloneqq T_\triangle - \alpha T_-$ for sufficiently small $\alpha > 0$. Here $\triangle$ and $-$ are the triangle and the edge graphs respectively. Recall that the homomorphism density function T_F of simple graph F, defined over unweighted graphs, simply computes the density of the simple graph F in the unweighted graph. Thus, minimizing $\mathcal{H}$ can be roughly thought of as an attempt to minimize the number of triangles in a graph while simultaneously maximizing the number of edges. The linear combinations of homomorphism densities also appear in the study of exponential random

graph models (ERGMs) which is usually defined as a probability measure on finite graphs with density proportional to $\exp(-\mathcal{H})$ where $\mathcal{H}$ is a linear combination of homomorphism density function [CV11]. Hence, in either case the behavior of the Metropolis algorithm to simulate samples from the Gibbs measure is of interest.

We simulate the relaxed Metropolis chain sampling algorithm for $\mathcal{H}$ with $\alpha = 1/4$, $n = 16$, $r = 16$, $\sigma = 1$, $\gamma_r = 1/4r$ and $\beta = 1/4$. In particular, $\mathcal{H}(\,\cdot\,) \coloneqq T_{\triangle} - \frac{1}{4}T_{-}$. The Fréchet-like derivative of the Hamiltonian is given by $D\mathcal{H}(\gamma)(x,y) = 3\int_{[0,1]}\gamma(x,z)\gamma(z,y)\,\mathrm{d}z - 1/4$, for $(x,y) \in [0,1]^{(2)}$, which is an affine transformation of the homomorphism density of 2-stars in the graphon.

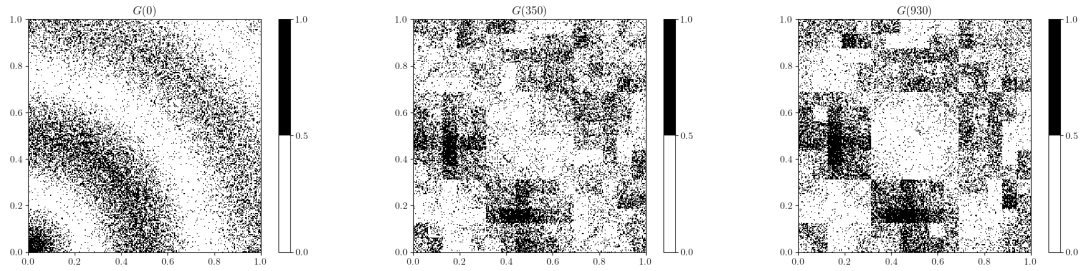
The limit of the adjacency matrix process of the relaxed Metropolis chain (see Section 2.2.2) as $r \to \infty$, followed by $n \to \infty$, and finally $\sigma \to 0$, is given by the a curve $\gamma\colon \mathbb{R}_+ \to \mathcal{W}_{[0,1]}$ given by

$$\gamma(t)(x,y) = \gamma(0)(x,y) - \beta\int_0^t D\mathcal{H}(\gamma(s))(x,y)\mathbb{1}_{G_{\gamma(s)}}\{\}\,\mathrm{d}s, \qquad (x,y) \in [0,1]^{(2)}, \tag{6.30}$$

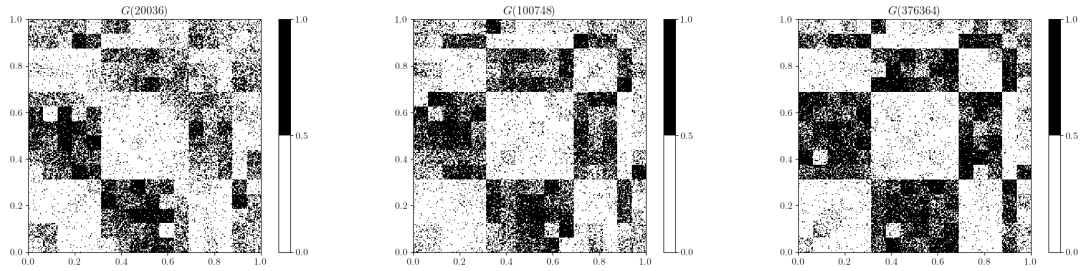
where the starting point $\gamma(0) \in \mathcal{W}_{[0,1]}$ is the L^2-limit of the community edge density kernel of the initialization of the Metropolis chain as $n \to \infty$. The set function G_u for any $u \in \mathcal{W}_{[0,1]}$, defined in equation (6.25), ensures that the velocity field does not point outside the domain of $\mathcal{W}$ when any coordinate of the flow hits the boundary $[0,1]$.

Since the drift is a constant multiple the Fréchet-like derivative of $\mathcal{H}$, the curve γ is a time reparametrization of the gradient flow of $\mathcal{H}$. In Figure 6.3 we see that the iteration sequence of the MCMC chain has a close resemblance with the curves shown in Figure 4.2 and Figure 6.1 which are forward Euler discretizations of the gradient flow of $\mathcal{H}$ on $\big(\widehat{\mathcal{W}}, \delta_2\big)$. After sufficiently many iterations, we see that the community density kernel corresponding to the graph $G^{(n)}_{r,3.7\times10^5}$ is close to the graph the one corresponding to a complete bipartite graph as one would expect from Mantel's theorem.

In Proposition 6.18 we show that if the Hamiltonian is strongly convex, the curve γ converges to the minimizer of $\mathcal{H}$ with an exponential rate. Homomorphism density functions, although semiconvex, are not generally strongly convex. To remedy, one may add to the Hamil-



(a) $G(0)$ (b) $G(3.5 \times 10^2)$ (c) $G(9.3 \times 10^2)$



(d) $G(2.0 \times 10^4)$ (e) $G(1.0 \times 10^5)$ (f) $G(3.7 \times 10^5)$

Figure 6.3: A relaxed Metropolis chain algorithm simulation for $\mathcal{H} = T_{\triangle} - \frac{1}{4}T_{-}$.

tonian a multiple of the scalar entropy function (see Section 4.6) that makes it strongly convex enough to guarantee an exponential rate of convergence. That is, for $\nu > 0$ large enough, the following new Hamiltonian $\mathcal{H}_\nu$ defined as $\mathcal{H}_\nu(\gamma) \coloneqq T_\triangle(\gamma) - \frac{1}{4}T_-(\gamma) + \nu \int_{[0,1]^2} h(\gamma(x,y))\,\mathrm{d}x\,\mathrm{d}y$, where $h(p) = p\log p + (1-p)\log(1-p)$ for $p \in (0,1)$ and zero if $p \in \{0,1\}$, is strongly convex and the corresponding gradient flow curve converges exponentially fast. In fact, in this particular example, $\mathcal{H}_\nu$ as defined above is strongly convex for any $\nu > 9/2$. In particular, if we set $\nu = 5$, following Proposition 6.18, we obtain an exponential rate of convergence to the minimizer with rate $\beta\lambda$ with respect to the δ_2 metric, where the semiconvexity constant $\lambda > \nu - 9/2 = 1/2$. However, to compare our current simulation in Section 4.6.4 and Section 6.1.4, we do not add the entropy regularization here.

6.3 Scaling limits of iterated products of matrices as curves on IEAs

We introduced the class of iterative algorithms of the multiplicative type in Chapter 2.2.3 and derived the continuous-time limit of iterated products of matrices in Chapter 3.3 for a fixed dimension. In this section of the chapter we are interested in taking the dimension of the obtained matrix-valued process to infinity.

To recall, the iterated product of matrices at iteration $\lfloor mt \rfloor$ for any $t \in \mathbb{R}_+$ is

$$P_{n,m}(t) \coloneqq \prod_{k=1}^{\lfloor mt \rfloor} \left(I_n + X_{n,k}^{(m)}\right), \qquad X_{n,k}^{(n)} \coloneqq \frac{\mu_n}{m} M_{n,k}^{(m)} + \frac{\sigma_n}{\sqrt{m}} G_{n,k}^{(m)}. \tag{6.31}$$

where $\mathbb{E}\left[M_{n,k}^{(m)}\right] = A_{n,k}^{(m)}$ and $G_{n,k}^{(m)}$ is an independent matrix with entries i.i.d. as $N(0,1)$ for every $k \in [m]$. Following Theorem 3.8, for every fixed $n \in \mathbb{N}$ we have that as $m \to \infty$, the sequence of curves $(P_{n,m})_{m\in\mathbb{N}}$ converges uniformly over compact time intervals to the curve $\mathrm{Texp}[Y_n]$, where

$$Y_n(t) = \mu_n \int_0^t A_n(s)\,\mathrm{d}s + \sigma_n B_n(t), \qquad t \in \mathbb{R}_+,$$

where B_n is an $n \times n$ matrix with i.i.d. Brownian motions coordinates.

It makes sense to consider the limit of $\mathrm{Texp}[Y_n]$ for a suitable class of semimartingale as $n \to \infty$. This raises an immediate question – namely, in what sense do we take limits of

(curves of) $n \times n$ matrices as $n \to \infty$?

We will return to this question, but before that let us remark that the scaling μ_n in equation (6.31) becomes important here. To make things easier to analyze, let us drop the σ_n term and consider the iterated product where $M_{n,k}^{(m)} = A_n \in \mathcal{M}_n$ for every $k \in [m]$, $m \in \mathbb{N}$ and $n \in \mathbb{N}$. Here, as $m \to \infty$, the product $P_{n,m}$ in equation (6.31) converges to e^{A_n}. Note that the coordinates of e^{A_n} are of the order $O(\mathrm{e}^n)$, if the entries of A_n are $O(1)$. This forces us to choose a suitable scaling μ_n. For instance, if we consider the $\mu_n = n^{-1}$ in dimension, then $\mathrm{e}^{A_n/n} = I_n + O(n^{-1})$ for large $n \in \mathbb{N}$. Thus, the entrywise limit of $\mathrm{e}^{A_n/n}$ becomes trivial. Therefore, a more natural object to consider is $n\mathcal{E}_n$ where $\mathcal{E}_n \coloneqq \mathrm{e}^{A_n/n} - I_n$. And, indeed the entries of $n\mathcal{E}_n$ remain bounded as $n \to \infty$ – and therefore, one can hope to take the limit of $n\mathcal{E}_n$ in some sense as $n \to \infty$. It is also instructive to consider the case when $\left(M_{n,k}^{(m)}\right)_{k\in[m]}$ are the same matrices G_n with entries, say, i.i.d. $N(0,1)$ for all $m \in \mathbb{N}$ and $n \in \mathbb{N}$. In this case, one can see that $n\mathcal{E}_n = G_n + O\big(\frac{1}{n}\big)$. Therefore, it may have a non-trivial limit.

Now, we come to meaning of limit as we increase the dimensionality of the matrices. Consider the case in the above paragraph, that is, where $\left(M_{n,k}^{(m)}\right)_{k\in[m]}$ are the same matrices G_n with entries i.i.d. $N(0,1)$ for all $m \in \mathbb{N}$ and $n \in \mathbb{N}$, and consider the matrix $n\mathcal{E}_n = n\big(\mathrm{e}^{G_n/n} - I_n\big) = G_n + O(\frac{1}{n})$. Intuitively, it makes sense to say that, as $n \to \infty$, the matrix $n\mathcal{E}_n$ converges to an 'infinite matrix' or more precisely to an infinite exchangeable array (IEA) whose entries are i.i.d. Gaussian (see Section 2.6 for definition and details). We can now define a notion of limit of a sequence of matrices $(X_n \in \mathcal{M}_n)_{n\in\mathbb{N}}$ as $n \to \infty$. We say that $(X_n)_{n\in\mathbb{N}}$ converges to an IEA X if for every $r \in \mathbb{N}$, the $r \times r$ exchangeable matrix $X_n\{r\} \coloneqq (X_n(n_i, n_j))_{(i,j)\in[r]^2}$ where $n_1, \ldots, n_r$ is chosen uniformly at random from all r-subsets of $[n]$ converges weakly to $X[r] \coloneqq (X_e)_{e\in[r]^2}$.

Now let $M_{n,k}^{(m)} = 0$ and let $G_{n,k}^{(m)} = G_n$ be a matrix with i.i.d. $N(0,1)$ coordinates for every $k \in [m]$, $m \in \mathbb{N}$ and $n \in \mathbb{N}$. Let $\sigma_n = n^{-1/2}$ and consider $n^{1/2}\mathcal{E}_n$ where $\mathcal{E}_n \coloneqq \mathrm{e}^{G_n/\sqrt{n}} - I_n$. Notice that, in this case, because of the Central Limit Theorem (CLT), the coordinates of $\frac{1}{n^{(k-1)/2}} G_n^k$ are $O(1)$. In particular, the coordinates of $n^{1/2}\mathcal{E}_n = \sum_{k=1}^{\infty} \frac{1}{k!} n^{-(k-1)/2} G_n^k$ remain $O(1)$ as $n \to \infty$. Therefore, one may expect a non-trivial limit for $\big(n^{1/2}\mathcal{E}_n\big)_{n\in\mathbb{N}}$ even with

this scaling. In other words, $\left(n^{1/2}\mathcal{E}_n\right)_{n\in\mathbb{N}}$ converges to an IEA with i.i.d. entries with zero mean and $O(1)$ variance. In the view of Theorem 3.8, the goal of this section is to consider the limits of $\mathrm{Texp}[Y_n]$ – with suitable scaling and centering – for a semimartingale $Y_n(t) = \mu_n \int_0^t A_n(s)\,\mathrm{d}s + \sigma_n B_n(t)$, $t \in [0,1]$.

6.3.1 Choice of scaling

In this section, extending the arguments done above, we will argue about the possible interesting choices of scalings of μ_n and σ_n. To continue, we need some definitions and notations. Unlike defined in Section 2.5, in this section we consider a *kernel* w to be a measurable function $w\colon [0,1]^2 \to \mathbb{R}$ that is square integrable. Let $(U_i)_{i\in\mathbb{N}}$ be a collection of i.i.d. $\mathrm{Uni}([0,1])$ random variables and let w be a kernel. Just as defined in Definition 2.9, we can construct a $r\times r$ exchangeable matrix $w\{r\} \coloneqq (W(U_i,U_j))_{(i,j)\in[r]^2}$ and an IEA $w\{\infty\} \coloneqq (w(U_i,U_j))_{(i,j)\in\mathbb{N}^2}$. If X_n is a $n\times n$ matrix, we will write $X_n\{r\}$ for $K(X_n)\{r\}$. Finally, for two kernels u, v, we define the kernel $u\odot v$ as $(u\odot v)(x,y) \coloneqq \int_0^1 u(x,z)v(z,y)\,\mathrm{d}z$ for a.e. $(x,y)\in[0,1]^2$.

Let $u\colon t \mapsto \int_0^t w(s)\,\mathrm{d}s$ be an absolutely continuous curve of kernels. We can extend the definition of J_k for $k\in\mathbb{N}$ to the kernels by setting

$$J_k(u)(t) \coloneqq \int_{\Delta_k(t)} w(s_k)\odot w(s_{k-1})\odot w(s_1)\cdot \mathrm{d}s_k\cdots\mathrm{d}s_1,$$

and similarly define

$$\Gamma(u)(t) \coloneqq \sum_{k=1}^{\infty} J_k(u)(t). \tag{6.32}$$

Let us now try to make sense of the limit of $\mathrm{Texp}[Y_n]$ when $\sigma_n = n^{-1/2}$ and $A_n \equiv 0$.

Example 6.2. Consider the semi-martingale $Y_n = \frac{B_n}{\sqrt{n}}$, where B_n is a matrix whose coordinates are i.i.d. BMs. Let us look at $\mathrm{Texp}[Y_n]$. It is instructive to consider the case when $n=1$, that is, let B be a standard BM. Note that

$$J_k(B)(t) = \int_0^t \mathrm{d}B(s_k)\int_0^{s_k}\mathrm{d}B(s_{k-1})\ldots\int_0^{s_2}\mathrm{d}B(s_1), \qquad t\in[0,1].$$

It is well-known that $J_k(B)(t) = \frac{1}{k!}H_k(B(t))$ for every $t \in \mathbb{R}_+$, where H_k is the k-th Hermite polynomial [RY04, Proposition 3.8]. It follows that $\mathrm{Texp}[B](t) = \exp\left(B_t - \frac{t}{2}\right)$ for every $t \in \mathbb{R}_+$.

Now consider the case $n > 1$. Fix $k \in \mathbb{N}$ and fix coordinates $(i,j) \in [n]^2$. Note that for

$$J_k\left(\frac{B_n}{\sqrt{n}}\right)(t)(i,j) = \frac{1}{n^{1/2}} \cdot \frac{1}{n^{(k-1)/2}} \sum_{\alpha \in [n]^{k-1}} U_{\alpha,k}(t),$$

where if $\alpha = (i_1, \ldots, i_{k-1}) \in [n]^{k-1}$, then

$$U_{\alpha,k}(t) = \int_0^t \mathrm{d}B(s_k)(i, i_{k-1}) \int_0^{s_k} \mathrm{d}B(s_{k-1})(i_{k-1}, i_{k-2}) \ldots \int_0^{s_2} \mathrm{d}B(s_1)(i_1, j).$$

Notice that $(U_{\alpha,k}(t))_{\alpha \in [n]^{k-1}}$ are i.i.d. random variables with distribution $\frac{1}{k!}H_k(B(t))$ where H_k is k-th Hermite polynomial. Moreover, $U_{\alpha,k}$ and $U_{\beta,k}$ are independent for $\alpha \neq \beta$. Therefore, for every $(i,j) \in [n]^2$, $\sqrt{n}J_k\left(\frac{B_n}{\sqrt{n}}\right)(t)(i,j)$ is again a Gaussian with variance $\frac{t^k}{k!}$ since $\mathrm{Var}\left[\frac{1}{k!}H_k(B(t))\right] = \frac{t^k}{k!}$. Moreover, observe that $J_k\left(\frac{B_n}{\sqrt{n}}\right)(t)(i,j)$ and $J_k\left(\frac{B_n}{\sqrt{n}}\right)(t)(k,l)$ are independent if $(i,j) \neq (k,l)$.

It follows that $\mathrm{Texp}[Y_n](t)$ converges an IEA I_∞ as $n \to \infty$ where $I_\infty(i,j) \coloneqq \mathbb{1}\{i = j\}$ for all $(i,j) \in \mathbb{N}^2$. Noting that $\left\{n^{1/2}J_k\left(\frac{B_n}{\sqrt{n}}\right)(i,j)\right\}_{k \in \mathbb{N}}$ is a collection of independent random variables for every $(i,j) \in [n]^2$, we conclude that every fixed coordinate of $n^{1/2}\mathcal{E}_n$, where $\mathcal{E}_n(t) \coloneqq \mathrm{Texp}[Y_n](t) - I_n$ converges to a Gaussian random variable with mean 0 and variance $(\mathrm{e}^t - 1)$ as $n \to \infty$. As the coordinates of $n^{1/2}\mathcal{E}_n$ are independent, it follows that $\left(n^{1/2}\mathcal{E}_n\right)_{n \in \mathbb{N}}$ converges to an infinite exchangeable array where each coordinate is a time-changed BM. Therefore, for large n, we see that $\mathrm{Texp}[Y_n] \approx I_n + \frac{1}{\sqrt{n}}B_n(\mathrm{e}^t - 1)$ in law.

Now let us now try to make sense of the limit of $\mathrm{Texp}[Y_n]$ when $\mu_n = n^{-1}$ and there is no diffusion term B_n.

Example 6.3. Let $t \mapsto Y_n(t) = \frac{1}{n}\int_0^t A_n(s)\,\mathrm{d}s$ be an absolutely continuous curve. Notice that

$$K(J_k(Y_n)(t)) = \frac{1}{n}J_k(K(Y_n))(t), \qquad k \in \mathbb{N}.$$

Let $\mathcal{E}_n \coloneqq \operatorname{Texp}[Y_n](t) - I_n$. Suppose that there exists some curve of kernels $t \mapsto w(t)$ such that $\sup_{s\in[0,t]} \|K(A_n(s)) - w(s)\|_2 \to 0$ as $n \to \infty$. Then, $t \mapsto nK(J_k(Y_n)(t)) = J_k(K(Y_n)(t))$ also uniformly converges in $L^2([0,1]^2)$ norm to $J_k(w)$. In particular, $nK(J_k(Y_n)(t))$ converges to $J_k(u)(t)$. Therefore, we get that $K(n\mathcal{E}_n)(U_1, U_2)$ converges – in probability – to $\Gamma(u)(t)(U_1, U_2)$ where $\Gamma(u)(t)$ is the kernel $\Gamma(u)(t) \coloneqq \sum_{k=1}^{\infty} J_k(u)(t)$.

Notice, however, that $K(n\mathcal{E}_n)(U_1, U_2)$ is a random coordinate of $n\mathcal{E}_n$. More generally, it follows that if $\{U_i\}_{i\in\mathbb{N}}$ is a collection of i.i.d. Uni$[0,1]$ random variable and $r \in \mathbb{N}$ is fixed, then $r \times r$ random submatrix $(n\mathcal{E}_n)[r]$ of $n\mathcal{E}_n$ converges – in probability – to $r \times r$ matrix $\Gamma(u)(t)\{r\}$. In other words, $(n\mathcal{E}_n)_{n\in\mathbb{N}}$ converges to IEA X defined as $X(t) = \Gamma(u)(t)\{\infty\}$ as $n \to \infty$.

Remark 6.19 (Possible choices of μ_n and σ_n)**.** Notice the difference in the scaling of $n \times n$ matrices in Example 6.2 and Example 6.3. In Example 6.3 if we consider $Y_n = \frac{1}{\sqrt{n}} \int_0^t A_n(s)\,\mathrm{d}s$, then as $n \to \infty$ the coordinates of $J_k(Y_n)(t)$ blow up for $k \geq 2$ while the coordinates of $J_1(Y_n)(t)$ are going to 0. The $\frac{1}{n}$ scaling in this case therefore is necessary. On the other hand, in Example 6.2 if we rather consider $Y_n \coloneqq \frac{B_n}{n}$ then $\left(n^{1/2}\mathcal{E}_n\right)_{n\in\mathbb{N}}$ will converge to the zero IEA, while $(n\mathcal{E}_n)_{n\in\mathbb{N}}$ converges to IEA whose coordinates are independent BM. This explains our choice of $\mu_n = \sigma_n = \frac{1}{n}$ and $\mu_n = \sigma_n^2 = \frac{1}{n}$ as mentioned in the beginning of this section.

6.3.2 Deriving the IEA scaling limit

With the choices of the interesting and non-trivial scalings discussed in Section 6.3.1, we are now ready to state the main result of this section.

Theorem 6.20 (IEA convergence [STH+24])**.** *Let A_n be a continuous curve of $n\times n$ matrices and let Y_n be a $\mathcal{M}_n$-valued semimartingale such that*

$$\mathrm{d}Y_n(t) = \mu_n A_n(t)\,\mathrm{d}t + \sigma_n\,\mathrm{d}B_n(t),$$

and define $\mathcal{E}_n(t) \coloneqq \operatorname{Texp}[Y_n](t) - I_n$ *for* $t \in [0,1]$. *Let* $w \in C([0,1], L^2([0,1]^2))$ *satisfy*

$$\lim_{n\to\infty} \sup_{s\in[0,t]} \|K(A_n)(s) - w(s)\|_2 = 0, \qquad t \in [0,1].$$

Define $u\colon t \mapsto u(t) \coloneqq \int_0^t w(s)\,\mathrm{d}s$. *Then the following statements hold true.*

1. *If* $\mu_n = \sigma_n = n^{-1}$, *then* $n\mathcal{E}_n(t)$ *converges, as* $n \to \infty$, *to an IEA* X *satisfying*

$$X(t) = \Gamma(u)(t)\{\infty\} + B_\infty(t), \qquad t \in [0,1],$$

   where B_∞ *is an IEA with i.i.d. Brownian motion coordinates and is independent of* $\Gamma(u)\{\infty\}$.

2. *If* $\mu_n = \sigma_n^2 = n^{-1}$, *then*

$$n^{1/2}\mathcal{E}_n = n^{1/2}(\operatorname{Texp}[\sigma_n B_n] - I_n) + O\big(n^{-1/2}\big),$$

   and it converges, as $n \to \infty$, *to an IEA* X *satisfying*

$$X(t) = B_\infty(\mathrm{e}^t - 1), \qquad t \in [0,1].$$

3. *If* $\mu_n = \sigma_n^2 = n^{-1}$, *then*

$$n^{1/2}\big(n^{1/2}\mathcal{E}_n - n^{1/2}(\operatorname{Texp}[\sigma_n B_n] - I_n)\big),$$

   converges, as $n \to \infty$, *to an IEA* X *satisfying*

$$X(t) = \Gamma(u)(t)\{\infty\} + Z(t), \qquad t \in [0,1],$$

   where Z *is an zero mean IEA with explicit covariance described in the proof.*

The proof of Theorem 6.20 is provided in Appendix D.3.1.

Notice that limiting IEA in case 2 in above theorem is independent of the the limiting curve of kernels w. In other words, the limit is trivial in some sense. It makes sense to do another centering and scaling to obtain a non-trivial limit in this case – that is what we do in case 3 in the above theorem.

Remark 6.21 (IEA approximation)**.** Following Theorem 6.20, we can infer the following law approximations when $n \in \mathbb{N}$ is large.

1. When $\mu_n = \sigma_n = n^{-1}$, we have that

$$\mathrm{Texp}[Y_n](t) = I_n + \frac{1}{n}\Gamma(u)(t)\{n\} + \frac{1}{n}B_n(t) + \frac{1}{n}E_n(t),$$

   where the coordinates of $E_n(t)$ have $O(1/n)$ variance.

2. When $\mu_n = \sigma_n^2 = n^{-1}$, we have

$$\mathrm{Texp}[Y_n](t) = I_n + \frac{1}{n}\Gamma(u)(t)\{n\} + \frac{1}{\sqrt{n}}B_n(\mathrm{e}^t - 1) + \frac{1}{n}Z_n(t) + \frac{1}{n}E_n(t),$$

   where once again $E_n(t)$ has entrywise variance of order $O(\frac{1}{n})$ and $Z_n(t)$ has Gaussian entries with explicit covariance that is non-zero only for elements in the same row of same column.

Remark 6.22. Following Remark 6.21, let $h_0 \in \mathbb{R}^n$ and let $h_t \coloneqq \mathrm{Texp}[Y_n](t)h_0$ for $t \in \mathbb{R}_+$.

1. When $\mu_n = \sigma_n = n^{-1}$, it is easy to show (see Appendix D.3.2) that the coordinates of $\frac{1}{n}B_n(t)h_0$ are i.i.d. Gaussian with variance $n^{-2}\|h_0\|_2^2$, while the coordinates of $\frac{1}{n}E_n(t)h_0$ are also Gaussian with variance of the same order $O\big(n^{-2}\|h_0\|_2^2\big)$. In particular, for large n, we have the following approximation for h_t

$$h_t \approx h_0 + \frac{1}{n}\Gamma(u)(t)\{n\}h_0 + O\big(n^{-1}\|h_0\|_2\big),$$

   where the above error in the approximation is coordinatewise. We also see that the coordinates of $h_0 + \frac{1}{n}\Gamma(u)(t)\{n\}h_0$ are $O\big(\mathrm{e}^{Ct}\|h_0\|_2\big)$.

2. When $\mu_n = \sigma_n^2 = n^{-1}$, similar to the previous case, we obtain

$$h_t = h_0 + \frac{1}{n}\Gamma(u)(t)\{n\}h_0 + \frac{1}{\sqrt{n}}B_n(\mathrm{e}^t - 1)h_0 + \frac{1}{n}Z_n(t)h_0 + \frac{1}{n}E_n(t)h_0,$$

   where $E_n(t)$ has Gaussian coordinates with variance $O(1/n)$. In particular, just like the previous case the entries of $\frac{1}{n}E_n(t)h_0$ have variance of order $O(n^{-2}\|h_0\|_2^2)$. However,

unlike the previous case, we notice that the coordinates of $\frac{1}{\sqrt{n}}B_n(\mathrm{e}^t-1)h_0$ are i.i.d. mean 0 Gaussian with variance $(\mathrm{e}^t-1)\|h_0\|_2^2$. And, similarly, the coordinates of $\frac{1}{n}Z_n(t)h_0$ are zero mean Gaussians with variance and covariance between its coordinates growing with t. This yields, the following approximation of h_t for large n,

$$h_t \approx h_0 + \frac{1}{n}\Gamma(u)(t)\{n\}h_0 + \sqrt{\mathrm{e}^t-1}\|h_0\|_2\xi + \|h_0\|_2\eta_t + O\big(n^{-1}\|h_0\|_2\big),$$

where $\xi \in \mathbb{R}^n$ is a vector of i.i.d. standard Gaussian random variables and η_t is also a vector of Gaussian with each coordinate having variance of the order $O\big((\mathrm{e}^{Ct}-1)^2 + C^2t^4\mathrm{e}^{2Ct}\big)$ and absolute covariance between its coordinates of the order $O\big(C^2t^4\mathrm{e}^{2Ct}\big)$. As previously, the approximation error $O(n^{-1}\|h_0\|_2)$ is coordinatewise and we once again note that the coordinates of $h_0 + \frac{1}{n}\Gamma(U)(t)\{n\}h_0$ are $O\big(\mathrm{e}^{Ct}\|h_0\|_2\big)$.

The moral of the above discussion is that in the first case, we can approximate h_t as $h_0 + \frac{1}{n}\Gamma(U)(t)\{n\}h_0$ up to a vanishing error. In the second case, we still have the approximation of h_t as $h_0 + \frac{1}{n}\Gamma(u)(t)\{n\}h_0$ plus a mean zero Gaussian noise. The signal has a magnitude of $\Theta\big(\mathrm{e}^{Ct}\big)$. For $t = o(1)$, noise is of the order $O(Ct)$ with correlation of the order $O(Ct^2)$. For $t = \Omega(1)$, the noise is of the order $\Theta\big((1+Ct^2)\mathrm{e}^{Ct}\big)$, and the absolute correlation is of the order $\Theta\Big(\frac{O(C^2t^4)}{1+O(C^2t^4)}\Big)$. This manifests itself via the fact that the noise has a variance that is non-vanishing in dimension – but the noise in each coordinate can be described explicitly.

6.3.3 Deriving the operator scaling limit

Another closely related notion of convergence is the convergence in the sense of operators. Let W be a bounded kernel. One can associate with w a Hilbert-Schmidt integral operator T_w on $L^2([0,1])$ as

$$(T_w f)(x) \coloneqq \int_0^1 w(x,y)f(y)\,\mathrm{d}y, \qquad f \in L^2([0,1]), \quad x \in [0,1].$$

Using the correspondence between A_n and $K(A_n)$, we can therefore, associate a Hilbert-Schmidt operator with every $A_n \in \mathcal{M}_n$. Let $L_n^2([0,1]) \coloneqq$

$\{f \in L^2([0,1]) \mid f \text{ is constant a.e. on } (i-1/n, i/n]\}$. Note that L^2_n is a linear subspace. Let $\mathcal{P}_n$ be the projection operator on $L^2_n([0,1]^2)$. Note that $\mathcal{P}_n$ is the integral operator corresponding to the kernel $nK(I_n)$ where I_n is the $n \times n$ identity matrix. Note that for any $A_n \in \mathcal{M}_n$, the operator $T_{K(A_n)}$ on $L^2([0,1])$ and $T_{K(A_n)}$ commutes with $\mathcal{P}_n$. Recall that a sequence of operators $(T_n)_{n\in\mathbb{N}}$ on L^2 are said to converge to T, in strong sense, if $\|T_n f - Tf\|_2 \to 0$ for every $f \in L^2([0,1])$. Naturally, we say that $(A_n \in \mathcal{M}_n)_{n\in\mathbb{N}}$ converges to an operator T on $L^2([0,1]^2)$ if $\big(T_{K(A_n)}\big)_{n\in\mathbb{N}}$ converges to T in strong sense. Note that even if $\big(T_{K(A_n)}\big)_{n\in\mathbb{N}}$ converges to some operator T, the limiting operator T need not be a Hilbert-Schmidt operator. For example, $A_n \coloneqq nI_n$ converges (in the above sense) to the identity operator $\mathrm{id}_{L^2([0,1])}$ on $L^2([0,1])$ which is not compact and hence not a Hilbert-Schmidt operator. Another important observation to make is that if $A_n, B_n \in \mathcal{M}_n$ then $T_{K(A_n)}T_{K(B_n)} = T_{K(A_n)\cdot K(B_n)} = T_{K\left(\frac{1}{n}A_nB_n\right)}$. Notice that if $(A_n \in \mathcal{M}_n)_{n\in\mathbb{N}}$ is a sequence of symmetric matrices and $(K(A_n))_{n\in\mathbb{N}}$ converges in $L^2([0,1]^2)$ to a symmetric kernel W, then $n\mathrm{e}^{A_n/n}$ converges to e^{T_W} where e^T is the exponential of self-adjoint compact operator defined via functional calculus.

Theorem 6.23 (Operator convergence [STH+24])**.** *For every $n \in \mathbb{N}$, let Y_n be a semi-martingale such that $\mathrm{d}Y_n(t) = \mu_n A_n(t)\,\mathrm{d}t + \sigma_n\,\mathrm{d}B_n(t)$, where A_n is continuous. Set $\mathcal{E}_n \coloneqq \mathrm{Texp}[Y_n] - I_n$. Suppose that $(A_n)_{n\in\mathbb{N}}$ converges to a curve of operators T uniformly on compact intervals of time. Let $\sup_{s\in[0,t]}\|T(s)\|_{\mathrm{op}} \le C_t$ for every $t \in [0,1]$. Then, the following statements hold as $n \to \infty$:*

1. *If $\mu_n = \sigma_n = \frac{1}{n}$, then $n\mathcal{E}_n$ converges in operator norm to the curve of operator, $T_{\Gamma(u)}$, uniformly over compact subsets of time.*

2. *If $\mu_n = \sigma_n^2 = \frac{1}{n}$, then $n^{1/2}\mathcal{E}_n$ converges in operator norm to the constant curve of zero operator, uniformly over compact subsets of time.*

3. *If $\mu_n = \sigma_n^2 = \frac{1}{n}$, then $n^{1/2}\big(n^{1/2}\mathcal{E}_n - n^{1/2}(\mathrm{Texp}[\sigma_n B_n] - I_n)\big)$ converges in operator norm to the curve of operator $T_{\Gamma(u)}$, uniformly over compact subsets of time.*

The proof of Theorem 6.23 is provided in Appendix D.3.2.

If $t \mapsto K(A_n)(t)$ converges to a curve of kernels w in the cut metric, uniformly over compact subsets of time t as $n \to \infty$, then A_n converges to operators $t \mapsto T(t) \coloneqq T_{w(t)}$ over compact intervals of t as $n \to \infty$ [Lov12, Lemma 8.12], i.e., $\left\|T_{K(A_n)}(t) - T_w(t)\right\|_{\mathrm{op}} \leq \|K(A_n)(t) - w(t)\|_{\square}^{1/4}$. Combining this with the fact that $T_{K(nI_n)}$ converges – in strong topology – to $\mathrm{id}_{L^2([0,1])}$ as $n \to \infty$. Thus, we obtain the following corollary of Theorem 6.23.

Corollary 6.24. *Let Y_n be as in Theorem 6.23. Assume that $K(A_n)$ converges in the cut-norm to a continuous curve of kernels W as $n \to \infty$. Then the following hold as $n \to \infty$.*

1. *If $\mu_n = \sigma_n = \frac{1}{n}$, then $n\,\mathrm{Texp}[Y_n]$ converges in – strong topology – to $\mathrm{Texp}[T_u](t)$, where $T_u(t) \coloneqq \int_0^t T_{w(s)}\,\mathrm{d}s$.*

2. *If $\mu_n = \sigma_n^2 = \frac{1}{n}$, then $n\,\mathrm{Texp}[Y_n]$ converges in – strong topology – to the identity operator $\mathrm{id}_{L^2([0,1])}$ on $L^2([0,1])$.*

In Chapter 7, we will use the results obtained in this section to argue how the application of the iterated product of matrices acts on an initial state vector. We will show that this setup provides us with enough machinery to understand how neurons in a linear residual neural network evolve across the depth of the network for every input as both the depth (m) and the width (n) of the network go to infinity. This viewpoint will allow us to formulate a control problem for the risk minimization problem of the infinite network.

6.4 Conclusion

Through each subsection of this chapter, we reach the scaling limit characterizations of the three classes of algorithms detailed in Section 2.2. The overarching philosophy and advantage of these scaling limits is to establish a framework of well-defined mathematical tools and principles applicable to problems in optimization, sampling, and large-scale computation that involve complex mean-field interactions. The reader will observe that the permutation

symmetry in two-dimensional exchangeable systems enables us to describe their scaling limits in terms of objects (and their corresponding curves) across a hierarchy of mathematical constructs - namely, graphons, measure-valued graphons, and exchangeable arrays. Depending on the level of detail required for a given application, the appropriate level of description will suffice to draw conclusions from the scaling limit. With this theoretical exposition, in Chapter 7, we will set up a simple machine learning application in which the scaling limits derived in Section 6.3, in terms of exchangeable arrays and graphons, will be of relevance.

Chapter 7

SCALING LIMIT OF LARGE LINEAR RESIDUAL NETWORKS

In this chapter, we investigate the dynamics of input data as it propagates through the layers of a linear residual network, characterized by random weight matrices. The transformation of an input data point through the network layers can be modeled as an application of iterated products of matrices where each matrix is the linear transformation corresponding to the network weights. By examining the asymptotics of iterated products of random matrices under depth maximal update parameterization, following Theorem 3.8 we establish a scaling limit for these products which can be described via a non-commutative exponential of a matrix-valued stochastic process (see Definition 3.5). Extending this analysis to networks where both width and depth approach infinity, in Theorem 7.1, we demonstrate that the neurons become independent, described by a Gaussian process whose mean and variance are explicitly linked to the network weights. This framework simplifies the evolution of neurons, leading to the phenomenon of propagation of chaos, where neuron interactions diminish, and they behave independently in the limit.

In Section 7.1, we introduce the problem in the context of the popular scalings used to initialize neural networks. In Section 7.2, we provide the setup and discuss the implications of scaling the depth and width of the network. In Section 7.3, we present the main results of this chapter, followed by a discussion of related work on the scaling limits of infinite depth networks. In Section 7.4, we support our theoretical results with numerical simulations, demonstrating a fine agreement. Finally, in Chapter 7.6, we discuss the potential directions for future research and some limitations of our work.

7.1 Introduction

Over the decades, deep learning has garnered significant interest in machine learning, mainly due to its recent empirical successes. Increasing the number of parameters by increasing the width and depth of neural networks has consistently led to notable improvements in model performance [KMH+20, HBM+22, ZKHB22, KH23]. The probabilistic, statistical, and optimization behavior of deep networks has been analyzed under various asymptotic setups, infinite width, infinite depth, and both [PC21, AK22].

As the size of a network increases, it is crucial to select a suitable scale-free parameterization that allows one to take the width and depth limits. Previous research has characterized such scaling limits for networks with increasing widths and increasing depths [YYZH24, BNL+24]. This scaling is known as the *depth maximal update parameterization*, or simply Depth-μP. This scaling has been shown to support stable and non-trivial feature learning [YYZH24]. The key to such parametrization is to have network of various sizes converge to a well-defined limit and maintain the same rate of feature learning, thereby allowing ease of hyperparameter transfer and a consistent interpretability.

We focus on the Depth-μP scaling and aim to investigate how neuron preactivations evolve along the depth of a fixed network as both the depth and the width approach infinity. Specifically, we examine residual neural networks with linear activations and aim to characterize the evolution of neurons as they propagate from the input layer to the output layer for a given set of weight matrices.

This scaling limit description of wide and deep neural networks has been extensively studied at the network's initialization, where the weights are initialized with i.i.d. samples from the standard normal distribution [HN20, LNR21, SCF+21, CLS23, HY23a]. Specifically, works such as [CLS23, SCF+21] characterize the limit of a randomly initialized infinite-width and infinite-depth Residual network as a Gaussian process with a kernel that satisfies a particular partial differential equation. Similar investigations are conducted in [GARA18, NXL+18, NLL+24] for various popular architectures of neural networks. Indeed, there is an

extensive body of research focused on deriving scaling limits for the lazy neural tangent kernel regime [HN19], neural network Gaussian process limits [Ant19, Yai20, LS21, ZVCRP21, SEVM$^+$22, Han22], as well as the rich feature learning regime [JHJ$^+$23, BP24, BNL$^+$24].

Since we consider residual networks, the identification of a neuron stays the same across all the layers. To understand the description of the ensemble of neurons, we consider the class of networks where each layer is parameterized by a scaled drift matrix perturbed by independent Gaussian noise. The scaling of weight matrices used is akin to Depth-μP. Keeping the number of neurons (width) in each layer fixed, we first characterize the infinite depth scaling limit by a matrix-valued stochastic process. The evolution of the neurons can be obtained by applying this matrix process to an input of the network. We take a further limit of this process to describe the evolution of the neurons as the width goes to infinity. See Figure 1.4 in Chapter 1 as an illustration. As a result, we show that the neurons become asymptotically independent, and the evolution of a tagged neuron can be characterized by a Gaussian process whose mean and variance can be written explicitly in terms of the weights of the limiting network.

Since we consider linear activations, the problem concerns understanding the corresponding limit of the iterated matrix products of triangular arrays. Since each layer has a residual connection, the corresponding linear operator is a random perturbation of identity by a (possibly random) drift and a random noise matrix. Limits of iterated matrix products have been extensively studied because of their ubiquitous appearance [Bel54, Ber84, FK60, Fur63, Joh94, Tut65, Wat84, TN13, EH18, HW20, HN20]. However, most of these analyses are limited to matrix products for a given sequence of random matrices and do not consider triangular array of matrices which is crucial for applications. Another crucial departure from the previous works is that we describe the scaling limit of the running product of matrices, which is essential to describe the evolution of the neurons at intermediate layers. Finally, most of the previous analysis is limited to fixed dimension of the matrices. As the dimension grows to infinity, the action of a matrix on a vector (appropriately normalized) exhibits propagation of chaos [Szn91, CD22]. Thus, passing to the

infinite depth limit allows for a simpler description of the evolution of neurons.

With this work, we provide a precise description of the scaling limit of the evolution of neurons along the depth of the network with given weights in infinite width and infinite depth regime. The limit of this evolution can described by a Gaussian process of every neuron. The mean of this process is characterized as a *non-commutative exponential* of a curve W of operators that is the limit of the drift in the weight matrices of the network. The variance of this process can also be described explicitly in terms of W. This description allows us to describe the output of the network as a function of depth. As the training aims to minimize some risk function, one can alternatively view the training task as an optimization problem over the mean curves of Gaussian processes, which, as we show, depends non-locally on a curve of operators. This characterization effectively yields an optimal control problem where the input space is the space of continuous curves of operators.

Notation We define an empty product of matrices as identity and always interpret $\prod$ of a finite collection of matrices indexed by time as denoting ordered multiplication going from left to right with increasing time indices. For any $n \in \mathbb{N}$, $A_n \in \mathbb{R}^{[n]^2}$ and any $H_n \in \mathbb{R}^n$, we define their embeddings on $L^2([0,1]^2)$ and $L^2([0,1])$ as the kernel $K(A_n)$ and the function $K(H_n)$, respectively, defined as: $K(A_n)(u,v) = A_{n,(\lceil nu \rceil, \lceil nv \rceil)}$ and $K(H_n)(u) = H_{n,\lceil nu \rceil}$ for every $u, v \in [0,1]$.

7.2 Background and Setup

We consider the simple setup of a linear residual neural network. As also introduced in Chapter 1.4.2, the feedforward computation of the network initialized under the Depth-μP scaling determines the scale of the network weight matrices, the scaling $n^{-1/2}m^{-1/2}$ at every layer, and the output scaling of n^{-1}, where $m \in \mathbb{N}$ is the number of hidden layers (depth) of the network, and $n \in \mathbb{N}$ is the number of neurons in each layer (width). Let $d \in \mathbb{N}$ be the input dimension. We assume that elements an input $x \in \mathbb{R}^d$ are $O(1)$. Therefore, the

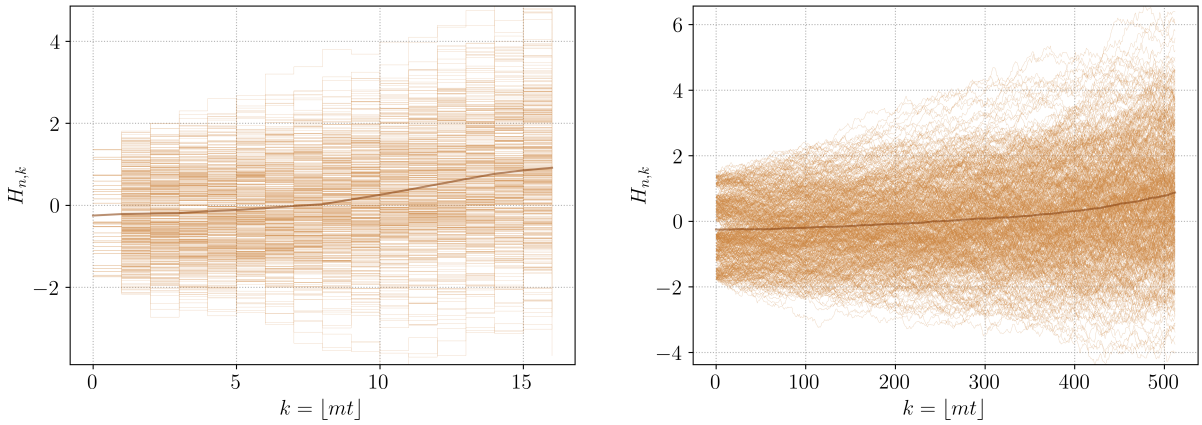


Figure 7.1: Evolution of neurons $H_{n,k}$ from input ($k = 0$) to output layer ($k = m$) for a fixed input, in linear Residual neural networks with depth $m \in \{16, 512\}$ and width $n = 512$. As m increases, we expect the evolution of neurons (in light brown) to converge to a stochastic process. The curve in dark brown represents the mean of the neurons along depth.

network computes its output as:

$$H_{n,0} = Jx, \quad H_{n,k} = H_{n,k-1} + \frac{1}{\sqrt{nm}}\theta_{n,k}^{(m)} H_{n,k-1}, \quad k \in [m], \quad \hat{y}(x) = \frac{1}{n}\sum_{i=1}^{n} H_{n,m,i}. \tag{7.1}$$

Here, the matrix $J \in \mathbb{R}^{n\times d}$ is a fixed sampling matrix with (possibly random) entries of the order $\Theta(1)$. Examples of such matrices are the tiling matrix that repeats the input x for n/d many times if $n \mod d = 0$, and the matrix with elements i.i.d. as $N(0, d^{-1})$. This operation ensures that the elements of $H_{n,0}$ are again $O(1)$. See Figure 7.1, where we illustrate the evolution of the neurons for a single fixed input for a trained linear residual Neural Network (NN) on a classification task.

The trainable matrices $\Theta_n^{(m)} = \left(\theta_{n,k}^{(m)}\right)_{k\in[m]}$ for every $m \in \mathbb{N}$ are the $n \times n$ dimensional weight matrices corresponding to every layer. At the time of initialization, these matrices are set to have i.i.d. $N(0, 1)$ entries, that is $\theta_{n,k}^{(m)} = G_{n,k}^{(m)}$, where the elements in $G_{n,k}^{(m)}$ are all i.i.d. standard Gaussian for every $k \in [m]$ and every $m \in \mathbb{N}$.

At the initialization, notice that, at every layer $k \in [m]$, the change in every neuron

$i \in [n]$, i.e., $H_{n,k,i} - H_{n,k-1,i}$ is random with a variance of $O(m^{-1})$. This observation ensures the network output stays $O(1)$. As the network is trained, the matrices $\Theta_n^{(m)}$ change to have a non-zero mean, which allows the network to learn via an empirical risk minimization process. Notice that if the non-zero drift matrices have $O(1)$ entries, the mean of the entries of $\theta_{n,k}^{(m)} H_{n,k-1}$ are of order $O(n)$, implying that the resulting change in every neuron at every layer is of the order $O\big(n^{1/2}m^{-1/2}\big)$. Because there are m many layers, the net effect of the drift is of the order $O\big(n^{1/2}m^{1/2}\big)$. Since we want to take n and m to infinity, such a choice of scaling the drift cannot be made since it leads to divergence of the neurons. For feature learning, we require that the total change in every neuron must be of the order $O(1)$, which forces the bias term to scale as $O\big(n^{-1/2}m^{-1/2}\big)$. Therefore, a reasonable way through which these weight matrices $\Theta_n^{(m)}$ can be modeled is

$$\theta_{n,k}^{(m)} = \frac{1}{\sqrt{nm}} M_{n,k}^{(m)} + G_{n,k}^{(m)}, \qquad k \in [m],\ m \in \mathbb{N}, \tag{7.2}$$

where $\big(M_{n,k}^{(m)}\big)_{k\in[m]}$ are random bias matrices with means $\big(A_{n,k}^{(m)}\big)_{k\in[m]}$ respectively. We note that although updating the network weights via empirical risk minimization can change the entire distribution of the weights and not just the mean, our description of the limit shall only consider network parameterizations restricted to the above form. Particularly, each element of the weight matrix in this model has a fixed variance of one.

Using the above model, the layer $\lfloor mt \rfloor$, for $t \in [0,1]$, computes $H_{n,\lfloor mt \rfloor} = P_{n,m}(t) \cdot H_{n,0}$ where

$$P_{n,m}(t) \coloneqq \prod_{k=1}^{\lfloor mt \rfloor} \Big(I_n + X_{n,k}^{(m)} \Big), \qquad X_{n,k}^{(n)} \coloneqq \frac{1}{nm} M_{n,k}^{(m)} + \frac{1}{\sqrt{nm}} G_{n,k}^{(m)}. \tag{7.3}$$

It is important for us to consider different finite sequences of matrices for every $m \in \mathbb{N}$. Since we want to scale the depth to infinity, it makes sense for us to be able to correspond layers in networks of a different depth. That is, since we are parameterizing the depth with $t \in [0,1]$, there needs to be a correspondence between layer $\lfloor m_1 t \rfloor$ of a network with m_1 layers and layer $\lfloor m_2 t \rfloor$ of a network with m_2 layers. In particular, for every $t \in [0,1]$, we essentially want $A_{n,\lfloor m_1 t \rfloor}^{(m_1)} \approx A_{n,\lfloor m_2 t \rfloor}^{(m_2)}$ when m_1 and m_2 are sufficiently large. To keep this

correspondence between layers consistent for networks of different depths, we are required to consider triangular sequence of matrices for every fixed width n, such that the piecewise-constant curve $t \mapsto A^{(m)}_{n,\lfloor mt \rfloor}$ has a uniform limit $t \mapsto A_n(t)$ as we take $m \to \infty$. See Figure 1.4 where we illustrate this.

For a finite network with depth m and width n, in order to understand the evolution of the neurons at any layer $\lfloor mt \rfloor$ for some $t \in [0,1]$, we need first to understand the evolution of the product of iterated matrices $P_{m,n}$ defined in equation (7.3). We do this in two steps.

7.2.1 Taking the depth limit

We first take $m \to \infty$ keeping $n \in \mathbb{N}$ fixed and show that the matrix processes $(P_{n,m})_{m\in\mathbb{N}}$ admits a limit that is described as a process on $n \times n$ matrices that starts with the identity matrix. This description allows us to approximate the evolution of neurons in a finite depth finite width network to an infinite depth and finite width network obtained as a limit.

To get an intuitive idea about the limit, observe that $P_{n,m}$ satisfies $P_{n,m}(k/m) - P_{n,m}((k-1)/m) = X^{(m)}_{n,k} P_{n,m}((k-1)/m)$ for every $k \in [m]$. Summing over for $k \in [\lfloor mt \rfloor]$, we get

$$P_{n,m}(t) = I_n + \sum_{k=1}^{\lfloor mt \rfloor} X^{(m)}_{n,k} P_{n,m}((k-1)/m), \qquad t \in [0,1].$$

This suggests that if $Y_{n,m}(t) := \sum_{k=1}^{\lfloor mt \rfloor} X^{(m)}_{n,k}$ admits a limit $Y_n(t)$ as $m \to \infty$ for every $t \in [0,1]$, then $(P_{n,m})_{m\in\mathbb{N}}$ should converge to a curve P_n satisfying $\mathrm{d}P_n(t) = \mathrm{d}Y_n(t)P_n(t)$. In the particular case when $M^{(m)}_{n,k} = A_n$ for some deterministic $n \times n$ matrix A_n, and $G^{(m)}_{n,k} \equiv 0$ for all $k \in [m]$, it is easy to see that $Y_n(t) = tA_n$ for every $t \in [0,1]$, and therefore P_n, in this particular case, solves the matrix-valued ordinary differential equation (ODE): $P_n'(t) = n^{-1}A_nP_n(t)$ with the initialization $P_n(0) = I_n$. The solution, in this case, is well-known to be the matrix exponential $t \mapsto \mathrm{e}^{\int_0^t A_n/n \,\mathrm{d}s}$.

