Stochastic Analysis on Graphons

Raghavendra Tripathi

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

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Soumik Pal, Chair

Krysztof Burdzy

Zaid Harchaoui

Stefan Steinerberger

Program Authorized to Offer Degree:
Mathematics

University of Washington

Abstract

Stochastic Analysis on Graphons

Raghavendra Tripathi

Chair of the Supervisory Committee:
Thesis Advisor Soumik Pal
Department of Mathematics

We say that a function on the space of symmetric matrices is *invariant* if it is invariant under the simultaneous permutation of rows and columns of the input matrix by the same permutation. Homomorphism densities of finite simple graphs offer a rich class of examples of invariant functions that prominently feature in several areas of mathematics. In this thesis, we study two natural classes of interrelated problems. The first problem concerns the optimization of invariant functions in large dimensions. The second problem concerns the analysis of the dynamics on graphs/matrices where the evolution of coordinates depends on the full graph/matrix via an invariant function as the dimension goes to infinity. An important theme of the present thesis is that due to the symmetry of invariant functions, their optimization and dynamics can be reduced to optimization and dynamics on the space of graphons as the dimension of the underlying space goes to infinity. The rich geometry and the analytical properties of the space of graphons make the problems on the space of graphons more tractable.

We develop a notion of gradient flow on the space of graphons following the general theory of gradient flows in metric spaces. We show that under mild differentiability assumption, any invariant function on the space of graphons admits a gradient flow which is an absolutely continuous curve with respect to the invariant L^2 metric. Furthermore, under appropriate convexity and differentiability assumptions, we show that the Euclidean gradient flows of invariant functions converge to the gradient flow of a suitable function on

the space of graphons.

We then consider a class of symmetric $n \times n$ matrix-valued diffusions where the drift is given by an invariant function. Such diffusions arise, for example, as the scaling limits of stochastic gradient descent of an invariant function. We establish a propagation of chaos phenomenon for such matrix-valued processes. That is, we show that any finite collection of coordinates of such processes becomes conditionally independent as $n \to \infty$ and that a uniformly random coordinate of such processes satisfies a novel graphon McKean-Vlasov SDE, in $n \to \infty$ limit. As a consequence of this, we obtain that these matrix-valued processes converge to a deterministic curve on the space of graphons.

We also construct a Metropolis chain, with a novel relaxation step, whose state space is the stochastic block model with r communities and n individuals in each community. We show that fixed r, under appropriate scaling of parameters, the $r \times r$ matrix of connection probabilities between communities converges to a diffusion of the previous type. In particular, as $r \to \infty$, the connection probability between communities becomes conditionally independent. This allows us to prove that the trajectory of this Metropolis chain is concentrated near a deterministic curve of graphons. This allows us to approximate the gradient flow of function on graphons by suitable Markov chains on stochastic block models.

Towards the end of the thesis, we also consider the scaling limit of the iterated product of matrices that are small perturbations of the identity matrix as the dimension of these matrices goes to infinity. In the fixed dimension, the scaling limit of the iterated product of such matrices is described by a non-commutative exponential of a matrix-valued semimartingale. Suppose that the bounded variation part of these semimartingales converges to some graphon as the dimension of these matrices goes to infinity. Then, we show that non-commutative exponentials converge to an infinite exchangeable array whose coordinates are Gaussians and whose mean and covariances can be described explicitly in terms of the limiting graphon.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

मन्दः कवियशःप्रार्थी गमिष्याम्युपहास्यताम्। प्रांशुलभ्ये फले लोभादुद्बाहुरिव वामनः॥1/3

Will I, a dunce, be subjected to mockery if I were to desire a poet's fame,
like a dwarf overstretching arms for a fruit obtainable only by the tall?

अथवा कृतवाग्द्वारे वंशेऽस्मिन्पूर्वसूरिभिः। मणौ वज्रसमुत्कीर्णे सूत्रस्येवास्ति मे गतिः॥1/4

Or, thanks to the doors of speech created by earlier sages in this dynasty,
my course will be easy like threading a diamond pierced by a diamond.

Raghuvaṃśam by Kālidāsa

Desirous of the poet's fame, Kalidasa's apprehensions in narrating the stories of Raghavs are assuaged by the groundwork of previous sages. I write this thesis with a similar desire and trepidation. Undoubtedly, this endeavor would have been impossible without the foundational work of the *Pūrvasūrāḥ*, which served as both the bedrock and inspiration for this work, along with the invaluable support of countless individuals. I take this opportunity to thank them all.

To begin, I must thank my thesis advisor Soumik Pal. He has been incredibly patient and kind throughout this journey. His constant support and generosity made navigating through difficult times possible. Under his guidance, I have grown both as a mathematician and as a person. When the research seemed stagnant, he pointed out that the progress in research is like Brownian local time. While insisting and imploring me to do examples, he shared with me the story of *Śhwetaketu* from *Chāndogyopaniṣad*. I will forever cherish these and many more such memories.

During the research presented in this thesis, I had the privilege to collaborate with Siva Athreya, Soumik Pal, Zaid Harchaoui, Sewoong Oh, and Raghav Somani. Each of them has left a lasting impression on my way of thinking, my research style, and my mathematical maturity, for which I am deeply grateful. I am particularly grateful to Siva Athreya for

hosting me in Bangalore and for his kind words and encouragement. I must also thank Zaid Harchaoui and Sewoong Oh for their funding support throughout my graduate school.

I would also like to express my appreciation to my committee members, Krzysztof Burdzy, Zaid Harchaoui, Bamdad Hosseini, Sewoong Oh, and Stefan Steinerberger, for graciously agreeing to serve on my committee and creating a supportive environment. Their insightful suggestions for improvement were invaluable.

In the past five years, the world has undergone unprecedented and unpredictable changes. During these times, learning mathematics and conducting research was made possible through the extraordinary dedication of department staff, the kindness and excellence of faculty members, the generosity of the donors to the department, and the supportive and vibrant community of students at the University of Washington. I am deeply thankful to all those who directly or indirectly supported me throughout this journey.

I owe a special thanks to Krzysztof Burdzy for his invaluable help in learning stochastic processes during my first year of graduate school and for his general kindness and chocolates. My discussions with Stefan Steinerberger have been immensely beneficial; they not only contributed to my research but also exposed me to a breadth of beautiful mathematics. Stefan's infectious enthusiasm for mathematics will stay with me forever. I also thank Zhen-Qing Chen for his course on Gaussian Free Fields and Chris Hoffman for his course on Ergodic theory.

This acknowledgment would be incomplete without mentioning the friends at the University of Washington and in Seattle who made the difficult times bearable. I thank Thessalonika for being a wonderful host and helping me on the very first day I arrived in Seattle. I am particularly grateful to Arkamouli Debnath, Soham Ghosh, Junaid Hasan, Garret Mulcahy, Andrea Ottolini, Hadrian Quan, and Anjali Yadav.

Andrea has been a trusted friend and a reliable proofreader. I am grateful for the opportunity to collaborate with Andrea; he has been my go-to person for help and advice, both in mathematics and beyond.

A special thank you goes to Junaid. We lived next door to each other for four years, and

every day is filled with memories. Together, we cooked, enjoyed tea, discussed mathematics, and mastered the art of making pizzas.

My mathematical journey truly began at the Indian Institute of Science (IISc), Bangalore. I owe whatever mathematical knowledge I possess to the faculty members and friends at IISc, whom I sincerely thank. In particular, Manjunath Krishnapur is to blame for my becoming a probabilist. His courses and guidance, both inside and outside the classroom, have profoundly influenced my journey. I cherish our discussions in *Prakruthi* (Prakṛti) canteen. Without his suggestion and encouragement, I wouldn't have applied to graduate school. As he puts it, I was let go on good behavior, and I hope I haven't completely wasted my parole.

I also fondly remember and thank R. Basu, G. Bharali, S.K. Iyer, A. Khare, G. Misra, E.K. Narayanan, and S. Thangavelu for their courses from which I benefited immensely and their encouragement and support. Any mention of IISc would be incomplete without the wonderful friends I made at IISc. In particular, I want to thank Shubham Rastogi, Poornendu Singh, Mayuresh Londhe, and Debashree Behera (Kajal).

I sincerely thank my teachers at the University of Delhi (DU). In particular, I must express my gratitude to Mukund Madhav Mishra, who encouraged me to pursue mathematics. I am glad that I listened to him. As the powerful play goes on, I can contribute a verse to it because he taught me the rhyme.

Finally, I must thank my family and friends who, despite their unceasing complaints (mostly justified), have been my strongest supporters. I cannot thank my parents, my aunt and uncle, and my elder sister enough. With them in my life, every day is a blessing. I will not even attempt to express my feelings for Ankur, Dhiraj, Durgesh, Nitesh, Roshan, Shivam, and Siddharth, as any such attempt is bound to fail.

In such a long and arduous journey I have come across numerous people who have left deep impressions on me. It is simply impossible to name each one of them. Nonetheless, I sincerely acknowledge all those who remain unnamed and who have enriched me.

DEDICATION

to my family

Chapter 1

INTRODUCTION

Graphs and matrices are arguably the most ubiquitous objects in mathematics. Many natural phenomena in the real world and mathematics are modeled by a (possibly weighted) graph. We think of a (symmetric) matrix as the adjacency matrix of a weighted graph. Therefore, in the following discussion graphs and matrices will be interchangeable. For concreteness and simplicity, throughout this chapter, a matrix will refer to a symmetric matrix with entries in $[0, 1]$, unless stated otherwise.

In this thesis, we study two natural classes of interrelated problems: optimization and dynamics on the space of graphs and matrices. The first problem concerns the optimization of functions – with some symmetry– on the space of graphs/matrices in very large dimensions. The second problem concerns the analysis of dynamics on graphs/matrices where the evolution of coordinates depends on the full matrix in a symmetric fashion. We will describe the symmetry more precisely later. These two problems are very intimately connected. For instance, a rich and interesting class of dynamics on the graphs or matrices arises from optimization problems. Algorithms like stochastic gradient descent yield a natural class of dynamics on Euclidean spaces. Another important class of examples in this regard is Markov chains to sample from a stationary distribution. Often sampling from a Gibbs measure is used as a technique to find approximate minimizers of some function on Euclidean space. Of course, one may consider other dynamics that may not necessarily arise from optimization, and in later chapters, we will consider general classes of evolution.

Besides symmetry, another important theme of our work is that we consider the limiting behavior of the dynamics or optimization problems as the dimension of the underlying space grows to infinity. This philosophy is motivated by practical considerations where many problems of interest are inherently high dimensional. In general, high-dimensional problems are harder to analyze. On the other hand, these problems often possess some useful

symmetries – which essentially reduce the problem to some other space. The underlying philosophy here is that because of the symmetry, complex high-dimensional dynamics are essentially controlled by some feature or statistics and not by the precise detail of the full system. This allows us to take the dimension of the underlying space to infinity. In the limit, it is enough to describe the evolution of this feature or statistics. Very often, the space in which these features live has a rich geometric and analytical structure. This makes the analysis of the limiting description of this feature or statistics more tractable.

This philosophy has been successfully used in many areas of probability in the past. The most pertinent example for us is the interacting system of particles under the so-called mean-field interaction that we explain below. In the following section, we explain these ideas in more detail using some simple and well-known examples from interacting particle systems. In this thesis, we deal with the evolution of graphs, but the following discussion on interacting particle systems serves as a useful analogy.

1.1 Interacting particle systems: Symmetry and propagation of chaos

Roughly speaking, in this section we will see that a symmetric function on Euclidean space can be thought of as a function on the space of measures. Therefore, to describe the evolution of a particle system with symmetric interaction (referred to as mean-field interaction), it is sufficient to describe the evolution of the empirical measure of the particles in the system. An important consequence of this symmetric interaction is that while every particle interacts with every other particle in the system, the effect of any given particle on another particle is mild. This leads to a phenomenon called *propagation of chaos*– which effectively means that, in the limit, particles behave as if they were independent. Because of this heuristic, the empirical measure of the system of particles is roughly the same as the law of a random particle. This reduces the study of the ensemble behavior to the study of a random particle or its law. The evolution of the law of a single particle is, in turn, described by a partial differential equation (PDE).

Example 1 (Symmetric functions are functions of measures)**.** Let $V, W : \mathbb{R} \to \mathbb{R}$ be two sufficiently smooth functions. Assume for simplicity that W is symmetric, that is, $W(x) =$

$W(-x)$ and $W(x) \geq 0$. Throughout this chapter, we will tacitly assume the necessary integrability assumptions on V, W wherever needed. Consider a function $H_n : \mathbb{R}^n \to \mathbb{R}$ as

$$H_n(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} V(x_i) + \frac{1}{n^2} \sum_{i,j=1}^{n} W(x_i - x_j) \; .$$

Let S_n denote the set of permutation σ on $[n]$. Observe that the function $H_n : \mathbb{R}^n \to \mathbb{R}$ is permutation-invariant, that is,

$$H_n(x_{\sigma(1)}, \ldots, x_{\sigma(n)}) = H(x_1, \ldots, x_n) \; ,$$

for any permutation $\sigma \in S_n$ and any $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Because of this symmetry, we can treat the function H_n as a function on $\mathcal{P}(\mathbb{R})$, the space of the probability measures on $\mathbb{R}$. Philosophically, the function H_n depends on $x = (x_1, \ldots, x_n)$ only via a feature of x, namely, its empirical measure. More precisely, define a function $\mathcal{H} : \mathcal{P}(\mathbb{R}) \to \mathbb{R}$ by

$$\mathcal{H}(\mu) = \int V(x)\mu(\mathrm{d}x) + \int \int W(x - y)\mu(\mathrm{d}x)\mu(\mathrm{d}y) \; .$$

Notice that if $x = (x, \ldots, x_n) \in \mathbb{R}^n$ and $\mu_x \coloneqq \frac{1}{n} \sum_{j=1}^{n} \delta_{x_i}$ is the empirical measure generated by x, then $H_n(x) = \mathcal{H}(\mu_x)$. The key takeaway of this example is the following. The permutation-invariance of the function H_n allows us to treat it as a function on the space of probability measures and conversely a function $\mathcal{H}$ on $\mathcal{P}(\mathbb{R})$ gives rise to symmetric functions $H_n : \mathbb{R}^n \to \mathbb{R}$ for each n.

Now consider the problem of minimizing H_n on $\mathbb{R}^n$ for large n. A commonly used technique to minimize the function H_n would be to consider the gradient flow of H_n:

$$\dot{x}(t) = -n \nabla H_n(x(t)) \; ,$$

where $x(t) \in \mathbb{R}^n$ and $\dot{x}(\cdot)$ denote the time derivative of x. Notice that we have scaled the gradient $-\nabla H_n(x(t))$ by a factor of n. This changes the speed of the gradient flow and it is important while considering the large n limit. In more detail, the gradient flow of H_n is described by a system of n ordinary differential equations:

$$\dot{x}_i(t) = -V'(x_i(t)) - \frac{2}{n} \sum_{j=1}^{n} W'(x_i(t) - x_j(t)) \quad \text{i.e.,}$$

$$\dot{x}_i(t) = -V'(x_i(t)) - 2 \int W'(x_i(t) - y)\mu_n(t)(\mathrm{d}y) \; , \tag{1.1}$$

for $i = 1, \dots, n$ and where $\mu_n(t)$ denotes the empirical measure of $x(t) = (x_1(t), \dots, x_n(t))$.

In practice, while simulating (1.1), the exact values for the gradient $\nabla H_n(x(t))$ are either unavailable or computationally expensive to obtain. Therefore, one often uses noisy estimates for $\nabla H_n(x(t))$ [187]. This can be modeled by adding independent Brownian noise to each coordinate of (1.1). That is, one often considers the process the $\mathbb{R}^n$ valued process $X_n(\cdot)$ defined as

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

where $B_n(t)$ is standard n-dimensional Brownian motion and $\beta > 0$ is a fixed parameter. Or in more detail, for $i = 1, \dots, n$ the coordinate $X_{i,n}$ satisfies the stochastic differential equation (SDE)

$$\mathrm{d}X_{i,n}(t) = -V'(X_{i,n})\,\mathrm{d}t - 2\int W'(X_{i,n}(t) - y)\mu_n(t)(\mathrm{d}y)\ \,\mathrm{d}t + \sqrt{\frac{2}{\beta}}\,\mathrm{d}B_{i,n}(t)\ , \tag{1.2}$$

where $B_{i,n}$ is a standard 1-dimensional Brownian motion and $\mu_n(t)$ is the empirical measure of the vector $X_n(t)$.

We will think of the evolution of X_n as a noisy version of the gradient flow of H_n. One can also view the above equation as describing the evolution of n identical particles under mean-field interaction. The term mean-field interaction here means that a given particle say $X_{i,n}$ in the ensemble depends on the full ensemble only via the empirical measure of the full ensemble. The mean-field evolution of particle systems has a long and rich history and such a system of particles has been studied since Kac [123] and McKean [157]–even without optimization context.

As pointed out in the first example, H_n is only a function of the empirical measure. It makes sense, therefore, to ask how the empirical measure of $x(t)$ in (1.1) or the empirical measure of $X_n(t)$ in (1.2) evolves as $n \to \infty$. For concreteness, let us consider the evolution of the empirical measure $\mu_n(t) \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{X_{i,n}(t)}$. Note that $\mu_n(t)$ is a random measure. The following question is natural to ask: suppose that $\mu_n(0) = \frac{1}{n}\sum_{i=1}^n \delta_{X_{i,n}(0)}$ converges to some deterministic measure $\mu(0)$ as $n \to \infty$. Does it follows that $\mu_n(t)$ converges to probability measure $\mu(t)$ as $n \to \infty$? If so, how is the curve $t \mapsto \mu(t)$ related to the function $\mathcal{H}$ on the space of probability measures on $\mathbb{R}$?

Let us denote by $\mu_n(t)$ the empirical measure of the vector $X_n(t)$ defined in (1.2). Let $\widetilde{\mu}_n(t) \coloneqq \mathbb{E}[\mu_n(t)]$ be the expected empirical measure of $X_n(t)$. A simple application of Ito's formula shows that the expected empirical measure $\widetilde{\mu}_n(t)$ satisfies the following PDE

$$\partial_t \mu(t) = \nabla(\mu(t) \cdot v(\mu(t))) + \frac{1}{\beta}\Delta\mu(t) \ , \tag{1.3}$$

where

$$v(\mu)(x) = V'(x) + 2\int W'(x-y)\mu(\mathrm{d}y) \ .$$

It is perhaps not very surprising that $\mu_n(t)$ converges (say in probability) to a curve $\mu(t)$ that satisfies (1.3). This amounts to showing that the empirical measure $\mu_n(t)$ concentrates near its expectation $\widetilde{\mu}_n(t)$. For instance, this would be true if the coordinates of $X_n(t)$ were independent and identically distributed (i.i.d.). A powerful consequence of the symmetry in the interaction (that is, mean-field interaction) is that, as $n \to \infty$, the coordinates of $X_n(t)$ indeed become independent asymptotically. Therefore, the random measures $\mu_n(t)$ converge (almost surely) to a deterministic measure $\mu(t)$ that satisfies (1.3), this phenomenon is referred to as *propagation of chaos*. We explain this in the following example.

Example 2. Assume that at time $t = 0$ we initialize $X_n(0)$ so that each coordinate is i.i.d. with some distribution, say μ_0. Note that at any time $t > 0$, the coordinates of $X_n(t)$ are correlated. However, because of the mean-field interaction, any collection of finitely many coordinates of $X_n(t)$ becomes independent as $n \to \infty$. Therefore, it is enough to study the evolution of one particle in the limit. Let I be a uniformly random coordinate chosen from $[n]$. Using (1.2), we can describe the evolution of a randomly chosen coordinate of $X_{I,n}$ satisfies

$$\mathrm{d}X_{I,n}(t) = -V'(X_{I,n})\,\mathrm{d}t - \int W'(X_{I,n}(t) - y)\mu_n(t)(\mathrm{d}y)\ \mathrm{d}t + \sqrt{\frac{2}{\beta}}\,\mathrm{d}B(t) \ , \tag{1.4}$$

where $B(t)$ is a standard 1-dimensional Brownian motion, and $\mu_n(t)$ is the empirical measure of $X_n(t)$.

If the coordinates of $X_n(t)$ were i.i.d. then $\mu_n(t)$ will be approximately equal to the law of a random coordinate. In the current setup, the coordinates are not independent but because of the propagation of chaos phenomenon, this heuristic remains valid. And,

consequently $X_{I,n}$ converges, as $n \to \infty$, to the solution of the so-called McKean-Vlasov SDE give as

$$\begin{aligned} \mathrm{d}X(t) &= -V'(X(t))\,\mathrm{d}t - 2\int W'(X(t)-y)\mu_t(\mathrm{d}y) + \sqrt{\frac{2}{\beta}}\,\mathrm{d}B(t)\,, \\ \mu_t &= \mathrm{Law}(X(t))\,. \end{aligned} \tag{1.5}$$

It is easy to see by an application of Ito's formula that $\mu(t) = \mathrm{Law}(X(t))$ satisfies (1.3). In other words, the evolution described by (1.3) essentially describes the evolution of the law of one particle. The convergence of $\mu_n(t)$ to $\mu(t)$ now becomes an analogue of the law of large numbers.

The moral of the above example is that *symmetry gives rise to propagation of chaos*. The propagation of chaos reduces the dynamics of a complex system of particles to study the dynamics of a single random particle. We refer an interested reader to [52, 51] for a modern and exhaustive discussion of the propagation of chaos phenomenon and McKean-Vlasov SDE.

1.2 From particle gradient flow to gradient flow of measures

We have seen that, under mean-field interaction, the evolution of the empirical measure of the particle system can be approximately described by (1.3). We also argued that the process $X_n(t)$ can be thought of as the gradient flow of H_n in the presence of noise. As H_n corresponds to a function $\mathcal{H}$ on the space of measures, it is natural to wonder if the PDE (1.3) can be interpreted as a gradient flow of $\mathcal{H}$ (or some perturbation of $\mathcal{H}$) on the space of probability measures? A powerful and deep result due to [122] shows that the answer is yes.

The space of probability measures on $\mathbb{R}$ with the finite second moment, denoted by $\mathcal{P}_2(\mathbb{R})$, can be equipped with the so-called Wasserstein metric $\mathbb{W}_2$ [212]. The space $(\mathcal{P}_2(\mathbb{R}), \mathbb{W}_2)$ is referred to as the Wasserstein space. The Wasserstein space admits – at least formally – a Riemannian structure [91, 174, 173]. Using this formalism, one can define a notion of gradient flow on $\mathcal{P}_2(\mathbb{R})$. We refer the reader to [193, 192] for a gentle introduction to gradient flows on Wasserstein space and [5] for a general treatment of gradient flow in metric

spaces.

It turns out that we can indeed interpret (1.3) as the gradient flow of a function $\mathcal{F}_\beta : \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$ defined as

$$\mathcal{F}_\beta(\mu) = \mathcal{H}(\mu) + \frac{1}{\beta}\mathcal{E}(\mu) \,,$$

where $\mathcal{E}(\mu) = \int \rho(x) \log(\rho(x))\, \mathrm{d}x$ if μ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$ and $\rho(x) = \frac{\mu(\mathrm{d}x)}{\mathrm{d}x}$ is the density of μ with respect to the Lebesgue measure and $\mathcal{E}(\mu) \equiv +\infty$ otherwise. The function $\mathcal{E}$ is called the entropy. The presence of the entropy forces the minimizer of $\mathcal{F}_\beta$ to have a density with respect to the Lebesgue measure. This can be contrasted with the presence of the Brownian noise in (1.2) that diffuses the vector $X_n(t)$.

Note that when $\beta = \infty$, the function $\mathcal{F}_\infty = \mathcal{H}$. Furthermore, the flow defined by the PDE

$$\partial_t \mu_t = \nabla \cdot (v(\mu_t) \cdot \mu_t),$$

describes the gradient flow of $\mathcal{H}$ on $\mathcal{P}_2(\mathbb{R})$. Without any additional convexity assumption on $\mathcal{H}$, the gradient flow is hard to analyze. The function $\mathcal{E}$ is a strongly convex function on $\mathcal{P}_2(\mathbb{R})$. Thus, $\mathcal{F}_\beta$ can be seen as a regularization of $\mathcal{H}$. Even if $\mathcal{H}$ is only convex, the function $\mathcal{F}_\beta$ is a strongly convex function and therefore admits a unique minimizer. This makes the flow (1.3) more well-behaved.

The point of this example is that, under permutation symmetry, a function on Euclidean space corresponds to a function on the Wasserstein space. Furthermore, the gradient flow of a function on the Wasserstein space can be approximated by the empirical measure of a particle system.

1.3 Sampling and Optimization

As we mentioned earlier, our perspective on interacting particle systems is that these particle systems are inspired by optimization problems. Optimization on Euclidean spaces is very intimately connected with sampling. In this section, we explain this connection.

For a fixed $n \in \mathbb{N}$, consider the SDE described by (1.2). Under appropriate growth assumptions on V and W, there is a unique stationary measure ρ_n for the SDE (1.2) with

density (with respect to the Lebesgue measure on $\mathbb{R}^n$) proportional to $e^{-\beta H_n(x)}$ [186]. Such probability measures are called Gibbs measure. For large $\beta > 0$, this stationary measure is concentrated near minimizers of H_n. In other words, finding approximate minimizers of H_n is equivalent to sampling from certain Gibbs measures. From this perspective, the SDE (1.2) is seen as a process to sample from the Gibbs measure with density $\propto e^{-\beta H_n(x)}$. In practice, people use many other variants for instance Metropolis chain to sample from Gibbs measures.

Furthermore, under mild continuity and convexity assumption on V, W, the measure ρ_n converges, as $n \to \infty$, to a measure $\rho \in \mathcal{P}_2(\mathbb{R})$ that is the unique minimizer of $\mathcal{H}$. As mentioned in the previous example, the flow described by the PDE in (1.3) also converges to the minimizer of $\mathcal{H}$, ρ, as $t \to \infty$. This completes a full circle of the ideas, namely, symmetric functions on $\mathbb{R}^n$ correspond to a function on $\mathcal{P}(\mathbb{R})$ and the minimization of such a function on $\mathbb{R}^n$ also corresponds to a minimization problem on $\mathcal{P}(\mathbb{R})$, that is, the minimizer of the symmetric function on $\mathbb{R}^n$ converges to the minimizer of the corresponding function on $\mathcal{P}(\mathbb{R})$ and the schemes like (noisy) gradient flow on $\mathbb{R}^n$ correspond to the gradient flow (of suitable regularization) of the corresponding function on $\mathcal{P}(\mathbb{R})$. One may often exploit the richness of the geometry of infinite dimensional space (Wasserstein space in this case) to get more insights into the optimization problems in finite-dimensional Euclidean space.

1.4 Symmetric functions on graphs and problems

The objects of interest for us are the functions on graphs or symmetric matrices that are invariant under the conjugation action of the permutation group. In this section, we describe this symmetry and ask natural questions about the dynamics of graphs or matrices with symmetric interaction.

1.4.1 What are invariant functions

Let X be a non-empty subset of $\mathbb{R}$ and let $\mathcal{M}_n(X)$ be the set of $n \times n$ symmetric matrices taking values in the set X. For a permutation $\sigma \in S_n$, and $A \in \mathcal{M}_n(X)$ we define A^σ to be the matrix such that $A^\sigma(i,j) = A(\sigma(i), \sigma(j))$. That is, the rows and columns of A are permuted by the same permutation σ to obtain A^σ. We are interested in the functions

$H_n : \mathcal{M}_n(X) \to \mathbb{R}$ such that $H_n(A^\sigma) = H_n(A)$ for all $A \in \mathcal{M}_n(X)$. Throughout this thesis, we call such functions to be *invariant functions*.

To understand the motivation behind invariant functions, let us consider the symmetric matrices with entries $\{0, 1\}$, that is, matrices in $\mathcal{M}_n(\{0, 1\})$. Let $G = ([n], E)$ be a graph on the vertex set $[n]$. Any such graph is represented by a matrix $A_G \in \mathcal{M}_n(\{0, 1\})$ called the adjacency matrix of G where $A_G(i, j) = 1$ if $\{i, j\} \in E$ and 0 otherwise. Alternatively, any matrix $A \in \mathcal{M}_n(\{0, 1\})$ gives a graph $G_A = ([n], E)$ on the vertex set $[n]$ where $\{i, j\} \in E$ if and only if $A(i, j) = A(j, i) = 1$. With this setup, we notice that two graphs $G_1 = ([n], E_1)$ and $G_2 = ([n], E_2)$ are isomorphic if and only if there are adjacency matrices are related as $A^\sigma_{G_1} = A_{G_2}$ for some $\sigma \in S_n$. That is, an invariant function $H_n : \mathcal{M}_n(\{0, 1\}) \to \mathbb{R}$ is a function that is invariant under graph isomorphism. In this sense, such a function is an honest function of the graph and does not depend on the labeling of the vertices of the graph. As an example, let $H_n : \mathcal{M}_n(\{0, 1\}) \to \mathbb{R}$ be the function

$$H_n(A) = \frac{1}{n^3} \operatorname{Tr}\left[A^3\right] .$$

If we think of A as the adjacency matrix of a graph G, then notice that $\operatorname{Tr}\left[A^3\right]$ is the number of homomorphisms of K_3 (complete graph on 3 vertices) into G. This shows that $\operatorname{Tr}\left[A^3\right]$ is indeed invariant and hence so is H_n. Note that, in general, H_n is not invariant under different permutations applied to the rows and columns of the matrix A. In other words, H_n is not invariant under the permutation of its n^2 many coordinates. Therefore, H_n is not a function of the empirical measure of entries of A.

More generally, let $X = [0, 1]$. Then, we can think of the matrices $A \in \mathcal{M}_n(X)$ as the weighted adjacency matrix of a graph with edge weights in $[0, 1]$. In this situation, an edge-weight 0 means that the edge is not present. In later chapters, we often work with $X = [-1, 1]$ (which can be replaced with any compact interval of $\mathbb{R}$), but for the current discussion, we set $X = [0, 1]$. With this setup, the invariant functions on $\mathcal{M}_n(X)$ are precisely the functions that do not depend on the labeling of the vertices– and hence are the true functions of the underlying graph (or isomorphism class of graphs).

1.4.2 Optimization and dynamics on graphs

Following our discussion of symmetric functions on $\mathbb{R}^n$, we can now ask several questions about the optimization of symmetric functions on graphs that we deal with in this thesis.

For concreteness, we will fix an example to illustrate and explain the questions that we ask. In this discussion, we will use $\mathcal{M}_n$ to denote $\mathcal{M}_n([0,1])$. The following invariant function on $\mathcal{M}_n$ will serve as our constant example throughout this chapter. Define

$$R_n(A) = \frac{1}{n^3}\operatorname{Tr}\left[A^3\right] - \frac{\alpha}{n^2}\sum_{i,j} A(i,j)\ , \tag{1.6}$$

for some fixed $\alpha > 0$. If A were the adjanceny matrix of a simple graph $G = ([n], E)$, then $\operatorname{Tr}\left[A^3\right]$ counts the number of homomorphism of K_3 into G. In other words, $\operatorname{Tr}\left[A^3\right]$ counts roughly the number of triangles in G.[1] The normalization n^3 can be thought of as the number of maps from K_3 into G. Therefore, we refer to $\frac{1}{n^3}\operatorname{Tr}\left[A^3\right]$ as *triangle density* function. It can be interpreted as the probability that a random map $K_3 \to G$ is a homomorphism. Similarly, we call $\frac{1}{n^2}\sum_{i,j} A(i,j)$ as the *edge density* function. When A is a matrix with entries in $[0,1]$, we will continue to call these functions triangle density function and edge-density function. Of course, in this case, we need to interpret $\operatorname{Tr}\left[A^3\right]$ as the weighted sum of triangles in the weighted graph corresponding to the matrix A, where each triangle has a weight that is given by the product of the weights of the edges in that triangle.

1.4.3 Problems

Equation (1.6) defines the functions R_n on $\mathcal{M}_n$ for each n. As we already noted, R_n is not a function of the empirical measure of the entries of the matrices. However, these functions R_n are clearly restrictions of a single function on the 'space of weighted graphs with edge-weights in $[0,1]$', say, $\mathcal{G}_\infty(X)$. Compare this with Example 1 where we noted that a symmetric function on Euclidean space corresponds to a function of empirical measure. We then interpreted symmetric functions on Euclidean spaces as functions on the space of

[1] We are being a little imprecise here. The function $\operatorname{Tr}\left[A^3\right]$ counts every triangle 3 times, that is, each triangle is counted with all possible relabelling of the vertices. But we will ignore this small issue.

probability measures. In the current setting, we will see that $\mathcal{G}_\infty(X)$ can be thought of as a dense subset of a suitable space of functions called the space of graphons and denoted as $\widehat{\mathcal{W}}$. Furthermore, the functions R_n can be interpreted as (the restriction of) some function, say R on $\widehat{\mathcal{W}}$. We give a brief introduction to graphons in Section 1.5. A more detailed discussion can be found in Chapter 2.

We now ask several questions analogous to the interacting particle systems and optimization on Euclidean spaces.

Question 1.4.1. *Suppose we are interested in minimizing the function R_n on $\mathcal{M}_n$ for large n. For a n, the function R_n is a smooth function of the entries and naturally one can run gradient flow for R_n. That is, define a symmetric matrix-valued flow by*

$$\frac{\mathrm{d}}{\mathrm{d}t}A_n(i,j)(t) = -\nabla_{i,j}R_n(A_n(t)), \qquad (i,j) \in [n]^2\,, \tag{1.7}$$

where $\nabla_{i,j}R_n(A)$ is the partial derivative of R_n with respect to the (i,j)-th coordinate.

Is there any way to take the limit of the gradient flow $t \mapsto A_n(t)$ defined by (1.7) as $n \to \infty$ to obtain a limiting curve on the space $\widehat{\mathcal{W}}$? If so, can this limiting curve be interpreted as a gradient flow of R on $\widehat{\mathcal{W}}$? Can we obtain (an approximate) minimizer of R_n from the minimizer of R on $\widehat{\mathcal{W}}$?

A careful reader will note that the flow defined by (1.7) is not well-defined for all time t. This is because a priori the function R_n is only defined on $\mathcal{M}_n([0,1])$ and the flow (1.7) may take a coordinate outside $[0,1]$ in finite time. To ensure that the gradient flow is always inside $\mathcal{M}_n$, we need to add a correction term that constrains the coordinates to remain inside $[0,1]$. This is an important point and we carefully handle this in later chapters. However, in the current discussion, we ignore this for the clarity of exposition.

Recall that in Example 2, we mentioned that the Euclidean gradient flows are often modified by independent additive noise in each coordinate. This gave rise to a particle system with mean-field interaction in (1.2). Following the same idea, let us consider a noisy analogue of (1.7) given by

$$\frac{\mathrm{d}}{\mathrm{d}t}A_n(i,j)(t) = -\nabla_{i,j}R_n(A_n(t)) + \frac{2}{\sqrt{\beta}}\,\mathrm{d}B_n(i,j)(t), \qquad (i,j) \in [n]^2\,, \tag{1.8}$$

where $\beta > 0$ is a constant and B_n is an $n \times n$ symmetric matrix such that $\{B_n(i,j) : i \leq j\}$ is a collection of i.i.d. Brownian motions.

We can interpret (1.8) as a random perturbation of the gradient flow of R_n from an optimization point of view. More generally, one can think of (1.8) as interacting particle systems. Note that in this setting the particles are labeled by the coordinate (on and above the diagonal) of the matrix (or the edge of the graph). Thus, there are $n(n-1)/2$ particles. More importantly, the interaction between these particles is not mean-field. Therefore, the classical theory of mean-field interacting particle systems does not directly apply in this case.

Question 1.4.2. *Observe that the evolution of a random coordinate of* (1.8) *depends on all other coordinates is only via an invariant function of the full matrix A_n and an independent Brownian noise. However, it is not a function of the empirical measures of the coordinates. It is natural to ask if this symmetry is enough to guarantee the propagation of chaos– in some appropriate sense.*

Assuming that there is a propagation of chaos, the evolution of a random coordinate should be described by a McKean-Vlasov SDE albeit the role of the measure (law of a random coordinate) needs to be replaced by a graphon (an element of the space $\widehat{\mathcal{W}}$). This leads us to the following question: if there is a propagation of chaos, can we describe the evolution of a randomly chosen coordinate of A_n in the limit? Furthermore, analogous to Example 2, can we describe the limiting dynamics of (1.8) *as a deterministic flow on $\widehat{\mathcal{W}}$?*

1.5 Setup and Our contribution

The bulk of this thesis is devoted to answering the above questions. In this section, we provide a high-level answer to these questions and summarize our key contributions. We begin with a brief set-up about the space of graphons, denoted $\widehat{\mathcal{W}}$. This would be analogous to the space of measure in the previous discussion. We refer the reader to [97] for a gentle introduction to graphons and [150] for a textbook exposition.