More generally, notice that

$$Y_{n,m}(t) = \frac{1}{nm}\sum_{k=1}^{\lfloor mt \rfloor} A^{(m)}_{n,k} + \frac{1}{nm}\sum_{k=1}^{\lfloor mt \rfloor}\Big(M^{(m)}_{n,k} - A^{(m)}_{n,k}\Big) + \frac{1}{\sqrt{nm}}\sum_{k=1}^{\lfloor mt \rfloor} G^{(m)}_{n,k}, \quad t \in [0,1],$$

where recall that $A_{n,k}^{(m)} = \mathbb{E}\Big[M_{n,k}^{(m)}\Big]$ for every $k \in [m]$. Since $\Big(M_{n,k}^{(m)}\Big)_{k\in[m]}$ are independent, the second term goes to zero as $m \to \infty$. While the first term gives a deterministic drift approximately $\frac{1}{n}\int_0^t A_{n,\lfloor ms\rfloor}^{(m)}\,\mathrm{d}s$ and the last term is given an independent diffusion in each coordinate. This suggests that under appropriate assumptions $(Y_{n,m})_{m\in\mathbb{N}}$ converges to the solution of an SDE:

$$\mathrm{d}Y_n(t) = \frac{1}{n}A_n(t)\,\mathrm{d}t + \frac{1}{\sqrt{n}}\,\mathrm{d}B_n(t), \tag{7.4}$$

where $A_n(t) \coloneqq \lim_{m\to\infty} A_{n,\lfloor mt\rfloor}^{(m)}$ for $t \in [0,1]$ is a deterministic curve of $n\times n$ matrices, and B_n is a $n\times n$ matrix containing i.i.d. Brownian motions. However, because the curve Y_n need not be commutative, i.e., $Y_n(s)Y_n(s')$ need not be equal to $Y_n(s')Y_n(s)$ for $s, s' \in [0,1]$. Therefore, the limit of the iterated matrix products need *not* be the pointwise matrix exponential $t \mapsto \mathrm{e}^{\int_0^t \mathrm{d}Y_n(s)}$.

The solution to the above problem comes from what we call the *non-commutative exponential*, denoted as $\mathrm{Texp}[\,\cdot\,]$. We define the non-commutative exponential of any continuous semi-martingale Y_n in Definition 3.5. In general, $\mathrm{Texp}[Y_n]$ is not a local map, i.e., at any time $t \in \mathbb{R}_+$, $\mathrm{Texp}[Y_n](t)$ can depend on the entire past $(Y_n(s))_{s\in[0,t]}$.

7.2.2 Taking the width limit

We now take a limit as the width n goes to infinity. Notice that if the network has n neurons in every layer, the state of the neurons for the infinite depth network, at depth $t \in [0,1]$ is $\mathrm{Texp}[Y_n](t)H_{n,0}$. Expanding $\mathrm{Texp}[Y_n]$ from equation (7.4), we obtain

$$\mathrm{Texp}[Y_n](t) = I_n + \Gamma\left(\frac{1}{n}\int_0^{\cdot} A_n(s)\,\mathrm{d}s\right)(t) + \Gamma\left(\frac{1}{\sqrt{n}}B_n\right)(t) + \frac{1}{n}Z_n(t) + \frac{1}{n}E_n(t), \tag{7.5}$$

where the matrix Z_n has Gaussian entries, which have an $O(1)$ covariance between the entries in the same row or column, and the entries of E_n have variance $O(n^{-1})$. In particular, for a vector $H_{n,0} \in \mathbb{R}^n$, we notice that the entries of $\big(I_n + \Gamma\big(n^{-1}\int_0^{\cdot} A_n(s)\,\mathrm{d}s\big)(t)\big)H_{n,0}$ are $O(1)$ and are deterministic given A_n and $H_{n,0}$. While $\Gamma\big(n^{-1/2}B_n\big)(t)$ has i.i.d. Gaussian coordinates with variance of order $O(n^{-1})$ and hence $\Gamma\big(n^{-1/2}B_n\big)(t)H_{n,0}$ has entries that

are i.i.d. Gaussian with variance of order $O(1)$. The matrix Z_n also has Gaussian entries; however, the entries of Z_n have non-trivial correlations along the rows and columns. We further show that $n^{-1}Z_nH_{n,0}$ has Gaussian coordinates with $O(1)$ variance in each coordinate and a correlation of order $O(n^{-1})$ between two distinct coordinates. Finally, the entries of E_n have $O(n^{-1})$ variance and therefore the coordinates of $n^{-1}E_nH_{n,0}$ are mean 0 and variance $O(n^{-1})$.

The conclusion is that for large width (and infinite depth), the evolution of two randomly chosen neurons can be asymptotically described as independent Gaussian processes (correlation between them is of order $O(n^{-1})$). Furthermore, the evolution of a fixed randomly chosen neuron can be entirely described by a deterministic drift and a time dependent order $O(1)$ Gaussian fluctuation. Particularly, as $n \to \infty$, the piecewise constant interpolations of the evolution of a randomly chosen neuron, converges weakly in the space of càdlàg paths with respect to the supremum norm on $[0, 1]$. The drift and the Gaussian fluctuation can be written explicitly in terms of the network weights.

In Section 7.3, we describe our main result, where the aforementioned arguments are formally presented in Theorem 7.1, along with their immediate implications.

7.3 Main Results

In this section, we describe the limit of this evolution as the width of the infinite depth network, and show that the neurons exhibit a propagation of chaos behaviour. Let Y_n be as defined in Theorem 3.8. Theorem 3.8 shows that the repeated application of a transformation on the initial set of neurons $H_{n,0} \in \mathbb{R}^n$ can be condensed into a single linear operation given by the non-commutative exponential of the process Y_n at any depth $t \in [0, 1]$. We now describe the action of $\mathrm{Texp}[Y_n](t)$ on the input $H_{n,0}$ as $n \to \infty$.

To achieve this, we need consistency in A_n and $H_{n,0}$ as the dimension $n \to \infty$. Let $(H_{n,0} \in \mathbb{R}^n)_{n\in\mathbb{N}}$ be a sequence that converges (after embedding to $L^2([0, 1])$) to $h_0 \in L^2([0, 1])$ as $n \to \infty$. In particular, $H_{n,0}$ defined as in equation 7.1 for the tiling matrix J satisfies this condition after rearrangements. We are now ready to state our theorem.

Theorem 7.1 (Evolution of neurons in an infinite width network [STH+24]). *Let $t \mapsto w(t)$ be a continuous curve in $L^2([0,1]^2)$, and let $h_0 \in L^2([0,1]^2)$. Let $H_{n,0}$ be as defined as in equation (7.1). Assume that*

$$\lim_{n\to\infty} \sup_{t\in[0,1]} \|K(A_n(t)) - w(t)\|_2 = 0, \qquad \lim_{n\to\infty} \|K(H_{n,0}) - h_0\|_2 = 0. \tag{7.6}$$

Let $u(t) \coloneqq \int_0^t w(s)\,\mathrm{d}s$. Set $H_n(t) \coloneqq \mathrm{Texp}[Y_n](t)H_{n,0}$. Then, the following holds as $n \to \infty$:

1. *A coordinate of the process H_n chosen uniformly at random converges weakly – with respect to the supremum norm – to the solution H of the SDE:*

$$\mathrm{d}H(t) = (\Gamma(u)(t)h_0)(V)\,\mathrm{d}t + \|h_0\|_2(\mathrm{e}^t - 1)^{1/2}\,\mathrm{d}B(t) + \mathrm{d}\eta(t),$$

   starting at $H(0) = h_0(V)$, where $V \sim \mathrm{Uni}([0,1])$ and B is a standard BM and η is a time-changed BM, independent of B, with variance $\int_0^t \|\Gamma(u)(s)h_0\|_2^2\,\mathrm{d}s$ at time $t \in [0,1]$.

2. *For any finite collection of independently and uniformly chosen coordinates of the process H_n, their joint distribution converges weakly to the product measure of the law of H.*

The proof of Theorem 7.1 is provided in Appendix E.

Remark 7.2. In practice, the input of the network often comes from $\mathbb{R}^d$ by the map $H_{n,0} = Jx$, where J samples the coordinates of $x \in \mathbb{R}^d$ to populate $H_{n,0} \in \mathbb{R}^n$. We can take $h_0 \in L^2([0,1])$ as $h_0(u) = x_{\lceil du \rceil}$. Given $(U_i)_{i\in\mathbb{N}}$ i.i.d. as $\mathrm{Uni}([0,1])$, we can define $H_{n,0}$ to be a certain rearrangement of $(h_0(U_i))_{i\in[n]}$ such that $K(H_{n,0})$ converges to h_0 in $L^2([0,1])$ as $n \to \infty$.

Remark 7.3. Following Theorem 7.1 and Assumption 3.5, $\|\Gamma(u)(s)h_0\|_2 = O\big(\|h_0\|_2(\mathrm{e}^{Cs} - 1)\big)$, which shows that the variance of every neuron is $\|h_0\|_2^2\Big((\mathrm{e}^t - 1) + O\Big(\int_0^t (\mathrm{e}^{Cs} - 1)^2\,\mathrm{d}s\Big)\Big)$.

In the next section, we present numerical illustrations to demonstrate our result in a simple setting.

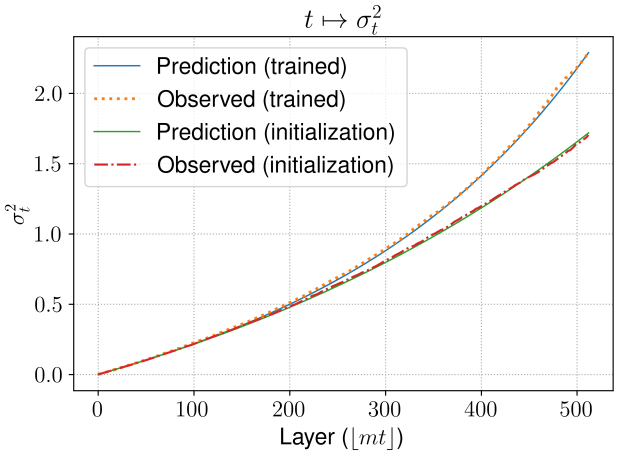
7.4 Numerical Illustrations

To illustrate our results empirically, we consider a simple synthetic dataset consisting of N data points, each sample from $N(0, I_d)$. The label associated with each point is simply the sign of the first coordinate of the point. We train a width n, depth m linear residual neural network using the mean square loss function using stochastic gradient descent. We take $N = 6 \times 10^4$, $n = 512$, $m = 512$, $d = 32$, and use the tiling matrix J as described in Section 7.2.

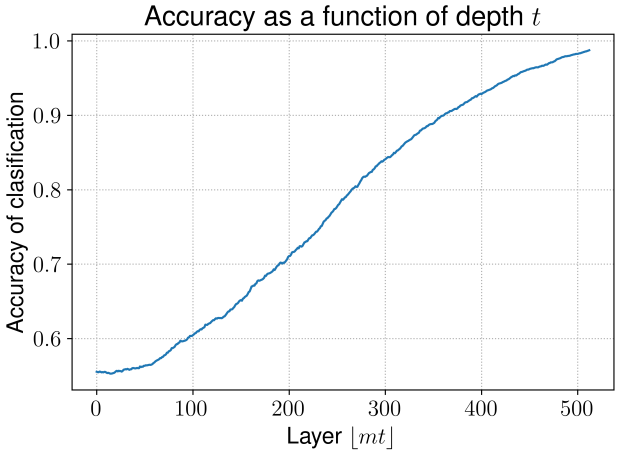
In Figure 7.1, we fix an input and show the evolutions of the n many neurons along the depth in blue. The network outputs the average of the neurons, and the red curve shows this process of the evolution of the average along the depth. The label of the chosen input is $+1$, which also agrees with the sign of the mean process at $t = 1$.

We compute the average variance of all neurons across each data point for every layer, with the mean and initialization subtracted. In Figure 7.2a, we plot this variance as a function of depth, both at the initialization and after the network is trained with stochastic gradient descent (with learning rate $\propto n$ as per [BNL$^+$24]) to achieve maximum training accuracy. Following Theorem 7.1, since there are no drift matrices at the initialization, taking $C = 0$, we expect a variance of $\mathrm{e}^t - 1$. Later, when the network is trained, we expect a variance of $(\mathrm{e}^t - 1) + O\Big(\int_0^t (\mathrm{e}^{Cs} - 1)^2 \,\mathrm{d}s\Big)$ for some $C > 0$. In Figure 7.2a, we plot the observed average variance and the predicted variance from the expression above for $C = 0$ and a suitable $C > 0$ that closely agrees with the observations if we consider the variance without the Big-O notation in our results. We find that at the initial layers, i.e., when t is small, i.e., $t \in o(1)$, the variance increases linearly, and for $t \in \Omega(1)$, the variance increases super-linearly as expected.

Since the limiting process that describes the network's output is a continuous stochastic process, one might question the advisability of evaluating the network's output prematurely, i.e., before the final layer. This approach can be helpful during inference time, particularly when one wants to assess an approximate output with minimal computation or is considering



(a) Noise variance along the depth of the network.



(b) Classification accuracy as a function of depth

Figure 7.2: Noise variance of neurons and classification accuracy along the depth ($n = 512$, $m = 512$).

discarding the latter layers beyond a certain threshold depth to store only a smaller network in memory. In Figure 7.2b, we plot the training classification accuracy of the fully trained network as a function of depth for the above example.

7.5 Related Work

Exploration of μP and mean-field scaling limits in neural networks employs an optimal transport framework to understand training dynamics across various architectures [SMN18, CB18, MMM19, AOY19, CCFRF22]. This includes describing the stochastic gradient descent algorithm's scaling limit through McKean-Vlasov equations [AOY19, NP20] and extending the framework to encompass first-order optimization and MCMC algorithms over large exchangeable matrices, using insights from extremal graph theory [HOP+22, APST23]. Authors in [CCFRF22] demonstrate that Gradient Descent on linear DNNs in infinite-width limits follows a well-defined continuous-time trajectory within an infinite-dimensional space,

reparameterized by ℓ_2-separately chained exchangeable arrays.

Residual neural networks have been studied under various theoretical angles [HY23b, CCRX21, Gul20]. Recent studies have explored training dynamics in infinitely deep and wide residual neural networks under mean-field scaling, with non-linear activations and without noise. Authors in [DCLW22] demonstrate that in such networks, gradient descent evolves into a gradient flow characterized by a PDE that converges to a zero-loss solution. Complementing this, authors in [BPV24] explore gradient flow convergence using a mean-field model and the conditional Optimal Transport distance, demonstrating convergence to a global minimizer under appropriate initial conditions. Additionally, authors in [CLL$^+$24] provide upper bounds on the generalization error of these networks.

Authors in [LTE19] show that stochastic gradient algorithms can be approximated in the weak sense by continuous-time SDEs which allows the application of optimal control theory providing a general methodology for understanding and improving optimization algorithms. We note that such works characterize the training dynamics of algorithms over the parameter space, whereas our interest in this work is to provide a description of the neurons in a fixed network. Authors in [GZ22] use optimal control to provide Hamilton-Jacobi-Bellman asymptotics for the evolution of neurons under the Neural ODE that was introduced by [CRBD18].

Recently, under Depth-μP, the depth and width scaling limits of residual neural network dynamics have been analyzed using dynamical mean field theory (DMFT), proving that this parameterization ensures the convergence of finite-size network dynamics towards a stable and non-trivial limit described by DMFT equations [LNR21, BNL$^+$24].

7.6 The control viewpoint

The description of the scaling limit of a fixed network allows us to state the optimal control problem associated with its risk minimization problem, where the control is the $L^2([0,1]^2)$ absolutely continuous curve w. To see this, note that following Theorem 7.1, the output of

the network at depth $t \in [0,1]$ is

$$y_w(x;t) = \int_0^1 (\mathrm{Texp}[u](t)h_0)(q)\,\mathrm{d}q, \qquad u(t) = \int_0^t w(s)\,\mathrm{d}s, \tag{7.7}$$

where $h_0(p) = x_{\lceil dp \rceil}$ for $p \in [0,1]$. Therefore, the problem of risk minimization reduces to finding an $L^2([0,1]^2)$ absolutely continuous curve $w\colon s \mapsto w(s)$ of $L^2([0,1]^2)$ kernels such that a certain risk function $R\colon w \mapsto \mathbb{E}_{X,Y\sim\mathcal{D}}[\ell(y_w(X;1),Y)]$ is minimized for a given data distribution $\mathcal{D}$ supported on $\mathbb{R}^d \times \mathbb{R}$, for some loss function $\ell\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$. This characterization of the learning problem can be viewed as an optimal control problem where w is the control, and one wants to minimize cost functionals like R simply, or J defined as

$$J(w) \coloneqq \mathbb{E}_{X,Y\sim\mathcal{D}}\left[\int_0^1 \mathrm{e}^{-\alpha t}\ell(y_w(X;t),Y)\,\mathrm{d}t\right],$$

for some $\alpha \in \mathbb{R}$. It is reasonable to expect that if instead of R, one optimizes for J, one can obtain a better accuracy curve than shown in Figure 7.2b. Such ideas have been put into practice in various machine learning applications [SLJ+15, TMK16, SFG+22, LLL+24]. An intuitive interpretation of the control problem is to consider the layers of the network as inducing a field over an appropriate space that steers the signal from the input layer to the network's output. Deriving the stationary conditions will involve the Hamilton-Jacobi-Bellman equation and the Pontryagin maximum principle over the metric space of graphons. For further details on potential approaches to progress on this problem, we refer the reader to the motivating analyses in [WHL18, WE19] and the references therein.

Note that in the infinite depth and infinite width limit, the evolution of neurons in a fixed network is described by a Gaussian process. However, the results in this chapter do not say anything about how this driving process changes during the training. In particular, at the initialization of a network, the drift matrices $\left\{A_{n,k}^{(m)}\right\}_{k\in[m]}$ are all zero implying that in the limit, the curve w is zero. Under the setup we consider, it shall be interesting to analyze how the curve w change under the influence of a training algorithm like SGD (see Definition 2.2).

Beyond the model studied in this chapter, posing the optimization problem on a suitable space of Gaussian processes and investigating the resulting control problem may require the

use of Malliavin calculus [Nua06]. Such an analysis would enable a deeper understanding of the properties of the minimizer, including insights into their stability. Moreover, identifying the conditions under which algorithms like SGD converge to these minimizers is crucial for addressing the broader set of questions raised in this work. This represents a significant direction for future research.

7.7 Conclusion

In this chapter, we adopt a simple yet relevant setup from machine learning and apply the mathematical developments presented in this thesis. We examine mean-field-like interactions that emerge from iterated matrix multiplication in the context of forward computation in residual neural networks. A recurring theme is the propagation of chaos, an emergent phenomenon that arises as the system size approaches infinity, simplifying our understanding by abstracting the problem and recasting it into analytical spaces more amenable to rigorous calculus and analysis. For the context of machine learning, we provide discussion and outline future work stemming from this research in Chapter 8.

Chapter 8

DISCUSSIONS

In this dissertation, we have provided a methodology to derive scaling limits of various dynamics over large systems of matrices. We will now discuss some important issues and insights stemming from this line of work.

8.1 Momentum-Based Iterative Algorithms

In Chapter 2.2.1, we considered algorithms that use the first-order information of the optimization objective to descend. This wide class of algorithms does not capture popular methods like Nesterov's momentum algorithm [Nes83, SBC16], the Adam optimizer [KB14, MLPA22], and other symplectic optimization algorithms [WWJ16]. These algorithms use momentum to accelerate the optimization process and have their continuous-time limits. It is therefore an interesting direction to extend the theory of gradient flows on graphons and McKean-Vlasov SDEs on MVGs to provide scaling limits for such algorithms.

8.2 Theory of Gradient Flows for Measure-Valued Graphons (MVGs)

In Chapter 6.1, we observed that under the asymptotic zero-noise setting, both the stochastic gradient descent (SGD) and Metropolis sampling algorithm suitably converge to the gradient flow of graphons. In Chapter 5, we studied the metric space of MVGs and introduced the invariant L^2 metric, with respect to which the limiting curve of MVG turned out to be an absolutely continuous curve. An immediate direction that arises from this is to develop a theory of gradient flows on MVGs and conjecture that SGD and Metropolis sampling algorithms may also converge to a gradient flow on MVGs. If true, this result would suggest that not only does the system macroscopically converge to a curve of steepest descent, but

the microscopic aspects, such as the distribution of the system's coordinates, also vary in a manner that minimizes the objective in some steepest sense.

8.3 Training Process of an Infinitely Deep and Wide Network

In Chapter 7, we considered a particular model of large noise in the random matrix model. Our study in this chapter, as well as in the setup of Chapter 2.2.3, is currently limited to parameterizing only the drift component of the weight matrices, which restricts the domain of matrix-valued processes available for analysis. A promising extension could involve exploring width limits for the non-commutative exponential of more general matrix-valued processes, incorporating a non-identity covariance in the diffusion term.

8.4 Optimal Control and Training Dynamics

In Chapter 7.6, we showed how the parameterization of the infinitely wide and deep network can be used to frame an optimal control problem. Particularly, framing the problem as an optimal control problem allows us to obtain the stationarity condition using the corresponding Hamilton-Jacobi-Bellman equations.

Additionally, in Chapter 7, the evolution of neurons in a fixed network is described by a Gaussian process in the infinite depth and infinite width limit. However, this result does not address how this driving process changes during training. Investigating the behavior of training dynamics may necessitate the use of Malliavin calculus [Nua06]. This represents a significant future direction.

8.5 Neural Networks with Non-Linearity, Convolution, and Attention

In Chapter 7, we confined our work to linear activations, positioning our research around understanding iterated products of scaled random matrices. A practically relevant extension is to generalize our findings to include non-linear activations, such as the rectified linear unit (ReLU).

In addition to networks containing fully connected layers, an interesting direction is to examine architectures that incorporate convolution operations and attention mechanisms [VSP+17], which are heavily used in the surge of transformer-based architectures, such as those in large language models.

With these discussions, we conclude our thesis, introducing a broad framework and formal language that can be employed to study systems of complex interacting coordinates and their applications in machine learning.

BIBLIOGRAPHY

[ADW23] Romain Abraham, Jean-François Delmas, and Julien Weibel. Probability-graphons: Limits of large dense weighted graphs. *arXiv preprint arXiv:2312.15935*, 2023. 104

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures. Second Edition.* Lectures in Mathematics. ETH Zürich. Birkhäuser Verlag AG, Basel, 2008. 6, 21, 70, 72, 75, 76, 79, 80, 84, 87, 133, 236, 242, 245, 246, 278

[AHS23] Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*, 2023. 18

[AK22] Benny Avelin and Anders Karlsson. Deep limits and a cut-off phenomenon for neural networks. *Journal of Machine Learning Research*, 23(191):1–29, 2022. 167

[Ald81] David J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981. 47, 73, 103

[Ald82] David J Aldous. On exchangeability and conditional independence. *Exchangeability in probability and statistics (Rome, 1981)*, pages 165–170, 1982. 47, 73, 103

[Ald85] David J. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour, XIII–1983. Lecture Notes in Math.*, volume 1117, pages 1–198. Springer-Berlin, 1985. 127

[Ant19] Joseph M Antognini. Finite size corrections for neural network gaussian processes. *arXiv preprint arXiv:1908.10030*, 2019. 168

[AOY19] Dyego Araújo, Roberto I Oliveira, and Daniel Yukimura. A mean-field limit for certain deep neural networks. arXiv preprint arXiv:1906.00193, 2019. 3, 14, 72, 177

[APST23] Siva Athreya, Soumik Pal, Raghav Somani, and Raghavendra Tripathi. Path convergence of markov chains on large graphs. *arXiv preprint arXiv:2308.09214*, 2023. 24, 57, 104, 112, 116, 147, 177

[Aus08] Tim Austin. On exchangeable random variables and the statistics of large graphs and hypergraphs. *Probability Surveys*, 5:80–145, 2008. 73, 127

[Aus12] Tim Austin. Exchangeable random arrays. In *Notes for IAS workshop*, 2012. 73

[Aus15] Tim Austin. Exchangeable random measures. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 51(3):842 – 861, 2015. 73

[Bag20] Sina Baghal. A matrix concentration inequality for products. *arXiv preprint arXiv:2008.05104*, 2020. 36

[BB07] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. 4

[BBW19] Shankar Bhamidi, Amarjit Budhiraja, and Ruoyu Wu. Weakly interacting particle systems on inhomogeneous random graphs. *Stochastic Processes and their Applications*, 129(6):2174–2206, 2019. 125

[BC21] Francis Bach and Lenaïc Chizat. Gradient descent on infinitely wide neural networks: Global convergence and generalization. arXiv preprint arXiv:2110.08084, 2021. 3, 14, 72

[BCCH18] Christian Borgs, Jennifer T. Chayes, Henry Cohn, and Nina Holden. Sparse exchangeable graphs and their limits via graphon processes. *Journal of Machine Learning Research*, 18(210):1–71, 2018. 43

[BCL+08] Christian Borgs, Jennifer T Chayes, László Lovász, Vera T Sós, and Katalin Vesztergombi. Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing. *Advances in Mathematics*, 219(6):1801–1851, 2008. 38, 42, 73, 74, 85

[BCL+12] Christian Borgs, Jennifer T Chayes, László Lovász, Vera T Sós, and Katalin Vesztergombi. Convergent sequences of dense graphs II. multiway cuts and statistical physics. *Annals of Mathematics*, pages 151–219, 2012. 38, 73

[BCM21] Alessandra Bianchi, Francesca Collet, and Elena Magnanini. Limit theorems for exponential random graphs. *arXiv preprint arXiv:2105.06312*, 2021. 126

[BCN18] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. 1, 8

[BCN20] Gianmarco Bet, Fabio Coppini, and Francesca R Nardi. Weakly interacting oscillators on dense random graphs. *arXiv preprint arXiv:2006.07670*, 2020. 126

[BCW20] Erhan Bayraktar, Suman Chakraborty, and Ruoyu Wu. Graphon mean field systems. *arXiv preprint arXiv:2003.13180*, 2020. 126

[BEFLY21] Omri Ben-Eliezer, Eldar Fischer, Amit Levi, and Yuichi Yoshida. Ordered Graph Limits and Their Applications. In James R. Lee, editor, *12th Innovations*

in Theoretical Computer Science Conference (ITCS 2021), volume 185 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 42:1–42:20, Dagstuhl, Germany, 2021. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. 73

[Bel54] Richard Bellman. Limit theorems for non-commutative operations. I. *Duke Journal of Mathematics*, 1954. 36, 168

[Ben99] Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pages 1–68. Springer, 1999. 8, 28

[Ber84] Marc A Berger. Central limit theorem for products of random matrices. *Transactions of the American Mathematical Society*, 285(2):777–803, 1984. 36, 168

[BG20] Bhaswar B. Bhattacharya and Shirshendu Ganguly. Upper tails for edge eigenvalues of random graphs. *SIAM Journal on Discrete Mathematics*, 34(2):1069–1083, 2020. 73

[BNL+24] Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In *The Twelfth International Conference on Learning Representations*, 2024. 19, 167, 168, 176, 178

[Bon71] J Adrian Bondy. Pancyclic graphs i. *Journal of Combinatorial Theory, Series B*, 11(1):80–84, 1971. 100, 152

[Bor09] Vivek S. Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009. 8, 28

[BP24] Blake Bordelon and Cengiz Pehlevan. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. *Advances in Neural Information Processing Systems*, 36, 2024. 168

[BPV24] Raphaël Barboni, Gabriel Peyré, and François-Xavier Vialard. Understanding the training of infinitely deep and wide resnets with conditional optimal transport. *arXiv preprint arXiv:2403.12887*, 2024. 178

[BQ16] Yves Benoist and Jean-François Quint. *Random walks on reductive groups*. Springer, 2016. 36

[Bré20] P. Brémaud. *Point Process Calculus in Time and Space: An Introduction with Applications*. Probability Theory and Stochastic Modelling. Springer International Publishing, 2020. 60

[BSM+22] Frederik Benzing, Simon Schug, Robert Meier, Johannes Von Oswald, Yassir Akram, Nicolas Zucchet, Laurence Aitchison, and Angelika Steger. Random initialisations performing above chance and how to find them. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022. 18

[Bub15] Sébastien Bubeck. Convex Optimization: Algorithms and Complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015. 1, 8

[But16] John Charles Butcher. *Numerical methods for ordinary differential equations*. John Wiley & Sons, Hoboken, NJ, 2016. 246

[Cau47] Augustin-Louis Cauchy. Méthode générale pour la résolution des systemes d'équations simultanées. *Comptes Rendus de l'Académie des Science*, 25:536–538, 1847. 8, 9

[CB18] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 3040–3050, Red Hook, NY, USA, 2018. Curran Associates Inc. 3, 14, 72, 177

[CCFRF22] Lénaïc Chizat, Maria Colombo, Xavier Fernández-Real, and Alessio Figalli. Infinite-width limit of deep linear neural networks. arXiv preprint arXiv:2211.16980, 2022. 14, 177

[CCP19] José Antonio Carrillo, Katy Craig, and Francesco S Patacchini. A blob method for diffusion. *Calculus of Variations and Partial Differential Equations*, 58(2):1–53, 2019. 3, 14, 72

[CCRX21] Alain-Sam Cohen, Rama Cont, Alain Rossier, and Renyuan Xu. Scaling properties of deep residual networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2039–2048. PMLR, 18–24 Jul 2021. 178

[CD22] Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: A review of models, methods and applications. *Kinetic and Related Models*, 15(6):1017–1173, 2022. 13, 124, 168

[Cha17] Sourav Chatterjee. *Large deviations for random graphs: École d'Été de Probabilités de Saint-Flour XLV-2015*, volume 2197. Springer, New York, 2017. 32

[Che16] Bobbie G. Chern. *Large deviations approximation to normalizing constants in exponential models*. PhD thesis, Stanford University, 2016. 10, 143, 144

[Chi22] Lénaïc Chizat. Mean-field langevin dynamics : Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022. 67

[CHKT21] Chi-Fang Chen, Hsin-Yuan Huang, Richard Kueng, and Joel A Tropp. Concentration for random product formulas. *PRX Quantum*, 2(4):040305, 2021. 36

[CLL+24] Yihang Chen, Fanghui Liu, Yiping Lu, Grigorios Chrysos, and Volkan Cevher. Generalization of scaled deep resnets in the mean-field regime. In *The Twelfth International Conference on Learning Representations*, 2024. 178

[CLS23] Nicola Muca Cirone, Maud Lemercier, and Cristopher Salvi. Neural signature kernels as infinite-width-depth-limits of controlled resnets. In *International Conference on Machine Learning*, pages 25358–25425. PMLR, 2023. 167

[Cop22] Fabio Coppini. A note on Fokker–Planck equations and graphons. *Journal of Statistical Physics*, 187(2):1–12, 2022. 125

[CRBD18] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018. 178

[CV11] Sourav Chatterjee and S. R. S. Varadhan. The large deviation principle for the Erdős-Rényi random graph. *European Journal of Combinatorics*, 32(7):1000–1017, 2011. Homomorphisms and Limits. 126, 153

[DCLW22] Zhiyan Ding, Shi Chen, Qin Li, and Stephen J Wright. Overparameterization of deep resnet: zero loss and mean-field analysis. *Journal of machine learning research*, 23(48):1–65, 2022. 178

[DGKR15] Peter Diao, Dominique Guillot, Apoorva Khare, and Bala Rajaratnam. Differential calculus on graphon space. *Journal of Combinatorial Theory, Series A*, 133:183–227, 2015. 73, 81

[DGL16] Sylvain Delattre, Giambattista Giacomin, and Eric Luçon. A note on dynamical models on random graphs and fokker–planck equations. *Journal of Statistical Physics*, 165(4):785–798, 2016. 125

[Dia09] Persi Diaconis. The Markov chain Monte Carlo revolution. *Bulletin of the AMS*, 46(2):179–205, 2009. 30

[DJ08] Persi Diaconis and Svante Janson. Graph limits and exchangeable random graphs. *Rendiconti di Matematica e delle sue Applicazioni*, 28(1):33–61, 2008. 73, 103, 104, 106, 115, 127

[DM83] Eugene B Dynkin and Avishai Mandelbaum. Symmetric statistics, Poisson point processes, and multiple Wiener integrals. *The Annals of Statistics*, pages 739–745, 1983. 12, 62

[DM22] Paul Dupuis and Georgi S Medvedev. The large deviation principle for interacting dynamical systems on random graphs. *Communications in Mathematical Physics*, 390(2):545–575, 2022. 125

[DMN+21] Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, and Hoi-To Wai. On the stability of random matrix product with markovian noise: Application to linear stochastic approximation and td learning. In *Conference on Learning Theory*, pages 1711–1752. PMLR, 2021. 36

[Dob79] L. Dobrushin, R. Vlasov equations. *Functional Analysis and its applications*, 13:115–123, 1979. 13, 124

[DSS+20] Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Gromov-Wasserstein optimal transport to align single-cell multi-omics data. *bioRxiv*, 2020. 44

[EG18] Ronen Eldan and Renan Gross. Exponential random graphs behave like mixtures of stochastic block models. *The Annals of Applied Probability*, 28(6):3698 – 3735, 2018. 144

[EH18] Jordan Emme and Pascal Hubert. Limit laws for random matrix products. *Mathematical Research Letters*, 25(4):1205–1212, 2018. 36, 59, 63, 168

[EK09] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*. John Wiley & Sons, USA, 2009. 217, 219

[ESSN22] Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks. In *International Conference on Learning Representations*, 2022. 18

[FDRC20] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear Mode Connectivity and the Lottery Ticket Hypothesis. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3259–3269. PMLR, 13–18 Jul 2020. 18

[FK60] Harry Furstenberg and Harry Kesten. Products of random matrices. *The Annals of Mathematical Statistics*, 31(2):457–469, 1960. 36, 168

[FK99] Alan Frieze and Ravi Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999. 42

[FKS89] J. Friedman, J. Kahn, and E. Szemerédi. On the second eigenvalue of random regular graphs. In *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*, STOC '89, page 587–598, New York, NY, USA, 1989. Association for Computing Machinery. 66

[Fur63] Harry Furstenberg. Noncommuting random products. *Transactions of the American Mathematical Society*, 108(3):377–428, 1963. 36, 168

[Fur02] Alex Furman. Random walks on groups and random transformations. In *Handbook of dynamical systems*, volume 1, pages 931–1014. Elsevier, 2002. 36

[Gär88] J. Gärtner. On the McKean-Vlasov limit for interacting diffusions. *Math. Nachr.*, 137:197–248, 1988. 13, 124

[GARA18] Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow gaussian processes. *arXiv preprint arXiv:1808.05587*, 2018. 167

[Grö19] Thomas Hakon Grönwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, pages 292–296, 1919. 210, 214, 219, 230, 263, 272, 275, 277

[Gul20] Talha Cihad Gulcu. Stronger convergence results for deep residual networks: network width scales linearly with training data size. *Information and Inference: A Journal of the IMA*, 11(2):497–532, 11 2020. 178

[GZ22] Borjan Geshkovski and Enrique Zuazua. Turnpike in optimal control of pdes, resnets, and beyond. *Acta Numerica*, 31:135–263, 2022. 178

[Han22] Boris Hanin. Random fully connected neural networks as perturbatively solvable hierarchies. *arXiv preprint arXiv:2204.01058*, 2022. 168

[HBM+22] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc., 2022. 167

[HLL83] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983. 30

[HN19] Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. *arXiv preprint arXiv:1909.05989*, 2019. 168

[HN20] Boris Hanin and Mihai Nica. Products of many large random matrices and

gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, 2020. 167, 168

[HNWW21] De Huang, Jonathan Niles-Weed, and Rachel Ward. Streaming k-pca: Efficient guarantees for oja's algorithm, beyond rank-one updates. In *Conference on Learning Theory*, pages 2463–2498. PMLR, 2021. 63

[Hoo79] D. N. Hoover. Relations on probability spaces and arrays of random variables, 1979. Preprint. Institute for Advanced Studies. 127

[Hoo82] David N Hoover. Row-column exchangeability and a generalized model for probability. *Exchangeability in probability and statistics (Rome, 1981)*, pages 281–291, 1982. 47, 73, 103, 127

[HOP+22] Zaid Harchaoui, Sewoong Oh, Soumik Pal, Raghav Somani, and Raghavendra Tripathi. Stochastic optimization on matrices and a graphon mckean-vlasov limit. arXiv preprint arXiv:2210.00422, 2022. 24, 51, 89, 130, 131, 177

[Huf77] R.E. Huff. The Radon-Nikodỳm property for Banach-spaces — a survey of geometric aspects. In Klaus-Dieter Bierstedt and Benno Fuchssteiner, editors, *Functional Analysis: Surveys and Recent Results*, volume 27 of *North-Holland Mathematics Studies*, pages 1–13. North-Holland, Germany, 1977. 89, 233

[Hun14] John K Hunter. Notes on partial differential equations. *Lecture Notes, https://www.math.ucdavis.edu/~hunter/pdes/pde_notes.pdf, Department of Mathematics, University of California*, 2014. 234

[HW20] Amelia Henriksen and Rachel Ward. Concentration inequalities for random matrix products. *Linear Algebra and its Applications*, 594:81–94, 2020. 36, 59, 168

[Hwa80] Chii-Ruey Hwang. Laplace's method revisited: Weak convergence of probability measures. *The Annals of Probability*, 8(6):1177–1182, 1980. 4, 10

[HY23a] Soufiane Hayou and Greg Yang. Width and depth limits commute in residual networks. In *International Conference on Machine Learning*, pages 12700–12723. PMLR, 2023. 167

[HY23b] Soufiane Hayou and Greg Yang. Width and depth limits commute in residual networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 12700–12723. PMLR, 23–29 Jul 2023. 178

[Jab14] Pierre-Emmanuel Jabin. A review of the mean field limits for vlasov equations. *Kinetic and Related Models*, 7(4):661–711, 2014. 13, 124

[Jan13] Svante Janson. Graphons, cut norm and distance, couplings and rearrangements. *NYJM Monographs*, 4, 2013. 83, 97, 145, 243

[Jan16] Svante Janson. Graphons and cut metric on sigma-finite measure spaces. arXiv preprint arXiv:1608.01833, 2016. 43

[JHJ+23] Samy Jelassi, Boris Hanin, Ziwei Ji, Sashank J. Reddi, Srinadh Bhojanapalli, and Sanjiv Kumar. Depth dependence of μp learning rates in relu mlps. *arXiv preprint arXiv:2305.07810*, 2023. 168

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The Variational Formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998. 100

[Joh94] Dudley Paul Johnson. Central limit theorems for random evolutions. *Stochastic Processes and their Applications*, 53(2):221–232, 1994. 36, 168

[Kac56] Mark Kac. Foundations of kinetic theory. In *Proceedings of The third Berkeley symposium on mathematical statistics and probability*, volume 3, pages 171–197, 1956. 13, 124

[Kal89] Olav Kallenberg. On the representation theorem for exchangeable arrays. *Journal of Multivariate Analysis*, 30(1):137–154, 1989. 47, 73, 103, 104, 117

[Kal05] O. Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Probability and Its Applications. Springer, New York, NY, 2005. 47, 104

[Kal21] O. Kallenberg. *Foundations of Modern Probability*. Probability Theory and Stochastic Modelling. Springer International Publishing, 2021. 258, 260, 263, 265

[KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 181

[KC12] Harold Joseph Kushner and Dean S. Clark. *Stochastic approximation methods for constrained and unconstrained systems*, volume 26. Springer Science & Business Media, 2012. 8, 28

[KH+09] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Toronto, ON, Canada, 2009. 140

[KH23] Tobit Klug and Reinhard Heckel. Scaling laws for deep learning based image reconstruction. In *The Eleventh International Conference on Learning Representations*, 2023. 167

[KK19] Dávid Kunszenti-Kovács. Uniqueness of banach space valued graphons. *Journal of Mathematical Analysis and Applications*, 474(1):413–440, 2019. 107

[KKLS14] Dávid Kunszenti-Kovács, László Lovász, and Balázs Szegedy. Multigraph limits, unbounded kernels, and banach space decorated graphs. *arXiv preprint arXiv:1406.7846*, 2014. 105, 114, 255

[KLRS07] Lukasz Kruk, John Lehoczky, Kavita Ramanan, and Steven Shreve. An explicit

formula for the Skorokhod map on [0, a]. *The Annals of Probability*, 35(5):1740 – 1768, 2007. 38, 143, 216, 261

[KMH+20] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 167

[KMS20] Tarun Kathuria, Satyaki Mukherjee, and Nikhil Srivastava. On concentration inequalities for random matrix products. *arXiv preprint arXiv:2003.06319*, 2020. 36, 59

[KMT75] János Komlós, Péter Major, and Gábor Tusnády. An approximation of partial sums of independent rv'-s, and the sample df. i. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32:111–131, 1975. 223

[KR18] Rajeeva L Karandikar and Bhamidi V Rao. *Introduction to stochastic calculus*. Springer, 2018. 228

[KRRS17] Richard Kenyon, Charles Radin, Kui Ren, and Lorenzo Sadun. Multipodal structure and phase transitions in large constrained graphs. *Journal of Statistical Physics*, 168:233–258, 2017. 126

[KS91] I. Karatzas and S. E. Shreve. *Brownian motion and stochastic calculus.*, volume 113 of *Graduate Texts in Mathematics*. Springer, second edition, 1991. 213, 261, 262, 263, 268, 274, 276

[KW52] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952. 28

[KY03] Harold Kushner and G. George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003. 8, 28

[Led01] François Ledrappier. Some asymptotic properties of random walks on free groups. In *CRM Proceedings and Lecture Notes*, pages 117–152. American Mathematical Society, 2001. 36

[Lin94] Ernest Lindelöf. Sur l'application de la méthode des approximations successives aux équations différentielles ordinaires du premier ordre. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 116(3):454–457, 1894. 9, 71

[LLL+24] Xianming Li, Zongxi Li, Jing Li, Haoran Xie, and Qing Li. 2d matryoshka sentence embeddings. arXiv preprint arXiv:2402.14776, 2024. 179

[LNR21] Mufan Li, Mihai Nica, and Dan Roy. The future is log-gaussian: Resnets and their infinite-depth-and-width limit at initialization. *Advances in Neural Information Processing Systems*, 34:7852–7864, 2021. 167, 178

[Lov12] László Lovász. *Large Networks and Graph Limits*, volume 60 of *Colloquium publications*. American Mathematical Society, Providence, RI, 2012. iv, 6, 41, 42, 73, 74, 85, 91, 92, 94, 95, 98, 102, 112, 136, 164, 232, 233, 243, 244, 255, 258, 264, 270, 275, 277, 278

[LP17] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Society, Providence, RI, 2017. 32

[LRW19] Daniel Lacker, Kavita Ramanan, and Ruoyu Wu. Local weak convergence for sparse networks of interacting processes. *arXiv preprint arXiv:1904.02585*, 2019. 126

[LS06] László Lovász and Balázs Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957, 2006. 38, 42, 73, 251, 252, 253

[LS07] László Lovász and Balázs Szegedy. Szemerédi's lemma for the analyst. *Geometric And Functional Analysis*, 17:252–270, 2007. 74, 236

[LS10] László Lovász and Balázs Szegedy. Limits of compact decorated graphs. arXiv preprint arXiv:1010.5155, 2010. 104, 105, 106, 107, 115, 255

[LS21] Qianyi Li and Haim Sompolinsky. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Physical Review X*, 11(3):031059, 2021. 168

[LTE19] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic Modified Equations and Dynamics of Stochastic Gradient Algorithms I: Mathematical Foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019. 29, 178

[LWLZ18] Chris Junchi Li, Mengdi Wang, Han Liu, and Tong Zhang. Near-optimal stochastic approximation for online principal component estimation. *Mathematical Programming*, 167:75–97, 2018. 63

[LZ17] László Miklós Lovász and Yufei Zhao. On derivatives of graphon parameters. *Journal of Combinatorial Theory Series A*, 145(C):364–368, January 2017. 73

[Man07] Willem Mantel. Problem 28. *Wiskundige Opgaven*, 10(2):60–61, 1907. 100, 152

[MB11] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. 8, 28

[McC97] Robert J. McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997. 99

[McK75] H. P. McKean. Fluctuations in the kinetic theory of gases. *Communications on Pure and Applied Mathematics*, 28(4):435–455, 1975. 13, 124

[Mém11] Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Found. Comput. Math*, 11:417–487, 2011. 43

[MLPA22] Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and scaling rules for adaptive gradient algorithms. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 51, 181

[MMM19] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2388–2464, 2019. 3, 14, 72, 177

[MRR+53] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953. 1

[Mun00] James R Munkres. *Topology*. Prentice Hall Upper Saddle River, NJ, 2000. 233, 266

[Nes83] Yurii Nesterov. A method of solving a convex programming problem with convergence rate o (1/k** 2). *Doklady Akademii Nauk SSSR*, 269(3):543, 1983. 181

[Net19] Praneeth Netrapalli. Stochastic gradient descent and its variants in machine learning. *Journal of the Indian Institute of Science*, 99(2):201–213, 2019. 1, 67

[NLL+24] Lorenzo Noci, Chuning Li, Mufan Li, Bobby He, Thomas Hofmann, Chris J Maddison, and Dan Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. *Advances in Neural Information Processing Systems*, 36, 2024. 167

[NP20] Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks. arXiv preprint arXiv:2001.11443, 2020. 3, 14, 72, 177

[NRS20] J. Neeman, C. Radin, and L. Sadun. Phase transitions in finite random networks. *Journal of Statistical Physics*, 181:305–328, 2020. 144

[NRS23a] J. Neeman, C. Radin, and L. Sadun. Typical large graphs with given edge and triangle densities. *PTRF*, 186:1167–1223, 2023. 144

[NRS23b] Joe Neeman, Charles Radin, and Lorenzo Sadun. Typical large graphs with given edge and triangle densities. *Probability Theory and Related Fields*, pages 1–57, 2023. 126, 137, 138

[Nua06] D. Nualart. *The Malliavin Calculus and Related Topics*. Probability and Its Applications. Springer Berlin Heidelberg, 2006. 180, 182

[NXL+18] Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Greg Yang, Jiri Hron, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. *arXiv preprint arXiv:1810.05148*, 2018. 167

[Oja82] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15:267–273, 1982. 1, 63

[OPST23] Sewoong Oh, Soumik Pal, Raghav Somani, and Raghavendra Tripathi. Gradient flows on graphons: Existence, convergence, continuity equations. *Journal of Theoretical Probability*, Jul 2023. 24, 78, 80, 84, 87, 89

[OR19] Roberto I Oliveira and Guilherme H Reis. Interacting diffusions on random graphs with diverging average degrees: Hydrodynamics and large deviations. *Journal of Statistical Physics*, 176(5):1057–1087, 2019. 126

[ORS20] Roberto I Oliveira, Guilherme H Reis, and Lucas M Stolerman. Interacting diffusions on sparse graphs: hydrodynamics from local weak limits. *Electronic Journal of Probability*, 25:1–35, 2020. 126

[PBSP18] Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):5–31, 2018. 141

[PC21] Geoff Pleiss and John P Cunningham. The limitations of large width in neural networks: A deep gaussian process perspective. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3349–3363. Curran Associates, Inc., 2021. 167

[Pet11] Valentin V Petrov. *Sums of independent random variables*, volume 82. Springer, Berlin, Heidelberg, 2011. 227

[PR17] Oleg Pikhurko and Alexander Razborov. Asymptotic structure of graphs with the minimum number of triangles. *Combinatorics, Probability and Computing*, 26(1):138–160, 2017. 126

[Raz08] Alexander A Razborov. On the minimal density of triangles in graphs. *Combinatorics, Probability and Computing*, 17(4):603–618, 2008. 126

[Ric10] Matthew Richey. The evolution of markov chain monte carlo methods. *The American Mathematical Monthly*, 117(5):pp. 383–413, 2010. 30

[RJPLH15] Carmona René, Fouque Jean-Pierre, and Sun Li-Hsien. Mean field games and systemic risk. *Communications in Mathematical Sciences*, 13(4):911–933, 2015. 69

[RM51] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. 1, 28

[RRS14] C. Radin, K. Ren, and L. Sadun. The asymptotics of large constrained graphs. *Journal of Physics A; Mathematical and Theoretical*, 47(17), 2014. 144

[RS23] Charles Radin and Lorenzo Sadun. Optimal graphons in the edge-2star model. arXiv preprint arXiv:2305.00333, 2023. 144

[RV13] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. arXiv preprint arXiv:1306.2872, 2013. 226

[RVE18] Grant M Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7146–7155, 2018. 3, 14, 72

[RY04] D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2004. 132, 158, 267, 279

[SABP22] Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pages 3515–3530. PMLR, 2022. 14

[San15] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Springer International Publishing, Cham, 2015. 2, 5, 21, 42, 70, 72, 77, 88, 96, 233, 236

[San17] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017. 72

[SBC16] Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016. 181

[SCF+21] Cristopher Salvi, Thomas Cass, James Foster, Terry Lyons, and Weixin Yang. The signature kernel is the solution of a goursat pde. *SIAM Journal on Mathematics of Data Science*, 3(3):873–899, 2021. 167

[SEVM+22] Kai Segadlo, Bastian Epping, Alexander Van Meegen, David Dahmen, Michael Krämer, and Moritz Helias. Unified field theoretical approach to deep and recurrent neuronal networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(10):103401, 2022. 168

[SFG+22] Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. *Advances in Neural Information Processing Systems*, 35:17456–17472, 2022. 179

[SLJ+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 179

[Sło94] Leszek Słomiński. On approximation of solutions of multidimensional SDE's with reflecting boundary conditions. *Stochastic processes and their Applications*, 50(2):197–219, 1994. 209

[Sło01] Leszek Słomiński. Euler's approximations of solutions of SDEs with reflecting boundary. *Stochastic processes and their applications*, 94(2):317–337, 2001. 214, 265

[SMN18] Mei Song, Andrea Montanari, and P Nguyen. A mean field view of the landscape of two-layers neural networks. *Proceedings of the National Academy of Sciences*, 115:E7665–E7671, 2018. 3, 14, 72, 177

[SS20a] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural

networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020. 3, 14, 72

[SS20b] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020. 3, 14, 72

[STH+24] Raghav Somani, Raghav Tripathi, Zaid Harchaoui, Sewoong Oh, and Soumik Pal. Scaling limits of large linear residual networks. *NeurIPS 2024*, 2024. Under review. 24, 62, 159, 163, 175

[Szn84] Alain-Sol Sznitman. Nonlinear reflecting diffusion process, and the propagation of chaos and fluctuations associated. *Journal of Functional Analysis*, 56(3):311–336, 1984. 124

[Szn91] Alain-Sol Sznitman. Topics in propagation of chaos. In Paul-Louis Hennequin, editor, *Ecole d'Eté de Probabilités de Saint-Flour XIX — 1989*, pages 165–251, Berlin, Heidelberg, 1991. Springer Berlin Heidelberg. 2, 13, 124, 168

[Tan79] H. Tanaka. Probabilistic treatment of the Boltzmann equation of Maxwellian molecules. *Z. Wahrsch. Verw. Gebiete*, 46(1):67–105, 1978/79. 13, 124

[TMK16] Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 2464–2469. IEEE, 2016. 179

[TN13] Behrouz Touri and Angelia Nedić. Product of random stochastic matrices. *IEEE Transactions on Automatic Control*, 59(2):437–448, 2013. 36, 168

[TR20] Belinda Tzen and Maxim Raginsky. A mean-field theory of lazy training in two-layer neural nets: entropic regularization and controlled McKean-Vlasov dynamics. arXiv preprint arXiv:2002.01987, 2020. 3, 14, 72

[Tut65] VN Tutubalin. On limit theorems for the product of random matrices. *Theory of Probability & Its Applications*, 10(1):15–27, 1965. 36, 168

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018. 30

[Vil03] Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003. 2, 5, 21, 72, 111, 232

[Vil12] C. Villani. Optimal transportation, dissipative PDE's and functional inequalities. Unpublished lecture notes. Accessed from `https://cedricvillani.org/sites/dev/files/old_images/2012/08/B04.MFranca.pdf`, 2012. 2, 13, 124

[vLA87] P. J. M. van Laarhoven and E. H. L. Aarts. *Simulated annealing: theory and applications*, volume 37 of *Mathematics and its Applications*. D. Reidel Publishing Co., Dordrecht, 1987. 4, 10

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 183

[Wat84] Joseph C Watkins. A central limit problem in random evolutions. *The Annals of Probability*, pages 480–513, 1984. 36, 168

[WE19] Qianxiao Li Weinan E, Jiequn Han. A mean-field optimal control formulation of deep learning. *Research in the Mathematical Sciences*, 6(1):1–41, 2019. 179

[WHL18] E Weinan, Jiequn Han, and Qianxiao Li. A mean-field optimal control formulation of deep learning. *arXiv preprint arXiv:1807.01083*, 2018. 179

[WWJ16] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016. 181

[XRV17] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017. 140

[Yai20] Sho Yaida. Non-gaussian processes and neural networks at finite widths. In *Mathematical and Scientific Machine Learning*, pages 165–192. PMLR, 2020. 168

[YYZH24] Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Feature learning in infinite depth neural networks. In *The Twelfth International Conference on Learning Representations*, 2024. 19, 167

[ZKHB22] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022. 167

[ZVCRP21] Jacob Zavatone-Veth, Abdulkadir Canatar, Ben Ruben, and Cengiz Pehlevan. Asymptotics of representation learning in finite bayesian neural networks. *Advances in neural information processing systems*, 34:24765–24777, 2021. 168

Appendix A

PROOFS OF THEOREMS IN CHAPTER 3

In this section, we will provide the proofs of theorems in Chapter 3.

A.1 Proofs of Chapter 3.1

In this Section, we will provide a proof of Theorem 3.1.

Recall the projected noisy SGD iterates defined in Definition 2.2, starting from $X_{n,0} \in \mathcal{M}_n$, rewritten for convenience:

$$X_{n,k+1} = P\Big(X_{n,k} - n^2\tau_{n,k}\nabla R_n(X_{n,k}) - \tau_{n,k}\Delta M_{n,k} + \tau_{n,k}^{1/2}G_{n,k}\Big), \tag{PNSGD}$$

for $k \in \mathbb{R}_+$, where $(G_{n,k})_{k\in\mathbb{Z}_+}$ is any $n \times n$ real symmetric matrix valued martingale difference sequence with each element containing centered and independent entries up to matrix symmetry, as defined in Chapter 2.2.1, and

$$\Delta M_{n,k} \coloneqq n^2 g_n(X_{n,k};\xi_{k+1}) - n^2\nabla R_n(X_{n,k}), \qquad k \in \mathbb{Z}_+.$$

Observe that $(\Delta M_{n,k})_{k\in\mathbb{Z}_+}$ is an $n\times n$ symmetric matrix valued martingale difference sequence with respect to the filtration $(\mathcal{F}_k)_{k\in\mathbb{Z}_+}$ where $\mathcal{F}_k \coloneqq \sigma\big(\{X_{n,0},\xi_{i+1},G_{n,i}\}_{i\in\{0\}\cup[k-1]}\cup\{\xi_{k+1}\}\big)$ for $k \in \mathbb{Z}_+$. Without the martingale difference term $\tau_{n,k}\Delta M_{n,k}$, equation (PNSGD) reduces to the projected GD iterates with additive noise, $(Y_{n,k})_{k\in\mathbb{Z}_+}$ starting at $Y_{n,0} = X_{n,0}$, described in (PNGD), re-written below

$$Y_{n,k+1} = P\Big(Y_{n,k} - n^2\tau_{n,k}\nabla R_n(Y_{n,k}) + \tau_{n,k}^{1/2}G_{n,k}\Big), \qquad k \in \mathbb{Z}_+. \tag{PNGD}$$

Let $w_k^{(n)} \coloneqq K(X_{n,k})$ and $v_k^{(n)} \coloneqq K(Y_{n,k})$ for all $k \in \mathbb{Z}_+$, and let $w^{(n)}$ and $v^{(n)}$ be piecewise constant interpolations of $\big(w_k^{(n)}\big)_{k\in\mathbb{Z}_+}$ and $\big(v_k^{(n)}\big)_{k\in\mathbb{Z}_+}$ respectively with the step size sequence

$\boldsymbol{\tau}_n$. Using Grönwall's inequality and an obvious coupling between the processes (PNSGD) and (PNGD), we show in Lemma A.1 that the two processes are close as $|\boldsymbol{\tau}_n| \to 0$.

Lemma A.1. *Let $R\colon \mathcal{W} \to \mathbb{R}$ be such that the Fréchet-like derivative $\phi = DR$ exists. Suppose Assumptions 3.1, and 3.2 hold. Let $n \in \mathbb{N}$. Let X_n and Y_n be the piecewise constant interpolations (see Definition 2.5) of $(X_{n,k})_{k\in\mathbb{Z}_+}$ and $(Y_{n,k})_{k\in\mathbb{Z}_+}$ respectively, as defined in* (PNSGD) *and* (PNGD)*, with step size sequence $\boldsymbol{\tau}_n \coloneqq (\tau_{n,k})_{k\in\mathbb{Z}_+}$. Then, there exists a universal constant $C > 0$ such that for any $T > 0$ we have*

$$\mathbb{E}\left[\sup_{s\in[0,T]} \left\|w^{(n)}(s) - v^{(n)}(s)\right\|_2^2\right] \leq C\sigma^2 T|\boldsymbol{\tau}_n| \exp\left[C\kappa_2^2 T^2\right].$$

Proof. Let X_n and Y_n be the piecewise constant interpolations of $(X_{n,j})_{j\in\mathbb{Z}_+}$ and $(Y_{n,j})_{j\in\mathbb{Z}_+}$ respectively as defined in Definition 2.5. Define $\Delta\colon \mathbb{R}_+ \to \mathbb{R}_+$ as

$$\Delta(t) \coloneqq \mathbb{E}\left[\sup_{s\in[0,t]} \|X_n(s) - Y_n(s)\|_{\mathrm{F}}^2\right], \qquad t \in \mathbb{R}_+. \tag{A.1}$$

Let $k \in \mathbb{Z}_+$ be such that $t \in [t_{n,k}, t_{n,k+1})$. Then, using [Sło94, Theorem 1],

$$\begin{aligned}\Delta(t) \leq C\,\mathbb{E}\left[\left(\sum_{j=0}^{k-1} \tau_{n,j}\left\|n^2\nabla R_n(X_{n,j}) - n^2\nabla R_n(Y_{n,j})\right\|_{\mathrm{F}}\right)^2\right] \\ + C\,\mathbb{E}\left[\sum_{j=0}^{k-1} \tau_{n,j}^2\|\Delta M_{n,j}\|_{\mathrm{F}}^2\right],\end{aligned} \tag{A.2}$$

where $C > 0$ is some universal constant. From Assumption 3.1, since ϕ is κ_2-Lipschitz as a map from $L^2\big([0,1]^{(2)}\big)$ to $L^2\big([0,1]^{(2)}\big)$, following Lemma 2.14 and the fact that $\|A_n\|_{\mathrm{F}}^2 = n^2\|K(A_n)\|_2^2$ for all $A_n \in \mathcal{M}_n$, we see that the map $\nabla R_n\colon \mathcal{M}_n \to \mathbb{R}^{[n]^2}$ satisfies

$$\left\|n^2\nabla R_n(A_n) - n^2\nabla R_n(B_n)\right\|_{\mathrm{F}}^2 \leq \kappa_2^2\|A_n - B_n\|_{\mathrm{F}}^2, \qquad \forall\ A_n, B_n \in \mathcal{M}_n. \tag{A.3}$$

Using the Cauchy-Schwarz inequality, and equation (A.3), we first bound the second term in equation (A.2) as

$$\mathbb{E}\left[\left(\sum_{j=0}^{k-1} \tau_{n,j}\left\|n^2\nabla R_n(X_{n,j}) - n^2\nabla R_n(Y_{n,j})\right\|_{\mathrm{F}}\right)^2\right]$$

$$\leq \mathbb{E}\left[\sum_{j=0}^{k-1}\left(\tau_{n,j}^{1/2}\right)^2 \cdot \sum_{j=0}^{k-1}\tau_{n,j}\left\|n^2\nabla R_n(X_{n,j}) - n^2\nabla R_n(Y_{n,j})\right\|_{\mathrm{F}}^2\right]$$
$$\leq \kappa_2^2 t\,\mathbb{E}\left[\sum_{j=0}^{k-1}\tau_{n,j}\|X_{n,j} - Y_{n,j}\|_{\mathrm{F}}^2\right] \leq \kappa_2^2 t\int_0^t \Delta(s)\,\mathrm{d}s, \tag{A.4}$$

where the last inequality follows by observing that if $s \in [t_{n,j}, t_{n,j+1})$ for some $j \in \mathbb{Z}_+$, then

$$\mathbb{E}\big[\|X_n(s) - Y_n(s)\|_{\mathrm{F}}^2\big] = \mathbb{E}\big[\|X_{n,j} - Y_{n,j}\|_{\mathrm{F}}^2\big] \leq \Delta(s).$$

Using Assumption 3.2, first note that

$$\|\Delta M_{n,j}\|_{\mathrm{F}}^2 = \left\|n^2 g_n(X_{n,k};\xi_{k+1}) - n^2\nabla R_n(X_{n,k})\right\|_{\mathrm{F}}^2$$
$$= n^2\left\|K\big(n^2 g_n(X_{n,k};\xi_{k+1}) - n^2\nabla R_n(X_{n,k})\big)\right\|_2^2 \leq n^2\sigma^2. \tag{A.5}$$

We use the above to bound the first term in equation (A.2) as

$$\mathbb{E}\left[\sum_{j=0}^{k-1}\tau_{n,j}^2\|\Delta M_{n,j}\|_{\mathrm{F}}^2\right] \leq n^2\sigma^2 t|\boldsymbol{\tau}_n|, \tag{A.6}$$

where $|\boldsymbol{\tau}_n|$ is defined in Chapter 2.2.1 as $\sup_{j\in\mathbb{Z}_+}\tau_{n,j}$.

Plugging back (A.4) and (A.6) in equation (A.2) we get

$$\Delta(t) \leq Cn^2\sigma^2 t|\boldsymbol{\tau}_n| + C\kappa_2^2 t\int_0^t \Delta(s)\,\mathrm{d}s, \tag{A.7}$$

and applying Grönwall's inequality [Grö19], we obtain $\Delta(t) \leq Cn^2\sigma^2 t|\boldsymbol{\tau}_n|\exp[C\kappa_2^2 t^2]$. □

Our next step is to show that sequence of iterates defined in (PNGD) is close to the solution of the SDE (3.1) which we reproduce below

$$\begin{aligned}\mathrm{d}X_n(t) = -n^2\nabla R_n(X_n(t)) + \Sigma_n(X_n(t)) \circ \mathrm{d}B_n(t)\\ - \mathrm{d}L_n^+(t) + \mathrm{d}L_n^-(t),\end{aligned} \qquad t \in \mathbb{R}_+, \tag{RSDE}$$

where B_n is an $n \times n$ symmetric matrix valued process whose entries are independent Brownian motions up to matrix symmetry, and $X_n(0) = Y_{n,0} = X_{n,0} \in \mathcal{M}_n$. The tuple (X_n, L_n^+, L_n^-) solves the Skorokhod problem with respect to the set $\mathcal{M}_n$ (see Chapter 2.4).

In Lemma A.2 we compare (PNGD) with a discretization of the SDE (3.1). This is obtained by coupling the discrete noise in (PNGD) with the Brownian motion driving the SDE (RSDE). Combining these we conclude the convergence of (PNSGD) to the SDE (RSDE) as $|\boldsymbol{\tau}_n| \to 0$.

Lemma A.2. *Let $n \in \mathbb{N}$. Let B_n be an $n \times n$ symmetric matrix valued process whose coordinates are i.i.d. Brownian motion (up to matrix symmetry) defined on some probability space. Let X_n be the strong solution of SDE* (RSDE) *with initial condition $X_n(0) = X_{n,0}$ (see* (PNGD)*). Then, there exists a càdlàg process $\widetilde{Y}_n$ on $\mathcal{M}_n$, defined on the same probability space as B_n, such that it has the same law as Y_n, the piecewise constant interpolation (see Definition 2.5) of $\big(Y_{n,k}\big)_{k\in\mathbb{Z}_+}$ obtained from* (PNGD)*. Moreover, for any $T \in \mathbb{R}_+$,*

$$\lim_{|\boldsymbol{\tau}_n|\to 0} \mathbb{E}\left[\sup_{s\in[0,T]} \left\| K(X_n(s)) - K\Big(\widetilde{Y}_n(s)\Big)\right\|_2^2\right] = 0.$$

Proof. Let B_n be as given in the assumption and let X_n be the strong solution of the SDE (RSDE). Since the discrete noise in (PNGD) is Gaussian (see Assumption 3.3), there is an obvious way to couple it with the Brownian motion driving the SDE in (RSDE). Given B_n and the step size sequence $\boldsymbol{\tau}_n = (\tau_{n,k} > 0)_{k\in\mathbb{Z}_+}$, define the discrete time $n \times n$ symmetric matrix valued martingale difference sequence $\big(\widetilde{Z}_{n,k}\big)_{k\in\mathbb{Z}_+}$ as

$$\widetilde{Z}_{n,k} \coloneqq \tau_{n,k}^{-1/2}(B_n(t_{n,k+1}) - B_n(t_{n,k})), \qquad k \in \mathbb{Z}_+. \tag{A.8}$$

Note that the entries in $\widetilde{Z}_{n,k}$ are distributed as $N(0,1)$ up to matrix symmetry for every $k \in \mathbb{Z}_+$. Starting from $\widetilde{Y}_{n,0} = X_{n,0}$, we now define an auxiliary process $\big(\widetilde{Y}_{n,k}\big)_{k\in\mathbb{Z}_+}$, on the same probability space as B_n, iteratively as

$$\widetilde{Y}_{n,k+1} = P\Big(\widetilde{Y}_{n,k} - n^2\tau_{n,k}\nabla R_n\Big(\widetilde{Y}_{n,k}\Big) + \tau_{n,k}^{1/2}\Sigma_n\Big(\widetilde{Y}_{n,k}\Big) \circ \widetilde{Z}_{n,k}\Big), \qquad k \in \mathbb{Z}_+, \tag{A.9}$$

Following Assumption 3.3, $\widetilde{Y}_{n,k}$ has the same law as $Y_{n,k}$ for each $k \in \mathbb{Z}_+$. Let $\widetilde{Y}_n \colon \mathbb{R}_+ \to \mathcal{M}_n$ be piecewise constant interpolation of $\Big(\widetilde{Y}_{n,k}\Big)_{k\in\mathbb{Z}_+}$. The particular choice of $\big(\widetilde{Z}_{n,k}\big)_{k\in\mathbb{Z}_+}$ in

equation (A.8) allows us to couple $\widetilde{Y}_n$ with the strong solution of the SDE (3.1). Let $\widetilde{G}_{n,j} \coloneqq \Sigma_n\Big(\widetilde{Y}_{n,j}\Big) \circ \widetilde{Z}_{n,j}$ for all $j \in \mathbb{Z}_+$. The curve $\widetilde{Y}_n$ can be written as

$$\widetilde{Y}_n(t) = \widetilde{Y}_{n,0} - \sum_{j=0}^{k-1} n^2 \tau_{n,j} \nabla R_n\Big(\widetilde{Y}_{n,j}\Big) + \sum_{j=0}^{k-1} \tau_{n,j}^{1/2} \widetilde{G}_{n,j} + \sum_{j=0}^{k-1} \tau_{n,j}\big(L_{n,j}^- - L_{n,j}^+\big), \tag{A.10}$$

for $t \in [t_{n,k}, t_{n,k+1})$. Here $\big(L_{n,j}^{\pm}\big)_{j\in\mathbb{Z}_+}$ is chosen so that the piecewise constant interpolation (see Definition 2.5) of $\big(Y_{n,k}, L_{n,k}^-, L_{n,k}^+\big)_{k\in\mathbb{Z}_+}$ solves the Skorokhod problem with respect to $\mathcal{M}_n$ (see Chapter 2.4).

Also consider three auxiliary processes Q_n, $\overline{Q}_n$, and $\widehat{Q}_n$ taking values over $n \times n$ real symmetric matrices, defined as

$$Q_n(t) \coloneqq X_n(0) - \int_0^t n^2 \nabla R_n(X_n(s))\,\mathrm{d}s + \int_0^t \Sigma_n(X_n(s)) \circ \mathrm{d}B_n(s), \tag{A.11}$$

$$\widehat{Q}_n(t) \coloneqq X_n(0) - \int_0^t n^2 \nabla R_n\Big(\widetilde{Y}_n(s)\Big)\,\mathrm{d}s + \int_0^t \Sigma_n\Big(\widetilde{Y}_n(s)\Big) \circ \mathrm{d}B_n(s), \tag{A.12}$$

$$\overline{Q}_n(t) \coloneqq X_n(0) - \sum_{j=0}^{k-1} n^2 \tau_{n,j} \nabla R_n\Big(\widetilde{Y}_{n,j}\Big) + \sum_{j=0}^{k-1} \tau_{n,j}^{1/2} \widetilde{G}_{n,j}, \tag{A.13}$$

for every $k \in \mathbb{Z}_+$ and all $t \in [t_{n,k}, t_{n,k+1})$. Observe that the curves X_n and $\widetilde{Y}_n$ can be obtained by applying the Skorokhod map to the curves Q_n and $\overline{Q}_n$ pointwise respectively. Let $\widehat{Y}_n \colon \mathbb{R}_+ \to \mathcal{M}_n$ be obtained from $\widehat{Y}_n$ by applying the Skorokhod map. First observe that using the Lipschitzness of the Skorokhod map, ϕ and Σ_n (see Assumption 3.1, Assumption 3.3, Chapter 2.4 and equation (A.3)), we obtain

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in[0,T]} \left\|\widehat{Y}_n(t) - X_n(t)\right\|_{\mathrm{F}}^2\right] &\le 16\,\mathbb{E}\left[\sup_{t\in[0,T]} \left\|\widehat{Q}_n(t) - Q_n(t)\right\|_{\mathrm{F}}^2\right] \\
&\le\ 16\,\mathbb{E}\left[\sup_{t\in[0,T]} \left\|\int_0^t n^2 \nabla R_n(X_n(s)) - n^2 \nabla R_n\Big(\widetilde{Y}_n(s)\Big)\,\mathrm{d}s\right\|_{\mathrm{F}}^2\right] \\
&\qquad + 16\,\mathbb{E}\left[\sup_{t\in[0,T]} \left\|\int_0^t \Big(\Sigma_n(X_n(s)) - \Sigma_n\Big(\widetilde{Y}_n(s)\Big)\Big) \circ \mathrm{d}B_n(s)\right\|_{\mathrm{F}}^2\right] \\
&\le 16\kappa_2^2\,\mathbb{E}\left[\int_0^T \left\|X_n(s) - \widetilde{Y}_n(s)\right\|_{\mathrm{F}}^2 \mathrm{d}s\right]
\end{aligned}$$

$$+ 64\,\mathbb{E}\left[\int_0^T \left\|\Sigma_n(X_n(s)) - \Sigma_n\left(\widetilde{Y}_n(s)\right)\right\|_{\mathrm{F}}^2 \mathrm{d}s\right]$$
$$\leq 80\kappa_2^2 \int_0^T \mathbb{E}\left[\sup_{s\in[0,t]}\left\|X_n(s) - \widetilde{Y}_n(s)\right\|_{\mathrm{F}}^2\right] \mathrm{d}s, \tag{A.14}$$

where the second last inequality follows from Doob's maximal inequality [KS91, page 14, Theorem 3.8.iv] and the fact that for all $A_n \in \mathcal{M}_n$, $\|A_n\|_{\mathrm{F}}^2 = n^2\|K(A_n)\|_2^2$. For any $t \in [0,T]$, define $k_t \coloneqq \arg\min_{j\in\mathbb{Z}_+}\{t \geq t_{n,j}\}$. Using the Lipschitzness of Skorokhod map (see Chapter 2.4) we obtain

$$\mathbb{E}\left[\sup_{s\in[0,T]}\left\|\widetilde{Y}_n(t) - \widehat{Y}_n(t)\right\|_{\mathrm{F}}^2\right] \leq 16\,\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\overline{Q}_n(t) - \widehat{Q}_n(t)\right\|_{\mathrm{F}}^2\right]$$
$$\leq 32\,\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\int_0^t n^2\nabla R_n\left(\widetilde{Y}_n(s)\right)\mathrm{d}s - \sum_{j=0}^{k_t-1} n^2\tau_{n,j}\nabla R_n\left(\widetilde{Y}_{n,j}\right)\right\|_{\mathrm{F}}^2\right]$$
$$+ 32\,\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\sum_{j=0}^{k_t-1}\tau_{n,j}^{1/2}\Sigma_n\left(\widetilde{Y}_{n,j}\right)\circ\widetilde{Z}_{n,j} - \int_0^t \Sigma_n\left(\widetilde{Y}_n(s)\right)\circ \mathrm{d}B_n(s)\right\|_{\mathrm{F}}^2\right], \tag{A.15}$$

where the last inequality follows from Assumption 3.3.

We now bound the first term from the above inequality (A.15). To this end observe that

$$\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\int_0^t n^2\nabla R_n\left(\widetilde{Y}_n(s)\right)\mathrm{d}s - \sum_{j=0}^{k_t-1} n^2\tau_{n,j}\nabla R_n\left(\widetilde{Y}_{n,j}\right)\right\|_{\mathrm{F}}^2\right]$$
$$= \mathbb{E}\left[\sup_{t\in[0,T]}\left\|n^2(t-t_{n,k_t})\nabla R_n\left(\widetilde{Y}_{n,k}\right)\right\|_{\mathrm{F}}^2\right] \leq |\boldsymbol{\tau}_n|^2\,\mathbb{E}\left[\sup_{t\in[0,T]}\left\|n^2\nabla R_n\left(\widetilde{Y}_{n,k}\right)\right\|_{\mathrm{F}}^2\right]$$
$$= n^2|\boldsymbol{\tau}_n|^2\,\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\phi\left(\widetilde{Y}^{(n)}(t)\right)\right\|_2^2\right] \leq n^2|\boldsymbol{\tau}_n|^2 M_2^2, \tag{A.16}$$

for some constant $M_2 \in \mathbb{R}_+$ by Assumption 3.1.

We now bound the second term in the inequality (A.15). Using the coupling defined in (A.8) and noting that $\widetilde{Y}(s) = \widetilde{Y}_{n,j}$ for $s \in [t_{n,j}, t_{n,j+1})$ (see Definition 2.5), we obtain that

$$\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\sum_{j=0}^{k_t-1}\tau_{n,j}^{1/2}\Sigma_n\left(\widetilde{Y}_{n,j}\right)\circ\widetilde{Z}_{n,j} - \int_0^t \Sigma_n\left(\widetilde{Y}_n(s)\right)\circ \mathrm{d}B_n(s)\right\|_{\mathrm{F}}^2\right]$$
$$= \mathbb{E}\left[\sup_{t\in[0,T]}\left\|\Sigma_n\left(\widetilde{Y}_{n,k_t}\right)\circ\left(B_n(t) - B_n(t_{n,k_t})\right)\right\|_{\mathrm{F}}^2\right] \leq M_\infty^2 n^2 C_{1,T}|\boldsymbol{\tau}_n|\log\frac{1}{|\boldsymbol{\tau}_n|}, \tag{A.17}$$

where the last inequality follows from Assumption 3.3 and [Sło01, Lemma A.4] for $C_{1,T} \in \mathbb{R}_+$.

Now define $\Delta\colon \mathbb{R}_+ \to \mathbb{R}_+$ as

$$\Delta(t) \coloneqq \mathbb{E}\left[\sup_{s\in[0,t]} \left\| X_n(s) - \widetilde{Y}_n(s) \right\|_{\mathrm{F}}^2\right], \qquad t \in \mathbb{R}_+.$$

Using the triangle inequality by combining equations (A.14), (A.15), (A.16) and (A.17), we get

$$\Delta(T) \leq 32n^2|\boldsymbol{\tau}_n|^2 M_2^2 + 32n^2 M_\infty^2 C_{1,T}|\boldsymbol{\tau}_n| \log\frac{1}{|\boldsymbol{\tau}_n|} + 80\kappa_2^2 \int_0^T \Delta(t)\,\mathrm{d}t. \tag{A.18}$$

Applying Grönwall's inequality [Grö19], we get

$$\Delta(T) \leq 32n^2\left(|\boldsymbol{\tau}_n|^2 M_2^2 + M_\infty^2 C_{1,T}|\boldsymbol{\tau}_n| \log\frac{1}{|\boldsymbol{\tau}_n|}\right)\exp\bigl[80\kappa_2^2 T\bigr]. \tag{A.19}$$

Taking limit as $|\boldsymbol{\tau}_n| \to 0$ on the above bound, completes the proof. □

We combine Lemma A.1 and A.2 to conclude the proof of Theorem 3.1. Moreover, we also the following non-asymptotic error rate

$$\mathbb{E}\left[\sup_{s\in[0,T]} \left\| w^{(n)}(s) - K(X_n)(s) \right\|_2^2\right] \leq Cn^2(M + \sigma^2 T)|\boldsymbol{\tau}_n| \log\frac{1}{|\boldsymbol{\tau}_n|}\exp\bigl[C\kappa_2^2 T\bigr]$$

for some constants $C, M < \infty$.

A.1.1 Convergence of Projected SGD

In the absence of "large noise" (i.e., when $\Sigma_n \equiv 0$), the SDE (RSDE) reduces to the SDE

$$\mathrm{d}X_n(t) = -n^2 \nabla R_n(X_n(t))\,\mathrm{d}t + \mathrm{d}L_n^-(t) - \mathrm{d}L_n^+(t), \qquad X_n(0) = X_{n,0}, \tag{A.20}$$

As we describe in Chapter 3.1.1, it is show in Chapter 4 that if the solution of

$$\mathrm{d}X_n(t) = -n^2 \nabla R_n(X_n(t)) \circ \mathbb{1}\{G_n(X_n(t))\}\,\mathrm{d}t, \tag{A.21}$$

exists, where $G_n(A)$ is the subset of $[n]^2$ defined as

$$\begin{aligned} G_n(A) \coloneqq &\bigl\{(i,j) \in [n]^2 \bigm| |A(i,j)| < 1\bigr\} \\ &\cup \bigl\{(i,j) \in [n]^2 \bigm| A(i,j) = 1, \partial_{i,j} R_n(A) > 0\bigr\} \\ &\cup \bigl\{(i,j) \in [n]^2 \bigm| A(i,j) = -1, \partial_{i,j} R_n(A) < 0\bigr\}, \end{aligned} \tag{A.22}$$

for all $A \in \mathcal{M}_n$, then the solution X_n is a gradient flow on $\mathcal{M}_n$ in a suitable sense. In this section, we will argue that the solutions X_n of equation (A.20) and (A.21) are equal. To this end, we define processes $L_n^{\pm}$ as

$$\begin{aligned} L_n^+(t) &\coloneqq -\int_0^t n^2 \nabla R_n(X_n(s)) \circ \mathbb{1}_{\{X_n(s)=+1, \nabla R_n(X_n(s))<0\}}\{\}\,\mathrm{d}s, \\ L_n^-(t) &\coloneqq +\int_0^t n^2 \nabla R_n(X_n(s)) \circ \mathbb{1}_{\{X_n(s)=-1, \nabla R_n(X_n(s))>0\}}\{\}\,\mathrm{d}s, \end{aligned} \tag{A.23}$$

for $t \in \mathbb{R}_+$, and equation (A.21) can be rewritten as

$$\mathrm{d}X_n(t) = -n^2 \nabla R_n(X_n(t)) \circ \mathbb{1}_{G_n(X_n(t))}\{\} + \mathrm{d}L_n^-(t) - \mathrm{d}L_n^+(t), \tag{A.24}$$

and the processes L_n^+ and L_n^- satisfy the following conditions:

1. The processes X_n, L_n^+ and L_n^- are adapted processes.

2. The processes L_n^- and L_n^+ are non-decreasing processes.

3. For every $(i,j) \in [n]^2$,

$$\begin{aligned} \int_0^\infty \mathbb{1}\big\{X_{n,(i,j)}(t) > -1\big\}\,\mathrm{d}L^-_{n,(i,j)}(t) = 0, \quad \text{and} \\ \int_0^\infty \mathbb{1}\big\{X_{n,(i,j)}(t) < +1\big\}\,\mathrm{d}L^+_{n,(i,j)}(t) = 0. \end{aligned}$$

Following Chapter 2.4, these conditions ensure that the processes L_n^+ and L_n^- are unique and (X_n, L_n^+, L_n^-) solves the Skorokhod problem with respect to the set $\mathcal{M}_n$.

A.2 Proofs of Chapter 3.2

In this Section we will provide the proof of Theorem 3.4. The subsections under this section will contain several lemmas that are required for Theorem 3.4 as well as other arguments in Chapter 3.2.

The proof of Theorem 3.4 is long and requires several lemmas (see Appendix A.2). Therefore, we first give an outline of the proof before presenting the details. Fix $n \in \mathbb{N}$. For every

$k \in \mathbb{Z}_+$, let $\mathcal{F}_k^r$ be the sigma algebra generated by $\Big\{q_{n,\ell}^{(r)} \,\Big|\, \ell \in \{0\} \cup [k]\Big\}$. Let $t_{r,n} = \lfloor tn^4/\gamma_r \rfloor$ as defined earlier. For $i, j \in [n]$, notice that

$$q_{n,(i,j)}^{(r)}(t) - q_{n,(i,j)}^{(r)}(0) = \sum_{\ell=0}^{t_{r,n}-1} \mathbb{E}\Big[\Delta q_{n,\ell,(i,j)}^{(r)} \,\Big|\, \mathcal{F}_\ell^r\Big] \mathbb{1}_{(0,1)}\Big\{q_{n,\ell,(i,j)}^{(r)}\Big\} + \sum_{\ell=0}^{t_{r,n}-1} \Delta M_{n,\ell,(i,j)}^{(r)} + L_{n,(i,j)}^{(r,0)}(t) - L_{n,(i,j)}^{(r,1)}(t),$$

for every $t \in \mathbb{R}_+$, where $\Delta M_{n,\ell}^{(r)} = \Delta q_{n,\ell}^{(r)} - \mathbb{E}\Big[\Delta q_{n,\ell}^{(r)} \,\Big|\, \mathcal{F}_\ell^r\Big]$ for all $\ell \in \mathbb{Z}_+$ and

$$L_{n,(i,j)}^{(r,0)}(t) = \sum_{\ell=0}^{t_{r,n}-1} \mathbb{E}\Big[\Delta q_{n,\ell,(i,j)}^{(r)} \,\Big|\, \mathcal{F}_\ell^r\Big] \mathbb{1}_{\{0\}}\Big\{q_{n,\ell,(i,j)}^{(r)}\Big\},$$
$$L_{n,(i,j)}^{(r,1)}(t) = \sum_{\ell=0}^{t_{r,n}-1} \mathbb{E}\Big[\Delta q_{n,\ell,(i,j)}^{(r)} \,\Big|\, \mathcal{F}_\ell^r\Big] \mathbb{1}_{\{1\}}\Big\{q_{n,\ell,(i,j)}^{(r)}\Big\},$$

for $t \in \mathbb{R}_+$, where $L_{n,(i,j)}^{(r,0)}(t)$ is $\frac{1}{r^2}$ times the number of times the process $q_{n,(i,j)}^{(r)}$ visits $\{0\}$ before time t and similarly for $L_{n,(i,j)}^{(r,1)}(t)$. Note that $\Big(M_{n,k}^{(r)} \coloneqq \sum_{\ell=0}^{k-1} \Delta M_{n,\ell}^{(r)}\Big)_{k \in \mathbb{Z}_+}$ is a $\mathcal{M}_n$-valued martingale and we define a piecewise constant interpolation of this martingale process $M_n^{(r)}$ defined as $M_n^{(r)}(t) = M_{n,t_{r,n}}^{(r)}$ for $t \in \mathbb{R}_+$. Let Sko be the Skorokhod map (see Chapter 2.4), then for any $t \in \mathbb{R}_+$, and any $(i,j) \in [n]^{(2)}$,

$$q_{n,(i,j)}^{(r)}(t) = \mathrm{Sko}\left(q_{n,(i,j)}^{(r)}(0) + \sum_{\ell=0}^{t_{r,n}-1} \mathbb{E}\Big[\Delta q_{n,\ell,(i,j)}^{(r)} \,\Big|\, \mathcal{F}_\ell^r\Big] \mathbb{1}_{(0,1)}\Big\{q_{n,\ell,(i,j)}^{(r)}\Big\} + M_{n,(i,j)}^{(r)}(t)\right). \quad \text{(A.25)}$$

1. Since the Skorokhod map Sko is a 4-Lipschitz map [KLRS07], to show that $q_n^{(r)}(t)$ converges uniformly to $X_n(t)$ as $r \to \infty$, it is sufficient to show that

$$\sum_{\ell=0}^{t_{r,n}-1} \mathbb{E}\Big[\Delta q_{n,\ell,(i,j)}^{(r)} \,\Big|\, \mathcal{F}_\ell^r\Big] \mathbb{1}_{(0,1)}\Big\{q_{n,\ell,(i,j)}^{(r)}\Big\} \to \int_0^t b_{n,(i,j)}(X_n(s)) \mathbb{1}_{(0,1)}\big\{X_{n,(i,j)}(s)\big\}\,\mathrm{d}s,$$
$$\text{and} \quad M_n^{(r)}(t) \to \sigma B_n(t),$$

uniformly over compact time intervals as $r \to \infty$.

2. In Lemma A.5, we show that the quadratic variation of the martingale $M_n^{(r)}$ in the time interval $[0,t]$ converges to $t\sigma^2$ for every $t \in \mathbb{R}_+$ as $r \to \infty$. The key ingredient is the fact that a simple symmetric reflected random walk spends negligible amount of time at the boundary.

3. Using Lemma A.5 and [EK09, Theorem 1.4, Chapter 7] we conclude that the process $M_n^{(r)}$ converges to the process (weakly) σB_n where B_n is an $n \times n$ symmetric matrix with i.i.d. Brownian motions. Using Skorokhod representation theorem both $M_n^{(r)}$ and B_n can be defined on some common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, on which we get almost sure convergence.

4. On the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ obtained in Step 3, we define versions of the processes X_n and $q_n^{(r)}$ using (3.12) and (A.25) respectively.

5. It remains to show that the first condition in Step 1 holds. To this end, we first show that for every fixed $\epsilon > 0$, $\mathbb{E}\Big[\Delta q^{(r)}_{n,\ell,(i,j)} \,\Big|\, \mathcal{F}^r_\ell\Big] \mathbb{1}_{(\epsilon,1-\epsilon)}\Big\{q^{(r)}_{n,\ell,(i,j)}\Big\} - b_{n,(i,j)}\Big(q^{(r)}_{n,\ell}\Big)\mathbb{1}_{(\epsilon,1-\epsilon)}\Big\{q^{(r)}_{n,\ell,(i,j)}\Big\} \to 0$ as $r \to \infty$. This is achieved by a sequence of reductions in Lemma A.9.

6. Finally, we show that $\sum_{\ell=0}^{t_{r,n}-1} \mathbb{1}_{(\delta,1-\delta)^c}\Big\{q^{(r)}_{n,\ell,(i,j)}\Big\} \to 0$ as $r \to \infty$. Since $\frac{1}{\gamma_r}\mathbb{E}\Big[\Delta q^{(r)}_{n,\ell} \,\Big|\, \mathcal{F}_\ell\Big]$ is uniformly bounded. We conclude that the first condition in Step 1 holds. This completes the proof.

We are now ready to provide the formal proof of Theorem 3.4.

Proof of Theorem 3.4. Let $q_n^{(r)}$ be defined by (A.25). In the following we will keep n fixed and therefore drop it from the subscript wherever necessary. Throughout, we will keep $t \in \mathbb{R}_+$ fixed. The map Sko in the discussion will refer to the Skorokhod map defined on the space $D([0,t], \mathbb{R}^{[n]^{(2)}})$, the space of right continuous paths with left limits from $[0,t]$ to $\mathbb{R}^{[n]^{(2)}}$. Define

$$b^{(r)}_{n,(i,j)}\Big(q^{(r)}_{n,\ell}\Big) = \frac{1}{\gamma_r n^{-4}}\,\mathbb{E}\Big[\Delta q^{(r)}_{n,\ell,(i,j)} \,\Big|\, \mathcal{F}^r_\ell\Big]\mathbb{1}_{(0,1)}\Big\{q^{(r)}_{n,\ell,(i,j)}\Big\},$$

and let b_n be in Definition in 3.2. Recall that both $b_n^{(r)}$ and b_n are uniformly bounded by some constant C. Also, recall that B_n is an $n \times n$ symmetric matrix with i.i.d. Brownian coordinates. Define the stochastic processes $Y_n^{(r)}, \widetilde{Y}_n^{(r)}, \widetilde{Z}_n^{(r)}, Z_n^{(r)} \colon \mathbb{R}_+ \to \mathbb{R}^{[n]^{(2)}}$ as:

$$\begin{aligned}
Y_{n,(i,j)}^{(r)}(t) &= q_{n,(i,j)}^{(r)}(0) + \sum_{\ell=0}^{t_{r,n}-1} \gamma_r n^{-4} b_{n,(i,j)}^{(r)}\Big(q_{n,\ell}^{(r)}\Big) + M_{n,(i,j)}^{(r)}(t), \\
\widetilde{Y}_{n,(i,j)}^{(r)}(t) &= q_{n,(i,j)}^{(r)}(0) + \sum_{\ell=0}^{t_{r,n}-1} \gamma_r n^{-4} b_{n,(i,j)}^{(r)}\Big(q_{n,\ell}^{(r)}\Big) + \sigma B_{n,(i,j)}(t), \qquad t \in \mathbb{R}_+, (i,j) \in [n]^{(2)}. \\
\widetilde{Z}_{n,(i,j)}^{(r)}(t) &= q_{n,(i,j)}^{(r)}(0) + \int_0^t b_{n,(i,j)}^{(r)}\big(q_n^{(r)}(s)\big)\,\mathrm{d}s + \sigma B_{n,(i,j)}(t), \\
Z_{n,(i,j)}^{(r)}(t) &= q_{n,(i,j)}^{(r)}(0) + \int_0^t b_{n,(i,j)}\big(q_n^{(r)}(s)\big)\,\mathrm{d}s + \sigma B_{n,(i,j)}(t),
\end{aligned}$$

Notice that $Y_n^{(r)}$ is the "unconstrained version" of the process $q_n^{(r)}$ in the sense that $\mathrm{Sko}\Big(Y_n^{(r)}\Big)(s) = q_n^{(r)}(s)$ for every $s \in \mathbb{R}_+$. Finally, let $\Big(X_n, L_n^{(0)}, L_n^{(1)}\Big)$ be the process that satisfies the Skorokhod SDE

$$\mathrm{d}X_n(t) = b_n(X(t))\,\mathrm{d}t + \sigma\,\mathrm{d}B_n(t) + \mathrm{d}L_n^{(0)}(t) - \mathrm{d}L_n^{(1)}(t), \qquad X_{n,(i,j)}(0) = q_{n,(i,j)}^{(r)}(0).$$

Denote the corresponding unconstrained process

$$\widetilde{X}_n(t) = q_n^{(r)}(0) + \int_0^t b_n(X_n(s))\,\mathrm{d}s + \sigma\,\mathrm{d}B_n(t).$$

Note that $\mathrm{Sko}\Big(\widetilde{X}_n\Big) = X_n$. Recall that the goal is to show that $q_n^{(r)}$ converges to the process X_n as $r \to \infty$. Using the fact that the Skorokhod map is Lipschitz (see Chapter 2.4), it is sufficient to show that $Y_n^{(r)}$ converges to $\widetilde{X}_n$. To this end, set

$$\begin{aligned}
\Delta^{(r)}(t) &\coloneqq \mathbb{E}\left[\sup_{s\in[0,t]}\left\|Y_n^{(r)}(s) - \widetilde{X}_n(s)\right\|_{\mathrm{F}}^2\right], \qquad \Delta_1^{(r)}(t) \coloneqq \mathbb{E}\left[\sup_{s\in[0,t]}\left\|Y_n^{(r)}(s) - \widetilde{Y}_n^{(r)}(s)\right\|_{\mathrm{F}}^2\right], \\
\Delta_2^{(r)}(t) &\coloneqq \mathbb{E}\left[\sup_{s\in[0,t]}\left\|\widetilde{Z}_n^{(r)}(s) - \widetilde{Y}_n^{(r)}(s)\right\|_{\mathrm{F}}^2\right], \qquad \Delta_3^{(r)}(t) \coloneqq \mathbb{E}\left[\sup_{s\in[0,t]}\left\|\widetilde{Z}_n^{(r)}(s) - Z_n^{(r)}(s)\right\|_{\mathrm{F}}^2\right], \\
\Delta_4^{(r)}(t) &\coloneqq \mathbb{E}\left[\sup_{s\in[0,t]}\left\|\widetilde{X}_n^{(r)}(s) - Z_n^{(r)}(s)\right\|_{\mathrm{F}}^2\right], \qquad \text{for all } t \in \mathbb{R}_+.
\end{aligned}$$

Since $(a+b+c+d)^2 \le 4(a^2+b^2+c^2+d^2)$ for all $(a,b,c,d) \in \mathbb{R}^4$, $\Delta^{(r)}(t) \le 4\sum_{i=1}^4 \Delta_i^{(r)}(t) \le 4\sum_{i=1}^3 \Delta_i^{(r)}(t) + 64t\int_0^t \Delta^{(r)}(s)\,\mathrm{d}s$, where the final inequality is due to the 4-Lipschtizness of the Skorokhod map. In particular, for $t \in [0,T]$, we have $\Delta^{(r)}(t) \le 4\sum_{i=1}^3 \Delta_i^{(r)}(t) + 64T\int_0^t \Delta^{(r)}(s)\,\mathrm{d}s$. Note that $t \mapsto \Delta_i^{(r)}(t)$ is increasing. Therefore, using Grönwall's inequality [Grö19], we obtain

$$\Delta^{(r)}(T) \le 4\left(\sum_{i=1}^3 \Delta_i^{(r)}(T)\right)\exp\bigl(64T^2\bigr). \tag{A.26}$$

It is therefore sufficient to show that $\Delta_i^{(r)}(t) \to 0$ as $r \to \infty$ for $i \in [3]$. This is done in following steps. Using Lemma A.5 below and Theorem [EK09, Theorem 1.4, Chapter 7], we know that the process $M_n^{(r)}$ converges to σB_n uniformly on compact subsets of time. In particular, for fixed $t > 0$ we have that $\Delta_1^{(r)}(t) \to 0$ as $r \to \infty$. For $\Delta_2^{(r)}$, we notice that the error is actually the error from the Riemann sum approximation. Hence, $\Delta_2^{(r)}(t) \le C_n\gamma_r^2 \to 0$ as $r \to \infty$. To see this, first recall that $q^{(r)}(s)$ is piecewise constant on the interval of length $\gamma_r n^{-4}$, that is, $q_n^{(r)}(s) = q_{n,\lfloor s/(\gamma_r n^{-4})\rfloor}^{(r)}$. Now observe that

$$\begin{aligned}\left|\widetilde{Y}_{n,(i,j)}^{(r)}(t) - \widetilde{Z}_{n,(i,j)}^{(r)}(t)\right| &= \left|\sum_{\ell=0}^{t_{r,n}-1} \gamma_r n^{-4} b_{(i,j)}^{(r)}\Bigl(q_{n,\ell}^{(r)}\Bigr) - \int_0^t b_{(i,j)}^{(r)}\bigl(q_n^{(r)}(s)\bigr)\right| \\ &= \left|\bigl(t - \gamma_r n^{-4}(t_{r,n}-1)\bigr) b_{(i,j)}^{(r)}\Bigl(q_{n,t_{r,n}-1}^{(r)}\Bigr)\right| \le C\gamma_r n^{-4},\end{aligned}$$

where the inequality in the last line follows from the fact that $b^{(r)}$ is uniformly bounded. Squaring both sides and summing over all $(i,j) \in [n]^{(2)}$ we conclude that $\Delta_2^{(r)}(t) \le C_n\gamma_r^2$.

We now show that $\Delta_3^{(r)}(t) \to 0$ as $r \to \infty$. To do this, we fix $\epsilon > 0$, $\delta > 0$ (we assume that $\delta \ll \epsilon$ and $\delta + \epsilon \ll 1$). Define

$$A_\epsilon \coloneqq \bigl\{M \in \mathcal{M}_n \,\big|\, \epsilon \le M_{(i,j)} \le 1-\epsilon \quad \forall \quad (i,j) \in [n]^{(2)}\bigr\}, \quad B_\epsilon \coloneqq \mathcal{M}_n \setminus A_\epsilon. \tag{A.27}$$

Start observing that $\sup_{s\in[0,t]}\left\|\widetilde{Z}_n^{(r)}(s) - Z_n^{(r)}(s)\right\|_{\mathrm{F}}^2$ is at most

$$t\int_0^t \left\|b_n^{(r)}\bigl(q_n^{(r)}(s)\bigr) - b_n\bigl(q_n^{(r)}(s)\bigr)\right\|_{\mathrm{F}}^2 \mathrm{d}s$$

$$\leq t\int_0^t \left\|b_n^{(r)}\big(q_n^{(r)}(s)\big) - b_n\big(q_n^{(r)}(s)\big)\right\|_{\mathrm{F}}^2 \mathbb{1}_{A_\epsilon}\big\{q_n^{(r)}(s)\big\}\,\mathrm{d}s$$
$$+ t\int_0^t \left\|b_n^{(r)}\big(q_n^{(r)}(s)\big) - b_n\big(q_n^{(r)}(s)\big)\right\|_{\mathrm{F}}^2 \mathbb{1}_{B_\epsilon}\big\{q_n^{(r)}(s)\big\}\,\mathrm{d}s. \tag{A.28}$$

From Lemma A.9 below $b_n^{(r)}\Big(q_n^{(r)}(s)\Big)$ and $b_n\Big(q_n^{(r)}(s)\Big)$ are close in $\|\cdot\|_{\mathrm{F}}^2$ by $\frac{Cn^2}{r^4} + 4n^2\beta_{r,n}^2 e_r^3 \max\{|\lambda|, L\}$ with probability at least $1 - p_{r,\epsilon}$, when $q_n^{(r)}(s) \in A_\epsilon$ where $p_{r,\epsilon} = \frac{4n^2}{r^4} + 2n^2\mathrm{erf}\Big(\frac{n^2\epsilon}{4\sigma\sqrt{2\gamma_r}}\Big)$ and $e_r = O(\gamma_r^2 \log r)$. On the other hand, we notice that $b_n^{(r)}$ and b_n are both uniformly bounded by in $\|\cdot\|_\infty$. Using these two facts we conclude that

$$\mathbb{E}\bigg[\int_0^t \left\|b_n^{(r)}\big(q_n^{(r)}(s)\big) - b_n\big(q_n^{(r)}(s)\big)\right\|_{\mathrm{F}}^2 \mathbb{1}_{A_\epsilon}\big\{q_n^{(r)}(s)\big\}\,\mathrm{d}s\bigg]$$
$$\leq C\frac{t}{\gamma_r}p_{r,\epsilon} + \frac{Cn^2}{r^4} + 4n^2\beta_{r,n}^2 e_r^3 \max\{|\lambda|, L\}. \tag{A.29}$$

Since $b_n^{(r)}$ and b_n are uniformly bounded, the second term in (A.28) is bounded as

$$\int_0^t \left\|b_n^{(r)}\big(q_n^{(r)}(s)\big) - b_n\big(q_n^{(r)}(s)\big)\right\|_{\mathrm{F}}^2 \mathbb{1}_{B_\epsilon}\big\{q_n^{(r)}(s)\big\}\,\mathrm{d}s \leq CD^{(r)}(t),$$

for some constant $C > 0$, where $D^{(r)}(t) \coloneqq \int_0^t \mathbb{1}_{B_\epsilon}\Big\{q_n^{(r)}(s)\Big\}\,\mathrm{d}s$.

We now approximate the indicator function, $\mathbb{1}_{B_\epsilon}\{\cdot\}$ by a smooth function $\psi_{\epsilon,\delta} \in C^\infty([0,1])$. That is, Let $\psi_{\epsilon,\delta}$ be a smooth function such that $\psi_{\epsilon,\delta} \equiv 1$ on the set $I_\epsilon \coloneqq [0,\epsilon) \cup (1-\epsilon, 1]$ and $0 \leq \psi_{\epsilon,\delta} \leq 1$ and $\mathrm{supp}(\psi) \subset [0, \epsilon+\delta) \cup (1-\epsilon-\delta, 1]$. Recall that by our assumption $\epsilon + \delta \ll 1$. Then, $D^{(r)}(t) \leq \int_0^t \psi_{\epsilon,\delta}\Big(q_n^{(r)}(s)\Big)\,\mathrm{d}s$.

Recall that $q_n^{(r)}(s) = \mathrm{Sko}\big(Y^{(r)}\big)(s)$ for every $s \in \mathbb{R}_+$. Also recall that both $\psi_{\epsilon,\delta}$ and Sko are Lipschitz functions. Therefore, the composition $\psi_{\epsilon,\delta} \circ \mathrm{Sko}$ is is also a Lipschitz function, say, with Lipschitz constant $L_{\epsilon,\delta}$. Define $\Psi_{\epsilon,\delta}(s) \coloneqq \psi_{\epsilon,\delta}\Big(\mathrm{Sko}\Big(Y_n^{(r)}\Big)(s)\Big)$ and $\widetilde{\Psi}_{\epsilon,\delta}(s) \coloneqq \psi_{\epsilon,\delta}\Big(\mathrm{Sko}\Big(\widetilde{Z}_n^{(r)}\Big)(s)\Big)$. Now observe that

$$\Psi_{\epsilon,\delta}(s) \leq \widetilde{\Psi}_{\epsilon,\delta}(s) + L_{\epsilon,\delta}\left\|\widetilde{Z}_n^{(r)}(s) - Y_n^{(r)}(s)\right\|_{\mathrm{F}}^2$$
$$\leq \widetilde{\Psi}_{\epsilon,\delta}(s) + 2L_{\epsilon,\delta}\bigg(\left\|\widetilde{Z}_n^{(r)}(s) - \widetilde{Y}_n^{(r)}(s)\right\|_{\mathrm{F}}^2 + \left\|Y_n^{(r)}(s) - \widetilde{Y}_n^{(r)}(s)\right\|_{\mathrm{F}}^2\bigg).$$

Therefore, we obtain

$$\mathbb{E}\big[D^{(r)}(t)\big] \leq \mathbb{E}\bigg[\int_0^t \widetilde{\Psi}_{\epsilon,\delta}(s)\,\mathrm{d}s\bigg] + 2L_{\epsilon,\delta}t\Big(\Delta_1^{(r)}(t) + \Delta_2^{(r)}(t)\Big). \tag{A.30}$$

Note that $\mathbb{E}\Big[\int_0^t \widetilde{\Psi}_{\epsilon,\delta}(s)\,\mathrm{d}s\Big] = \mathbb{E}\Big[F_{\epsilon,\delta}\Big(\widetilde{Z}^{(r)}\Big)\Big]$ for some bounded continuous function $F_{\epsilon,\delta}\colon C([0,t],\mathcal{M}_{n,+}) \to \mathbb{R}$. Also, recall that $\widetilde{Z}^{(r)}$ satisfies the SDE $\widetilde{Z}_n^{(r)}(t) = q_n(0) + \int_0^t f(s)\,\mathrm{d}s + \sigma B_n(t)$, where $f\colon s \mapsto b_n^{(r)}\Big(q_n^{(r)}(s)\Big)$ is a bounded function. Set

$$\mathcal{E} = \exp\Big(\frac{1}{\sigma}\int_0^t b_n^{(r)}\big(q_n^{(r)}(s)\big)\,\mathrm{d}B_n(s) - \frac{1}{2\sigma^2}\int_0^t b_n^{(r)}\big(q_n^{(r)}(s)\big)\,\mathrm{d}s\Big).$$

Using Girsanov's theorem and the Cauchy–Schwarz inequality we obtain

$$\mathbb{E}\Big[F_{\epsilon,\delta}\Big(\widetilde{Z}_n^{(r)}\Big)\Big]^2 = \mathbb{E}[F_{\epsilon,\delta}(B)\mathcal{E}]^2 \le \mathbb{E}\big[F_{\epsilon,\delta}^2(B)\big]\,\mathbb{E}\big[\mathcal{E}^2\big].$$

Finally using the fact that $b_n^{(r)}$ is uniformly bounded, we obtain that $\mathbb{E}[\mathcal{E}^2] \le C_{n,t,\sigma}$. On the other hand, we notice that by definition

$$\mathbb{E}\big[F_{\epsilon,\delta}^2(B)\big] = \mathbb{E}\left[\left(\int_0^t \psi_{\epsilon,\delta}(\mathrm{RBM}(s))\,\mathrm{d}s\right)^2\right] \le t\int_0^t \mathbb{P}\{\mathrm{RBM}(s)\in I_\epsilon\}\,\mathrm{d}s =: C(\epsilon,\delta,n,t)^2, \tag{A.31}$$

where the equality follows from the Cauchy-Schwarz and the fact that $\psi_{\epsilon,\delta}^2 \le \mathbb{1}_{B_{\epsilon+\delta}}\{\cdot\}$. Combining equations (A.29), (A.30) and (A.31) we obtain that $\Delta_3^{(r)}(t) \le \frac{t}{\gamma_r}p_r + C(\epsilon,\delta,n,t) + 2L_{\epsilon,\delta}t\Big(\Delta_1^{(r)}(t) + \Delta_2^{(r)}(t)\Big)$.

Putting this back in equation (A.26) we conclude that

$$\Delta^{(r)}(t) \le \left(\frac{t}{\gamma_r}p_r + C(\epsilon,\delta,n,t) + (2L_{\epsilon,\delta}t + C)\Big(\Delta_1^{(r)}(t) + \Delta_2^{(r)}(t)\Big)\right)\mathrm{e}^{Ct^2}.$$

Recall that $\frac{p_r}{\gamma_r} \to 0$ and $\Delta_1^{(r)}(t) \to 0$ and $\Delta_2^{(r)}(t) \le \gamma_r^2 \to 0$ as $r \to \infty$. Therefore, we conclude that $\limsup_{r\to\infty}\Delta^{(r)}(t) \le C(\epsilon,\delta,n,t)\mathrm{e}^{Ct^2}$. Since $C(\epsilon,\delta,n,t) \to 0$ as $\epsilon,\delta \to 0$, this concludes the proof. □

A.2.1 Properties of drift function

Lemma A.3. *Let Y be a standard normal random variable on $\mathbb{R}^{[r]^{(2)}}$. For any $v \in \mathbb{R}^{[r]^{(2)}}$ and $t > 0$ we have*

$$\mathbb{E}_Y\big[Y\exp\big(-t\langle v, Y\rangle_{\mathrm{F}}^+\big)\big] = -2tv\exp\big(t^2\|v\|_{\mathrm{F}}^2\big)\overline{\Phi}(\sqrt{2}t\|v\|_{\mathrm{F}}).$$

Proof. Let Y be as above. Let $\pi\colon y \mapsto \langle v, y\rangle_{\mathrm{F}}$ and let $X = \pi(Y)$. Note that $X = \langle v, Y\rangle \sim \mathcal{N}(0, 2\|v\|_{\mathrm{F}}^2)$. The factor of 2 is due to symmetry. Observe that

$$\mathbb{E}\big[Y \exp\big(-t\langle v, Y\rangle_{\mathrm{F}}^{+}\big)\big] = \mathbb{E}_X\big[\exp\big(-tX^{+}\big)\mathbb{E}_Y[Y \mid \langle v, Y\rangle = X]\big]$$
$$= \frac{v}{\|v\|_{\mathrm{F}}^2}\,\mathbb{E}\big[X \exp\big(-tX^{+}\big)\big] = \frac{\sqrt{2}v}{\|v\|_{\mathrm{F}}}\,\mathbb{E}\Big[Z \exp\Big(-\sqrt{2}t\|v\|_{\mathrm{F}} Z^{+}\Big)\Big],$$

where $Z \sim N(0,1)$ is standard normal random variable. he proof follows by observing that $\mathbb{E}[Z\exp(-\alpha Z^{+})] = -\alpha \exp\big(\frac{1}{2}\alpha^2\big)\overline{\Phi}(\alpha)$ and taking $\alpha = \sqrt{2}t\|v\|_{\mathrm{F}}$. □

A.2.2 Quadratic variation of the Martingale

The proof of the following lemma follows from a standard argument using Donsker's invariance theorem and the Lipschitzness of Skorokhod map and is skipped.