1.5.1 A very brief introduction to graphons

A *kernel* $W : [0,1]^2 \to [0,1]$ is a symmetric measurable function, that is, $W(x,y) = W(y,x)$. Very roughly, a graphon is an equivalence class of kernels that we define precisely in Chapter 2. When we wish to emphasize the whole equivalence of a kernel W, we will denote it by $[W]$. For the current purpose, we will work with a representative of this equivalence class and hence identify a kernel with its equivalence class when we mean a graphon. The space of graphons $\widehat{\mathcal{W}}$ has a natural metric called cut-metric, denoted, $\delta_\square$. Equipped with the cut-metric, the space of graphons becomes a compact space. However, this metric does not have a good geometry. We, therefore, work with another stronger metric δ_2 that is called invariant L^2 metric for reasons that will be clear in the next chapter. This situation can be compared with optimal transport. The space of probability measures on a compact set is compact under weak convergence. But for most geometrical considerations in optimal transport, one works with 2-Wasserstein metric, $\mathbb{W}_2$. Analogously, in our setup, we will often consider curves on the space of graphons. Absolute continuity of the curves will be always with respect to δ_2 metric while our convergences will be often in weaker metric $\delta_\square$.

To relate the space of graphons with our questions, we need to understand how a kernel or a graphon is related to graphs. To understand this, let us first mention that a kernel W gives rise to an infinite sequence of random graphs as follows. Let $U_1, U_2, \dots$ be a collection of i.i.d. Uniform $[0,1]$ random variables. For any $n \geq 2$, we define a graph on the vertex set $[n]$ as follows. Create an edge between i and j with probability $W(U_i, U_j)$ independently (given $U_1, \dots, U_n$) for each pair of distinct vertices $i \neq j$. Denote this graph by $\mathbb{G}(n, W)$. Note that the edges in $\mathbb{G}(n, W)$ have independent Bernoulli distribution given $U_1, \dots, U_n$ (but they are not identically distributed). If we take $W(x,y) \equiv p$, then $\mathbb{G}(n, W)$ is the same as Erdös-Rényi graph. This construction also hints at the aforementioned equivalence relation that defines kernel. It is natural to identify two kernels W and W' if $\mathbb{G}(n, W)$ and $\mathbb{G}(n, W')$ have the same law for all $n \geq 2$. While this is true, this equivalence relation can be expressed in a more analytically amenable way as we do in the next chapter.

Now we describe how a graph or a symmetric matrix with entries in $[0,1]$ can be identified with a kernel or a graphon. Let A be a symmetric $n \times n$ matrix with entries in $[0,1]$. We

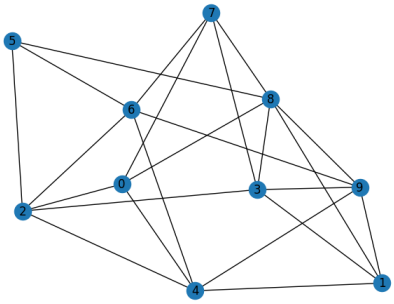
associate to A, a kernel W_A where $W_A(x,y) = A(\lceil nx \rceil, \lceil ny \rceil)$. That is, we divide the $[0,1]^2$ into an $n \times n$ grid and W_A is a step function whose value on the (i,j) cell is the same as $A(i,j)$. Naturally, given a graph G on n vertices, we associate to G a kernel W_A where A is the adjacency matrix of G.

Note that matrices of different dimensions can correspond to the same kernel. For instance, for each $n \geq 2$ let J_n be the $n \times n$ matrix such that $J_n(i,j) = 1$ for all $i,j \in [n]$. Note that $W_{J_n} \equiv 1$. This is analogous to the fact that if $v_n \in \mathbb{R}^n$ is the vector such that $v_n(i) = 1$ for all $i \in [n]$, then the empirical measure $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{v_n(i)}$ corresponding to v_n are the same. The graphon theory is best suited to understand the limit of the graphs and matrices as $n \to \infty$.

In figure 1.1, we illustrate the correspondence of going from a graph to a kernel and from a kernel to a matrix. In figure 1.1, we begin with a graph G which is an instance of an Erdös-Rényi graph $E(10, 0.5)$. The graph G is shown in Figure 1.1a and we show the adjacency matrix A of G next to it. Figure 1.1c shows the kernel W_A, where the black pixels denote the value 1 and the white pixels denote the value 0. Note that in drawing the kernel, the cell at the bottom-left corner denotes the entry corresponding to $A(1,1)$. The axis of symmetry in the kernel is the line $x = y$. We draw 10 i.i.d. Unif$[0,1]$ samples $U_1, \ldots, U_{10}$. We create a graph G' on the vertex set $[n]$ where the edge $\{i,j\}$ is present if $W_A(U_i, U_j) = 1$. That is, G' is a sample from $\mathbb{G}(n, W)$. This is plotted in Figure 1.1d and the corresponding adjacency matrix and kernel are plotted in Figure 1.1e and 1.1f respectively.

Figure 1.1 also illustrates an important point worth remarking. Let G be a graph on n vertices with the adjacency matrix A and let W_A be the kernel corresponding to A. The graph on the vertex set $[n]$ described by $\mathbb{G}(n, W_A)$ is a random graph and it need not be isomorphic to G. However, for a given kernel W, the random graph $\mathbb{G}(n, W)$ converges, almost surely, to W in $\delta_\square$ metric as $n \to \infty$. This is reminiscent of the fact that for a given vector $x = (x_1, \ldots, x_n) \in R^n$ with empirical measure $\mu_n \coloneqq n^{-1}\sum_{i=1}^n \delta_{x_i}$, if we generate n i.i.d. samples, say $y_1, \ldots, y_n$, from μ_n then $(y_1, \ldots, y_n)$ need not be equal to a permutation of x. However, if $\mu \in \mathcal{P}(\mathbb{R})$ is fixed and y_is are i.i.d. samples from μ then $\mu_n \coloneqq n^{-1}\sum_{i=1}^n \delta_{y_i}$ converges, almost surely, to μ in weak sense.

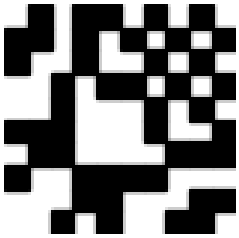
The key takeaway of the above discussion is there is a compact space $(\widehat{\mathcal{W}}, \delta_\square)$ such that



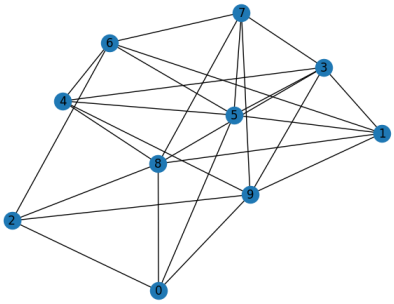
(a) G : a sample of $E(10, 0.5)$

$$\begin{bmatrix}
0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\
1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0
\end{bmatrix}$$

(b) Adjacency matrix A of G



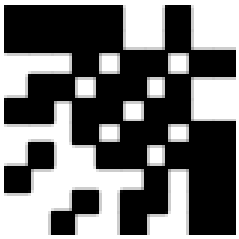
(c) Kernel W_A



(d) G' : a sample of $\mathbb{G}(10, W_A)$

$$\begin{bmatrix}
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\
0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\
0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0
\end{bmatrix}$$

(e) Adjacency matrix A' of G'



(f) Kernel $W_{A'}$

Figure 1.1: Graph to matrix to kernel to graph to matrix to kernel

all weighted graphs with edge-weights in $[0,1]$, and all symmetric matrices with entries in $[0,1]$ sit inside it. Naturally, a function on $\widehat{\mathcal{W}}$ therefore gives rise to a function on the space $\mathcal{M}_n$, the space of symmetric matrices (with entries in $[0,1]$) for each $n \geq 2$. Given a sequence of processes $X_n : [0,\infty) \to \mathcal{M}_n$, one can naturally ask if it has a limit in $\delta_\square$ metric as $n \to \infty$. The processes that we are concerned with are often motivated by optimization problems as explained earlier. In the following section, we relate this background to our questions and we give a chapterwise outline of the rest of the thesis.

1.5.2 Our contributions and an overview of the thesis

Recall that for our current discussion $\mathcal{M}_n$ denotes the set of symmetric $n \times n$ matrices with entries in $[0,1]$. As we already observed, a matrix $A \in \mathcal{M}_n$ corresponds to a kernel W and hence to a graphon. We will use W_A or $K(A)$ to denote the kernel corresponding to a matrix A. Also recall the function $R_n : \mathcal{M}_n \to \mathbb{R}$ defined in Equation (1.6) that we reproduce below

$$R_n(A) = \frac{1}{n^3} \operatorname{Tr}\left[A^3\right] - \frac{\alpha}{n^2} \sum_{i,j} A(i,j) \ .$$

Let us define a function R on the space of kernels $\mathcal{W}$ as

$$R(W) = \int_{[0,1]^3} W(x_1,x_2)W(x_2,x_3)W(x_3,x_1)\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3 - \alpha \int_{[0,1]} W(x,y)\,\mathrm{d}x\,\mathrm{d}y \ . \quad (1.9)$$

It is easily seen that for any $A \in \mathcal{M}_n$, we have $R_n(A) = R(W_A)$. That is, we can interpret the sequence of functions R_n as (the restriction of) a single function on a common space $\widehat{\mathcal{W}}$. One might think that we did not quite use the permutation invariance of R_n. To understand this, note that we have defined the function R on the space of kernels $\mathcal{W}$. But the space of graphons, $\widehat{\mathcal{W}}$, is obtained by suitably quotienting the space $\mathcal{W}$. A function $R : \mathcal{W} \to \mathbb{R}$ descends to a function on $\widehat{\mathcal{W}}$ precisely if R is constant on the equivalence class defining $\widehat{\mathcal{W}}$. In that case, one can define a function on $\widehat{\mathcal{W}}$ (by abuse of notation we use the same symbol R to denote a function on $\widehat{\mathcal{W}}$ now) $R([W]) = R(W)$. This equivalence relation is defined precisely in the next chapter, but it entails that the functions R_n on $\mathcal{M}_n$ can be obtained as the restrictions of some function R on $\widehat{\mathcal{W}}$ precisely when R_n are permutation invariant.

In Chapter 2, we discuss the space of graphons in more detail. In the later chapters, we also need a generalization of graphons called measure-valued graphons. This is also discussed in Chapter 2. Thus, this chapter serves as the background for the remainder of the thesis. While the bulk of this chapter is standard and can be found in [150], it also includes some novel results (taken from [167, 9]) about the space of graphons, space of measure-valued graphons and its relationship with infinite exchangeable arrays. We feel that these results will be of general interest to the community working with graphons. With this background, in Chapter 3, we present a literature survey highlighting different directions of research.

Our main contributions are primarily presented in Chapter 4 to Chapter 7. Finally, in Chapter 8, we provide a summary of the thesis. After giving a summary of the thesis, we describe some possible applications of our current work and some questions and directions that naturally emerge out of our work that need further work. We end this chapter with a brief outline of our results in Chapter 4 to Chapter 7 below.

Chapter 4, broadly speaking, concerns the Question 1.4.1. Following and specializing the theory for the gradient flows in metric spaces in [5], we show that the space of graphons $\widehat{\mathcal{W}}$ equipped with the δ_2 metric admits a notion of gradient flow. More precisely, we introduce a notion of differentiability that we call *Fréchet-like differentiability* for functions defined on $\widehat{\mathcal{W}}$. Under some mild convexity and Fréchet-like differentiability assumptions on a function R, we show that R admits a unique gradient flow curve (more precisely, a curve of maximal slope) starting at any prescribed point. This curve is absolutely continuous with respect to δ_2 metric. Furthermore, given a function R on $\widehat{\mathcal{W}}$ and a kernel $W_0 \in \mathcal{W}$, we can describe a curve

$$W_t(x,y) = W_0 - \int_0^t \mathcal{D}_{\mathcal{W}} R(W_s)(x,y)\,\mathrm{d}s\ ,$$

where $\mathcal{D}_{\mathcal{W}} R$ is the Fréchet-like derivative of R. Then, the curve $t \mapsto W_t$ is absolutely continuous in $L^2([0,1]^2)$ and the this naturally descends to an absolutely continuous curve $(\widehat{\mathcal{W}}, \delta_2)$. This curve defines the unique gradient flow of R on $\widehat{\mathcal{W}}$.

Now consider the Euclidean gradient flow of R_n on $\mathcal{M}_n$ described by

$$\frac{\mathrm{d}}{\mathrm{d}t} A_n(i,j)(t) = -n^2 \nabla_{i,j} R_n(A_n(t)), \qquad (i,j) \in [n]^2\ . \tag{1.10}$$

Furthermore, the kernel valued curve corresponding $t \mapsto A_n(t)$, that is $t \mapsto K(A_n)$ converges, under appropriate assumption on the initial condition, to the gradient flow of R on the space of graphons described above.

We should again emphasize that we need to modify the $\mathcal{D}_{\mathcal{W}}$ if $W(x,y) \in \{0,1\}$ in order to ensure that $W_t(x,y) \in [0,1]$ for all time $t \geq 0$. However, we ignore this detail in the current discussion. Similarly, in (1.10) to ensure that $A_n(i,j)(t) \in [0,1]$ at all time, we need to modify the drift term in a way so that whenever $A_n(i,j)(t) \in \{0,1\}$ the drift forces this coordinate to lie in $[0,1]$.

This part of Chapter 4 is based on the paper *Gradient Flows on Graphons: Existence, Convergence, Continuity Equations*[167] with Sewoong Oh, Soumik Pal and Raghav Somani. As an illustration of this theory, let us mention that for our example function R in Equation 1.9, the gradient flow curve is given by

$$W_t(x,y) = W_0(x,y) - 3\int_0^t W_t(x,z)W_t(z,y)\,\mathrm{d}z + \alpha t\ .$$

Gradient flow of R and Mantel Turán theorem

In this section, we go over our example function R in (1.9) in explain the key outcomes of the theory developed in Chapter 4 in the context of this example. We begin with a celebrated theorem of Mantel [155] (a special case of Turán's theorem).

Theorem 1.5.1. *The maximum number of edges in an n-vertex triangle-free graph is $n^2/4$ and it is achieved by a balanced, complete bipartite graph $K_{n/2,n/2}$.*

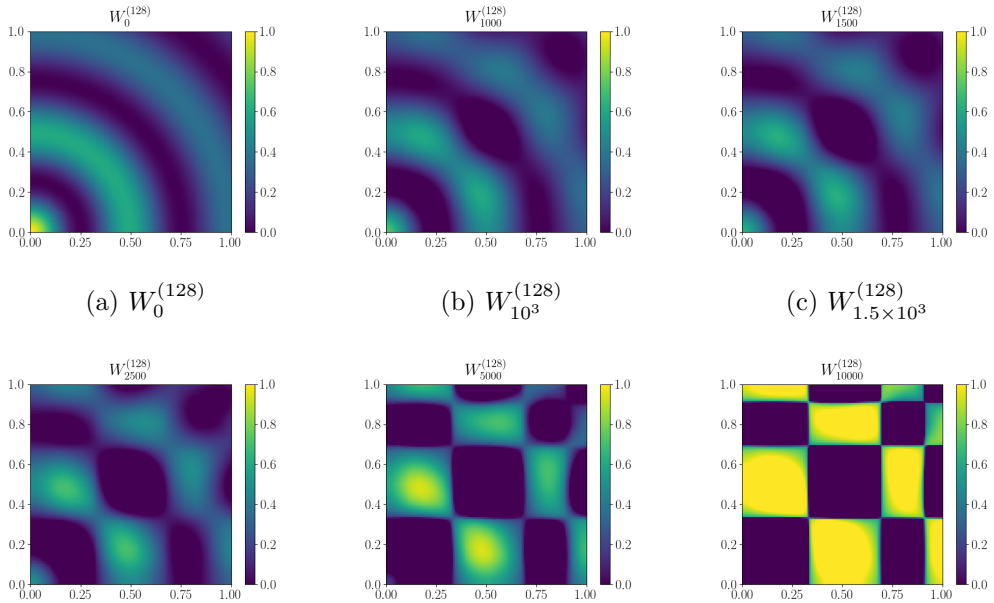
This suggests that if one maximizes the edge density subject to the condition that triangle density is 0, then the maximizer should correspond to a complete bipartite graph. Now observe that the minimizing function

$$R_n(A) = \frac{1}{n^3}\operatorname{Tr}\left[A^3\right] - \frac{\alpha}{n^2}\sum_{i,j} A(i,j)\ \ ,$$

with a small α, say $\alpha = 1/10$, is akin to minimizing triangle density while also maximizing the edge density as much as possible. In particular, if we run the gradient flow of R for a sufficiently long time, we should expect a complete bipartite graph to emerge. In the view of

Theorem 4.1.1 (where we show that the Euclidean gradient flow approximates the gradient flow on $\widehat{\mathcal{W}}$) it is enough to run a Euclidean gradient flow for sufficiently large n.

For numerical simulation, we set $n = 128$ and consider a time discretization with step size $\tau = 10^{-3}$ and use the forward Euler method starting from an initial kernel $\left[W_0^{(n)}\right] \in \widehat{\mathcal{W}}_n$ as shown in Figure 1.2a. Figure 1.2 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 1.2f 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.



(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 1.2: A gradient descent simulation of R

Such optimization problems often arise from extremal combinatorics and large deviations of exponential random graphs. Understanding the structure of minimizers in such problems is often challenging. While our current theory does not say anything about the structure of minimizer(s), it provides a convenient computational tool with to obtain approximate minimizers. This can be especially useful for producing counterexamples.

Chapter 5 is based on the paper *Stochastic optimization on matrices and a graphon McKean-Vlasov limit* [102] with Zaid Harchaoui, Sewoong Oh, Soumik Pal and Raghav Somani. In this part, we study the limit of the noisy version of the gradient flow described in (1.2) that is,

$$\frac{\mathrm{d}}{\mathrm{d}t}A_n(i,j)(t) = -n^2\nabla_{i,j}R_n(A_n(t)) + \frac{2}{\sqrt{\beta}}\,\mathrm{d}B_n(i,j)(t), \qquad (i,j) \in [n]^2\,, \tag{1.11}$$

where $\beta > 0$ is a constant and B_n is an $n \times n$ symmetric matrix such that $\{B_n(i,j) : i \leq j\}$ is a collection of i.i.d. Brownian motions.

Note that the drift in the above equation is slightly different from that in (1.8). To get a non-trivial limit as $n \to \infty$, we need to scale the drift ∇R_n by n^2. We now answer the Question 1.4.2, that is, what happens to this process as $n \to \infty$? Consider the kernel valued process $t \mapsto K(A_n(t))$. We show that there is a deterministic curve of kernels $t \mapsto W(t)$ that is absolutely continuous with respect to the usual L^2 norm on $L^2([0,1]^2)$ such that for any fixed finite time $T > 0$ we have $\sup_{t\in[0,T]}\|K(A_n)(t) - W(t)\|_\square \to 0$, in probability, as $n \to \infty$. This answers part of Question 1.4.2, namely, it shows that the matrix-valued random process converges to a deterministic curve of kernels.

It is useful to recall the analogy with the mean-field interacting particle systems. We think of (1.11) as describing the evolution of $n(n-1)/2$ many particles. Here we think of every entry of the matrix as a particle (but our matrix is assumed to be symmetric). However, as we already pointed out, the system described by (1.11) is not mean-field. In particular, the evolution of the system is not quite captured by the empirical measure of the coordinates. The graphon or kernel corresponding to the matrix A_n captures the state of the evolution in this setting. The statement that $t \mapsto K(A_n)(t)$ converges to a deterministic curve $t \mapsto W(t)$ as $n \to \infty$ is analogous to the fact that the empirical measure process of mean-field particle system converges to a deterministic curve of measures as the ensemble size grows to infinity. In a mean-field particle system, this phenomenon is accompanied by the propagation of chaos.

Our current setting has a close parallel to this. To explain this, let us define an infinite exchangeable array (IEA)see [124, Chapter 7]. An IEA $\mathbf{X}$ 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. For a definitive modern account of exchangeable arrays and their connections with the limits of large graphs, the reader is referred to [74, 13, 14, 15].

We show that the coordinates of $k \times k$ submatrix of A_n, chosen uniformly at random from all submatrices of size k, become asymptotically independent as $n \to \infty$. This allows us to show that A_n converges to an IEA. This means that a randomly (uniformly at random) chosen $k \times k$ submatrix of A_n converges (weakly) to the principle $k \times k$ submatrix of the IEA as $n \to \infty$. This is essentially the propagation of chaos phenomenon. To further the analogy with the mean-field particle system, we also show that the evolution of a randomly chosen coordinate of A_n as $n \to \infty$ is described by an SDE analogous to McKean-Vlasov SDE where the measure is replaced by a kernel. That is, in the limit, we obtain an SDE on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ supporting i.i.d. Uniform$[0,1]$ random variables U, V and a standard Brownian motion B such that on the set $U = x, V = y$, we have

$$\begin{aligned} \mathrm{d}X_{u,v}(t) &= W_0(x,y) - \mathcal{D}_{\mathcal{W}}(R)(W(s))(x,y) + \frac{2}{\sqrt{\beta}}B(t), \\ W(s)(u,v) &= \mathbb{E}[X(t) \mid U = u, V = v]\,. \end{aligned} \tag{1.12}$$

The well-posedness of such a system is part of our work. We refer to this coupled system of SDE and kernel as graphon McKean-Vlasov SDE. Morally, the evolution of $X_{U,V}$ describes the evolution of a uniformly random coordinate of A_n in the limit as $n \to \infty$. We once again emphasize that in order to ensure that the coordinates of $A_n(t)$ are in $[0,1]$ at all times, we need to modify (1.11). This is done via coordinatewise Skorokhod reflection of (1.11). This reflection makes the analysis more delicate and the process with and without reflections are qualitatively different. We detail this point further in Chapter 5.

We now move to Chapter 6. Chapter 6 is based on the paper *Path convergence of Markov chains on large graphs* with Siva Athreya, Soumik Pal, and Raghav Somani. While the cut metric is very useful and naturally suited for studying the convergence of simple graphs it is not well-suited for the convergence of weighted graphs. This weakness of cut-convergence also manifests in the fact that a sequence of kernels W_n converging to W in metric does not necessarily imply the convergence of $\int_{[0,1]^2} W_n^2(x,y)\,\mathrm{d}x\,\mathrm{d}y$ to $\int W^2(x,y)\,\mathrm{d}x\,\mathrm{d}y$. If one is

interested in studying the dynamics of matrices under permutation-invariant interactions, it is natural to search for a stronger mode of convergence. This is one issue that we address in Chapter 6. More importantly, we construct a variant of the Metropolis chain on a suitable space of graphs that approximates the gradient flow of R.

To understand this, suppose we are interested in minimizing R_n on the space of graphs for very large n and suppose that we can solve or approximate the gradient flow curve $t \mapsto W_t$ of R on the space of graphons. To obtain an approximate minimizer of R_n, one can sample $\mathbb{G}(n, W(t))$ for large t. Under appropriate assumptions, it follows that $\mathbb{G}(n, W(t))$ is an approximate minimizer of R_n. This naturally suggests the following question.

Question 1.5.2. *Let R be a nice function on the space of graphons. Can we construct a Markov chain on $\mathcal{G}_n$, the space of graphs on n vertices, that mimics the gradient flow of R when n is large?*

This is another question that we address in Chapter 6. In this chapter, We construct a Metropolis-Hastings type chain on the stochastic block models. In other words, our state space is a graph on nr vertices where the vertices are divided into n communities and each community has r individuals. We show that as $r, n \to \infty$ suitably, the paths of the Markov chain are close (in cut-metric), with high probability, to the deterministic curve of kernels described by the graphon McKean-Vlasov SDE described in (1.12).

In fact, in Chapter 6, we consider a general class of dynamics on the symmetric matrices and establish its convergence to an IEA. However, this time, our convergence is in a much stronger sense. In particular, the limiting objects are identified with kernels taking values in the space of measures (and correspondingly measure-valued graphons). The measure-valued graphons as tools to study the convergence of weighted graphs were introduced by [151]. We introduce a metric on the space of measure-valued graphons analogous to the cut-metric on graphons and further extend many results connecting IEA and graphons to measure-valued graphon settings.

In Chapter 7, we consider a slightly different but related theme. We study the scaling limit of the iterated product of matrices. Such problems arise in many different contexts that we discuss in detail in Chapter 7. Philosophically, we can say that this chapter deals

with an evolution of matrices where the updates are multiplicative and not additive. We will explain this point shortly.

To state the problem, let us consider a triangular array of $n \times n$ (possibly random) matrices $\left(\left(A_{n,k}^{(m)}\right)_{k\in[m]}\right)_{m\in\mathbb{N}}$ and define the following iterated product of matrices

$$P_n^{(m)}(k) \coloneqq \left(I_n + \frac{\mu_n}{m}A_{n,k}^{(m)}\right)\ldots\left(I_n + \frac{\mu_n}{m}A_{n,2}^{(m)}\right)\left(I_n + \frac{\mu_n}{m}A_{n,1}^{(m)}\right), \qquad k \in [m], \tag{1.13}$$

where μ_n is a dimension-dependent scaling factor. We set $P_n^{(m)}(0) = I_n$. Our goal is to establish the scaling limit for $P_n^{(m)}$ in equation (1.13) as $m, n \to \infty$. In the following, we first explain the scaling limit as $m \to \infty$ and then consider the limit as $n \to \infty$. The role of μ_n becomes important only when we consider the limit as $n \to \infty$. Therefore, in the following discussion, we fix $\mu_n = 1$. Such iterated products of matrices arise naturally in many different contexts including random walks on groups, Oja's algorithm, and Neural networks. Chapter 7 is based on the paper *Scaling Limits of Large Linear Residual Networks* with Zaid Harchaoui, Sewoong Oh, Soumik Pal and Raghav Somani. In Chapter 7, we do not assume matrices $A_{n,k}^{(m)}$ to be symmetric. First, let us consider the scaling limit of $P_n^{(m)}$, fixing the dimension $n \in \mathbb{N}$, as $m \to \infty$. As n is fixed, we will drop μ_n for simplicity in the following discussion. Note that $P_n^{(m)}$ satisfies following difference equation

$$P_n^{(m)}(k+1) - P_n^{(m)}(k) = \frac{1}{m}A_{n,k+1}^{(m)}P_n^{(m)}(k), \qquad k \in [m-1]. \tag{1.14}$$

Note that in (1.14), we see that $P_n^{(m)}(k+1)$ is obtained by multiplying pre-multiplying $P_n^{(m)}(k)$ with $I + \frac{1}{m}A_{n,k+1}^{(m)}$. This should be compared with gradient flow equation (1.8). If we consider a time discretization of (1.8), we see that given a state $A_n(t)$, the next step $A_n(t+\tau)$ is obtained by adding $-\tau\nabla R(A_n(t))$ with $A_n(t)$. In this sense, we can say that the theme of Chapter 7 is to deal with the multiplicative dynamics of matrices.

We now return to the discussion of the scaling limit of $P_n^{(m)}$. In the view of (1.14), it is reasonable to expect that $P_n^{(m)}$ admits a scaling limit. 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 $\frac{\mathrm{d}}{\mathrm{d}t}P_n(t) = A_n(t)P_n(t)$ as $m \to \infty$. If $A_n(t) \equiv A_n$ is a constant curve, then the solution to this differential equation is $P_n(t) = \mathrm{e}^{tA_n}$. For a

more general curve A_n, one may guess the solution of the above differential equation to be $P_n(t) = \mathrm{e}^{\int_0^t A_n(s)\,\mathrm{d}s}$. However, this is incorrect – unless $A_n(s)$ and $A_n(s')$ commute for all $s, s' \in [0,t]$.

Under very general conditions on the triangular array $A^{(m)}_{n,k}$, we show that $P^{(m)}_n$ admits a scaling limit, as $m \to \infty$ while n is fixed. This limit is described by a non-commutative exponential $\mathrm{Texp}[Y_n]$ of a semimartingale Y_n where Y_n is the scaling limit of $A^{(m)}_{n,\lfloor mt \rfloor}$ as $m \to \infty$. Next, we explore the suitable scaling limit of $\mathrm{Texp}[Y_n]$, as $n \to \infty$, where Y_n is a semimartingale of the form

$$Y_n(t) = \mu_n \int_0^t A_n(s)\,\mathrm{d}s + \sigma_n B_n(t)\ ,$$

where B_n is an $n \times n$ matrix whose coordinates are i.i.d. Brownian motion. Note that we do not assume any symmetry here. In Chapter 7 we study the two important cases $\mu_n = \sigma_n = n^{-1}$ and $\mu_n = \sigma_n^2 = n^{-1}$. Somewhat surprisingly, we observe a propagation of chaos phenomenon in $\mathrm{Texp}[Y_n]$ as well. To study this limit, we often assume that $t \mapsto K(A_n(t))$ converges to a curve of L^2 kernels $t \mapsto W(t)$. As a consequence, we show that, under the $\mu_n = \sigma_n = n^{-1}$ regime, $n(\mathrm{Texp}[Y_n] - I_n)$ converges to an IEA whose coordinates are conditionally independent Gaussian whose mean and variance can be completely characterized in terms of the limiting curve $t \mapsto W(t)$. As a consequence, we also obtain the convergence of the curve of kernels $t \mapsto K(\mathrm{Texp}[Y_n](t))$ to a deterministic curve of kernels that is described by a suitable non-commutative exponential of the curve $t \mapsto W(t)$ that we define in Chapter 7.

In $\mu_n = \sigma_n^2 = n^{-1}$ regime the situation is more delicate. In this case the variance of a coordinate of $n(\mathrm{Texp}[Y_n] - I_n)$ blows up as $n \to \infty$. On the other hand, $\sqrt{n}(\mathrm{Texp}[Y_n] - I_n)$ converges to an IEA where the coordinates become i.i.d. time changed Brownian motions. Note that this limit is independent of the curve $t \mapsto W(t)$. In other words, this limit is trivial– in the sense that it does not depend on the data $t \mapsto A_n(t)$. We therefore consider a further centering and scaling of $\sqrt{n}(\mathrm{Texp}[Y_n] - I_n)$, namely, we consider $\sqrt{n}(\sqrt{n}(\mathrm{Texp}[Y_n] - I_n) - G_n)$ where G_n is a square matrix whose coordinates are (almost) i.i.d. time changed of Brownian motion. We show that $\sqrt{n}(\sqrt{n}(\mathrm{Texp}[Y_n] - I_n) - G_n)$ converges to an IEA whose coordinates are Gaussian processes but coordinates of this IEA

have a non-trivial correlation. We describe the mean and covariance of this limiting IEA in terms of the curve $t \mapsto W(t)$. This second case has important applications in the study of residual neural networks. The mean of the limiting IEAs in the above two cases agree and we interpret the mean in terms of quantum homomorphism density of some quantum graphs that we describe in 7.

.

Chapter 2

PRELIMINARIES

There are three central objects in this thesis: graphons, measure-valued graphons, and infinite exchangeable arrays. In this chapter, we provide a gentle introduction to these three topics and their interrelationship. We also state some preliminary results that are used throughout the thesis.

We also provide a background on the theory of gradient flows in metric spaces in Section 2.1.4 while keeping in mind the metric space of graphons.

In Section 2.2 and Section 2.3.1, we state and prove some novel results from [167] and [9]. These results are also used in the following chapters, but we believe these will be useful to the general audience as well.

2.1 Background on graphons

In Section 2.1.1, we introduce the required metric on graphons and other properties related to graphons. The material in this section is mostly borrowed from [150, 118]. In Section 2.1.4, we introduce the necessary terminology to talk about the gradient flow on a metric space. The material in that section is adapted from [5].

2.1.1 Graphons and Metrics on Graphons

Recall that a kernel $W\colon [0,1]^2 \to [-1,1]$ is a Borel measurable, symmetric function. On the space of kernels, $\mathcal{W} \subset L^2\big([0,1]^2\big)$, we have the usual L^2 norm, $\|\cdot\|_2$, that is, $\|W\|_2^2 := \int_{[0,1]^2} |W(x,y)|^2 \,\mathrm{d}x\,\mathrm{d}y$. We also define the cut norm, denoted $\|\cdot\|_\square$, on $\mathcal{W}$ as follows.

Definition 2.1.1 (Cut norm)**.** *The* cut norm $\|\cdot\|_\square \colon \mathcal{W} \to \mathbb{R}_+$ *is defined as*

$$\begin{aligned}\|W\|_\square &\coloneqq \sup_{S,T\subseteq[0,1]} \left| \int_{S\times T} W(x,y)\,\mathrm{d}x\,\mathrm{d}y \right| \\ &= \sup_{\|f\|_\infty,\|g\|_\infty\leq 1} \left| \int_{[0,1]^2} W(x,y)f(x)g(y)\,\mathrm{d}x\,\mathrm{d}y \right|,\end{aligned}$$

for all $W \in \mathcal{W}$ where S and T are Borel measurable subsets of $[0,1]$, and f and g are Borel measurable functions on $[0,1]$, and $\|\cdot\|_\infty$ denotes the usual L^∞ norm.

The cut norm was first introduced in [87] in the context of matrices and was later extended to kernels in [39]. The cut norm is used to define a metric called the cut metric, $\delta_\square$, on the space of graphons that we define now.

In the following, we use $\mathcal{T}$ to denote the set of all Lebesgue measure-preserving maps $\varphi \colon [0,1] \to [0,1]$, and $\mathcal{I}$ to denote the set of all invertible Lebesgue measure-preserving maps $\varphi \colon [0,1] \to [0,1]$. Given a kernel $W \in \mathcal{W}$ and a Lebesgue measure preserving map $\varphi \in \mathcal{T}$, one can define $W^\varphi \in \mathcal{W}$ as $W^\varphi(x,y) \coloneqq W(\varphi(x),\varphi(y))$ for Lebesgue a.e. $x, y \in [0,1]$. From hereon in the text, we will always refer to Lebesgue measure preserving transformations, and Lebesgue almost everywhere (a.e.) unless explicitly specified otherwise. The kernels $W_1, W_2 \in \mathcal{W}$ are said to be *weakly isomorphic* if there exists a kernel $U \in \mathcal{W}$ and Lebesgue measure preserving transforms $\varphi_1, \varphi_2 \colon [0,1] \to [0,1]$ such that $W_i = U^{\varphi_i}$, for $i \in [2]$.

Definition 2.1.2 (Graphons and the space of graphons)**.** *The space of graphons $\widehat{\mathcal{W}}$ is defined to be the set of equivalence classes in $\mathcal{W}/\cong$ where $W_1 \cong W_2$ if W_1 and W_2 are weakly isomorphic. Naturally, the elements in $\widehat{\mathcal{W}}$ are called graphons. We will often make an abuse of notation and use W to refer to a kernel as well as the graphons corresponding to the equivalence class of W. When we wish to be precise about the distinction between a kernel and a graphon, we will denote a graphon by $[W]$.*

The cut norm defined above induces a metric on the space of graphons called the cut metric. We now define this metric but before that, we need some notations.

For any $n \in \mathbb{N}$, let $Q_n \coloneqq \{Q_{n,i}\}_{i\in[n]}$ be a partition of $[0,1]$ into intervals of equal Lebesgue measure, for example $Q_{n,1} \coloneqq [0,1/n]$ and $Q_{n,i} \coloneqq ((i-1)/n, i/n]$ for $i \in [n] \setminus \{1\}$. 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 $\mathcal{I}_n$.

Definition 2.1.3 (Cut metric [39, Section 3.2])**.** *The* cut metric $\delta_\square\colon \widehat{\mathcal{W}} \times \widehat{\mathcal{W}} \to \mathbb{R}_+$ *is defined as*

$$\begin{aligned}\delta_\square([W_0],[W_1]) &\coloneqq \inf_{\varphi_0,\varphi_1\in\mathcal{T}} \|W_0^{\varphi_0} - W_1^{\varphi_1}\|_\square \\ &= \inf_{\varphi\in\mathcal{I}} \|W_0 - W_1^{\varphi}\|_\square = \lim_{k\to\infty} \min_{\tilde{\pi}\in\mathcal{I}_k} \big\|W_0 - W_1^{\tilde{\pi}}\big\|_\square,\end{aligned} \tag{2.1}$$

for all $[W_0],[W_1] \in \widehat{\mathcal{W}}$*, where the latter two equalities are due to [39, Lemma 3.5].*

Note that $\delta_\square$ can be naturally extended to kernels, but it only defines a pseudometric on $\mathcal{W}$. In fact, $\delta_\square(W_1,W_2) = 0$ if and only if the kernels W_1, W_2 are weakly isomorphic. In other words, graphons can also be defined as the class of kernels identified up to zero distance in the cut metric.

More generally, one can start with a norm on the space of kernel that is invariant, that is, $\|W\| = \|W^\varphi\|$. Any such norm can used to induce a metric on the space of $\widehat{\mathcal{W}}$. In this spirit, we also define the so-called invariant L^2 metric on $\widehat{\mathcal{W}}$ that will be needed later.