Lemma A.4 (Time at boundary of reflected Random Walk)**.** *Let $\ell_r = \lceil n^{-4}\sigma^2\gamma_r r^4\rceil$. Fix $x \in \{i/r^2 \mid i = 0, \ldots, r^2\}$. Let S denote the symmetric random walk with step size $\frac{1}{r^2}$ reflected at $\{0,1\}$ starting at x. Then,*

$$\lim_{r\to\infty} \frac{1}{\gamma_r r^4} \sum_{k=1}^{\ell_r} \mathbb{1}_{\{0,1\}}\{S_k\} = 0, \quad \textit{in probability.}$$

We now compute the quadratic variation of the martingale $M_n^{(r)}(t)$ defined in Chapter 3.2.

Lemma A.5 (Martingale Quadratic Variation)**.** *For $n \in \mathbb{N}$, $r \in \mathbb{N}$ and $t \in \mathbb{R}_+$, let $M_n^{(r)}(t) \coloneqq \sum_{\ell=0}^{t_{r,n}-1} \Delta M_{n,\ell}^{(r)}$ where $t_{r,n} = \lfloor tn^4/\gamma_r\rfloor$. Then, the quadratic variation of $M_n^{(r)}$ in the time interval $[0,t]$ converges to $t\sigma^2 I_n$ for all $t \in \mathbb{R}_+$. That is, the following convergence holds in probability:*

$$\lim_{r\to\infty} \sum_{\ell=0}^{t_{r,n}-1} \mathbb{E}\Big[\Big(\Delta M_{n,\ell,(i,j)}^{(r)}\Big)\Big(\Delta M_{n,\ell,(i',j')}^{(r)}\Big) \,\Big|\, \mathcal{F}_\ell\Big] = t\sigma^2 \mathbb{1}\{i = i', j = j'\},$$

for all $(i,j),(i',j') \in [n]^{(2)}$.

Proof. We first notice that for each $k \in \mathbb{N}$ we have $\mathbb{E}\left[\left\|\widetilde{q}_{n,k+1}^{(r)} - q_{n,k}^{(r)}\right\|_2^2 \,\middle|\, \mathcal{F}_k\right] \le n^2\gamma_r^2$. Let $\mathcal{G}_k$ be the sigma algebra generated by $\mathcal{F}_k \vee \left\{p_{n,k+1}^{(r)}\right\}$. Recall that given $p_{n,k+1}^{(r)}$, the iterate $q_{n,k+1}^{(r)}$

is obtained by running ℓ_r steps of independent symmetric random walk with step size $\frac{1}{r^2}$ (with reflection at $\{0,1\}$) starting at $p^{(r)}_{n,k+1}$. Fix $(i,j) \in [n]^{(2)}$. Let S_k denote the symmetric random walk with step size $1/r^2$ run for m steps starting at $p^{(r)}_{n,k+1,(i,j)}$. Now observe that

$$\begin{aligned}\mathbb{E}\left[\left(q^{(r)}_{n,k+1,(i,j)} - p^{(r)}_{n,k+1,(i,j)}\right)^2 \,\middle|\, \mathcal{G}_k\right] &= \frac{1}{r^4}\sum_{m=1}^{\ell_r} \mathbb{1}_{(0,1)}\{S_{k,m}\} + \frac{1}{2r^4}\sum_{m=1}^{\ell_r} \mathbb{1}_{\{0,1\}}\{S_{k,m}\} \\ &= \frac{\ell_r}{r^4} - \frac{1}{2r^4}\sum_{m=1}^{\ell_r} \mathbb{1}_{\{0,1\}}\{S_{k,m}\}.\end{aligned}$$

Set $h^{(r)}_k = \frac{1}{2r^4}\sum_{m=1}^{\ell_r} \mathbb{1}_{\{0,1\}}\{S_m\}$. Note that $\lim_{r\to\infty}\sum_{m=0}^{t_{r,n}-1} \frac{\ell_r}{r^4} = t\sigma^2$. It follows that

$$\lim_{r\to\infty}\left|\sum_{\ell=0}^{t_{r,n}-1} \mathbb{E}\left[\left(\Delta M^{(r)}_{k,\ell,(i,j)}\right)^2\right] - t\sigma^2\right| \le \lim_{r\to\infty} n^2\gamma_r^2 t_{r,n} + \lim_{r\to\infty}\sum_{\ell=0}^{t_{r,n}-1} h^{(r)}_k.$$

It is clear that $n^2\gamma_r^2 t_{r,n} \to 0$ as $r\to\infty$, and $\lim_{r\to\infty}\sum_{k=0}^{t_{r,n}-1} h^{(r)}_k = 0$ by Lemma A.4.

For simplicity define $\widehat{\Delta} q^{(r)}_{n,k,(i,j)} \coloneqq q^{(r)}_{n,k+1,(i,j)} - p^{(r)}_{n,k+1,(i,j)}$. If $\{i,j\} \neq \{i',j'\}$ then $\widehat{\Delta} q^{(r)}_{n,k,(i,j)}$ and $\widehat{\Delta} q^{(r)}_{n,k,(i',j')}$ are independent given $\mathcal{G}_k$. In particular,

$$\begin{aligned}\mathbb{E}\left[\widehat{\Delta} q^{(r)}_{n,k,(i,j)}\widehat{\Delta} q^{(r)}_{n,k,(i',j')} \,\middle|\, \mathcal{G}_k\right] &= \mathbb{E}\left[\widehat{\Delta} q^{(r)}_{n,k,(i,j)} \,\middle|\, \mathcal{G}_k\right]\mathbb{E}\left[\widehat{\Delta} q^{(r)}_{n,k,(i',j')} \,\middle|\, \mathcal{G}_k\right] \\ &\le \frac{1}{r^4}\sum_{m=1}^{\ell_r} \mathbb{1}_{\{0,1\}}\{S_{k,m}\},\end{aligned}$$

where $S_{k,m}$ is as above. Using Lemma A.4 we conclude that

$$\lim_{r\to\infty}\left|\sum_{\ell=0}^{t_{r,n}-1} \mathbb{E}\left[\Delta M^{(r)}_{k,\ell,(i,j)}\Delta M^{(r)}_{k,\ell,(i',j')}\right]\right| \le \lim_{r\to\infty}\sum_{\ell=0}^{t_{r,n}-1} h^{(r)}_k = 0.$$

This completes the proof. □

A.2.3 Analysis away from the boundary

In the following, we denote by $S = (S_k)_{k\in\mathbb{Z}_+}$ a standard simple symmetric random walk. Recall the KMT embedding theorem [KMT75] which states that one can couple S with some Brownian motion B such that

$$\mathbb{P}\left\{\max_{0\le k\le T}\frac{|S_k - B(k)|}{r^2} \ge C\frac{\log T + x}{r^2}\right\} \le \mathrm{e}^{-x},$$

for any $T \in \mathbb{N}$. Taking $T = s_r$ (and ℓ_r respectively), we obtain that for r sufficiently large we have

$$\mathbb{P}\bigg\{\max_{0\le k\le s_r}\frac{|S_k - B(k)|}{r^2} \ge \frac{C\log r}{r^2}\bigg\} \le \mathbb{P}\bigg\{\max_{0\le k\le \ell_r}\frac{|S_k - B(k)|}{r^2} \ge \frac{C\log r}{r^2}\bigg\} \le \frac{1}{r^4}.$$

Further observe that for a fixed $\delta > 0$, we have that

$$\mathbb{P}\bigg\{\max_{0\le t\le s_r/r^4}|B(t)| \ge \delta\bigg\} \le \mathbb{P}\bigg\{\max_{0\le t\le \ell_r/r^4}|B(t)| \ge \delta\bigg\} \le 2\overline{\Phi}\bigg(\frac{\delta}{n^{-2}\sigma\sqrt{\gamma_r}}\bigg).$$

We combine these observations to obtain the following lemma.

Lemma A.6. *Let $\widetilde{S}_k = \frac{1}{r^2}S_k$ for every $k \in \mathbb{Z}_+$. Let $\epsilon > 0$ be fixed. Then, for all $r \in \mathbb{N}$ sufficiently large, we have*

$$\mathbb{P}\bigg\{\max_{k\le s_r}\big|\widetilde{S}_k\big| \ge \epsilon/2\bigg\} \le \mathbb{P}\bigg\{\max_{k\le \ell_r}\big|\widetilde{S}_k\big| \ge \epsilon/2\bigg\} \le \frac{1}{r^4} + 2\overline{\Phi}\bigg(\frac{\epsilon}{4n^{-2}\sigma\sqrt{\gamma_r}}\bigg).$$

Lemma A.7. *Let $\epsilon > 0$ fixed. Let $\ell \in \mathbb{Z}_+$ be such that $q^{(r)}_{n,\ell} \in A_\epsilon$. Then, for r sufficiently large, we have*

$$\Big\|\mathbb{E}\Big[\Delta q^{(r)}_{n,\ell}\,\Big|\,\mathcal{F}_\ell\Big] - \mathbb{E}\Big[\widetilde{\Delta} q^{(r)}_{n,\ell}\,\Big|\,\mathcal{F}_\ell\Big]\Big\|_{\mathrm{F}}^2 \le 2\bigg(\frac{n^2}{r^4} + 2n^2\overline{\Phi}\bigg(\frac{\epsilon}{4n^{-2}\sigma\sqrt{\gamma_r}}\bigg)\bigg),$$

with probability at least $1 - \frac{n^2}{r^4} - 2n^2\overline{\Phi}\Big(\frac{\epsilon}{4n^{-2}\sigma\sqrt{\gamma_r}}\Big)$.

Proof. Let $\epsilon > 0$, n, ℓ be fixed. Let $\widehat{\Delta} q^{(r)}_{n,\ell} := q^{(r)}_{n,\ell+1} - \widetilde{q}^{(r)}_{n,\ell+1}$. Begin by observing that

$$\Big\|\mathbb{E}\Big[\Delta q^{(r)}_{n,\ell}\,\Big|\,\mathcal{F}_\ell\Big] - \mathbb{E}\Big[\widetilde{\Delta} q^{(r)}_{n,\ell}\,\Big|\,\mathcal{F}_\ell\Big]\Big\|_{\mathrm{F}}^2 = \Big\|\mathbb{E}\Big[\widehat{\Delta} q^{(r)}_{n,\ell}\,\Big|\,\mathcal{F}_\ell\Big]\Big\|_2^2.$$

Let $E_{r,\ell}$ be the event that $\widetilde{q}^{(r)}_{n,\ell+1} \in A_{\epsilon/2}$. Using Lemma A.6 and union bound we conclude that

$$\mathbb{P}\Big\{\widetilde{E}_{r,\ell}\Big\} \ge 1 - \frac{n^2}{r^4} - 2n^2\overline{\Phi}\bigg(\frac{\epsilon}{4n^{-2}\sigma\sqrt{\gamma_r}}\bigg).$$

Given $p^{(r)}_{n,\ell+1}$, we observe that $\widehat{\Delta} q^{(r)}_{n,\ell}$ has the same distribution as symmetric random walk with step-size $\frac{1}{r^2}$ (reflected at boundary $\{0,1\}$) run for $\ell_{r,n}$ steps. Let us denote this j-th step of this walk by $S_{k,j}$. Also define a simple random walk with step-size $\frac{1}{r^2}$ (without reflection)

$\widetilde{S}_k$ starting with the same initial condition as S_k. Given $\widetilde{q}^{(r)}_{n,\ell+1} \in A_{\epsilon/2}$, we can couple the walk S_k and $\widetilde{S}_k$ so that $S_{k,j} = \widetilde{S}_{k,j}$ for all $j \leq T$ where $T = \min\Big\{i \in \mathbb{Z}_+ \;\Big|\; \Big|\widetilde{S}_{k,i} - \widetilde{S}_{k,0}\Big| \geq \epsilon/2\Big\}$. That is, we couple the two walks so that they are equal till they move at least $\epsilon/2$ distance from the starting position. Now notice that

$$\mathbb{E}\Big[\widehat{\Delta} q^{(r)}_{n,\ell} \;\Big|\; \mathcal{G}_\ell\Big] = \mathbb{E}\Big[\widehat{\Delta} q^{(r)}_{n,\ell} \mathbb{1}_{T \leq \ell_r}\{\,\cdot\,\} \;\Big|\; \mathcal{G}_\ell\Big] + \mathbb{E}\Big[\widehat{\Delta} q^{(r)}_{n,\ell} \mathbb{1}_{T > \ell_r}\{\,\cdot\,\} \;\Big|\; \mathcal{G}_\ell\Big].$$

using the bound $\widehat{\Delta} q^{(r)}_{n,\ell} \leq 1$ and Lemma A.6 we have that

$$\Big\|\mathbb{E}\Big[\widehat{\Delta} q^{(r)}_{n,\ell} \mathbb{1}_{T \leq \ell_r}\{\,\cdot\,\} \;\Big|\; \mathcal{G}_\ell\Big]\Big\|^2_{\mathrm{F}} \leq \frac{n^2}{r^4} + 2n^2 \overline{\Phi}\bigg(\frac{\epsilon}{4n^{-2}\sigma\sqrt{\gamma_r}}\bigg).$$

On the other hand, using the fact that $\mathbb{E}\Big[\widetilde{S}_{k,\ell_r} \;\Big|\; \mathcal{G}_\ell\Big] = 0$, we obtain

$$\begin{aligned}
\Big\|\mathbb{E}\Big[\widehat{\Delta} q^{(r)}_{n,\ell} \mathbb{1}_{T > \ell_r}\{\,\cdot\,\} \;\Big|\; \mathcal{G}_\ell\Big]\Big\|^2_{\mathrm{F}} &= \Big\|\mathbb{E}\Big[\widetilde{S}_{k,\ell_r} \mathbb{1}_{T > \ell_r}\{\,\cdot\,\} \;\Big|\; \mathcal{G}_\ell\Big]\Big\|^2_{\mathrm{F}} \\
&= \Big\|\mathbb{E}\Big[\widetilde{S}_{k,\ell_r} \mathbb{1}_{T > \ell_r}\{\,\cdot\,\} \;\Big|\; \mathcal{G}_\ell\Big] - \mathbb{E}\Big[\widetilde{S}_{k,\ell_r} \;\Big|\; \mathcal{G}_\ell\Big]\Big\|^2_{\mathrm{F}} \\
&= \Big\|\mathbb{E}\Big[\widetilde{S}_{k,\ell_r} \mathbb{1}_{T \leq \ell_r}\{\,\cdot\,\} \;\Big|\; \mathcal{G}_\ell\Big]\Big\|^2_{\mathrm{F}} \\
&\leq n^{-2}\sigma^2\gamma_r\bigg(\frac{n^2}{r^4} + 2n^2\overline{\Phi}\bigg(\frac{\epsilon}{4n^{-2}\sigma\sqrt{\gamma_r}}\bigg)\bigg).
\end{aligned}$$

Thus, we conclude that for r sufficiently large we have

$$\Big\|\mathbb{E}\Big[\widehat{\Delta} q^{(r)}_{n,\ell} \;\Big|\; \mathcal{F}_\ell\Big]\Big\|^2_{\mathrm{F}} \leq 2\bigg(\frac{n^2}{r^4} + 2n^2\overline{\Phi}\bigg(\frac{\epsilon}{4n^{-2}\sigma\sqrt{2\gamma_r}}\bigg)\bigg),$$

completing the proof. □

Lemma A.8. *Let $n \in \mathbb{N}$, for any $k \in \mathbb{Z}_+$, let $q^{(r)}_{n,k}$ and $\widetilde{q}^{(r)}_{n,k+1}$ be as defined in Step 1 of our algorithm in Chapter 3.2. Then, there exists a universal constant $c > 0$ such that for all $r \in \mathbb{N} \setminus \big[\big\lceil \mathrm{e}^{cn/4} \big\rceil\big]$, we have*

$$\Big\|K\Big(\widetilde{q}^{(r)}_{n,k+1}\Big) - K\Big(q^{(r)}_{n,k}\Big)\Big\|^2_2 \leq 4\gamma_r^2 + (16/c)\gamma_r^2 \log r =: e_r \leq (32/c)\gamma_r^2 \log r,$$

with probability at least $1 - 2r^{-4}$.

Proof. For every $i, j \in [n]$, let $(S_{i,j,k})_{k\in\mathbb{Z}_+}$ denote the 1-dimensional symmetric random walk with step-size $\frac{1}{r^2}$ starting at 0. Let these random walks be independent up to the double index symmetry for indices $(i,j) \in [n]^{(2)}$. Recall that $s_r = \lceil \gamma_r^2 r^4 \rceil$, and note that for any $i, j \in [n]$, given $q^{(r)}_{n,k,(i,j)}$ we have

$$\widetilde{q}^{(r)}_{n,k+1,(i,j)} - q^{(r)}_{n,k,(i,j)} \stackrel{\mathrm{d}}{=} \mathrm{Sko}(S_{s_r}).$$

Since the Skorokhod map is 4-Lipschitz, we conclude that

$$\mathbb{P}\left\{ \left\| K\left(\widetilde{q}^{(r)}_{n,k+1}\right) - K\left(q^{(r)}_{n,k}\right) \right\|_2^2 \geq e_r \right\} \leq \mathbb{P}\left\{ \frac{1}{n^2} \sum_{(i,j)\in[n]^{(2)}} \left(S_{i,j,s_r}\right)^2 \geq e_r/4 \right\}. \tag{A.32}$$

We will now show that the quantity $\frac{1}{n^2}\sum_{i,j\in[n]^{(2)}} \left(S_{i,j,s_r}\right)^2$ is concentrated near its expectation, that is $s_r \cdot \left(\frac{1}{r^2}\right)^2$. From the Hanson-Wright concentration inequality [RV13],

$$\begin{aligned}
&\mathbb{P}\left\{ \left| \frac{1}{n^2} \sum_{i,j\in[n]^{(2)}} \left(S_{i,j,s_r}\right)^2 - \gamma_r^2 \right| > t_r \right\} \leq \mathbb{P}\left\{ \left| \frac{1}{n^2} \sum_{i,j\in[n]^{(2)}} \left(S_{i,j,s_r}\right)^2 - s_r r^{-4} \right| > t_r \right\} \\
&\leq 2\exp\left(-c\min\left\{ \frac{t_r^2}{\left(\frac{1}{r^2}\right)^4 n s_r^2}, \frac{t_r}{\left(\frac{1}{r^2}\right)^2 s_r} \right\} \right) \leq 2\exp\left(-c\min\left\{ \frac{t_r^2}{n\gamma_r^4}, \frac{t_r}{\gamma_r^2} \right\} \right),
\end{aligned} \tag{A.33}$$

for every $t_r \geq 0$, for some universal constant $c > 0$. Let us consider $t_r \geq n\gamma_r^2$. Then, the above probability becomes $2\exp(-ct_r/\gamma_r^2)$. Moreover, for $r \geq \mathrm{e}^{cn/4}$ if we choose $t_r = (4/c)\gamma_r^2 \log r$, we have that for all $(i,j) \in [n]^{(2)}$,

$$\frac{1}{n^2} \sum_{i,j\in[n]^{(2)}} \left(S_{i,j,s_r}\right)^2 \leq \gamma_r^2 + (4/c)\gamma_r^2 \log r,$$

with probability at least $1 - 2r^{-4}$, for $e_r \coloneqq 4\gamma_r^2 + (16/c)\gamma_r^2 \log r \leq (32/c)\gamma_r^2 \log r$. □

Lemma A.9. *Let $\epsilon > 0$ be fixed, e_r be as defined in Lemma A.8, and let $q^{(r)}_{n,\ell} \in A_\epsilon$ where A_ϵ is defined in* (A.27)*. Then,*

$$\left\| \mathbb{E}\left[\Delta q^{(r)}_{n,\ell} \,\middle|\, \mathcal{F}_\ell \right] - \gamma_r n^{-4} b_n\left(q^{(r)}_{n,\ell}\right) \right\|_{\mathrm{F}}^2 \leq \frac{Cn^2}{r^4} + 4n^2\beta_{r,n}^2 e_r^3 \max\{\lambda, L\} + 2n^2\overline{\Phi}\left(\frac{\epsilon}{4n^{-2}\sigma\sqrt{\gamma_r}} \right),$$

with probability at least $1 - \frac{2}{r^4} - \frac{2n^2}{r^4} - 4n^2\overline{\Phi}\left(\frac{\epsilon}{4n^{-2}\sigma\sqrt{\gamma_r}}\right)$.

Proof. Let $I \coloneqq \left\| K\left(\widetilde{q}^{(r)}_{n,\ell+1}\right) - K\left(q^{(r)}_{n,\ell}\right)\right\|_2^2$ and let $A_{r,\ell}$ be the event that $\{I \le e_r\}$. In the following we will work on this event. Set

$$J \coloneqq \frac{\exp\left(-\beta_{r,n}\left[\mathcal{H}\left(K\left(\widetilde{q}^{(r)}_{n,\ell}\right)\right) - \mathcal{H}\left(K\left(q^{(r)}_{n,\ell}\right)\right)\right]^+\right)}{\exp\left(-\beta_{r,n}\left\langle D\mathcal{H}\left(K\left(q^{(r)}_{n,\ell}\right)\right), K\left(\widetilde{q}^{(r)}_{n,\ell+1}\right) - K\left(q^{(r)}_{n,\ell}\right)\right\rangle^+\right)},$$

From our assumption on $(\gamma_r)_{r\in\mathbb{N}}$, we have that for sufficiently large r, $\beta\gamma_r \log^2 r \le 1$. Notice that by Assumption 3.4, we have

$$1 - 2\beta_{r,n}\lambda I \le \exp(-\beta_{r,n}\lambda I) \le J \le \exp(\beta_{r,n} L I) \le 1 + 2\beta_{r,n} L I,$$

if $\beta_{r,n}\lambda I, \beta_{r,n} L I \le 1$, i.e., when r is sufficiently large. Define $b^{(r)}_n$ at $q^{(r)}_{n,\ell}$ as

$$\mathbb{E}\left[\left(\widetilde{q}^{(r)}_{n,\ell+1} - q^{(r)}_{n,\ell}\right)\exp\left(-\beta_{r,n}\left\langle D\mathcal{H}\left(K\left(q^{(r)}_{n,\ell}\right)\right), K\left(\widetilde{q}^{(r)}_{n,\ell+1}\right) - K\left(q^{(r)}_{n,\ell}\right)\right\rangle^+\right)\,\middle|\,\mathcal{F}_k\right].$$

Then, on the event $A_{r,\ell}$ we have $\left\|\mathbb{E}\left[\widetilde{\Delta} q^{(r)}_{n,\ell}\,\middle|\,\mathcal{F}_k\right] - b^{(r)}_n\left(q^{(r)}_{n,\ell}\right)\right\|_{\mathrm{F}}^2 \le 4n^2\beta^2_{r,n} e_r^3 \max\{\lambda, L\}$.

Let $E_{r,\ell}$ be the event as in the proof of Lemma A.6. On this event, we have

$$\left\|\mathbb{E}\left[\Delta q^{(r)}_{n,\ell}\,\middle|\,\mathcal{F}_\ell\right] - \mathbb{E}\left[\widetilde{\Delta} q^{(r)}_{n,\ell}\,\middle|\,\mathcal{F}_\ell\right]\right\|_{\mathrm{F}}^2 \le C\left(\frac{n^2}{r^4} + 2n^2\overline{\Phi}\left(\frac{\epsilon}{4n^{-2}\sigma\sqrt{\gamma_r}}\right)\right). \tag{A.34}$$

Moreover, on the event $E_{r,\ell}$ we also have that given $\mathcal{F}_k$, the coordinates of the $n\times n$ symmetric matrix $\left(\widetilde{q}^{(r)}_{n,\ell+1} - q^{(r)}_{n,\ell}\right)$ are i.i.d. and have the same distribution as $\widetilde{S}_{s_r}$.

Let $\widetilde{Y}_r$ be $n \times n$ matrix with independent entries such that $\widetilde{Y}_{r,(i,j)}$ be increment of the symmetric random walk (without reflection) of step-size r^{-2} starting from $q^{(r)}_{n,\ell,(i,j)}$ run for $s_r = \lceil \gamma_r^2 r^4 \rceil$ steps. Let B_n be an $n \times n$ symmetric matrix of standard Brownian motions. On the event $E_{r,\ell} \cap A_{r,\ell}$, we use the Berry-Esseen lemma (see [Pet11, Theorem 16]) with a union bound to obtain $\mathbb{W}_2^2\left(\widetilde{Y}_r, B_n(\gamma_r^2)\right) \le \frac{Cn^2}{r^4}$, for some universal constant $C > 0$.

Let $\nabla H_n\left(q^{(r)}_{n,\ell}\right) = V$. Define a function $G(Y) \coloneqq Y\exp\left(-\beta_{r,n}\langle V, Y\rangle^+_{\mathrm{F}}\right)$. Note that G is a bounded Lipschitz function of Y. Observe that $b^{(r)}_n\left(q^{(r)}_{n,\ell}\right) = \mathbb{E}\left[G\left(\widetilde{Y}_r\right)\right]$. On the other hand, we know that $\gamma_r n^{-4} b_n\left(q^{(r)}_{n,\ell}\right) = \mathbb{E}[G(B_n(\gamma_r))]$. We conclude that on the event $E_{r,\ell} \cap A_{r,\ell}$ we have

$$\left\|\mathbb{E}\left[\widetilde{\Delta} q^{(r)}_{n,\ell}\,\middle|\,\mathcal{F}_\ell\right] - \gamma_r n^{-4} b_n\left(q^{(r)}_{n,\ell}\right)\right\|_{\mathrm{F}}^2 \le \frac{Cn^2}{r^4} + 4n^2\beta^2_{r,n} e_r^3 \max\{\lambda, L\}.$$

The conclusion follows by using (A.34) and noticing that the

$$\mathbb{P}\{E_{r,\ell} \cap A_{r,\ell}\} \geq 1 - \frac{2}{r^4} - \frac{2n^2}{r^4} - 4n^2\overline{\Phi}\bigg(\frac{\epsilon}{4n^{-2}\sigma\sqrt{\gamma_r}}\bigg).$$

□

A.3 Proofs of Chapter 3.3

Proof of Proposition 3.7. The existence and uniqueness of the solution follows from the standard arguments for vector valued SDEs [KR18, Section 7.6]. We skip the details.

Let Y_n be a semimartingale. We now show that $Z_n(t) \coloneqq \mathrm{Texp}[Y_n](t)$ satisfies the SDE

$$Z_n(t) = I_n + \int_0^t \mathrm{d}Y_n(s)Z_n(s).$$

To this end, recall that $J_0(Y_n) \equiv I_n$ by definition. Note that $J_1(Y_n) = Y_n$ and for $k \in \mathbb{N}$ we have

$$J_k(Y_n)(t) = \int_0^t \mathrm{d}Y_n(s)J_{k-1}(Y_n)(s), \qquad t \in [0,1].$$

It follows that

$$\begin{aligned}
\mathrm{Texp}[Y_n](t) &= I_n + \sum_{k=1}^{\infty} J_k(Y_n)(t) \\
&= I_n + \sum_{k=1}^{\infty} \int_0^t \mathrm{d}Y_n(s)J_{k-1}(s) \\
&= I_n + \int_0^t \mathrm{d}Y_n(s)\bigg(\sum_{k=1}^{\infty} J_{k-1}(Y_n)(s)\bigg) \\
&= I_n + \int_0^t \mathrm{d}Y_n(s)\,\mathrm{Texp}[Y_n](s), \qquad t \in \mathbb{R}_+.
\end{aligned}$$

This completes the proof. □

Proof of Theorem 3.8. Recall that

$$P_{n,m}(t) \coloneqq \prod_{k=1}^{\lfloor mt \rfloor}\Big(I_n + X_{n,k}^{(m)}\Big), \qquad t \in [0,1].$$

Set $H_p = \prod_{k=1}^{p}\left(I_n + X_{n,k}^{(m)}\right)$ and $H_0 = P_{n,m}(0) = I_n$. Notice that $P_{n,m}$ is a piecewise constant interpolation of H_p. Observe that

$$\begin{aligned} P_{n,m}(t) - P_{n,m}(0) &= \sum_{j=1}^{\lfloor mt \rfloor}(H_j - H_{j-1}) \\ &= \sum_{j=1}^{\lfloor mt \rfloor} X_{n,j}^{(m)} H_{j-1} \\ &= \frac{\mu_n}{m}\sum_{j=1}^{\lfloor mt \rfloor} A_{n,k}^{(m)} H_{j-1} + \sum_{j=1}^{\lfloor mt \rfloor} M_j H_{j-1}, \end{aligned}$$

where $M_j = \frac{\mu_n}{m}\left(X_{n,j}^{(m)} - \mathbb{E}\left[X_{n,j}^{(m)}\right]\right) + \frac{\sigma_n}{\sqrt{m}} G_{n,j}^{(m)}$ for all $j \in [\lfloor mt \rfloor]$. Note that $(M_j H_{j-1})_{j\in[\lfloor mt \rfloor]}$ is a martingale difference sequence.

Consider the process

$$P_n(t) = I_n + \mu_n \int_0^t A_n(s) P_n(s)\,\mathrm{d}s + \sigma_n \int_0^t \mathrm{d}B_n(s) P_n(s), \qquad t \in \mathbb{R}_+,$$

where B_n is a matrix of i.i.d. BMs. We now couple the process P_n with $P_{n,m}$. To do so, we couple the Brownian motion B_n with Gaussian increments $\left(G_{n,k}^{(m)}\right)_{k\in[m]}$ such that $\frac{1}{\sqrt{m}} G_{n,k}^{(m)} = B_n((k+1)/m) - B_n(k/m)$ for every $k \in [m]$ and $m \in \mathbb{N}$. With this coupling, we obtain

$$\begin{aligned} \|P_{n,m}(t) - P_n(t)\|_{\mathrm{F}}^2 \le 3t\mu_n^2 \int_0^t \left\| \widetilde{A}_n^{(m)}(s) P_{n,m}(s) - A_n(s) P_n(s) \right\|_{\mathrm{F}}^2 \mathrm{d}s \\ + 3\sigma_n^2 \left\| \int_0^t \mathrm{d}B_n(s)(P_{n,m}(s) - P_n(s)) \right\|_{\mathrm{F}}^2 + \frac{3\mu_n^2}{m^2} \left\| \sum_{j=1}^{\lfloor mt \rfloor} Z_{n,j}^{(m)} H_{j-1} \right\|_{\mathrm{F}}^2, \end{aligned}$$

where $Z_{n,j}^{(m)} = M_{n,j}^{(m)} - \mathbb{E}\left[M_{n,j}^{(m)}\right]$ for all $j \in [m]$ and all $m \in \mathbb{N}$. We now set $\Delta_m(t) := \sup_{s\in[0,t]} \|P_{n,m}(s) - P_n(s)\|_{\mathrm{F}}^2$. And, obtain

$$\begin{aligned} \Delta_m(t) \le 3t\mu_n^2 \int_0^t \left\| \widetilde{A}_n^{(m)}(s) \right\|_{\mathrm{F}}^2 \Delta_m(s)\,\mathrm{d}s + 3t\mu_n^2 \int_0^t \zeta_{n,m}(s) \|P_n(s)\|_{\mathrm{F}}^2\,\mathrm{d}s \\ + 3\sigma_n^2 \sup_{s\in[0,t]} \left\| \int_0^s \mathrm{d}B_n(r)(P_{n,m}(r) - P_n(r)) \right\|_{\mathrm{F}}^2 \end{aligned}$$

$$+ \sup_{s\in[0,t]} \frac{3\mu_n^2}{m^2} \sum_{j,j'=1}^{\lfloor ms\rfloor} \operatorname{tr}\Big[H_{j-1}^\top Z_{n,j}^{(m)\top} Z_{n,j'}^{(m)} H_{j'-1}\Big],$$

where $\zeta_{n,m}(t) \coloneqq \sup_{s\in[0,t]} \Big\|\widetilde{A}_n^{(m)}(s) - A_n(s)\Big\|_{\mathrm{F}}^2$. Finally, since $\Big\|\widetilde{A}_n^{(m)}(s)\Big\|_{\mathrm{F}}^2 \le Cn^2$ for all $s \in [0,1]$, for some constant $C > 0$. Since $\Big(Z_{n,j}^{(m)}\Big)_{j\in[m]}$ are all independent for every $m \in \mathbb{N}$, and $\mathbb{E}\Big[Z_{n,j}^{(m)\top} Z_{n,j}^{(m)}\Big] \preccurlyeq nDI_n$ for all $j \in [m]$ and every $m \in \mathbb{N}$, taking expectations and using Doob's maximal inequality, we get

$$\begin{aligned}
\mathbb{E}[\Delta_m(t)] &\le 3(Ctn^2\mu_n^2 + 4\sigma_n^2) \int_0^t \mathbb{E}[\Delta_m(s)]\,\mathrm{d}s + 3t\mu_n^2\zeta_{n,m}(t) \int_0^t \mathbb{E}\big[\|P_n(s)\|_{\mathrm{F}}^2\big]\,\mathrm{d}s \\
&\quad + 24nD\mu_n^2 m^{-1} \int_0^t \mathbb{E}\big[\|P_n(s)\|_{\mathrm{F}}^2\big]\,\mathrm{d}s + 24nD\mu_n^2 m^{-1} \int_0^t \mathbb{E}[\Delta_m(s)]\,\mathrm{d}s \\
&= 3(Ctn^2\mu_n^2 + 4\sigma_n^2 + 8nD\mu_n^2 m^{-1}) \int_0^t \mathbb{E}[\Delta_m(s)]\,\mathrm{d}s \\
&\quad + 3(t\mu_n^2\zeta_{n,m}(t) + 8nD\mu_n^2 m^{-1}) \int_0^t \mathbb{E}\big[\|P_n(s)\|_{\mathrm{F}}^2\big]\,\mathrm{d}s.
\end{aligned}$$

Now we apply Grönwall inequality [Grö19] to get

$$\mathbb{E}[\Delta_m(t)] \le 3(t\mu_n^2\zeta_{n,m}(t) + 8nD\mu_n^2 m^{-1}) \int_0^t \mathbb{E}\big[\|P_n(s)\|_{\mathrm{F}}^2\big]\,\mathrm{d}s \cdot \mathrm{e}^{3t(Ctn^2\mu_n^2 + 4\sigma_n^2 + 8nD\mu_n^2 m^{-1})}.$$

The claim now follows from the assumption that $\zeta_{n,m}(t) \to 0$ as $m \to \infty$. □

Appendix B

PROOFS OF THEOREMS IN CHAPTER 4

In this section, we will provide the proofs of theorems in Chapter 4.

B.1 Proofs of Chapter 4.2

Lemma B.1. *Let $\omega_1, \dots, \omega_n \in \widehat{\mathcal{W}}$. Then there exist $w_1, \dots, w_n \in \mathcal{W}$ such that $[w_i] = \omega_i$ and $\|w_i - w_{i+1}\|_2 = \delta_2(\omega_i, \omega_{i+1})$ for every $i \in [n-1]$.*

Proof. Let $(\Omega, \mathcal{F}, \mu)$ be a probability space over some Polish space Ω equipped with the usual Borel sigma algebra $\mathcal{F}$. For a kernel w on Ω, that is, $w\colon \Omega \times \Omega \to \mathbb{R}$ we define the norm $\|\cdot\|_{2,\Omega,\mu}$ as

$$\|w\|_{2,\Omega,\mu}^2 \coloneqq \int_{\Omega^2} |w(x,y)|^2 \mu(\mathrm{d}x)\mu(\mathrm{d}y).$$

Also, Let $w \in \mathcal{W}$ be a kernel. Let $\varphi\colon (\Omega, \mathcal{F}, \mu) \to ([0,1], \mathcal{B}([0,1]), \lambda_{[0,1]})$ be a measure preserving map (i.e., $\mu(\varphi^{-1}(B)) = \lambda_{[0,1]}(B)$ for all Borel sets $B \subseteq [0,1]$). We can define w^φ as a kernel on $\Omega^{(2)}$ as

$$w^\varphi(\omega_1, \omega_2) \coloneqq w(\varphi(\omega_1), \varphi(\omega_2)), \quad \text{for } \mu\text{-a.e.} \quad \omega_1, \omega_2 \in \Omega. \tag{B.1}$$

Let $\pi, \rho\colon [0,1]^2 \to [0,1]$ be the usual coordinate projection maps, that is, $\pi\colon (x,y) \mapsto x$ and $\rho\colon (x,y) \mapsto y$. Using equation (B.1), we can define w^π and w^ρ as kernels on $\Omega = [0,1]^2$ for every kernel $w \in \mathcal{W}$. For example,

$$w^\pi((x_1,y_1),(x_2,y_2)) \coloneqq w(x_1,x_2), \qquad (x_1,y_1),(x_2,y_2) \in [0,1]^2.$$

It is easy to see that w^π is symmetric on $[0,1]^2 \times [0,1]^2$.

In the following discussion, we always equip $[0,1]$ with the Borel sigma-algebra and the Lebesgue measure, often without explicitly mentioning.

Let $u_i \in \omega_i$ for $i \in [n]$. From [Lov12, Theorem 8.13] there exist probability measures μ_i on $[0,1]^2$ for $i \in [n-1]$ such that each μ_i is a coupling of Lebesgue measures satisfying

$$\delta_2(\omega_i, \omega_{i+1}) = \left\| u_i^\pi - u_{i+1}^\rho \right\|_{2,[0,1]^2,\mu_i}. \tag{B.2}$$

Let $\pi_i \colon [0,1]^n \to [0,1]$ be the usual projection map on the i-th coordinate. By the gluing lemma [Vil03, Lemma 7.6], there exists a measure $\tilde{\mu}$ on $[0,1]^n$ such that $(\pi_i, \pi_{i+1})_\sharp \tilde{\mu} = \mu_i$. Therefore we have

$$\left\| u_i^\pi - u_{i+1}^\rho \right\|_{2,[0,1]^2,\mu_i} = \left\| u_i^{\pi_i} - u_{i+1}^{\pi_{i+1}} \right\|_{2,[0,1]^n,\tilde{\mu}}. \tag{B.3}$$

Let $\eta \colon [0,1] \to ([0,1]^n, \tilde{\mu})$ be a measure preserving bijection and let $\varphi_i \coloneqq \pi_i \circ \eta$. Then $\varphi_i \colon [0,1] \to [0,1]$ is measure preserving and therefore we obtain

$$\left\| u_i^{\pi_i} - u_{i+1}^{\pi_{i+1}} \right\|_{2,[0,1]^n,\tilde{\mu}} = \left\| u_i^{\varphi_i} - u_{i+1}^{\varphi_{i+1}} \right\|_{2,[0,1]}. \tag{B.4}$$

Combining equations (B.2), (B.3) and (B.4), and taking $w_i = u_i^{\varphi_i}$ for all $i \in [n]$, yields $\delta_2(\omega_i, \omega_{i+1}) = \|w_i - w_{i+1}\|_2$. This completes the proof. □

If $(w_t)_{t\in[0,1]} \in \mathrm{AC}(\mathcal{W}, d_2)$. It is easily seen that $(\omega_t \coloneqq [w_t])_{t\in[0,1]} \in \mathrm{AC}(\widehat{\mathcal{W}}, \delta_2)$. Lemma B.2 shows that the converse is also true.

Lemma B.2. *Let $\omega = (\omega_t)_{t\in[0,1]} \in \mathrm{AC}(\widehat{\mathcal{W}}, \delta_2)$. Then there exists $W = (w_t)_{t\in[0,1]} \in \mathrm{AC}(\mathcal{W}, d_2)$ such that $\omega_t = [w_t]$, and $\delta_2(\omega_t, \omega_s) = \|w_t - W_s\|_2$ for all $s, t \in [0,1]$.*

Proof of Lemma B.2. Following Remark 4.9, assume (possibly after a reparametrization) that the curve ω is Lipschitz with Lipschitz constant $L \geq 0$. Let $n \in \mathbb{N}$. From Lemma B.1, there exists $w_{i,n} \in \mathcal{W}$ such that $[w_{i,n}] = \omega_{i/n}$ for all $i \in \{0\} \cup [n]$, and

$$\|w_{i,n} - w_{i+1,n}\|_2 = \delta_2(\omega_{i/n}, \omega_{(i+1)/n}),$$

for all $i \in [n-1]$. For each $n \in \mathbb{N}$, let us define the curve $w^{(n)} = \left(w_t^{(n)}\right)_{t\in[0,1]}$ as

$$w_t^{(n)} \coloneqq (1 - nt + i)w_{i,n} + (nt - i)w_{i+1,n},$$

when $t \in [i/n, (i+1)/n]$ for some $i \in [n-1]$. Note that $w^{(n)}$ is also Lipschitz with constant L and therefore the family $\{w^{(n)}\}_{n\in\mathbb{N}}$ is equicontinuous w.r.t. d_2.

Since $\mathcal{W} \subseteq L^2\big([0,1]^{(2)}\big)$ is bounded in $L^2\big([0,1]^{(2)}\big)$, it is weak-$*$ precompact [San15, Box 1.2]. Since $\{w^{(n)}\}_{n\in\mathbb{N}}$ is equicontinuous w.r.t. d_2, it will also be equicontinuous w.r.t. the weak-$*$ topology. It follows from Ascoli's theorem [Mun00, Theorem 47.1] (possibly after passing to a subsequence and relabeling) that $\big(w^{(n)}\big)_{n\in\mathbb{N}}$ converges uniformly in weak-$*$ to some curve $(w_t)_{t\in[0,1]} \subseteq L^2\big([0,1]^{(2)}\big)$. It is easy to see that w_t is symmetric and $|w_t| \leq 1$ a.e. on $[0,1]^{(2)}$ and hence $w_t \in \mathcal{W}$ for every $t \in [0,1]$.

To conclude our proof, we show that $(w_t)_{t\in[0,1]}$ is Lipschitz in $\|\cdot\|_2$ and that $\delta_2([w_t], \omega_t) = 0$ for all $t \in [0,1] \cap \mathbb{Q}$ (therefore $[w_t] = \omega_t$ for rational t). Since ω is also Lipschitz, it follows that $\omega_t = [w_t]$ for all $t \in [0,1]$.

To see that $(w_t)_{t\in[0,1]}$ is Lipschitz, observe that for any $s, t \in [0,1]$,

$$\left\langle w_t^{(n)} - w_s^{(n)}, w_t - w_s \right\rangle \to \|w_t - w_s\|_2^2.$$

Using Cauchy–Schwarz inequality, we obtain

$$\|w_t - w_s\|_2 \leq \liminf_{n\to\infty} \left\| w_t^{(n)} - w_s^{(n)} \right\|_2 \leq L|t-s|.$$

We now show that $\delta_2([w_t], \omega_t) = 0$ for all $t \in [0,1] \cap \mathbb{Q}$. To this end, fix a $t \in [0,1] \cap \mathbb{Q}$ and let $t = p/q$ for some $p, q \in \mathbb{N}$. To see this, note that it follows from the proof of [Lov12, Lemma 14.16] that $\delta_2([w_t], \omega_t) \leq \liminf_{n\to\infty} \delta_2([w_{np,nq}], \omega_t) = 0$. Note that the hypothesis in [Lov12, Lemma 14.16] states that $[w_{np,nq}] \to [w_t]$ in cut-sense, but the proof only requires $w_{np,nq} \to w_t$ in weak-$*$ sense. □

Corollary B.3. *If $\omega \in \mathrm{AC}(\widehat{\mathcal{W}}, \delta_2)$, then $|\omega'|(t) = \|w_t'\|_2$ for a.e. $t \in (0,1)$, where $(w_t)_{t\in[0,1]} \in \mathrm{AC}(\mathcal{W}, d_2)$ is obtained as in Lemma B.2.*

Proof of Corollary B.3. Let ω and $(w_t)_{t\in[0,1]}$ be as above. Recall that $(w_t)_{t\in[0,1]}$ is an absolutely continuous curve in $(\mathcal{W}, d_2)$. Since every absolutely continuous curve in a Hilbert space is differentiable a.e. (Radon–Nikodým property) [Huf77, page 30, Theorem 5], there exists

a family $w'_t \in L^2\big([0,1]^{(2)}\big)$, for a.e. $t \in [0,1]$, such that $w_t - w_0 = \int_0^t w'_s \,\mathrm{d}s$ holds pointwise a.e. on $[0,1]^{(2)}$. It follows from Lebesgue differentiation theorem [Hun14, Theorem 6.32] that $\left\| \frac{w_t - w_s}{t-s} - w'_t \right\|_2 \to 0$ as $s \to t$. We know from Lemma B.2 that $\delta_2(\omega_t, \omega_s) = \|w_t - w_s\|_2$. Thus, it follows that $|\omega'|(t) = \lim_{s\to t} \frac{\delta_2(\omega_t,\omega_s)}{|t-s|} = \|w'_t\|_2$ for a.e. $t \in (0,1)$. □

Proof of Lemma 4.12. Let $(w_t)_{t\in[r,s]} \subseteq \mathrm{AC}(\mathcal{W}, d_2)$ be such that $w_r \in [u]$ and $w_s \in [v]$. Applying Jensen's inequality, we obtain

$$\int_r^s \|w'_t\|_2 \,\mathrm{d}t \geq \left\| \int_r^s w'_t \,\mathrm{d}t \right\|_2 = \|w_s - w_r\|_2 \geq \delta_2([u],[v]). \tag{B.5}$$

Following Definition 2.12, there exists $\varphi_1, \varphi_2 \in \mathcal{T}$ such that

$$\delta_2([u],[v]) = \|u^{\varphi_1} - v^{\varphi_2}\|_2. \tag{B.6}$$

Therefore, we can define an curve $(w_t)_{t\in[r,s]} \in \mathrm{AC}(\mathcal{W}, d_2)$ as $w_r \coloneqq u^{\varphi_1}$, $w_s \coloneqq v^{\varphi_2}$ and $w_t \coloneqq ((s-t)w_r + (t-r)w_s)/(s-r)$ for $t \in (r,s)$. Since for any $r \leq a < b \leq s$,

$$\|w_b - w_a\|_2 = \frac{\|w_s - w_r\|_2}{s-r} \cdot (b-a) = \frac{\|u^{\varphi_1} - v^{\varphi_2}\|_2}{s-r} \cdot (b-a), \tag{B.7}$$

therefore $(w_t)_{t\in[r,s]} \in \mathrm{AC}(\mathcal{W}, d_2)$ and $w'_t = (u^{\varphi_1} - v^{\varphi_2})/(s-r)$ exists for all $t \in (r,s)$. With this choice of $(w_t)_{t\in[r,s]} \in \mathrm{AC}(\mathcal{W}, d_2)$, from equation (B.6) we get

$$\int_r^s \|w'_t\|_2 \,\mathrm{d}t = \|u^{\varphi_1} - v^{\varphi_2}\|_2 = \delta_2([u],[v]). \tag{B.8}$$

Combining equation (B.5) and equation (B.8) completes the proof. □

Proof of Proposition 4.13. Recall that (see the remark after the Definition 4.6) for any $\omega = (\omega_t)_{t\in[0,1]} \in \mathrm{AC}(\widehat{\mathcal{W}}, \delta_2)$ such that $\omega_0 = [u]$, and $\omega_1 = [v]$, we have

$$\ell(\omega) \geq \delta_2(\omega_0, \omega_1) = \delta_2([u],[v]). \tag{B.9}$$

Given $[u], [v] \in \widehat{\mathcal{W}}$, it suffices to construct a curve $\omega^* = (\omega^*_t)_{t\in[0,1]} \in \mathrm{AC}(\widehat{\mathcal{W}}, \delta_2)$ such that $\omega^*_0 = [u]$, $\omega^*_1 = [v]$, and $\ell(\omega^*) \leq \delta_2([u],[v])$. Without any loss of generality, we can choose

$u, v \in \mathcal{W}$ such that $\delta_2([u],[v]) = \|u-v\|_2$. Define ω^* as $\omega_t^* \coloneqq [w_t]$ where $w_t \coloneqq (1-t)u + tv$ for all $t \in [0,1]$. The curve $\omega^* \in \mathrm{AC}(\widehat{\mathcal{W}}, \delta_2)$ since

$$\delta_2([w_s],[w_r]) \leq \|w_s - w_r\|_2 = \|u-v\|_2 \cdot (s-r), \tag{B.10}$$

for all $0 \leq r < s \leq 1$. Now observe that

$$\begin{aligned}
\ell(\omega^*) &= \sup\left\{ \sum_{k=0}^{n-1} \delta_2\big([w_{t_k}], \big[w_{t_{k+1}}\big]\big) \;\middle|\; n \in \mathbb{N}, 0 = t_0 < t_1 \cdots < t_n = 1 \right\} \\
&\leq \sup\left\{ \sum_{k=0}^{n-1} \|u-v\|_2 (t_{k+1} - t_k)\,\mathrm{d}t \;\middle|\; n \in \mathbb{N}, 0 = t_0 < t_1 \cdots < t_n = 1 \right\} \\
&= \|u-v\|_2 = \delta_2(\omega_0^*, \omega_1^*).
\end{aligned} \tag{B.11}$$

This completes the proof. □

Proof of Lemma 4.15. From Lemma B.1 we obtain $\varphi, \varphi_0, \varphi_1 \in \mathcal{T}$ such that

$$\delta_2([w],[w_0]) = \|w^\varphi - w_0^{\varphi_0}\|_2, \quad \text{and} \quad \delta_2([w],[w_1]) = \|w^\varphi - w_1^{\varphi_1}\|_2. \tag{B.12}$$

Defining $w_t \coloneqq (1-t)w_0^{\varphi_0} + t w_1^{\varphi_1}$ and $\vartheta_t \coloneqq [w_t]$ for $t \in [0,1]$, we get that that

$$\begin{aligned}
\delta_2^2([w],[w_t]) &\leq \|w^\varphi - (1-t)w_0^{\varphi_0} - t w_1^{\varphi_1}\|_2^2 \\
&= (1-t)\|w^\varphi - w_0^{\varphi_0}\|_2^2 + t\|w^\varphi - w_1^{\varphi_1}\|_2^2 - t(1-t)\|w_0^{\varphi_0} - w_1^{\varphi_1}\|_2^2 \\
&= (1-t)\delta_2^2([w],[w_0]) + t\delta_2^2([w],[w_1]) - t(1-t)\|w_0^{\varphi_0} - w_1^{\varphi_1}\|_2^2 \\
&\leq (1-t)\delta_2^2([w],[w_0]) + t\delta_2^2([w],[w_1]) - t(1-t)\delta_2^2([w_0],[w_1]).
\end{aligned}$$

Therefore, $\delta_2^2([w], \cdot\,)/2$ is 1-semiconvex along ϑ w.r.t. δ_2. □

B.2 Proofs of Chapter 4.3

In this section, we will provide all proofs for statements made in Chapter 4.3.

B.2.1 Proofs of Chapter 4.3.2

Lemma B.4. *If* $R\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ *is sequentially* $\delta_\square$*-lower semicontinuous, then*

1. *for every* $\tau > 0$ *and* $[u] \in \widehat{\mathcal{W}}$*, we have* $\inf_{\widehat{\mathcal{W}}} \Phi_R(\tau, [u]; \cdot) > -\infty$*, where* Φ_R *is defined in equation* (4.11)*, and*

2. *if* $([u_n])_{n\in\mathbb{N}} \subset \widehat{\mathcal{W}}$ *with* $\sup_{n\in\mathbb{N}} R([u_n]) < \infty$*, then* $([u_n])_{n\in\mathbb{N}}$ *admits a* $\delta_\square$*-converging subsequence.*

Proof. Since $(\widehat{\mathcal{W}}, \delta_\square)$ is a compact metric space [LS07], from the Weierstrass Theorem [San15, Box 1.1], both $\arg\min_{\widehat{\mathcal{W}}} R$ and $\arg\min_{\widehat{\mathcal{W}}} \Phi_R(\tau, [u]; \cdot)$ exist for all $\tau > 0$ and $[u] \in \widehat{\mathcal{W}}$. Thus the minima are greater than $-\infty$, and every sequence admits a $\delta_\square$-converging subsequence. □

Proof of Theorem 4.17. From Lemma 4.1 we know that the topology induced by δ_2 is sequentially $\delta_\square$-lower semicontinuous. This with Lemma B.4 shows that the assumptions in [AGS08, Proposition 2.2.3] are satisfied, guaranteeing that $\mathrm{GMM}_{\delta_2,\delta_\square}(\Phi_R, [u_0])$ is non-empty. If $|\partial R|$ is $\delta_\square$-lower semicontinuous and R is $\delta_\square$-continuous on the sublevel sets of $|\partial R|$, then it follows from [AGS08, Theorem 2.3.1] that every element $\omega \in \mathrm{GMM}_{\delta_2,\delta_\square}(\Phi_R, [u_0])$, for $[u_0] \in \text{eff-Dom}(R)$, is a curve of maximal slope. □

B.2.2 Proofs of Chapter 4.3.3

Lemma B.5 shows that Fréchet-like derivatives behave nicely under the Lebesgue measure-preserving transforms and hence is a well-defined map from $\widehat{\mathcal{W}}$ to $\widehat{L}^\infty([0,1]^2)$. That is, we can project $D_{\mathcal{W}}R$ to obtain $D_{\widehat{\mathcal{W}}}R\colon \text{eff-Dom}(R) \to \widehat{L}^\infty([0,1]^2)$ as $D_{\widehat{\mathcal{W}}}R([v]) \coloneqq [D_{\mathcal{W}}R(v)]$ for $v \in \mathcal{W}$.

Lemma B.5. *Let* $R\colon \mathcal{W} \to \mathbb{R} \cup \{\infty\}$ *be an invariant function. Let* $v, v' \in \text{eff-Dom}(R)$ *such that* $v' = v^\varphi$ *for some* $\varphi \in \mathcal{T}$*. Suppose that the Fréchet-like derivatives* $D_{\mathcal{W}}R(v)$ *and*

$D_{\mathcal{W}}R(v')$ exist. If $\phi = D_{\mathcal{W}}R(v)$ and $\phi' = D_{\mathcal{W}}R(v')$, then $\phi' = \phi^{\varphi}$ a.e. In particular, this implies that $D_{\mathcal{W}}R(v) \in L^{\infty}\big([0,1]^{(2)}\big)$ if it exists, is unique.

Proof. Let the sequence $(v_n)_{n\in\mathbb{N}} \subset \mathcal{W}$ be such that $\|v_n - v\|_2 \to 0$ as $n \to \infty$, then we have $\|v_n^{\varphi} - v'\|_2 \to 0$ as $n \to \infty$. We first show that

$$\lim_{n\to\infty} \frac{\langle \phi' - \phi^{\varphi}, v_n^{\varphi} - v' \rangle}{\|v_n - v\|_2} = 0. \tag{B.13}$$

To this end, recall that R is invariant and hence $R(v) = R(v^{\varphi})$ and $R(v_n^{\varphi}) = R(v_n)$. Therefore, we have

$$\begin{aligned}
&\lim_{n\to\infty} \frac{\langle \phi' - \phi^{\varphi}, v_n^{\varphi}\rangle - \langle \phi' - \phi^{\varphi}, v^{\varphi}\rangle}{\|v_n - v\|_2} \\
&= \lim_{n\to\infty} \left[\frac{R(v_n) - R(v) - \langle \phi, v_n - v\rangle}{\|v_n - v\|_2} - \frac{R(v_n^{\varphi}) - R(v^{\varphi}) - \langle \phi', v_n^{\varphi} - v^{\varphi}\rangle}{\|v_n^{\varphi} - v^{\varphi}\|_2} \right] \\
&= \lim_{n\to\infty} \frac{R(v_n) - R(v) - \langle \phi, v_n - v\rangle}{\|v_n - v\|_2} - \lim_{n\to\infty} \frac{R(v_n^{\varphi}) - R(v^{\varphi}) - \langle \phi', v_n^{\varphi} - v^{\varphi}\rangle}{\|v_n^{\varphi} - v^{\varphi}\|_2} = 0,
\end{aligned}$$

where the last equality holds because each limit individually goes to 0 by the definition of Fréchet differentiability and our assumption that R has Fréchet-like derivative at v and v'.

We now show that $\phi' - \phi^{\varphi} = 0$ a.e. Let $A^+ \coloneqq \{\phi' - \phi^{\varphi} > 0\}$ and $A^- \coloneqq \{\phi' - \phi^{\varphi} < 0\}$. It suffices to show that $|A^+| + |A^-| = 0$. We only prove that A^+ has measure 0, the proof for A^- follows similarly. Let $A \coloneqq \{v = 1\} \cap A^+$ and $B \coloneqq \{v < 1\} \cap A^+$. We claim that both A and B have measure 0. We prove this by contradiction. Suppose, for contradiction, that $|B| > 0$. Define the set $B^{\varphi} \coloneqq \{(x,y) \in [0,1]^2 \mid (\varphi(x), \varphi(y)) \in B\}$ and note that $|B| = |B^{\varphi}|$ and hence $|B^{\varphi}|$ has positive measure. Set $v_n \coloneqq v + \frac{1}{n}\chi_{B^{\varphi}}$ and note that $\|v_n - v\|_2 = \frac{|B|}{n} \to 0$ as $n \to \infty$. By equation (B.13) we conclude that

$$0 = \langle \phi' - \phi^{\varphi}, \chi_{B^{\varphi}} \rangle = \int_B (\phi' - \phi^{\varphi})(x,y)\,\mathrm{d}x\,\mathrm{d}y > 0,$$

which is a contradiction. Therefore, we must have that B has 0 measure. Repeating the same argument with $v_n \coloneqq v - \frac{1}{n}\chi_{A^{\varphi}}$ shows that A has measure 0. Since $A^+ = A \cup B$, it follows that A^+ has measure zero.

To conclude the second part, suppose that ϕ and ϕ' are two Fréchet-like derivatives of R at v. Then, (taking $\varphi = \mathrm{id}$) we must have that $\phi = \phi'$ a.e. Hence, $D_{\mathcal{W}}R(v)$ is a unique element in $L^{\infty}\big([0,1]^{(2)}\big)$. □

Proof of Lemma 4.20. Fixing $[v] \in \text{eff-Dom}(R)$, we verify using Lemma B.5 that η_R is well defined on $\widehat{\mathcal{W}}$. If $v_2 = v_1^{\varphi}$ for $v_1 \in [v]$ for some $\varphi \in \mathcal{T}$, and $\phi_1 = D_{\mathcal{W}}R(v_1)$ then $D_{\mathcal{W}}R(v_2) = \phi_1^{\varphi} =: \phi_2$, and

$$\begin{aligned}\sup_{w\in\mathcal{W}} \frac{(\langle \phi_1, v_1 - w\rangle)^+}{\|v_1 - w\|_2} &= \sup_{w\in\mathcal{W}} \frac{(\langle \phi_1^{\varphi}, v_1^{\varphi}\rangle - \langle \phi_1^{\varphi}, w^{\varphi}\rangle)^+}{\|v_1^{\varphi} - w^{\varphi}\|_2} \\ &= \sup_{w\in\mathcal{W}} \frac{(\langle \phi_2, v_2\rangle - \langle \phi_2, w\rangle)^+}{\|w - v_2\|_2}.\end{aligned} \tag{B.14}$$

We will now break the proof of the claim into two parts:

1. For any $\varepsilon > 0$, let us consider $[w] \in \widehat{\mathcal{W}}$ such that $\delta_2([v],[w]) < \delta_\varepsilon/2$ for some $\delta_\varepsilon > 0$ such that if $\varepsilon \to 0$, then $\delta_\varepsilon \to 0$. From Definition 2.11, there exists $\varphi \in \mathcal{I}$ such that $\delta_2([v],[w]) < \|w^{\varphi} - v\|_2 \le \delta_2([v],[w]) + \delta_\varepsilon/2$, i.e.,

$$\delta_\varepsilon/2 > \delta_2([v],[w]) \ge \|w^{\varphi} - v\|_2 - \delta_\varepsilon/2 > 0. \tag{B.15}$$

From assumption if we choose $w^{\varphi} \in \mathcal{W}$, since $\|w^{\varphi} - V\|_2 < \delta_\varepsilon$ we get

$$-\varepsilon \le \frac{R(w^{\varphi}) - R(v) - (\langle \phi, w^{\varphi}\rangle - \langle \phi, v\rangle)}{\|w^{\varphi} - v\|_2} \le \varepsilon, \tag{B.16}$$

where $\phi = D_{\mathcal{W}}R(v)$. Using equations (B.16) and equation (B.15), we get

$$\begin{aligned}\frac{(R([v]) - R([w]))^+}{\delta_2([v],[w])} &\le \frac{(R([v]) - R([w]))^+}{\|w^{\varphi} - v\|_2 - \delta_\varepsilon/2} \\ &\le \frac{(\langle \phi, v\rangle - \langle \phi, w^{\varphi}\rangle + \varepsilon\|w^{\varphi} - v\|_2)^+}{\|w^{\varphi} - v\|_2 - \delta_\varepsilon/2} \\ &\le \left(\frac{(\langle \phi, v\rangle - \langle \phi, w^{\varphi}\rangle)^+}{\|w^{\varphi} - v\|_2} + \varepsilon\right)\frac{\|w^{\varphi} - v\|_2}{\|w^{\varphi} - v\|_2 - \delta_\varepsilon/2} \\ &\le (\eta_R([v]) + \varepsilon)\frac{\|w^{\varphi} - v\|_2}{\|w^{\varphi} - v\|_2 - \delta_\varepsilon/2},\end{aligned} \tag{B.17}$$

for some $v \in [v]$. Taking $\varepsilon \to 0$ in equation (B.17) we get

$$|\partial R|([v]) \le \eta_R([v]). \tag{B.18}$$

2. When $\eta_R([v]) > 0$, for all $\varepsilon \in (0, \eta_R([v]))$, by the definition of $\eta_R([v])$, for any $v \in [v]$ and $\phi = D_{\mathcal{W}} R(v)$, there exists $w \in \mathcal{W}$ such that

$$0 < \varepsilon < \eta_R([v]) \leq \frac{\langle \phi, v \rangle - \langle \phi, w \rangle}{\|v - w\|_2} + \varepsilon. \tag{B.19}$$

Let $w_t \coloneqq (1-t)v + tw$ for all $t \in [0,1]$. Since $\mathcal{W}$ is a convex subset of $L^2([0,1]^{(2)})$, the curve $(w_t)_{t\in[0,1]} \subseteq \mathcal{W}$. Since $\|w_t - v\|_2 \to 0$ as $t \to 0$, by assumption we have

$$\begin{aligned}
&\lim_{t\to 0} \frac{R(w_t) - R(v) - (\langle \phi, w_t \rangle - \langle \phi, v \rangle)}{\|w_t - v\|_2} = 0 \\
\implies &\lim_{t\to 0} \frac{R(w_t) - R(v) - t(\langle \phi, w \rangle - \langle \phi, v \rangle)}{t\|w - v\|_2} = 0 \\
\implies &\lim_{t\to 0} \frac{R(v) - R(w_t)}{t\|w - v\|_2} = \frac{\langle \phi, v \rangle - \langle \phi, w \rangle}{\|v - w\|_2} \geq \eta_R([v]) - \varepsilon > 0 \\
\implies &\lim_{t\to 0} \frac{R(v) - R(w_t)}{\|w_t - v\|_2} = \lim_{t\to 0} \frac{(R(v) - R(w_t))^+}{t\|w - v\|_2} \geq \eta_R([v]) - \varepsilon \\
\implies &\lim_{t\to 0} \frac{(R([v]) - R([w_t]))^+}{\delta_2([w_t], [v])} \geq \eta_R([v]) - \varepsilon.
\end{aligned} \tag{B.20}$$

Therefore, the curve $([w_t])_{t\in[0,1]} \to [v]$ along which equation (B.20) holds for every $\varepsilon > 0$. When $\eta_R([v]) = 0$, equation (B.20) trivially holds for $\varepsilon = 0$.

Combining the two parts, we find that $|\partial R|([v]) = \eta_R([v])$.

For any $n \in \mathbb{N}$ and $\delta_n > 0$, let $A_{\delta_n} \coloneqq \{|v| < \delta_n\} \cup \{v = 1, \phi > 0\} \cup \{v = -1, \phi < 0\}$. Note that for any $t_n > 0$ and $\delta_n > 0$, define $w_n \coloneqq v - t_n \phi \mathbb{1}_{A_{\delta_n}}\{\,\cdot\,\}$ and

$$\frac{(\langle \phi, v \rangle - \langle \phi, w_n \rangle)^+}{\|v - w_n\|_2} = \big\|\phi \mathbb{1}_{A_{\delta_n}}\{\}\big\|_2. \tag{B.21}$$

Let $(\delta_n)_{n\in\mathbb{N}}$ be a sequence in $(0,1)$ such that $\lim_{n\to\infty} \delta_n = 1$. Since $\phi \in L^\infty\big([0,1]^2\big)$, for every $\delta_n > 0$, there exists $t_n > 0$ such that $w_n = v - t_n \phi \mathbb{1}_{A_{\delta_n}}\{\} \in \mathcal{W}$ for each $n \in \mathbb{N}$. It follows from equation (B.21) that

$$\eta_R([v]) \geq \limsup_{n\to\infty} \big\|\phi \mathbb{1}_{A_{\delta_n}}\{\,\cdot\,\}\big\|_2 = \|\phi \mathbb{1}_{G_v}\{\,\cdot\,\}\|_2, \tag{B.22}$$

where the last equality follows from the dominated convergence theorem and the fact that $\mathbb{1}_{A_{\delta_n}}\{\,\cdot\,\} \to \mathbb{1}_{G_v}\{\,\cdot\,\}$ a.e. as $\delta_n \to 1$.

For any $w \in \mathcal{W}$, define $w_0 = w$ on G_v and $w_0 = v$ otherwise. Note that

$$\begin{aligned}\langle \phi, v - w\rangle &= \int_{G_v} \phi(v-w)\,\mathrm{d}\lambda_{[0,1]^2} + \int_{[0,1]^2\setminus G_v} \phi(v-w)\,\mathrm{d}\lambda_{[0,1]^2} \\ &= \int_{G_v} \phi(v-w_0)\,\mathrm{d}\lambda_{[0,1]^2} + \int_{[0,1]^2\setminus G_v} \phi(v-w)\,\mathrm{d}\lambda_{[0,1]^2} \\ &\le \int_{G_v} \phi(v-w_0)\,\mathrm{d}\lambda_{[0,1]^2} = \int \phi(v-w_0)\,\mathrm{d}\lambda_{[0,1]^2} \\ &= \langle \phi, v - w_0\rangle, \end{aligned} \tag{B.23}$$

where the inequality above follows from the fact that $\phi(v-w) \le 0$ on $[0,1]^2 \setminus G_v$. Using that $\|v - w_0\|_2 \le \|v - w\|_2$, we obtain

$$\frac{(\langle \phi, v\rangle - \langle \phi, w\rangle)^+}{\|v-w\|_2} \le \frac{(\langle \phi, v\rangle - \langle \phi, w_0\rangle)^+}{\|v-w_0\|_2}.$$

It therefore follows that

$$\eta_R([v]) \coloneqq \sup_w \frac{(\langle \phi, v\rangle - \langle \phi, w\rangle)^+}{\|v-w\|_2}, \tag{B.24}$$

where the supremum is taken over $w \in \mathcal{W}$ such that $w = v$ on $[0,1]^2 \setminus G_v$. For any such w, we obtain by the Cauchy–Schwarz inequality that $\langle \phi, v - w\rangle \le \|\phi \mathbb{1}_{G_v}\{\,\cdot\,\}\|_2 \|v-w\|_2$. Therefore, it follows from that

$$\eta_R([v]) \le \|\phi \mathbb{1}_{G_v}\{\,\cdot\,\}\|_2. \tag{B.25}$$

Combining equations (B.22) and (B.25), the conclusion follows. □

Lemma B.6. *Let $R\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$. Let R be Fréchet differentiable. Let us consider $\omega \in \mathrm{AC}(\widehat{\mathcal{W}}, \delta_2)$, and let $(w_t)_{t\in[0,1]} \in \mathrm{AC}(\mathcal{W}, d_2)$ be its representative curve such that $w_t' = -\eta_R([w_t])n_t$ for a.e. $t \in [0,1]$ for some $n_t \in L^\infty\big([0,1]^{(2)}\big)$ satisfying $\|n_t\|_2 = 1$ and $\langle \phi_t, n_t\rangle = \eta_R([w_t])$. Then, ω is a curve of maximal slope on $(\widehat{\mathcal{W}}, \delta_2)$.*

Proof. Since $(w_t)_{t\in[0,1]} \in \mathrm{AC}(\mathcal{W}, d_2)$, the metric derivative of $(w_t)_{t\in[0,1]}$ with respect to d_2 at any $t \in (0,1)$ is given by

$$\lim_{h\to 0} \frac{\|w_{t+h} - w_t\|_2}{|h|} = \|w_t'\|_2 = \|\eta_R([w_t])N_t\|_2 = |\eta_R([w_t])|. \tag{B.26}$$

That is, the metric derivative of $(w_t)_{t\in[0,1]} \in \mathrm{AC}(\mathcal{W}, d_2)$ equals the upper gradient. Moreover, by the absolute continuity of the curve and from Definition 4.18,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}R(w_t) &= \lim_{h\to 0}\frac{R(w_{t+h}) - R(w_t)}{h}\\ &= \lim_{h\to 0}\frac{\langle \phi_t, w_{t+h}\rangle - \langle \phi_t, w_t\rangle + o(\|w_{t+h} - h_t\|_2)}{h}\\ &= \langle \phi_t, w_t'\rangle + 0 = -\langle \phi_t, n_t\rangle \eta_R([w_t]) = -\eta_R^2([w_t]),\end{aligned} \tag{B.27}$$

where $\phi_t = D_{\mathcal{W}}R(w_t)$. Thus, $(w_t)_{t\in[0,1]}$ satisfies Definition 4.5 and is a curve of maximal slope on $(\mathcal{W}, d_2)$, and ω is a curve of maximal slope on $(\widehat{\mathcal{W}}, \delta_2)$. □

Lemma B.7. *Let $R\colon \mathcal{W} \to \mathbb{R}\cup\{\infty\}$ be a λ-semiconvex invariant function for some $\lambda \in \mathbb{R}$ such that the Fréchet-like derivative, $\phi(w) = D_{\mathcal{W}}R(w)$, exists for all $w \in$ eff-Dom(R). Let $(w_t)_{t\in[0,1]} \in \mathrm{AC}(\mathcal{W}, d_2)$ be an absolutely continuous curve satisfying $w_t' = -\phi(w_t)\mathbb{1}_{G_{w_t}}\{\,\cdot\,\} = -D_{\mathcal{W}}R(w_t)\mathbb{1}_{G_{w_t}}\{\}$ for a.e. $t \in [0,1]$. Then, $([w_t])_{t\in[0,1]}$ is the unique* minimizing movement curve *(MM) satisfying the following evolution variational inequality (EVI)*

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}d_2^2(w_t, v) + \frac{\lambda}{2}\|w_t - v\|_2^2 + R(w_t) \le R(v), \tag{B.28}$$

for every $v \in$ eff-Dom(R).

Proof. The curve $(w_t)_{t\in[0,1]}$ is a curve of maximal slope follows from Lemma B.6. We now show that it satisfies the EVI. For $t \in \mathbb{R}_+$, let $\phi_t \coloneqq D_{\mathcal{W}}R(w_t)$. Fix $u \in \mathcal{W}$ and define the function $g_u\colon \mathcal{W} \to \mathbb{R}\cup\{\infty\}$ by $g_u(v) \coloneqq R(v) - \lambda\|u - v\|_2^2/2$, for $v \in$ eff-Dom(R). We first observe that $D_{\mathcal{W}}g_{w_t}(w_t) = \phi_t$. To see this, note that

$$\begin{aligned}&\lim_{s\to t}\frac{R(w_s) - R(w_t) - \langle \phi_t, w_s - w_t\rangle}{\|w_s - w_t\|_2} = 0\\ \implies &\lim_{s\to t}\frac{g_{w_t}(w_s) - g_{w_t}(w_t) - \langle \phi_t, w_s - w_t\rangle}{\|w_s - w_t\|_2} = 0.\end{aligned} \tag{B.29}$$

The conclusion, that is $D_{\mathcal{W}}g_{w_t}(w_t) = \phi_t$, now follows from the uniqueness of Fréchet-like derivatives (Lemma B.5). Since R is λ-semiconvex, g_u is convex, i.e.,

$$g_{w_t}(v) \ge g_{w_t}(w_t) + \langle \phi_t, v - w_t\rangle\ , \quad v \in \text{eff-Dom}(R). \tag{B.30}$$

From equation (B.23) and using the fact that $w_t' = \phi_t$, we obtain

$$\begin{aligned}\langle \phi_t, v - w_t \rangle \leq \big\langle \phi_t \mathbb{1}_{G_{w_t}}\{\,\cdot\,\}, v - w_t \big\rangle &= \left\langle -\frac{\mathrm{d}}{\mathrm{d}t} w_t, v - w_t \right\rangle \\ &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|w_t - v\|_2^2 = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} d_2^2(w_t, v), \end{aligned} \tag{B.31}$$

where the second equality follows from the reflexivity of $L^2([0,1]^2)$. Plugging equations (B.29) and (B.31) in equation (B.30) and rearranging, we get

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} d_2^2(w_t, v) + \frac{\lambda}{2}\|w_t - v\|_2^2 + R(w_t) \leq R(v), \tag{B.32}$$

for all $v \in \text{eff-Dom}(R)$. Using equation (B.32) it follows from [AGS08, Theorem 4.0.4] that the curve $([w_t])_{t\in[0,1]}$ is the unique curve in $\mathrm{MM}_{\delta_2,\delta_\square}(\Phi_R, [w_0])$. □

Proof of Theorem 4.23. Fix $[w_0] \in \text{eff-Dom}(R)$ and define

$$w_t \coloneqq w_0 - \int_0^t \phi(w_s)\mathbb{1}_{G_{w_s}}\{\,\cdot\,\}\,\mathrm{d}s\,, \quad t \in (0,1],$$

where the above integral is a pointwise integral, i.e., for a.e. $(x,y) \in [0,1]^2$,

$$w_t(x,y) \coloneqq w_0(x,y) - \int_0^t \phi(w_s)(x,y)\mathbb{1}_{G_{w_s}}\{(x,y)\}\,\mathrm{d}s\,, \quad t \in (0,1].$$

By construction, we have $(w_t)_{t\in[0,1]} \in \mathrm{AC}(\mathcal{W}, d_2)$ and $w_t' = -\phi(w_t)\mathbb{1}_{G_{w_t}}\{\,\cdot\,\}$ for all $t \in [0,1]$. It follows from Lemma B.7 that $(w_t)_{t\in[0,1]}$ is a minimizing movement. It follows from the definition of minimizing movements (see Chapter 4.3.2 and [AGS08, Definition 2.0.6]) that there exists a sequence of discrete solutions in $\widehat{\mathcal{W}}$ that converges to $([w_t])_{t\in[0,1]}$ in $\delta_\square$. Since $\mathcal{W}$ is closed in $\|\cdot\|_\square$, $w_t \in \mathcal{W}$ for all $t \in [0,1]$.

Set $\omega_t = [w_t]$ for $t \in [0,1]$. Then $\omega \in \mathrm{AC}(\widehat{\mathcal{W}}, \delta_2)$. From Lemma 4.20, we know that for any $t \in [0,1]$, $\eta_R([w_t]) = \big\|\phi(w_t)\mathbb{1}_{G_{w_t}}\{\,\cdot\,\}\big\|_2$ and therefore we have $w_t' = -\eta_R(w_t)n_t$ where $n_t \coloneqq \phi(w_t)\mathbb{1}_{G_{w_t}}\{\,\cdot\,\}/\big\|\phi(w_t)\mathbb{1}_{G_{w_t}}\{\,\cdot\,\}\big\|_2$. It follows from Lemma B.6 that ω is a curve of maximal slope on $(\widehat{\mathcal{W}}, \delta_2)$. □

B.3 Proofs of Chapter 4.4

Proof of Proposition 4.27. Note that for any sequence of graphons $([w_n])_{n\in\mathbb{N}}$ such that $([w_n])_{n\in\mathbb{N}} \xrightarrow{\delta_\square} [w]$ for some $[w] \in \widehat{\mathcal{W}}$, by Lemma 4.1 we have

$$\liminf_{n\to\infty} \delta_2([u_n],[w_n]) \geq \delta_2([u],[w]). \tag{B.33}$$

We now construct a recovery sequence of graphons $([w_n^*] \in \widehat{\mathcal{W}}_n)_{n\in\mathbb{N}} \subset \widehat{\mathcal{W}}$ such that

$$\lim_{n\to\infty} \delta_2([u_n],[w_n^*]) = \delta_2([u],[w]), \qquad \text{and} \qquad \lim_{n\to\infty} \delta_\square([w_n^*],[w]) = 0. \tag{B.34}$$

To do so, we first obtain $\varphi,\psi \in \mathcal{T}$ from Definition 2.12 and [Jan13, Theorem 6.16] such that

$$\delta_2([u],[w]) = \left\|u^\varphi - w^\psi\right\|_2. \tag{B.35}$$

Since $\delta_\square([u_n],[u]) \to 0$ as $n\to\infty$, using [Lov12, Theorem 11.59] we can find $(\varphi_n \in \mathcal{I}_n)_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty} \|u_n^{\varphi_n} - u^\varphi\|_\square = 0. \tag{B.36}$$

We now define a sequence of kernels $(z_n \in \mathcal{W}_n)_{n\in\mathbb{N}}$ as

$$z_n \coloneqq u_n^{\varphi_n} - \mathbb{E}[u^\varphi \mid \mathcal{F}_n],$$

where $\mathcal{F}_n = \sigma\{Q_n \times Q_n\}$ for every $n\in\mathbb{N}$. Note that

$$\begin{aligned}\|z_n\|_\square &\leq \|u_n^{\varphi_n} - u^\varphi\|_\square + \|u^\varphi - \mathbb{E}[u^\varphi \mid \mathcal{F}_n]\|_\square \\ &\leq \|u_n^{\varphi_n} - u^\varphi\|_\square + \|u^\varphi - \mathbb{E}[u^\varphi \mid \mathcal{F}_n]\|_2.\end{aligned}$$

Also note that for any $v\in\mathcal{W}$, the martingale sequence $(\mathbb{E}[v \mid \mathcal{F}_n] \in \mathcal{W}_n)_{n\in\mathbb{N}}$ converges to $v\in\mathcal{W}$ in $L^2\big([0,1]^{(2)}\big)$ as $n\to\infty$. Using L^2 convergence of the martingales $\mathbb{E}[u^\varphi \mid \mathcal{F}_n]$ and equation (B.36) we conclude that

$$\|z_n\|_\square \to 0, \qquad \text{as} \quad n\to\infty. \tag{B.37}$$

The sequence of kernels $(w_n^* \in \mathcal{W}_n)_{n\in\mathbb{N}}$ can now be defined as

$$w_n^* \coloneqq \mathbb{E}\big[w^\psi \;\big|\; \mathcal{F}_n\big] + z_n.$$

It now follows that for any $n \in \mathbb{N}$,

$$\begin{aligned}\big\|w_n^* - w^\psi\big\|_\square &\le \big\|\mathbb{E}\big[w^\psi \;\big|\; \mathcal{F}_n\big] - w^\psi\big\|_\square + \|z_n\|_\square \\ &\le \big\|\mathbb{E}\big[w^\psi \;\big|\; \mathcal{F}_n\big] - w^\psi\big\|_2 + \|z_k\|_\square.\end{aligned}$$

Using L^2 convergence of the martingales, and equation (B.37) we obtain $\|(w_n^*)^{\psi_n} - w^\psi\|_\square \to 0$ and therefore have

$$\limsup_{n\to\infty} \delta_\square([w_n^*],[w]) = 0. \tag{B.38}$$

Moreover,

$$\begin{aligned}\|u_n^{\varphi_n} - w_n^*\|_2^2 &= \big\|\mathbb{E}[u^\varphi \mid \mathcal{F}_n] - \mathbb{E}\big[w^\psi \;\big|\; \mathcal{F}_n\big]\big\|_2^2 \\ &\le \big\|u^\varphi - w^\psi\big\|_2^2 = \delta_2^2([u],[w]) \qquad \text{(using equation (B.35))},\end{aligned} \tag{B.39}$$

where the last inequality follows from [Lov12, Equation 9.7]. From equation (B.39) and Lemma 4.1 we obtain

$$\lim_{n\to\infty} \delta_2([u_n],[w_n^*]) = \delta_2([u],[w]). \tag{B.40}$$

Now, by the definition of $u_{n,\tau}^+$, we have

$$R\big(\big[u_{n,\tau}^+\big]\big) + \frac{1}{2\tau}\delta_2^2\big([u_n],\big[u_{n,\tau}^+\big]\big) \le R([w_n^*]) + \frac{1}{2\tau}\delta_2^2([u_n],[w_n^*]). \tag{B.41}$$

Taking $\liminf_{n\to\infty}$ on both sides of equation (B.41), and from equation (B.33), equation (B.34) and the $\delta_\square$-continuity of R, we get

$$\begin{aligned}&R\big(\big[u_{\infty,\tau}^+\big]\big) + \frac{1}{2\tau}\delta_2^2\big([u],\big[u_{\infty,\tau}^+\big]\big) \\ &\le \liminf_{n\to\infty} R\big(\big[u_{n,\tau}^+\big]\big) + \liminf_{n\to\infty} \frac{1}{2\tau}\delta_2^2\big([u_n],\big[u_{n,\tau}^+\big]\big) \\ &\le \liminf_{n\to\infty} R([w_n^*]) + \liminf_{n\to\infty} \frac{1}{2\tau}\delta_2^2([u_n],[w_n^*]) = R([w]) + \frac{1}{2\tau}\delta_2^2([u],[w]).\end{aligned} \tag{B.42}$$

Since $[w] \in \widehat{\mathcal{W}}$ was arbitrary, this completes the proof. □

Proof of Theorem 4.28. By increasing the constant G suitably, we may assume that

$$\max\left\{\sup_{n\geq 2}|\partial R|([u_{n,0}]), |\partial R|([u_0])\right\} \leq G < \infty. \tag{B.43}$$

Fix $T > 0$ and let $\boldsymbol{\tau}_m$ be a sequence of positive time steps such that $|\boldsymbol{\tau}_m| = T/m$. Since $R\colon \widehat{\mathcal{W}} \to \mathbb{R}\cup\{\infty\}$ is $\delta_\square$-continuous and $\delta_2([u],\,\cdot\,)\colon \widehat{\mathcal{W}} \to \mathbb{R}\cup\{\infty\}$ is $\delta_\square$-lower semicontinuous, the functional $\Phi_R(\tau,[u];\,\cdot\,)$ is $\delta_\square$-lower semicontinuous. From Lemma 4.15, it follows that Φ_R satisfies [AGS08, Assumption 4.0.1] and hence by [AGS08, Proposition 4.0.4] we have that

$$\omega^{(n)}(t) = \delta_\square\text{-}\lim_{m\to\infty}\left(J^{(n)}_{t/m}\right)^m([u_{n,0}])\,, \qquad \omega(t) = \delta_\square\text{-}\lim_{m\to\infty}\left(J_{t/m}\right)^m([u_0]),$$

exist and are unique for all $n \in \mathbb{N}$ and $t \in [0,T]$.

Let $\overline{[u_{n,\boldsymbol{\tau}_m}]}\colon [0,T] \to \widehat{\mathcal{W}}$ be the discrete solution of the implicit Euler method with the sequence $\boldsymbol{\tau}_m$ and initial point $[u_{n,0}]$, for each $n \in \mathbb{N}$. Inductively applying the Proposition 4.27, we obtain $\overline{[u_{\boldsymbol{\tau}_m}]}\colon [0,T] \to \widehat{\mathcal{W}}$ such that $\overline{[u_{\boldsymbol{\tau}_m}]}$ is the discrete solution of the implicit Euler method with the sequence $\boldsymbol{\tau}_m$ and initial point $[u_0] \in \widehat{\mathcal{W}}$. Passing to a subsequence and relabeling, we may assume that $\left(\overline{[u_{n,\boldsymbol{\tau}_m}]}\right)_{n\in\mathbb{N}} \xrightarrow{\delta_\square} \overline{[u_{\boldsymbol{\tau}_m}]}$ uniformly on $[0,T]$ as $n\to\infty$, that is, for any fixed sequence of step sizes $\boldsymbol{\tau}_m$, we have $\delta_\square\Big(\overline{[u_{k,\boldsymbol{\tau}_m}]}(t), \overline{[u_{\boldsymbol{\tau}_m}]}(t)\Big) \to 0$ uniformly over $t \in [0,T]$ as $n\to\infty$. For every $t \in [0,T]$ we have

$$\delta_2\Big(\overline{[u_{n,\boldsymbol{\tau}_m}]}(t), \omega^{(n)}(t)\Big) < \gamma(|\boldsymbol{\tau}_m|,\lambda,t;T,|\partial R_n|([u_{n,0}])), \tag{B.44}$$

$$\delta_2\Big(\overline{[u_{\boldsymbol{\tau}_m}]}(t), \omega(t)\Big) < \gamma(|\boldsymbol{\tau}_m|,\lambda,t;T,|\partial F|([u_0])), \tag{B.45}$$

for every $n \in \mathbb{N}$ where

$$\gamma\colon \{(\tau,\lambda,t,T) \in \mathbb{R}_{++}\times\mathbb{R}\times\mathbb{R}_{++}\times\mathbb{R}_{++} \mid \tau\lambda > -1, \tau \leq T, t \leq T\} \times (0,\infty) \to \mathbb{R}_+$$

is defined as

$$\gamma(\tau,\lambda,t;T,G) \coloneqq \begin{cases} \frac{\tau G}{\sqrt{2}}\,, & \text{if } \lambda = 0,\\ \frac{1+2|\lambda|T}{1+\lambda\tau}\cdot\frac{\tau G}{\sqrt{2}}\cdot\exp\big(-\ln\big(\frac{1+\lambda\tau}{\tau}\big)t\big)\,, & \text{if } \lambda < 0,\\ \sqrt{1+2\lambda T}\cdot\frac{\tau G}{\sqrt{2}}\cdot\exp\big(-\ln\big(\frac{1+\lambda\tau}{\tau}\big)t\big)\,, & \text{if } \lambda > 0, \end{cases} \tag{B.46}$$

by [AGS08, Equation 4.0.6, Theorem 4.0.9, Theorem 4.0.10], and the uniform bound in equation (B.43). Note that γ is independent of n. Using the triangle inequality, we get

$$\begin{aligned}\delta_{\square}\Big(\omega_t^{(n)}, \omega_t\Big) &\leq \delta_{\square}\Big(\omega_t^{(n)}, \overline{[u_{n,\boldsymbol{\tau}_m}]}(t)\Big) + \delta_{\square}\Big(\overline{[u_{n,\boldsymbol{\tau}_m}]}(t), \overline{[u_{\boldsymbol{\tau}_m}]}(t)\Big) \\ &\qquad + \delta_{\square}\Big(\overline{[u_{\boldsymbol{\tau}_m}]}(t), \omega_t\Big) \\ &\leq \delta_{2}\Big(\omega_t^{(n)}, \overline{[u_{n,\boldsymbol{\tau}_m}]}(t)\Big) + \delta_{\square}\Big(\overline{[u_{n,\boldsymbol{\tau}_m}]}(t), \overline{[u_{\boldsymbol{\tau}_m}]}(t)\Big) \\ &\qquad + \delta_{2}\Big(\overline{[u_{\boldsymbol{\tau}_m}]}(t), \omega_t\Big) \\ &\leq 2\gamma(|\boldsymbol{\tau}_m|, \lambda, t; T, G) + \delta_{\square}\Big(\overline{[u_{n,\boldsymbol{\tau}_m}]}(t), \overline{[u_{\boldsymbol{\tau}_m}]}(t)\Big), \end{aligned} \tag{B.47}$$

for all $n \in \mathbb{N}$ and $t \in [0, T]$ by equations (B.44) and (B.45).

It is clear that $\gamma(|\boldsymbol{\tau}_m|, \lambda, t; T, G) \to 0$ uniformly on $[0, T]$ as $|\boldsymbol{\tau}_m| \to 0$. Therefore, we conclude from equation (B.47) that $\delta_{\square}\Big(\omega_t^{(n)}, \omega_t\Big) \to 0$ uniformly on $t \in [0, T]$ as $n \to \infty$. $\square$

Remark B.8. The proof of the Theorem 4.28 can be carried as long as we have the uniform estimates in equation (B.44). In particular, if $\nabla R_n \circ M_n$ are uniformly Lipschitz, and there is a constant $m \in \mathbb{R}_+$ such that

$$R_n(Y_n) \leq R_n(X_n) + \langle \nabla R_n(X_n), Y_n - X_n \rangle + \frac{m}{2} \|Y_n - X_n\|_{\mathrm{F}}^2,$$

for all $X_n, Y_n \in \mathcal{M}_n$, where $R_n \coloneqq (R \circ K) \mid_{\mathcal{M}_n}$, then [But16, Theorem 212A] guarantees a uniform estimate in (B.44) and therefore the conclusion of the theorem remains valid.

B.4 Proofs of Chapter 4.5

Proof of Lemma 4.31. For any $\varphi \in \mathcal{T}$, given $\{U_i\}_{i \in \mathbb{N}}$, let us define $V_i \coloneqq \varphi(U_i)$ for all $i \in \mathbb{N}$. Since $\varphi_\sharp \lambda_{[0,1]} = \lambda_{[0,1]}$, $\{V_i\}_{i \in \mathbb{N}}$ is a set of i.i.d. uniform random variables in $[0, 1]$. Using this, observe that for any $(i, j) \in [n]^{(2)}$,

$$\begin{aligned} v_n^{\varphi}(X_n^{\varphi})(i, j) &= \mathbb{E}[v^{\varphi}(U_i, U_j) \mid X_n^{\varphi}((U_\ell)_{\ell=1}^n)] \\ &= \mathbb{E}[v(\varphi(U_i), \varphi(U_j)) \mid X_n(\varphi(U_\ell)_{\ell=1}^n)] \\ &= \mathbb{E}[v(V_i, V_j) \mid X_n((V_\ell)_{\ell=1}^n)] = v_n(X_n)(i, j), \end{aligned} \tag{B.48}$$

holds, completing the proof. □

Proof of Proposition 4.32. By definition, $\rho_{n,t}$ is the law of the random symmetric matrix $X_{n,t}((U_i)_{i=1}^n)$. Fix some $R_n \in C^1(\mathcal{M}_n, \mathbb{R})$ vanishing at the boundary of $\mathcal{M}_n$. By change of variables

$$\int_{\mathcal{M}_n} R_n(z)\rho_{n,t}(z)\,\mathrm{d}z = \int_{[0,1]^n} R_n\Big((w_t(u_i,u_j))_{(i,j)\in[n]^{(2)}}\Big)\,\mathrm{d}\lambda_{[0,1]^n}(u), \tag{B.49}$$

where $\lambda_{[0,1]^n}$ is the Lebesgue measure on $[0,1]^n$, and $u = (u_i)_{i=1}^n$. Note that $v_{n,t} \in L^2(\rho_{n,t})$. Taking time derivative on both sides of equation (B.49) for t in the set of full measure where w_t' is defined,

$$\begin{aligned}
&\partial_t \int_{\mathcal{M}_n} R_n(z)\rho_{n,t}(z)\,\mathrm{d}z \\
&= \int_{[0,1]^n} \partial_t R_n\Big((w_t(u_i,u_j))_{(i,j)\in[n]^{(2)}}\Big)\,\mathrm{d}\lambda_{[0,1]^n}(u) \\
&= \int_{[0,1]^n} \Big\langle (\nabla R \circ X_{n,t})(u), (w_t'(u_i,u_j))_{(i,j)\in[n]^{(2)}}\Big\rangle\,\mathrm{d}\lambda_{[0,1]^n}(u) \\
&= \int_{\mathcal{M}_n} \left\langle \nabla R_n(z), \int_{\{u\in[0,1]^n | X_{n,t}=z\}} (w_t'(u_i,u_j))_{(i,j)\in[n]^{(2)}}\,\mathrm{d}\mu_z(u)\right\rangle \mathrm{d}z,
\end{aligned} \tag{B.50}$$

where $\{\mu_z\}_{z\in\mathcal{M}_n}$ is the disintegration of $\lambda_{[0,1]^n}$, with respect to the function $X_{n,t}$. By definition of $v_{n,t}$, the above expression is exactly equal to $\int_{\mathcal{M}_n}\langle \nabla R_n(z), v_{n,t}(z)\rho_{n,t}(z)\rangle\,\mathrm{d}z$. This completes the proof. □

B.5 Proofs of Chapter 4.6

Proof of Lemma 4.33. By change of variables, note that

$$\begin{aligned}
R([w]) &= \mathbb{E}_{X_n\sim\rho_n([w])}[R_n(X_n)] \\
&= \mathbb{E}_{\{Z_i\sim\mathrm{Uni}[0,1]\}_{i=1}^n}\Big[R_n\Big((w(Z_i,Z_j))_{(i,j)\in[n]^{(2)}}\Big)\Big].
\end{aligned} \tag{B.51}$$

For two different graphons $[v],[w] \in \widehat{\mathcal{W}}$, consider their representative kernels v and w. Since kernels are identified a.e. on $[0,1]^{(2)}$, we may assume $w(x,x) = v(x,x) = 0$ for a.e. $x \in [0,1]$.

Now use the same sequence of Uni$[0,1]$ random variables $(Z_i)_{i=1}^n$ to obtain a coupling of $\rho_n([v])$ and $\rho_n([w])$. This is used implicitly in the following derivation and, hence, we will skip referring to the random variables $\{Z_i\}_{i=1}^n$ from here on. Define $X_n \coloneqq (w(Z_i, Z_j))_{(i,j)\in[n]^{(2)}}$ and $Y_n \coloneqq (v(Z_i, Z_j))_{(i,j)\in[n]^{(2)}}$. As a consequence of this coupling,

$$\begin{aligned}\mathbb{E}\big[\|Y_n - X_n\|_2^2\big] &= \sum_{i=1, j\neq i}^{n} \mathbb{E}\big[(v(Z_i, Z_j) - w(Z_i, Z_j))^2\big] \\ &= n(n-1)\int_{[0,1]^2} (v(x,y) - w(x,y))^2 \,\mathrm{d}x\,\mathrm{d}y = n(n-1)\|v - w\|_2^2.\end{aligned}$$

The first equality is using the fact that the diagonal terms of $Y_n - X_n$ are all zeroes. Hence $\mathbb{E}[\|Y_n - X_n\|_2] \leq n\|v - w\|_2$, by the Jensen's inequality.

If either assumptions of the Lemma hold, then for any $\varepsilon > 0$,

$$\begin{aligned}&|R_n(Y_n) - R_n(X_n) - \langle \nabla R_n(X_n), Y_n - X_n\rangle| \\ &\leq \varepsilon\|Y_n - X_n\|_2 \mathbb{1}\{\|Y_n - X_n\|_2 \leq \delta_\varepsilon\} + C_0 \mathbb{1}\{\|Y_n - X_n\|_2 > \delta_\varepsilon\}. \end{aligned} \tag{B.52}$$

Taking expectations on both sides,

$$\begin{aligned}&|\mathbb{E}[R_n(Y_n)] - \mathbb{E}[R_n(X_n)] - \mathbb{E}[\langle \nabla R_n(X_n), Y_n - X_n\rangle]| \\ &\leq \varepsilon\,\mathbb{E}[\|Y_n - X_n\|_2] + C_0\,\mathbb{E}[\mathbb{1}\{\|Y_n - X_n\|_2 > \delta_\varepsilon\}] \\ &\leq \varepsilon k\|v - w\|_2 + C_0\,\mathbb{P}\{\|Y_n - X_n\|_2 > \delta_\varepsilon\} \\ &\leq \varepsilon k\|v - w\|_2 + \frac{C_0}{\delta_\varepsilon^2} n^2 \|v - w\|_2^2,\end{aligned}$$

by Markov's inequality. As $\|v - w\|_2$ approaches zero, it is now clear that, for any $\varepsilon' > 0$,

$$\limsup_{v\in\mathcal{W},\, \|v-w\|_2 \to 0} \frac{|\mathbb{E}[R_n(Y_n)] - \mathbb{E}[R_n(X_n)] - \mathbb{E}[\langle \nabla R_n(X_n), Y_n - X_n\rangle]|}{\|v - w\|_2} \leq \varepsilon'. \tag{B.53}$$

Since ε' is arbitrary, the above lim sup must be zero.

By the definition of Fréchet-like derivatives (Definition 4.18), we want to obtain some $\phi \in L^\infty\big([0,1]^{(2)}\big)$ such that

$$\mathbb{E}[\langle \nabla R_n(X_n), Y_n - X_n\rangle] = \langle \phi, v - w\rangle. \tag{B.54}$$

Let $U_n \coloneqq Y_n - X_n$ (also similarly measurable with respect to $\{Z_i\}_{i=1}^n$), and let us denote by $A(Z) \coloneqq \nabla R_n(X_n) = \nabla R_n\Big((w(Z_i, Z_j))_{(i,j)\in[n]^{(2)}}\Big)$. Observe that

$$\begin{aligned}
\mathbb{E}[\langle \nabla R_n(X_n), Y_n - X_n\rangle] &= \sum_{i,j=1}^n \mathbb{E}\Big[(A(Z))_{i,j}(U_n(Z))_{i,j}\Big] \\
&= \sum_{i,j=1}^n \mathbb{E}\Big[\mathbb{E}\Big[(A(Z))_{i,j}(U_n(Z))_{i,j} \;\Big|\; Z_i, Z_j\Big]\Big] \\
&= \sum_{i=1,j\neq i}^n \mathbb{E}\Big[\mathbb{E}\Big[(A(Z))_{i,j}u(Z_i, Z_j) \;\Big|\; Z_i, Z_j\Big]\Big] \\
&= \sum_{i=1,j\neq i}^n \mathbb{E}\Big[\mathbb{E}\Big[(A(Z))_{i,j} \;\Big|\; Z_i, Z_j\Big]u(Z_i, Z_j)\Big] \\
&= \sum_{i=1,j\neq i}^n \int_{[0,1]^2} \mathbb{E}\Big[(A(Z))_{i,j} \;\Big|\; (Z_i, Z_j) = (x, y)\Big]u(x, y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{[0,1]^2}\left(\sum_{i,j=1}^n \mathbb{E}\Big[(A(Z))_{i,j} \;\Big|\; (Z_i, Z_j) = (x, y)\Big]\right)u(x, y)\,\mathrm{d}x\,\mathrm{d}y. \qquad \text{(B.55)}
\end{aligned}$$

Notice that, including the term $i = j$ above makes no difference in the integral. Therefore, if we choose $\phi \in L^\infty\big([0,1]^2\big)$ to be defined as

$$\phi(x, y) \coloneqq \sum_{i,j=1}^n \mathbb{E}\left[\Big(\nabla R_n\Big((w(Z_i, Z_j))_{(i,j)\in[n]^{(2)}}\Big)\Big)_{i,j} \;\middle|\; Z_i = x, Z_j = y\right], \qquad \text{(B.56)}$$

for $(x, y) \in [0,1]^{(2)}$, then the required equality in equation (B.54) is satisfied. And the action of the Fréchet-like derivative ϕ on a kernel, say $u \in \mathcal{W}$, is

$$\begin{aligned}
&\mathbb{E}[\langle \nabla R_n \circ X_n, G_n[u]\rangle] \\
&\coloneqq \mathbb{E}\Big[\Big\langle \nabla R_n\Big((w(Z_i, Z_j))_{(i,j)\in[n]^{(2)}}\Big), (u(Z_i, Z_j))_{(i,j)\in[n]^{(2)}}\Big\rangle\Big].
\end{aligned} \qquad \text{(B.57)}$$

□

Proof of Lemma 4.34. Since every geodesic is also a generalized geodesic, it suffices to prove the result for a generalized geodesic. Since R_n is λ-semiconvex, for some $\lambda \in \mathbb{R}$, it follows that for any $X_0, X_1 \in \mathcal{M}_n$,

$$R_n(X_t) \leq (1-t)R_n(X_0) + tR_n(X_1) - \frac{\lambda}{2}t(1-t)\|X_1 - X_0\|_2^2, \quad \forall\ t \in [0,1], \qquad \text{(B.58)}$$

along the curve $(X_t \coloneqq (1-t)X_0 + tX_1)_{t\in[0,1]}$.

Let $[w_0], [w_1]$ be two graphons and let $\omega = ([w_t])_{t\in[0,1]}$ be a generalized geodesic between $\omega_0 = [w_0]$ and $\omega_1 = [w_1]$. It follows from Definition 4.14 that ω has a representation in the space of kernels given by the line segment $(w_t = (1-t)w_0 + tw_1)_{t\in[0,1]}$, where the kernels w_0 and w_1 are such that $\|w_0 - w_1\|_2 \geq \delta_2([w_0], [w_1])$. Now use the same set $\{Z_i\}_{i=1}^n$ of i.i.d. Uni$[0,1]$ random variables as above to get a process $(X_{t,k})_{n\in\mathbb{N}}$ of random matrices with distributions $(\rho_n([w_t]))_{n\in\mathbb{N}}$ respectively, for each $t \in [0,1]$. Note that $X_{t,k} = (1-t)X_{0,k} + tX_{1,k}$, $t \in [0,1]$, $n \in \mathbb{N}$. Hence applying equation (B.58) to this line segment and then taking expectations with respect to the joint law of $(Z_i)_{i=1}^n$, we get

$$R([w_t]) \leq (1-t)R([w_0]) + tR([w_1]) - \frac{\lambda}{2}t(1-t)\,\mathbb{E}\big[\|X_{1,k} - X_{0,k}\|_2^2\big], \quad t \in [0,1].$$

Now by equation (B.52),

$$\mathbb{E}\big[\|X_{1,k} - X_{0,k}\|_2^2\big] = n(n-1)\|w_1 - w_0\|_2^2 \geq n(n-1)\delta_2^2([w_1], [w_0]).$$

Putting it back together we get that for $t \in [0,1]$,

$$R(\omega_t) \leq (1-t)R(\omega_0) + tR(\omega_1) - \frac{n(n-1)\lambda}{2}t(1-t)\delta_2^2(\omega_1, \omega_0). \tag{B.59}$$

Therefore R is $n(n-1)\lambda$-semiconvex along the generalized geodesic ω. □

Appendix C

PROOFS OF THEOREMS IN CHAPTER 5

In this section, we will provide the proofs of theorems in Chapter 5.

C.1 Proofs of Chapter 5.1

Proof of Proposition 5.13. We first show that $\mathbb{W}_{\blacksquare}$ and $\Delta_{\blacksquare}$ are equal. Let U, V be measure valued graphons and let φ be some bounded Lipschitz function. Using the definition of the cut norm and using Fubini's theorem,

$$\begin{aligned}\|\Gamma(\psi,U)-\Gamma(\psi,V)\|_{\square} &= \sup_{S,T}\left|\int_{S\times T}(\Gamma(\psi,U)-\Gamma(\psi,V))(x,y)\,\mathrm{d}x\,\mathrm{d}y\right| \\ &= \sup_{S,T}\left|\int\psi(\zeta)(\mathcal{L}_{S\times T}U)(\mathrm{d}\zeta)-\int\psi(\zeta)(\mathcal{L}_{S\times T}V)(\mathrm{d}\zeta)\right|,\end{aligned}$$

where $(\mathcal{L}_{S\times T}W) \coloneqq \int_{S\times T} W(x,y)\,\mathrm{d}x\,\mathrm{d}y$ for any $W \in \mathfrak{W}$, and Borel measurable sets $S, T \subseteq [0,1]$. Taking supremum over all Lipschitz functions ψ on $[-1,1]$ with $\|\psi\|_{\mathrm{Lip}} \leq 1$ on both side and interchanging the order of two suprema in the right, we obtain $\sup_{\psi}\|\Gamma(\psi,U)-\Gamma(\psi,V)\|_{\square} = \sup_{S,T}\mathbb{W}_1(\mathcal{L}_{S\times T}U, \mathcal{L}_{S\times T}V)$. Since U, V were arbitrary, the desired result now follows by replacing V with V^{φ} and taking infimum over all $\varphi \in \mathcal{T}$. It follows that $\mathbb{W}_{\blacksquare}$ and $\Delta_{\blacksquare}$ are equal. The fact that $\mathbb{W}_{\blacksquare}$ is a metric on $\widehat{\mathfrak{W}}$ follows by mimicking the standard proof of cut-metric being a metric on graphons (see [LS06]). We briefly outline the idea of the proof. Note that $\mathbb{W}_{\blacksquare}$ and $\Delta_{\blacksquare}$ do not satisfy positivity on $\mathfrak{W}$, that is, $\mathbb{W}_{\blacksquare}(U,V)$ can be 0 even though $U \neq V$. It suffices to show that $\mathbb{W}_{\blacksquare}(U,V) = 0$ if and only if $U \cong V$ in $\widehat{\mathfrak{W}}$, that is, $T_{\mathrm{d}}(F,U) = T_{\mathrm{d}}(F,V)$ for every decorated graph F. This follows from Theorem 5.15. □

Proof of Lemma 5.14. For any $\psi \in \mathcal{L}$ and $W_1, W_2 \in \mathfrak{W}$ define

$$V_\psi(x,y) = \int_{-1}^{1} \psi(\xi) W_1(x,y)(\mathrm{d}\xi) - \int_{-1}^{1} \psi(\xi) W_2(x,y)(\mathrm{d}\xi),$$

for $(x,y) \in [0,1]^{(2)}$. For any $S, T \subseteq [0,1]$, by the Kantorovich duality and Proposition 5.13, we observe that

$$\begin{aligned}
&\mathbb{W}_1\bigg(\int_{S\times T} W_1(x,y)\,\mathrm{d}x\,\mathrm{d}y, \int_{S\times T} W_2(x,y)\,\mathrm{d}x\,\mathrm{d}y\bigg) \\
&= \sup_{\psi\in\mathcal{L}}\bigg|\int_{S\times T} V_\psi(x,y)\,\mathrm{d}x\,\mathrm{d}y\bigg| \le \sup_{\psi\in\mathcal{L}}\int_{[0,1]^2} |V_\psi(x,y)|\,\mathrm{d}x\,\mathrm{d}y \le \int_{[0,1]^2} \sup_{\psi\in\mathcal{L}} |V_\psi(x,y)|\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{[0,1]^2} \mathbb{W}_1(W_1(x,y), W_2(x,y))\,\mathrm{d}x\,\mathrm{d}y \le \int_{[0,1]^2} \mathbb{W}_2(W_1(x,y), W_2(x,y))\,\mathrm{d}x\,\mathrm{d}y \\
&\le \bigg(\int_{[0,1]^2} \mathbb{W}_2^2(W_1(x,y), W_1(x,y))\,\mathrm{d}x\,\mathrm{d}y\bigg)^{1/2},
\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. The conclusion follows by replacing W_1, W_2 by $W_1^{\varphi_1}$ and $W_2^{\varphi_2}$ respectively, and taking infimum over $\varphi_1, \varphi_2 \in \mathcal{T}$. □

Lemma C.1. *Let $\mathcal{D} \subseteq \mathcal{C}$ be a subset that is closed under finite products. Suppose that the linear span $A(\mathcal{D})$ generated by $\mathcal{D}$ is dense in $\mathcal{C}$ in the sup norm. Let $(W_n)_{n\in\mathbb{N}} \in \mathfrak{W}$ and let $W \in \mathfrak{W}$. Then, the following are equivalent.*

1. *$\lim_{n\to\infty} T_{\mathrm{d}}(F, W_n) = T_{\mathrm{d}}(F, W)$ for every decorated graph F.*

2. *$\lim_{n\to\infty} T_{\mathrm{d}}(F, W_n) = T_{\mathrm{d}}(F, W)$ for every $\mathcal{D}$-decorated graph F.*

Proof. Obviously (1) implies (2). To see the converse, first note that if $\lim_{n\to\infty} T_{\mathrm{d}}(F, W_n) = T_F(W)$ for every $\mathcal{D}$-decorated graph F then $\lim_{n\to\infty} T_{\mathrm{d}}(F, W_n) = T_{\mathrm{d}}(F, W)$ for every $A(\mathcal{D})$-decorated graph F. Now let (F, f) be a $\mathcal{C}$-decorated graph and let $\epsilon > 0$ be fixed. Then, there exists an $A(\mathcal{D})$-decoration (F, g) of the skeleton F such that $\max_{i,j\in E(F)} \|f_{i,j} - g_{i,j}\|_\infty \le \epsilon$. Let $C > 0$ be a finite constant such that $\max_{i,j} \|f_{i,j}\|_\infty \le C$. It follows that $\max_{i,j} \|g_{i,j}\|_\infty \le C' = (1 + C)$. Using the Counting Lemma for decorated graphs [LS06, Lemma 10.26], for any $U \in \widehat{\mathfrak{W}}$ we have $|T_{\mathrm{d}}((F, g), U) - T_{\mathrm{d}}((F, f), U)| \le 4|E(F)|C'\epsilon$. Thus

$$|T_{\mathrm{d}}((F, f), W_n) - T_{\mathrm{d}}((F, f), W)|$$

$$\begin{aligned}
&\leq |T_{\mathrm{d}}((F,f),W_n) - T_{\mathrm{d}}((F,g),W_n)| + |T_{\mathrm{d}}((F,g),W) - T_{\mathrm{d}}((F,f),W)| \\
&\quad + |T_{\mathrm{d}}((F,g),W_n) - T_{\mathrm{d}}((F,g),W)| \\
&\leq 2C'\epsilon + |T_{\mathrm{d}}((F,g),W_n) - T_{\mathrm{d}}((F,g),W)|.
\end{aligned}$$

Since g is an $A(\mathcal{D})$-decoration and $|T_{\mathrm{d}}((F,g),W_n) - T_{\mathrm{d}}((F,g),W)| \to 0$ as $n \to \infty$. It follows that $\lim_{n\to\infty}|T_{\mathrm{d}}((F,f),W_n) - T_{\mathrm{d}}((F,f),W)| \leq 2C'\epsilon$. Taking $\epsilon \to 0$ completes the proof. □

Lemma C.2. *Let $(W_n)_{n\in\mathbb{N}} \in \mathfrak{W}$ and let $W \in \mathfrak{W}$. Then, $(W_n)_{n\in\mathbb{N}} \to W$ in $\widehat{\mathfrak{W}}$ if and only if $(T_F(W_n))_{n\in\mathbb{N}} \to T_F(W)$ for every finite simple graph F.*

Proof. Let (F,f) be a decorated graph. Define $\varphi_F := \otimes_{\{i,j\}\in E(F)} f(\{i,j\})$. Hence, $T_{\mathrm{d}}((F,f),W) = \langle \varphi_F, T_F(W)\rangle$, where T_F is as in Definition 5.5. It follows that if $T_F(W_n) \to T_F(W)$ weakly for a skeleton F, then $T_{\mathrm{d}}(F,W_n) \to T_{\mathrm{d}}(F,W)$ for any decoration (F,f). Conversely, the linear span of $\{\varphi_F \colon f \text{ is a decoration of } F\}$ is dense in $C(I_F)$ by the Stone-Weierstrass theorem. Thus, $T_{\mathrm{d}}((F,f),W_n) \to T_{\mathrm{d}}((F,f),W)$ implies that $\langle \varphi, T_F(W_n)\rangle \to \langle \varphi, T_F(W)\rangle$ for any $\varphi \in C(I_F)$. □

Lemma C.3. *If $\lim_{n\to\infty} \Delta_{\blacksquare}(W_n,W) = 0$ then $\lim_{n\to\infty} W_n = W$ in $\widehat{\mathfrak{W}}$ (see Definition 5.4).*

Proof. Assume that $\lim_{n\to\infty} \Delta_{\blacksquare}(W_n,W) = 0$. We want to show that $\lim_{n\to\infty} T_{\mathrm{d}}(F,W_n) = T_{\mathrm{d}}(F,W)$ for every decorated graph F. Since the set of Lipschitz continuous functions is dense in $\mathcal{C}$, by Lemma C.1 it is enough to show that $\lim_{n\to\infty} T_{\mathrm{d}}(F,W_n) = T_{\mathrm{d}}(F,W)$ for every Lipschitz decorated graph F.