Definition 2.1.4 (Invariant L^2 metric [37, 118])**.** *The* invariant L^2 metric $\delta_2\colon \widehat{\mathcal{W}}\times\widehat{\mathcal{W}} \to \mathbb{R}_+$ *is defined as*

$$\delta_2([W_0],[W_1]) \coloneqq \inf_{\varphi_1,\varphi_2\in\mathcal{T}} \|W_0^{\varphi_1} - W_1^{\varphi_2}\|_2 = \inf_{\varphi\in\mathcal{I}} \|W_0 - W_1^{\varphi}\|_2, \tag{2.2}$$

for all $[W_0],[W_1] \in \widehat{\mathcal{W}}$*, where* $\|\cdot\|_2\colon L^2\big([0,1]^2\big) \to \mathbb{R}_+$ *is the usual* L^2*-norm, and the second equality is a consequence of [150, Theorem 8.13].*

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 [152], [150, 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 2.1.5. *[150, 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*

$$\liminf_{n\to\infty} \delta_2([U_n],[V_n]) \geq \delta_2([U],[V]).$$

As we mentioned in the Introduction, the gradient flow on the space of graphons will be with respect to the invariant L^2 metric, but the convergence statements will be with respect to the topology generated by the cut metric.

Let us mention in passing that the invariant L^2 metric is closely related to the popular Gromov-Wasserstein metric [158] used to compare two metric measure spaces or their sample equivalents [72]. 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. In this vein also see the unpublished article [202] which constructs gradient flows on the space of metric measure spaces in a spirit that is quite similar to ours.

2.1.2 Block Graphons, Matrices and Graphons

As we briefly explained in the Introduction, symmetric matrices or equivalently edge-weighted graphs can be thought of as sitting inside the space of kernels and hence inside the space of graphons. As we frequently make this identification, we recall this notion again and explain it in more detail below.

For any $n \in \mathbb{N}$, define the set of kernels $\mathcal{W}_n \subset \mathcal{W}$ which contain kernels that are constant a.e. over sets in $Q_n \times Q_n$. The set $\mathcal{W}_n$ can be naturally identified with a convex subset of the finite-dimensional vector space of symmetric $n \times n$ matrices. Since this identification will be used often, we make it a definition.

Definition 2.1.6 (Kernels and finite symmetric matrices)**.** *For any $n \in \mathbb{N}$, and a symmetric matrix $A \in \mathcal{M}_n$, the kernel $K(A)$ corresponds to the element in $\mathcal{W}_n$ which takes the constant value $A_{i,j}$ on $Q_{n,i} \times Q_{n,j}$ for $i, j \in [n]$. The inverse map from $\mathcal{W}_n$ to $\mathcal{M}_n$ is denoted by M_n.*

Kernels in $\mathcal{W}_n$ can be thought of as adjacency matrices of vertex-labeled graphs with edge weights. For a finite graph $G = (V, E)$ with vertices labeled as $[n]$ and associated weights with every edge with weight $w(\{i, j\}) \in [-1, 1]$ for $\{i, j\} \in E$, we can construct its adjacency matrix $A \in \mathcal{M}_n$ by defining

$$A_{i,j} \coloneqq w(\{i, j\})\mathbb{1}\{\{i, j\} \in E\}, \quad i, j \in [n].$$

Thus, we can also map vertex-labeled graphs with weights associated with its edges to kernels by considering $K(A) \in \mathcal{W}_n$. Naturally, we can also define $\widehat{\mathcal{W}}_n \coloneqq \mathcal{W}_n/\cong$. In other words, we now identify a symmetric matrix or edge-weighted graph with a kernel.

On the other hand, we explained in the Introduction, that given a kernel $W : [0, 1]^2 \to [0, 1]$ we can construct a sequence of (simple) random graphons $\mathbb{G}(n, W)$ on the vertex set $[n]$ by making an edge between i and j with probability $W(U_i, U_j)$ (independently given $\{U_i : i \in \mathbb{N}\}$). The adjacency matrix of such a graph is exchangeable. Furthermore, as we hinted in the Introduction, the random graphs $\mathbb{G}(n, W_1)$ and $\mathbb{G}(n, W_1)$ have the same law for each $n \geq 2$ if W_1 and W_2 are weakly isomorphic. Therefore, one can naturally define $\mathbb{G}(n, [W])$ for a graphon $[W]$.

However, the above construction does not work for the kernels $W \in \mathcal{W}$ that can take values in $[-1, 1]$. To address this issue we now explain how to sample a sequence of exchangeable matrices (or equivalently edge-weighted graphs) given a graphon $[W] \in \widehat{\mathcal{W}}$.

Starting with a graphon $[W] \in \widehat{\mathcal{W}}$, for any $n \in \mathbb{N}$ we can sample a random graph $G_n[W]$ of size n as follows. Consider any representative element $W \in [W]$ and sample n i.i.d. elements $\{U_i\}_{i=1}^n$ uniformly at random from $[0, 1]$, and assign edge weight $W(U_i, U_j)$ to edge $\{i, j\}$ for all $(i, j) \in [n]^{(2)}$. By an abuse of notation, we also denote the exchangeable symmetric $n \times n$ weighted adjacency matrix of this random graph by $G_n[W]$. The distinction will be apparent from the context. In either case, $G_n[W]$ is measurable with respect to $\sigma(\{U_i\}_{i=1}^n)$.

2.1.3 Homomorphism densities and cut-metric

Recall the function R on $\mathcal{W}$ defined as

$$R(W) = \int_{[0,1]^2} W(x_1, x_2)W(x_2, x_3)W(x_3, x_1)\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3 - \alpha \int W(x, y)\,\mathrm{d}x\,\mathrm{d}y\,,$$

where $\alpha > 0$. We defined the function R in the Introduction. It is straightforward to verify $R(W) = R(W^\varphi)$ for any Lebesgue measure-preserving transform $\varphi : [0,1] \to [0,1]$. In particular, R projects to a well-defined function on the space of graphons $\widehat{\mathcal{W}}$. This is a particular example of a rich and interesting class of functions on the space of graphons, namely, *homomorphism densities*. These functions have been central in the development of the theory of graphons.

Let F be a finite simple graph on the vertex set $[m]$ for some $m \in \mathbb{N}$. Let $E(F)$ denote the set of edges in F. Let $W \in \mathcal{W}$ be a kernel. We define the homomorphism density of F into W, denoted by $t(F, W)$, as

$$t(F, W) = \int_{[0,1]^m} \prod_{\{i,j\} \in E(F)} W(x_i, x_j) \prod_{k=1}^{m} \mathrm{d}x_k \ .$$

It is easily verified that $t(F, W) = t(F, W^\varphi)$ for any Lebesgue measure preserving transform $\varphi : [0,1] \to [0,1]$. Therefore, $t(F, \cdot)$ defines a function on $\widehat{\mathcal{W}}$.

Let G be a graph on the vertex set $[n]$. Let A be the adjacency matrix of G and let W_A be the kernel corresponding to the adjacency matrix of G as explained in the previous section. Let F be as above and assume that $m \leq n$. Notice that

$$t(F, G) \coloneqq t(F, W_A) = \frac{1}{n^m} \sum_{1 \leq p_1, \ldots, p_m \leq n} \prod_{\{i,j\} \in E(F)} A(p_i, p_j) \ .$$

Fix an m-tuple $(p_1, \ldots, p_m) \in [n]^m$. Notice that the product $\prod_{\{i,j\} \in E(F)} A(p_i, p_j) = 1$ if and only if there is an $\{p_i, p_j\} \in E(G)$ whenever $\{i, j\} \in E(F)$. In other words, $\prod_{\{i,j\} \in E(F)} A(p_i, p_j)$ is exactly the indicator function of the condition that the subgraph of G induced by the vertices $\{p_1, \ldots, p_m\}$ is homomorphic to F. Thus, we can interpret $t(F, W_A)$ as the probability that a random map from $[m]$ into $[n]$ is a homomorphism of F into G. This justifies the name homomorphism density for $t(F, \cdot)$. While working with graphs, it is sometimes useful to consider a modification of the homomorphism density function, denoted by $t_{\mathrm{inj}}(F, \cdot)$, which gives the probability that a random *inejctive* map from $[m] \to [n]$ is a homomorphism of F into G. In other words,

$$t_{\mathrm{inj}}(F, G) = \frac{1}{\binom{n}{m}} \sum_{\{p_1, \ldots, p_m\} \in \binom{[n]}{m}} \prod_{\{i,j\} \in E(F)} A(p_i, p_j) \ ,$$

where $\binom{[n]}{m}$ denotes all subsets of $[n]$ with cardinality m. For a fixed F, the two types of homomorphism densities are asymptotically the same. That is, $|t(F,A) - t(F,A)| \leq \frac{1}{n}\binom{m}{2}$ for any $A \in \mathcal{M}_n$. Therefore, in the limiting regime as $n \to \infty$, it does not matter which definition one uses and we will use whichever will be convenient.

Homomorphism density functions not only provide a rich class of examples of invariant functions, but they are also foundational to the theory of graphons. We refer the reader to [150, Chapter 5] for more details. We end this section with two important properties of the homomorphism density functions. Firstly, it is easily seen from the definition that if W_1 and W_2 and weakly isomorphic then $t(F,W_1) = t(F,W_2)$ for any finite simple graph F. It turns out that the converse is also true [150, Section 10.7, 13.2]. In other words, W_1, W_2 are weakly isomorphic if and only if $t(F,W_1) = t(F,W_2)$ for all finite simple graphs F. Historically, the graphons were first defined as the kernels identified up to the equivalence and it was later shown to be equivalent to being weakly isomorphic. Secondly, let us say that a sequence of graphons W_n (or more precisely $[W_n]$) converges to some graphon W in the *homomorphism density sense* if $\lim_{n\to\infty} t(F,W_n) = t(F,W)$ for every finite simple graph. One can use this notion of convergence to define a topology on the space of graphons $\widehat{\mathcal{W}}$. One of the most important insights coming from the graphon theory is that the topology generated by the convergence in the homomorphism density on $\widehat{\mathcal{W}}$ is the same as the topology generated by the cut metric. More succinctly $\lim_{n\to\infty} \delta_\square(W_n, W) = 0$ if and only if $\lim_{n\to\infty} t(F,W_n) = t(F,W)$ for every finite simple graph F (see [150, Theorem 11.5]). This equivalence is often useful in practice.

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 functions on $[0,1]^2$ that are constant 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 $\widehat{L}^p\big([0,1]^2\big) := L^p\big([0,1]^2\big)/\!\cong$ and $\widehat{L}^p_n\big([0,1]^2\big) := L^p_n\big([0,1]^2\big)/\!\cong$ 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 $n \in \mathbb{N}$ from Definition 2.1.6 even when the kernels are in $L^p\big([0,1]^2\big)$ for $p \in [1, \infty]$.

Unless otherwise explicitly stated (as we do in Chapter 7), the kernels and matrices that we work with are symmetric. To emphasize that $W : [0,1]^2 \to [-1,1]$ is a symmetric function, we often use the notation $W : [0,1]^{(2)} \to [-1,1]$. In other words, we use the notation $[0,1]^{(2)}$ to denote the set $\{(x,y) \in [0,1]^2 : x \leq y\}$. Note that symmetric measurable functions on $[0,1]^2$ are in one-to-one correspondence with functions on $[0,1]^{(2)}$. This is primarily for a notational convenience.

2.1.4 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 [5]. 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 [5] but adapted to our particular setting.

Definition 2.1.7 (Absolutely continuous curves)**.** *For a metric space (X,d), and any time horizon $T \in \mathbb{R}_+$, a curve $\omega = (\omega_t)_{t\in[0,T]}$ in X is* absolutely continuous *with respect to the metric d if there exists $m \in L^1([0,T])$ such that for all $0 \leq r < s \leq T$*

$$d(\omega_r, \omega_s) \leq \int_r^s m(t)\,\mathrm{d}t. \tag{2.3}$$

The set of all absolutely continuous curves will be denoted as $\mathrm{AC}(X,d)$.

Definition 2.1.8 (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*

$$\big|\omega'\big|(t) \coloneqq \lim_{s\to t} \frac{d(\omega_s, \omega_t)}{|s-t|}, \tag{2.4}$$

provided this limit exists.

If $\omega \in \mathrm{AC}(X,d)$, then the limit in equation (2.4) exists for a.e. $t \in (0,T)$ and $|\omega'| \in L^1([0,T])$ [5, Theorem 1.1.2]. In other words, every absolutely continuous curve

in a metric space has a 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 $F\colon X \to \mathbb{R}\cup\{\infty\}$. In a metric space, the usual notion of derivative can not be defined. However, the following [5, Definition 1.2.4] acts as a substitute in many situations of interest.

Definition 2.1.9 (Local slope)**.** *The* local slope $|\partial F|(v)$ *of* $F\colon X \to \mathbb{R}\cup\{+\infty\}$ *on a metric space* (X,d)*, at* $v \in \text{eff-Dom}(F)$ *is defined as*

$$|\partial F|(v) \coloneqq \limsup_{w\in X,\, d(v,w)\to 0} \frac{(F(v)-F(w))^+}{d(v,w)}. \tag{2.5}$$

The definition below is narrower than the one in [5, Definition 1.3.2] since we restrict our choice of *upper gradient* in that definition to the local slope [5, Theorem 1.2.5].

Definition 2.1.10 (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* $F\colon X \to \mathbb{R}\cup\{+\infty\}$ *with respect to its local slope, if* $F\circ\omega = G$ *a.e. for some non-increasing map* G *on* $(0,T)$*, and*

$$G'(t) \le -\frac{1}{2}|\omega'|^2(t) - \frac{1}{2}|\partial F|^2(\omega_t), \qquad \text{a.e.} \quad t \in (0,T). \tag{2.6}$$

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 [5, Section 1.3] for the intuition.

Definition 2.1.11 (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 2.1.11 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 $\ell(\omega) \ge d(x,y)$. 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 2.1.12 (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 2.1.13 (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 $0 \le r \le s \le 1$,*

$$d(\omega_r, \omega_s) = d(\omega_0, \omega_1)(s - r). \tag{2.7}$$

Note that if a curve ω satisfies equation (2.7), then ω is 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 2.1.14. *It is also worth pointing out that not only every geodesic but every absolutely continuous curve can be reparametrized so that it becomes Lipschitz [192, 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 2.1.15 (λ-semiconvexity along curves w.r.t. a metric)**.** *On a metric space (X, d), a function $F\colon X \to \mathbb{R} \cup \{\infty\}$ 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*

$$F(\omega_t) \le (1 - t)F(\omega_0) + tF(\omega_1) - \frac{1}{2}\lambda t(1 - t)d^2(\omega_0, \omega_1), \tag{2.8}$$

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

Definition 2.1.16 (λ-geodesic semiconvexity w.r.t. a metric)**.** *On a metric space (X, d), a function $F\colon X \to \mathbb{R} \cup \{\infty\}$ is* λ-geodesically semiconvex *with respect to the metric d, if for any $v_0, v_1 \in \text{eff-Dom}(F)$ there exists a constant speed geodesic $\omega = (\omega_t)_{t\in[0,T]}$ on (X, d) (Definition 2.1.13) with $\omega_0 = v_0$ and $\omega_1 = v_1$ such that F is λ-semiconvex on ω with respect to the metric d for some $\lambda \in \mathbb{R}$ (Definition 2.1.15).*

2.2 Some Preliminary results on the space of graphons

In this section, we prove some results that are used in the construction of gradient flows on the space of graphons, but are also of independent interest. The two key results in this section are Lemma 2.2.1 and Proposition 2.2.5. 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 2.2.1 shows that the converse is also true. Proposition 2.2.5 states that $(\widehat{\mathcal{W}}, \delta_2)$ is a geodesic metric space. All the results in this Section are taken from [167].

Lemma 2.2.1. *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 $s, t \in [0, 1]$.*

The proof of Lemma 2.2.1 requires a strengthening of [150, Theorem 8.13], [117, Theorem 6.16] that we state and prove below. Before we begin the proof, we define some notations. 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, $W : \Omega \times \Omega \to \mathbb{R}$ we define the norm $\|\cdot\|_{2,\Omega,\mu}$ as

$$\|W\|^2_{2,\Omega,\mu} \coloneqq \int_{\Omega^2} |W(x, y)|^2 \mu(\mathrm{d}x)\mu(\mathrm{d}y).$$

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

$$W^\varphi(\omega_1, \omega_2) \coloneqq W(\varphi(\omega_1), \varphi(\omega_2)), \quad \text{for } \mu\text{-a.e.} \quad \omega_1, \omega_2 \in \Omega. \tag{2.9}$$

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 (2.9), 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.

Lemma 2.2.2. *Let $\omega_1, \ldots, \omega_n \in \widehat{\mathcal{W}}$. Then there exist $W_1, \ldots, W_n \in \mathcal{W}$ 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 $U_i \in \omega_i$ for $i \in [n]$. From [150, 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

$$\delta_2(\omega_i, \omega_{i+1}) = \left\|U_i^{\pi} - U_{i+1}^{\rho}\right\|_{2,[0,1]^2,\mu_i}. \tag{2.10}$$

Let $\pi_i \colon [0,1]^n \to [0,1]$ be the usual projection map on the i-th coordinate. By the gluing lemma [212, 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

$$\left\|U_i^{\pi} - U_{i+1}^{\rho}\right\|_{2,[0,1]^2,\mu_i} = \left\|U_i^{\pi_i} - U_{i+1}^{\pi_{i+1}}\right\|_{2,[0,1]^n,\tilde{\mu}}. \tag{2.11}$$

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

$$\left\|U_i^{\pi_i} - U_{i+1}^{\pi_{i+1}}\right\|_{2,[0,1]^n,\tilde{\mu}} = \left\|U_i^{\varphi_i} - U_{i+1}^{\varphi_{i+1}}\right\|_{2,[0,1]}. \tag{2.12}$$

Combining equations (2.10), (2.11) and (2.12), 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. □

Proof of Lemma 2.2.1. Following Remark 2.1.14, assume (possibly after a reparametrization) that the curve ω is Lipschitz with Lipschitz constant $L \geq 0$. Let $n \in \mathbb{N}$. From Lemma 2.2.2, 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 $W^{(n)} = \big(W_t^{(n)}\big)_{t\in[0,1]}$ as

$$W_t^{(n)} := (1 - nt + i)W_{i,n} + (nt - i)W_{i+1,n},$$

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 [192, 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 [160, 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 $s, t \in [0,1]$,

$$\Big\langle W_t^{(n)} - W_s^{(n)}, W_t - W_s \Big\rangle \to \|W_t - W_s\|_2^2.$$

Using Cauchy–Schwarz inequality, we obtain

$$\|W_t - W_s\|_2 \leq \liminf_{n\to\infty} \Big\| W_t^{(n)} - W_s^{(n)} \Big\|_2 \leq L|t - s|.$$

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 [150, 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 [150, 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 2.2.3. *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 2.2.1.*

Proof. 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) [111, 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 [112, 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 2.2.1 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. $t \in (0,1)$. □

Lemma 2.2.4. *The invariant L^2 metric between two graphons $[U], [V] \in \widehat{\mathcal{W}}$ satisfies*

$$\delta_2([U],[V]) = \min \int_r^s \left\|W_t'\right\|_2 \mathrm{d}t \ , \tag{2.13}$$

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 $W_s \in [V]$.

Proof. 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

$$\int_r^s \left\|W_t'\right\|_2 \mathrm{d}t \ge \left\| \int_r^s W_t' \,\mathrm{d}t \right\|_2 = \|W_s - W_r\|_2 \ge \delta_2([U],[V]) \ . \tag{2.14}$$

Following Definition 2.1.4, there exists $\varphi_1, \varphi_2 \in \mathcal{T}$ such that

$$\delta_2([U],[V]) = \|U^{\varphi_1} - V^{\varphi_2}\|_2. \tag{2.15}$$

Therefore, we can define an curve $(W_t)_{t\in[r,s]} \in \mathrm{AC}(\mathcal{W}, d_2)$ as $W_r := U^{\varphi_1}$, $W_s := V^{\varphi_2}$ and $W_t := ((s-t)W_r + (t-r)W_s)/(s-r)$ for $t \in (r,s)$. Since for any $r \le a < b \le s$,

$$\|W_b - W_a\|_2 = \frac{\|W_s - W_r\|_2}{s-r} \cdot (b-a) = \frac{\|U^{\varphi_1} - V^{\varphi_2}\|_2}{s-r} \cdot (b-a), \tag{2.16}$$

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 (2.15) we get

$$\int_r^s \left\|W_t'\right\|_2 \mathrm{d}t = \|U^{\varphi_1} - V^{\varphi_2}\|_2 = \delta_2([U],[V]). \tag{2.17}$$

Combining equation (2.14) and equation (2.17) completes the proof. □

As a consequence of the above discussion, 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 2.2.5. *The space $(\widehat{\mathcal{W}}, \delta_2)$ is a geodesic metric space.*

Proof. Recall that (see the remark after the Definition 2.1.11) 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

$$\ell(\omega) \geq \delta_2(\omega_0, \omega_1) = \delta_2([U], [V]). \tag{2.18}$$

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

$$\delta_2([W_s], [W_r]) \leq \|W_s - W_r\|_2 = \|U - V\|_2 \cdot (s - r), \tag{2.19}$$

for all $0 \leq r < s \leq 1$. Now observe that

$$\begin{aligned}
\ell(\omega^*) &= \sup\left\{ \sum_{k=0}^{n-1} \delta_2\big([W_{t_k}], \big[W_{t_{k+1}}\big]\big) \,\middle|\, n \in \mathbb{N}, 0 = t_0 < t_1 \cdots < t_n = 1 \right\} \\
&\leq \sup\left\{ \sum_{k=0}^{n-1} \|U - V\|_2 (t_{k+1} - t_k)\, \mathrm{d}t \,\middle|\, n \in \mathbb{N}, 0 = t_0 < t_1 \cdots < t_n = 1 \right\} \\
&= \|U - V\|_2 = \delta_2(\omega_0^*, \omega_1^*).
\end{aligned} \tag{2.20}$$

This completes the proof. □

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

Definition 2.2.6 (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 2.2.2, we obtain $\varphi, \varphi_0, \varphi_1 \in \mathcal{T}$ such that*

$$\delta_2([W], [W_0]) = \|W^{\varphi} - W_0^{\varphi_0}\|_2, \quad \text{and} \quad \delta_2([W], [W_1]) = \|W^{\varphi} - W_1^{\varphi_1}\|_2. \tag{2.21}$$

Define the curve $\vartheta \coloneqq ([W_t])_{t\in[0,1]}$, where $W_t \coloneqq (1-t)W_0^{\varphi_0} + tW_1^{\varphi_1}$ 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 to 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.*

Next, we show that there exists a generalized geodesic ϑ between two graphons such that the function $\delta_2^2([W], \cdot\,)$ is convex along ϑ for every $[W] \in \widehat{\mathcal{W}}$. This is used in the proof of Theorem 4.1.1.

Lemma 2.2.7. *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.*

Proof. From Lemma 2.2.2 we obtain $\varphi, \varphi_0, \varphi_1 \in \mathcal{T}$ such that

$$\delta_2([W],[W_0]) = \|W^{\varphi} - W_0^{\varphi_0}\|_2, \quad \text{and} \quad \delta_2([W],[W_1]) = \|W^{\varphi} - W_1^{\varphi_1}\|_2. \tag{2.22}$$

Defining $W_t \coloneqq (1-t)W_0^{\varphi_0} + tW_1^{\varphi_1}$ and $\vartheta_t \coloneqq [W_t]$ for $t \in [0,1]$, we get that that

$$\begin{aligned}
\delta_2^2([W],[W_t]) &\le \|W^{\varphi} - (1-t)W_0^{\varphi_0} - tW_1^{\varphi_1}\|_2^2 \\
&= (1-t)\|W^{\varphi} - W_0^{\varphi_0}\|_2^2 + t\|W^{\varphi} - W_1^{\varphi_1}\|_2^2 - t(1-t)\|W_0^{\varphi_0} - W_1^{\varphi_1}\|_2^2 \\
&= (1-t)\delta_2^2([W],[W_0]) + t\delta_2^2([W],[W_1]) - t(1-t)\|W_0^{\varphi_0} - W_1^{\varphi_1}\|_2^2 \\
&\le (1-t)\delta_2^2([W],[W_0]) + t\delta_2^2([W],[W_1]) - t(1-t)\delta_2^2([W_0],[W_1]).
\end{aligned}$$

Therefore, $\delta_2^2([W], \cdot\,)/2$ is 1-semiconvex along ϑ w.r.t. δ_2. □

2.3 Space of Measure-valued graphons

We now introduce another protagonist of our story, namely, the measure-valued graphons. As the name suggests these are kernels (identified up to some equivalence relation) that take values in the space of measures instead of $[-1,1]$.

Definition 2.3.1 (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 $\mathfrak{W}$.*

We will define the space of measure-valued graphons (MVGs), $\widehat{\mathfrak{W}}$ as a suitable quotient of $\mathfrak{W}$. This comes equipped with a notion of convergence introduced in [153, 140]. We call this convergence the usual convergence. In Section 2.4, we introduce two novel metrics on

$\widehat{\mathfrak{W}}$ analogous to the cut metric on the space of graphons. In Theorem 2.4.7 we show that these metrics induce the same topology on the measure-valued graphons and agree with the usual convergence.

In Section 2.5, we establish the correspondence between the distributions of IEAs and probability distributions on $\widehat{\mathfrak{W}}$ (see Theorem 2.5.3). We further use this connection to make precise the notion of convergence of symmetric exchangeable matrices to IEAs.

2.3.1 Definitions and notations

As mentioned already, the convergence of the homomorphism density functions $t(F, \cdot)$ 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 [153]). 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 2.3.2 (Decorated graph [153, 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 2.3.1) and F a decorated graph. Following [153, Section 2.5] one can define the (decorated) homomorphism density $t_{\mathrm{d}}(F, W)$ of F in W as

$$t_{\mathrm{d}}(F, W) \coloneqq \int_{[0,1]^m} \left(\prod_{\{j,k\} \in E(F)} \int_{[-1,1]} F_{j,k}(\zeta) W(x_j, x_k)(\mathrm{d}\zeta) \right) \prod_{i=1}^{m} \mathrm{d}x_i. \tag{2.23}$$

Definition 2.3.3 (Measure-valued graphon [153, 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*

$\widehat{\mathfrak{W}} := \mathfrak{W}/\sim$ *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 chapter.*

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 [139, 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 $\widehat{\mathfrak{W}} := \mathfrak{W}/\cong$. 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.

Definition 2.3.4 (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* $w(x,y) := \int_{[-1,1]} \zeta \, W(x,y)(\mathrm{d}\zeta)$ *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 2.4.5.*

2.3.2 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

$$t_{\mathrm{d}}(F, M) := \frac{1}{n^m} \sum_{i_1, \ldots, i_m} \prod_{\{j,k\} \in E(F)} F_{j,k}(M_{i_j, i_k}), \tag{2.24}$$

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 ealier. 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 for 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 2.3.5. *Since the map $W \mapsto w \coloneqq \mathbb{E}[W]$ is Lipschitz (see Definition 2.3.4). 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.*

2.3.3 Sampling matrices and graphs from MVGs

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 2.3.6 shows that almost surely $\mathbb{G}(n, W) \to W$ as $n \to \infty$ (see Subsection 2.3.2).

Lemma 2.3.6. *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 2.3.6 follows essentially the same idea as the proof of [140, Theorem 3.8]. An immediate consequence of Lemma 2.3.6 is that every MVG can be obtained as the limit of finite weighted graphs. MVGs were introduced as the limits of finite weighted in [153, Section 2.5].

Proof of Lemma 2.3.6. Let (F, f) be a decorated graph and let $\mathbb{G}(n, W)$ be as defined in Lemma 2.3.6. 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 [150, 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 the 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 [140, Theorem 3.4]. □

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 Uniform$[0,1]$ 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

$$\mu(n,W)(i,j) \coloneqq W(U_i, U_j), \qquad (i,j) \in [n]^{(2)}. \tag{2.25}$$

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 2.3.7. *Let $W \in \widehat{\mathfrak{W}}$. For $n \in \mathbb{N}$, let $\mu(n, W)$ be defined as in (2.25). Then $\mu(n, W) \to W$ in $\widehat{\mathfrak{W}}$ as $n \to \infty$, $\mathbb{P}$-almost surely.*

Lemma 2.3.7 follows directly from [140, 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 2.3.7.

2.4 Topology and metrics on measure-valued graphons

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 2.3.3.

Definition 2.4.1 (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*

$$t(F,W) \coloneqq \int_{[0,1]^{V(F)}} \bigotimes_{\{i,j\}\in E(F)} W(x_i, x_j) \prod_{v\in V(F)} \mathrm{d}x_v. \tag{2.26}$$

The measure in (2.26) 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

$$\int_{I_F} \varphi(\zeta) t(F,W)(\mathrm{d}\zeta) = \int_{[0,1]^{V(F)}} \left(\int_{I_F} \varphi(\zeta) \bigotimes_{\{i,j\}\in E(F)} W(x_i, x_j)(\mathrm{d}\zeta) \right) \prod_{k\in V(F)} \mathrm{d}x_k. \tag{2.27}$$

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

$$\int_{I_F} \varphi(\zeta) t(F,W)(\mathrm{d}\zeta) = \mathbb{E}\left[\left\langle \varphi, \bigotimes_{\{i,j\}\in E(F)} W(U_i,U_j)\right\rangle\right], \tag{2.28}$$

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 6.1.1), $t(F,w)$ 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 2.4.2 (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 2.3.2, the definition can be naturally extended to symmetric matrices and graphons. We skip the details to avoid repetitions.

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

$$\Gamma(\psi,W)(x,y) \coloneqq \int_{-1}^{1} \psi(\zeta) W(x,y)(\mathrm{d}\zeta). \tag{2.29}$$

Definition 2.4.3 (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 (2.29).

Remark 2.4.4. *Recall from Section 2.3.2 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. $(x,y) \in [0,1]^{(2)}$.

Graphons		**Measure-Valued Graphons**	
Cut norm on $\mathcal{W}$	$\|\cdot\|_{\square}$	$\|\cdot\|_{\blacksquare}$	Generalized Cut norm on $\mathfrak{W}$
L^2 metric on $\mathcal{W}$	d_2	D_2	$\mathbb{W}_2$ metric on $\mathfrak{W}$
Invariant L^2 metric on $\widehat{\mathcal{W}}$	δ_2	Δ_2	Invariant $\mathbb{W}_2$ metric on $\widehat{\mathfrak{W}}$
Cut metric on $\widehat{\mathcal{W}}$	$\delta_{\square}$	$\Delta_{\blacksquare}$ $\mathbb{W}_{\blacksquare}$	Generalized Cut metric on $\widehat{\mathfrak{W}}$ Wasserstein Cut metric on $\widehat{\mathfrak{W}}$
Curve on $\widehat{\mathcal{W}}$	$w\colon t \mapsto w(t)$	$W\colon t \mapsto W(t)$	Curve on $\widehat{\mathfrak{W}}$

Table 2.1: Table contains notations used for graphons and measure-valued graphons. Each row contains the corresponding notation used in both these settings in the article.

Definition 2.4.5 (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 [212, Section 1.2.1, Corollary 1.16]) which considers all Lipschitz functions.

Definition 2.4.6 (Wasserstein Cut metric on $\widehat{\mathfrak{W}}$)**.** *Define* $\mathbb{W}_{\blacksquare}\colon \widehat{\mathfrak{W}} \times \widehat{\mathfrak{W}} \to \mathbb{R}_+$ *as*

$$\mathbb{W}_{\blacksquare}(W_1, W_2) \coloneqq \inf_{\varphi_1,\varphi_2\in\mathcal{T}} \sup_{S,T\subseteq[0,1]} \mathbb{W}_1\left(\int_{S\times T} W_1^{\varphi_1}(x,y)\,\mathrm{d}x\,\mathrm{d}y, \int_{S\times T} W_2^{\varphi_2}(x,y)\,\mathrm{d}x\,\mathrm{d}y\right).$$

With this setup, we can now state Theorem 2.4.7. The proof of this theorem relies on several lemmas which are proved in Section 2.4.1. We, therefore, also defer the proof of Theorem 2.4.7

Theorem 2.4.7. *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. $\mathbb{W}_{\blacksquare}(W_n, W) \to 0$*.*

Perhaps surprisingly, we show that the metrics in Definitions 2.4.5 and 2.4.6 are exactly equal.

Proposition 2.4.8. *Let* $\mathbb{W}_{\blacksquare}$ *and* $\Delta_{\blacksquare}$ *be as defined above. Then,* $\mathbb{W}_{\blacksquare}$ *and* $\Delta_{\blacksquare}$ *are metrics on* $\widehat{\mathfrak{W}}$*. Furthermore,* $\mathbb{W}_{\blacksquare} = \Delta_{\blacksquare}$*.*

Proof. 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 $(\mathcal{L}_{S\times T}W) := \int_{S\times T} W(x,y)\,\mathrm{d}x\,\mathrm{d}y$ 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 sides 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 [151]). 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 2.4.7. □

Remark 2.4.9. *It follows from [150, Theorem 17.9] that $\widehat{\mathfrak{W}}$ is compact (with respect to the usual topology). 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 2.4.7, $(\widehat{\mathfrak{W}}, \mathbb{W}_\blacksquare)$ (or $(\widehat{\mathfrak{W}}, \Delta_\blacksquare)$) is a compact metric space.*

It is clear from Definition 2.4.5 and Theorem 2.4.7 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 3. 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 $\big\|m_k(W_n^{\varphi_{n,k}}) - m_k(W)\big\|_\square \to 0$ 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

'labelings' (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 4. 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. 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 6. 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 $n \to \infty$.

We close this section with the following definition and Lemma. For $W_1, W_2 \in \mathfrak{W}$, define the Wasserstein-2 metric D_2 on $\mathfrak{W}$ as

$$D_2^2(W_1, W_2) \coloneqq \int_{[0,1]^2} \mathbb{W}_2^2(W_1(x,y), W_2(x,y))\,\mathrm{d}x\,\mathrm{d}y, \tag{2.30}$$

where $\mathbb{W}_2$ is the Wasserstein-2 metric on $\mathcal{P}([-1,1])$.

Definition 2.4.10 (Invariant Wasserstein-2 metric on $\widehat{\mathfrak{W}}$)**.** *Define* $\Delta_2 \colon \widehat{\mathfrak{W}} \times \widehat{\mathfrak{W}} \to \mathbb{R}_+$ *as*

$$\Delta_2^2(W_1, W_2) \coloneqq \inf_{\varphi_1, \varphi_2 \in \mathcal{T}} D_2^2(W_1^{\varphi_1}, W_2^{\varphi_2}), \qquad W_1, W_2 \in \widehat{\mathfrak{W}}.$$

The following is easy to see.

Lemma 2.4.11. $\Delta_\blacksquare \leq \Delta_2$.

Proof. 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 2.4.8, we observe that

$$\begin{aligned}
&\mathbb{W}_1\left(\int_{S\times T} W_1(x,y)\,\mathrm{d}x\,\mathrm{d}y, \int_{S\times T} W_2(x,y)\,\mathrm{d}x\,\mathrm{d}y\right) \\
&= \sup_{\psi\in\mathcal{L}}\left|\int_{S\times T} V_\psi(x,y)\,\mathrm{d}x\,\mathrm{d}y\right| \leq \sup_{\psi\in\mathcal{L}} \int_{[0,1]^2} |V_\psi(x,y)|\,\mathrm{d}x\,\mathrm{d}y \leq \int_{[0,1]^2} \sup_{\psi\in\mathcal{L}} |V_\psi(x,y)|\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{[0,1]^2} \mathbb{W}_1(W_1(x,y), W_2(x,y))\,\mathrm{d}x\,\mathrm{d}y \leq \int_{[0,1]^2} \mathbb{W}_2(W_1(x,y), W_2(x,y))\,\mathrm{d}x\,\mathrm{d}y \\
&\leq \left(\int_{[0,1]^2} \mathbb{W}_2^2(W_1(x,y), W_1(x,y))\,\mathrm{d}x\,\mathrm{d}y\right)^{1/2},
\end{aligned}$$

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 $\varphi_1, \varphi_2 \in \mathcal{T}$.

□

2.4.1 Remaining proofs

Lemma 2.4.12. *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 $\max_{i,j\in E(F)} \|f_{i,j} - g_{i,j}\|_\infty \leq \epsilon$. Let $C > 0$ be a finite constant such that $\max_{i,j} \|f_{i,j}\|_\infty \leq C$. It follows that $\max_{i,j} \|g_{i,j}\|_\infty \leq C' = (1 + C)$. Using the Counting Lemma for decorated graphs [151, Lemma 10.26], for any $U \in \widehat{\mathfrak{W}}$ we have $|t_{\mathrm{d}}((F, g), U) - t_{\mathrm{d}}((F, f), U)| \leq 4|E(F)|C'\epsilon$. Thus

$$
\begin{aligned}
&|t_{\mathrm{d}}((F, f), W_n) - t_{\mathrm{d}}((F, f), W)| \\
&\leq |t_{\mathrm{d}}((F, f), W_n) - t_{\mathrm{d}}((F, g), W_n)| + |t_{\mathrm{d}}((F, g), W) - t_{\mathrm{d}}((F, f), W)| \\
&\qquad + |t_{\mathrm{d}}((F, g), W_n) - t_{\mathrm{d}}((F, g), W)| \\
&\leq 2C'\epsilon + |t_{\mathrm{d}}((F, g), W_n) - t_{\mathrm{d}}((F, g), W)|.
\end{aligned}
$$

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 2.4.13. *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 $t(F, \cdot)$ is as in Definition 2.26. 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 2.4.14. *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 2.3.3).*

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 2.4.12 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 [151, 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 2.4.15. *Let $\mathcal{L}$ be the space of all Lipschitz functions 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 $\Delta_{\blacksquare}^{\mathcal{F}}(U,V) := \inf_{\varphi_1,\varphi_2\in\mathcal{T}} \sup_{\psi\in\mathcal{F}} \|\Gamma(\psi,U^{\varphi_1}) - \Gamma(\psi,V^{\varphi_2})\|_{\square}$, 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. It is an immediate consequence of the 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

$$\left|\sup_{\psi\in\mathcal{L}} \|\Gamma(\psi,U) - \Gamma(\psi,V)\|_{\square} - \sup_{\psi\in\mathcal{F}_\epsilon} \|\Gamma(\psi,U) - \Gamma(\psi,V)\|_{\square}\right| < \epsilon. \tag{2.31}$$

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 2.4.7. Equivalence of (1) and (2) follows from Lemma 2.4.13. Lemma 2.4.14 shows that (3) (or equivalently (4)) implies (1). It remains to show that (1) implies (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 2.4.15 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 [153, 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. □

2.5 Infinite exchangeable arrays, graphons and measure-valued graphons

In this section, we recall the definition of infinite exchangeable arrays (IEAs) from the Introduction and explain how an IEA naturally gives rise to a graphon and a measure-valued graphon via Aldous-Hoover representation. With some examples, we show why graphons do not adequately describe an IEA. Finally, we show (Theorem 2.5.3) that IEAs are in one-to-one correspondence with random measure-valued graphons.