To this end, fix a Lipschitz decorated graph F. Let $L > 0$ be such that $\max_{\{i,j\}\in E(F)} \|f_{i,j}\|_{\mathrm{BL}} \leq L$. Now observe that for any $W \in \widehat{\mathfrak{W}}$ we have $T_{\mathrm{d}}(F,W) = \int_{[0,1]^{V(F)}} \prod_{\{i,j\}\in E(F)} \Gamma(F_{i,j},W)(x_i,x_j) \prod_{v\in V(F)} \mathrm{d}x_v$. It follows from above and the Counting Lemma for decorated graphs [LS06, Lemma 10.24] that

$$\begin{aligned}
|T_{\mathrm{d}}(F,W_n) - T_{\mathrm{d}}(F,W)| &\leq 4 \sum_{\{i,j\}\in E(F)} \|\Gamma(F_{i,j},W_n) - \Gamma(F_{i,j},W)\|_{\square} \\
&\leq 4L \sum_{\{i,j\}\in E(F)} \|W_n - W\|_{\blacksquare} = 4L|E(F)|\|W_n - W\|_{\blacksquare}.
\end{aligned}$$

Replacing W_n by $W_n^{\varphi_n}$ and W by W^{φ} for any $\varphi_n, \varphi \in \mathcal{T}$ and taking infimum we obtain $|T_{\mathrm{d}}(F, W_n) - T_{\mathrm{d}}(F, W)| \leq L|E(F)|\Delta_{\blacksquare}(W_n, W)$. Since $\Delta_{\blacksquare}(W_n, W) \to 0$ as $n \to \infty$, it follows that $T_{\mathrm{d}}(F, W_n) \to T_{\mathrm{d}}(F, W)$ as $n \to \infty$. □

Lemma C.4. *Let $\mathcal{L}$ be the space of all Lipschitz function on $[-1, 1]$ with bounded Lipschitz norm at most 1. For every $\epsilon > 0$, there exists a finite set $\mathcal{F}_\epsilon \subseteq \mathcal{L}$ such that $\left|\Delta_{\blacksquare}(U, V) - \Delta_{\blacksquare}^{\mathcal{F}_\epsilon}(U, V)\right| \leq \epsilon$, for all $U, V \in \widehat{\mathfrak{W}}$, where $\Delta_{\blacksquare}^{\mathcal{F}}(U, V) \coloneqq \inf_{\varphi_1, \varphi_2 \in \mathcal{T}} \sup_{\psi \in \mathcal{F}} \|\Gamma(\psi, U^{\varphi_1}) - \Gamma(\psi, V^{\varphi_2})\|_{\square}$, for any subset $\mathcal{F} \subseteq \mathcal{C}$. Moreover, the set F_ϵ can be chosen so that $|F_\epsilon| \leq 3 \cdot 2^{16/\epsilon^2}$.*

Proof of Lemma C.4. It is an immediate consequence of Arzéla-Ascoli theorem that $\mathcal{L}$ is compact in $\mathcal{C}$. Let $\epsilon > 0$ be given. By the compactness of $\mathcal{L}$, there exists a finite subset $\mathcal{F}_\epsilon \subseteq \mathcal{L}$ such that union of $\epsilon/2$ balls centered at $\psi \in \mathcal{F}_\epsilon$ cover $\mathcal{L}$. In other words, for every $\psi \in \mathcal{L}$ there exists $\psi_0 \in \mathcal{F}_\epsilon$ such that $\|\psi - \psi_0\|_\infty < \epsilon/2$. For any $U, V \in \mathfrak{W}$, by triangle inequality, we obtain

$$|\|\Gamma(\psi, U) - \Gamma(\psi, V)\|_{\square} - \|\Gamma(\psi_0, U) - \Gamma(\psi_0, V)\|_{\square}| \leq \|\Gamma(\psi - \psi_0, U)\|_{\square} + \|\Gamma(\psi - \psi_0, V)\|_{\square},$$

that is strictly bounded by ϵ. It follows that

$$\left|\sup_{\psi \in \mathcal{L}} \|\Gamma(\psi, U) - \Gamma(\psi, V)\|_{\square} - \sup_{\psi \in \mathcal{F}_\epsilon} \|\Gamma(\psi, U) - \Gamma(\psi, V)\|_{\square}\right| < \epsilon. \tag{C.1}$$

Since the above inequality holds for every $U, V \in \mathfrak{W}$, the desired conclusion follows by replacing U and V by U^{φ_1} and V^{φ_2} respectively and taking infimum over $\varphi_1, \varphi_2 \in \mathcal{T}$.

For the second part of the claim, we will construct the finite set of bounded Lipschitz functions, denoted as F_ϵ, as follows: We divide the domain $[-1, 1]$ into $4/\epsilon$ contiguous intervals, each of length $\epsilon/4$. Given our interest in functions with a Lipschitz constant bounded by 1, we also partition the range $[-1, 1]$ into $4/\epsilon$ contiguous intervals of length $\epsilon/4$. Observe that any continuous function with bounded Lipschitz norm bounded by 1, can be approximated to within ϵ in the supremum norm using piecewise linear and continuous functions whose local slopes are taken from the set $\{-1, 0, 1\}$ over the divided domain. Therefore, to

define F_ϵ, it suffices to consider the set of piecewise linear and continuous functions that have local slopes in the set $\{-1, 0, 1\}$ over the aforementioned partition. By our construction, the size of this set is at most $3 \cdot 2^{16/\epsilon^2}$. □

Proof of Theorem 5.15. Equivalence of part 1 and part 2 follows from Lemma C.2. Lemma C.3 shows that part 3 (or equivalently part 4) implies part 1. It remains to show that part 1 implies part 3. Suppose $(W_n)_{n\in\mathbb{N}} \to W$ in $\widehat{\mathfrak{W}}$ as $n \to \infty$. We want to show that $\lim_{n\to\infty} \Delta_\blacksquare(W_n, W) = 0$.

We will argue by contradiction. Suppose, for contradiction, that there exists some $\epsilon > 0$ and some subsequence $(n_k)_{k=1}^\infty$ such that $\Delta_\blacksquare(W_{n_k}, W) \geq \epsilon$. By Lemma C.4 there exists a finite family of functions $\mathcal{F} \subseteq \mathcal{L}$ such that $\Delta_\blacksquare(U, V) \leq \Delta_\blacksquare^{\mathcal{F}}(U, V) + \frac{\epsilon}{2}$, for all $U, V \in \widehat{\mathfrak{W}}$. Since $\mathcal{F}$ is finite and $(W_{n_k})_{k\in\mathbb{N}} \to W$ as $k \to \infty$ in $\widehat{\mathfrak{W}}$, it follows from [LS10, Lemma 3.2, Lemma 3.7] that $\lim_{k\to\infty} \Delta_\blacksquare^{\mathcal{F}}(W_{n_k}, W) = 0$. This implies that $\limsup_{k\to\infty} \Delta_\blacksquare(W_{n_k}, W) \leq \epsilon/2$ which is a contradiction. □

Proof of Lemma 5.18. The proof of Lemma 5.18 follows essentially the same idea as the proof of [KKLS14, Theorem 3.8].

Let (F, f) be a decorated graph and let $\mathbb{G}(n, W)$ be as defined in Lemma 5.18. Recall that $T_{\mathrm{d}}(F, \mathbb{G}(n, W)) = \frac{1}{n^k} \sum_{i_1,\ldots,i_k} \prod_{\{j,l\}\in E(F)} f_{j,l}(\zeta_{i_j,i_l})$, where $\zeta_{u,v}$s are independent and distributed as $W(U_u, U_v)$ for all (u, v). In particular, $\mathbb{E}[T_{\mathrm{d}}(F, \mathbb{G}(n, W))] = T_{\mathrm{d}}(F, W)$ for each $n \in \mathbb{N}$. It suffices to show that $T_{\mathrm{d}}(F, \mathbb{G}(n, W))$ concentrates around its mean for all decorated graphs (F, f). To this end, fix a decorated graph (F, f) and set $d_n(F) := |T_{\mathrm{d}}(F, \mathbb{G}(n, W)) - \mathbb{E}[T_{\mathrm{d}}(F, \mathbb{G}(n, W))]|$. Using a 4-th moment bound, following the same argument as in [Lov12, Equation 11.5], we obtain $\mathbb{P}\{d_n(F) \geq \epsilon\} \leq \frac{C}{\epsilon^2 n^2}$. Using Borel-Cantelli Lemma we conclude that $T_{\mathrm{d}}(F, \mathbb{G}(n, W)) \to T_{\mathrm{d}}(F, W)$ almost surely. To conclude the proof, we observe that set of all finite simple graphs is countable and $\mathcal{C} = C[-1, 1]$ is a separable space. We, therefore, can find a countable dense subset of decorated graphs for which almost sure convergence of homomorphism densities holds. The proof is complete using a standard approximation argument similar to [KKLS14, Theorem 3.4]. □

C.2 Proofs of Chapter 5.2

Proof of Theorem 5.21. Recall that the Aldous-Hoover representation provides a one-to-one correspondence (see (5.2)) between IEAs and random MVGs, in other words, between $\mathcal{P}_{\mathrm{e}}(\mathcal{S})$ and $\mathcal{P}(\widehat{\mathfrak{W}})$. Also note that $\mathcal{P}_{\mathrm{e}}(\mathcal{S})$ and $\mathcal{P}(\widehat{\mathfrak{W}})$ are both compact and metrizable (and hence Hausdorff). To show that $\mathcal{P}_{\mathrm{e}}(\mathcal{S})$ and $\mathcal{P}(\widehat{\mathfrak{W}})$ are homeomorphic, it suffices to show that the $\mathbf{X} \mapsto W_{\mathbf{X}}$ is continuous. Let $\mathbf{X}_n$ be a sequence of exchangeable arrays such that $\mathbf{X}_n \to \mathbf{X}$ weakly as $n \to \infty$ for some exchangeable array $\mathbf{X}$. Let W_n and W be the corresponding (random) measure valued graphons. We want to show that $W_n \to W$ weakly, that is, $\mathbb{E}[T_{\mathrm{d}}(F, W_n)] \to \mathbb{E}[T_{\mathrm{d}}(F, W)]$ for every decorated graph F. To see this, fix a decorated graph F. Since $\mathbf{X}_n \to \mathbf{X}$ weakly as $n \to \infty$, it follows that $T_{\mathrm{d}}(F, \mathbf{X}_n) \to T_{\mathrm{d}}(F, \mathbf{X})$ as $n \to \infty$. Observe that $\mathbb{E}[T_{\mathrm{d}}(F, W_n)] = T_{\mathrm{d}}(F, \mathbf{X}_n)$ and $T_{\mathrm{d}}(F, \mathbf{X}) = \mathbb{E}[T_{\mathrm{d}}(F, W)]$ by Proposition 5.20. Hence, $\mathbb{E}[T_{\mathrm{d}}(F, W_n)] \to \mathbb{E}[T_{\mathrm{d}}(F, W)]$ as $n \to \infty$. This completes the proof. □

Appendix D
PROOFS OF CHAPTER 6

In this chapter we will provide all proofs of statements in Chapter 6.

D.1 Proofs of Chapter 6.1

Proof of Lemma 6.1. Fix $(i,j) \in \mathbb{N}^{(2)}$ and note that $f(U,U_i,U_j,U_{i,j}) = f(U,U_j,U_i,U_{i,j})$ since $\zeta_{i,j} = \zeta_{j,i}$ and $U_{i,j} = U_{j,i}$. Therefore, $\mathbb{E}[f(U,U_i,U_j,U_{i,j}) \,|\, U,U_i,U_j] = \mathbb{E}[f(U,U_j,U_i,U_{i,j}) \,|\, U,U_i,U_j]$, and,

$$\begin{aligned} w_u(x,y) &= \mathbb{E}[f(U,U_i,U_j,U_{i,j}) \,|\, U=u, U_i=x, U_j=y] \\ &= \mathbb{E}[f(U,U_j,U_i,U_{i,j}) \,|\, U=u, U_i=x, U_j=y] = g_u(y,x), \end{aligned}$$

for a.e. $(x,y) \in [0,1]^{(2)}$. Since the maps f, $\mathbb{E}$ and $[\,\cdot\,]$ are all measurable, their composition is also measurable. Because U is a random variable, $[w_U]$ is also a random variable obtained as a composition of measurable maps.

To see equation (6.6), start with the Aldous-Hoover representation $\zeta_{i,j} = f(U,U_i,U_j,U_{i,j})$ for every $(i,j) \in \mathbb{N}^{(2)}$. Condition on $\{U=u\}$ throughout for $u \in [0,1]$. For any finite simple graph F, with k vertices,

$$\begin{aligned} T_F\left(K\left((\zeta_{i,j})_{(i,j)\in[n]^{(2)}}\right)\right) &= \frac{1}{n^{\downarrow k}} \sum_{i_1,i_2,\ldots,i_k} \prod_{\{j,l\}\in E(F)} \zeta_{i_j i_l} \\ &= \frac{1}{n^{\downarrow k}} \sum_{i_1,i_2,\ldots,i_k} \prod_{\{j,l\}\in E(F)} f(u,U_{i_j},U_{i_l},U_{i_j,i_l}), \end{aligned} \tag{D.1}$$

where the summation runs over the $n^{\downarrow k} \coloneqq n!/(n-k)!$ many injections from $[k]$ to $[n]$, and $T_F \colon \mathcal{W} \to \mathbb{R}$ is the homomorphism density function of F (see equation (2.7)). Notice that

$$\mathbb{E}\Big[T_F\left(K\left((\zeta_{i,j})_{(i,j)\in[n]^{(2)}}\right)\right)\Big] = \int_{[0,1]^k} \prod_{\{j,l\}\in E(F)} \mathbb{E}[f(u,u_j,u_l,V)]\,\mathrm{d}u_1\cdots\mathrm{d}u_k = T_F(w_u),$$

where w_u is defined in equation (6.5). Hence, the lemma will be true if we show that the strong law of large numbers holds. That the weak law of large numbers holds, can be seen by a variance computation. That the convergence is a.e. follows from Borel-Cantelli lemma [Kal21, Theorem 4.18]. We skip the standard argument. The conclusion holds following the inverse counting lemma [Lov12, Lemma 10.32]. □

D.1.1 Proofs on the existence of solution of McKean-Vlasov SDE

To argue about the existence of a unique solution of the system of SDEs (Graphon-MKV), we construct a sequence of stochastic processes $\left(X^{(k)}, \gamma^{(k)}\right)_{k\in\mathbb{Z}_+}$ on $C\big([0,\infty), [-1,1]^{\mathbb{N}^{(2)}} \times \mathcal{W}\big)$ iteratively. Start by defining $\left(X^{(0)}, \gamma^{(0)}\right)$ as $X^{(0)}_{i,j}(t) \equiv w_0(U_i, U_j)$, $\gamma^{(0)}(t) \equiv w_0$, for all $(i,j) \in \mathbb{N}^{(2)}$, and $t \in \mathbb{R}_+$. The induction proceeds by showing that whenever $\left(X^{(k)}, \gamma^{(k)}\right)$ for $k \in \mathbb{Z}_+$ is well defined, $X^{(k)}$ is an infinite exchangeable array (Lemma D.1 below) and, $\gamma^{(k)}$ is a deterministic process of kernels (Lemma D.2). Note that these claims are clearly true for $k = 0$. Then, inductively, define the process $X^{(k+1)}$ as the strong solution to the coordinatewise reflected SDE:

$$\begin{aligned} \mathrm{d}X^{(k+1)}_{i,j}(t) = b\Big(X^{(k)}_{i,j}(t), \gamma^{(k)}(t)\Big)(U_i, U_j)\,\mathrm{d}t + \Sigma\left(\gamma^{(k)}(t)\right)(U_i, U_j)\,\mathrm{d}B_{i,j}(t) \\ + \mathrm{d}L^{(k+1)-}_{i,j}(t) - \mathrm{d}L^{(k+1)+}_{i,j}(t), \end{aligned} \tag{D.2}$$

for $t \in \mathbb{R}_+$, with the same initial condition $X^{(k+1)}_{i,j}(0) = w_0(U_i, U_j)$ for all $(i,j) \in \mathbb{N}^{(2)}$. As usual, $L^{(k+1)-}_{i,j}$ and $L^{(k+1)+}_{i,j}$ are processes such that $\left(X^{(k+1)}_{i,j}, L^{(k+1)+}_{i,j}, L^{(k+1)-}_{i,j}\right)$ solves the Skorokhod problem with respect to $[-1,1]$ (see Section 2.4) for every $(i,j) \in \mathbb{N}^{(2)}$. Since the drift and diffusion functions ϕ and Σ are deterministic and Lipschitz (Assumption 3.1), given $\mathcal{F}_0$, every process $X^{(k)}$ for $k \in \mathbb{N}$ exists uniquely in the strong sense.

In fact, given $\mathcal{F}_0$, the entries of the array $X^{(k+1)}$ are independent and distributed as reflected Brownian motions (RBMs) with Lipschitz (but time-varying) drifts and diffusion coefficients. In particular, the kernel $\gamma^{(k+1)}$ is constructed from the array $X^{(k+1)}$ (which over the entire probability space is exchangeable, as we show next in Lemma D.1) as described

in equation (6.5) in Lemma 6.1, and is therefore defined as

$$\gamma^{(k+1)}(t)(x,y) \coloneqq \mathbb{E}\Big[X_{1,2}^{(k+1)}(t) \,\Big|\, U_1 = x, U_2 = y\Big], \qquad t \in \mathbb{R}_+. \tag{D.3}$$

The kernel $\gamma^{(k+1)}(t)$ is well-defined for a.e. $(x,y) \in [0,1]^{(2)}$ and all $t \in \mathbb{R}_+$. The induction hence continues.

Lemma D.1. *Suppose that, for some $k \in \mathbb{Z}_+$, there is a unique in law solution to the SDE* (D.2) *for $X^{(k+1)}$ and that $\gamma^{(k+1)}$ is a deterministic process of kernels. Then the process $X^{(k+1)}$ is an infinite exchangeable array taking values in $\mathcal{E} = C[0,\infty)$, equipped with the usual locally uniform metric.*

Proof. To argue the exchangeability, let $\sigma\colon \mathbb{N} \to \mathbb{N}$ be a finite permutation of the natural numbers $\mathbb{N}$. Note that σ fixes every large enough natural number. We need to argue that $\big(X_{i,j}^{(k+1)}\big)_{(i,j)\in\mathbb{N}^{(2)}}$ has the same law as $\big(X_{\sigma_i,\sigma_j}^{(k+1)}\big)_{(i,j)\in\mathbb{N}^{(2)}}$ in the sense of equality of the two probability measures on $(C[0,\infty))^{\mathbb{N}^{(2)}}$.

Let $\widetilde{U}_i \coloneqq U_{\sigma_i}$, for all $i \in \mathbb{N}$. Then $\big(\widetilde{U}_i\big)_{i\in\mathbb{N}}$ is again a sequence of i.i.d. Uni$[0,1]$ random variables. Let $Y_{i,j}^{(k+1)} \equiv X_{\sigma_i,\sigma_j}^{(k+1)}$ for every $(i,j) \in \mathbb{N}^{(2)}$. Since $Y_{i,j}^{(k+1)}(0) = w_0(U_{\sigma_i}, U_{\sigma_j}) =: w_0(\widetilde{U}_i, \widetilde{U}_j)$. It follows that $\big(Y_{i,j}^{(k+1)}(0)\big)_{(i,j)\in\mathbb{N}^{(2)}}$ has the same distribution as $\big(X_{i,j}^{(k+1)}(0)\big)_{(i,j)\in\mathbb{N}^{(2)}}$. Moreover for every $(i,j) \in \mathbb{N}^{(2)}$, the process $Y^{(k+1)}$ satisfies the SDEs

$$\begin{aligned}
\mathrm{d}Y_{i,j}^{(k+1)}(t) &= b\Big(X_{\sigma_i,\sigma_j}^{(k)}(t), \gamma^{(k)}(t)\Big)(U_{\sigma_i}, U_{\sigma_j})\,\mathrm{d}t + \Sigma\big(\gamma^{(k)}(t)\big)(U_{\sigma_i}, U_{\sigma_j})\,\mathrm{d}B_{\sigma_i,\sigma_j}(t) \\
&\qquad + \mathrm{d}L_{\sigma_i,\sigma_j}^{(k+1)-}(t) - \mathrm{d}L_{\sigma_i,\sigma_j}^{(k+1)+}(t) \\
&= b\Big(Y_{i,j}^{(k)}(t), \gamma^{(k)}(t)\Big)(\widetilde{U}_i, \widetilde{U}_j)\,\mathrm{d}t + \Sigma\big(\gamma^{(k)}(t)\big)(\widetilde{U}_i, \widetilde{U}_j)\,\mathrm{d}B_{\sigma_i,\sigma_j}(t) \\
&\qquad + \mathrm{d}L_{\sigma_i,\sigma_j}^{(k+1)-}(t) - \mathrm{d}L_{\sigma_i,\sigma_j}^{(k+1)+}(t),
\end{aligned}$$

for $(i,j) \in \mathbb{N}^{(2)}$ and $t \in \mathbb{R}_+$. Note that, $\gamma^{(k)}$ does not get affected by the permutation σ.

Relabeling $\widetilde{B}_{i,j} \coloneqq B_{\sigma_i,\sigma_j}$, $\widetilde{L}_{i,j}^{(k+1)-} \coloneqq L_{\sigma_i,\sigma_j}^{(k+1)-}$ and $\widetilde{L}_{i,j}^{(k+1)+} \coloneqq L_{\sigma_i,\sigma_j}^{(k+1)+}$ for every $(i,j) \in \mathbb{N}^{(2)}$, leaves their joint law unchanged, and we get

$$\mathrm{d}Y_{i,j}^{(k+1)}(t) = b\Big(Y_{i,j}^{(k)}(t), \gamma^{(k)}(t)\Big)(\widetilde{U}_i, \widetilde{U}_j)\,\mathrm{d}t + \Sigma\big(\gamma^{(k)}(t)\big)(\widetilde{U}_i, \widetilde{U}_j)\,\mathrm{d}\widetilde{B}_{i,j}(t)$$

$$+ \mathrm{d}\widetilde{L}_{i,j}^{(k+1)-}(t) - \mathrm{d}\widetilde{L}_{i,j}^{(k+1)+}(t),$$

for every $(i,j) \in \mathbb{N}^{(2)}$ and $t \in \mathbb{R}_+$. Since $X^{(k+1)}$ and $Y^{(k+1)}$ follow the same system of recursive SDEs (D.2), their equivalence in law follows from the uniqueness in law of the SDE. □

Lemma D.2. *Under the same assumption as in Lemma D.1 and Assumption 6.1, the kernel-valued map $t \mapsto \gamma^{(k)}(t)$, is deterministic and absolutely continuous. Moreover, for each $t \in \mathbb{R}_+$, we have*

$$\lim_{n\to\infty} \delta_{\square}\left(\left[K\left(\left(X_{i,j}^{(k)}(t)\right)_{(i,j)\in[n]^{(2)}}\right)\right], \left[\gamma^{(k)}(t)\right]\right) = 0, \qquad a.s. \tag{D.4}$$

Proof. By definition, for $(x,y) \in [0,1]^{(2)}$, and $t \in \mathbb{R}_+$,

$$\gamma^{(k)}(t)(x,y) \coloneqq \mathbb{E}\Big[X_{1,2}^{(k)}(t) \,\Big|\, U_1 = x, U_2 = y\Big].$$

This is a deterministic kernel for every $t \in \mathbb{R}_+$. To see (D.4), repeat the proof of Lemma 6.1. Notice that, there is no random variable U as in Lemma 6.1 (also see Remark 6.2). This is now a consequence of Kolmogorov's zero-one law [Kal21, Theorem 4.13]. For $n \in \mathbb{N}$, let $\mathcal{G}_n$ be the sigma algebra generated by U_n and the i.i.d. standard Brownian motions $B_{i,j}$s for the set of indices $\big\{(i,j) \in \mathbb{N}^{(2)} \,\big|\, j = n\big\}$. This is a sequence of independent sigma algebras. Consider its tail sigma algebra $\mathcal{T} \coloneqq \cap_{n\in\mathbb{N}} \vee_{\ell\geq n} \mathcal{G}_\ell$. This is a trivial sigma algebra by the Kolmogorov zero-one law.

Consider, for any finite simple graph F and $t \in \mathbb{R}_+$, the limiting homomorphism densities $\lim_{n\to\infty} T_F\big(K\big(\big(X_{i,j}^{(k)}(t)\big)_{(i,j)\in[n]^{(2)}}\big)\big)$, as in equation (D.1). These limiting homomorphism densities do not depend on finitely many elements in $\{X_{i,j}^{(k)}(t)\}_{(i,j)\in\mathbb{N}^{(2)}}$ or $\{U_i\}_{i\in\mathbb{N}}$. In particular, such limits are measurable with respect to the tail sigma algebra $\mathcal{T}$. Exactly as in the proof of Lemma 6.1, it follows that

$$\lim_{n\to\infty} \delta_{\square}\left(\left[K\left(\left(X_{i,j}^{(k)}(t)\right)_{(i,j)\in[n]^{(2)}}\right)\right], \left[\gamma^{(k)}(t)\right]\right) = 0.$$

In particular, the graphon $\left[\gamma^{(k)}(t)\right]$ is measurable with respect to $\mathcal{T}$, and thus constant a.e.

Finally, the absolute continuity of $t \mapsto \gamma(t)$ follows from the path continuity of the process $X_{1,2}^{(k)}$ and our assumptions on b and Σ. □

Proposition D.3. *Assume that the drift functions* $b\colon [-1,1] \times \mathcal{W} \to L^\infty\big([0,1]^{(2)}\big)$ *satisfies Assumption 6.1, and the diffusion coefficient function* $\Sigma\colon \mathcal{W} \to L^\infty\big([0,1]^{(2)}\big)$ *is bounded and* κ_2*-Lipschitz in* $\|\cdot\|_2$ *(Assumption 3.3). Then the sequence of processes taking values in* $C([0,\infty), [-1,1] \times \mathcal{W})$ *given by* $\Big(\big(X_{1,2}^{(k)}(t), \gamma^{(k)}(t)\big)_{t\in\mathbb{R}_+}\Big)_{k\in\mathbb{Z}_+}$*, converges locally uniformly in the* 2*-product metric of* $[-1,1]$ *and* $(\mathcal{W}, d_2)$*, to a pathwise unique process* $\big(X_{1,2}(t), \gamma(t)\big)_{t\in\mathbb{R}_+}$ *starting from* $\gamma(0) = w_0 \in \mathcal{W}$ *and* $X_{1,2}(0) = w_0(U_1, U_2)$*. That is, for every* $t \in \mathbb{R}_+$*,*

$$\lim_{k\to\infty} \sup_{s\in[0,t]} \left[\left| X_{1,2}^{(k)}(s) - X_{1,2}(s) \right|^2 + \left\| \gamma^{(k)}(s) - \gamma(s) \right\|_2^2 \right] = 0, \qquad a.s. \tag{D.5}$$

In particular, the limiting processes $X_{1,2}$ *is continuous and* γ *is absolutely continuous and deterministic.*

Proof. The proof is a standard Picard iteration based proof of existence of solutions of SDEs. See, for example, the proof of [KS91, Theorem 2.9, page 289]. Hence, we will skip some of the details and refer the reader to the above cited reference.

We will take $k \to \infty$ and produce a limit. Start by noticing that the process $X_{1,2}^{(k+1)}\colon \mathbb{R}_+ \to [-1,1]$ is the result of applying the Skorokhod map [KLRS07] pathwise to the "noise before reflection" process $Y_{1,2}^{(k+1)}$ obtained as the unique strong solution to the SDE:

$$\mathrm{d}Y_{1,2}^{(k+1)}(t) = b\Big(X_{1,2}^{(k)}(t), \gamma^{(k)}(t)\Big)(U_1, U_2)\,\mathrm{d}t + \Sigma\left(\gamma^{(k)}(t)\right)(U_1, U_2)\,\mathrm{d}B_{1,2}(t), \tag{D.6}$$

for $t \in \mathbb{R}_+$, with initial conditions $Y_{1,2}^{(k+1)}(0) = X_{1,2}^{(k+1)}(0) = w_0(U_1, U_2)$ for all $k \in \mathbb{Z}_+$.

Fix $t \in \mathbb{R}_+$ and consider $\sup_{s\in[0,t]}\Big|X_{1,2}^{(k+1)}(s) - X_{1,2}^{(k)}(s)\Big|$ for any $k \in \mathbb{N}$. Since the Skorokhod map is 4-Lipschitz in the local uniform norm (see Section 2.4), the above distance is bounded by $4\sup_{s\in[0,t]}\Big|Y_{1,2}^{(k+1)}(s) - Y_{1,2}^{(k)}(s)\Big|$. Now for every fixed $k \in \mathbb{N}$, from equation (D.6)

we have

$$\begin{aligned}
&Y_{1,2}^{(k+1)}(t) - Y_{1,2}^{(k)}(t) \\
&= \int_0^t \Big(b\Big(X_{1,2}^{(k-1)}(t), \gamma^{(k-1)}(t)\Big)(U_1, U_2) - b\Big(X_{1,2}^{(k)}(t), \gamma^{(k)}(t)\Big)(U_1, U_2)\Big)\,\mathrm{d}s \\
&\quad - \int_0^t \big(\Sigma\big(\gamma^{(k-1)}\big)(U_1, U_2) - \Sigma\big(\gamma^{(k)}\big)(U_1, U_2)\big)\,\mathrm{d}B_{1,2}(s).
\end{aligned} \tag{D.7}$$

Define $\Delta, M \colon \mathbb{R}_+ \to \mathbb{R}$ for $t \in \mathbb{R}_+$ as

$$\begin{aligned}
\Delta(t) &\coloneqq \int_0^t \Big(b\Big(X_{1,2}^{(k-1)}(t), \gamma^{(k-1)}(t)\Big)(U_1, U_2) - b\Big(X_{1,2}^{(k)}(t), \gamma^{(k)}(t)\Big)(U_1, U_2)\Big)\,\mathrm{d}s, \\
M(t) &\coloneqq \int_0^t \big(\Sigma\big(\gamma^{(k-1)}\big)(U_1, U_2) - \Sigma\big(\gamma^{(k)}\big)(U_1, U_2)\big)\,\mathrm{d}B_{1,2}(s).
\end{aligned}$$

Note that, for a kernel $A \in \mathcal{W}$, we have $\|A\|_2^2 = \mathbb{E}[A^2(U_1, U_2)]$, for U_1, U_2 i.i.d. as $\mathrm{Uni}[0,1]$. Using Jensen's inequality and interchanging expectation with integral and Assumption 6.1,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{s\in[0,t]} \Delta^2(s)\right] \\
&\leq t\,\mathbb{E}\left[\int_0^t \left|b\Big(X_{1,2}^{(k-1)}(t), \gamma^{(k-1)}(t)\Big)(U_1, U_2) - b\Big(X_{1,2}^{(k)}(t), \gamma^{(k)}(t)\Big)(U_1, U_2)\right|^2 \mathrm{d}s\right] \\
&= t\int_0^t \left\|b\Big(X_{1,2}^{(k-1)}(t), \gamma^{(k-1)}(t)\Big) - b\Big(X_{1,2}^{(k)}(t), \gamma^{(k)}(t)\Big)\right\|_2^2 \mathrm{d}s \\
&\leq 2\kappa^2 t\int_0^t \left\|\gamma^{(k-1)}(s) - \gamma^{(k)}(s)\right\|_2^2 \mathrm{d}s + 2L^2 t\int_0^t \mathbb{E}\Big[\left|X^{(k-1)}(s) - X^{(k)}(s)\right|^2\Big]\,\mathrm{d}s.
\end{aligned} \tag{D.8}$$

For M, we use the fact that it is a stochastic integral of a bounded integrand with respect to a Brownian motion, and hence a continuous martingale. By an application of Doob's maximal inequality [KS91, Theorem 3.8.iv, page 14], we get that,

$$\mathbb{E}\left[\sup_{s\in[0,t]} M^2(s)\right] \leq 4\int_0^t \mathbb{E}\Big[\left|\Sigma\big(\gamma^{(k-1)}(s)\big)(U_1, U_2) - \Sigma\big(\gamma^{(k)}(s)\big)(U_1, U_2)\right|^2\Big]\,\mathrm{d}s.$$

Using the assumption that Σ is κ_2-Lipschitz in $\|\cdot\|_2$ and the same argument as above,

$$\mathbb{E}\left[\sup_{s\in[0,t]} M^2(s)\right] \leq 4\kappa_2^2\int_0^t \left\|\gamma^{(k-1)}(s) - \gamma^{(k)}(s)\right\|_2^2 \mathrm{d}s. \tag{D.9}$$

Now, taking absolute values on both sides on (D.7), we immediately get,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{s\in[0,t]}\left|X_{1,2}^{(k+1)}(s)-X_{1,2}^{(k)}(s)\right|^2\right]\\
\leq&16\,\mathbb{E}\left[\sup_{s\in[0,t]}\left|Y_{1,2}^{(k+1)}(s)-Y_{1,2}^{(k)}(s)\right|^2\right]\leq 32\,\mathbb{E}\left[\sup_{s\in[0,t]}\Delta^2(s)+\sup_{s\in[0,t]}M^2(s)\right]\\
\leq&64(\kappa^2t+2\kappa_2^2)\int_0^t\left\|\gamma^{(k-1)}(s)-\gamma^{(k)}(s)\right\|_2^2\mathrm{d}s\\
&+64L^2t\int_0^t\mathbb{E}\Big[\left|X^{(k-1)}(s)-X^{(k)}(s)\right|^2\Big]\,\mathrm{d}s. \qquad \text{(D.10)}
\end{aligned}$$

Using the fact that the operator γ, given by a conditional expectation (D.3), and, therefore, must have a smaller L^2 norm

$$\sup_{s\in[0,t]}\left\|\gamma^{(k+1)}(s)-\gamma^{(k)}(s)\right\|_2^2\leq\mathbb{E}\left[\sup_{s\in[0,t]}\left|X_{1,2}^{(k+1)}(s)-X_{1,2}^{(k)}(s)\right|^2\right].$$

Combining the last two bounds above, one gets the recursive bound

$$\begin{aligned}
&\mathbb{E}\left[\sup_{s\in[0,t]}\left|X_{1,2}^{(k+1)}(s)-X_{1,2}^{(k)}(s)\right|^2+\sup_{s\in[0,t]}\left\|\gamma^{(k+1)}(s)-\gamma^{(k)}(s)\right\|_2^2\right]\\
&\leq 128((\kappa^2+L^2)t+4\kappa_2^2)\int_0^t\mathbb{E}\Big[\left|X^{(k-1)}(s)-X^{(k)}(s)\right|^2\Big]\,\mathrm{d}s.
\end{aligned}$$

The rest of the argument follows exactly as in [KS91, page 290] by applications of Grönwall's lemma [Grö19] and the Borel-Cantelli lemma [Kal21, Theorem 4.18]. We skip the similar argument for pathwise uniqueness. See the proof of [KS91, Proposition 2.13, page 291]. □

Proof of Theorem 6.3. Start with the countably many i.i.d. Uni$[0,1]$ random variables $(U_i)_{i\in\mathbb{N}}$ and an independent infinite (symmetric) array of i.i.d. standard Brownian motions $(B_{i,j})_{(i,j)\in\mathbb{N}^{(2)}}$ and construct the deterministic process γ in Proposition D.3.

Given γ and $(U_i)_{i\in\mathbb{N}}$ and following the system of SDEs (Graphon-MKV), the diffusions $X_{i,j}$s are independent (but not identically distributed) reflected Brownian motions with deterministic bounded time-dependent drifts for $(i,j)\in\mathbb{N}^{(2)}$. So, they exist in a pathwise or strong sense exactly as the process $X_{1,2}$ does in Proposition D.3 and satisfies the constraint (Graphon-MKV) since γ is a fixed point of the Picard iterations.

It is obvious from the symmetry of the construction that the infinite array $(X_{i,j})_{(i,j)\in\mathbb{N}^{(2)}}$ is exchangeable with $\mathcal{E} = C[0,\infty)$, the set of continuous functions from $[0,\infty)$ to $\mathbb{R}$.

For the limit (6.7) we will make use of the following result from [Lov12, Proposition 8.12], which states that for any $v \in \mathcal{W}$,

$$\|v\|_{\square}^4 \leq T_{C_4}(v) \leq 4\|v\|_{\square}. \tag{D.11}$$

Here C_4 is the cyclic graph with four vertices and $T_{C_4}(v)$ is the homomorphism density function of the simple graph C_4. We will apply this for the choice of $v_n(t) \coloneqq K\left((X_{i,j}(t))_{(i,j)\in[n]^2}\right) - K\Big((\gamma(t)(U_i,U_j))_{(i,j)\in[n]^2}\Big)$. Thus,

$$\begin{aligned} H_n(t) \coloneqq T_{C_4}(v_n(t)) &= \frac{1}{n^{\downarrow 4}} \sum_{i_1,i_2,\ldots,i_4} \prod_{l=1}^{4} \big(X_{i_l,i_{l+1}}(t) - \gamma(t)(U_{i_l},U_{i_{l+1}})\big) \\ &= \frac{1}{n^{\downarrow 4}} \sum_{i_1,i_2,\ldots,i_4} \prod_{l=1}^{4} \big(X_{i_l,i_{l+1}}(t) - \mathbb{E}\big[X_{i_l,i_{l+1}}(t) \,\big|\, \mathcal{F}_0\big]\big), \end{aligned}$$

with the convention that, when $l = 4$, $l+1 \equiv 1$. The above sum is over all injections in $[n]^{[4]}$.

Notice that $H_n(0) = 0$. The fact that for each $t \in \mathbb{R}_+$, $\lim_{n\to\infty} H_n(t) = 0$ almost surely follows similarly to the proof of Lemma 6.1. We now show that $t \mapsto H_n(t)$ is equicontinuous. From which, using a standard argument, we can show that almost surely, $H_n(t) \to 0$ for each $t \in \mathbb{R}_+$, that is,

$$\lim_{n\to\infty} \delta_{\square}\left(\left[K\Big((X_{i,j}(s))_{(i,j)\in[n]^{(2)}}\Big)\right], [\gamma(s)]\right) = 0, \qquad \text{a.s.} \quad \forall \;\; s \in [0,t].$$

To show that $(H_n)_{n\in\mathbb{N}}$ is equicontinuous, we first observe that for any $s_1, s_2 \in [0,t]$,

$$\begin{aligned} &|H_n(s_2) - H_n(s_1)| \\ &\leq 16\left\|K\Big((X_{i,j}(s_2))_{(i,j)\in[n]^{(2)}}\Big) - K\Big((X_{i,j}(s_1))_{(i,j)\in[n]^{(2)}}\Big)\right\|_2 \\ &\qquad\qquad + 16\|\gamma(s_2) - \gamma(s_1)\|_2, \end{aligned} \tag{D.12}$$

where the inequality follows by an application of the counting lemma [Lov12, Lemma 10.23, Exercise 10.27], the triangle inequality and using the fact that the cut norm $\|\cdot\|_{\square}$ is upper bounded by the L^2 norm $\|\cdot\|_2$.

Using the Lipschitzness of the Skorokhod map (see equation (2.6)), we therefore obtain

$$\begin{aligned}
&\left\|K\Big((X_{i,j}(s_2))_{(i,j)\in[n]^{(2)}}\Big) - K\Big((X_{i,j}(s_1))_{(i,j)\in[n]^{(2)}}\Big)\right\|_2^2 \\
&\leq \frac{2^4}{n^2}\sum_{(i,j)\in[n]^{(2)}} |Y_{i,j}(s_2) - Y_{i,j}(s_1)|^2 \\
&\leq \frac{2^5}{n^2}\sum_{(i,j)\in[n]^{(2)}}\left|\int_{s_1}^{s_2} b(X_{1,j}(u),\gamma(u))(U_i,U_j)\,\mathrm{d}u\right|^2 \\
&\qquad + \frac{2^5}{n^2}\sum_{(i,j)\in[n]^{(2)}}\left|\int_{s_1}^{s_2}\Sigma(\gamma(u))(U_i,U_j)\,\mathrm{d}B_{i,j}(u)\right|^2 \\
&\leq 2^5 M_\infty^2|s_2-s_1|^2 + \frac{2^5}{n^2}\sum_{(i,j)\in[n]^2}\left|\int_{s_1}^{s_2}\Sigma(\gamma(u))(U_i,U_j)\,\mathrm{d}B_{i,j}(u)\right|^2. \qquad \text{(D.13)}
\end{aligned}$$

Now let $|s_2 - s_1| \leq \delta$ for some $\delta > 0$. Set for all $(i,j) \in [n]^{(2)}$,

$$\eta_{i,j} \coloneqq \sup_{\substack{s_1,s_2\in[0,t],\\ |s_2-s_1|\leq\delta}}\left|\int_{s_1}^{s_2}\Sigma(\gamma(u))(U_i,U_j)\,\mathrm{d}B_{i,j}(u)\right|^2.$$

From [Sło01, Lemma A.4], there exist constants $C_{1,t}, C_{2,t} \in \mathbb{R}_+$ depending of t, such that for all $(i,j) \in [n]^{(2)}$,

$$\mathbb{E}[\eta_{i,j}] \leq M_\infty^2 C_{1,t}\delta\left|\log\frac{1}{\delta}\right|, \qquad \text{and} \qquad \mathbb{E}\big[\eta_{i,j}^2\big] \leq M_\infty^4 C_{2,t}^2\delta^2\log^2\frac{1}{\delta}. \qquad \text{(D.14)}$$

Since, $\eta_{i,j}$s are independent and have finite variance, it follows from the Chebyshev's inequality [Kal21, Lemma 5.1] that

$$\mathbb{P}\left\{\left|\frac{1}{n^2}\sum_{(i,j)\in[n]^{(2)}}\eta_{i,j} - \mathbb{E}[\eta_{i,j}]\right| \geq \max_{(i,j)\in[n]^{(2)}}\mathrm{Var}^{1/2}(\eta_{i,j})\right\} \leq \frac{1}{n^2}.$$

Using the Borel-Cantelli lemma [Kal21, Theorem 4.18], it follows that almost surely,

$$\frac{1}{n^2}\sum_{(i,j)\in[n]^{(2)}}\eta_{i,j} \leq M_\infty^2(C_{1,t}+C_{2,t})\delta\left|\log\frac{1}{\delta}\right|, \qquad \text{(D.15)}$$

for all $n \in \mathbb{N}$, sufficiently large. Combining equations (D.12) and (D.15), we obtain that almost surely, for all $n \in \mathbb{N}$ sufficiently large, we have

$$\sup_{\substack{s_1,s_2\in[0,t],\\ |s_2-s_1|\leq\delta}}|H_n(s_2) - H_n(s_1)| \leq 2^8 M_\infty\left(\delta + (C_{1,t}+C_{2,t})^{1/2}\delta^{1/2}\log^{1/2}\frac{1}{\delta}\right) + 16\omega(\delta),$$

where $\omega(\delta) \coloneqq \sup_{s_1,s_2\in[0,t],|s_2-s_1|\le\delta}\|\gamma(s_2)-\gamma(s_1)\|_2$ is the modulus of continuity of the curve $t \mapsto \gamma(t)$. Since $s \mapsto \gamma(s)$ is continuous in $(\mathcal{W}, d_2)$ (and independent of n), it follows that, almost surely, $(H_n)_{n\in\mathbb{N}}$ is equicontinuous. Since $(H_n)_{n\in\mathbb{N}}$ is equicontinuous uniformly bounded almost surely, the proof is complete by a standard application of Arzelà-Ascoli theorem [Mun00, Theorem 47.1]. □

Proof of Proposition 6.4. Given $(U_1, U_2) = (x, y)$, the process $X_{1,2}$ is a diffusion with a Lipschitz drift and a constant diffusion coefficient. Using (Graphon-MKV) and Itô's formula, we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\gamma(t)(x,y) = &-\frac{\mathrm{d}}{\mathrm{d}t}\phi(\gamma(t))(x,y)\\ &+\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}\big[L^-_{1,2}(t)\,\big|\,U_1=x,U_2=y\big]-\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}\big[L^+_{1,2}(t)\,\big|\,U_1=x,U_2=y\big].\end{aligned} \tag{D.16}$$

Now consider the reflecting diffusion Z which solves the SDE

$$\mathrm{d}Z(s) = \Psi(s;\beta)\,\mathrm{d}s + \mathrm{d}B(s) + \mathrm{d}L^-(s) - \mathrm{d}L^+(s), \qquad s\in\mathbb{R}_+, \tag{D.17}$$

starting at $Z(0) = w_0(x, y)$, such that (Z, L^+, L^-) solves the Skorokhod problem with respect to the set $[-1, 1]$, and $\Psi(s;\beta) \coloneqq -\frac{1}{\beta^2}b(\gamma(s/\beta^2))(x, y)$ for all $s \in \mathbb{R}_+$ (see Section 2.4). By reparametrizing $s = \beta^2 t$ and setting $Z(s) = X_{1,2}(t)$, we get back our reflected diffusion $X_{1,2}$ in law following

$$\begin{aligned}\mathrm{d}Z(\beta^2t) &= -\frac{1}{\beta^2}\phi(\gamma(t))(x,y)\,\mathrm{d}(\beta^2t) + \mathrm{d}B(\beta^2t) + \mathrm{d}L^-(\beta^2t) - \mathrm{d}L^+(\beta^2t),\\ \implies X_{1,2}(t) &= -\phi(\gamma(t))\,\mathrm{d}t + \beta\,\mathrm{d}B(t) + \mathrm{d}L^-(\beta^2t) - \mathrm{d}L^+(\beta^2t), \qquad t\in\mathbb{R}_+,\end{aligned}$$

where the processes $(L^+(\beta^2t))_{t\in\mathbb{R}_+}$ and $(L^-(\beta^2t))_{t\in\mathbb{R}_+}$ constrain the process $X_{1,2}$ in the interval $[-1, 1]$ (see Section 2.4). Here the equality is in law. We use the fact that the solution of both the above SDEs agree in law since the distribution of $B(\beta^2t)$ and $\beta B(t)$ coincide for all $\beta \in \mathbb{R}_+$. Let $p_s^{(\pm1)}(w_0, \phi\circ\gamma, \beta)(x, y)$ denote the transition density of the solution of SDE (D.17) at time $s \in \mathbb{R}_+$ at the boundary ±1, then the transition density of the process $X_{1,2}$ at time t at the boundary ±1 is $p_{\beta^2t}^{(\pm1)}(w_0, \phi\circ\gamma, \beta)(x, y)$.

Using [RY04, Exercise (1.12), page 407] and equation (D.16), we deduce that

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}\big[L^{\pm}_{i,j}(t)\big] = p^{(\pm 1)}_{\beta^2 t}(w_0, b\circ(X_{1,2},\gamma),\beta)(x,y), \tag{D.18}$$

which gives us the desired result. □

Proof of Theorem 6.6. Consider a probability space satisfying the assumptions of Proposition 6.3 and an infinite exchangeable array of diffusions $(X_{i,j})_{(i,j)\in\mathbb{N}^{(2)}}$ on it. For $k \in [n]$ and any $t \in \mathbb{R}_+$, consider the sampled $k \times k$ symmetric matrix $\gamma(t)[k]$ whose (i,j)-th element is $\gamma(t)(U_i, U_j)$, $(i,j) \in [k]^{(2)}$. Consider also the corresponding $k \times k$ matrix of diffusions $X^{(k)}(\cdot) \coloneqq \big(X_{(i,j)}\big)_{(i,j)\in[k]^{(2)}}$.

Now consider $K(X_n(t))$ from a solution of SDEs (6.1). One may construct a sampled $k \times k$ matrix from this kernel as well. We estimate the cut distance of this sampled matrix from $\gamma(t)[k]$ by coupling this sampled matrix with $K\big(X^{(k)}\big)$ in a particular way.

Notice that, for any $(i,j) \in [k]^{(2)}$ and $(m_i, m_j) \in [n]^{(2)}$, if $U_i \in ((m_i-1)/n, m_i/n]$ and $U_j \in ((m_j-1)/n, m_j/n]$, then $K(X_n(t))(U_i,U_j) \equiv X_{n,m_i,m_j}(t)$. Let $E_k(n)$ denote the event that that no two $U_i, U_{i'}$, for distinct $i, i' \in [k]^{(2)}$, falls in the same interval $((m-1)/n, m/n]$. Under this event every entry of the sampled diffusions will be run by independent standard Brownian motions. Before we use this property to proceed with our coupling, let us show that $E_k(n)$ happens with high probability as k is fixed and $n \to \infty$. Order the uniform random variables as $U_{(1)} < U_{(2)} < \ldots < U_{(k)}$. Clearly $E^c_k(n)$ implies that there is at least one pair $(U_{(i)}, U_{(i+1)})$ for $i \in [k-1]$, such that $U_{(i+1)} - U_{(i)} \le 1/n$. Hence $\mathbb{P}\{E^c_k(n)\} \le \mathbb{P}\big\{\min_{i\in[k-1]}\big(U_{(i+1)} - U_{(i)}\big) \le \frac{1}{n}\big\}$. But $\min_{i\in[k-1]}\big(U_{(i+1)} - U_{(i)}\big)$ has a density at zero and hence the above probability is $O(1/n)$, which goes to zero as $n \to \infty$. Thus $\lim_{k\to\infty}\lim_{n\to\infty}\mathbb{P}\{E_k(n)\} = 1$.

On the event $E_k(n)$, every m_i, $i \in [k]$, is distinct. Consider the corresponding independent Brownian motion $B_{i,j}$ from the diffusion $X_{i,j}$ from equation (Graphon-MKV). Since (6.1) admits a strong solution, construct a solution where the entry processes $X_{n,m_i,m_j}(\cdot)$ is driven by $B_{i,j}$, $(i,j) \in [k]^{(2)}$, while the rest of the entries of X_n are driven by a disjoint subset of $(B_{i,j})_{(i,j)\in\mathbb{N}^2}$. Thus, one couples $K(X_n)(\cdot)(U_i,U_j)$ with $X_{i,j}$ which are both driven by the same

Brownian motion and having a starting value of $w_0^{(n)}(U_i, U_j)$ and $w_0(U_i, U_j)$, respectively. Our subsequent analysis will be on the event $E_k(n)$ and it is unimportant how the coupling is done on $E_k^c(n)$.

Define, $\widetilde{X}_{n,i,j}(t) \coloneqq K(X_n(t))(U_i, U_j),\ (i,j) \in [k]^2$. The evolution of $\widetilde{X}_{n,1,2}$, for example, can be described by the SDE

$$\begin{aligned}\mathrm{d}\widetilde{X}_{n,1,2}(t) = b\Big(\widetilde{X}_{n,1,2}(t), K(X_n(t))\Big)(U_1, U_2)\,\mathrm{d}t + \Sigma(K(X_n(t)))(U_1, U_2)\,\mathrm{d}B_{1,2}(t)\\ + \mathrm{d}L^-_{n,1,2}(t) - \mathrm{d}L^+_{n,1,2}(t),\end{aligned}$$

with the initial condition $\widetilde{X}_{n,1,2}(0) = w_0^{(n)}(U_1, U_2)$. Since $X_{1,2}$ is also driven by the same Brownian motion, by using the Lipschitz property of the Skorokhod map and the triangle inequality, it follows that for any $(U_1, U_2) = (u_1, u_2)$ on the event $E_k(n)$, $\sup_{s\in[0,t]}\Big|\widetilde{X}_{n,1,2}(s) - X_{1,2}(s)\Big|^2$ is at most

$$\begin{aligned}48\int_0^t \Big|b(X_{1,2}(s), \gamma(s))(u_1, u_2) - b\Big(\widetilde{X}_{n,1,2}(s), K(X_n(s))\Big)(u_1, u_2)\Big|^2\,\mathrm{d}s\\ + 48\sup_{s\in[0,t]}\left|\int_0^s (\Sigma(\gamma(r))(u_1, u_2) - \Sigma(K(X_n(r)))(u_1, u_2))\,\mathrm{d}B_{1,2}(r)\right|^2\\ + 48\Big|\widetilde{X}_{n,1,2}(0) - X_{1,2}(0)\Big|^2.\end{aligned} \tag{D.19}$$

We can now use Assumption 6.1 and 6.2 on the first term in (D.19) to get

$$\begin{aligned}&\Big|b(X_{1,2}(s), \gamma(s))(u_1, u_2) - b\Big(\widetilde{X}_{n,1,2}(s), K(X_n(t))\Big)(u_1, u_2)\Big|^2\\ &\le 2L^2\Big|X_{1,2}(s) - \widetilde{X}_{n,1,2}(s)\Big|^2 + 2\kappa_\square^2\|\gamma(s) - K(X_n(s))\|_\square^2, \qquad s \in \mathbb{R}_+.\end{aligned} \tag{D.20}$$

Define for $s \in [0,t]$,

$$M^{(n)}(s) \coloneqq \int_0^s (\Sigma(\gamma(r))(u_1, u_2) - \Sigma(K(X_n(r)))(u_1, u_2))\,\mathrm{d}B_{1,2}(r),$$

which makes the second term in (D.19) equal to $48\sup_{s\in[0,t]} M^2(s)$. Using Markov's inequality followed by Doob's maximal inequality [KS91, page 14, Theorem 3.8.iv], we obtain

$$\mathbb{P}\left\{\sup_{s\in[0,t]} M^{(n)}(s)^2 \ge 2\lambda_k\,\mathbb{E}\big[M^{(n)}(t)^2\big]\right\} \le \big(2\lambda_k\,\mathbb{E}\big[M^{(n)}(t)^2\big]\big)^{-1}\,\mathbb{E}\left[\sup_{s\in[0,t]} M^{(n)}(s)^2\right]$$

$$\leq \left(2\lambda_k \, \mathbb{E}\big[M^{(n)}(t)^2\big]\right)^{-1} \mathbb{E}\big[M^{(n)}(t)^2\big] = 2\lambda_k^{-1}, \quad \text{(D.21)}$$

for every $\lambda_k > 0$. Let $(\lambda_k)_{k\in\mathbb{N}}$ satisfy $\lim_{k\to\infty} \lambda_k = \infty$. The choice of λ_k will be made later.

Therefore, with probability at least $1 - 2\lambda_k^{-1}$,

$$\begin{aligned}
\sup_{s\in[0,t]} M^{(n)}(s)^2 &\leq 2\lambda_k \, \mathbb{E}\big[M^{(n)}(t)^2\big] \\
&= 2\lambda_k \int_0^t \left|\Sigma(\gamma(s))(u_1,u_2) - \Sigma(K(X_n(s)))(u_1,u_2)\right|^2 \mathrm{d}s \qquad \text{(D.22)} \\
&\leq 2\lambda_k \kappa_\square^2 \int_0^t \|\gamma(s) - K(X_n(s))\|_\square^2 \, \mathrm{d}s.
\end{aligned}$$

By the abuse of notation, we redefine the event $E_k(n)$ to intersect with the event where the above bound holds. By a union bound, we still have $\lim_{k\to\infty} \lim_{n\to\infty} \mathbb{P}\{E_k(n)\} = 1$.

Using equations (D.20) and (D.22) in equation (D.19) we get

$$\begin{aligned}
\sup_{s\in[0,t]} \left|\widetilde{X}_{n,1,2}(s) - X_{1,2}(s)\right|^2 &\leq 48 \left|w_0^{(n)}(U_1,U_2) - w_0(U_1,U_2)\right|^2 \\
&+ 96\kappa_\square^2(\lambda_k+1) \int_0^t \|\gamma(s) - K(X_n(s))\|_\square^2 \, \mathrm{d}s \qquad \text{(D.23)} \\
&+ 96L^2 \int_0^t \left|X_{1,2}(s) - \widetilde{X}_{n,1,2}(s)\right|^2 \mathrm{d}s.
\end{aligned}$$

Replacing the role of $(1,2)$ by any other $(i,j) \in [k]^{(2)}$, and summing over, we get

$$\begin{aligned}
&\sup_{s\in[0,t]} \frac{1}{k^2} \sum_{(i,j)\in[k]^{(2)}} \left|\widetilde{X}_{n,i,j}(s) - X_{i,j}(s)\right|^2 \\
&\quad\leq \frac{48}{k^2} \sum_{(i,j)\in[k]^{(2)}} \left|w_0^{(n)}(U_i,U_j) - w_0(U_i,U_j)\right|^2 \\
&\qquad + 96\kappa_\square^2(\lambda_k+1) \int_0^t \|\gamma(s) - K(X_n(s))\|_\square^2 \, \mathrm{d}s \qquad \text{(D.24)} \\
&\qquad + 96L^2 \int_0^t \frac{1}{k^2} \sum_{(i,j)\in[k]^{(2)}} \left|X_{i,j}(s) - \widetilde{X}_{n,i,j}(s)\right|^2 \mathrm{d}s.
\end{aligned}$$

By the triangle inequality,

$$\begin{aligned}
&\sup_{s\in[0,t]}\left\|K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right)-K\left(\left(\gamma(s)(U_i,U_j)\right)_{(i,j)\in[k]^{(2)}}\right)\right\|_{\square}^2 \\
&\quad\le 2\sup_{s\in[0,t]}\left\|K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right)-K\left(\left(X_{i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right)\right\|_{\square}^2 \\
&\qquad+2\sup_{s\in[0,t]}\left\|K\left(\left(\gamma(s)(U_i,U_j)\right)_{(i,j)\in[k]^{(2)}}\right)-K\left(\left(X_{i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right)\right\|_{\square}^2.
\end{aligned}\tag{D.25}$$

Then notice that the kernel

$$\frac{1}{2}K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right)-\frac{1}{2}K\left(\left(\gamma(s)(U_i,U_j)\right)_{(i,j)\in[k]^{(2)}}\right)$$

has entries in $[-1,1]$ and is sampled from the kernel $\frac{1}{2}K(X_n(s))-\frac{1}{2}\gamma(s)$. By [Lov12, Lemma 10.6], the difference

$$\left\|K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right)-K\left(\left(\gamma(s)(U_i,U_j)\right)_{(i,j)\in[k]^{(2)}}\right)\right\|_{\square}^2-\|K(X_n(s))-\gamma(s)\|_{\square}^2$$

lies in the interval $\left[-24/k-36/k^2,64k^{-1/4}+256k^{-1/2}\right]$ with probability at least $1-4\mathrm{e}^{-k^{1/2}/10}$, for all $n\ge k$. Using this in (D.25) we get

$$\begin{aligned}
&\sup_{s\in[0,t]}\left\|K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right)-K\left(\left(X_{i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right)\right\|_{\square}^2 \\
&\quad\ge\frac{1}{2}\|K(X_n(s))-\gamma(s)\|_{\square}^2-320k^{-1/4} \\
&\qquad-\sup_{s\in[0,t]}\left\|K\left(\left(\gamma(s)(U_i,U_j)\right)_{(i,j)\in[k]^{(2)}}\right)-K\left(\left(X_{i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right)\right\|_{\square}^2.
\end{aligned}\tag{D.26}$$

with probability at least $1-4\mathrm{e}^{-k^{1/2}/10}$. By an abuse of notation, we redefine the event $E_k(n)$ to intersect with the event where the above bound holds. We still have $\lim_{k\to\infty}\lim_{n\to\infty}\mathbb{P}\{E_k(n)\}=1$.

We first lower bound twice the left hand side of equation (D.24) using equation (D.26)

as

$$
\begin{aligned}
2 \sup_{s\in[0,t]} &\left\| K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_2^2 \\
&\geq \sup_{s\in[0,t]} \left\| K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_2^2 \\
&\quad + \sup_{s\in[0,t]} \left\| K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_\square^2 \\
&\geq \sup_{s\in[0,t]} \left\| K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_2^2 \\
&\quad + \frac{1}{2}\|K(X_n(s)) - \gamma(s)\|_\square^2 - 320k^{-1/4} \\
&\quad - \sup_{s\in[0,t]} \left\| K\left((\gamma(s)(U_i,U_j))_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_\square^2.
\end{aligned}
\tag{D.27}
$$

Here we used the fact that the L^2 norm is lower bounded by the cut norm. Using equation (D.27) back in equation (D.24) (multiplied by 2), and rearranging terms we get

$$
\begin{aligned}
&\sup_{s\in[0,t]} \left\| K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_2^2 \\
&\qquad + \frac{1}{2}\sup_{s\in[0,t]} \|K(X_n(s)) - \gamma(s)\|_\square^2 \\
&\leq \sup_{s\in[0,t]} \left\| K\left((\gamma(s)(U_i,U_j))_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_\square^2 \\
&\qquad + 320k^{-1/4} + \frac{96}{k^2}\sum_{(i,j)\in[k]^{(2)}} \left| w_0^{(n)}(U_i,U_j) - w_0(U_i,U_j)\right|^2 \\
&\qquad + 192L^2 \int_0^t \left\| K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_2^2 \mathrm{d}s \\
&\qquad + 192\kappa_\square^2(\lambda_k+1)\int_0^t \|\gamma(s) - K(X_n(s))\|_\square^2 \,\mathrm{d}s.
\end{aligned}
\tag{D.28}
$$

Now let

$$
\begin{aligned}
A_k &\coloneqq \sup_{s\in[0,t]} \left\| K\left((\gamma(s)(U_i,U_j))_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_\square^2, \\
B_k(n) &\coloneqq \frac{96}{k^2}\sum_{(i,j)\in[k]^{(2)}} \left| w_0^{(n)}(U_i,U_j) - w_0(U_i,U_j)\right|^2 + 320k^{-1/4}.
\end{aligned}
$$

Applying Grönwall's inequality [Grö19] and noticing that the first term on the left of equation (D.28) is always non-negative, gives us that on the event $E_k(n)$,

$$\begin{aligned}\sup_{s\in[0,t]}\left\|K\Big(\big(\widetilde{X}_{n,i,j}(s)\big)_{(i,j)\in[k]^{(2)}}\Big)-K\Big(\big(X_{i,j}(s)\big)_{(i,j)\in[k]^{(2)}}\Big)\right\|_2^2 \\ +\sup_{s\in[0,t]}\|K(X_n(s))-\gamma(s)\|_\square^2\le 2\,(A_k+B_k(n))\exp\big(192(L^2+2\kappa_\square^2(\lambda_k+1))t\big),\end{aligned} \tag{D.29}$$

for every $n\ge k$. Note that

$$\mathbb{E}\left[\left|w_0^{(n)}(U_i,U_j)-w_0(U_i,U_j)\right|^2\right]=\left\|w_0^{(n)}-w_0\right\|_2^2\to 0,$$

as $n\to\infty$, by assumption (6.9). By a variance bound it follows that

$$\lim_{k\to\infty}\lim_{n\to\infty}B_k(n)=0,$$

in probability. Also, $\lim_{k\to\infty}A_k=0$ by Proposition 6.3. Since $\lim_{k\to\infty}\lim_{n\to\infty}\mathbb{P}\{E_k(n)\}=1$,

$$\begin{aligned}\lim_{n\to\infty}\sup_{s\in[0,t]}\|K(X_n(s))-\gamma(s)\|_\square=0,\qquad\text{and}\\ \lim_{k\to\infty}\lim_{n\to\infty}\sup_{s\in[0,t]}\frac{1}{k^2}\left\|\big(K(X_n(s))(U_i,U_j)\big)_{(i,j)\in[k]^{(2)}}-\big(X_{i,j}(s)\big)_{(i,j)\in[k]^{(2)}}\right\|_{\mathrm{F}}^2=0,\end{aligned}$$

in probability, by choosing $(\lambda_k)_{k\in\mathbb{N}}$ (depending on $(A_k,\lim_{n\to\infty}B_k(n))_{k\in\mathbb{N}}$) that increases sufficiently slowly to infinity as $k\to\infty$. This proves our claim.

Remark D.4. To get a non-asymptotic error rate, we need to control on A_k and $B_k(n)$. Observe that $B_k(n)$ depends on the initial condition and in general it can be arbitrarily slow. However, assuming that the initial condition is i.i.d., one can use Chebyshev's inequality to obtain $\mathbb{P}\big\{B_k(n)\ge 66k^{-1/4}\big\}\le k^{-3/2}$.

On the other hand, it follows from the arguments in Proposition 6.3 that there exists a constant M_t (depending only on t) such that for any $\delta>0$ we have $\mathbb{P}\big\{A_k\ge M_t(\delta\log(1/\delta))^{1/4}\big\}\le k^{-2}+t\delta^{-1}\mathrm{e}^{\frac{128}{\delta\log(1/\delta)}}\mathrm{e}^{-k\delta\log(1/\delta)/2}$.

In particular, choosing $\delta=64\sqrt{k^{-1}\log k}$ and $\lambda_k=\log(k)/(16\cdot 384t(L^2+2\kappa_\square^2))$, we have the left hand side of (D.29) bounded by $M_tk^{-1/16}\log^{3/2}k$ with probability at least

$1 - \frac{k^2}{n} - 4k^{-\frac{1}{\kappa^2 t}} - 2te^{-\sqrt{k}/20} - 2k^{-3/2}$, where $\kappa = 32\sqrt{6}(L^2 + 2\kappa_{\square}^2)^{1/2}$. Since t is fixed, we can choose k to be a suitable function of n, say $k = n^{2/7}$, to get a non-asymptotic rate of convergence. Moreover, using the remark after the proof of Lemma A.2, we can get a non-asymptotic rate of convergence with finite n and $|\boldsymbol{\tau}_n|$.

□

D.2 Proofs of Chapter 6.2

Proof of Theorem 6.11. Consider the probability space satisfying Assumption of Proposition 6.10 and an infinite exchangeable array of diffusions $(X_{i,j})_{(i,j)\in\mathbb{N}^{(2)}}$ on it. For $k \in [n]$ and any $t \in \mathbb{R}_+$, consider the sampled $k \times k$ symmetric measure-valued matrix $\Gamma(t)[k]$ defined as $\Gamma(t)[k](i,j) = \Gamma(t)(U_i, U_j)$ for $(i,j) \in [k]^{(2)}$. Consider also the corresponding $k \times k$ matrix of diffusions $X[k](\cdot) \coloneqq (X_e)_{e\in[k]^{(2)}}$. Now consider $\mathcal{K}(X_n(t))$, the measure-valued finite dimensional kernel from a solution of SDE (6.1). One may construct a sampled $k \times k$ measure-valued matrix from this measure-valued finite dimensional kernel as well. We estimate the cut distance of this sampled measure-valued matrix from $\Gamma(t)[k]$ by coupling this sampled matrix with $\mathcal{K}(X[k])$ in a particular way.

Divide $[0,1]$ into n contiguous intervals of equal length. Let $E_k(n)$ denote the event that that $U_i \in ((m_i - 1)/n, m_i/n]$ where each m_i, $i \in [k]$, is distinct. On this event, we can couple $X_{n,m_i,m_j}(\cdot)$ and $X_{i,j}$ so that they are driven by the same copies of independent Brownian motion and having starting laws $W_0^{(n)}(U_i, U_j)$ and $W_0(U_i, U_j)$ respectively. Our subsequent analysis will be on the event $E_k(n)$ and it is unimportant how the coupling is done on $E_k^c(n)$. For any $i \neq j$ we have $\mathbb{P}\{|U_i - U_j|\} \leq \frac{1}{n}$. Since there are at most $\binom{k}{2}$ distinct pairs $(i,j) \in [k]^2$, a simple union bound yields that $\mathbb{P}\{E_k^c(n)\} \leq k^2/n$.

Define, $\widetilde{X}_{n,(i,j)}(t) \coloneqq K(X_n(t))(U_i, U_j),\ (i,j) \in [k]^2$. The evolution of $\widetilde{X}_{n,(1,2)}$, for example,

can be described by the SDE

$$\begin{aligned}\mathrm{d}\widetilde{X}_{n,(1,2)}(t) = b\Big(\widetilde{X}_{n,(1,2)}(t), \mathcal{K}(X_n(t))\Big)(U_1,U_2)\,\mathrm{d}t \\ + \Sigma\Big(\widetilde{X}_{n,(1,2)}(t), \mathcal{K}(X_n(t))\Big)(U_1,U_2)\,\mathrm{d}B_{(1,2)}(t) \\ + \mathrm{d}L^-_{n,(1,2)}(t) - \mathrm{d}L^+_{n,(1,2)}(t),\end{aligned}$$

with the initial condition $\mathrm{Law}\Big(\widetilde{X}_{n,(1,2)}(0)\Big) = W_0^{(n)}(U_1,U_2)$. Define

$$M^{(n)}(s) \coloneqq \int_0^s \Big(\Sigma(X_{(1,2)}(s),\Gamma(r))(u_1,u_2) - \Sigma\Big(\widetilde{X}_{n,(1,2)}(s), \mathcal{K}(X_n(r))\Big)(u_1,u_2)\Big)\,\mathrm{d}B_{(1,2)}(r),$$

for $s \in [0,t]$. Note that

$$\mathbb{P}\left\{\sup_{s\in[0,t]} M^{(n)}(s) \geq \sqrt{\lambda_k\,\mathbb{E}[M^{(n)}(t)^2]}\right\} = \mathbb{P}\left\{\sup_{s\in[0,t]} \exp\big(uM^{(n)}(s)\big) \geq \exp(\lambda_k)\right\},$$

where $u = \sqrt{\lambda_k/\,\mathbb{E}[M^{(n)}(t)^2]}$. Using Markov's inequality followed by Doob's maximal inequality [KS91, page 14, Thoerem 3.8.iv], we obtain that with probability at least $1-4\mathrm{e}^{-\lambda_k/2}$,

$$\sup_{s\in[0,t]} M^{(n)}(s)^2 \leq 2\lambda_k\left[\kappa_{\blacksquare}^2\int_0^t \|\Gamma(s)-\mathcal{K}(X_n(s))\|_{\blacksquare}^2 + L^2\Big|X_{(1,2)}(s)-\widetilde{X}_{n,(1,2)}(s)\Big|^2\,\mathrm{d}s\right], \quad \text{(D.30)}$$

where the parameter $\lambda_k \to \infty$ will be chosen later. Redefining the event $E_k(n)$ to intersect with the event where the above bound holds. Since $X_{(1,2)}$ is also driven by the same Brownian motion on this event, using (D.30) and the Lipschitz property of the Skorokhod map, triangle inequality and Assumption 6.3 (replacing $(1,2)$ by any other $e \in [k]^{(2)}$ and summing over),

$$\begin{aligned}\sup_{s\in[0,t]}\Big\|\mathcal{K}\Big(\widetilde{X}_n[k](s)\Big) - \mathcal{K}(X[k](s))\Big\|_{\blacksquare}^2 \leq \frac{48}{k^2}\sum_{e\in[k]^{(2)}}\Big|\widetilde{X}_{n,e}(0)-X_e(0)\Big|^2 \\ + 96(\lambda_k+1)\kappa_{\blacksquare}^2\int_0^t \|\Gamma(s)-\mathcal{K}(X_n(s))\|_{\blacksquare}^2\,\mathrm{d}s \\ + 96(\lambda_k+1)L^2\int_0^t \sup_{s\in[0,t]}\Big\|\mathcal{K}\Big(\widetilde{X}_n[k](s)\Big) - \mathcal{K}(X[k](s))\Big\|_{\blacksquare}^2\,\mathrm{d}s.\end{aligned} \quad \text{(D.31)}$$

We now want to replace $\Big\|\mathcal{K}\Big(\widetilde{X}_n[k](s)\Big) - \mathcal{K}(\Gamma(s)[k])\Big\|_{\blacksquare}^2$ by $\|\mathcal{K}(X_n(s)) - \Gamma(s)\|_{\blacksquare}^2$ up to some error that goes to zero as $k \to \infty$. This is achieved by exploiting the first sampling

lemma [Lov12, Lemma 10.6] for cut norm. The first sampling lemma is not available directly to us for the $\|\cdot\|_{\blacksquare}$. However, we notice that using the first sampling lemma [Lov12, Lemma 10.6] and equation (C.1) we obtain that for every $\epsilon > 0$ there exists a constant $F_\epsilon < \infty$ such that

$$\left| \left\| \mathcal{K}\Big(\widetilde{X}_n[k](s)\Big) - \mathcal{K}(\Gamma(s)[k]) \right\|_{\blacksquare}^2 - \|\mathcal{K}(X_n(s)) - \Gamma(s)\|_{\blacksquare}^2 \right| \leq \frac{1}{k^{1/4}} + \epsilon,$$

with probability at least $F_\epsilon \mathrm{e}^{-\sqrt{k}/10}$. Moreover, from Lemma C.4, we can choose $\epsilon_k = \frac{64}{k^{1/4}}$ so that $F_{\epsilon_k} \leq \mathrm{e}^{\sqrt{k}/40}$. In particular, setting $C_k = \frac{65}{k^{1/4}}$ and $c_k = \sqrt{k}/20$ we can repeat the same proof as in Theorem 6.6 to obtain

$$\begin{aligned} \sup_{s\in[0,t]} \left\| \mathcal{K}\Big(\widetilde{X}_n[k](s)\Big) - \mathcal{K}(X[k](s)) \right\|_{\blacksquare}^2 &\geq \frac{1}{2}\|\mathcal{K}(X_n(s)) - \Gamma(s)\|_{\blacksquare}^2 - C_k \\ &\quad - \sup_{s\in[0,t]} \|\mathcal{K}(\Gamma(s)[k]) - \mathcal{K}(X[k](s))\|_{\blacksquare}^2. \end{aligned} \tag{D.32}$$

with probability at least $1 - \mathrm{e}^{-c_k}$. Once again we redefine the event $E_k(n)$ to intersect with the event where the above bound holds and note that we still have $\mathbb{P}\{E_k(n)\} \geq 1 - 4\mathrm{e}^{-\lambda_k/2} - \frac{k^2}{n} - \mathrm{e}^{-c_k}$.

We can now repeat the same argument as in the proof of Theorem 6.6. After doing some rearrangement and applying Grönwall's inequality [Grö19] we obtain that on the event $E_k(n)$,

$$\begin{aligned} &\sup_{s\in[0,t]} D_2^2\Big(\mathcal{K}\Big(\widetilde{X}_n[k](s)\Big), \mathcal{K}(X[k](s))\Big) + \sup_{s\in[0,t]} \|\mathcal{K}(X_n(s)) - \Gamma(s)\|_{\blacksquare}^2 \\ &\qquad \leq 2\left(A_k + B_k(n)\right) \exp\Big(192(L^2 + 2\kappa_{\blacksquare}^2)(\lambda_k + 1)t\Big), \end{aligned} \tag{D.33}$$

where $A_k = \sup_{s\in[0,t]} \left\| \widetilde{A}_k(s) \right\|_{\blacksquare}^2$ and

$$\begin{aligned} \widetilde{A}_k(s) &\coloneqq \mathcal{K}(\Gamma(s)[k]) - \mathcal{K}(X[k](s)), \\ B_k(n) &\coloneqq C_k + \frac{96}{k^2} \sum_{e\in[k]^{(2)}} \left| \widetilde{X}_{n,e}(0) - X_e(0) \right|^2. \end{aligned} \tag{D.34}$$

Note that $\lim_{n\to\infty} \mathbb{E}\left[\left| \widetilde{X}_{n,(i,j)}(0) - X_{(i,j)}(0) \right|^2 \right] = 0$, by the assumption. Using a variance bound and the fact that $\lim_{k\to\infty} C_k \to 0$ it follows that $\lim_{k\to\infty} \lim_{n\to\infty} B_k(n) = 0$, in probability. By Lemma 5.18 and Lemma 5.17 we have $\left\| \widetilde{A}_k(s) \right\|_{\blacksquare} \to 0$ in probability for each fixed

$s \in [0,t]$ as $k \to \infty$. It can be shown following the proof of Proposition 6.3 that $\left(\widetilde{A}_k\right)_{k\in\mathbb{N}}$ is equicontinuous over $[0,t]$, almost surely, for sufficiently large k. Therefore, we conclude that $A_k \to 0$ in probability as $k \to \infty$. Since $\lim_{k\to\infty}\lim_{n\to\infty}\mathbb{P}\{E_k(n)\} = 1$,

$$\lim_{n\to\infty}\sup_{s\in[0,t]}\|\mathcal{K}(X_n(s)) - \Gamma(s)\|_{\blacksquare} = 0,\ \lim_{k\to\infty}\lim_{n\to\infty}\sup_{s\in[0,t]} D_2^2\Big(\mathcal{K}\Big(\widetilde{X}_n[k](s)\Big), \mathcal{K}(X[k](s))\Big) = 0,$$

in probability, by choosing $(\lambda_k)_{k\in\mathbb{N}}$ (depending on $(A_k, \lim_{n\to\infty} B_k(n))_{k\in\mathbb{N}}$) that increases sufficiently slowly to infinity as $k \to \infty$. Moreover, it is clear that one can choose $k = o(\sqrt{n})$. This completes the proof. □

D.2.1 Proofs of Chapter 6.2.2

Proof of Proposition 6.16. We first prove a slightly stronger result, that is, we show that Γ^σ converges to Γ in the MVG sense. The desired result therefore follows immediately. The proof closely resembles the proof of Theorem 6.11. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be as above and $X^\sigma, X, \Gamma^\sigma, \Gamma$ be as above with the initial condition $\Gamma^\sigma(0)(x,y) = \Gamma(0)(x,y) = \delta_{\gamma_0(x,y)}$. Using the Lipschitzness of Skorokhod map as in the proof of Theorem 6.11, we observe that for any (i,j) we have

$$\left|X^\sigma_{(i,j)}(t) - X_{(i,j)}(t)\right|^2 \le Ct\int_0^t |b(\gamma^\sigma(s))(U_i,U_j) - b(\gamma(s))(U_i,U_j)|^2 + C\sigma^2\left|B_{(i,j)}(t)\right|^2.$$

Summing over $e \in [k]^2$ and diving by $\frac{1}{k^2}$ we obtain that for each $k \in \mathbb{N}$ we have

$$\|\mathcal{K}(X^\sigma[k](t)) - \mathcal{K}(X[k](t))\|_2^2 \le I_k(t) + J_k(t),$$
$$\text{where}\quad I_k(t) = Ct\int_0^t \frac{1}{k^2}\sum_{(i,j)\in[k]^2} |b(\gamma^\sigma(s))(U_i,U_j) - b(\gamma(s))(U_i,U_j)|^2\,\mathrm{d}s,$$
$$\text{and}\quad J_k(t) \coloneqq C\sigma^2\frac{1}{k^2}\sum_{e\in[k]^2}|B_e(t)|^2.$$

By Doob's maximal inequality [KS91, page 14, Theorem 3.8.iv] and Markov's inequality we get $\mathbb{P}\{\sup_{s\in[0,t]} J_k(s) \ge 2C\sigma^2 t\} \le \frac{C}{4k^2}$. Using our assumption on b, we conclude that (compare with (D.31))

$$\sup_{s\in[0,t]}\|\mathcal{K}(X^\sigma[k](s)) - \mathcal{K}(X[k](s))\|_{\blacksquare}^2 \le C\beta^2\kappa_{\blacksquare}^2 t\int_0^t \|\Gamma(s) - \Gamma^\sigma(s)\|_{\blacksquare}^2\,\mathrm{d}s + 2C\sigma^2 t, \qquad \text{(D.35)}$$

with probability at least $1 - \frac{C}{4k^2}$. Note that compared to equation (D.31) in the proof of Theorem 6.11, the above inequality is much simpler. The reason being that the drift function b depends only on MVG and not on $X_{(i,j)}(t)$. Secondly, the our initial condition ensures that $X^{\sigma}_{(i,j)}(0) = X_{(i,j)}(0)$. At this step, we use the same argument as in the proof of Theorem 6.11 to replace $\|\mathcal{K}(X^{\sigma}[k](s)) - \mathcal{K}(X[k](s))\|^2_{\blacksquare}$ with $\|\Gamma(s) - \Gamma^{\sigma}(s)\|^2_{\blacksquare}$ up to a small error $C_k = 64k^{-1/4}$ with probability at least $1 - \mathrm{e}^{-c_k}$ where $c_k = \sqrt{k}/20$. Combining all this and using Grönwall's inequality [Grö19] as in the proof of Theorem 6.11 we conclude that

$$\sup_{s\in[0,t]} D_2^2(\mathcal{K}(X^{\sigma}[k](s)), \mathcal{K}(X[k](s))) + \sup_{s\in[0,t]} \|\Gamma^{\sigma}(s) - \Gamma(s)\|^2_{\blacksquare}$$
$$\leq (C_k + 2C\sigma^2 t)\mathrm{e}^{C\beta^2\kappa^2_{\blacksquare}t^2}, \text{ with probability at least } 1 - \mathrm{e}^{-c_k} - \frac{C}{4k^2}.$$

Letting $k \to \infty$, we conclude that $\sup_{s\in[0,t]}\|\Gamma^{\sigma}(s) - \Gamma(s)\|^2_{\blacksquare} \leq 2C\sigma^2 t\mathrm{e}^{C\beta^2\kappa^2_{\blacksquare}t^2}$. The desired claim now follows from the fact that $\Gamma \to \mathbb{E}[\Gamma]$ is a contraction. □

Proof of Proposition 6.17. To get a non-asymptotic error rate, we need to control on A_k and $B_k(n)$ in equation (D.33). Observe that $B_k(n)$ depends on the initial condition and in general it can be arbitrarily slow. However, assuming that the initial condition is i.i.d., one can use Chebyshev's inequality to obtain $\mathbb{P}\big\{B_k(n) \geq 66k^{-1/4}\big\} \leq k^{-3/2}$.

On the other hand, combining the arguments in Proposition 6.3 and [Lov12, Proposition 8.12], it can be shown that there exists a constant M_t (depending only on t) such that for any $\delta > 0$ we have $\mathbb{P}\big\{A_k \geq M_t(\delta\log(1/\delta))^{1/4}\big\} \leq k^{-2} + t\delta^{-1}\mathrm{e}^{\frac{128}{\delta\log(1/\delta)}}\mathrm{e}^{-k\delta\log(1/\delta)/2}$.

To obtain above bound for A_k, we argue as follows. For each fixed $\psi \in \mathcal{L}$, moment computation yields $T_{C_4}(\Gamma(\psi, \widetilde{A}_k(s)))$ is sub-gaussian with norm at most $\frac{1}{\sqrt{k}}$. In particular, for any $\delta > 0$ we have $\mathbb{P}\Big\{T_{C_4}(\Gamma(\psi, \widetilde{A}_k(s)) \geq \sqrt{\delta\log(1/\delta)}\Big\} \leq \mathrm{e}^{-\frac{k\delta\log(1/\delta)}{2}}$. Note that the right side is independent of ψ. For any subset $F \subseteq \mathcal{L}$ define $\Delta_{k,F}(s) \coloneqq \sup_{\psi\in F}\Big|t(C_4, \Gamma(\psi, \widetilde{A}_k(s)))\Big|$. Fix $\epsilon > 0$ and note that by Lemma C.4 there exists a finite set $F \subseteq \mathcal{L}$ such that $|F| \leq \mathrm{e}^{\frac{32}{\epsilon^2}}$ and $|\Delta_{k,\mathcal{L}}(s) - \Delta_{k,F}(s)| \leq \epsilon$. Taking $\epsilon = \frac{1}{2}\sqrt{\delta\log(1/\delta)}$, we get $\mathbb{P}\Big\{\Delta_{k,\mathcal{L}}(s) \geq \sqrt{\delta\log(1/\delta)}\Big\} \leq \mathbb{P}\Big\{\Delta_{k,F}(s) \geq 2^{-1}\sqrt{\delta\log(1/\delta)}\Big\} \leq \mathrm{e}^{\frac{128}{\delta\log(1/\delta)}}\mathrm{e}^{-k\delta\log(1/\delta)/2}$.

Repeating the proof of Proposition 6.3, we obtain that $(\Delta_{k,\mathcal{L}})_{k\in\mathbb{N}}$ is equicontinuous with high probability. That is, for any fixed $\delta > 0$ we have

$\mathbb{P}\Big\{\sup_{|s_1-s_2|\le\delta, s_1,s_2\in[0,t]}|\Delta_{k,\mathcal{L}}(s_1)-\Delta_{k,\mathcal{L}}(s_2)|\ge M_t\sqrt{\delta\log(1/\delta)}\Big\} \le \frac{1}{k^2}$. It now follows from a δ-net argument that $\mathbb{P}\big\{\sup_{s\in[0,t]}\Delta_{k,\mathcal{L}}(s)\ge M_t(\delta\log(1/\delta))^{1/2}\big\} \le k^{-2} + t\delta^{-1}\mathrm{e}^{\frac{128}{\delta\log(1/\delta)}}\mathrm{e}^{-k\delta\log(1/\delta)/2}$. Following [Lov12, Proposition 8.12], we have $\left\|\widetilde{A}_k(s)\right\|_{\blacksquare}^4 \le \Delta_{k,\mathcal{L}}(s)$. This yields the desired conclusion. In particular, choosing $\delta = 64\sqrt{k^{-1}\log k}$ and $\lambda_k = \log(k)/(16\cdot 384t(L^2+2\kappa_{\blacksquare}^2))$, we have the left hand side of equation (D.33) bounded by $M_t k^{-1/16}\log^{3/2}k$ with probability at least $1-\frac{k^2}{n}-4k^{-\frac{1}{\kappa^2 t}}-2t\mathrm{e}^{-\sqrt{k}/20}-2k^{-3/2}$, where $\kappa = 32\sqrt{6}(L^2+2\kappa_{\blacksquare}^2)^{1/2}$.

Since t is fixed, we can choose k to be a suitable function of n, say $k = n^{2/7}$. The proof is now complete with the help of Proposition 6.16 and a triangle inequality. □

Proof of Proposition 6.18. Notice that $\beta D\mathcal{H}$ is the Fréchet-like derivative evaluation map of $\beta\mathcal{H}$. The proof immediately follows from Remark 4.25 following [AGS08, Remark 4.0.5, part (d)], [AGS08, Corollaray 4.0.6] and Assumption 3.4. □

D.3 Proofs of Chapter 6.3

In this section, we will provide proofs of the statements made in Chapter 6.3.

D.3.1 Proofs of Chapter 6.3.2

In this section, we will provide the proof arguments for Theorem 6.20. We first give a brief intuition behind the proof. The general philosophy is to rewrite $n\mathcal{E}_n$ (or $n^{1/2}\mathcal{E}_n$ depending on the case) as the sum of two matrices. The first matrix has entrywise variance of order $O(1)$ while the second one has (entrywise) variance going to 0 as $n\to\infty$. The proof now follows by showing that the first matrix (with entrywise $O(1)$ variance) converges to the appropriate IEA. We should remark that $n\mathcal{E}_n$ is an infinite sum where each term has complicated dependence with each other. This makes the problem of identifying the terms with vanishing variance non-trivial. We explain this philosophy more concretely below.

Case 1: We first consider the case $\mu_n = \sigma_n = n^{-1}$. Begin by noticing that

$$\operatorname{Texp}[Y_n] = \operatorname{Texp}\left[\int_0^{\cdot} \mu_n A_n(s)\,\mathrm{d}s\right] + \sigma_n B_n + \sum_{k=2}^{\infty} \widetilde{J}_k,$$

where $\widetilde{J}_k$ is the sum of all k-fold integrals that contain at least one (scaled) BM. Note that

$$n(\operatorname{Texp}[Y_n] - I_n) = n\left(\operatorname{Texp}\left[\int_0^{\cdot} \mu_n A_n(s)\,\mathrm{d}s\right] - I_n\right) + B_n + n\sum_{k=2}^{\infty} \widetilde{J}_k. \tag{D.36}$$

On the other hand,

$$nK\left(\operatorname{Texp}\left[\int_0^{t} \mu_n A_n(s)\,\mathrm{d}s\right] - I_n\right) = (\mathrm{e}^t - 1)\,\mathbb{E}\left[\prod_{s\in N_t} K(A_n(s)) \,\middle|\, N_t \geq 1\right].$$

From the assumption that $K(A_n)$ converges to some kernel $w(t)$ in $L^2([0,1]^2)$, we obtain $(\mathrm{e}^t - 1)\,\mathbb{E}\big[\prod_{s\in N_t} K(A_n)(s) \,\big|\, N_t \geq 1\big]$ converges in L^2 to $\Gamma(u)(t) \coloneqq (\mathrm{e}^t - 1)\,\mathbb{E}\big[\prod_{s\in N_t} w(s) \,\big|\, N_t \geq 1\big]$. A randomly chosen $r \times r$ submatrix of $\left(\operatorname{Texp}\left[\int_0^t \mu_n A_n(s)\,\mathrm{d}s\right] - I_n\right)$ therefore converges to $\Gamma(u)(t)\{r\}$ for i.i.d. $\mathrm{Uni}([0,1])$ random variables $\{U_i\}_{i\in\mathbb{N}}$.

It is reasonable to believe that, as $n \to \infty$, the sum $n\big(\operatorname{Texp}\big[\int_0^{\cdot} \mu_n A_n(s)\,\mathrm{d}s\big] - I_n\big) + B_n$ converges to the appropriate IEA $X(t) = \Gamma(u)(t)\{\infty\} + B(t)$ where B is an IEA with all Brownian motions. On the other hand, we note that $\sum_{k=2}^{\infty} \widetilde{J}_k$ is a Gaussian random variable with mean 0, and show that its variance is $O\big(\frac{1}{n}\big)$.

Case 2: Consider the case $\mu_n = \sigma_n^2 = n^{-1}$. Just as above, let us rewrite

$$\operatorname{Texp}[Y_n] = \operatorname{Texp}\left[\int_0^{\cdot} \mu_n A_n(s)\,\mathrm{d}s\right] + \sum_{k=1}^{\infty} \widetilde{J}_k.$$

Now notice that the same heuristic as above shows that $\sqrt{n}\big(\operatorname{Texp}\big[\int_0^{\cdot} \mu_n A_n(s)\,\mathrm{d}s\big] - I_n\big) = O\Big(\frac{1}{\sqrt{n}}\Big)$. On the other hand, rewrite $\sum_{k=1}^{\infty} \widetilde{J}_k = (\operatorname{Texp}[\sigma_n B_n] - I_n) + \sum_{k=2}^{\infty} \widehat{J}_k$, where $\widehat{J}_k$ is the sum of all k-fold integrals which contain at least one BM but not all are BMs. Following [RY04, page

151], we now notice that $\sqrt{n}(\text{Texp}[\sigma_n B_n](t) - I_n)$ has $O(1)$ variance and it converges to an IEA with entries distributed as B_{e^t-1}, where B is a one dimensional BM. We show that $\sum_{k=2}^{\infty} \sqrt{n}\widehat{J}_k(t)$ has variance of order $O\big(\frac{1}{n}\big)$. And, therefore, we conclude that $n^{1/2}\mathcal{E}_n$ converges to an IEA whose coordinates are i.i.d. and have the same distribution as a Brownian motion.

Case 3: In the same setting as $\mu_n = \sigma_n^2 = n^{-1}$. Notice that the limiting IEA is obtained as the limit of $\sqrt{n}(\text{Texp}[\sigma_n B_n] - I_n)$. And, this limit is trivial – in the sense that – the limit does not depend on the deterministic sequence of matrices A_n. This is, however, expected. Notice that with this choice of scaling the noise is much larger than the 'signal' or the deterministic term. To see the effect of the 'signal', one can consider the limit of the matrix

$$\begin{aligned} n(\mathcal{E}_n - (\text{Texp}[\sigma_n B_n] - I_n)) &= n(\text{Texp}[Y_n] - \text{Texp}[\sigma_n B_n]) \\ &= n\left(\text{Texp}\left[\int_0^{\cdot} \mu_n A_n(s)\,\mathrm{d}s\right] - I_n\right) + \sum_{k=2}^{\infty} n\widehat{J}_k. \end{aligned}$$

As we mentioned earlier, the first term remain $O(1)$ as $n \to \infty$ and we understand the limit of this term. We further decompose the $n\sum_{k=2}^{\infty} \widehat{J}_k$ as follows. For $k \geq 2$, write $\widehat{J}_k = \widehat{J}_{k,0} + \widehat{J}_{k,1}$, where $\widehat{J}_{k,0}$ is the sum of all k-fold integrals with exactly one BM at either the first or the last integral. We then show that $\sum_{k=2}^{\infty} n\widehat{J}_{k,0}$ is a zero mean Gaussian with $O(1)$ variance, while the remaining term $\sum_{k=3}^{\infty} n\widehat{J}_{k,1}$ is mean 0 Gaussian with vanishing variance. We therefore conclude that $n(\text{Texp}[Y_n] - \text{Texp}[\sigma_n B_n])$ converges to an IEA with independent Gaussian coordinates. Note that this limiting the mean of this IEA is same as the IEA obtained in the case 1, but the variances are different.

It is clear from the above heuristic that we will need to compute the variances of infinite sum of Gaussian random variables which may be dependent. To do this, we need the following lemma.

Let $k \geq 2$ and let $\pi = (z_k, z_{k-1}, \ldots, z_1, z_0)$ be a $(k+1)$-tuple where each $z_i \in [n]$. For

$p \le k$, define

$$I_{k,p,\pi}(t) \coloneqq \sum_{\alpha \in \binom{k}{p}} \int_{\Delta_k(t)} \mathrm{d}U_{\alpha,\pi}(\mathbf{s}), \qquad t \in [0,1].$$

where $\mathrm{d}U_{\alpha,\pi}(\mathbf{s}) = \prod_{i=1}^{k} \mathrm{d}U_{\alpha,i,(z_i,z_{i-1})}(s_i)$, and $\mathrm{d}U_{\alpha,i}(s_i) = \begin{cases} \mathrm{d}B_n(s_i), & \text{if } i \in \alpha, \\ A_n \,\mathrm{d}s_i, & \text{if } i \notin \alpha \end{cases}$. Also define $I_{k,p}(t)$ as

$$I_{k,p,(x,y)}(t) \coloneqq \sum_{\pi \text{ s.t. } (z_k,z_0)=(x,y)} I_{k,p,\pi}, \qquad (x,y) \in [n]^2, \qquad t \in [0,1].$$

Lemma D.5. *For $k_1, k_2 \in \mathbb{N}$, $p \le k_1 \wedge k_1$, $\pi_1 \in [n]^{k_1+1}$, $\pi_2 \in [n]^{k_2+1}$, $t \in \mathbb{R}_+$, and $\alpha \in \binom{k_1}{p}$ and $\beta \in \binom{k_2}{p}$ such that $\pi_1(\alpha_i) = \pi_2(\beta_i)$ for all $i \in [p]$. Then*

$$\left| \int_{\Delta_{k_1}(t)} \int_{\Delta_{k_2}(t)} \mathbb{E}[\mathrm{d}U_{\alpha,\pi_1}(\mathbf{s}) \,\mathrm{d}U_{\beta,\pi_2}(\boldsymbol{\tau})] \right| \le C^{k_1-p} C^{k_2-p} \cdot |\Delta(p; k_1, k_2, \alpha, \beta; t)|, \tag{D.37}$$

where $\Delta(p; k_1, k_2, \alpha, \beta)$ is the $k_1 + k_2 - p$ dimensional space defined by

$$\Delta(p; k_1, k_2, \alpha, \beta, t) \coloneqq \{(\mathbf{s}, \boldsymbol{\tau}) \in \Delta_{k_1}(t) \times \Delta_{k_2}(t) \mid s_{\alpha_i} = \tau_{\beta_i} \ \ \forall \ \ i \in [p]\}.$$

Proof of Lemma D.5. Following the condition on π_1 and π_2, $\mathbb{E}\Big[\mathrm{d}U_{\alpha,i,(z_{\alpha_i},z_{\alpha_i-1})}(s_{\alpha_i}) \,\mathrm{d}U_{\beta,i,(\tilde{z}_{\beta_i},\tilde{z}_{\beta_i-1})}(\tau_{\beta_i})\Big] = \delta_{s_{\alpha_i}=\tau_{\beta_i}}$ for all $i \in [p]$. Therefore, in the following we assume that π_1, π_2 are such that $(z_{\alpha_i}, z_{\alpha_i-1}) = (\tilde{z}_{\beta_i}, \tilde{z}_{\beta_i-1})$ for all $i \in [p]$. Therefore, the $(k_1 + k_2)$-dimensional Lebesgue integral over $\Delta_{k_1}(t) \times \Delta_{k_2}(t)$ gets reduced to a $(k_1 + k_2 - p)$-dimensional Lebesgue integral over the resulting constraint set $\Delta(p; k_1, k_2, \alpha, \beta; t)$. Since that the absolute value of the coordinates of A_n are bounded by $C \ge 0$, we get

$$\left| \int_{\Delta_{k_1}(t)} \int_{\Delta_{k_2}(t)} \mathbb{E}[\mathrm{d}U_{\alpha,\pi_1}(\mathbf{s}) \,\mathrm{d}U_{\beta,\pi_2}(\boldsymbol{\tau})] \right| \le C^{k_1-p} C^{k_2-p} \cdot |\Delta(p; k_1, k_2, \alpha, \beta; t)|.$$

□

We now make the following claim that bounds the volume of the set $|\Delta(p; k_1, k_2, \alpha, \beta; t)|$.

Claim D.1. *We denote by* $\delta_\alpha(1) = \alpha_1 - 1, \delta_\alpha(2) = \alpha_2 - \alpha_1 - 1$ *and similarly* $\delta_\alpha(i) = \alpha_i - \alpha_{i-1} - 1$ *for* $i \in [p]$. *Also define* $\delta_\alpha(p+1) = k_1 - \alpha_p$. *Note that* $\sum_{i=1}^{p+1} \delta_\alpha(i) = k_1 - p$. *And, similarly we define* $\delta_\beta(i)$ *as well. Then,*

$$|\Delta(p; k_1, k_2, \alpha, \beta; t)| \le \frac{t^p}{p!} \frac{t^{k_1-p}}{\delta_\alpha(1)! \dots \delta_\alpha(p+1)!} \frac{t^{k_2-p}}{\delta_\beta(1)! \dots \delta_\beta(p+1)!}.$$

Proof of Claim D.1. For each $j \in [p+1]$, define two collections of i.i.d. Uni$([0,1])$ random vectors, say $X^j = (X_1^j, \dots, X_{\delta_\alpha(j)}^j)$ and $Y^j = (Y_1^j, \dots, Y_{\delta_\beta(j)}^j)$. Let $U = (U_1, \dots, U_p)$ be another vector where U_i are i.i.d. Uni$([0,1])$ random variables. We also set $U_0 = 0$ and $U_{p+1} = t$.

For a vector $v \in \mathbb{R}^n$, we say $v \in \mathcal{I}_n(a,b)$ if $b \ge v_n \ge v_{n-1} \ge v_1 \ge a$. Given a vector $u = (u_1, \dots, u_p)$ define the events

$$E_1(u) \coloneqq \{X^j \in \mathcal{I}_{\delta_\alpha(j)}(u_{j-1}, u_j) \quad \forall\, j \in [p+1]\},$$
$$E_2(u) \coloneqq \{Y^j \in \mathcal{I}_{\delta_\alpha(j)}(u_{j-1}, u_j) \quad \forall\, j \in [p+1]\},$$

where $u_0 = 0$ and $u_{p+1} = t$. Now notice that

$$\begin{aligned} |\Delta(p; k_1, k_2, \alpha, \beta, t)| &= \mathbb{P}\{E_1(U) \cap E_2(U)\} \\ &\le \frac{t^{k_1-p}}{\delta_{\alpha(1)}! \dots \delta_{\alpha(p+1)}!} \frac{t^{k_2-p}}{\delta_{\beta(1)}! \dots \delta_{\beta(p+1)}!} \int_{\Delta_p(t)} \mathrm{d}u_1 \dots \mathrm{d}u_p. \end{aligned}$$

□

We use the Lemma D.5 to compute the variances of the error terms in Case 1 to 3 above.

Lemma D.6. *For every* $(x,y) \in [n]^2$,

1. $\mathrm{Var}\Big[n \sum_{k=2}^{\infty} \widetilde{J}_{k,(x,y)}(t)\Big] = O\big(\frac{1}{n}\big)$,
2. $\mathrm{Var}\Big[n^{1/2} \sum_{k=2}^{\infty} \widehat{J}_{k,(x,y)}(t)\Big] = O\big(\frac{1}{n}\big)$, *and*
3. $\mathrm{Var}\Big[n \sum_{k=3}^{\infty} \widehat{J}_{k,1,(x,y)}(t)\Big] = O\big(\frac{1}{n}\big)$,

where each statement corresponds to error terms in Case 1, Case 2 and Case 3 respectively.

Proof.

1. Notice that

$$\sum_{k=2}^{\infty} \widetilde{J}_k(t) = \sum_{p=1}^{\infty} H_{n,p}(t), \qquad H_{n,p}(t) \coloneqq \sum_{k=p\vee 2}^{\infty} \mu_n^{k-p}\sigma_n^p I_{k,p,(x,y)}(t). \tag{D.38}$$

The benefit of such rearrangement is that the random variables $H_{n,p}(t)$ for all $p \in \mathbb{N}$ are independent, that is, H_{n,p_1} and H_{n,p_2} are independent Gaussians. This allows us to compute the variance of $n\sum_{k=1}^{\infty} \widetilde{J}_k(t)$ by adding $\mathrm{Var}[H_{n,p}(t)]$ over $p \in \mathbb{N}$. In order to compute the variance of $H_{n,p}(t)$, we need to compute the covariance between $I_{k_1,p}$ and $I_{k_2,p}$ for $k_1, k_2 \geq p$. Then,

$$\begin{aligned}
\mathrm{Var}\left[n\sum_{k=2}^{\infty} \widetilde{J}_{k,(x,y)}(t)\right] &\leq \mathrm{Var}\left[n\sum_{p=1}^{\infty} H_{n,p}(t)\right] \\
&= n^2 \sum_{p=1}^{\infty} \mathrm{Var}[H_{n,p}(t)] \\
&= n^2 \sum_{p=1}^{\infty} \mathbb{E}\left[\left(\sum_{k=p}^{\infty} \mu_n^{k-p}\sigma_n^p I_{k,p,(x,y)}(t)\right)^2\right] \\
&= n^2 \sum_{p=1}^{\infty} \mathbb{E}\left[\sum_{k_1,k_2=p}^{\infty} \mu_n^{k_1-p}\mu_n^{k_2-p}\sigma_n^{2p} I_{k_1,p,(x,y)}(t) I_{k_2,p,(x,y)}(t)\right] \\
&= n^2 \sum_{p=1}^{\infty} \sum_{k_1,k_2=p}^{\infty} \mu_n^{k_1-p}\mu_n^{k_2-p}\sigma_n^{2p} \sum_{\pi_1}\sum_{\pi_2} \mathbb{E}[I_{k_1,p,\pi_1}(t) I_{k_2,p,\pi_2}(t)]
\end{aligned}$$

The final two summations in the last expression above, can be rearranged as be written as

$$\sum_{\pi_1}\sum_{\pi_2} \mathbb{E}[I_{k_1,p,\pi_1}(t) I_{k_2,p,\pi_2}(t)] = \sum_{\alpha\in\binom{k_1}{p}} \sum_{\beta\in\binom{k_2}{p}} \sum_{\pi_1,\pi_2} \int_{\Delta_{k_1}(t)} \int_{\Delta_{k_2}(t)} \mathbb{E}[\mathrm{d}U_{\alpha,\pi_1}(\mathbf{s})\,\mathrm{d}U_{\beta,\pi_2}(\boldsymbol{\tau})],$$

where for every $\alpha \in \binom{k_1}{p}$ and $\beta \in \binom{k_2}{p}$, the above sum over $\pi_1 \in [n]^{k_1+1}$ and $\pi_2 \in [n]^{k_2+1}$ are such that $\pi_1(\alpha_i) = \pi_2(\beta_i)$ for all $i \in [p]$. Notice that, without this constraint on π_1, π_2, this summation potentially has $n^{k_1-1}n^{k_2-1}$ summands, but due to the constraint some terms will be zero and can be dropped.