2.5.1 IEAs and Aldous-Hoover

Recall that an IEA is an infinite array $\mathbf{X} = (X_{i,j})_{i,j\in\mathbb{N}}$ of random variables which is symmetric $X_{i,j} = X_{j,i}$ and $\mathbf{X}^\sigma = (X_{\sigma(i),\sigma(j)})_{i,j\in\mathbb{N}}$ has the same law as $\mathbf{X}$ for every finite permutation

σ of $\mathbb{N}$. In the following discussion, we will often assume that IEA $\mathbf{X}$ takes values in $[-1,1]$.

IEAs are intimately related to graphons and measure-valued graphons. We now describe this relation. In the following 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. Uniform$[0,1]$ random variables on some probability space. An IEA $\mathbf{X}$ is said to be directed by an Aldous-Hoover function $f : [0,1]^4 \to [0,1]$, if $\mathbf{X} \stackrel{d}{=} f(U, U_i, U_j, U_{i,j})$. Aldous-Hoover representation theorem states that every exchangeable array $\mathbf{X}$ can is directed by some Borel measurable function $f : [0,1]^4 \to [-1,1]$ that is symmetric in the middle two coordinates, that is, $f(\cdot, x, y, \cdot) = f(\cdot, y, x, \cdot)$ [124].

Suppose $\mathbf{X}$ is an IEA that is directed by the Aldous-Hoover function f. We can naturally define a (random) graphon

$$w^{(u)}(x,y) \coloneqq \mathbb{E}[f(U, U_1, U_2, U_{1,2}) \,|\, U = u, (U_1, U_2) = (x,y)] \,. \tag{2.32}$$

Analogous to equation (2.32), one can define a (random) measure-valued kernel $W \in \mathfrak{W}$ (and hence an MVG) as

$$W^{(u)}(x,y) \coloneqq \mathrm{Law}(X_{i,j} \,|\, U = u, (U_i, U_j) = (x,y)) \,. \tag{2.33}$$

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 the 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 the Aldous-Hoover function (see [126]). Note that pure IEAs give rise to graphons and measure-valued graphons. In fact, [74, Theorem 5.3] shows that $\{0,1\}$-valued IEAs are in one-to-one correspondence with random graphons.

We now give examples of IEAs and their Aldous-Hoover representation which in turn yields the corresponding (random) MVG. These examples emphasize that graphons do not capture general IEAs while MVGs do.

Example 7 (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 (2.32) and (2.33) in Section 6.1.1) 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 8 (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

$$w^{(u)}_{\mathbf{X}}(x,y) = \alpha\Phi(u) + \beta(\Phi(x) + \Phi(y)), \qquad W^{(u)}_{\mathbf{X}}(x,y) = \mathrm{Law}\Big(\Big(\mathcal{N}\Big(w^{(u)}_{\mathbf{X}}(x,y), \gamma^2\Big)\Big)\Big),$$

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 9 (Vertex and edge dependence - Stochastic Block Model (SBM))**.** We now describe an exchangeable array that can be seen as the limit of a sequence of SBMs. 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 the 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 $(A_n)_{n\in\mathbb{N}}$.

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 $X_{i,j} \coloneqq f(U,U_i,U_j,U_{i,j})$ for $i,j \in \mathbb{N}$. The corresponding graphon and measure-valued graphon are $w^{(u)}$ and $W^{(u)}$ defined as

$$\begin{aligned} w^{(u)}(x,y) &= p_{bb}\mathbb{1}_B\{(x,y)\} + p_{rr}\mathbb{1}_R\{(x,y)\} + p_{rb}\mathbb{1}_D\{(x,y)\}, \\ W^{(u)}(x,y) &= \mathrm{Ber}(p_{bb})\mathbb{1}_B\{(x,y)\} + \mathrm{Ber}(p_{rr})\mathbb{1}_R\{(x,y)\} + \mathrm{Ber}(p_{rb})\mathbb{1}_D\{(x,y)\}, \end{aligned}$$

for a.e. $(x,y) \in [0,1]^2$. This example can be generalized to distributions other than Bernoulli straightforwardly.

2.5.2 Correspondence between IEAs and MVGs

As we saw in the previous section IEAs give rise to random measure-valued graphons. In this section, we show that in fact, this correspondence is a homeomorphism.

We begin with some definitions and notations. 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

$$\mathcal{S} := \Big\{ \mathbf{x} \in \mathbb{R}^{\mathbb{N}^2} \;\Big|\; x_{i,j} = x_{j,i} \in [-1,1],\;\; x_{i,i} = 0 \quad \forall\, i,j \in \mathbb{N} \Big\}.$$

Equip $\mathcal{S}$ with the product topology. Note that $\mathcal{S}$ 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 $\mathbf{x}^\sigma(i,j) := \mathbf{x}(\sigma(i),\sigma(j))$ 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, $\mathcal{P}_{\mathrm{e}}(\mathcal{S}) := \{\rho \in \mathcal{P}(\mathcal{S}) \mid \rho = \mathrm{Law}(\mathbf{X}), \mathbf{X} \text{ is an IEA}\}$. 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.

Definition 2.5.1 (Homomorphism density of IEAs w.r.t. decorated graphs)**.** *Let* $\mathbf{X}$ *be an IEA. For every decorated graph* F*, define* $t_{\mathrm{d}}(F,\mathbf{X}) := \mathbb{E}\Big[\prod_{\{i,j\}\in E(F)} F_{i,j}(X_{i,j})\Big]$.

The following assertion is immediate from the definition and Theorem 2.5.3. For the importance of it, we record it as a Proposition. We skip the proof.

Proposition 2.5.2. *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 2.5.3 (Homeomorphism Theorem)**.** *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* $\mathcal{P}_{\mathrm{e}}(\mathcal{S})$*.*

Proof. Recall that the Aldous-Hoover representation provides a one-to-one correspondence (see (2.33)) 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 2.5.2. 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. □

2.5.3 Convergence of exchangeable matrices 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 2.5.3). 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 2.3.2). 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 as follows. Let $\{U_i\}_{i \in \mathbb{N}}$ be a family of i.i.d. Uniform$[0, 1]$ random variables independent of A. Define an IEA $\mathbf{X}$ such that $X_{i,j} := A_{\lceil nU_i \rceil, \lceil nU_j \rceil}$.

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 $t^{(0)}_{\mathrm{finite}}(F, \cdot)$ over exchangeable matrices as $t^{(0)}_{\mathrm{finite}}(F, A) := t(F, \mathbf{X})$.

On the other hand, analogous to Definition 2.5.1 we have the following definition for homomorphism density into exchangeable matrices.

Definition 2.5.4 (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 2.5.5. *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 2.3.2, we can define $t_{\text{d}}(F, A)$ as in equation (2.24). 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 2.5.6 (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 10. Let $V\colon \mathbb{R}^{(2)} \to [-1, 1]$ be a C^2 function such that $V(x, y) = V(y, x)$ and $\left\|\nabla^2 V\right\|_\infty \leq 1$, 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. Uniform$[0,1]$. 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 11. 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 12. 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 $n \to \infty$.

2.6 Discussion

We end this chapter with a quick map of how the materials of this chapter are used later in the thesis. Of course, the introductory material on the space of graphons and its topology is used throughout. The material in Section 2.1.4 is crucial in the construction of the gradient flows on the space of graphons that is done in Chapter 4. In Chapter 5–Chapter 7 we study the limit of matrix-valued processes under different setups as the dimension goes to infinity. The motif in the current work is that while such high-dimensional matrix-valued processes

are hard to describe, under appropriate assumptions, they exhibit a propagation of chaos. This entails that a finite collection of coordinates often becomes asymptotically uncorrelated as the dimension goes to infinity. This allows us to regard an IEA as the limit of a matrix-valued process. The material developed in Section 2.5 defines this notion of convergence and explains how it is related to convergence in the sense of graphons. While the connection between IEA and graphons is well-known, as we explain in Section 2.5 an IEA carries much more information than a graphon. In other words, the correspondence between IEA and (random) graphons is lossy. More philosophically, an IEA provides the microscopic picture of the ensemble while the graphons provide a macroscopic picture. The idea of measure-valued graphons discussed in 2.3 is essentially an attempt to capture more information about an IEA than graphons. While the measure-valued graphons were introduced in [153], the metric introduced in 2.3 and its connection with IEAs and their convergence is novel. The setup of measure-valued graphons is important and probably more naturally suited to studying the evolution of random weighted graphs or matrix-valued processes, in this thesis we only use this setup in Chapter 6 (see 6.1.1) to illustrate this point.

Chapter 3

LITERATURE REVIEW

Graphons, introduced as limit objects for dense graphs, were first discussed in [39, 40, 151]. Since then, the subject has rapidly grown and has found connections with various areas of mathematics, including extremal graph theory, exchangeability [3, 14, 117, 118], exponential random graph models (ERGMs) [58, 57, 56], machine learning [190], economics, and game theory [175, 12, 47], among others. It has also inspired similar attempts to understand the limits of sparse graphs; see, for instance, [38, 34, 31], and even for the limit of bounded degree graphs [36].

While the subject of graphons is relatively young, it has seen tremendous progress. In this chapter, we survey some important developments and current research directions in this field. As the primary focus of our thesis is optimization on graphons and the limits of dynamics on large matrices or graphs resulting in processes or curves on the space of graphons, our attention will be more on related themes. However, we will also describe other important research directions that may not directly relate to our work.

3.1 Graphon driven interacting particle systems

Recall from Chapter 1 that the classical mean-field interacting particle system studies the evolution of the empirical measure of the system of n particles that evolve like

$$\mathrm{d}X_{i,n}(t) = b(X_{i,n}(t), \mu_n(t))\,\mathrm{d}t + \mathrm{d}B_i(t), \qquad i = 1, \dots, n,$$

where B_i are independent Brownian motions and $\mu_n(t) = n^{-1}\sum_{i=1}^{n} \delta_{X_{i,n}(t)}$.

More recently, graphon-driven interacting particle systems have been investigated by various authors. The idea here is that particles $X_{i,n}$ live on the vertices of a graph G_n. Each particle $X_{i,n}$ is driven by an independent Brownian motion and it interacts only with its neighbors that are determined by the underlying graph G_n. In other words, the evolution

of this particle system looks like

$$\mathrm{d}X_{i,n}(t) = b(X_{i,n}(t), \mu_{i,n}(t))\,\mathrm{d}t + \mathrm{d}B_i(t), \qquad i = 1, \ldots, n,$$

where B_i are independent Brownian motions and $\mu_{i,n}(t) = \frac{1}{d_i}\sum_{j\sim i}\delta_{X_{i,n}(t)}$, here $j \sim i$ means the vertex j and i are neighbors in the graph G_n and d_i is the number of neighbors of i.

The main problem here is understanding the limiting behavior of such a system where G_n is a sequence of (possibly random) graphs that converge to a (deterministic) graphon, say W, as $n \to \infty$. The study of such an interacting system of particles was initiated in [28] for the case when G_n is an Erdös-Rényi graph. When the underlying graph G_n is the complete graph, we recover the classical mean-field interacting particle systems described in Chapter 1.

Similar models, varying in generality, have been studied extensively since then in [68, 138, 27, 144, 22, 66, 67, 28, 77, 71]. In many of these cases, if the underlying graph sequence that governs the interaction of particles is sufficiently dense, one observes the emergence of the classical McKean-Vlasov equations in the limit. It's important to note that this differs from the McKean-Vlasov equation introduced in [102] and discussed extensively in Chapter 5. In essence, the limiting description of the particle system interacting via a dense graph becomes independent of the precise nature of the underlying graphs when they are dense enough. Similar questions have also been explored in the context of sparse underlying graph sequences [143, 170, 27, 20, 171, 144].

It's important to note that the particle system studied in this thesis involves the evolution of edges rather than vertices. Specifically, we consider a symmetric evolution where the graph itself evolves. Previous works have primarily focused on particle systems where the underlying graph controls interactions among particles but the underlying graph itself does not evolve in time.

Some recent works have explored particle systems where interaction is determined by an evolving underlying graph, where the evolution of the graph itself depends on the particle positions [19, 99, 21]. This is closer in spirit to our work. This field presents numerous open problems; interested readers can find an excellent discussion on some of these in [215].

3.2 Evolution of dense graphs and their limits

In [70], the author considers Markov processes with cádlág paths on graphons, which can be obtained by projecting Markov processes with cádlág paths on infinite exchangeable arrays. Unfortunately, such processes on the space of graphons cannot be diffusive and are thus, in some sense, trivial. However, this does not rule out the possibility of obtaining diffusions on graphons. For instance, [8] demonstrates the existence of naturally occurring diffusions on the space of graphons that arise as the limit of processes in population dynamics models. The convergence of natural graph dynamics to processes on graphons has been further explored in [10]. Similar constructions have also been obtained in the sparse graph regime, as seen in [11, 69].

We briefly explain the idea from [10] for the convenience of the reader. Consider a collection of n individuals where each individual is labeled with Type 0 or 1. Let $G^n(0)$ be a graph on the vertex set $[n]$ where i and j are connected if and only if i and j have the same type. Now we evolve this graph as follows: every individual (that is vertex) independently at rate 1 picks another individual (from the whole population not necessarily from its neighbors) and adapts its type. Define the graph $G^n(s)$ analogous to $G^n(0)$ at time $s \geq 0$. Let $H^n(s) = \frac{1}{n}G^n(sn)$. Note that if the fraction of Type 0 individual in the population converges to some constant $p \in [0,1]$. Then, $G^n(0)$ converges to a kernel W such that $W(x,y) = 1$ if $(x,y) \in [0,p] \times [0,p] \cup [1-p,1] \times [1-p,1]$ and 0 otherwise. One can ask if $H^n(s)$ converges weakly, say in $\delta_\square$ metric, to a process on the space of graphons. In other words, if F is a finite graph, can one show that the homomorphism density of F into $H^n(s)$, denoted $t_F(H^n(s))$, admit a weak limit as $n \to \infty$ for each s? This example is particularly simple because the homomorphism density $t_F(H^n(s))$ is completely determined by the fraction of Type 0 individual $Y^n(s)$ at time ns.

Let us denote by $X^n(s)$ the number of individuals of Type 0 in the population at time s. And, let $Y^n(s) = \frac{1}{n}X^n(sn)$ be the fraction of individual of Type 0 at (scaled) time s. It is well-known that $Y^n(s)$ converges weakly to the Wright-Fisher diffusion $Y(s)$ on $[0,1]$, that is,

$$Y(s) = p + \int_0^s \sqrt{Y(t)(1-Y(t))}\,\mathrm{d}B(t),$$

where $B(t)$ is the 1-dimensional standard Brownian motion. This in particular yields that if F has k edges then

$$t_F(H^n(s)) = Y^n(s)^k + (1 - Y^n(s))^k$$

converges weakly to $Y(s)^k + (1 - Y(s))^k$ as $n \to \infty$. One can therefore conclude that the graph $H^n(s)$ converges weakly to a process on graphons that is diffusive.

This is more in line with our work. We study the limits of certain processes on graphs and matrices as the dimension goes to infinity. In Chapter 6, we consider a Metropolis chain on large graphs and ultimately take its limit to obtain a curve on the space of graphon. One crucial difference, however, is that the symmetry conditions that we impose on our evolution force the limiting curve to be deterministic.

3.3 Exponential Random Graph Model (ERGM)

For this section, define $\widehat{\mathcal{W}_0}$ to be the space of graphons which take value in $[0, 1]$. Recall that for a finite simple graph F and a graphon W, the homomorphism density of F into W, denoted $t(F, W)$, is defined as

$$t(F, W) = \int_{[o,1]^{|V(F)|}} \prod_{\{ij\} \in E(F)} W(x_i, x_j) \prod_{\nu=1}^{|V(F)|} \mathrm{d}x_\nu,$$

where $V(F)$ and $E(F)$ denote the set of vertices and the set of edges in F, respectively.

Let $F_1, \ldots, F_k$ be finite simple graphs and let $\beta = (\beta_1, \ldots, \beta_k) \in \mathbb{R}^k$. An ERGM is a probability measure $\mathcal{P}$ on the space of all simple graphs on n, $\mathcal{G}_n$, given by

$$\mathcal{P}_n(G) = \exp\left(n^2 \sum_{i=1}^{k} \beta_i t(F_i, G) - n^2 \psi_n(\beta) \right).$$

Exponential random graphs have been studied in the statistical physics and networks community for a long time. We refer the reader to [84, 86, 176, 56]. In the study of large deviations of ERGMs, one problem of interest is to approximate $\lim_{n\to\infty} \psi_n(\beta)$. It is known that (see [56, Theorem 7.1])

$$\psi(\beta) \coloneqq \lim_{n\to\infty} \psi_n(\beta) = \sup_{W \in \widehat{\mathcal{W}_0}} \left(T(W) - \frac{1}{2} I(W) \right),$$

where $T(\cdot) = \sum_{i=1}^{k} \beta_i t(F_i, \cdot)$ and $I(W) = \int_{[0,1]^2} W(x, y) \log(W(x, y))\ \mathrm{d}x\, \mathrm{d}y$.

The minimizer W^* of $F(W) \coloneqq T(W) - \frac{1}{2}I(W)$ yields important information about the 'typical graph' drawn from $\mathcal{P}_n$ for large n. In particular, the weak law of large number ([56, Theorem 7.2]) states that for any $\eta > 0$ there exists $C > 0, \gamma > 0$ such that for any n we have

$$\mathbb{P}(\delta_\square(W^*, G_n) > \eta) \leq Ce^{-n^2\gamma}$$

where G_n is a random graph drawn from the ERGM $\mathcal{P}_n$.

However, there are only a few ERGMs where the constant $\psi(\beta) = \lim_{n\to\infty} \psi_n(\beta)$ and the minimizer of W^* are known explicitly. For instance, if all β_i are non-negative, then it is known that [58, Theorem 4.1] that any minimizer of F must be a constant graphons $W^*(x, y) \equiv p$. This reduces the problem of finding $\psi(\beta)$ and the minimizer(s) to a simple calculus problem. The parameter space for β for which the minimizer(s) are constant is called the *replica symmetric regime*. In the replica symmetry regime the typical exponential random graph 'looks like' an Erdös-Rényi graph. Understanding the replica symmetry regime is an important and challenging problem that is not fully resolved yet.

A significant amount of work has gone into understanding the replica symmetry regime and its phase transition in a particular ERGM called the *edge-triangle model* where $T(\cdot) = \beta_1 t(K_2, \cdot) + \beta_2 t(K_3, \cdot)$ where K_2 and K_3 are complete graphs on 2 and 3 vertices respectively. A curious and important insight that has emerged in this case is that the minimizer of F need not be unique even in the replica symmetry regime. We refer the reader to [181, 180, 161, 162, 182, 30] for more detail. Very little is known about the minimizers outside the replica symmetry regime.

Numerical approximations for the $\psi(\beta) = \lim_{n\to\infty} \psi_n(\beta)$ have been investigated in the PhD thesis [60] where the author computes the maximum likelihood estimate for a model of exponential random graphs analyzed in [57]. Since graphs are discrete, the optimization is more amenable to analytical tools on the limiting graphon space. But the space of graphons is infinite-dimensional and the author uses gradient descent on matrices to approximate the gradient descent on graphons–albeit without rigorous justifications. Our results show that the method is consistent as the discretization gets finer and finer and provides a mathematical justification to the algorithm in [60].

3.4 Constrained optimization on graphons

Extremal graph theory problem often involves maximizing or minimizing certain homomorphism densities while fixing some others. For instance, as discussed in 1 the problem of minimizing $t(K_3, G) - \frac{1}{10}t(K_2, G)$ can be seen as a relaxation of extremal graph theory problem that asks to maximize the number of edges in a graph while having no triangles. The extremal graph theory problem in this case is to maximize $t(K_2, G)$ over all graphs such that $t(K_3, G) = 0$. Related to the edge-triangle model discussed above, there is an extremal problem (posed by Turán in 1941) that asks the typical graph with given $(t(K_2, G), t(K_3, G)) = (\epsilon, \tau)$ where (ϵ, τ) is given. The range of achievable (ϵ, τ) is also an interesting problem with rich and long history. We refer the interested reader to [183, 179] and the references therein. In recent times, there has been a renewed interest in studying the typical behavior of a graph n vertices with given edge and triangle density. This involves understanding the large deviation of the uniform measures on $\mathcal{G}_n(\epsilon, \tau)$, the set of graphs on n vertices with $t(K_2, G) = \epsilon$ and $t(K_3, G) = \tau$ in the feasible region. The large deviation rate function turns out to be the entropy. Therefore, studying the behavior of a typical graph involves minimizing the entropy over the set of graphons with given $t(K_2, W) = \epsilon$ and $t(K_3, W) = \tau$. For the edge-triangle model, this problem has been studied in [163].

As the theory developed in this thesis does not say anything about the structure of minimizers, strictly speaking, this problem falls outside the scope of our work. However, in this problem and related problems, one can use our theory as a computational tool to come up with heuristics. Let us mention that the homomorphism density constraints like $t(F, W) = \epsilon$ are highly non-convex. Therefore, defining gradient flows on these constrained sets is not straightforward and needs more work.

3.5 Other related works

As we already surveyed a few directions of active research that are closely related to the theme of this thesis, we now survey some other relevant literature.

3.5.1 Graphon estimation

As we have seen, given a kernel $W : [0,1]^2 \to [0,1]$, one can obtain a sequence of random graphs $\mathbb{G}(n, W)$ on the vertex set $[n]$ where we create an edge between every pair of vertices i and j independently with probability $W(U_i, U_j)$. When the kernel W is a constant, this procedure recovers the usual Erdös-Rényi graph, and when W is a piecewise constant one obtains the so-called stochastic block model.

The so-called graphon estimation problem involves estimating the function W from the samples of $\mathbb{G}(n, W)$ [169, 214, 63]. There is another variant of this problem that is somewhat less demanding. This is often dubbed as graphon value estimation and it involves determining the values of the so-called latent variables $W(U_i, U_j)$ from the sample of $\mathbb{G}(n, W)$ [93, 1, 57, 54]. Interestingly, the problem of graphon estimation can be traced back to the work of Kallenberg [127]– even before graphons were developed. We refer the reader to [172] for a short and interesting discussion of the topic and recent advances in this area we refer to [46] and references therein. Most approaches to graphon estimation rely on the compactness of the space of graphons under cut-metric. In particular, one often tries to produce a graphon estimator that is a step-function. This is done via first partitioning the vertex set of the sampled graph into communities. This is in turn done via the usual clustering algorithms. After this, the step graphon is estimated by computing the average edge densities between the communities obtained in the first step. This determines a step graphon. This has naturally led to several deep and beautiful results on the number of communities into which $\mathbb{G}(n, W)$ can be partitioned. Most of these results require certain smoothness assumptions on the kernel W (like Lipschitzness) and lead to L^2-error bounds with high probability guarantees. For a more exhaustive discussion on this and related topics, we refer to the PhD thesis [135] and references therein.

3.5.2 Large deviation of graphons

We already mentioned that one source of optimization problems on the space of graphons is the large deviation of exponential random graphs. We refer the reader to [55, 56] for the general overview of the subject. For the benefit of the reader, we explain it a little. Large

deviation of Erdös-Rényi was first studied by Chatterjee and Varadhan in [58]. The Erdös-Rényi graph $E(n,p)$ induces a probability measure $\mathbb{P}_n$ on the set of all graphs on n vertices $\mathcal{G}_n$. This in turn induces a discrete probability measure say $\widetilde{\mathbb{P}}_n$ on the space of graphons. Chatterjee and Varadhan in [58] studied the large deviation of for $\widetilde{\mathbb{P}}_n$ and established that it satisfies a large deviation principle with speed n^2 and rate function given by entropy function $\mathcal{E}$ defined as $\mathcal{E}_p(W) = \int_0^1 \int_0^1 \frac{W(x,y)}{p} \log\Big(\frac{W(x,y)}{p}\Big)\, \mathrm{d}x\, \mathrm{d}y$.

More generally, Let $Q \in \mathcal{M}_k$ be a symmetric $k \times k$ matrix with entries in $[0,1]$. Let $n = (n_1, \ldots, n_k) \in \mathbb{N}^k$ be a k-tuple of natural numbers. A stochastic block model $\mathrm{SBM}(Q,n)$ is a simple random graph on $n \times k$ many vertices which is defined as follows. There are k communities and community i has n_i individuals. More precisely, we label n_i many vertices with label i for each $i \in [k]$. Now for every pair of vertices u, v, we connect them, independently of everything else, with probability p_{ij} if u is in the community i and v is in the community j. The large deviations for this model are studied in [35, 98]. This probability measure indeed satisfies a large deviation with rate n^2 with an explicit rate function. Let $W \in \mathcal{W}_k$ be a step kernel. For each $n \geq 2$, the random graph $\mathbb{G}(n,W)$ induces a probability measure on $\mathcal{G}_n$ and in turn a discrete probability measure on the space of graphons. This also satisfies a large deviation principle with rate n^2.

Let us explain that if we consider the k-step kernel W corresponding to the matrix Q and consider the random graph $\mathbb{G}(mk,W)$. This is not the same as $\mathrm{SBM}(Q,n)$ as defined in the previous paragraph even when $n_i = m$ for all $i \in [k]$. This is because in the $\mathrm{SBM}(Q,n)$ the number of individuals in each community is fixed and is equal to m, but in the model $\mathbb{G}(mk,W)$ there are on an average m individuals in each community, but the number of individuals in each community is random with binomial distribution. But this does not matter for the large deviation at speed n^2. Heuristically, this can be explained as follows. Let $n = (n_1, \ldots, n_k)$ and $n' = (n'_1, \ldots, n'_k)$ be such that $n_1 + \ldots + n_k = n'_1 + \ldots + n'_k = m$. In order to obtain $\mathrm{SBM}(Q,n)$ or $\mathrm{SBM}(Q,n')$ from $\mathbb{G}(m,W)$ we only a factor of $e^{-O(n)}$. For large deviation at speed n^2, this does not matter. Therefore, at speed n^2, the large deviation rate function remains the same for both models. Recently, the large deviation for $\mathbb{G}(n,W)$ for general graphon W has been studied in [178] at various speeds. However, the large deviation with speed n^2 in the general case remains open.

3.5.3 Subgraph count fluctuations

Recall that the convergence in cut-metric is defined via the convergence of homomorphism density functions. In particular, the fact that the Erdös-Rényi graph $E(n,p)$ converges to a graphon $W_p \equiv p$ as $n \to \infty$ is equivalent to saying that for any finite simple graph F with m-edges, the homomorphism density $t(F, E(n,p))$ converges to $t(F, W_p) = p^m$ almost surely. This is straightforward to verify and can be established using a fourth-moment bound. Naturally, this leads to the study of fluctuations in these subgraph counts.

For the Erdös-Rényi graph the fluctuations of subgraph counts have been studied for a long time [18, 115, 116]. For more general graphons, the fluctuations of homomorphism densities in $\mathbb{G}(n,W)$ have been studied recently. The fluctuation of many classes of homomorphism densities (e.g. for cliques) is known to be Gaussian. However, in many cases, the fluctuation may become degenerate. We refer the reader to the recent works [29, 106, 83, 85]. We close this section with two recent papers [131, 53] and the references therein. The joint distribution of multiple subgraph densities are studied in[131, 53] and their asymptotic normality is established.

Chapter 4

GRADIENT FLOWS ON GRAPHONS

4.1 Introduction

For any metric space (Ω, d), denote its Borel sigma algebra as $\mathcal{B}(\Omega)$ and the set of all Borel probability measures as $\mathcal{P}(\Omega)$. 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

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

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 $[n] \coloneqq \{1, 2, \ldots, n\}$. Throughout this text, we will denote the symmetric group on the finite set $[n]$ as S_n. Consider the Cauchy problem

$$\dot{x}_i(t) = -\nabla_i f_n(x_1(t), \ldots, x_n(t)), \quad i \in [n], \qquad t \in \mathbb{R}_+, \tag{4.1}$$

with given initial conditions $(x_i(0))_{i\in[n]}$. Here ∇_i refers to the partial derivative with respect to the i-th variable and $\mathbb{R}_+$ denotes the set of all non-negative real numbers. The solution to this problem–which exists and is unique when, say, ∇f_n is Lipschitz– is often called the gradient flow of f_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 [212, 192] 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 (AKA gradient flows) developed for functions on metric spaces which may lack a differentiable structure [5]. The Wasserstein space is a prominent example that has been thoroughly studied [5, 193]. Recently there has been a surge in interest in the application of the above convergence of gradient flows in the context of single hidden layer neural networks, see [201, 62, 188, 200, 48, 7, 165, 196, 197, 208, 16].

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 $\{i, j\} \in E$ one has an associated variable $W_{i,j} = W_{j,i}$ that we take to be real-valued in this article. For all the applications we consider, we can take $W_{i,j} = 0$ if $\{i, j\} \notin E$. Thus our variables can be arranged in an $n \times n$ symmetric matrix $(W_{i,j})_{i,j\in[n]}$. Let the set of $n \times n$ real-valued symmetric matrices be denoted by $\mathcal{M}_n(\mathbb{R})$. Let $f_n \colon \mathcal{M}_n(\mathbb{R}) \to \mathbb{R} \cup \{\infty\}$ be a function of such matrices. The crucial difference from the previous set-up is that we want f_n to satisfy a permutation invariance property with respect to relabeling the vertices of G: for every $\pi \in S_n$,

$$f_n\left(\left(W_{\pi_i,\pi_j}\right)_{i,j\in[n]}\right) = f_n\left(\left(W_{i,j}\right)_{i,j\in[n]}\right).$$

That is, the function value does not change if we permute the rows and the columns of the symmetric matrix $(W_{i,j})_{i,j\in[n]}$ by the same permutation. In other words, such functions are invariant under graph isomorphisms of G. We call such functions *invariant*. One can now ask the same question as before. Consider the gradient flow Cauchy problem

$$\dot{W}_{i,j}(t) = -\nabla_{i,j} f_n\left(\left(W_{i,j}(t)\right)_{i,j\in[n]}\right), \quad i, j \in [n], \tag{4.2}$$

with given $(W_{i,j}(0))_{i,j\in[n]}$. Is there a suitable scaling limit as n goes to infinity? This chapter answers this question in affirmative under reasonable conditions on f_n.

We restrict ourselves to the case where the edge weights $(W_{i,j})_{i,j\in[n]}$ 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 $(W_{i,j})_{i,j\in[n]}$. 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 graphon [151, 39, 40]. This

is intimately connected with the theory of exchangeable arrays in probability theory [2, 3, 109, 126].

4.1.1 Setup and Results

Recall from Chapter 2 that we denote the set of all $n \times n$ symmetric matrices with entries in $[-1, 1]$ as $\mathcal{M}_n$. 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 $\widehat{\mathcal{W}}_n \subseteq \widehat{\mathcal{W}}$ (see Section 2.1 for details). Thus, any function $F\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ induces a sequence of functions $\left(F_n\colon \widehat{\mathcal{W}}_n \to \mathbb{R} \cup \{\infty\}\right)_{n\in\mathbb{N}}$, by restriction, and a sequence of invariant functions $(f_n)_{n\in\mathbb{N}}$ on such $n \times n$ symmetric matrices with entries in $[-1, 1]$.

The first pertinent question is that given $F\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ and $[U_0] \in \widehat{\mathcal{W}}$ in the proper effective domain $\text{eff-Dom}(F) \coloneqq \{[U] \in \widehat{\mathcal{W}} \mid F([U]) < \infty\}$ of F (see [5, equation 1.2.1]), under what assumptions on F a "gradient flow" of F on $(\widehat{\mathcal{W}}, \delta_2)$ exists starting at $[U_0]$. On a general metric space, a gradient flow curve (i.e., a curve of maximal slope [5, Definition 1.3.2]) is obtained by showing that the limits of the solutions of implicit Euler iterations (see Section 4.2) exist and satisfy added assumptions. The existence of the limit of the sequence of such implicit Euler iterations requires the *local slope* $|\partial F|$ of F, defined as

$$|\partial F|([V]) \coloneqq \limsup_{[W]\in\widehat{\mathcal{W}},\, \delta_2([W],[V])\to 0} \frac{(F([V]) - F([W]))^+}{\delta_2([W], [V])}. \tag{4.3}$$

In practice, however, evaluating the local slope is not easy. We introduce the concept of *Fréchet-like derivative* in Section 4.2.3 and show that for functions that have Fréchet-like derivative, the local slope $|\partial F|$ admits a more amenable expression in terms of L^2-norm of the Fréchet-like derivative. Moreover, under a semiconvexity assumption, just the existence of Fréchet-like derivative suffices for the existence of a curve of maximal slope (see Theorem 4.2.14).

Since $(\widehat{\mathcal{W}}, \delta_2)$ is a geodesic space, one can talk about λ-semiconvex functions for $\lambda \in \mathbb{R}$ over geodesics and generalized geodesics (see Section 2.1 and Section 2.2). Theorem 4.1.1 shows that under suitable assumptions on a semiconvex function F, the gradient flow of F on $(\widehat{\mathcal{W}}, \delta_2)$ can be obtained as the time-scaled limit of the Euclidean gradient flows of $(f_n)_{n\in\mathbb{N}}$ on $(\mathcal{M}_n)_{n\in\mathbb{N}}$ respectively. That is, suppose we have a sequence of block graphons

$([U_{n,0}] \in \widehat{\mathcal{W}}_n)_{n\in\mathbb{N}} \xrightarrow{\delta_\square} [U_0] \in \widehat{\mathcal{W}}$ and let $\omega = (\omega_t)_{t\in\mathbb{R}_+}$ be the gradient flow of F on $(\widehat{\mathcal{W}}, \delta_2)$ starting at $\omega_0 = [U_0] \in \widehat{\mathcal{W}}$. Then the gradient flow $\omega^{(n)} = \left(\omega_t^{(n)}\right)_{t\in\mathbb{R}_+}$ of F_n starting at $\omega_0^{(n)} = [U_{k,0}] \in \widehat{\mathcal{W}}_n$ converges, in $\delta_\square$ to ω uniformly over compact time intervals as $k \to \infty$. The next argument shows that for any $n \in \mathbb{N}$, $\omega^{(n)}$ is a time-scaling of the Euclidean gradient flow of f_n.

For any $n \in \mathbb{N}$, the Euclidean gradient flow of the function f_n over $\mathcal{M}_n$ can be approximated via the implicit Euler method. Starting from $X_{n,\tau} \in \mathcal{M}_n$ with a step size of $\tau > 0$, the next iterate of the implicit Euler method, say $X_{n,\tau,+}$, is obtained as

$$X_{k,\tau,+} \in \operatorname*{arg\,min}_{X_n \in \mathcal{M}_n} \left[f_n(X_n) + \frac{1}{2\tau} \|X_n - X_{n,\tau}\|_2^2 \right]. \tag{4.4}$$

Let K be the natural embedding map from $n \times n$ symmetric matrices to the space of block kernels $\mathcal{W}_n$ (Definition 2.1.6). Since the function f_n is permutation invariant, setting $[U_{n,\tau}] = [K(X_{n,\tau})]$, equation (4.4) is equivalent to obtaining

$$\begin{aligned} [U_n] &\in \operatorname*{arg\,min}_{[U_n]\in\widehat{\mathcal{W}}_n} \left[F_n([U_n]) + \frac{n^2}{2\tau} \min_{\substack{X_n\in\mathcal{M}_n,\\ [U_n]=[K(X_n)]}} \frac{1}{n^2} \sum_{i,j=1}^{n} \left| (X_n)_{i,j} - (X_{n,\tau})_{i,j} \right|^2 \right] \\ &= \operatorname*{arg\,min}_{[U_n]\in\widehat{\mathcal{W}}_n} \left[F_n([U_n]) + \frac{n^2}{2\tau} \delta_2^2([U_n], [U_{n,\tau}]) \right], \end{aligned} \tag{4.5}$$

via the substitution $[U_n] = [K(X_n)]$. Equation (4.5) is precisely the implicit Euler iteration for gradient flow on $(\widehat{\mathcal{W}}_n, \delta_2)$ with a step size of τ/n^2 (see Section 4.2.1). Since iterations of the form in equation (4.4) converge to the Euclidean gradient flow as $\tau \to 0$, its image via the map $X \mapsto [K(X)]$ in equation (4.5) converges to the gradient flow on $(\widehat{\mathcal{W}}_n, \delta_2)$ as $\tau \to 0$.

Theorem 4.1.1 (Convergence of Gradient Flows)**.** *Suppose $F\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ satisfies the following conditions:*

1. *F is continuous in $\delta_\square$.*

2. *F is λ-semiconvex (Definition 2.1.15) along generalized geodesics on $(\widehat{\mathcal{W}}, \delta_2)$ (Definition 2.2.6), 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* F *on each* $\widehat{\mathcal{W}}_n$*, starting at some* $\omega_0^{(n)} = [U_{n,0}]$ *for* $n \in \mathbb{N}$*. Assume that the sequence* $([U_{n,0}])_{n\in\mathbb{N}} \xrightarrow{\delta_\square} [U_0] \in \widehat{\mathcal{W}}$*, and* $|\partial F|([U_0]) < \infty$ *and* $\limsup_{n\to\infty} |\partial F|([U_{k,0}]) \le G < \infty$*, for some* $G \ge 0$*. Then,*

$$\limsup_{n\to\infty} \sup_{t\in[0,T]} \delta_\square\left(\omega_t^{(n)}, \omega_t\right) = 0, \tag{4.6}$$

for any $T \in \mathbb{R}_+$*, where* $\omega = (\omega_t)_{t\in\mathbb{R}_+}$ *is the unique minimizing movement curve [5, Definition 2.0.6, page 42] on* $\widehat{\mathcal{W}}$ *for the function* F *starting at* $\omega_0 = [U_0]$*. In addition, if the conditions for the existence of curves of maximal slope (Theorem 4.2.4 or Theorem 4.2.14) hold, then* ω *is a curve of maximal slope.*

Remark 4.1.2. *Condition 2 in Theorem 4.1.1 is satisfied if the invariant extension* $f\colon \mathcal{W} \to \mathbb{R}\cup\{\infty\}$ *of* F *is* λ*-semiconvex on* $(\mathcal{W}, d_2)$ *(see Section 2.1.1, and Definition 2.1.16).*

An important and recurring theme throughout this thesis is that many important curves in $\widehat{\mathcal{W}}$ are obtained as the projection of some curve defined in $\mathcal{W}$. In particular, the gradient flow of F is the natural image of an absolutely continuous curve in $(\mathcal{W}, d_2)$. More precisely, if f has a Fréchet-like derivative $D_{\mathcal{W}} f(W)$ for all $W \in \mathcal{W}$, then there gradient flow ω of F can be written as the image of the curve $([W_t])_{t\in\mathbb{R}_+}$ defined as

$$W_t(x,y) = W_0(x,y) - \int_0^t D_{\mathcal{W}} f(W_s)(x,y) \mathbb{1}_{G_{W_s}}\{(x,y)\}\,\mathrm{d}s,$$

for a.e. $(x,y) \in [0,1]^{(2)}$, where $\mathbb{1}_{G_{W_s}}$ is a 'boundary correction term' that makes sure that the curve remains inside $[-1,1]$. The reader is referred to Section 4.2.3 for details. In this sense, our work is in the vein of [92] where the authors view the Wasserstein geometry as a projection of an L^2 geometry.

4.1.2 An example and discussion

To elucidate our results, consider the scalar entropy function $\mathcal{E}\colon \widehat{\mathcal{W}} \to \mathbb{R}\cup\{\infty\}$ (see [58] for applications to large deviations of Erdős-Rényi random graphs):

$$\mathcal{E}([W]) \coloneqq \int_0^1 \int_0^1 h(W(x,y))\,\mathrm{d}x\,\mathrm{d}y, \qquad [W] \in \widehat{\mathcal{W}}, \tag{4.7}$$

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.

The function $\mathcal{E}$ is lower semicontinuous in the cut metric [58, Lemma 2.1]. However, for a given $\epsilon \in (0, 1/2)$, if we restrict the domain of $\mathcal{E}$ to be all $[W] \in \widehat{\mathcal{W}}$ such that $\epsilon \leq W \leq 1-\epsilon$, a.e., then $\mathcal{E}$ is $\delta_\square$-continuous on this restricted domain. The function $\mathcal{E}$ can be shown to be convex along generalized geodesics and its local slope can be computed easily (see Section 4.5 for all the details). The Fréchet-like derivative $D_{\mathcal{W}}\mathcal{E}$ of $\mathcal{E}$ at any such graphon $[W]$ is given by another graphon that is "coupled" with $[W]$ (see Definition 4.2.8)

$$D_{\mathcal{W}}\mathcal{E}(W)(x,y) = \log\left(\frac{W(x,y)}{1-W(x,y)}\right), \qquad (x,y) \in [0,1]^{(2)}. \tag{4.8}$$

Thus, starting from a graphon $[U_0]$ such that $\epsilon \leq U_0 \leq 1-\epsilon$, a.e., the gradient flow ω evolves every coordinate of the graphon ω_t at time $t \in \mathbb{R}_+$ by the velocity $-D_{\widehat{\mathcal{W}}}\mathcal{E}(\omega_t)$ (see Section 4.2.3). But the expression in equation (4.8) is positive for any $W(x,y) > 1/2$ and negative for any $W(x,y) < 1/2$. Thus, the gradient flow converges, as $t \to \infty$, to the constant graphon $\omega_\infty \equiv 1/2$, a.e., which is the unique minimizer of the function $\mathcal{E}$. Restrict the domain of the scalar entropy function on $n \times n$ symmetric matrices A with entries in $[\epsilon, 1-\epsilon]$. Define $\mathcal{E}_n(A) \coloneqq n^{-2}\sum_{i=1}^n \sum_{j=1}^n h(A_{i,j})$. Then Theorem 4.1.1 further says the following. If $\omega^{(n)}$ is an Euclidean gradient flow of $\mathcal{E}_n$, and if $\delta_\square\text{-}\lim_{n\to\infty} \omega^{(n)}(0) = [U_0] \in \widehat{\mathcal{W}}$, then $(\omega^{(n)})_{n\in\mathbb{N}}$ converges, uniformly in the cut metric on compact sets $[0,T]$ for $T > 0$, as $k \to \infty$, to the curve ω described.

It is important to note that the curve ω is actually obtained as the natural image of the curve $t \mapsto W_t$ obtained by solving

$$W_t(x,y) = W_0(x,y) - \int_0^t \log\left(\frac{W_s(x,y)}{1-W_s(x,y)}\right) \mathrm{d}s, \quad \text{a.e. } (x,y) \in [0,1]^{(2)}, \tag{4.9}$$

for $t \in \mathbb{R}_+$, where the above integral is defined pointwise (See Section 4.5 for details). It is a recurring theme in this chapter that many important curves in $\widehat{\mathcal{W}}$ can be seen as the natural image of a curve in $\mathcal{W}$.

More examples have been worked out in Section 4.5. This includes the case when F is a homomorphism density function. Our examples also cover the gradient flow of any linear combination of the scalar entropy function and homomorphism density functions that are of

particular interest in the study of exponential random graph models (see [56, 57, 133, 80, 96]) and in the large deviation principle of dense random graphs [58, 56, 154, 65] where one is interested in optimizing the so-called rate function.

It is well-known in optimal transport that an absolutely continuous curve in the Wasserstein space has an associated continuity equation [192, Theorem 5.14]. Proposition 4.4.3 gives a partial analogue of this result in the current setting. That is, a curve of gradient flow (which is absolutely continuous) has an associated family of continuity equations, not just one continuity equation.

4.1.3 Notations recall

Throughout the chapter, we use the symbols A, X, Y, and Z (and their variations with sub/superscripts, hats, tildes, and primes) to represent matrices. For example, X_n, A_n would stand for $n \times n$ (symmetric) matrices.

We use the letters U, V, W to represent *kernels*, that is, symmetric real-valued Borel measurable functions on the unit square. The set of all kernels is denoted by $\mathcal{W}$ and $\widehat{\mathcal{W}}$ denotes the set of graphons.

For a kernel V, we always use $[V]$ to represent its corresponding graphon when we wish to emphasize the equivalence class. Note that a $n \times n$ matrix, say A_n, can be naturally embedded in the space of kernels (see Section 2.1 for details). We use $K(A_n)$ to denote the kernel corresponding to the matrix A_n.

4.2 Gradient Flows on Graphons

The goal of this section is to prove the existence of gradient flow (see Theorem 4.2.14) and the convergence of Euclidean gradient flows to the gradient flow on graphons (Theorem 4.1.1). In Section 4.3 we give the proof of Theorem 4.1.1. This section can be read directly after Proposition 4.2.5.

In Section 4.2.1 we define an implicit Euler scheme following [5] that allows one to prove the existence of gradient flow in Theorem 4.2.4. However, the conditions in Theorem 4.2.4 are hard to verify. Later we prove an alternate existence theorem for gradient flow. However,

the main highlight of Section 4.2.1 is Proposition 4.2.5 where we prove a Γ-convergence result for Euler iterates on finite-dimensional matrices (or step-kernels). This result is crucial in the proof of Theorem 4.1.1 and we believe that it would be of independent interest as well.

In Section 4.2.3 we introduce a notion of 'Fréchet-like derivative' which in turn allows us to prove the existence of gradient flow for function F that have Fréchet-like derivative. We do verify this condition for a large class of functions in Section 4.5. As pointed out in the introduction, the existence of a Fréchet-like derivative not only allows us to prove the existence of gradient flow, but it also allows a kernel representation of the gradient flow (see Theorem 4.2.14) which is extremely useful in practice.

Finally, Section 4.4 complements the discussion by proving that the gradient flows in $(\widehat{\mathcal{W}}, \delta_2)$ can also be described by a family of continuity equations.

4.2.1 Implicit Euler method, Generalized Minimizing Movements

Here we introduce an implicit Euler scheme and use to show the existence of gradient flow. We begin with the setup and definitions. Given $F\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$, a step size $\tau > 0$ and $[U] \in \widehat{\mathcal{W}}$, define a functional $\Phi_F(\tau, [U]; \cdot)\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$, called *penalized functional*, given by

$$\Phi_F(\tau, [U]; [V]) \coloneqq F([V]) + \frac{1}{2\tau}\delta_2^2([V], [U])\,, \tag{4.10}$$

and a set-valued *resolvent operator* J_τ on $\widehat{\mathcal{W}}$ as

$$J_\tau([U]) \coloneqq \arg\min_{\widehat{\mathcal{W}}} \Phi_F(\tau, [U]; \cdot), \quad \text{for } [U] \in \widehat{\mathcal{W}}. \tag{4.11}$$

For a sequence $\boldsymbol{\tau} \coloneqq (\tau_n)_{n\in\mathbb{N}}$ of positive time steps with $|\boldsymbol{\tau}| \coloneqq \sup_{n\in\mathbb{N}} \tau_n < \infty$, we can associate a partition of the time interval $(0, \infty)$ as

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

if $t_{\boldsymbol{\tau}}^0 = 0$ and $\lim_{n\to\infty} t_{\boldsymbol{\tau}}^n = \infty$. Given such a sequence $\boldsymbol{\tau}$ and $[U_{\boldsymbol{\tau},0}] \in \widehat{\mathcal{W}}$, we can obtain a sequence $([U_{\boldsymbol{\tau},n}])_{n\in\mathbb{N}}$ by iteratively solving for $[U_{\boldsymbol{\tau},n}]$, by setting

$$[U_{\boldsymbol{\tau},n}] \in J_{\tau_n}([U_{\boldsymbol{\tau},n-1}]), \tag{4.12}$$

provided $J_{\tau_n}([U_{\boldsymbol{\tau},n-1}])$ is non-empty for every $n \in \mathbb{N}$. Note that the iterates $([U_{\boldsymbol{\tau},n}])_{n\in\mathbb{N}}$ tries to minimize the function F at each step, but it is penalized against taking big jumps. In practice, therefore, it makes sense to treat the curve obtained by joining these iterates as a proxy for gradient flow. We need the following definition to makes this idea precise.

Definition 4.2.1 (Discrete solution)**.** *Given the sequence* $([U_{\boldsymbol{\tau},n}])_{n\in\mathbb{N}}$ *as above in equation* (4.12)*, interpolate the discrete points by a piece-wise constant left-continuous function* $\overline{[U_{\boldsymbol{\tau}}]}\colon [0,\infty) \to \widehat{\mathcal{W}}$*, defined as*

$$\overline{[U_{\boldsymbol{\tau}}]}(0) \coloneqq [U_{\boldsymbol{\tau},0}]\,, \qquad \overline{[U_{\boldsymbol{\tau}}]}(t) \coloneqq [U_{\boldsymbol{\tau},n}]\,, \quad t \in (t_{n-1}, t_n]. \tag{4.13}$$

We call $\overline{[U_{\boldsymbol{\tau}}]}$ *to be a* discrete solution *corresponding to the partition* $P_{\boldsymbol{\tau}}$*.*

Discrete solutions are simply a way of creating a curve from the iterates of resolvent operator. One should expect that as one take $|\boldsymbol{\tau}_n| \to 0$, the discrete solutions yield a curve that is often a good candidate for gradient flow (a.k.a. curves of maximal slope) in an arbitrary metric space setting. Such curves are called generalized minimizing movements that we define below.

Definition 4.2.2 (Generalized minimizing movements)**.** *For a function* F*, its corresponding functional* Φ_F *as defined in equation* (4.10)*, 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_F *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* $(\overline{[U_{\boldsymbol{\tau}_k}]})_{n\in\mathbb{N}}$ *defined as in Definition 4.2.1 such that for all* $t \in \mathbb{R}_+$*,*

$$\begin{gathered}\lim_{k\to\infty} F([U_{\boldsymbol{\tau}_k,0}]) = F([U_0]), \quad \limsup_{k\to\infty} \delta_2([U_{\boldsymbol{\tau}_k,0}], [U_0]) < \infty, \\ \delta_{\square}\text{-}\lim_{k\to\infty} \overline{[U_{\boldsymbol{\tau}_k}]}(t) = \omega_t.\end{gathered} \tag{4.14}$$

There is a related definition of *minimizing movement* (MM) curves that can be found in [5, Definition 2.0.6] where the conditions in equation (4.14) 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 $\text{MM}_{\delta_2,\delta_\square}(\Phi_F, [U_0])$ and $\text{GMM}_{\delta_2,\delta_\square}(\Phi_F, [U_0])$ respectively. From their definitions it can be verified that the set of minimizing movements is contained in the set of generalized minimizing movements. See [5, 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 (4.14) and [5, equation 2.0.10] are trivially satisfied.

Lemma 4.2.3. *If $F\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ is sequentially $\delta_\square$-lower semicontinuous, then*

1. *for every $\tau > 0$ and $[U] \in \widehat{\mathcal{W}}$, we have $\inf_{\widehat{\mathcal{W}}} \Phi_F(\tau, [U]; \cdot) > -\infty$, where Φ_F is defined in equation (4.10), and*

2. *if $([U_n])_{n\in\mathbb{N}} \subset \widehat{\mathcal{W}}$ with $\sup_{n\in\mathbb{N}} F([U_n]) < \infty$, 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 [152], from the Weierstrass Theorem [192, Box 1.1], both $\arg\min_{\widehat{\mathcal{W}}} F$ and $\arg\min_{\widehat{\mathcal{W}}} \Phi_F(\tau, [U]; \cdot)$ 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. □

From Lemma 2.1.5 we know that the topology induced by δ_2 is sequentially $\delta_\square$-lower semicontinuous. This with Lemma 4.2.3 shows that the assumptions in [5, Proposition 2.2.3] are satisfied, guaranteeing that $\text{GMM}_{\delta_2,\delta_\square}(\Phi_F, [U_0])$ is non-empty. If $|\partial F|$ is $\delta_\square$-lower semicontinuous and F is $\delta_\square$-continuous on the sublevel sets of $|\partial F|$, then it follows from [5, Theorem 2.3.1] that every element $\omega \in \text{GMM}_{\delta_2,\delta_\square}(\Phi_F, [U_0])$, for $[U_0] \in \text{eff-Dom}(F)$, is a curve of maximal slope. For the sake of clarity, we record the above discussion as a theorem.

Theorem 4.2.4 (Existence of curves of maximal slope-I)**.** *Suppose $F\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ satisfies the following conditions.*

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

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

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

Then every curve $\omega \in \mathrm{GMM}_{\delta_2,\delta_\square}(\Phi_F, [U_0])$ for $[U_0] \in \text{eff-Dom}(F)$ is a curve of maximal slope.

In practice, it is difficult to compute $|\partial F|$ or to ascertain its $\delta_\square$-lower semicontinuity. This makes it difficult to apply Theorem 4.2.4 on natural examples. Later in Theorem 4.2.14 we show that, when f 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 $|\partial F|$.

4.2.2 Γ-convergence of penalized functional

Recall that the goal of this chapter 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

$$J_\tau^{(n)}([U]) \coloneqq \operatorname*{arg\,min}_{\widehat{\mathcal{W}}_n} \Phi_F(\tau, [U]; \cdot) = \operatorname*{arg\,min}_{\widehat{\mathcal{W}}_n} \left\{ F + \frac{1}{2\tau}\delta_2^2([U], \cdot) \right\}, \tag{4.15}$$

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

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

Proposition 4.2.5. *Fix some $\delta_\square$-continuous function $F\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ 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 $\left[U_{n,\tau}^+\right] \in \arg\min_{\widehat{\mathcal{W}}_n} \Phi_F(\tau, [U_n]; \cdot)$. 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 $\left[U_{\infty,\tau}^+\right] \in \arg\min_{\widehat{\mathcal{W}}} \Phi_F(\tau, [U]; \cdot)$.*

Proof. 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 2.1.5 we have

$$\liminf_{n\to\infty} \delta_2([U_n], [W_n]) \geq \delta_2([U], [W]). \tag{4.16}$$

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

$$\lim_{n\to\infty} \delta_2([U_n],[W_n^*]) = \delta_2([U],[W])\,, \qquad \text{and} \qquad \lim_{n\to\infty} \delta_\square([W_n^*],[W]) = 0. \tag{4.17}$$

To do so, we first obtain $\varphi, \psi \in \mathcal{T}$ from Definition 2.1.4 and [117, Theorem 6.16] such that

$$\delta_2([U],[W]) = \left\| U^\varphi - W^\psi \right\|_2. \tag{4.18}$$

Since $\delta_\square([U_n],[U]) \to 0$ as $n \to \infty$, using [150, Theorem 11.59] we can find $(\varphi_n \in \mathcal{I}_n)_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty} \|U_n^{\varphi_n} - U^\varphi\|_\square = 0. \tag{4.19}$$

We now define a sequence of kernels $(Z_n \in \mathcal{W}_n)_{n\in\mathbb{N}}$ as

$$Z_n := U_n^{\varphi_n} - \mathbb{E}[U^\varphi \mid \mathcal{F}_n],$$

where $\mathcal{F}_n = \sigma\{Q_n \times Q_n\}$ for every $n \in \mathbb{N}$. Note that

$$\begin{aligned} \|Z_n\|_\square &\le \|U_n^{\varphi_n} - U^\varphi\|_\square + \|U^\varphi - \mathbb{E}[U^\varphi \mid \mathcal{F}_n]\|_\square \\ &\le \|U_n^{\varphi_n} - U^\varphi\|_\square + \|U^\varphi - \mathbb{E}[U^\varphi \mid \mathcal{F}_n]\|_2. \end{aligned}$$

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 (4.19) we conclude that

$$\|Z_n\|_\square \to 0, \qquad \text{as} \quad n \to \infty. \tag{4.20}$$

The sequence of kernels $(W_n^* \in \mathcal{W}_n)_{n\in\mathbb{N}}$ can now be defined as

$$W_n^* := \mathbb{E}\Big[W^\psi \,\Big|\, \mathcal{F}_n\Big] + Z_n.$$

It now follows that for any $n \in \mathbb{N}$,

$$\begin{aligned} \left\|W_n^* - W^\psi\right\|_\square &\le \left\|\mathbb{E}\Big[W^\psi \,\Big|\, \mathcal{F}_n\Big] - W^\psi\right\|_\square + \|Z_n\|_\square \\ &\le \left\|\mathbb{E}\Big[W^\psi \,\Big|\, \mathcal{F}_n\Big] - W^\psi\right\|_2 + \|Z_n\|_\square. \end{aligned}$$

Using L^2 convergence of the martingales, and equation (4.20) we obtain $\|(W_n^*)^{\psi_n} - W^{\psi}\|_{\square} \to 0$ and therefore have

$$\limsup_{n\to\infty} \delta_{\square}([W_n^*],[W]) = 0. \tag{4.21}$$

Moreover,

$$\begin{aligned}\|U_n^{\varphi_n} - W_n^*\|_2^2 &= \left\|\mathbb{E}[U^{\varphi} \mid \mathcal{F}_n] - \mathbb{E}\Big[W^{\psi} \;\Big|\; \mathcal{F}_n\Big]\right\|_2^2 \\ &\le \left\|U^{\varphi} - W^{\psi}\right\|_2^2 = \delta_2^2([U],[W]) \qquad \text{(using equation (4.18))},\end{aligned} \tag{4.22}$$

where the last inequality follows from [150, Equation 9.7]. From equation (4.22) and Lemma 2.1.5 we obtain

$$\lim_{n\to\infty} \delta_2([U_n],[W_n^*]) = \delta_2([U],[W]). \tag{4.23}$$

Now, by the definition of $U_{n,\tau}^+$, we have

$$F\big(\big[U_{n,\tau}^+\big]\big) + \frac{1}{2\tau}\delta_2^2\big([U_k],\big[U_{n,\tau}^+\big]\big) \le F([W_n^*]) + \frac{1}{2\tau}\delta_2^2([U_n],[W_n^*]). \tag{4.24}$$

Taking $\liminf_{n\to\infty}$ on both sides of equation (4.24), and from equation (4.16), equation (4.17) and the $\delta_{\square}$-continuity of F, we get

$$\begin{aligned}&F\big(\big[U_{\infty,\tau}^+\big]\big) + \frac{1}{2\tau}\delta_2^2\big([U],\big[U_{\infty,\tau}^+\big]\big) \\ &\le \liminf_{n\to\infty} F\big(\big[U_{n,\tau}^+\big]\big) + \liminf_{n\to\infty} \frac{1}{2\tau}\delta_2^2\big([U_n],\big[U_{n,\tau}^+\big]\big) \\ &\le \liminf_{n\to\infty} F([W_n^*]) + \liminf_{n\to\infty} \frac{1}{2\tau}\delta_2^2([U_n],[W_n^*]) = F([W]) + \frac{1}{2\tau}\delta_2^2([U],[W]).\end{aligned} \tag{4.25}$$

Since $[W] \in \widehat{\mathcal{W}}$ was arbitrary, this completes the proof. □

4.2.3 Fréchet-like derivatives and Existence of gradient flow

In Section 4.2.3 we introduce the notion of Fréchet-like differentiability. The most important result in Section 4.2.3 is Lemma 4.2.9 which relates the Fréchet-like derivative with the local slope of the function. This plays a crucial role in Section 4.2.4 where we show the existence of gradient flow in Theorem 4.2.14.

Fréchet-like derivatives and local slope

Recall that given a function $F\colon \widehat{\mathcal{W}} \to \mathbb{R}\cup\{\infty\}$, we can define an invariant function $f\colon \mathcal{W} \to \mathbb{R}\cup\{\infty\}$ such that $f = F \circ [\,\cdot\,]$.

Definition 4.2.6 (Fréchet-like derivative on $\mathcal{W}$)**.** *Suppose $f\colon \mathcal{W} \to \mathbb{R}\cup\{\infty\}$ is an invariant function. Let $V \in$ eff-Dom(f). The* Fréchet-like derivative *at V is given by any $\phi \in L^\infty\big([0,1]^{(2)}\big)$ that satisfies the following condition,*

$$\lim_{W\in\mathcal{W},\, \|W-V\|_2\to 0} \frac{f(W) - f(V) - (\langle \phi, W\rangle - \langle \phi, V\rangle)}{\|W-V\|_2} = 0, \tag{4.26}$$

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

In [75], the authors consider Gâteuax and Fréchet derivatives of functions on graphons with respect to the cut metric. However, as they remark [75, Remark 2.18, page 195], such a notion of Fréchet derivative is too weak to cover natural functions such as homomorphism densities.

The next lemma 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 $D_{\mathcal{W}}f$ to obtain $D_{\widehat{\mathcal{W}}}F\colon$ eff-Dom$(F) \to \widehat{L}^\infty([0,1]^2)$ as $D_{\widehat{\mathcal{W}}}F([V]) \coloneqq [D_{\mathcal{W}}f(V)]$ for $V \in \mathcal{W}$.

Lemma 4.2.7. *Let $f\colon \mathcal{W} \to \mathbb{R}\cup\{\infty\}$ be an invariant function. Let $V, V' \in$ eff-Dom(f) such that $V' = V^\varphi$ for some $\varphi \in \mathcal{T}$. Suppose that the Fréchet-like derivatives $D_{\mathcal{W}}f(V)$ and $D_{\mathcal{W}}f(V')$ exist. If $\phi = D_{\mathcal{W}}f(V)$ and $\phi' = D_{\mathcal{W}}f(V')$, then $\phi' = \phi^\varphi$ a.e. In particular, this implies that $D_{\mathcal{W}}f(V) \in L^\infty\big([0,1]^{(2)}\big)$ 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

$$\lim_{n\to\infty} \frac{\langle \phi' - \phi^\varphi, V_n^\varphi - V'\rangle}{\|V_n - V\|_2} = 0\,. \tag{4.27}$$

To this end, recall that f is invariant and hence $f(V) = f(V^\varphi)$ and $f(V_n^\varphi) = f(V_n)$. Therefore, we have

$$\begin{aligned}
&\lim_{n\to\infty} \frac{\langle \phi' - \phi^\varphi, V_n^\varphi\rangle - \langle \phi' - \phi^\varphi, V^\varphi\rangle}{\|V_n - V\|_2} \\
&= \lim_{n\to\infty} \left[\frac{f(V_n) - f(V) - \langle \phi, V_n - V\rangle}{\|V_n - V\|_2} - \frac{f(V_n^\varphi) - f(V^\varphi) - \langle \phi', V_n^\varphi - V^\varphi\rangle}{\|V_n^\varphi - V^\varphi\|_2}\right] \\
&= \lim_{n\to\infty} \frac{f(V_n) - f(V) - \langle \phi, V_n - V\rangle}{\|V_n - V\|_2} - \lim_{n\to\infty} \frac{f(V_n^\varphi) - f(V^\varphi) - \langle \phi', V_n^\varphi - V^\varphi\rangle}{\|V_n^\varphi - V^\varphi\|_2} = 0,
\end{aligned}$$

where the last equality holds because each limit individually goes to 0 by the definition of Fréchet differentiability and our assumption that f 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 (4.27) we conclude that

$$0 = \left\langle \phi' - \phi^\varphi, \chi_{B^\varphi}\right\rangle = \int_B \big(\phi' - \phi^\varphi\big)(x, y)\,\mathrm{d}x\,\mathrm{d}y > 0,$$

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 f at V. Then, (taking $\varphi = \mathrm{id}$) we must have that $\phi = \phi'$ a.e. Hence, $D_{\mathcal{W}} f(V)$ is a unique element in $L^\infty\big([0, 1]^{(2)}\big)$. □

Let $F\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ and let $f\colon \mathcal{W} \to \mathbb{R} \cup \{\infty\}$ be the invariant extension of F. Lemma 4.2.7 justifies saying F has Fréchet-like derivative if f has a Fréchet-like derivative. Note that the Lemma 4.2.7 says that not only can the Fréchet-like derivative be thought of as a graphon, but also the two graphons $[D_{\mathcal{W}} f(V)]$ 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 4.2.8 (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 $(W_1, W_2, \dots, W_r) \in \mathcal{W}^r$ as*

$$\bigodot_{i=1}^{r} [W_i] \coloneqq \left\{ (W_i^{\varphi})_{i=1}^{r} \,\middle|\, \varphi \in \mathcal{T} \right\}. \tag{4.28}$$

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 4.2.8 to tuples with elements in $L^{\infty}\big([0,1]^{(2)}\big) \supset \mathcal{W}$. That is, we can consider sets of the form

$$\bigodot_{i=1}^{r} [V_i] \coloneqq \left\{ (V_i^{\varphi})_{i=1}^{r} \,\middle|\, \varphi \in \mathcal{T} \right\}, \tag{4.29}$$

with initial labeling $\big(V_i \in L^{\infty}\big([0,1]^{(2)}\big)\big)_{i=1}^{r}$. Therefore, from Lemma 4.2.7, if $V \in [V] \in \widehat{\mathcal{W}}$, and $\phi = D_{\mathcal{W}} f(V)$, 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

$$G_V \coloneqq \{|V| < 1\} \cup \{V = 1, \phi > 0\} \cup \{V = -1, \phi < 0\}. \tag{4.30}$$

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 F in terms of its Fréchet-like derivative.

Lemma 4.2.9. *Let $F \colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ be a function and $f \colon \mathcal{W} \to \mathbb{R} \cup \{\infty\}$ its invariant extension. Assume that for each $[V] \in \widehat{\mathcal{W}}$ the Fréchet-like derivative $D_{\mathcal{W}} f(V)$ exists for all $V \in [V]$, then the local slope (Definition 2.1.9) of F at $[V]$ satisfies*

$$|\partial F|([V]) = \eta_F([V]) \coloneqq \sup_{W \in \mathcal{W}} \frac{(\langle \phi, V \rangle - \langle \phi, W \rangle)^+}{\|V - W\|_2} = \|\phi \mathbb{1}_{G_V}\|_2, \tag{4.31}$$

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

Proof. Fixing $[V] \in \text{eff-Dom}(F)$, we verify using Lemma 4.2.7 that η_F 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 $\phi_1 = D_{\mathcal{W}} f(V_1)$ then $D_{\mathcal{W}} f(V_2) = \phi_1^{\varphi} =: \phi_2$,

and

$$\begin{aligned}\sup_{W\in\mathcal{W}}\frac{(\langle\phi_1,V_1-W\rangle)^+}{\|V_1-W\|_2}&=\sup_{W\in\mathcal{W}}\frac{(\langle\phi_1^\varphi,V_1^\varphi\rangle-\langle\phi_1^\varphi,W^\varphi\rangle)^+}{\|V_1^\varphi-W^\varphi\|_2}\\&=\sup_{W\in\mathcal{W}}\frac{(\langle\phi_2,V_2\rangle-\langle\phi_2,W\rangle)^+}{\|W-V_2\|_2}.\end{aligned}\tag{4.32}$$

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 2.1.3, there exists $\varphi\in\mathcal{I}$ such that $\delta_2([V],[W])<\|W^\varphi-V\|_2\leq\delta_2([V],[W])+\delta_\varepsilon/2$, i.e.,

$$\delta_\varepsilon/2>\delta_2([V],[W])\geq\|W^\varphi-V\|_2-\delta_\varepsilon/2>0.\tag{4.33}$$

From assumption if we choose $W^\varphi\in\mathcal{W}$, since $\|W^\varphi-V\|_2<\delta_\varepsilon$ we get

$$-\varepsilon\leq\frac{f(W^\varphi)-f(V)-(\langle\phi,W^\varphi\rangle-\langle\phi,V\rangle)}{\|W^\varphi-V\|_2}\leq\varepsilon,\tag{4.34}$$

where $\phi=D_{\mathcal{W}}(V)$. Using equations (4.34) and equation (4.33), we get

$$\begin{aligned}\frac{(F([V])-F([W]))^+}{\delta_2([V],[W])}&\leq\frac{(F([V])-F([W]))^+}{\|W^\varphi-V\|_2-\delta_\varepsilon/2}\\&\leq\frac{(\langle\phi,V\rangle-\langle\phi,W^\varphi\rangle+\varepsilon\|W^\varphi-V\|_2)^+}{\|W^\varphi-V\|_2-\delta_\varepsilon/2}\\&\leq\left(\frac{(\langle\phi,V\rangle-\langle\phi,W^\varphi\rangle)^+}{\|W^\varphi-V\|_2}+\varepsilon\right)\frac{\|W^\varphi-V\|_2}{\|W^\varphi-V\|_2-\delta_\varepsilon/2}\\&\leq(\eta_F([V])+\varepsilon)\frac{\|W^\varphi-V\|_2}{\|W^\varphi-V\|_2-\delta_\varepsilon/2},\end{aligned}\tag{4.35}$$

for some $V\in[V]$. Taking $\varepsilon\to 0$ in equation (4.35) we get

$$|\partial F|([V])\leq\eta_F([V]).\tag{4.36}$$

2. When $\eta_F([V])>0$, for all $\varepsilon\in(0,\eta_F([V]))$, by the definition of $\eta_F([V])$, for any $V\in[V]$ and $\phi=D_{\mathcal{W}}(V)$, there exists $W\in\mathcal{W}$ such that

$$0<\varepsilon<\eta_F([V])\leq\frac{\langle\phi,V\rangle-\langle\phi,W\rangle}{\|V-W\|_2}+\varepsilon.\tag{4.37}$$

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

$$\begin{aligned}
&\lim_{t\to 0} \frac{f(W_t) - f(V) - (\langle \phi, W_t\rangle - \langle \phi, V\rangle)}{\|W_t - V\|_2} = 0 \\
\implies &\lim_{t\to 0} \frac{f(W_t) - f(V) - t(\langle \phi, W\rangle - \langle \phi, V\rangle)}{t\|W - V\|_2} = 0 \\
\implies &\lim_{t\to 0} \frac{f(V) - f(W_t)}{t\|W - V\|_2} = \frac{\langle \phi, V\rangle - \langle \phi, W\rangle}{\|V - W\|_2} \geq \eta_F([V]) - \varepsilon > 0 \\
\implies &\lim_{t\to 0} \frac{f(V) - f(W_t)}{\|W_t - V\|_2} = \lim_{t\to 0} \frac{(f(V) - f(W_t))^+}{t\|W - V\|_2} \geq \eta_F([V]) - \varepsilon \\
\implies &\lim_{t\to 0} \frac{(F([V]) - F([W_t]))^+}{\delta_2([W_t],[V])} \geq \eta_F([V]) - \varepsilon. \qquad (4.38)
\end{aligned}$$

Therefore, the curve $([W_t])_{t\in[0,1]} \to [V]$ along which equation (4.38) holds for every $\varepsilon > 0$. When $\eta_F([V]) = 0$, equation (4.38) trivially holds for $\varepsilon = 0$.

Combining the two parts, we find that $|\partial F|([V]) = \eta_F([V])$.