Let $\pi_1 = (z_{k_1}, \dots, z_0)$, and $\pi_2 = (\tilde{z}_{k_2}, \dots, \tilde{z}_0)$. Notice that the above expectation is 0 unless $U_{\alpha,i,(z_{\alpha_i}, z_{\alpha_i - 1})} = U_{\beta,i,(\tilde{z}_{\beta_i}, \tilde{z}_{\beta_i - 1})}$ for $i \in [p]$. And, in this case, $\mathbb{E}\Big[\mathrm{d}U_{\alpha,i,(z_{\alpha_i}, z_{\alpha_i - 1})}(s_{\alpha_i})\, \mathrm{d}U_{\beta,i,(\tilde{z}_{\beta_i}, \tilde{z}_{\beta_i - 1})}(\tau_{\beta_i})\Big] = \delta_{s_{\alpha_i} = \tau_{\beta_i}}$ for all $i \in [p]$. Therefore, in the following we assume that π_1, π_2 are such that $(z_{\alpha_i}, z_{\alpha_i - 1}) = (\tilde{z}_{\beta_i}, \tilde{z}_{\beta_i - 1})$ for all $i \in [p]$, leading to at most $n^{k_1 - 1} n^{k_2 - 1} n^{-p}$ many non-zero terms. This observation, and Lemma D.5 allows us to bound the absolute value of the above sum as

$$n^{k_1 - 1} n^{k_2 - 1} \cdot n^{-p} \cdot C^{k_1 - p} C^{k_2 - p} \sum_{\alpha \in \binom{k_1}{p}} \sum_{\beta \in \binom{k_2}{p}} |\Delta(p; k_1, k_2, \alpha, \beta; t)|.$$

Plugging back, and using the triangle inequality, we have

$$\begin{aligned}
&\mathrm{Var}\left[n \sum_{k=2}^{\infty} \widetilde{J}_{k,(x,y)}(t) \right] \\
&\leq n^2 \sum_{p=1}^{\infty} \sum_{k_1, k_2 = p}^{\infty} \mu_n^{k_1 - p} \mu_n^{k_2 - p} \sigma_n^{2p} n^{k_1 - 1} n^{k_2 - 1} n^{-p} t^{k_1 + k_2 - p} \frac{1}{p!} \frac{(C(p+1))^{k_1 - p}}{(k_1 - p)!} \frac{(C(p+1))^{k_2 - p}}{(k_2 - p)!} \\
&= n^2 \sum_{p=1}^{\infty} \frac{1}{p!} (n\sigma_n)^{2p} n^{-(p+2)} t^p \mathrm{e}^{2Ct(p+1)} \\
&\leq \mathrm{e}^{2Ct} \sum_{p=1}^{\infty} \frac{1}{p!} \big(n\sigma_n^2 t \mathrm{e}^{2Ct}\big)^p \\
&= \mathrm{e}^{2Ct} \big(\exp\big(n\sigma_n^2 t \mathrm{e}^{2Ct}\big) - 1\big) = O\left(\frac{1}{n}\right).
\end{aligned}$$

The last relation holds by noting that $\sigma_n = \frac{1}{n}$ and the Taylor approximation of the exponential.

2. The proof is similar to the proof of part 1, where we have $\sigma_n = n^{-1/2}$ instead of n^{-1} (and the prefactor n^2 replaced by n). This yields, that $\mathrm{Var}\Big[n^{1/2} \sum_{k=2}^{\infty} \widehat{J}_{k,(x,y)}(t)\Big] \leq \frac{1}{n} \mathrm{e}^{2Ct} \big(\exp\big(t\mathrm{e}^{2Ct}\big) - 1\big)$. We skip the details.

3. The proof is similar to the proof of part 1, where we have $\sigma_n = n^{-1/2}$ instead of n^{-1}, and $n^{k_1 - 1} n^{k_2 - 1} n^{-(p+1)}$ number of non-zero terms instead of $n^{k_1 - 1} n^{k_2 - 1} n^{-p}$ many. This yields, that $\mathrm{Var}\Big[n \sum_{k=3}^{\infty} \widehat{J}_{k,1,(x,y)}(t)\Big] \leq \frac{1}{n} \mathrm{e}^{2Ct} \big(\exp\big(t\mathrm{e}^{2Ct}\big) - 1\big)$. We skip the details.

This completes the proof. □

Lemma D.7. *For every* $((i_1, j_1), t_1), ((i_2, j_2), t_2) \in [n]^2 \times \mathbb{R}_+$, *the covariance between* $n \sum_{k=2}^{\infty} \widehat{J}_{k,0,(i_1,j_1)}(t_1)$ *and* $n \sum_{k=2}^{\infty} \widehat{J}_{k,0,(i_2,j_2)}(t_2)$ *is*

$$\begin{aligned} &C_n(((i_1, j_1), t_1), ((i_2, j_2), t_2)) \\ &\quad \coloneqq \mathbb{1}\{i_1 = i_2\} \int_0^{\min\{t_1,t_2\}} \frac{1}{n} \big(\Gamma_{n,1}(s)^\top \Gamma_{n,1}(s)\big)(j_1, j_2) \,\mathrm{d}s \\ &\qquad + \mathbb{1}\{j_1 = j_2\} \int_0^{t_1} \int_0^{t_2} \min\{s_1, s_2\} \frac{1}{n} \big(\Gamma_{n,2}(s_1; t_1) \Gamma_{n,2}(s_2; t_2)^\top\big)(i_1, i_2) \,\mathrm{d}s_2 \,\mathrm{d}s_1, \end{aligned} \tag{D.39}$$

where

$$\begin{aligned} \Gamma_{n,1}(s) &\coloneqq n\bigg(\operatorname{Texp}\bigg[\int_0^{\cdot} \mu_n A_n(r) \,\mathrm{d}r\bigg](s) - I_n\bigg), \\ \Gamma_{n,2}(s; t) &\coloneqq n \operatorname{Texp}\bigg[\int_0^{\cdot} \mu_n \tau_s(A_n)(r) \,\mathrm{d}r\bigg](t - s) \mu_n A_n(s), \end{aligned} \tag{D.40}$$

and $\tau_s(A_n)$ *is the curve* A_n *shifted by* s, *i.e.,* $\tau_s(A_n)(r - s) \coloneqq A_n(r)$ *for all* $r \in [s, t]$, *for* $s \in [0, t]$, *and all* $t \in \mathbb{R}_+$.

Proof of Lemma D.7. The term $n\widehat{J}_{k,0}$ for $k \geq 2$, has two kinds of terms. The first kind in which the BM appears at the position 1, and the second kind in which the BM appears at the position k.

For the terms of the first kind, notice the following:

$$\begin{aligned} &n \cdot \int_0^t \mu_n A_n(s_k) \,\mathrm{d}s_k \int_0^{s_k} \mu_n A_n(s_{k-1}) \,\mathrm{d}s_{k-1} \cdots \int_0^{s_3} \mu_n A_n(s_2) \cdot \sigma_n B_n(s_2) \,\mathrm{d}s_2 \\ &\quad = n \int_0^t J_{k-2}\bigg(\int_0^{\cdot} \mu_n \tau_{s_2}(A_n)(r) \,\mathrm{d}r\bigg)(t - s_2) \mu_n A_n(s_2) \sigma_n B(s_2) \,\mathrm{d}s_2, \end{aligned}$$

where $\tau_{s_2}(A_n)$ is nothing but the curve A_n shifted by s_2, i.e., $\tau_{s_2}(A_n)(s - s_2) \coloneqq A_n(s)$ for all $s \in [0, t - s_2]$. Summing over all such terms for $k \in \mathbb{Z}_+ \setminus \{0, 1\}$, we get that the above is equal to

$$\int_0^t n \operatorname{Texp}\bigg[\int_0^{\cdot} \mu_n \tau_{s_2}(A_n)(r) \,\mathrm{d}r\bigg](t - s_2) \mu_n A_n(s_2) \sigma_n B_n(s_2) \,\mathrm{d}s_2.$$

For the terms of the second kind, the argument is however simpler. Notice that

$$\begin{aligned} & n\cdot \int_0^t \sigma_n \,\mathrm{d}B_n(s_k) \int_0^{s_k} \mu_n A_n(s_{k-1})\,\mathrm{d}s_{k-1} \cdots \int_0^{s_2} \mu_n A_n(s_1)\,\mathrm{d}s_1 \\ & \quad = n\cdot \int_0^t \sigma_n \,\mathrm{d}B_n(s_k) J_{k-1}(\mu_n A_n)(s_k). \end{aligned}$$

Summing over all such terms for $k \in \mathbb{Z}_+ \setminus \{0,1\}$, we get that the above is equal to

$$\int_0^t \sigma_n \,\mathrm{d}B_n(s_k) \cdot n\bigg(\mathrm{Texp}\bigg[\int_0^{\cdot} \mu_n A_n(r)\,\mathrm{d}r\bigg](s_k) - I_n\bigg).$$

The sum of the two kinds of term finally is

$$\sigma_n \int_0^t \mathrm{d}B_n(s)\Gamma_{n,1}(s) + \sigma_n \int_0^t \Gamma_{n,2}(s) B_n(s)\,\mathrm{d}s. \tag{D.41}$$

Consider two pairs of indices $((i_1,j_1),t_1)$ and $((i_2,j_2),t_2)$ in $[n]^2\times\mathbb{R}_+$. Then the covariance between the two pair of coordinates is

$$\begin{aligned} & C_n(((i_1,j_1),t_1),((i_2,j_2),t_2)) \\ & = \mathbb{E}\bigg[\int_0^{t_1}\int_0^{t_2} \frac{1}{n}\sum_{k_1,k_2=1}^n \mathrm{d}B_{n,(i_1,k_1)}(s_1)\,\mathrm{d}B_{n,(i_2,k_2)}(s_2)\Gamma_{n,1,(k_1,j_1)}(s_1)\Gamma_{n,1,(k_2,j_2)}(s_2)\bigg] \\ & + \mathbb{E}\bigg[\int_0^{t_1}\int_0^{t_2} \frac{1}{n}\sum_{k_1,k_2=1}^n B_{n,(k_1,j_1)}(s_1)B_{n,(k_2,j_2)}(s_2)\Gamma_{n,2,(i_1,k_1)}(s_1;t_1)\Gamma_{n,2,(i_2,k_2)}(s_2;t_2)\,\mathrm{d}s_2\,\mathrm{d}s_1\bigg] \\ & = \mathbb{1}\{i_1=i_2\}\int_0^{\min\{t_1,t_2\}} \bigg(\frac{1}{n}\sum_{k=1}^n \Gamma_{n,1,(k,j_1)}(s)\Gamma_{n,1,(k,j_2)}(s)\bigg)\,\mathrm{d}s \\ & \qquad + \mathbb{1}\{j_1=j_2\}\int_0^{t_1}\int_0^{t_2} \min\{s_1,s_2\}\bigg(\frac{1}{n}\sum_{k=1}^n \Gamma_{n,2,(i_1,k)}(s_1;t_1)\Gamma_{n,2,(i_2,k)}(s_2;t_2)\bigg)\,\mathrm{d}s_2\,\mathrm{d}s_1 \\ & = \mathbb{1}\{i_1=i_2\}\int_0^{\min\{t_1,t_2\}} \frac{1}{n}\big(\Gamma_{n,1}(s)^\top\Gamma_{n,1}(s)\big)(j_1,j_2)\,\mathrm{d}s \\ & \qquad + \mathbb{1}\{j_1=j_2\}\int_0^{t_1}\int_0^{t_2} \min\{s_1,s_2\}\frac{1}{n}\big(\Gamma_{n,2}(s_1;t_1)\Gamma_{n,2}(s_2;t_2)^\top\big)(i_1,i_2)\,\mathrm{d}s_2\,\mathrm{d}s_1. \end{aligned}$$

This completes the proof. □

We are now ready to prove Theorem 6.20. Recall that by our assumption there exists a continuous curve $t \mapsto w(t)$ of kernels such that $\sup_{s\in[0,t]}\|K(A_n)(s) - w(s)\|_2 \to 0$ as

$n \to \infty$. Furthermore, we assume that $\sup_{s\in[0,t]}\|w(s)\|_\infty \le C$. Under these assumption, the kernel $\Gamma(u)(t) \coloneqq \sum_{k=1}^\infty J_k(u)(t)$ is well defined and $\|\Gamma(u)(t)\|_\infty \le \mathrm{e}^{Ct} - 1$. In the following, we will use the notation $\Gamma(t)$ instead of $\Gamma(u)(t)$ for simplicity. Let us also define the kernel $\Gamma_n(t) = nK\big(\sum_{k=1}^\infty J_k\big(\frac{A_n}{n}\big)\big) = \sum_{k=1}^\infty J_k(K(A_n))$. It follows from our assumption that $\sup_{s\in[0,t]}\|\Gamma_n(s) - \Gamma(s)\|_2 \to 0$ as $n \to \infty$. Analogous to C_n defined above, we define a kernel C_∞. Let

$$\Gamma_1(s) \coloneqq \Gamma(u)(s), \quad \Gamma_2(s;t) \coloneqq \mathrm{Texp}[u](t-s) \odot w(s), \qquad s \in [0,t], \quad t \in [0,1],$$

and define

$$\begin{aligned} &C_\infty(((x_1,y_1),t_1),((x_2,y_2),t_2)) \\ &\quad\coloneqq \mathbb{1}\{x_1 = x_2\} \int_0^{\min\{t_1,t_2\}} \big(\Gamma_1(s)^\top \odot \Gamma_1(s)\big)(y_1,y_2)\,\mathrm{d}s \\ &\qquad + \mathbb{1}\{y_1 = y_2\} \int_0^{t_1}\int_0^{t_2} \min\{s_1,s_2\}\big(\Gamma_2(s_1;t_1) \odot \Gamma_2(s_2;t_2)^\top\big)(x_1,x_2)\,\mathrm{d}s_2\,\mathrm{d}s_1. \end{aligned} \tag{D.42}$$

Let $S\colon t \mapsto \int_0^t T(s)\,\mathrm{d}s$ be an absolutely continuous curve of operators on $L^2([0,1])$. Recall that we can define $J_k(S)(t)$ as

$$J_k(S)(t) \coloneqq \int_{\Delta_k(t)} T(s_{k-1})\dots T(s_1)\,\mathrm{d}s_k \dots \mathrm{d}s_1.$$

Lemma D.8. *Let T_1 and T_2 be the curves of curve of operators and define $S_i(t) \coloneqq \int_0^t T_i(s)\,\mathrm{d}s$ for $i \in [2]$. Assume that $\sup_{s\in[0,t]}\|T_1(s)\|_{\mathrm{op}}, \sup_{s\in[0,t]}\|T_2(s)\|_{\mathrm{op}} \le C_t$ for some constant $C_t > 0$ for every $t \in \mathbb{R}_+$. Then,*

$$\|J_k(S_1)(t) - J_k(S_2)(t)\|_{\mathrm{op}} \le \eta(t) C_t^{k-1} k \frac{t^k}{k!}, \qquad k \ge 1,$$

where $\eta(t) = \sup_{s\in[0,t]}\|T_1(s) - T_2(s)\|_{\mathrm{op}}$ for $t \in \mathbb{R}_+$. In particular, we have

$$\sup_{s\in[0,t]} \|\mathrm{Texp}[S_1](s) - \mathrm{Texp}[S_2](s)\|_{\mathrm{op}} \le \eta(t) t(\mathrm{e}^{tC_t} - 1), \qquad t \in \mathbb{R}_+.$$

Proof. Observe that

$$\|J_k(S_1)(t) - J_k(S_2)\|_{\mathrm{op}} \le \int_{\Delta_k(t)} \|T_1(s_k)\dots T_1(s_1) - T_2(s_k)\dots T_2(s_1)\|_{\mathrm{op}}\,\mathrm{d}s_k \dots \mathrm{d}s_1$$

$$
\begin{aligned}
&\leq \int_{\Delta_k(t)} \sum_{j=1}^{k} \left\| \prod_{i=j+1}^{k} T_1(s_i) \cdot (T_1(s_j) - T_2(s_j)) \cdot \prod_{i=1}^{j-1} T_2(s_i) \right\|_{\mathrm{op}} \prod_{i=1}^{k} \mathrm{d}s_i \\
&\leq k C_t^{k-1} \eta(t) \int_{\Delta_k(t)} \mathrm{d}s_k \cdots \mathrm{d}s_1 \\
&\leq \eta(t) C_t^{k-1} k \frac{t^k}{k!}.
\end{aligned}
$$

Finally observe that

$$
\|\mathrm{Texp}[S_1](s) - \mathrm{Texp}[S_2](s)\|_{\mathrm{op}} \leq \sum_{k \geq 1} \|J_k(S_1)(s) - J_k(S_2)(s)\|_{\mathrm{op}} \leq \eta(s) t (\mathrm{e}^{tC_t} - 1).
$$

Taking supremum over $s \in [0,t]$ we get the final result. □

Lemma D.9. *Let w_1 and w_2 be two curves of kernels and let $u_i = \int_0^t w_i(s)\,\mathrm{d}s$ for $i \in [2]$. Assume that $\sup_{s\in[0,t]} \|w_i(s)\|_\infty \leq C_t$ for some $C_t > 0$ and for $i \in [2]$. Define,*

$$
\eta(t) = \sup_{s \in [0,t]} \|w_1(s) - w_2(s)\|_2.
$$

Then, for every fixed $0 \leq s \leq t \in \mathbb{R}_+$ we have

$$
\begin{aligned}
\|\Gamma_1(u_1)(t) - \Gamma_1(u_2)(t)\|_2 &\leq t C_t \mathrm{e}^{tC_t} \eta(t), \\
\|\Gamma_2(u_1)(s;t) - \Gamma_2(u_2)(s;t)\|_2 &\leq \big(tC_t(\mathrm{e}^{tC_t} - 1) + \mathrm{e}^{tC_t}\big)\eta(t).
\end{aligned}
$$

Proof of Lemma D.9. The proof for the continuity of Γ_1 follows exactly the same argument as in Lemma D.8, where we prove this result for a curve of operators on $L^2[0,1]$. The continuity of Γ_2 follows a similar argument that we give present here for completeness. Observe that

$$
\begin{aligned}
&\|\Gamma_2(u_1)(s;t) - \Gamma_2(u_2)(s;t)\|_2 \\
&\leq \|(\mathrm{Texp}[u_1](t-s) - \mathrm{Texp}[u_2](t-s)) \odot w_1(s)\|_2 + \|\mathrm{Texp}[u_2](t-s) \odot (w_1(s) - w_2(s))\|_2 \\
&\leq \|\mathrm{Texp}[u_1](t-s) - \mathrm{Texp}[u_2](t-s)\|_{\mathrm{op}} \|w_1(s)\|_2 + \|\mathrm{Texp}[u_2](t-s)\|_{\mathrm{op}} \|w_1(s) - w_2(s)\|_2 \\
&\leq tC_t(\mathrm{e}^{tC_t} - 1)\eta(t) + \mathrm{e}^{tC_t}\eta(t) = \big(tC_t(\mathrm{e}^{tC_t} - 1) + \mathrm{e}^{tC_t}\big)\eta(t),
\end{aligned}
$$

where the last line uses Lemma D.8. □

The proof of Theorem 6.20 in Case 1 and Case 2 now follows easily.

Proof of Theorem 6.20 Case 1 and Case 2. Case 1: Let $(\Omega, \mathcal{F}, \mathbb{P})$ supporting a collection of i.i.d. Brownian motions $B_\infty = (B_{i,j})_{(i,j)\in\mathbb{N}^2}$ and a collection of i.i.d. Uni($[0,1]$) random variables $\{U_j\}_{j\in\mathbb{N}}$. We define an IEA X on this probability space, by setting $X = \Gamma(u)(t)\{\infty\} + B_\infty$.

Let $r \in \mathbb{N}$ be fixed. Consider the $r \times r$ sampled submatrix $(n\mathcal{E}_n)\{r\}$ out of $n\mathcal{E}_n$. Note that with probability at least $1 - \frac{r^2}{n}$, the coordinates of $(n\mathcal{E}_n)\{r\}$ are distinct. In other words, $(n\mathcal{E}_n)\{r\}$ is a (uniformly) random $r \times r$ submatrix of $n\mathcal{E}_n$ with probability at least $1 - r^2/n$. On this event, we further assume that $(n\mathcal{E}_n)\{r\}(i,j)$ is driven by the same Brownian motion $B_{i,j}$ for every $(i,j) \in [r]^2$.

On the above event, we couple $(n\mathcal{E}_n)\{r\}$ with $X[r]$ where $X[r](i,j) \coloneqq X_{i,j}$ for $(i,j) \in [r]^2$. That is, $X[r]$ is the principle $r \times r$ submatrix of IEA X. Now observe that on this event, using Lemma D.6 we obtain

$$\mathbb{W}_2^2((n\mathcal{E}_n)\{r\}(t), X[r](t)) \le 2e_{n,r}(t) + 2\operatorname{Var}\left[n\sum_{k=2}^{\infty}\widetilde{J}_k(t)\right] \le 2e_{n,r}(t) + 2\left(\frac{r^2}{n}\right),$$

where

$$e_{n,r}(t) \coloneqq \|\Gamma_n(t)\{r\} - \Gamma(u)(t)\{r\}\|_{\mathrm{F}}^2.$$

Now notice that

$$\mathbb{E}\left[\sup_{s\in[0,t]} e_{n,r}(s)\right] \le 2r^2 \sup_{s\in[0,t]}\|m_n(s) - m(s)\|_2^2.$$

By our assumption we have that $\sup_{s\in[0,t]}\|m_n(s) - m(s)\|_2 \to 0$ as $n \to \infty$. It follows that $\sup_{s\in[0,t]}\mathbb{W}_2^2((n\mathcal{E}_n)\{r\}(t), X[r](t)) \to 0$ – in probability – as $n \to \infty$. This completes the proof in Case 1.

Case 2: Recall from the Case 2, we have that

$$n^{1/2}\mathcal{E}_n = \sqrt{n}\Gamma\left(\int_0^{\cdot}\mu_n A_n(s)\,\mathrm{d}s\right) + \sqrt{n}\Gamma\left(\frac{B_n}{\sqrt{n}}\right) + n^{1/2}\sum_{k=2}^{\infty}\widehat{J}_k,$$

where $\mathrm{Var}\Big[n^{1/2}\sum_{k=2}^{\infty}\widehat{J}_{k,(i,j)}(t)\Big] \leq \frac{C_t}{n}$. By our assumption, we have that $\big\|\sqrt{n}\big(\mathrm{Texp}\big[\int_0^{\cdot}\mu_n A_n(s)\,\mathrm{d}s\big](t) - I_n\big)\big\|_{\max} \leq \frac{C_t}{\sqrt{n}}$. Now observe that $\sqrt{n}\Gamma\Big(\frac{B_n}{\sqrt{n}}\Big)$ has i.i.d. coordinates with entries distributed as a time changed BM $t \mapsto B(\mathrm{e}^t - 1)$.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space supporting an IEA B_∞ and that we can define a copy of B_n for every $n \in \mathbb{N}$ on the same probability space such that $\sqrt{n}\Gamma\Big(\frac{B_n}{\sqrt{n}}\Big)(t) = B_\infty[n](\mathrm{e}^t - 1)$. With this coupling, it is immediate that

$$\mathbb{W}_2^2\big((\sqrt{n}\mathcal{E}_n)\{r\}(t), B_\infty[r](t)\big) \leq 8\bigg(\frac{r^2 C_t}{n}\bigg),$$

It follows that $\mathbb{W}_2^2((\sqrt{n}\mathcal{E}_n)\{r\}(t), B_\infty[r](t)) \to 0$ – with probability 1 – as $n \to \infty$. This completes the proof of Case 2. We should note that the condition $\big\|\sqrt{n}\big(\mathrm{Texp}\big[\int_0^{\cdot}\mu_n A_n(s)\,\mathrm{d}s\big](t) - I_n\big)\big\|_{\max} \leq \frac{C_t}{\sqrt{n}}$ is enough to guarantee this conclusion and this condition follows as long as the entries of $\sup_{s\in[0,t]}\|A_n(s)\|_{\max} \leq C$. In particular, for Case 2, we do not need $K(A_n)$ converging to a kernel w.

□

We are now ready to state the proof of Theorem 6.20 for Case 3.

Proof of Theorem 6.20 Case 3. The proof in Case 3 is also similar but with some technical differences. Therefore, we first present a heuristic argument before giving the rigorous proof. Recall from the decomposition in Case 3, we have that

$$\widehat{\mathcal{E}}_n \coloneqq n(\mathrm{Texp}[Y_n] - \mathrm{Texp}[\sigma_n B_n]) = \mathcal{E}_{n,\mathrm{det}} + \mathcal{E}_{n,0} + \mathcal{E}_{n,1}.$$

Set

$$\mathcal{E}_{n,\mathrm{det}} = n\bigg(\mathrm{Texp}\bigg[\int_0^{\cdot}\mu_n A_n(s)\,\mathrm{d}s\bigg] - I_n\bigg)$$
$$\mathcal{E}_{n,0} = \sum_{k=2}^{\infty} n\widehat{J}_{k,0}, \qquad \mathcal{E}_{n,1} = \sum_{k=2}^{\infty} n\widehat{J}_{k,1}.$$

The proof strategy is similar to the first two cases with some technical differences that we explain first. From Lemma D.7, we have that entries of $\mathcal{E}_{n,1}$ have variance $O(1/n)$. Now,

notice that $\mathcal{E}_{n,\det} = n\Gamma\big(\int_0^{\cdot} \mu_n A_n(s)\,\mathrm{d}s\big)$. We know from our assumption that $K(\mathcal{E}_{n,\det})$ converges in L^2 to $\Gamma(u)(t)$. In particular, a randomly chosen $r \times r$ submatrix $\mathcal{E}_{n,\det}\{r\}$ of $\mathcal{E}_{n,\det}$ converges to $\Gamma(t)\{r\}$ in probability. And, $\mathcal{E}_{n,0}$ is a matrix with Gaussian processes. However, unlike the previous cases, the entries of $\mathcal{E}_{n,0}$ are correlated. This makes the coupling more delicate. From Lemma D.7 that the covariance kernel of $\mathcal{E}_{n,0}$ is given by C_n.

Roughly, the idea of the proof is as follows. Let $\widetilde{\mathcal{E}}_n = \mathcal{E}_{n,\det} + \mathcal{E}_{n,0}$. Ignoring the $O(1/n)$ term $\mathcal{E}_{n,1}$, we notice that

$$\widetilde{\mathcal{E}}_n\{r\} = \mathcal{E}_{n,\det}\{r\} + G_{n,r},$$

where $G_{n,r}$ is an $r \times r$ matrix of mean zero Gaussian processes with covariance kernel given by $K_{n,r}$ such that $K_{n,r}(((i_1,j_1),t_1),((i_2,j_2),t_2)) = C_n(((x_{i_1},x_{j_1}),t_1),(x_{i_2},x_{j_2}),t_2))$, where $x_l = \lceil nU_l \rceil$ for every $l \in [r]$. An important observation to make here is that $\mathcal{E}_{n,\det}\{r\}$ and $G_{n,r}$ are conditionally independent given $\{U_i\}_{i\in[r]}$. From our assumptions, it follows that $K(C_n)$ converges in L^2 to a covariance kernel C_∞. On some probability space we construct a Gaussian process $G_\infty \coloneqq (G_{i,j})_{(i,j)\in\mathbb{N}^2}$ such that $G_{i_1,j_1}(t_1)$ and $G_{i_2,j_2}(t_2)$ have a covariance of $K_\infty(((i_1,j_1),t_1),((i_2,j_2),t_2)) = C_\infty(((U_{i_1},U_{j_1}),t_1),((U_{i_2},U_{j_2}),t_2))$ for every $(t_1,t_2) \in [0,1]^2$ and $(i_1,j_1),(i_2,j_2) \in [r]^2$. Since $K_{n,r}$ and $K_{\infty,r}$ are close and it is reasonable to believe that $G_{n,r}$ and $B[r] = (B_{i,j})_{(i,j)\in[r]^2}$ are close. As we have already argued in Case 1, the matrix $\mathcal{E}_{n,\det}(t)\{r\} \approx \Gamma(t)\{r\}$. We therefore conclude that $\widetilde{\mathcal{E}}_n\{r\}$ is close to $\Gamma\{r\} + B_\infty\{r\}$.

We now give a formal proof for completeness. Let $\mathcal{I} = [0,1]^2 \times \mathbb{R}_+$ and let $C\colon \mathcal{I}^2 \to \mathbb{R}$ be a covariance kernel defined as

$$Q(((x_1,y_1),t_1),((x_2,y_2),t_2)) \coloneqq \delta_{x_1=y_1,x_2=y_2} \min\{t_1,t_2\},$$

for $(x_1,y_1;t_1),(x_2,y_2;t) \in \mathcal{I}$. Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space on which we can define a Gaussian process B that is indexed for every $(x,y;t) \in \mathcal{I}$ with covariance kernel Q. Let $G(x,y;\,\cdot\,)$ be a process defined as

$$G(x,y;t) = \int_0^t \int_0^1 \mathrm{d}B(x,z;s)\Gamma_1(s)(z,y)\,\mathrm{d}z\,\mathrm{d}s + \int_0^t \int_0^1 \Gamma_2(s;t)(x,z)B(z,y;s)\,\mathrm{d}z\,\mathrm{d}s, \tag{D.43}$$

for every $(x, y; t) \in \mathcal{I}$. Notice that G has a covariance kernel given by C_∞ defined in Lemma D.7.

Possibly after extending the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ we assume that it supports a collection of i.i.d. Uni($[0,1]$) random variables $\{U_i\}_{i\in\mathbb{N}}$ independent of B. Define an IEA Y by setting

$$Y_{i,j} \coloneqq \Gamma(t)(U_i, U_j) + G(U_i, U_j; t), \tag{D.44}$$

for all $(i,j) \in \mathbb{N}^2$ and $t \in \mathbb{R}_+$.

On this probability space $(\Omega, \mathcal{F}, \mathbb{P})$ we now define a copy of $\widetilde{\mathcal{E}}_n\{r\}$. To do this, we first define a process G_n indexed by $\mathcal{I}$ as

$$G_n(x, y; t) \coloneqq \int_0^t \int_0^1 \mathrm{d}B(x, z; s) K(\Gamma_{n,1}(s))(z, y)\,\mathrm{d}z \tag{D.45}$$

$$+ \int_0^t \int_0^1 K(\Gamma_{n,2}(s; t))(x, z) B(z, y; s)\,\mathrm{d}z\,\mathrm{d}s. \tag{D.46}$$

Now define an $r \times r$ matrix $Y_{n,r}$ such that

$$Y_{n,r}(i,j)(t) \coloneqq \Gamma_n(t)(U_i, U_j) + G_n(U_i, U_j; t), \qquad (i,j) \in [r]^2.$$

With probability at least $1 - r^2/n$, we have that $\lceil nU_i \rceil$s are distinct for $i \in [r]$. And, note that on this event, given $U_1, \ldots, U_r$, we have that $Y_{n,r}$ has the same law as $\widetilde{\mathcal{E}}_n\{r\}$. In particular – with probability at least $1 - r^2/n$ – we obtain that

$$\begin{aligned}
\mathbb{W}_2^2(Y(t)[r], \widehat{\mathcal{E}}_n(t)\{r\}) &\leq 2\mathbb{W}_2^2(Y(t)[r], Y_{n,r}(t)) + 2C_t\frac{r^2}{n} \\
&\leq 2\|\Gamma_n(t)\{r\} - \Gamma(t)\{r\}\|_{\mathrm{F}}^2 + \|E_{n,r,1}(t)\|_{\mathrm{F}}^2 + \|E_{n,r,2}(t)\|_{\mathrm{F}}^2 + 2C_t\frac{r^2}{n},
\end{aligned}$$

where

$$\begin{aligned}
E_{n,r,1,(i,j)}^2(t) &\coloneqq \int_0^t \left(\int_0^1 (K(\Gamma_{n,1}(s))(z, U_j) - \Gamma_1(s)(z, U_j))\,\mathrm{d}z \right)^2 \mathrm{d}s \\
&\leq \int_0^t \int_0^1 |K(\Gamma_{n,1}(s))(z, U_j) - \Gamma_1(s)(z, U_j)|^2\,\mathrm{d}z\,\mathrm{d}s, \\
E_{n,r,2,(i,j)}^2(t) &\coloneqq \int_0^t \int_0^t \int_0^1 \min\{s_1, s_2\}\xi(U_i, z, s_1, t)\xi(U_i, z, s_2, t)\,\mathrm{d}z\,\mathrm{d}s_1\,\mathrm{d}s_2 \\
&\leq t \int_0^1 \left(\int_0^t \xi(U_i, z, s, t)\,\mathrm{d}s \right)^2 \mathrm{d}z \leq t^2 \int_0^t \int_0^1 |\xi(U_i, z, s, t)|^2\,\mathrm{d}z\,\mathrm{d}s,
\end{aligned}$$

where $\xi(U_i, z, s, t) = K(\Gamma_{n,2}(s;t))(U_i, z) - \Gamma_2(s;t)(U_i, z)$, for $(i,j) \in [r]^2$. Now observe that

$$\begin{aligned}\mathbb{E}\big[\|\Gamma_n(t)\{r\} - \Gamma(t)\{r\}\|_{\mathrm{F}}^2\big] &= r^2\|K(\Gamma_{n,1})(t) - \Gamma_1(t)\|_2^2 \\ \mathbb{E}\big[\|E_{n,r,1}\|_{\mathrm{F}}^2\big] &\leq r^2\int_0^t \|K(\Gamma_{n,1}(s)) - \Gamma_1(s)\|_2^2\,\mathrm{d}s, \\ \mathbb{E}\big[\|E_{n,r,2}\|_{\mathrm{F}}^2\big] &\leq r^2t^2\int_0^t \|K(\Gamma_{n,2}(s;t)) - \Gamma_2(s;t)\|_2^2\,\mathrm{d}s.\end{aligned}$$

Define

$$\eta_{n,r}(t) \coloneqq 2\,\mathbb{E}\big[\|\Gamma_n(t)\{r\} - \Gamma(t)\{r\}\|_{\mathrm{F}}^2 + \|E_{n,r,1}(t)\|_{\mathrm{F}}^2 + \|E_{n,r,2}(t)\|_{\mathrm{F}}^2\big] + 2C_t\frac{r^2}{n}.$$

By our assumption and Lemma D.9, it follows that $\eta_{n,r}(t) \to 0$ as $n \to \infty$. We conclude that $\mathbb{W}_2^2(Y(t)[r], \widehat{\mathcal{E}}_m(t)\{r\}) \to 0$ as $n \to \infty$ – in probability. □

D.3.2 Proofs of Chapter 6.3.3

We will prove Theorem 6.23 in this section. We start with some simple observations that intuitively explains why the result holds before giving the formal proof.

Let $f \in L^2([0,1])$. Define $f_n \in \mathbb{R}^n$ by setting $f_{n,i} = \int_{(i-1)/n}^{i/n} f(x)\,\mathrm{d}x / \int_{(i-1)/n}^{i/n}$ for $i \in [n]$. Note that $\frac{1}{n}\|f_n\|_2^2 \leq \|f\|_2^2$. Let X_n be an $n \times n$ matrix. Notice that

$$\|K(X_n)f\|_2^2 = \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{n}\sum_{j=1}^n X_{n,(i,j)}f_{n,j}\right)^2 \leq \|f\|_2^2\left(\frac{1}{n^2}\sum_{i,j=1}^n X_{n,(i,j)}^2\right). \tag{D.47}$$

In particular, if $X_{n,(i,j)}$s are mean 0 random variables with variance bounded by ς_n^2, then

$$\mathbb{E}\big[\|K(X_n)f\|_2^2\big] \leq \varsigma_n^2\|f\|_2^2.$$

Note that this is giving an upper bound on the operator norm of $K(X_n)$. In particular, if $\varsigma_n \to 0$ as $n \to \infty$ then $\|K(X_n)\|_{\mathrm{op}} \to 0$.

Notice that the above bound does not require any assumption on the correlations between the entries of X_n. Taking all $X_{n,(i,j)}$s to be equal, we see that – in general – one can not do better than this. However, when entries of the $X_{n,(i,j)}$ are uncorrelated this bound is clearly weak. Intuitively, when the entries in the row i, $X_{n,(i,j)}$, are uncorrelated (and the variances

bounded by say ς_n^2) we expect the variance of $\frac{1}{n}\sum_{j=1}^n X_{n,(i,j)} f_{n,j}$ to be at most $\frac{1}{n}\varsigma_n^2\|f\|_2^2$. In particular, $\mathbb{E}\big[\|K(X_n)f\|_2^2\big] \le \frac{\varsigma_n^2}{n}\|f\|_2^2$. Therefore, $K(X_n)$ converges to the zero operator as $n \to \infty$, even if $\varsigma_n^2 = O(1)$ as $n \to \infty$. In particular, if X_n is a matrix with i.i.d. Gaussian coordinates, then it converges to a non-trivial IEA, but the limit of X_n, in operator sense, is the zero operator.

With this discussion, the proof is immediate. We write

$$n\mathcal{E}_n \coloneqq \mathcal{E}_{n,\mathrm{det}} + \mathcal{E}^0_{n,\mathrm{err}} + \mathcal{E}^1_{n,\mathrm{err}},$$

where $\mathcal{E}_{n,\mathrm{det}}$ is an $n \times n$ matrix that converges to a deterministic kernel or operator in strong sense and $\mathcal{E}^0_{n,\mathrm{err}}$ is a matrix with i.i.d. Gaussian coordinates with mean 0 and bounded variance while $\mathcal{E}^1_{n,\mathrm{err}}$ is a matrix with mean 0 coordinates and variances bounded by $\frac{1}{n}$. It is clear from the above discussion that the proof follows if we could show that $\mathcal{E}_{n,\mathrm{det}}$ converges to the desired operator in strong sense.

Proof of Theorem 6.23. We begin the proof in Case 1. Set

$$\mathcal{E}_{n,\mathrm{det}} = n\left(\mathrm{Texp}\left[\int_0^{\cdot} \mu_n A_n(s)\,\mathrm{d}s\right]\right), \quad \mathcal{E}^0_{n,\mathrm{err}} = B_n, \quad \text{and} \quad \mathcal{E}^1_{n,\mathrm{err}} = n\sum_{k=2}^{\infty} \widetilde{J}_k.$$

Following equation (D.36), we write

$$n\,\mathrm{Texp}[Y_n] = \mathcal{E}_{n,\mathrm{det}} + \mathcal{E}^0_{n,\mathrm{err}} + \mathcal{E}^1_{n,\mathrm{err}}.$$

Let $f \in L^2([0,1])$. Observe that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{s\in[0,t]} \|(K(n\,\mathrm{Texp}[Y_n](s)) - \mathcal{E}_{n,\mathrm{det}}(s))f\|_2^2\right] \\
&\le 2\,\mathbb{E}\left[\sup_{s\in[0,t]} \left(\big\|K(\mathcal{E}^0_{n,\mathrm{err}}(s))f\big\|_2^2 + \big\|K(\mathcal{E}^1_{n,\mathrm{err}}(s))f\big\|_2^2\right)\right] \le \frac{D_t}{n}\|f\|_2^2,
\end{aligned}$$

where $D_t \ge 1$ is a constant that depends only on t.

Let T_n be the integral operator corresponding to A_n and set $\eta_n(t) \coloneqq \sup_{s\in[0,t]}\|T_n(s) - T(s)\|_{\mathrm{op}}$. Similarly define the integral curve S_n of T_n. It follows that

$$\sup_{s\in[0,t]} \|\mathrm{Texp}[S_n](s) - \mathrm{Texp}[S](s)\|_{\mathrm{op}} \le \eta_n(t) t \mathrm{e}^{tC_t} \to 0,$$

as $n \to \infty$.

The proof in the Case 2, follows exactly from the same argument, by noting that $\|\mathcal{E}_{n,\det}\|_\infty \leq \frac{1}{\sqrt{n}}$ in this case. We skip the details. $\square$

Appendix E

PROOFS OF CHAPTER 7

In this entire chapter, we will provide the proof of Theorem 7.1.

Recall from the definition of $\mathrm{Texp}[Y_n]$ that it is an infinite sum of the k-fold integrals of the form $J_k(Y_n)(t) \coloneqq \int_{\Delta_k(t)} \mathrm{d}Y_n(s_k) \dots \mathrm{d}Y_n(s_1)$. Since Y_n is a martingale of the form $\mathrm{d}Y_n = \frac{1}{n} A_n(t)\,\mathrm{d}t + \frac{1}{\sqrt{n}}\,\mathrm{d}B_n(t)$, the $J_k(Y_n)(t)$ further decomposes into 2^k terms each involving k-fold integrals of the products of the terms like $n^{-1}A_n(s)\,\mathrm{d}s$ and $n^{-1/2}\,\mathrm{d}B_n(s)$.

We collect all the integrals only involving the integrand $n^{-1}A_n(s)\,\mathrm{d}s$ to get $\Gamma\big(\frac{1}{n}\int_0^{\cdot} A_n(s)\,\mathrm{d}s\big)(t)$. Similarly, we can collect all integrals only involving the integrands of the form $n^{-1/2}\,\mathrm{d}B_n(s)$ to obtain $\Gamma\Big(\frac{1}{\sqrt{n}}B_n\Big)(t)$. This suggests that we can write $\mathrm{Texp}[Y_n](t)$ as in (7.5). That is, we have

$$\mathrm{Texp}[Y_n](t) = I_n + \Gamma\left(\frac{1}{n}\int_0^{\cdot} A_n(s)\,\mathrm{d}s\right)(t) + \Gamma\left(\frac{1}{\sqrt{n}}B_n\right)(t) + \frac{1}{n}Z_n(t) + \frac{1}{n}E_n(t), \qquad \text{(E.1)}$$

where Z_n is the sum of all k-fold integrals of the form defined in Definition 3.5 with exactly one $n^{-1/2}\,\mathrm{d}B_n$ term which appears at either the first or the last integral and E_n is the sum of remaining terms.

Note that both Z_n and E_n are mean zero Gaussian processes. The benefit of the above decomposition is that, we can make following observations:

1. $n\Gamma\big(\frac{1}{n}\int_0^{\cdot} A_n(s)\,\mathrm{d}s\big)(t)$ converges to $\Gamma(u)(t)$ in L^2 under our assumptions.

2. The $\Gamma\Big(\frac{1}{\sqrt{n}}B_n\Big)(t)$ have i.i.d. Gaussian coordinates with mean zero and variance $\frac{1}{n}(\mathrm{e}^t - 1)$.

3. We show that $Z_n(t)$ has $O(1)$ coordinates with some non-trivial correlation, but we can compute these correlations explicitly.

4. $E_n(t)$ has entries with variances of order $O(1/n)$.

Let $H_{n,0}$ be as in the assumption. It follows from the above observation that

$$\begin{aligned}\operatorname{Texp}[Y_n](t)H_{n,0} = H_{n,0} + \Gamma\left(\frac{1}{n}\int_0^{\cdot} A_n(s)\,\mathrm{d}s\right)(t)H_{n,0} + \Gamma\left(\frac{1}{\sqrt{n}}B_n\right)(t)H_{n,0} \\ + \frac{1}{n}Z_n(t)H_{n,0} + O\big(n^{-1}\big).\end{aligned} \tag{E.2}$$

Now the proof idea is as follows. Let V be a uniform $[0,1]$ random variable. Let $X_n(t) \coloneqq K(\operatorname{Texp}[Y_n](t)H_{n,0})(V)$. Note that $X(t)$ is precisely a coordinate of $\operatorname{Texp}[Y_n](t)H_{n,0}$ chosen uniformly at random. Ignoring the $O(1/n)$ contribution from E_n, we observe from (E.2) that

$$\begin{aligned}X_n(t) \approx h_0(V) + K\left(\Gamma\left(\frac{1}{n}\int_0^{\cdot} A_n(s)\,\mathrm{d}s\right)(t)H_{n,0}\right)(V) + K\left(\Gamma\left(\frac{1}{\sqrt{n}}B_n\right)(t)H_{n,0}\right)(V) \\ + K\left(\frac{1}{n}Z_n(t)H_{n,0}\right)(V).\end{aligned}$$

From our assumption, it follows that $K(\Gamma\big(\frac{1}{n}\int_0^{\cdot} A_n(s)\,\mathrm{d}s\big)(t)H_{n,0})$ converges in L^2 to $\Gamma(U)(t)h_0$ and hence $K(\Gamma\big(\frac{1}{n}\int_0^{\cdot} A_n(s)\,\mathrm{d}s\big)(t)H_{n,0})(V)$ converges in probability to $(\Gamma(U)(t)h_0)(V)$. On the other hand, we $\Gamma\Big(\frac{1}{\sqrt{n}}B_n\Big)(t)H_{n,0}$ and $\frac{1}{n}Z_n(t)H_{n,0}$ are independent mean 0 Gaussian processes. And, the covariance of $\Gamma\Big(\frac{1}{\sqrt{n}}B_n\Big)(t)H_{n,0} = (\mathrm{e}^t - 1)\|h_0\|_2 I_n$ and similarly the covariance of $\frac{1}{n}Z_n(t)H_{n,0}$ is $C_t(h)I_n + O\big(\frac{1}{n}\big)$. We combine these to conclude the first part of the proof.

For part 2, we use the same idea. But instead, we take and i.i.d. collection of $V_1, \ldots, V_k$ uniform $[0,1]$ random variables and consider the vector $X_{n,k}(t)$ where $X_{n,k}(t)(i) = K(\operatorname{Texp}[Y_n](t)H_{n,0})(V_i)$. The desired conclusion follows from the fact that the covariance of $\Gamma\Big(\frac{1}{\sqrt{n}}B_n\Big)(t)H_{n,0} = (\mathrm{e}^t - 1)\|h_0\|_2 I_n$ and similarly the covariance of $\frac{1}{n}Z_n(t)H_{n,0}$ is $C_t(h)I_n + O\big(\frac{1}{n}\big)$.

Lemma E.1. *Let C_n be as defined in Lemma D.7 for every $n \in \mathbb{N}$. Let $h_0 \in L^2([0,1])$ and $\{U_i\}_{i\in\mathbb{N}}$ be i.i.d.* $\mathrm{Uni}([0,1])$ *random variables. Define $H_{n,0} = h_0(U_i)$ for every $i \in [n]$ and $n \in \mathbb{N}$. Then for any $i_1, i_2 \in [n]$,*

$$\lim_{n\to\infty} \mathbb{E}\left[\left(\frac{1}{n}Z_n(t)H_{n,0}\right)_{i_1}\left(\frac{1}{n}Z_n(t)H_{n,0}\right)_{i_2}\right] = \mathbb{1}\{i_1 = i_2\}\int_0^t \|\Gamma(U)(s)h_0\|_2^2\,\mathrm{d}s. \tag{E.3}$$

Proof. Following the definition of C_n,

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{n}Z_n(t)H_{n,0}\right)_{i_1}\left(\frac{1}{n}Z_n(t)H_{n,0}\right)_{i_2}\right] \\
&= \frac{1}{n^2}\sum_{j_1,j_2=1}^{n} H_{n,j_1}\,\mathbb{E}\big[Z_{n,(i_1,j_1)}(t)Z_{n,(i_2,j_2)}(t)\big]H_{n,j_2} \\
&= \frac{1}{n^2}\sum_{j_1,j_2=1}^{n} H_{n,j_1}C_n(((i_1,j_1),t),((i_1,j_1),t))H_{n,j_2}
\end{aligned}$$

Separating the terms when $j_1 = j_2$ and otherwise, the above simplifies to

$$\begin{aligned}
&\frac{1}{n^2}\sum_{j_1=j_2=1}^{n} H_{n,j_1}C_n(((i_1,j_1),t),((i_1,j_1),t))H_{n,j_2} \\
&\qquad + \frac{1}{n^2}\sum_{j_1\neq j_2=1}^{n} H_{n,j_1}C_n(((i_1,j_1),t),((i_1,j_1),t))H_{n,j_2}
\end{aligned}$$

First, consider the case when $i_1 \neq i_2$. As we take the limit of $n \to \infty$, the first term goes to zero, whereas the second term is exactly zero. For the case when $i_1 = i_2$, the first term again goes to zero as $n \to \infty$, but the second term does not. Plugging in the expression for C_n in this case, we get that the above converges to

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{n^2}\sum_{j_1,j_2=1}^{n} H_{n,j_1}H_{n,j_2}\int_0^t \big(\Gamma_{n,1}(s)^\top\Gamma_{n,1}(s)\big)(j_1,j_2)\,\mathrm{d}s \\
&= \lim_{n\to\infty}\int_0^t \frac{1}{n}\sum_{k=1}^{n}\frac{1}{n^2}\sum_{j_1,j_2=1}^{n} H_{n,j_1}H_{n,j_2}\Gamma_{n,1,(k,j_1)}(s)\Gamma_{n,1,(k,j_2)(s)}\,\mathrm{d}s \\
&= \lim_{n\to\infty}\int_0^t \frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{n}\sum_{j_1=1}^{n}\Gamma_{n,1,(k,j_1)}(s)H_{n,j_1}\right)\left(\frac{1}{n}\sum_{j_2=1}^{n}\Gamma_{n,1,(k,j_2)}(s)H_{n,j_2}\right) \\
&= \lim_{n\to\infty}\int_0^t \frac{1}{n}\sum_{k=1}^{n} K(\Gamma_{n,1}H_{n,0})^2(k/n) \\
&= \int_0^t\int_0^1 (\Gamma(s)h_0)^2(z)\,\mathrm{d}z\,\mathrm{d}s = \int_0^t \|\Gamma(s)h_0\|_2^2\,\mathrm{d}s,
\end{aligned}$$

where the last statement holds using Lemma D.9. This completes the proof. □

Recall that by our assumption there exists a continuous curve $t \mapsto w(t)$ of kernels such that $\sup_{s\in[0,t]}\|K(A_n)(s) - w(s)\|_2 \to 0$ as $n \to \infty$. Furthermore, we assume that

$\sup_{s\in[0,t]}\|w(s)\|_\infty \le C$. Under these assumption, the kernel $\Gamma(u)(t) \coloneqq \sum_{k=1}^\infty J_k(u)(t)$ is well defined and $\|\Gamma(u)(t)\|_\infty \le \mathrm{e}^{Ct}-1$. In the following, we will use the notation $\Gamma(t)$ instead of $\Gamma(u)(t)$ for simplicity. Let us also define the kernel $\Gamma_n(t) = nK\big(\sum_{k=1}^\infty J_k\big(\frac{A_n}{n}\big)\big) = \sum_{k=1}^\infty J_k(K(A_n))$. It follows from our assumption that $\sup_{s\in[0,t]}\|\Gamma_n(s)-\Gamma(s)\|_2 \to 0$ as $n\to\infty$. Let

$$\Gamma_1(s) \coloneqq \Gamma(u)(s), \quad \Gamma_2(s;t) \coloneqq \mathrm{Texp}[U](t-s)\odot w(s), \qquad s\in[0,t], \quad t\in[0,1],$$

where on $L^2([0,1]^2)$, we define the product $u\odot v \in L^2([0,1]^2)$ as $(u\odot v)(x,y) = \int_0^1 u(x,z)v(z,y)\,\mathrm{d}z$.

We are now ready to prove Theorem 7.1.

Proof of Theorem 7.1. Let $H_n(t) \coloneqq \mathrm{Texp}[Y_n](t)H_{n,0}$ be as in the statement of the Theorem. Let $k\in\mathbb{N}$ be given and let $V_1,\dots,V_k$ be i.i.d. Uni$([0,1])$ random variables. Let $X_{n,k}(t)\in\mathbb{R}^k$ be the vector defined as

$$X_{n,k,i}(t) = H_n(V_i)(t), \qquad i\in[k].$$

Given $V_1,\dots,V_k$, we notice that $X_{n,k}$ is a Gaussian process with mean $M_{n,k}\in\mathbb{R}^k$ where

$$M_{n,k,i}(t) \coloneqq K\left(\Gamma\left(\frac{1}{n}\int_0^\cdot A_n(s)\,\mathrm{d}s\right)(t)H_{n,0}\right)(V_i), \qquad i\in[k],$$

and covariance $\mu_n+\rho_n+e_n$ where following Lemma E.1,

$$\mu_n(t)(i_1,i_2) = \mathbb{1}\{i_1=i_2\}\big(\mathrm{e}^t-1\big)\|H_{n,0}\|_2^2/n,$$
$$\rho_n(t)(i_1,i_2) = \mathbb{1}\{i_1=i_2\}\int_0^t \frac{1}{n}\left\|\frac{1}{n}\Gamma_{n,1}(s)H_{n,0}\right\|_2^2\mathrm{d}s + \mathbb{1}\{i_1\neq i_2\}\cdot O\big(n^{-1}\big), \qquad (i_1,i_2)\in[n]^2.$$
$$e_n(t)(i_1,i_2) = O\big(n^{-1}\big)\|H_{n,0}\|_2^2/n.$$

It follows from D.9 that ρ_n converges to ρ as $n\to\infty$ where

$$\rho(t)(i_1,i_2) = \mathbb{1}\{i_1=i_2\}\int_0^t \|\Gamma(U)(s)h_0\|_2^2\,\mathrm{d}s,$$

μ_n converges to μ as $n\to\infty$ where

$$\mu(t)(i_1,i_2) = \mathbb{1}\{i_1=i_2\}\big(\mathrm{e}^t-1\big)\|h_0\|_2^2,$$

and e_n converges to zero as $n \to \infty$. From D.8, we have that $K\big(\Gamma\big(\frac{1}{n}\int_0^{\cdot} A_n(s)\,\mathrm{d}s\big)(t)H_{n,0}\big)$ converges uniformly to $\Gamma(U)(t)h_0$ in $L^2([0,1])$ as $n \to \infty$. It follows that $M_{n,k}(t)$ converges uniformly, in probability, to the vector M_k where $M_{k,i} = (\Gamma(U)(t)h_0)(V_i)$ for $i \in [k]$.

Since $X_{n,k}$ is Gaussian process with mean $M_{n,k}$ and covariance $\mu_n + \rho_n + e_n$ and $M_{n,k}$ converges to M_k and $\mu_n + \rho_n + e_n$ converges to $\mu + \rho$. It follows that $X_{n,k}$ converges to a Gaussian process X_k with mean M_k and covariance $\mu + \rho$. Since $\mu + \rho$ is a multiple of identity, it follows that the coordinates of X_k are independent. This proves the second part. The first part follows simply by taking $k = 1$. □