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 $W_n \coloneqq V - t_n \phi \mathbb{1}_{A_{\delta_n}}$ and

$$\frac{(\langle \phi, V\rangle - \langle \phi, W_n\rangle)^+}{\|V - W_n\|_2} = \left\|\phi \mathbb{1}_{A_{\delta_n}}\right\|_2. \qquad (4.39)$$

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 (4.39) that

$$\eta_F([V]) \geq \limsup_{n\to\infty} \left\|\phi \mathbb{1}_{A_{\delta_n}}\right\|_2 = \|\phi \mathbb{1}_{G_V}\|_2, \qquad (4.40)$$

where the last equality follows from the dominated convergence theorem and the fact that $\mathbb{1}_{A_{\delta_n}} \to \mathbb{1}_{G_V}$ 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

$$\begin{aligned}\langle \phi, V - W\rangle &= \int_{G_V} \phi(V - W)\,\mathrm{d}\lambda_{[0,1]^2} + \int_{[0,1]^2 \setminus G_V} \phi(V - W)\,\mathrm{d}\lambda_{[0,1]^2} \\ &= \int_{G_V} \phi(V - W_0)\,\mathrm{d}\lambda_{[0,1]^2} + \int_{[0,1]^2 \setminus G_V} \phi(V - W)\,\mathrm{d}\lambda_{[0,1]^2} \\ &\leq \int_{G_V} \phi(V - W_0)\,\mathrm{d}\lambda_{[0,1]^2} = \int \phi(V - W_0)\,\mathrm{d}\lambda_{[0,1]^2} \\ &= \langle \phi, V - W_0\rangle, \end{aligned} \tag{4.41}$$

where the inequality above follows from the fact that $\phi(V - W) \leq 0$ on $[0,1]^2 \setminus G_V$. Using that $\|V - W_0\|_2 \leq \|V - W\|_2$, we obtain

$$\frac{(\langle \phi, V\rangle - \langle \phi, W\rangle)^+}{\|V - W\|_2} \leq \frac{(\langle \phi, V\rangle - \langle \phi, W_0\rangle)^+}{\|V - W_0\|_2}.$$

It therefore follows that

$$\eta_F([V]) \coloneqq \sup_W \frac{(\langle \phi, V\rangle - \langle \phi, W\rangle)^+}{\|V - W\|_2}, \tag{4.42}$$

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 $\langle \phi, V - W\rangle \leq \|\phi \mathbb{1}_{G_V}\|_2 \|V - W\|_2$. Therefore, it follows from that

$$\eta_F([V]) \leq \|\phi \mathbb{1}_{G_V}\|_2. \tag{4.43}$$

Combining equations (4.40) and (4.43), the conclusion follows. □

Remark 4.2.10. *We can define a similar expression for the set valued function G when* eff-Dom(f) *is a cubic domain. As an example, when* eff-Dom$(f) = \{W \in \mathcal{W} \mid a \leq W \leq b \text{ a.e.}\}$ *for some $-1 \leq a \leq b \leq 1$ (see Section 4.5.1 for a discussed example), we can define $G_V \subseteq [0,1]^{(2)}$ for any $V \in$* eff-Dom(f) *as*

$$G_V = \{a < V < b\} \cup \{V = b, \phi > 0\} \cup \{V = a, \phi < 0\}, \tag{4.44}$$

for $(V, \phi \coloneqq D_{\mathcal{W}} f(V)) \in [V] \odot [\phi]$. Lemma 4.2.9 continues to hold when $V \in$ eff-Dom$(f) \subset \mathcal{W}$ *whenever* eff-Dom(f) *is a cubic domain. In this case, the set valued function G is defined as described above and the proof of Lemma 4.2.9 can be modified accordingly.*

Remark 4.2.11. *Lemma 4.2.9 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 4.2.9 says that if ω is a gradient flow of F, then its local slope is given by the L^2-norm of its Fréchet-like derivative, i.e., $|\partial F|(\omega_t) = \left\|\phi(\omega_t)\mathbb{1}_{G_{\omega_t}}\right\|_2 = \left\|D_{\widehat{\mathcal{W}}}F(\omega_t)\mathbb{1}_{G_{\omega_t}}\right\|_2$ for all $t > 0$. Here for any $t > 0$,*

$$D_{\widehat{\mathcal{W}}}F(\omega_t)\mathbb{1}_{G_{\omega_t}} \coloneqq \left\{ \left(D_{\mathcal{W}} f(U_t)\mathbb{1}_{G_{U_t}} \right)^{\varphi} \in L^{\infty}\left([0,1]^{(2)}\right) \,\middle|\, \varphi \in \mathcal{T} \right\},$$

for $U_t \in \omega_t$. Since the L^2-norm is invariant under measure preserving transformations [117, 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 $W_t' = -D_{\mathcal{W}} f(W_t)\mathbb{1}_{G_{W_t}}$ 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 4.2.13 which in turn shows the existence of a gradient flow under suitable assumption (See Theorem 4.2.14).

Lemma 4.2.12. *Let $F\colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ be a function and $f\colon \mathcal{W} \to \mathbb{R} \cup \{\infty\}$ be its invariant extension. Let F 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 $W_t' = -\eta_F([W_t])N_t$ for a.e. $t \in [0,1]$ for some $N_t \in L^{\infty}\left([0,1]^{(2)}\right)$ satisfying $\|N_t\|_2 = 1$ and $\langle \phi_t, N_t\rangle = \eta_F([W_t])$. 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

$$\lim_{h\to 0} \frac{\|W_{t+h} - W_t\|_2}{|h|} = \left\|W_t'\right\|_2 = \|\eta_F([W_t])N_t\|_2 = |\eta_F([W_t])|. \tag{4.45}$$

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 4.2.6,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} f(W_t) &= \lim_{h\to 0} \frac{f(W_{t+h}) - f(W_t)}{h} \\
&= \lim_{h\to 0} \frac{\langle \phi_t, W_{t+h}\rangle - \langle \phi_t, W_t\rangle + o(\|W_{t+h} - W_t\|_2)}{h} \\
&= \langle \phi_t, W_t'\rangle + 0 = -\langle \phi_t, N_t\rangle \eta_F([W_t]) = -\eta_F^2([W_t]),
\end{aligned} \tag{4.46}$$

where $\phi_t = D_{\mathcal{W}} f(W_t)$. Thus, $(W_t)_{t\in[0,1]}$ satisfies Definition 2.1.10 and is a curve of maximal slope on $(\mathcal{W}, d_2)$, and ω is a curve of maximal slope on $(\widehat{\mathcal{W}}, \delta_2)$. □

4.2.4 *Existence of gradient flow*

We now prove the existence of a curve of maximal slope if F 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 $(\mathcal{W}, d_2)$.

Lemma 4.2.13. *Let $f\colon \mathcal{W} \to \mathbb{R}\cup\{\infty\}$ be a λ-semiconvex invariant function for some $\lambda \in \mathbb{R}$ such that the Fréchet-like derivative, $\phi(W) = D_{\mathcal{W}} f(W)$, exists for all $W \in$ eff-Dom(f). Let $(W_t)_{t\in[0,1]} \in \mathrm{AC}(\mathcal{W}, d_2)$ be an absolutely continuous curve satisfying $W_t' = -\phi(W_t)\mathbb{1}_{G_{W_t}} = -D_{\mathcal{W}} f(W_t)\mathbb{1}_{G_{W_t}}$ 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)*

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} d_2^2(W_t, V) + \frac{\lambda}{2}\|W_t - V\|_2^2 + f(W_t) \le f(V), \tag{4.47}$$

for every $V \in$ eff-Dom(f).

Proof. The curve $(W_t)_{t\in[0,1]}$ is a curve of maximal slope follows from Lemma 4.2.12. We now show that it satisfies the EVI. For $t \in \mathbb{R}_+$, let $\phi_t := D_{\mathcal{W}} f(W_t)$. Fix $U \in \mathcal{W}$ and define the function $g_U\colon \mathcal{W} \to \mathbb{R}\cup\{\infty\}$ by $g_U(V) := f(V) - \lambda\|U - V\|_2^2/2$, for $V \in$ eff-Dom(f). We first observe that $D_{\mathcal{W}} g_{W_t}(W_t) = \phi_t$. To see this, note that

$$\begin{aligned} &\lim_{s\to t} \frac{f(W_s) - f(W_t) - \langle \phi_t, W_s - W_t\rangle}{\|W_s - W_t\|_2} = 0 \\ \implies &\lim_{s\to t} \frac{g_{W_t}(W_s) - g_{W_t}(W_t) - \langle \phi_t, W_s - W_t\rangle}{\|W_s - W_t\|_2} = 0. \end{aligned} \tag{4.48}$$

The conclusion, that is $D_{\mathcal{W}} g_{W_t}(W_t) = \phi_t$, now follows from the uniqueness of Fréchet-like derivatives (Lemma 4.2.7). Since f is λ-semiconvex, g_U is convex, i.e.,

$$g_{W_t}(V) \ge g_{W_t}(W_t) + \langle \phi_t, V - W_t\rangle\ , \quad V \in \text{eff-Dom}(f). \tag{4.49}$$

From equation (4.41) and using the fact that $W_t' = \phi_t$, we obtain

$$\begin{aligned} \langle \phi_t, V - W_t\rangle \le \left\langle \phi_t \mathbb{1}_{G_{W_t}}, V - W_t \right\rangle &= \left\langle -\frac{\mathrm{d}}{\mathrm{d}t} W_t, V - W_t \right\rangle \\ &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|W_t - V\|_2^2 = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} d_2^2(W_t, V), \end{aligned} \tag{4.50}$$

where the second equality follows from the reflexivity of $L^2([0,1]^2)$. Plugging equations (4.48) and (4.50) in equation (4.49) and rearranging, we get

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}d_2^2(W_t, V) + \frac{\lambda}{2}\|W_t - V\|_2^2 + f(W_t) \leq f(V), \tag{4.51}$$

for all $V \in \text{eff-Dom}(f)$. Using equation (4.51) it follows from [5, Theorem 4.0.4] that the curve $([W_t])_{t\in[0,1]}$ is the unique curve in $\mathrm{MM}_{\delta_2,\delta_\square}(\Phi_F, [W_0])$. □

Theorem 4.2.14 (Existence of curve of maximal slope-II)**.** *Let $F\colon \widehat{\mathcal{W}} \to \mathbb{R}\cup\{\infty\}$ be a real valued function such that the Fréchet-like derivative $D_{\widehat{\mathcal{W}}}F([W])$ exists for all $[W] \in \text{eff-Dom}(F)$. Let $f\colon \mathcal{W} \to \mathbb{R}\cup\{\infty\}$ be its invariant extension. For $W_0 \in [W_0] \in \text{eff-Dom}(F)$ and $t \geq 0$ define*

$$W_t \coloneqq W_0 - \int_0^t \phi(W_s)\mathbb{1}_{G_{W_s}}\,\mathrm{d}s, \qquad t \in \mathbb{R}_+,$$

where the above integral is pointwise. If f is λ-semiconvex w.r.t. d_2, then the curve $t \mapsto \omega_t = [W_t]$ is a curve maximal slope for F starting at $[W_0] \in \text{eff-Dom}(F)$.

Proof. Fix $[W_0] \in \text{eff-Dom}(F)$ and define

$$W_t \coloneqq W_0 - \int_0^t \phi(W_s)\mathbb{1}_{G_{W_s}}\,\mathrm{d}s\,, \quad t \in (0,1],$$

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 $W_t' = -\phi(W_t)\mathbb{1}_{G_{W_t}}$ for all $t \in [0,1]$. It follows from Lemma 4.2.13 that $(W_t)_{t\in[0,1]}$ is a minimizing movement. It follows from the definition of minimizing movements (see Section 4.2.1 and [5, 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 4.2.9, we know that for any $t \in [0,1]$, $\eta_F([W_t]) = \left\|\phi(W_t)\mathbb{1}_{G_{W_t}}\right\|_2$ and therefore we have $W_t' = -\eta_f(W_t)N_t$ where $N_t \coloneqq \phi(W_t)\mathbb{1}_{G_{W_t}}/\left\|\phi(W_t)\mathbb{1}_{G_{W_t}}\right\|_2$. It follows from Lemma 4.2.12 that ω is a curve of maximal slope on $(\widehat{\mathcal{W}}, \delta_2)$. □

Remark 4.2.15. *An important consequence of the above Theorem is that if ω is a gradient flow of F then there exists an absolutely continuous curve $(W_t)_{t\in[0,T]} \in \mathrm{AC}(\mathcal{W}, d_2)$ such that $[W_t] = \omega_t$, $|\omega'|(t) = \left\| D_{\mathcal{W}} f(W_t) \mathbb{1}_{G_{W_t}} \right\|_2$ and $W_t' = -D_{\mathcal{W}} f(W_t) \mathbb{1}_{G_{W_t}}$, for each $t \in (0,T]$.*

Remark 4.2.16. *If F 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 F. When $\lambda > 0$, let ω^* be the unique minimizer of F. Then following [5, Remark 4.0.5, part (d)], a gradient flow ω of F 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 [5, Corollary 4.0.6], ω satisfies

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

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

4.2.5 Finite dimensional Fréchet-like derivatives and upper gradients

Recall the partition $Q_n \coloneqq \{Q_{n,i}\}_{i\in[n]}$ defined for any $n \in \mathbb{N}$ in Section 2.1.1. Given an invariant function $f\colon \mathcal{W} \to \mathbb{R} \cup \{\infty\}$, we can restrict its domain to kernels in $\mathcal{W}_n$ and still consider the Fréchet-like derivative $D_{\mathcal{W}_n} f\colon \mathcal{W}_n \cap \text{eff-Dom}(f) \to L_n^\infty\big([0,1]^{(2)}\big)$.

There are two equivalent ways of doing this. First, suppose the Fréchet-like derivative of f at V is given by $D_{\mathcal{W}} f(V) = \phi$. Then define $D_{\mathcal{W}_n} f(V) = \phi_n$ by conditional expectations as $\phi_n \coloneqq \mathbb{E}[\phi \,|\, \mathcal{F}_n]$, 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 [39, Section 3.3] and [150, Section 9.2.1] respectively. Since $\phi = D_{\mathcal{W}} f(V)$, by the *Tower Property* of conditional expectations we obtain that when $W, V \in \mathcal{W}_n$,

$$\begin{aligned} &\langle \phi_n, W\rangle - \langle \phi_n, V\rangle = \langle \phi, W\rangle - \langle \phi, V\rangle, \\ \implies &\lim_{\substack{W\in\mathcal{W}_n, \\ \|W-V\|_2\to 0}} \frac{f(W) - f(V) - (\langle \phi_n, W\rangle - \langle \phi_n, V\rangle)}{\|W - V\|_2} = 0. \end{aligned} \tag{4.52}$$

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 4.3.1 below.

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

$$\eta_{F,n}([V]) \coloneqq \sup_{W \in \mathcal{W}_n} \frac{(\langle \phi_n, V \rangle - \langle \phi_n, W \rangle)^+}{\|V - W\|_2}. \tag{4.53}$$

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

$$|\partial_n F|([V]) \coloneqq \limsup_{[W] \in \widehat{\mathcal{W}}_n,\, \delta_2([W],[V]) \to 0} \frac{(F([V]) - F([W]))^+}{\delta_2([W],[V])}, \tag{4.54}$$

for $[V] \in \widehat{\mathcal{W}}_n \cap \text{eff-Dom}(F)$. Then, by a similar argument as shown in the proof of Lemma 4.2.9, we have the following corollary.

Corollary 4.2.17. *Let $F \colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ be a function and $f \colon \mathcal{W} \to \mathbb{R} \cup \{\infty\}$ be its invariant extension. Assume that for $[V] \in \widehat{\mathcal{W}}_n \cap \text{eff-Dom}(F)$ the Fréchet-like derivative $D_{\mathcal{W}} f(V)$ exists for all $V \in [V]$. Then the local slope (Definition 2.1.9) of F at $[V]$ satisfies $|\partial_n F|([V]) = \eta_{F,n}([V])$.*

4.3 Convergence of finite dimensional gradient flows

In Section 4.1 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.

To set the stage, let $f \colon \mathcal{W} \to \mathbb{R} \cup \{\infty\}$ be an invariant function. Consider its restriction to $\mathcal{W}_n$, viewed as the space of $n \times n$ symmetric matrices: $f_n \coloneqq f \circ K$. This is a function on a closed convex subset of an Euclidean space. Suppose the function f_n is C^1 up to the boundary then we show that the Euclidean gradient and the Fréchet-like derivative are equal up to a scaling.

Lemma 4.3.1. *Let $n \in \mathbb{N}$. Let $f \colon \mathcal{W} \to \mathbb{R} \cup \{+\infty\}$ be an invariant function that is Fréchet differentiable according to Definition 4.2.6, that is, $D_{\mathcal{W}_k} f(V)$ exists for every $V \in \mathcal{W}_n$. If*

$f_n \coloneqq f \circ K$ is differentiable up to the boundary of $\mathcal{M}_n$, then

$$n^2(\nabla f_n \circ M_n)(V) = (M_n \circ D_{\mathcal{W}_n} f)(V), \qquad V \in \mathcal{W}_n,$$

where $(\nabla f_n)_{i,j} \coloneqq \partial_{i,j} f_n$ for all $(i,j) \in [n]^{(2)}$, and M_n is as defined in Definition 2.1.6.

Proof. Since f_n is assumed to be differentiable on a finite dimensional Euclidean space, ∇f_n is its Fréchet derivative as well. By composing with M_n, which scales distances by a factor, we get that $n^2(\nabla f_n \circ M_n)$ is a Fréchet-like derivative on $\mathcal{W}_n$. We have already shown in equation (4.52) that $D_{\mathcal{W}_n} f$ is also a Fréchet-like derivative on $\mathcal{W}_n$. We are done by arguing that Fréchet-like derivatives are unique by following an argument very similar to that of Lemma 4.2.7. □

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

$$\frac{\mathrm{d}}{\mathrm{d}t} M_n(W_n(t)) = -(\nabla f_n \circ M_n)(W_n(t)), \tag{4.55}$$

with initial condition $W_n(0) = W_{n,0} \in \mathcal{W}_n \cap \text{eff-Dom}(f)$, 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 4.3.1, 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 4.1.1. For completeness, we reproduce the theorem statement below.

Theorem 4.3.2 (Convergence of Gradient Flows)**.** *Suppose $F \colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ satisfies the following conditions:*

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

2. *F is λ-semiconvex (Definition 2.1.15) along generalized geodesics on $(\widehat{\mathcal{W}}, \delta_2)$ (Definition 2.2.6), 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 F 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 $|\partial F|([U_0]) < \infty$ and $\limsup_{n\to\infty} |\partial F|([U_{n,0}]) \le G < \infty$, for some $G \ge 0$. Then,

$$\limsup_{n\to\infty} \sup_{t\in[0,T]} \delta_\square\left(\omega_t^{(n)}, \omega_t\right) = 0, \tag{4.56}$$

for any $T \in \mathbb{R}_+$, where $\omega = (\omega_t)_{t\in\mathbb{R}_+}$ is the unique minimizing movement curve [5, Definition 2.0.6, page 42] on $\widehat{\mathcal{W}}$ for the function F starting at $\omega_0 = [U_0]$ [5, Theorem 4.0.4]. In addition, if the conditions for the existence of curves of maximal slope (Theorem 4.2.4 or Theorem 4.2.14) hold, then ω is also a curve of maximal slope.

Proof. By increasing the constant G suitably, we may assume that

$$\max\left\{\sup_{n\ge2} |\partial F|([U_{n,0}]), |\partial F|([U_0])\right\} \le G < \infty. \tag{4.57}$$

Fix $T > 0$ and let $\boldsymbol{\tau}_m$ be a sequence of positive time steps such that $|\boldsymbol{\tau}_m| = T/m$. Since $F\colon \widehat{\mathcal{W}} \to \mathbb{R}\cup\{\infty\}$ 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 $\Phi_F(\tau, [U]; \cdot)$ is $\delta_\square$-lower semicontinuous. From Lemma 2.2.7, it follows that Φ_F satisfies [5, Assumption 4.0.1] and hence by [5, Proposition 4.0.4] we have that

$$\omega^{(n)}(t) = \delta_\square\text{-}\lim_{m\to\infty} \left(J_{t/m}^{(n)}\right)^m([U_{n,0}]), \qquad \omega(t) = \delta_\square\text{-}\lim_{m\to\infty} \left(J_{t/m}\right)^m([U_0]),$$

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 (Definition 4.2.1) 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 4.2.5, 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 $\delta_\square\left(\overline{[U_{n,\boldsymbol{\tau}_m}]}(t), \overline{[U_{\boldsymbol{\tau}_m}]}(t)\right) \to 0$ uniformly over $t \in [0,T]$ as $n \to \infty$. For every $t \in [0,T]$ we

have

$$\delta_2\Big(\overline{[U_{n,\boldsymbol{\tau}_m}]}(t), \omega^{(n)}(t)\Big) < \gamma(|\boldsymbol{\tau}_m|, \lambda, t; T, |\partial F_n|([U_{n,0}])), \tag{4.58}$$

$$\delta_2\Big(\overline{[U_{\boldsymbol{\tau}_m}]}(t), \omega(t)\Big) < \gamma(|\boldsymbol{\tau}_m|, \lambda, t; T, |\partial F|([U_0])), \tag{4.59}$$

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

$$\gamma(\tau, \lambda, t; T, G) \coloneqq \begin{cases} \frac{\tau G}{\sqrt{2}}, & \text{if } \lambda = 0, \\ \frac{1+2|\lambda|T}{1+\lambda\tau} \cdot \frac{\tau G}{\sqrt{2}} \cdot \exp\big(-\ln\big(\frac{1+\lambda\tau}{\tau}\big)t\big), & \text{if } \lambda < 0, \\ \sqrt{1+2\lambda T} \cdot \frac{\tau G}{\sqrt{2}} \cdot \exp\big(-\ln\big(\frac{1+\lambda\tau}{\tau}\big)t\big), & \text{if } \lambda > 0, \end{cases} \tag{4.60}$$

by [5, Equation 4.0.6, Theorem 4.0.9, Theorem 4.0.10], and the uniform bound in equation (4.57). Note that γ is independent of n. Using the triangle inequality, we get

$$\begin{aligned} \delta_\square\Big(\omega_t^{(n)}, \omega_t\Big) &\leq \delta_\square\Big(\omega_t^{(n)}, \overline{[U_{n,\boldsymbol{\tau}_m}]}(t)\Big) + \delta_\square\Big(\overline{[U_{n,\boldsymbol{\tau}_m}]}(t), \overline{[U_{\boldsymbol{\tau}_m}]}(t)\Big) \\ &\quad + \delta_\square\Big(\overline{[U_{\boldsymbol{\tau}_m}]}(t), \omega_t\Big) \\ &\leq \delta_2\Big(\omega_t^{(n)}, \overline{[U_{n,\boldsymbol{\tau}_m}]}(t)\Big) + \delta_\square\Big(\overline{[U_{n,\boldsymbol{\tau}_m}]}(t), \overline{[U_{\boldsymbol{\tau}_m}]}(t)\Big) \\ &\quad + \delta_2\Big(\overline{[U_{\boldsymbol{\tau}_m}]}(t), \omega_t\Big) \\ &\leq 2\gamma(|\boldsymbol{\tau}_m|, \lambda, t; T, G) + \delta_\square\Big(\overline{[U_{n,\boldsymbol{\tau}_m}]}(t), \overline{[U_{\boldsymbol{\tau}_m}]}(t)\Big), \end{aligned} \tag{4.61}$$

for all $n \in \mathbb{N}$ and $t \in [0, T]$ by equations (4.58) and (4.59).

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 (4.61) that $\delta_\square\Big(\omega_t^{(n)}, \omega_t\Big) \to 0$ uniformly on $t \in [0, T]$ as $n \to \infty$. □

Remark 4.3.3. *The proof of the Theorem 4.1.1 can be carried as long as we have the uniform estimates in equation (4.58). In particular, if $\nabla f_n \circ M_n$ are uniformly Lipschitz, and there is a constant $m \in \mathbb{R}_+$ such that*

$$f_n(B) \leq f_n(A) + \langle \nabla f_n(A), B - A \rangle + \frac{m}{2}\|B - A\|_{\mathrm{F}}^2,$$

for all $A, B \in \mathcal{M}_n$, where $f_n \coloneqq (f \circ K)\,|_{\mathcal{M}_n}$, then [45, Theorem 212A] guarantees a uniform estimate in (4.58) and therefore the conclusion of the theorem remains valid.

4.4 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 [192, 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

$$\begin{aligned} X_n((u_\ell)_{\ell=1}^n) &:= (W(u_i,u_j))_{(i,j)\in[n]^{(2)}}, \\ v_n(z)(i,j) &= \mathbb{E}[v(U_i,U_j) \mid X_n((U_\ell)_{\ell=1}^n) = z], \end{aligned} \tag{4.62}$$

where $\{U_i\}_{i\in\mathbb{N}}$ are i.i.d. as Uni$[0,1]$. Intuitively, formula (4.62) 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 $\mathcal{F} = (\mathcal{F}_n := \sigma(X_n))_{n\in\mathbb{N}}$. 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 a lemma below.

Lemma 4.4.1. *For every $i,j \in \mathbb{N}$, the process $(v_k(X_n)(i,j))_{k=\max\{i,j\}}^{\infty}$ is a martingale with respect to the filtration $\mathcal{F}$.*

Lemma 4.4.2. *For any $\varphi \in \mathcal{T}$, we have $v_n^{\varphi}(X_k^{\varphi}) = v_n(X_n)$, for all $n \in \mathbb{N}$, where $v^{\varphi}(x,y) := v(\varphi(x),\varphi(y))$ for all $(x,y) \in [0,1]^{(2)}$.*

Proof. For any $\varphi \in \mathcal{T}$, given $\{U_i\}_{i\in\mathbb{N}}$, let us define $V_i := \varphi(U_i)$ 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 $(i,j) \in [n]^{(2)}$,

$$\begin{aligned} v_n^{\varphi}(X_n^{\varphi})(i,j) &= \mathbb{E}[v^{\varphi}(U_i,U_j) \mid X_n^{\varphi}((U_\ell)_{\ell=1}^n)] \\ &= \mathbb{E}[v(\varphi(U_i),\varphi(U_j)) \mid X_n(\varphi(U_\ell)_{\ell=1}^n)] \\ &= \mathbb{E}[v(V_i,V_j) \mid X_n((V_\ell)_{\ell=1}^n)] = v_n(X_n)(i,j), \end{aligned} \tag{4.63}$$

holds, completing the proof. □

Suppose we are given some $\omega = (\omega_t)_{t\in[0,1]} \in \mathrm{AC}(\widehat{\mathcal{W}}, \delta_2)$. From Lemma 2.2.1, 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 [111, page 30, Theorem 5] that there exists $v_t := W_t' \in L^2([0,1]^{(2)})$, 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 (4.62) with W_t replacing the role of W and v_t replacing v. The following Proposition 4.4.3 shows that the $\rho_{n,t} = \mathrm{Law}(X_{knt}((U_i)_{i=1}^n))$ satisfies continuity equation with the velocity $v^{n,t}$ defined as

$$v^{n,t}(z) := \mathbb{E}\Big[\big(W_t'(U_i,U_j)\big)_{(i,j)\in[n]^{(2)}} \,\Big|\, \big(W_t(U_i,U_j)\big)_{(i,j)\in[n]^{(2)}} = z\Big], \qquad z \in \mathcal{M}_n.$$

Proposition 4.4.3. *Let $n \in \mathbb{N}$, and $\rho^{n,t} = \mathrm{Law}\big(X^{n,t}((U_i)_{i=1}^n)\big)$. Then, for a.e. $t \in [0,1]$, the continuity equation $\partial_t \rho^{n,t} + \nabla \cdot \big(v^{n,t}\rho^{n,t}\big) = 0$ holds weakly with Dirichlet boundary conditions. That is, for any continuously differentiable test function f on $\mathcal{M}_n$, vanishing at the boundary,*

$$\partial_t\bigg(\int f(z)\,\mathrm{d}\rho^{n,t}(z)\bigg) - \int \nabla f(z) v^{n,t}(z)\,\mathrm{d}\rho^{n,t}(z) = 0, \qquad \text{a.e.} \quad t \in [0,1].$$

Proof. By definition, $\rho^{n,t}$ is the law of the random symmetric matrix $X^{n,t}((U_i)_{i=1}^n)$. Fix some $f \in C^1(\mathcal{M}_n, \mathbb{R})$ vanishing at the boundary of $\mathcal{M}_n$. By change of variables

$$\int_{\mathcal{M}_n} f(z)\rho^{n,t}(z)\,\mathrm{d}z = \int_{[0,1]^n} f\Big(\big(W_t(u_i,u_j)\big)_{(i,j)\in[n]^{(2)}}\Big)\,\mathrm{d}\lambda_{[0,1]^n}(u), \tag{4.64}$$

where $\lambda_{[0,1]^n}$ is the Lebesgue measure on $[0,1]^n$, and $u = (u_i)_{i=1}^n$. Note that $v^{n,t} \in L^2\big(\rho^{n,t}\big)$. Taking time derivative on both sides of equation (4.64) for t in the set of full measure where

W'_t is defined,

$$\begin{aligned}
&\partial_t \int_{\mathcal{M}_n} f(z)\rho^{n,t}(z)\,\mathrm{d}z \\
&= \int_{[0,1]^n} \partial_t f\Big((W_t(u_i,u_j))_{(i,j)\in[n]^{(2)}}\Big)\,\mathrm{d}\lambda_{[0,1]^n}(u) \\
&= \int_{[0,1]^n} \Big\langle \big(\nabla f \circ X^{n,t}\big)(u), \big(W'_t(u_i,u_j)\big)_{(i,j)\in[n]^{(2)}} \Big\rangle\,\mathrm{d}\lambda_{[0,1]^n}(u) \\
&= \int_{\mathcal{M}_n} \left\langle \nabla f(z), \int_{\{u\in[0,1]^n \mid X^{n,t}=z\}} \big(W'_t(u_i,u_j)\big)_{(i,j)\in[n]^{(2)}}\,\mathrm{d}\mu_z(u) \right\rangle \mathrm{d}z, \qquad (4.65)
\end{aligned}$$

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 $\int_{\mathcal{M}_n}\langle \nabla f(z), v^{n,t}(z)\rho^{n,t}(z)\rangle\,\mathrm{d}z$. This completes the proof. □

Proposition 4.4.3 shows that an absolutely continuous curve in $(\widehat{\mathcal{W}}, \delta_2)$ determines a family of continuity equations. In [102], 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.

As an example, consider the continuity equation for the scalar entropy. From equation (4.81) it follows that the velocity of gradient flow for scalar entropy function satisfies

$$v([W])(x,y) = -\log\left(\frac{W(x,y)}{1-W(x,y)}\right), \qquad (x,y)\in[0,1]^{(2)}, \qquad (4.66)$$

if $0 < W < 1$ a.e. Let $\{U_i\}_{i=1}^{\infty}$ be i.i.d. Uni$[0,1]$, then using equation (4.62) we can compute the velocities $(v_n)_{n\in\mathbb{N}}$ appearing in the continuity equation as

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

Unfortunately, for our other examples, the velocity fields do not have closed form expressions since they are averages over all possible embeddings of a labeled weighted graph in a graphon.

4.5 Examples of Gradient Flows on Graphons

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 $k \in \mathbb{N}$, sample $\{Z_i\}_{i\in[k]}$ i.i.d. from Uni$[0,1]$. Let $G_k[W] = (W(Z_i, Z_j))_{(i,j)\in[k]^{(2)}}$. Let $\rho_k([W]) = \mathrm{Law}(G_k[W])$ denote its law over $k \times k$ symmetric matrices. Consider the functions $F_k \colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ defined through the function composition $F_k = H_k \circ \rho_k$, for different choices of $H_k \colon \mathcal{P}(\mathcal{M}_k) \to \mathbb{R} \cup \{\infty\}$. Notice that the function F_k is well-defined over $\widehat{\mathcal{W}}$.

4.5.1 Linear functions

Let $H_k \colon \mathcal{P}(\mathcal{M}_k) \to \mathbb{R} \cup \{\infty\}$ be defined as a linear function such that F_k takes the form

$$F_k([W]) = \langle f_k, \rho_k([W])\rangle \coloneqq \int_{\mathcal{M}_k} f_k(z)\rho_k([W])(\mathrm{d}z), \tag{4.67}$$

for all $[W] \in \widehat{\mathcal{W}}$ such that $\mathrm{supp}(\rho_k([W])) \subseteq \text{eff-Dom}(f_k)$, where $f_k \colon \mathcal{M}_k \to \mathbb{R} \cup \{\infty\}$ satisfies $f_k \in L^1(\rho_k)$ and $\nabla f_k \in L^\infty(\rho_k)$. Let f_k be continuously differentiable on an open set containing $\mathrm{supp}(\rho_k)$. Suppose that f_k satisfies either Assumption 1 or Assumption 2 stated below.

Assumption 1. *For any $\varepsilon > 0$ there is some $\delta_\varepsilon > 0$ such that for all $X, X_0 \in \mathcal{M}_k$ satisfying $\|X - X_0\|_2 \le \delta_\varepsilon$, we have*

$$|f_k(X) - f_k(X_0) - \langle \nabla f_k(X_0), X - X_0\rangle| \le \varepsilon\|X - X_0\|_2,$$

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

Assumption 2. *There is a constant C_0 such that*

$$\sup_{X,X_0\in\mathcal{M}_k} |f_k(X) - f_k(X_0) - \langle \nabla f_k(X_0), X - X_0\rangle| \le C_0.$$

Then, F_k admits a Fréchet-like derivative as shown below. By change of variables, note that

$$\begin{aligned} F_k([W]) &= \mathbb{E}_{X_k\sim\rho_k([W])}[f_k(X_k)] \\ &= \mathbb{E}_{\{Z_i\sim\mathrm{Uni}[0,1]\}_{i=1}^k}\Big[f_k\Big((W(Z_i, Z_j))_{(i,j)\in[k]^{(2)}}\Big)\Big]. \end{aligned} \tag{4.68}$$

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 $(Z_i)_{i=1}^k$ to obtain a coupling of $\rho_k([V])$ and $\rho_k([W])$. This is used implicitly in the following derivation and, hence, we will skip referring to the random variables $\{Z_i\}_{i=1}^k$ from here on. Define $X_k \coloneqq (W(Z_i, Z_j))_{(i,j)\in[k]^{(2)}}$ and $Y_k \coloneqq (V(Z_i, Z_j))_{(i,j)\in[k]^{(2)}}$. As a consequence of this coupling,

$$\begin{aligned}\mathbb{E}\Big[\|Y_k - X_k\|_2^2\Big] &= \sum_{i=1, j\neq i}^{k} \mathbb{E}\big[(V(Z_i,Z_j) - W(Z_i,Z_j))^2\big] \\ &= k(k-1)\int_{[0,1]^2} (V(x,y) - W(x,y))^2 \,\mathrm{d}x\,\mathrm{d}y = k(k-1)\|V-W\|_2^2.\end{aligned}$$

The first equality is using the fact that the diagonal terms of $Y_k - X_k$ are all zeroes. Hence $\mathbb{E}[\|Y_k - X_k\|_2] \leq k\|V-W\|_2$, by the Jensen's inequality.

If either Assumption 1 or Assumption 2 hold, then for any $\varepsilon > 0$,

$$\begin{aligned}&|f_k(Y_k) - f_k(X_k) - \langle \nabla f_k(X_k), Y_k - X_k\rangle| \\ &\leq \varepsilon \|Y_k - X_k\|_2 \mathbb{1}\{\|Y_k - X_k\|_2 \leq \delta_\varepsilon\} + C_0 \mathbb{1}\{\|Y_k - X_k\|_2 > \delta_\varepsilon\}. \end{aligned} \tag{4.69}$$

Taking expectations on both sides,

$$\begin{aligned}&|\mathbb{E}[f_k(Y_k)] - \mathbb{E}[f_k(X_k)] - \mathbb{E}[\langle \nabla f_k(X_k), Y_k - X_k\rangle]| \\ &\leq \varepsilon \mathbb{E}[\|Y_k - X_k\|_2] + C_0 \mathbb{E}[\mathbb{1}\{\|Y_k - X_k\|_2 > \delta_\varepsilon\}] \\ &\leq \varepsilon k \|V-W\|_2 + C_0 \mathbb{P}\{\|Y_k - X_k\|_2 > \delta_\varepsilon\} \\ &\leq \varepsilon k \|V-W\|_2 + \frac{C_0}{\delta_\varepsilon^2} k^2 \|V-W\|_2^2,\end{aligned}$$

by Markov's inequality. As $\|V-W\|_2$ approaches zero, it is now clear that, for any $\varepsilon' > 0$,

$$\limsup_{V\in\mathcal{W},\, \|V-W\|_2 \to 0} \frac{|\mathbb{E}[f_k(Y_k)] - \mathbb{E}[f_k(X_k)] - \mathbb{E}[\langle \nabla f_k(X_k), Y_k - X_k\rangle]|}{\|V-W\|_2} \leq \varepsilon'. \tag{4.70}$$

Since ε' is arbitrary, the above lim sup must be zero.

By the definition of Fréchet-like derivatives (Definition 4.2.6), we want to obtain some $\phi \in L^\infty\big([0,1]^{(2)}\big)$ such that

$$\mathbb{E}[\langle \nabla f_k(X_k), Y_k - X_k\rangle] = \langle \phi, V - W\rangle. \tag{4.71}$$

Let $U_k \coloneqq Y_k - X_k$ (also similarly measurable with respect to $\{Z_i\}_{i=1}^k$), and let us denote by $A(Z) \coloneqq \nabla f_k(X_k) = \nabla f_k\Big((W(Z_i, Z_j))_{(i,j)\in[k]^{(2)}}\Big)$. Observe that

$$\begin{aligned}
\mathbb{E}[\langle \nabla f_k(X_k), Y_k - X_k\rangle] &= \sum_{i,j=1}^{k} \mathbb{E}\Big[(A(Z))_{i,j}(U_k(Z))_{i,j}\Big] \\
&= \sum_{i,j=1}^{k} \mathbb{E}\Big[\mathbb{E}\Big[(A(Z))_{i,j}(U_k(Z))_{i,j} \;\Big|\; Z_i, Z_j\Big]\Big] \\
&= \sum_{i=1,j\neq i}^{k} \mathbb{E}\Big[\mathbb{E}\Big[(A(Z))_{i,j}U(Z_i, Z_j) \;\Big|\; Z_i, Z_j\Big]\Big] \\
&= \sum_{i=1,j\neq i}^{k} \mathbb{E}\Big[\mathbb{E}\Big[(A(Z))_{i,j} \;\Big|\; Z_i, Z_j\Big]U(Z_i, Z_j)\Big] \\
&= \sum_{i=1,j\neq i}^{k} \int_{[0,1]^2} \mathbb{E}\Big[(A(Z))_{i,j} \;\Big|\; (Z_i, Z_j) = (x, y)\Big]U(x, y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{[0,1]^2} \left(\sum_{i,j=1}^{k} \mathbb{E}\Big[(A(Z))_{i,j} \;\Big|\; (Z_i, Z_j) = (x, y)\Big]\right) U(x, y)\,\mathrm{d}x\,\mathrm{d}y. \qquad (4.72)
\end{aligned}$$

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

$$\phi(x, y) \coloneqq \sum_{i,j=1}^{k} \mathbb{E}\left[\Big(\nabla f_k\Big((W(Z_i, Z_j))_{(i,j)\in[k]^{(2)}}\Big)\Big)_{i,j} \;\middle|\; Z_i = x, Z_j = y\right], \qquad (4.73)$$

for $(x, y) \in [0,1]^{(2)}$, then the required equality in equation (4.71) is satisfied. And the action of the Fréchet-like derivative ϕ on a kernel, say $U \in \mathcal{W}$, is

$$\begin{aligned}
&\mathbb{E}[\langle \nabla f_k \circ W_k, G_k[U]\rangle] \\
&\coloneqq \mathbb{E}\Big[\Big\langle \nabla f_k\Big((W(Z_i, Z_j))_{(i,j)\in[k]^{(2)}}\Big), (U(Z_i, Z_j))_{(i,j)\in[k]^{(2)}}\Big\rangle\Big].
\end{aligned} \qquad (4.74)$$

Lemma 4.5.1. *If f_k is λ-semiconvex on the convex set $\mathcal{M}_k$, then F_k is generalized geodesically $k(k-1)\lambda$-semiconvex on $\widehat{\mathcal{W}}$. In particular, it is also geodesically $k(k-1)\lambda$-semiconvex on $\widehat{\mathcal{W}}$.*

Proof. Since every geodesic is also a generalized geodesic, it suffices to prove the result for a generalized geodesic. Since f_k is λ-semiconvex, for some $\lambda \in \mathbb{R}$, it follows that for any

$X_0, X_1 \in \mathcal{M}_k$,

$$f_k(X_t) \leq (1-t)f_k(X_0) + tf_k(X_1) - \frac{\lambda}{2}t(1-t)\|X_1 - X_0\|_2^2, \quad \forall\, t \in [0,1], \tag{4.75}$$

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 2.2.6 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 $\{Z_i\}_{i=1}^k$ of i.i.d. Uni$[0,1]$ random variables as above to get a process $(X_{t,k})_{k\in\mathbb{N}}$ of random matrices with distributions $(\rho_k([W_t]))_{k\in\mathbb{N}}$ respectively, for each $t \in [0,1]$. Note that $X_{t,k} = (1-t)X_{0,k} + tX_{1,k}$, $t \in [0,1]$, $k \in \mathbb{N}$. Hence applying equation (4.75) to this line segment and then taking expectations with respect to the joint law of $(Z_i)_{i=1}^k$, we get

$$F_k([W_t]) \leq (1-t)F_k([W_0]) + tF_k([W_1]) - \frac{\lambda}{2}t(1-t)\mathbb{E}\Big[\|X_{1,k} - X_{0,k}\|_2^2\Big], \quad t \in [0,1].$$

Now by equation (4.69),

$$\mathbb{E}\Big[\|X_{1,k} - X_{0,k}\|_2^2\Big] = k(k-1)\|W_1 - W_0\|_2^2 \geq k(k-1)\delta_2^2([W_1],[W_0]).$$

Putting it back together we get that for $t \in [0,1]$,

$$F_k(\omega_t) \leq (1-t)F_k(\omega_0) + tF_k(\omega_1) - \frac{k(k-1)\lambda}{2}t(1-t)\delta_2^2(\omega_1,\omega_0). \tag{4.76}$$

Therefore F_k is $k(k-1)\lambda$-semiconvex along the generalized geodesic ω. □

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

Scalar Entropy function

Recall the scalar entropy function $\mathcal{E}\colon \widehat{\mathcal{W}} \to \mathbb{R}\cup\{\infty\}$, defined in equation (4.7) as

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

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, as defined in Section 4.1.2. Observe

that it can also be thought of as a linear function with $k = 2$. If

$$f_2\Big(\big(x_{(i,j)}\big)_{(i,j)\in[2]^{(2)}}\Big) \coloneqq h(x_{(1,2)}), \qquad x \in \mathcal{M}_2, \tag{4.77}$$

then from equations (4.67) and (4.68),

$$\begin{aligned} F_2([W]) = \langle f_2, \rho_2([W])\rangle &= \mathbb{E}_{\{Z_i\sim\mathrm{Uni}[0,1]\}_{i=1}^2}\Big[f_2\Big(\big(W(Z_i,Z_j)\big)_{(i,j)\in[2]^{(2)}}\Big)\Big] \\ &= \mathbb{E}_{\{Z_i\sim\mathrm{Uni}[0,1]\}_{i=1}^2}[h(W(Z_1,Z_2))] = \mathcal{E}([W]), \end{aligned} \tag{4.78}$$

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 [150, Lemma 8.22], it follows that $\mathcal{E}$ restricted to $\widehat{\mathcal{W}}_\epsilon$ is $\delta_\square$-continuous.

Since $h''(p) = 1/(p(1-p)) \ge 4$ 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

$$\begin{aligned} E(W_t) &= \int_0^1\int_0^1 h(W_t(x,y))\,\mathrm{d}x\,\mathrm{d}y \\ &\le \int_0^1\int_0^1 [(1-t)h(W_0(x,y)) + th(W_1(x,y))]\,\mathrm{d}x\,\mathrm{d}y \\ &\quad - \int_0^1\int_0^1 \frac{4}{2}t(t-1)(W_0-W_1)^2(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= (1-t)E(W_0) + tE(W_1) - \frac{1}{2}4t(1-t)\|W_0-W_1\|_2^2, \end{aligned} \tag{4.79}$$

which implies that E is 4-semiconvex on $(\mathcal{W}, d_2)$ (see Definition 2.1.16). Following Remark 4.1.2, $\mathcal{E}$ is 4-semiconvex along generalized geodesics on $(\widehat{\mathcal{W}}, \delta_2)$.

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

$$\mathbb{E}[\langle \nabla f_2 \circ W_2, G_2[U]\rangle] = \int_0^1\int_0^1 \log\Big(\frac{W(z_1,z_2)}{1-W(z_1,z_2)}\Big)U(z_1,z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2, \tag{4.80}$$

for all $U \in \mathcal{W}$. By the characterization of the Fréchet-like derivative in equation (4.74), we get that $\phi_{\mathcal{E}} = D_{\mathcal{W}} f_2(W)$ is given by

$$\phi_{\mathcal{E}}(x,y) = \log\Big(\frac{W(x,y)}{1-W(x,y)}\Big), \qquad \text{a.e.} \quad (x,y) \in [0,1]^{(2)}. \tag{4.81}$$

If $[W] \in \widehat{\mathcal{W}}_\epsilon$ for some $\epsilon \in (0, 1/2)$, then by Lemma 4.2.9 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 [150, 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 4.2.15, 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 $\epsilon \le W_0 \le 1 - \epsilon$ a.e. Consider the flow $W_t(x, y)$ obtained by solving

$$W_t'(x, y) = -\log\left(\frac{W_t(x, y)}{1 - W_t(x, y)}\right), \qquad \text{a.e.} \quad (x, y) \in [0, 1]^{(2)}, \tag{4.82}$$

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 $\omega_0 = [W_0]$.

Takeaway. It is worth emphasizing that, following equation (4.82), 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 4.2.16 that we are guaranteed an exponential rate of convergence of the gradient flow to the minimizer.

Homomorphism functions

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

$$T_H([W]) \coloneqq \mathbb{E}\left[\prod_{\{i,j\}\in E(H)} W(Z_i, Z_j)\right] = \int_{\mathcal{M}_k} f_k(z)\rho_k(\mathrm{d}z), \tag{4.83}$$

where H is a simple graph with $|V(H)| = k$, $E(H) = \{e_i\}_{i=1}^m$, $\{Z_i\}_{i\in[k]}$ are i.i.d. uniformly in $[0, 1]$, and $f_k\big((x_{(i,j)})_{(i,j)\in[k]^{(2)}}\big) = \prod_{\{i,j\}\in E(H)} x_{(i,j)}$. In particular, the function f_k is a monomial for every simple graph H. Let $t_H \colon \mathcal{W} \to \mathbb{R}$ be the invariant extension of T_H such that $t_H \coloneqq T_H \circ [\,\cdot\,]$. Since

$$(\nabla f_k(X))_{(p,q)} = \mathbb{1}_{E(H)}\{p, q\} \cdot \prod_{\{i,j\}\in E(H)\setminus\{\{p,q\}\}} X_{(i,j)}, \tag{4.84}$$

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 (4.74) is given by

$$\sum_{\ell=1}^{m} \mathbb{E}\left[\prod_{r=1}^{\ell-1} W(Z_{e_r}) \cdot U(Z_{e_\ell}) \cdot \prod_{r=\ell+1}^{m} W(Z_{e_r})\right], \tag{4.85}$$

where $Z_e \coloneqq \big(Z_{e(1)}, Z_{e(2)}\big)$ for all $e \in E(F)$. Following a similar approach to equation (4.72), we obtain that ϕ_{T_H} satisfies

$$\begin{aligned} \phi_{T_H}(x,y) &= \sum_{\ell=1}^{m} \mathbb{E}\left[\prod_{r=1, r\neq\ell}^{m} W(Z_{e_r}) \,\middle|\, Z_{e_\ell} = (x,y)\right] \\ &= \sum_{\ell=1}^{m} \mathbf{t}_{x,y}(H_{e_\ell}, W), \quad x, y \in [0,1], \end{aligned} \tag{4.86}$$

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* [150, 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 $\{m,n\} \in E(H_e)$, consider the integral

$$\begin{aligned} I_{m,n} \coloneqq \int_{[0,1]^k} (W(x_m, x_n) - W'(x_m, x_n)) \prod_{\{i,j\} \in E(H_e)\setminus\{m,n\}} W(x_i, x_j) \cdot \\ \prod_{\ell \in V(H_e)\setminus e} \mathrm{d}x_\ell \; f(x)g(y)\,\mathrm{d}x\,\mathrm{d}y. \end{aligned}$$

It follows from [150, Lemma 8.10] (or see the second last display in the proof of [150, Lemma 10.24]) that $|I_{m,n}| \leq \|W - W'\|_\square$. From the same references above, it follows that

$$\begin{aligned} &\left|\int_{[0,1]^2} (\mathbf{t}_{x,y}(H_e, W) - \mathbf{t}_{x,y}(H_e, W'))f(x)g(y)\,\mathrm{d}x\,\mathrm{d}y\right| \\ &\leq \sum_{\{m,n\} \in E(H_e)} |I_{m,n}| \leq |E(H_e)|\big\|W - W'\big\|_\square. \end{aligned}$$

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 (4.86) 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 [150, 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 $\deg(x) := \int_0^1 W(x,z)\,\mathrm{d}z$ for a.e. $x, y \in [0,1]$.

Obtaining the expression for the Fréchet-like derivative in equation (4.86), 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 f_k can be easily computed as

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

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 $(i,j) \in [k]^{(2)}$, every entry of the Hessian is uniformly bounded in $[-1,1]$ and hence $\|\mathrm{Hess}(f_k)\|_{\mathrm{op}} \leq k(k-1)/2$. Therefore, f_k is $(-k(k-1)/2)$-semiconvex w.r.t. d_2. It follows from Lemma 4.5.1 that homomorphism function T_H is $(-k^2(k-1)^2/2)$-semiconvex w.r.t. δ_2. In fact, since $\mathrm{Hess}(f_k)(\{i,j\},\{p,q\})$ is non-zero only when $\{i,j\},\{p,q\} \in E(H)$, it follows that $\|\mathrm{Hess}(f_k)\|_{\mathrm{op}} \leq \sqrt{m(m-1)} \leq m$. This would yield that T_H is $(-mk(k-1))$-semiconvex w.r.t. δ_2. This is useful when H is sparse.

Takeaway. 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 4.2.16. To alleviate this, one can regularize the objective function by adding the scalar entropy function. We discuss this in the next example in Section 4.5.1.

Linear combination of Scalar Entropy and Homomorphism function

Let $\beta \in \mathbb{R}$ and let H be a finite simple graph with $k \in \mathbb{N}$ 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 graph models (see Section 4.1). 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 4.5.1. The gradient flow of G, therefore, exists and the Fréchet-like derivative $\phi_G = D_{\mathcal{W}} G(W) = \phi_{\mathcal{E}} + \beta\phi_{T_H}$. Since $\mathcal{E}$ is 4-semiconvex and T_H is $(-k^2(k-1)^2/2)$-semiconvex w.r.t. δ_2, it follows that $G_{\beta,H}$ is $(4 - \beta k^2(k-1)^2/2)$-semiconvex w.r.t. δ_2. Therefore, $G_{\beta,H}$ is at least 0-semiconvex (i.e., convex) w.r.t. δ_2 when $\beta \leq 8/k^2(k-1)^2$.

Takeaway. Note that $\beta < 8/k^2(k-1)^2$ guarantees exponential rates of convergence of the gradient flow curve. See Remark 4.2.16.

4.5.2 Interaction energy

In the optimal transport literature, linear functionals of probability measures are often called *potential energy* [192, page 249]. Inspired by similar definitions, one can define *interaction energy*. Let $f_k \colon \mathcal{M}_k \times \mathcal{M}_k \to \mathbb{R}$ be a function defined on pairs of $k \times k$ symmetric matrices with entries in $[-1, 1]$. Given a graphon $[W] \in \widehat{\mathcal{W}}$, as before let $\rho_k = \mathrm{Law}(G_k[W])$. This defines a function $\mathbb{F}_k \colon \widehat{\mathcal{W}} \to \mathbb{R} \cup \{\infty\}$ given by

$$\mathbb{F}_k([W]) \coloneqq \int_{\mathcal{M}_k} \int_{\mathcal{M}_k} f_k(z, z') \rho_k(\mathrm{d}z) \rho_k(\mathrm{d}z').$$

Although it looks different than before, this is also a particular case of equation (4.74) as shown below. Define two independent sequences of i.i.d. $\mathrm{Uni}[0, 1]$ random variables: $(Z_1, \ldots, Z_k)$ and $(Z'_1, \ldots, Z'_k)$, and consider the corresponding matrices

$$X_k \coloneqq (W(Z_i, Z_j))_{(i,j)\in[k]^{(2)}}, \quad \text{and} \quad X'_k \coloneqq (W(Z'_i, Z'_j))_{(i,j)\in[k]^{(2)}}.$$

Then X_k and X'_k are i.i.d. samples from ρ_k and $\mathbb{F}_k([W]) = \mathbb{E}[f_k(X_k, X'_k)]$. On the other hand, one can concatenate $(Z_i)_{i=1}^k$ and $(Z'_i)_{i=1}^k$ to construct a single vector $(Z_1, \ldots, Z_k, Z'_1, \ldots, Z'_k)$ of dimension $2k$ and consider the corresponding $2k \times 2k$ symmetric matrix X_{2k} whose block diagonal components are X_k and X'_k. By defining $\bar{f}_k(X_{2k}) \coloneqq$

$f_k(X_k, X_k')$, we represent $\mathbb{F}_k([W]) = \mathbb{E}\big[\bar{f}_{2k}(X_{2k})\big] = \int \bar{f}_{2k}(w)\rho_{2k}(\mathrm{d}w)$ and equation (4.74) 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_k, W_k' be i.i.d. sampled $(k \times k)$ symmetric matrices from a graphon $[W]$. Consider the function

$$\mathbb{F}_k([W]) \coloneqq \frac{1}{2}\mathbb{E}\left[\left(\prod_{e\in E(H)} W(Z_e) - \prod_{e\in E(H)} W(Z_e')\right)^2\right] = \mathrm{Var}\left[\prod_{e\in E(H)} W(Z_e)\right], \tag{4.87}$$

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 $f_k(z, z') = (g_k(z) - g_k(z'))^2$, for any function g_k on $\mathcal{M}_k$ that is square-integrable w.r.t. ρ_k. Unfortunately this particular function $\mathbb{F}_k$ in (4.87) is continuous in δ_2 but not in $\delta_\square$. See [117, 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 k vertices, consider its simple subgraphs H_1 and H_2 with k_1 and k_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 $\{Z_i\}_{i=1}^{k_1} \cup \{Z_i'\}_{i=1}^{k_2}$ to get a single set of $k_1 + k_2 \geq k$ i.i.d. Uni$[0,1]$ random variables $\{U_i\}_{i=1}^{k_1+k_2}$ such that $\{U_i\}_{i=1}^{k_1} = \{Z_i\}_{i=1}^{k_1}$ and $\{U_i\}_{i=k_1+1}^{k} = \left\{Z_j'\right\}_{j\in V(H)\setminus V(H_1)}$. We can then define $I_{H_1,H_2}(\,\cdot\,; H)\colon \widehat{\mathcal{W}}_\epsilon \to \mathbb{R}\cup\{\infty\}$ as

$$\begin{aligned} I_{H_1,H_2}([W]; H) \coloneqq \log&\left(\mathbb{E}\left[\prod_{\{i,j\}\in E(H_1)} W(Z_i, Z_j)\right]\right) \\ &+ \log\left(\mathbb{E}\left[\prod_{\{i,j\}\in E(H_2)} W(Z_i', Z_j')\right]\right) \\ &- 2\log\left(\mathbb{E}\left[\prod_{\{i,j\}\in E(H)} W(U_i, U_j)\right]\right), \end{aligned} \tag{4.88}$$

for some $\epsilon \in (0,1)$. Each of the terms in the expression in equation (4.88) 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 [150, 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 [150, Section 4.2]. H_1H_2 can have multiple edges. Then, using [150, 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

$$\begin{aligned} T_{H_1H_1}([W])T_{H_2H_2}([W]) - T^2_{H_1H_2}([W]) \geq 0, \\ \text{i.e.,} \quad \frac{T_{H_1H_1}([W])T_{H_2H_2}([W])}{T^2_{H_1H_2}([W])} \geq 1. \end{aligned} \tag{4.89}$$

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. 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 (4.89), 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 4.5.1 shows that this function satisfy all our conditions. The computation for its Fréchet-like derivative also follows from Section 4.5.1 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 [156, proof of Theorem 2.2]). It is an interesting coincidence that they also appear in this context.

4.5.3 Internal energy

Similar to potential and interaction energies, one can define non-linear functions of ρ_k 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

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

Here we have used the standard abuse of notation in 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 $1 \leq l \leq k$, consider a function $G_{k,l} \colon \mathcal{M}_k \to [-1,1]^l$ that selects a particular subset of length l from the upper-diagonal elements of a $k \times k$ matrix. Consider the pushforward $\rho_{k,l}([W]) \coloneqq (G_{k,l})_\sharp (\rho_k([W]))$. Since ρ_k is generated by k i.i.d. Uni$[0,1]$ random variables, and $l \leq k$, it is easy to come up with examples where $\rho_{k,l}$ has a density. For example, take $W(x,y) = \sin(x+y)$, $k = 3, l = 2$, and $G_{k,l}(A) = (A_{1,2}, A_{1,3})$. Thus $(G_{k,l} \circ G_k)([W]) = (\sin(Z_1 + Z_2), \sin(Z_1 + Z_3))$. This random vector has a density. Thus, the function $U\left(\rho_{k,l}([W])\right)$ 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 $\rho_{k,l}([W])$.

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 [122].

Chapter 5

STOCHASTIC OPTIMIZATION ON MATRICES AND A GRAPHON MCKEAN-VLASOV

5.1 Introduction

The study of particle systems under mean-field interaction is a classical topic in probability theory [95]. It involves multidimensional diffusions that interact through their empirical distributions of the type

$$\mathrm{d}X_i(t) = b\left(X_i(t), \hat{\mu}^{(N)}(t)\right)\mathrm{d}t + \mathrm{d}B_i(t), \quad i \in [N], \quad t \in \mathbb{R}_+, \tag{5.1}$$

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. Prominent examples of such particle systems include the diffusion given by the SDE

$$\mathrm{d}X_i(t) = -\nabla V\left(X_i(t)\right)\mathrm{d}t - \frac{1}{N}\sum_{j=1}^{N} \nabla W\left(X_i(t) - X_j(t)\right)\mathrm{d}t + \mathrm{d}B_i(t), \quad t \in \mathbb{R}_+, \tag{5.2}$$

for $i \in [N]$, where V and W are differentiable convex functions on $\mathbb{R}^d$. However, any drift that is symmetric in the coordinates ("mean-field interactions") can be represented as (5.1) for some suitable function b. Often, the SDE (5.1) includes a reflection term to constrain the coordinate process to a subset of the Euclidean space [203]. The study of such systems originated from the probabilistic study of the Boltzmann and Vlasov equations due to Kac [123], McKean [157], Dobrushin [76], Tanaka [205] and many others. For modern surveys, see Sznitman [204], Villani [211], Chaintron and Diez [51] and Jabin [114].

Under suitable assumptions, as the number of particles goes 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. For example, for the system (5.2), the random process $\hat{\mu}^{(N)}$ converges weakly to the solution of granular media equation [49], as $N \to \infty$. 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 [95]: $\mathrm{d}X(t) = b\left(X(t), \mu(t)\right)\mathrm{d}t + \mathrm{d}B(t)$, $t \in \mathbb{R}_+$, where $\mu(t)$ is the law of $X(t)$.

In this work, 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 $\mathcal{M}_n^0$, the set of all $n \times n$ symmetric matrices with entries in $[0,1]$, given by

$$R_n(A) \coloneqq \frac{1}{2}\left(n^{-2}\sum_{i,j=1}^{n} A(i,j) - e\right)^2 + \frac{1}{2}\left(n^{-3}\operatorname{Tr}\left[A^3\right] - \tau\right)^2 + \mathcal{E}_n(A), \tag{5.3}$$

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 function R_n is invariant, 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

$$\mathrm{d}X_n(t) = -n^2\nabla R_n(X_n(t))\,\mathrm{d}t + \beta\,\mathrm{d}B_n(t) + \mathrm{d}L_n(t), \qquad t \in \mathbb{R}_+, \tag{5.4}$$

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 5.2.2 for details). One may ask what is an appropriate notion of limit of such a process as $n \to \infty$? Does (5.4) exhibit propagation of chaos?

Note that the function R_n in (5.4) 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. The same is true for any 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. As we have seen this particular class of symmetry is captured, not by empirical measures, but by graphons.

Analogous to the classical McKean-Vlasov theory, we show in this chapter that, under suitable assumptions, (5.4) exhibits a propagation of chaos. Furthermore, in $n \to \infty$ limit, the coordinates of $X_n(t)$ become conditionally independent and the evolution of a randomly chosen coordinate can be described by a novel graphon valued McKean-Vlasov equation. The existence and uniqueness of such a process are established in Proposition 5.4.5. Proposition 5.4.6 shows that the process $X_n(t)$ converges to a deterministic curve on the space of graphons, $\widehat{\mathcal{W}}$ (see Section 5.2).

While the evolution of matrices described by (5.4) can arise in many different contexts, our primary motivation for studying such evolution comes from the stochastic gradient descent on large graphs. Consider the function R_n defined in (5.3). Notice that if we regard $A \in \mathcal{M}_n^0$ 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 $n^{-3}\operatorname{Tr}\big[A^3\big]$ can be regarded as the density of triangles in the graph. Consider the problem of minimizing $\mathcal{E}_n(A)$ over all $A \in \mathcal{M}_n^0$ 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 [163, 132]. For certain feasible regime of (e, τ), this minimization problem has been studied in [163]. In some regimes where the minimizer is known to be unique, the minimizer is characterized. Similar results were obtained in [132] 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 [183, 179] and exponential random graph models [30, 58]. While minimizing $\mathcal{E}_n$ with such 'hard constraints' is difficult, notice that minimizing $R_n(A) := \Big(n^{-2}\sum_{i,j=1}^n A(i,j) - e\Big)^2/2 + \big(n^{-3}\operatorname{Tr}\big[A^3\big] - \tau\big)^2/2 + \mathcal{E}_n(A)$ can be considered as a relaxation of this problem. Our method provides a numerical scheme to obtain minimizers of such problems. Notice that (5.4) arise as the 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 (5.4) converges to a deterministic curve on the space of graphons. Under appropriate assumptions on R_n (see Section 5.6.3) 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. Thus, the evolution (5.4)

gives a way to numerically approximate the minimizer. More generally, the limiting curve converges to stationary points and thus (5.4) provides an algorithm to numerically approximate these stationary points that may be useful in obtaining reasonable guesses regarding the structure of the minimizers in such problems. We describe the projected gradient descent and projected stochastic gradient descent algorithms in more detail in the following paragraphs.

Projected Gradient Descent (GD) based algorithms are the workhorse in optimizing such functions [50, 44, 42]. However, in most cases, computing gradients can be computationally intensive. 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 [187, 134]. The details of this common Markov chain are described later in the section, and the reader can refer to the monographs [24, 141, 41, 159, 142] for a detailed overview. Roughly, if the current state is a symmetric matrix A, one jumps to a new state by taking a small step along the negative Euclidean gradient $-\nabla R_n(A)$, and potentially adding independent, centered, and variance-bounded noise to each matrix entry (up to symmetry). Each matrix entry is then projected onto the interval $[0, 1]$ to satisfy the entrywise constraint.

Gradient descent (GD), with small step sizes, approximates the Euclidean gradient flow obtained as a solution to Cauchy's problem

$$\dot{A}_{i,j}(t) = -\nabla_{i,j} R_n(A(t)), \qquad (i,j) \in [n]^2, \qquad t \in \mathbb{R}_+,$$

in the interior of $\mathcal{M}_n^0$. Here $\mathbb{R}_+$ denotes the set of non-negative real numbers which is used to index time, $\nabla_{i,j}$ refers to the partial derivative with respect to the (i,j)-th matrix entry. It is therefore natural to understand a suitable scaling limit of SGD on the space of such matrices.

We saw in the last chapter that under suitable assumptions on $(R_n)_{n\in\mathbb{N}}$, the implicit Euler update scheme approximates the gradient flow curve, in an appropriate sense, over $\widehat{\mathcal{W}}$, when the step size is taken to zero and n grows to infinity. Refer to Section 2 for the required exposition on graphons. In this work, we ask a similar question for SGD-based algorithms. We show that under an appropriate "small noise" assumption and consistency and other

suitable assumptions on the functions $(R_n)_{n\in\mathbb{N}}$, the SGD iterations converge appropriately to a limiting deterministic curve that is a gradient flow on the space of graphons. Moreover, when an extra Gaussian noise is added to each SGD iterate, the noisy SGD iterations also converge to a deterministic curve on graphons which admits a McKean-Vlasov description.

Recall that a function $R\colon \widehat{\mathcal{W}} \to \mathbb{R}$ over graphons naturally extends to a function over the set of kernels $\mathcal{W}$. For any $n \in \mathbb{N}$, the set of symmetric matrices $\mathcal{M}_n$, over which algorithms like GD and SGD operate, can be naturally identified with a subset, *finite dimensional kernels*, $\mathcal{W}_n \subset \mathcal{W}$ of the kernels (see Section 2 for details). As usual this identification/embedding will be denoted by K (as in kernel) and its inverse will be denoted by M_n (as in matrix). Using K, the restriction of the function R to $\mathcal{W}_n$ can be viewed as a function R_n on $\mathcal{M}_n$.

Define the projection operator $P\colon \mathbb{R} \to [-1,1]$ 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 on matrices and kernels. For every $n \in \mathbb{N}$, let $\boldsymbol{\tau}_n \coloneqq (\tau_{n,k})_{k\in\mathbb{Z}_+}$, be a sequence of positive step sizes (also known as the 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 assume $\boldsymbol{\tau}_n$ to have a divergent sum so to cover the whole non-negative real line $\mathbb{R}_+$, i.e., to satisfy $\lim_{k\to\infty} t_{n,k} = \infty$. 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 is assumed to be finite. We now describe our first iterative scheme.

Definition 5.1.1 (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 $V_{n,0} \in \mathcal{M}_n$ is defined to be a sequence of symmetric matrices $(V_{n,k})_{k\in\mathbb{Z}_+}$ given iteratively as*

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

Suppose R is such a function whose Fréchet-like derivative evaluation map is denoted by ϕ. If R_n is obtained from R by restricting R to $\mathcal{M}_n$ and the function R_n is differentiable up to the boundary of $\mathcal{M}_n$ for every $n \in \mathbb{N}$. We know that

$$n^2 \nabla R_n = M_n \circ \phi \circ K. \tag{5.5}$$

Simply put, n^2 times the Euclidean gradient of R_n at a matrix argument A can be identified as the Fréchet-like derivative ϕ of R at the kernel argument $K(A)$. The time in the Euclidean gradient in Definition 5.1.1 is therefore scaled by n^2 following the relation (5.5). The PGD algorithm is essentially the explicit Euler iteration scheme up to the projection.

We now define the stochastic optimization setup for R_n. In order to do so, we first fix some notations and make some assumptions on R and R_n. 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 $g\colon \mathcal{W} \times \Omega \to L^\infty\big([0,1]^{(2)}\big)$ where $L^\infty\big([0,1]^{(2)}\big)$ is the set of all bounded measurable functions $\phi : [0,1]^2 \to \mathbb{R}$ such that $\phi(x,y) = \phi(y,x)$. To emphasize that ϕ is symmetric, we denote the domain by $[0,1]^{(2)}$ which denotes the set $\{(x,y) \in [0,1]^2 : x \le y\}$. Define g_n on $\mathcal{M}_n \times \Omega$ as $g_n(A;\xi) = g(K(A);\xi)$ for every $n \in \mathbb{N}$ and $A \in \mathcal{M}_n$, and assume that

$$\nabla R_n = \mathbb{E}_{\xi\sim\mathcal{D}}[g_n(\,\cdot\,;\xi)]. \tag{5.6}$$

Under suitable assumptions (see Assumption 4) on the function g, the function R is invariant under measure preserving transformations and hence defines a function on $\widehat{\mathcal{W}}$. We are interested in stochastic analogues of the iteration scheme in Definition 5.1.1, for such a function R, possibly with a noise at each iteration. In other words, our interest lies in noisy variations of projected GD iterations (see Definition 5.1.1). In this setting, 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 $g_n(\,\cdot\,;\xi_{k+1})$. As a special case, g can be obtained from a function $\ell\colon \mathcal{W} \times \Omega \to \mathbb{R}$, as $g(\,\cdot\,;\xi) \coloneqq (D_{\mathcal{W}})\ell(\,\cdot\,;\xi)$ for all $\xi \in \Omega$, where $(D_{\mathcal{W}}\ell)(\,\cdot\,;\xi)$ is the Fréchet-like derivative (see Definition 4.2.6) of $\ell(\,\cdot\,;\xi)$. Such a stochastic approximation is known as Stochastic Gradient Descent (SGD).

2. **Large noise**: We can add an additive noise to iterates in equation (PGD) before the projection, as we describe in Definition 5.1.2 below.

We can now define the noisy analogs of (PGD), that is, *projected (noisy) SGD*. We will use the operator $\circ$ over symmetric matrices to denote the Hadamard (elementwise) product.

Definition 5.1.2 (Projected SGD with and without noise)**.** *Let $n \in \mathbb{N}$. Starting at $W_{n,0} \in \mathcal{M}_n$, the projected (noisy) SGD algorithm produces a sequence of iterates $(W_{n,k})_{k\in\mathbb{Z}_+}$ defined as*

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

Here $(G_{n,k})_{k\in\mathbb{Z}_+}$ is an $n \times n$ symmetric matrix valued martingale difference sequence independent of $(\xi_{k+1})_{k\in\mathbb{Z}_+}$. We only consider the noise $G_{n,k}$, for $k \in \mathbb{Z}_+$, of the form $G_{n,k} = \Sigma_n(W_{n,k}) \circ Z_{n,k}$ for some Σ_n that maps matrices in $\mathcal{M}_n$ to $n \times n$ symmetric matrices with non-negative entries and $(Z_{n,k})_{k\in\mathbb{Z}_+}$ is a sequence of independent $n \times n$ symmetric random matrices with standard normal entries (up to matrix symmetry).

Due to the natural identification of $\mathcal{M}_n$ with $\mathcal{W}_n$, the GD iterates $(V_{n,k})_{k\in\mathbb{Z}_+} \subset \mathcal{M}_n$ and the SGD iterates $(W_{n,k})_{k\in\mathbb{Z}_+} \subset \mathcal{M}_n$ in Definitions 5.1.1 and 5.1.2 respectively, can be viewed as kernel valued iterates $\big(V_k^{(n)}\big)_{k\in\mathbb{Z}_+} \subset \mathcal{W}_n$ and $\big(W_k^{(n)}\big)_{k\in\mathbb{Z}_+} \subset \mathcal{W}_n$, under the embeddings $V_k^{(n)} = K(V_{n,k})$ and $W_k^{(n)} = K(W_{n,k})$ respectively for $k \in \mathbb{Z}_+$. This allows us to interpret (PGD) and (PNSGD) as kernel-valued updates.

We consider piecewise constant interpolations of the iterates (see Definition 5.2.1) and in this chapter, we establish the existence of the scaling limit of these curves. We also characterize the limit under the absence of "large noise". Our limiting procedure takes two steps. First, for every fixed $n \in \mathbb{N}$, we take the step size, i.e., $|\boldsymbol{\tau}_n| \to 0$ to obtain a limiting SDE on $\mathcal{M}_n$. We then characterize the limit of the SDEs as $n \to \infty$ as an absolutely continuous curve on the space of graphons.

Theorem 5.1.3. *Let $n \in \mathbb{N}$ be fixed, and suppose Assumptions 3, 4 and 5 hold (see Section 5.2.1). Let $W_n \colon \mathbb{R}_+ \to \mathcal{M}_n$ be the piecewise constant interpolation (Definition 5.2.1) of noisy SGD iterates $(W_{n,k})_{k\in\mathbb{Z}_+}$ as defined in (PNSGD). Then, W_n converges weakly in*

the space of càdlàg processes to X_n as $|\boldsymbol{\tau}_n| \to 0$ that satisfies the SDE:

$$\mathrm{d}X_n(t) = -n^2 \nabla R_n(X_n(t))\,\mathrm{d}t + \Sigma_n(X_n(t)) \circ \mathrm{d}B_n(t) + \mathrm{d}L_n^-(t) - \mathrm{d}L_n^+(t), \tag{RSDE}$$

for $t \in \mathbb{R}_+$, starting at $X_n(0) = W_{n,0}$. Here B_n is an $n \times n$ symmetric matrix-valued process with coordinatewise independent standard Brownian motions up to matrix symmetry, and (X_n, L_n^+, L_n^-) solves the Skorokhod problem with respect to the set $\mathcal{M}_n$ (see Section 5.2.2).

Our main interest is in the limit of the kernel-valued stochastic process $X^{(n)}(\cdot) = K(X_n(\cdot))$ (Theorem 5.1.3), as $n \to \infty$. This limit is a deterministic curve in $\widehat{\mathcal{W}}$ that we now describe. Consider, for simplicity, the special case when each Σ_n is β times the identity matrix for some $\beta > 0$. On a probability space that supports a standard linear Brownian motion $B_{1,2}(\cdot)$ and a pair of independent Uni$[0,1]$ random variables (U_1, U_2) and given some $W_0 \in \mathcal{W}$, one can construct a unique solution of the following family of one-dimensional reflected diffusions. Given $(U_1, U_2) = (x, y)$, for some $(x, y) \in [0,1]^{(2)}$, let $X_{1,2}$ be a diffusion with state space $[-1, 1]$ with the initial condition $X_{1,2}(0) = W_0(x, y)$, and satisfying

$$\mathrm{d}X_{1,2}(t) = -\phi\left(\Gamma(t)\right)(x, y)\,\mathrm{d}t + \beta\,\mathrm{d}B_{1,2}(t) + \mathrm{d}L_{1,2}^-(t) - \mathrm{d}L_{1,2}^+(t), \tag{5.7}$$

for some $\beta \in \mathbb{R}_+$ and $t \in \mathbb{R}_+$. Here, ϕ is the Fréchet-like derivative of R in (5.5), $L_{1,2}^-$ and $L_{1,2}^+$ are the local time processes such that $(X_{1,2}, L_{1,2}^+, L_{1,2}^-)$ solves the Skorokhod problem with respect to $[-1, 1]$ (see Section 5.2.2). The kernel-valued process $\Gamma\colon \mathbb{R}_+ \to \mathcal{W}$ is given by

$$\Gamma(t)(u, v) \coloneqq \mathbb{E}[X_{1,2}(t) \mid (U_1, U_2) = (u, v)], \quad \forall\, (u, v) \in [0,1]^{(2)}, \tag{5.8}$$

and any $t \in \mathbb{R}_+$. In Proposition 5.4.5, we show that the coupled system $(X_{1,2}, \Gamma)$ exists in a strong sense and is pathwise unique and that the kernel-valued process $X^{(n)}$ in Theorem 5.1.3 converges to the curve Γ in the following sense an $n \to \infty$.

Theorem 5.1.4. *Suppose Assumptions 3, 5, and 6 hold (see Section 5.2.1). Then, for any sequence of initial kernels $\left(W_0^{(n)} \in \mathcal{W}_n\right)_{n\in\mathbb{N}}$ that converges in $L^2\left([0,1]^{(2)}\right)$ norm $\|\cdot\|_2$, i.e.,*

$$\lim_{n\to\infty} \left\| W_0^{(n)} - W_0 \right\|_2 = 0, \tag{5.9}$$

the process of random kernels $\big(X^{(n)}(t) = K(X_n(t))\big)_{t\in\mathbb{R}_+}$ *obtained from solutions of the SDE* (RSDE)*, converges locally uniformly in the cut norm, in probability, to the curve* $\Gamma\colon \mathbb{R}_+ \to \mathcal{W}$*, with* $\Gamma(0) = W_0$*, defined in equation* (5.8) *as* $n \to \infty$*.*

Remark 5.1.5. *The assumption* $\big\|W_0^{(n)} - W_0\big\|_2 \to 0$ *can not be weakened to* $\big\|W_0^{(n)} - W_0\big\|_\square \to 0$ *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,* $\big\|W_0^{(n)} - W_0\big\|_\square \to 0$ *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 [184, 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 5.1.6. *We should also remark that arranging for* $W_0^{(n)}$ *such that* $\big\|W_0^{(n)} - W_0\big\|_2 \to 0$ *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.*

In Section 5.5 a more general statement has been proved (see Proposition 5.4.6). It is worth noting that the presence of noise and the boundary $\{-1, 1\}$ in our problem makes it non-trivial. To see this, consider (RSDE) 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 (RSDE) 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.

5.1.1 Scaling limit without added noise

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

$$\mathrm{d}X_n(t) = -n^2 \nabla R_n(X_n(t))\,\mathrm{d}t + \mathrm{d}L_n^-(t) - \mathrm{d}L_n^+(t), \quad t \in \mathbb{R}_+, \quad X_n(0) = W_{n,0}, \tag{5.10}$$

such that (X_n, L_n^+, L_n^-) solves the Skorokhod problem on $\mathcal{M}_n$ (see Section 5.2.2 for details). Moreover, it is shown in Section 5.3 that the solution of (5.10) is the same as the solution of (5.11) given below. Furthermore, it is shown in [167, Theorem 4.4, Theorem 4.14] that if the solution $X_n \colon \mathbb{R}_+ \to \mathcal{M}_n$ of

$$\mathrm{d}X_n(t) = -n^2 \nabla R_n(X_n(t)) \circ \mathbb{1}_{G_n(X_n(t))}\,\mathrm{d}t, \qquad t \in \mathbb{R}_+, \tag{5.11}$$

exists, where $G_n(A)$ is the subset of $[n]^2$ (defined in equation (5.36) later in Section 5.3.1), then X_n is a gradient flow on $\mathcal{M}_n$ in a suitable sense. As we know, under reasonable assumptions on R, the sequence of solutions $(X_n)_{n\in\mathbb{N}}$ of equation (5.11) obtained for all natural numbers $n \in \mathbb{N}$, converge to an absolutely continuous curve $W \colon \mathbb{R}_+ \to \mathcal{W}$, which is a curve of maximal slope [5] (a.k.a. gradient flow) of R, as $n \to \infty$. This yields the following.

Theorem 5.1.7. *Suppose Assumptions 3 and 4 hold (see Section 5.2.1). 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 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 (5.10). 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 $W(t) \coloneqq W_0 - \int_0^t \phi(W(s))\mathbb{1}_{G_{W(s)}}$ for $t \in \mathbb{R}_+$, is the gradient flow for R.

We should mention that our method allows us to also obtain a non-asymptotic rate of convergence. We refer the reader to Remark 5.5.2 for details.

5.2 Assumptions and Preliminaries

Since we want to obtain continuous time scaling limits of the iterative schemes defined in Definition 5.1.1 and Definition 5.1.2, we will use piecewise constant interpolations.

Definition 5.2.1 (Piecewise constant interpolation)**.** *Given a sequence $(a_k)_{k\in\mathbb{Z}_+}$ over any domain, and a sequence of positive step sizes $\boldsymbol{\tau} = (\tau_k)_{k\in\mathbb{Z}_+}$, we can define a piecewise constant interpolation of $(a_k)_{k\in\mathbb{Z}_+}$ as a right-continuous curve $a\colon \mathbb{R}_+ \to \{a_k\}_{k\in\mathbb{Z}_+}$ as*

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

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

5.2.1 Assumptions

In this section we state all the required assumptions we need to prove our results (see Theorem 5.1.3 and Theorem 5.1.4).

Assumption 3. *We make following assumptions on R, g and ϕ:*

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 ϕ is κ_2-Lipschitz with respect to $\|\cdot\|_2$, for some constant $\kappa_2 \in \mathbb{R}_+$. That is,*

$$\|\phi(W_1) - \phi(W_2)\|_2 \leq \kappa_2 \|W_1 - W_2\|_2, \qquad \forall\ W_1, W_2 \in \mathcal{W}.$$

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 4. *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 $A \in \cup_{n\in\mathbb{N}}\mathcal{W}_n$,*

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

Assumption 5. *We assume the following on the "large noise" for every $n \in \mathbb{N}$.*

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 $U, V \in \mathcal{W}$,*

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

Assumption 6. *There exists a constant $\kappa_\square \in \mathbb{R}_+$ such that, for almost every $(x,y) \in [0,1]^{(2)}$, the map $\phi_{x,y} \coloneqq \phi(\,\cdot\,)(x,y)$ is $\kappa_\square$-Lipschitz in cut norm $\|\cdot\|_\square$. That is, for every $U, V \in \mathcal{W}$,*

$$|\phi_{x,y}(U) - \phi_{x,y}(V)| \leq \kappa_\square\|U - V\|_\square.$$

5.2.2 Preliminaries on the system of reflected diffusions

For $n \in \mathbb{N}$, consider the domain $\mathcal{M}_n$. Notice that $\mathcal{M}_n$ is a cube, and is closed with respect to the usual topology. Consider the SDE:

$$\mathrm{d}X_n(t) = -n^2\nabla R_n(X_n(t))\,\mathrm{d}t + \Sigma_n(X_n(t)) \circ \mathrm{d}B_n(t) + \mathrm{d}L_n^-(t) - \mathrm{d}L_n^+(t), \tag{5.12}$$

for $t \in [0,T]$ for some fixed $T \in \mathbb{R}_+$ and starting at $X_n(0) = X_{n,0} \in \mathcal{M}_n$. Here Σ_n is a map from $\mathcal{M}_n$ 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 $\left(B_{n,(i,j)}\right)_{(i,j)\in[n]^{(2)}}$ which are independent up to matrix symmetry, and the processes L_n^- and L_n^+ are local times at the boundary. 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 $(i,j) \in [n]^2$,

$$\begin{aligned} \int_0^\infty \mathbb{1}\{X_{n,(i,j)}(t) > -1\} \, \mathrm{d}L^-_{n,(i,j)}(t) = 0, \qquad \text{and} \\ \int_0^\infty \mathbb{1}\{X_{n,(i,j)}(t) < +1\} \, \mathrm{d}L^+_{n,(i,j)}(t) = 0. \end{aligned} \tag{5.13}$$

We say that (X_n, L_n^+, L_n^-) solves the Skorokhod problem with respect to the set $\mathcal{M}_n$. Following [137, Definition 1.2], the strong solution (X_n, L_n^+, L_n^-) of the Skorokhod problem exists and is unique if $n^2 \nabla R_n$ and Σ_n are Lipschitz with respect to $\|\cdot\|_{\mathrm{F}}$ (following Assumption 3, Assumption 5 and equation (5.5)).

5.2.3 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 [137, Corollary 1.6], i.e.,

$$\sup_{t\in[0,T]} |X_1(t) - X_2(t)| \le 4 \sup_{t\in[0,T]} |Y_1(t) - Y_2(t)|, \qquad \forall\, T \in \mathbb{R}_+. \tag{5.14}$$

5.3 Convergence of Projected Noisy Stochastic Gradient Descent

The goal of this section is to show that for each $n \in \mathbb{N}$, the projected noisy SGD iterates, defined in (PNSGD), converges weakly to the strong solution of the SDE (RSDE) as $|\boldsymbol{\tau}_n| \to 0$. This is done in two steps that we describe below.

Recall the projected noisy SGD iterates defined in Definition 5.1.2, starting from $W_{n,0} \in \mathcal{M}_n$, rewritten for convenience:

$$W_{n,k+1} = P\Big(W_{n,k} - n^2 \tau_{n,k} \nabla R_n(W_{n,k}) - \tau_{n,k} \Delta M_{n,k} + \tau_{n,k}^{1/2} G_{n,k}\Big), \tag{PNSGD}$$

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 Section 5.1, and

$$\Delta M_{n,k} \coloneqq n^2 g_n(W_{n,k}; \xi_{k+1}) - n^2 \nabla R_n(W_{n,k}), \qquad k \in \mathbb{Z}_+.$$

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 $\mathcal{F}_k \coloneqq \sigma\big(\{W_{n,0}, \xi_{i+1}, G_{n,i}\}_{i\in\{0\}\cup[k-1]} \cup \{\xi_{k+1}\}\big)$ 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, $(V_{n,k})_{k\in\mathbb{Z}_+}$ starting at $V_{n,0} = W_{n,0}$, described in (PNGD), re-written below

$$V_{n,k+1} = P\Big(V_{n,k} - n^2\tau_{n,k}\nabla R_n(V_{n,k}) + \tau_{n,k}^{1/2} G_{n,k}\Big), \qquad k \in \mathbb{Z}_+. \tag{PNGD}$$

Let $W_k^{(n)} \coloneqq K(W_{n,k})$ and $V_k^{(n)} \coloneqq K(V_{n,k})$ 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 5.3.1 that the two processes are close as $|\boldsymbol{\tau}_n| \to 0$.

Lemma 5.3.1. *Let $R\colon \mathcal{W} \to \mathbb{R}$ be such that the Fréchet-like derivative $\phi = D_{\mathcal{W}}R$ exists. Suppose Assumptions 3, and 4 hold. Let $n \in \mathbb{N}$. Let W_n and V_n be the piecewise constant interpolations (see Definition (5.2.1)) of $(W_{n,k})_{k\in\mathbb{Z}_+}$ and $(V_{n,k})_{k\in\mathbb{Z}_+}$ 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*

$$\mathbb{E}\left[\sup_{s\in[0,T]} \left\|W^{(n)}(s) - V^{(n)}(s)\right\|_2^2\right] \le C\sigma^2 T|\boldsymbol{\tau}_n| \exp\big[C\kappa_2^2 T^2\big].$$

Proof. Let W_n and V_n be the piecewise constant interpolations of $(W_{n,j})_{j\in\mathbb{Z}_+}$ and $(V_{n,j})_{j\in\mathbb{Z}_+}$ respectively as defined in Definition 5.2.1. Define $\Delta\colon \mathbb{R}_+ \to \mathbb{R}_+$ as

$$\Delta(t) \coloneqq \mathbb{E}\left[\sup_{s\in[0,t]} \|W_n(s) - V_n(s)\|_{\mathrm{F}}^2\right], \qquad t \in \mathbb{R}_+. \tag{5.15}$$

Let $k \in \mathbb{Z}_+$ be such that $t \in [t_{n,k}, t_{n,k+1})$. Then, using [198, Theorem 1],

$$\begin{aligned}\Delta(t) \leq C\mathbb{E}\left[\left(\sum_{j=0}^{k-1} \tau_{n,j} \left\|n^2 \nabla R_n(W_{n,j}) - n^2 \nabla R_n(V_{n,j})\right\|_{\mathrm{F}}\right)^2\right] \\ + C\mathbb{E}\left[\sum_{j=0}^{k-1} \tau_{n,j}^2 \|\Delta M_{n,j}\|_{\mathrm{F}}^2\right],\end{aligned} \tag{5.16}$$

where $C > 0$ is some universal constant. From Assumption 3, 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 equation (5.5) 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

$$\left\|n^2 \nabla R_n(A_n) - n^2 \nabla R_n(B_n)\right\|_{\mathrm{F}}^2 \leq \kappa_2^2 \|A_n - B_n\|_{\mathrm{F}}^2, \qquad \forall\ A_n, B_n \in \mathcal{M}_n. \tag{5.17}$$

Using the Cauchy-Schwarz inequality, and equation (5.17), we first bound the second term in equation (5.16) as

$$\begin{aligned}&\mathbb{E}\left[\left(\sum_{j=0}^{k-1} \tau_{n,j} \left\|n^2 \nabla R_n(W_{n,j}) - n^2 \nabla R_n(V_{n,j})\right\|_{\mathrm{F}}\right)^2\right] \\ &\leq \mathbb{E}\left[\sum_{j=0}^{k-1} \left(\tau_{n,j}^{1/2}\right)^2 \cdot \sum_{j=0}^{k-1} \tau_{n,j} \left\|n^2 \nabla R_n(W_{n,j}) - n^2 \nabla R_n(V_{n,j})\right\|_{\mathrm{F}}^2\right] \\ &\leq \kappa_2^2 t \mathbb{E}\left[\sum_{j=0}^{k-1} \tau_{n,j} \|W_{n,j} - V_{n,j}\|_{\mathrm{F}}^2\right] \leq \kappa_2^2 t \int_0^t \Delta(s)\,\mathrm{d}s,\end{aligned} \tag{5.18}$$

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

$$\mathbb{E}\Big[\|W_n(s) - V_n(s)\|_{\mathrm{F}}^2\Big] = \mathbb{E}\Big[\|W_{n,j} - V_{n,j}\|_{\mathrm{F}}^2\Big] \leq \Delta(s).$$

Using Assumption 4, first note that

$$\begin{aligned}\|\Delta M_{n,j}\|_{\mathrm{F}}^2 &= \left\|n^2 g_n(W_{n,k}; \xi_{k+1}) - n^2 \nabla R_n(W_{n,k})\right\|_{\mathrm{F}}^2 \\ &= n^2 \left\|K\big(n^2 g_n(W_{n,k}; \xi_{k+1}) - n^2 \nabla R_n(W_{n,k})\big)\right\|_2^2 \leq n^2 \sigma^2.\end{aligned} \tag{5.19}$$

We use the above to bound the first term in equation (5.16) as

$$\mathbb{E}\left[\sum_{j=0}^{k-1} \tau_{n,j}^2 \|\Delta M_{n,j}\|_{\mathrm{F}}^2\right] \leq n^2 \sigma^2 t |\boldsymbol{\tau}_n|, \tag{5.20}$$

where $|\boldsymbol{\tau}_n|$ is defined in Section 5.1 as $\sup_{j\in\mathbb{Z}_+}\tau_{n,j}$.

Plugging back (5.18) and (5.20) in equation (5.16) we get

$$\Delta(t) \leq Cn^2\sigma^2 t|\boldsymbol{\tau}_n| + C\kappa_2^2 t\int_0^t \Delta(s)\,\mathrm{d}s, \tag{5.21}$$

and applying Grönwall's inequality [100], we obtain $\Delta(t) \leq Cn^2\sigma^2 t|\boldsymbol{\tau}_n|\exp\left[C\kappa_2^2 t^2\right]$. □

Our next step is to show that the sequence of iterates defined in (PNGD) is close to the solution of the SDE (RSDE) which we reproduce below

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

where B_n is an $n\times n$ symmetric matrix-valued process whose entries are independent Brownian motions up to matrix symmetry, and $X_n(0) = V_{n,0} = W_{n,0} \in \mathcal{M}_n$. The tuple (X_n, L_n^+, L_n^-) solves the Skorokhod problem with respect to the set $\mathcal{M}_n$ (see Section 5.2.2).

In Lemma 5.3.2 we compare (PNGD) with a discretization of the SDE (RSDE). 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 5.3.2. *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 $X_n(0) = V_{n,0}$ (see (PNGD)). Then, there exists a càdlàg process $\widetilde{V}_n$ on $\mathcal{M}_n$, defined on the same probability space as B_n, such that it has the same law as V_n, the piecewise constant interpolation (see Definition 5.2.1) of $\left(V_{n,k}\right)_{k\in\mathbb{Z}_+}$ obtained from (PNGD). Moreover, for any $T\in\mathbb{R}_+$,*

$$\lim_{|\boldsymbol{\tau}_n|\to 0}\mathbb{E}\left[\sup_{s\in[0,T]}\left\|K(X_n(s)) - K\left(\widetilde{V}_n(s)\right)\right\|_2^2\right] = 0.$$

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 5), 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 $\left(\widetilde{Z}_{n,k}\right)_{k\in\mathbb{Z}_+}$ as

$$\widetilde{Z}_{n,k} \coloneqq \tau_{n,k}^{-1/2}(B_n(t_{n,k+1}) - B_n(t_{n,k})), \qquad k \in \mathbb{Z}_+. \tag{5.22}$$

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 $\widetilde{V}_{n,0} = V_{n,0}$, we now define an auxiliary process $\left(\widetilde{V}_{n,k}\right)_{k\in\mathbb{Z}_+}$, on the same probability space as B_n, iteratively as

$$\widetilde{V}_{n,k+1} = P\Big(\widetilde{V}_{n,k} - n^2\tau_{n,k}\nabla R_n\Big(\widetilde{V}_{n,k}\Big) + \tau_{n,k}^{1/2}\Sigma_n\Big(\widetilde{V}_{n,k}\Big)\circ\widetilde{Z}_{n,k}\Big), \qquad k \in \mathbb{Z}_+, \tag{5.23}$$

Following Assumption 5, $\widetilde{V}_{n,k}$ has the same law as $V_{n,k}$ for each $k \in \mathbb{Z}_+$. Let $\widetilde{V}_n\colon \mathbb{R}_+ \to \mathcal{M}_n$ be piecewise constant interpolation of $\left(\widetilde{V}_{n,k}\right)_{k\in\mathbb{Z}_+}$. The particular choice of $\left(\widetilde{Z}_{n,k}\right)_{k\in\mathbb{Z}_+}$ in equation (5.22) allows us to couple $\widetilde{V}_n$ with the strong solution of the SDE (RSDE). Let $\widetilde{G}_{n,j} \coloneqq \Sigma_n\Big(\widetilde{V}_{n,j}\Big)\circ\widetilde{Z}_{n,j}$ for all $j \in \mathbb{Z}_+$. The curve $\widetilde{V}_n$ can be written as

$$\widetilde{V}_n(t) = \widetilde{V}_{n,0} - \sum_{j=0}^{k-1} n^2\tau_{n,j}\nabla R_n(\widetilde{V}_{n,j}) + \sum_{j=0}^{k-1}\tau_{n,j}^{1/2}\widetilde{G}_{n,j} + \sum_{j=0}^{k-1}\tau_{n,j}\Big(L_{n,j}^- - L_{n,j}^+\Big), \tag{5.24}$$

for $t \in [t_{n,k}, t_{n,k+1})$. Here $\left(L_{n,j}^{\pm}\right)_{j\in\mathbb{Z}_+}$ is chosen so that the piecewise constant interpolation (see Definition 5.2.1) of $\left(V_{n,k}, L_{n,k}^-, L_{n,k}^+\right)_{k\in\mathbb{Z}_+}$ solves the Skorokhod problem with respect to $\mathcal{M}_n$ (see Section 5.2.2).

Also consider three auxiliary processes Y_n, $\overline{Y}_n$, and $\widehat{Y}_n$ taking values over $n\times n$ real symmetric matrices, defined as

$$Y_n(t) \coloneqq X_n(0) - \int_0^t n^2\nabla R_n(X_n(s))\,\mathrm{d}s + \int_0^t \Sigma_n(X_n(s))\circ\mathrm{d}B_n(s), \tag{5.25}$$

$$\widehat{Y}_n(t) \coloneqq X_n(0) - \int_0^t n^2\nabla R_n\Big(\widetilde{V}_n(s)\Big)\,\mathrm{d}s + \int_0^t \Sigma_n\Big(\widetilde{V}_n(s)\Big)\circ\mathrm{d}B_n(s), \tag{5.26}$$

$$\overline{Y}_n(t) \coloneqq X_n(0) - \sum_{j=0}^{k-1} n^2\tau_{n,j}\nabla R_n(\widetilde{V}_{n,j}) + \sum_{j=0}^{k-1}\tau_{n,j}^{1/2}\widetilde{G}_{n,j}, \tag{5.27}$$

for every $k \in \mathbb{Z}_+$ and all $t \in [t_{n,k}, t_{n,k+1})$. Observe that the curves X_n and $\widetilde{V}_n$ can be obtained by applying the Skorokhod map to the curves Y_n and $\overline{Y}_n$ pointwise respectively. Let $\widehat{V}_n\colon \mathbb{R}_+ \to \mathcal{M}_n$ 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, Assumption 5, Section 5.2.2 and equation (5.17)), we obtain

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\widehat{V}_n(t)-X_n(t)\right\|_{\mathrm{F}}^2\right] &\leq 16\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\widehat{Y}_n(t)-Y_n(t)\right\|_{\mathrm{F}}^2\right] \\
&\leq 16\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\int_0^t n^2\nabla R_n(X_n(s))-n^2\nabla R_n\left(\widetilde{V}_n(s)\right)\mathrm{d}s\right\|_{\mathrm{F}}^2\right] \\
&\quad + 16\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\int_0^t\left(\Sigma_n(X_n(s))-\Sigma_n\left(\widetilde{V}_n(s)\right)\right)\circ\mathrm{d}B_n(s)\right\|_{\mathrm{F}}^2\right] \\
&\leq 16\kappa_2^2\mathbb{E}\left[\int_0^T\left\|X_n(s)-\widetilde{V}_n(s)\right\|_{\mathrm{F}}^2\mathrm{d}s\right] \\
&\quad + 64\mathbb{E}\left[\int_0^T\left\|\Sigma_n(X_n(s))-\Sigma_n\left(\widetilde{V}_n(s)\right)\right\|_{\mathrm{F}}^2\mathrm{d}s\right] \\
&\leq 80\kappa_2^2\int_0^T\mathbb{E}\left[\sup_{s\in[0,t]}\left\|X_n(s)-\widetilde{V}_n(s)\right\|_{\mathrm{F}}^2\right]\mathrm{d}s, \qquad (5.28)
\end{aligned}$$

where the second last inequality follows from Doob's maximal inequality [129, 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 $k_t := \arg\min_{j\in\mathbb{Z}_+}\{t \geq t_{n,j}\}$. Using the Lipschitzness of Skorokhod map (see Section 5.2.2) we obtain

$$\begin{aligned}
\mathbb{E}\left[\sup_{s\in[0,T]}\left\|\widetilde{V}_n(t)-\widehat{V}_n(t)\right\|_{\mathrm{F}}^2\right] &\leq 16\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\overline{Y}_n(t)-\widehat{Y}_n(t)\right\|_{\mathrm{F}}^2\right] \\
&\leq 32\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\int_0^t n^2\nabla R_n\left(\widetilde{V}_n(s)\right)\mathrm{d}s-\sum_{j=0}^{k_t-1}n^2\tau_{n,j}\nabla R_n\left(\widetilde{V}_{n,j}\right)\right\|_{\mathrm{F}}^2\right] \\
&\quad + 32\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\sum_{j=0}^{k_t-1}\tau_{n,j}^{1/2}\Sigma_n\left(\widetilde{V}_{n,j}\right)\circ\widetilde{Z}_{n,j}-\int_0^t\Sigma_n\left(\widetilde{V}_n(s)\right)\circ\mathrm{d}B_n(s)\right\|_{\mathrm{F}}^2\right], \qquad (5.29)
\end{aligned}$$

where the last inequality follows from Assumption 5.

We now bound the first term from the above inequality (5.29). To this end observe that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\int_0^t n^2\nabla R_n\Big(\widetilde{V}_n(s)\Big)\,\mathrm{d}s-\sum_{j=0}^{k_t-1}n^2\tau_{n,j}\nabla R_n\Big(\widetilde{V}_{n,j}\Big)\right\|_{\mathrm{F}}^2\right]\\
&=\mathbb{E}\left[\sup_{t\in[0,T]}\left\|n^2(t-t_{n,k_t})\nabla R_n\Big(\widetilde{V}_{n,k}\Big)\right\|_{\mathrm{F}}^2\right]\leq|\boldsymbol{\tau}_n|^2\mathbb{E}\left[\sup_{t\in[0,T]}\left\|n^2\nabla R_n\Big(\widetilde{V}_{n,k}\Big)\right\|_{\mathrm{F}}^2\right]\\
&=n^2|\boldsymbol{\tau}_n|^2\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\phi\Big(\widetilde{V}^{(n)}(t)\Big)\right\|_2^2\right]\leq n^2|\boldsymbol{\tau}_n|^2M_2^2,
\end{aligned}\tag{5.30}$$

for some constant $M_2\in\mathbb{R}_+$ by Assumption 3.

We now bound the second term in the inequality (5.29). Using the coupling defined in (5.22) and noting that $\widetilde{V}(s)=\widetilde{V}_{n,j}$ for $s\in[t_{n,j},t_{n,j+1})$ (see Definition 5.2.1), we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\sum_{j=0}^{k_t-1}\tau_{n,j}^{1/2}\Sigma_n\Big(\widetilde{V}_{n,j}\Big)\circ\widetilde{Z}_{n,j}-\int_0^t\Sigma_n\Big(\widetilde{V}_n(s)\Big)\circ\mathrm{d}B_n(s)\right\|_{\mathrm{F}}^2\right]\\
&=\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\Sigma_n\Big(\widetilde{V}_{n,k_t}\Big)\circ(B_n(t)-B_n(t_{n,k_t}))\right\|_{\mathrm{F}}^2\right]\leq M_\infty^2n^2C_{1,T}|\boldsymbol{\tau}_n|\log\frac{1}{|\boldsymbol{\tau}_n|},
\end{aligned}\tag{5.31}$$

where the last inequality follows from Assumption 5 and [199, Lemma A.4] for $C_{1,T}\in\mathbb{R}_+$.

Now define $\Delta\colon\mathbb{R}_+\to\mathbb{R}_+$ as

$$\Delta(t)\coloneqq\mathbb{E}\left[\sup_{s\in[0,t]}\left\|X_n(s)-\widetilde{V}_n(s)\right\|_{\mathrm{F}}^2\right],\qquad t\in\mathbb{R}_+.$$

Using the triangle inequality by combining equations (5.28), (5.29), (5.30) and (5.31), we get

$$\Delta(T)\leq32n^2|\boldsymbol{\tau}_n|^2M_2^2+32n^2M_\infty^2C_{1,T}|\boldsymbol{\tau}_n|\log\frac{1}{|\boldsymbol{\tau}_n|}+80\kappa_2^2\int_0^T\Delta(t)\,\mathrm{d}t.\tag{5.32}$$

Applying Grönwall's inequality [100], we get

$$\Delta(T)\leq32n^2\left(|\boldsymbol{\tau}_n|^2M_2^2+M_\infty^2C_{1,T}|\boldsymbol{\tau}_n|\log\frac{1}{|\boldsymbol{\tau}_n|}\right)\exp\big[80\kappa_2^2T\big].\tag{5.33}$$

Taking limit as $|\boldsymbol{\tau}_n|\to0$ on the above bound, completes the proof. □

We combine Lemma 5.3.1 and 5.3.2 to conclude the proof of Theorem 5.1.3. Moreover, we also the following non-asymptotic error rate

$$\mathbb{E}\left[\sup_{s\in[0,T]}\left\|W^{(n)}(s)-K(X_n)(s)\right\|_2^2\right]\leq Cn^2(M+\sigma^2T)|\boldsymbol{\tau}_n|\log\frac{1}{|\boldsymbol{\tau}_n|}\exp\big[C\kappa_2^2T\big]$$

for some constants $C, M < \infty$.

5.3.1 Convergence of Projected Stochastic Gradient Descent

In the absence of "large noise" (i.e., when $\Sigma_n \equiv 0$), the SDE (RSDE) reduces to the SDE

$$\mathrm{d}X_n(t) = -n^2 \nabla R_n(X_n(t))\,\mathrm{d}t + \mathrm{d}L_n^-(t) - \mathrm{d}L_n^+(t), \qquad X_n(0) = W_{n,0}, \tag{5.34}$$

As we describe in Section 5.1.1, it is show in [167, Theorem 4.4, Theorem 4.14] that if the solution of

$$\mathrm{d}X_n(t) = -n^2 \nabla R_n(X_n(t)) \circ \mathbb{1}_{G_n(X_n(t))}\,\mathrm{d}t, \tag{5.35}$$

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

$$\begin{aligned} G_n(A) \coloneqq &\left\{(i,j) \in [n]^2 \;\middle|\; |A(i,j)| < 1\right\} \\ &\cup \left\{(i,j) \in [n]^2 \;\middle|\; A(i,j) = 1, \partial_{i,j} R_n(A) > 0\right\} \\ &\cup \left\{(i,j) \in [n]^2 \;\middle|\; A(i,j) = -1, \partial_{i,j} R_n(A) < 0\right\}, \end{aligned} \tag{5.36}$$

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 (5.34) and (5.35) are equal. To this end, we define processes $L_n^{\pm}$ as

$$\begin{aligned} L_n^+(t) &\coloneqq -\int_0^t n^2 \nabla R_n(X_n(s)) \circ \mathbb{1}_{\{X_n(s)=+1, \nabla R_n(X_n(s))<0\}}\,\mathrm{d}s, \\ L_n^-(t) &\coloneqq +\int_0^t n^2 \nabla R_n(X_n(s)) \circ \mathbb{1}_{\{X_n(s)=-1, \nabla R_n(X_n(s))>0\}}\,\mathrm{d}s, \end{aligned} \tag{5.37}$$

for $t \in \mathbb{R}_+$, and equation (5.35) can be rewritten as

$$\mathrm{d}X_n(t) = -n^2 \nabla R_n(X_n(t)) \circ \mathbb{1}_{G_n(X_n(t))} + \mathrm{d}L_n^-(t) - \mathrm{d}L_n^+(t), \tag{5.38}$$

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 $(i,j) \in [n]^2$,

$$\int_0^\infty \mathbb{1}\{X_{n,(i,j)}(t) > -1\}\, \mathrm{d}L^-_{n,(i,j)}(t) = 0, \quad \text{and}$$
$$\int_0^\infty \mathbb{1}\{X_{n,(i,j)}(t) < +1\}\, \mathrm{d}L^+_{n,(i,j)}(t) = 0.$$

Following Section 5.2.2, 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 $\mathcal{M}_n$. This proves Theorem 5.1.7.

5.4 The limit at infinity: infinite exchangeable array of diffusions

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 $\big(\zeta_{i,j} \coloneqq \zeta_{\{i,j\}} \in \mathcal{E}\big)_{(i,j)\in\mathbb{N}^{(2)}} =: \zeta$ 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 [4, 108, 109] 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 $\zeta_{i,j} = f(U, U_i, U_j, U_{i,j}) = f(U, U_j, U_i, U_{i,j})$ for $(i,j) \in \mathbb{N}^{(2)}$, where U, $(U_i)_{i\in\mathbb{N}}$, $\big(U_{i,j} = U_{\{i,j\}}\big)_{(i,j)\in\mathbb{N}^{(2)}}$ are i.i.d. Uni$[0,1]$ random variables. The function f is typically not unique. Following [13], we say that ζ is directed by f.

The relationship between exchangeable arrays and graphons follows from the Aldous-Hoover representation [74]. 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 5.4.1. *Let $\zeta \in [-1,1]^{\mathbb{N}^{(2)}}$ be an infinite exchangeable array directed by f. Consider the family of symmetric kernels $(g_u,\ u \in [0,1])$ defined by*

$$g_u(x,y) \coloneqq \mathbb{E}[f(u,x,y,V)], \qquad u \in [0,1], \quad (x,y) \in [0,1]^{(2)}, \tag{5.39}$$

where the above expectation is with respect to a Uni$[0,1]$ *random variable* V*. Then, for* $u \in [0,1]$*, given* $\{U = u\}$*,*

$$\lim_{n\to\infty} \delta_{\square}\left(K\Big((\zeta_{i,j} = f(u, U_i, U_j, U_{i,j}))_{(i,j)\in[n]^{(2)}}\Big), [g_u]\right) = 0, \qquad a.s. \tag{5.40}$$

Proof. 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,

$$\begin{aligned} g_u(x,y) &= \mathbb{E}[f(U, U_i, U_j, U_{i,j}) \,|\, U = u, U_i = x, U_j = y] \\ &= \mathbb{E}[f(U, U_j, U_i, U_{i,j}) \,|\, U = u, U_i = x, U_j = y] = g_u(y,x), \end{aligned}$$

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, $[g_U]$ is also a random variable obtained as a composition of measurable maps.

To see (5.40), 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,

$$\begin{aligned} h_F\left(K\Big((\zeta_{i,j})_{(i,j)\in[n]^{(2)}}\Big)\right) &= \frac{1}{n^{\downarrow k}} \sum_{i_1,i_2,\ldots,i_k} \prod_{\{j,l\}\in E(F)} \zeta_{i_j i_l} \\ &= \frac{1}{n^{\downarrow k}} \sum_{i_1,i_2,\ldots,i_k} \prod_{\{j,l\}\in E(F)} f(u, U_{i_j}, U_{i_l}, U_{i_j,i_l}), \end{aligned} \tag{5.41}$$

where the summation runs over the $n^{\downarrow k} := n!/(n-k)!$ many injections from $[k]$ to $[n]$, and $h_F \colon \mathcal{W} \to \mathbb{R}$ is the homomorphism density function of F [150, Section 7.2]. Notice that

$$\mathbb{E}\Big[h_F\left(K\Big((\zeta_{i,j})_{(i,j)\in[n]^{(2)}}\Big)\right)\Big] = \int_{[0,1]^k} \prod_{\{j,l\}\in E(F)} \mathbb{E}[f(u, u_j, u_l, V)]\, \mathrm{d}u_1 \cdots \mathrm{d}u_k = h_F(g_u),$$

where g_u is defined in (5.39). 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 [125, Theorem 4.18]. We skip the standard argument. The conclusion holds following the inverse counting lemma [150, Lemma 10.32]. □

Remark 5.4.2. *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 $[g_U]$ 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) $\{B_{i,j} = B_{\{i,j\}}\}_{(i,j)\in\mathbb{N}^{(2)}}$. 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 $\mathcal{F}_0$.

Let ϕ and Σ be two functions from $\mathcal{W}$ to $L^\infty\big([0,1]^{(2)}\big)$ that are both κ_2-Lipschitz functions on kernels with respect to the the L^2 norm $\|\cdot\|_2$ (Assumption 3 and 5). Our goal is to construct, on the above probability space with filtration $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$, an exchangeable array of reflected diffusions satisfying

$$\mathrm{d}X_{i,j}(t) = -\phi(\Gamma(t))(U_i,U_j)\,\mathrm{d}t + \Sigma\left(\Gamma(t)\right)(U_i,U_j)\,\mathrm{d}B_{i,j}(t) + \mathrm{d}L^-_{i,j}(t) - \mathrm{d}L^+_{i,j}(t), \tag{5.42}$$

with the initial condition $X_{i,j}(0) = W_0(U_i,U_j)$ for all $(i,j)\in\mathbb{N}^{(2)}$, for some $W_0\in\mathcal{W}$ and

$$\Gamma(t)(x,y) = \mathbb{E}[X_{1,2}(t)\mid U_1 = x, U_2 = y].$$

We construct a diffusion with more general drift as follows. Let $b\colon [-1,1]\times\mathcal{W}\to L^\infty\big([0,1]^{(2)}\big)$ be satisfy Assumption 7. Given $W_0\in\mathcal{W}$, let $X := \big(X_{i,j} := X_{\{i,j\}}\big)_{(i,j)\in\mathbb{N}^{(2)}}$, 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

$$\begin{aligned}\mathrm{d}X_{i,j}(t) = b(X_{i,j}(t),\Gamma(t))(U_i,U_j)\,\mathrm{d}t + \Sigma\left(\Gamma(t)\right)(U_i,U_j)\,\mathrm{d}B_{i,j}(t)\\ + \mathrm{d}L^-_{i,j}(t) - \mathrm{d}L^+_{i,j}(t),\end{aligned} \tag{5.43}$$

for $(i,j)\in\mathbb{N}^{(2)}$ and $t\in\mathbb{R}_+$. The processes $L^-_{i,j}$ and $L^+_{i,j}$ are such that $(X_{i,j},L^+_{i,j},L^-_{i,j})$ solves the Skorokhod problem with respect to $[-1,1]$ (see Section 5.2.2), i.e., $L^-_{i,j}$ and $L^+_{i,j}$

are non-decreasing processes that keep the processes $X_{i,j}$s in the closed interval $[-1,1]$. 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 $(B_{i,j})_{(i,j)\in\mathbb{N}^{(2)}}$, and given by

$$\Gamma(t)(x,y) \coloneqq \mathbb{E}[X_{1,2}(t) \mid U_1 = x, U_2 = y], \tag{5.44}$$

for $(x,y) \in [0,1]^{(2)}$ and $t \in \mathbb{R}_+$. Note that if the solution X of the system of SDEs (5.43) exists, then conditioned over the sigma algebra $\mathcal{F}_0$, the coordinate processes of X are all independent but not necessarily identically distributed. In particular, taking $b(z,W)(x,y) = -\phi(W)(x,y)$, we recover the system of diffusions in (5.42).

It is not obvious if an infinite-dimensional stochastic process satisfying (5.43) and (5.44) exists, although it is obvious that such a process, if it exists, will be an infinite exchangeable array taking values in $\mathcal{E} = C[0,\infty)$. In the rest of this section, under Assumption 7 we show that the process (X,Γ) is indeed well-defined. As will be made clear in Proposition 5.4.6, 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.

Assumption 7. *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* $\sup_{z\in[-1,1]}\|b(z,W_1) - b(z,W_2)\|_2 \le \kappa\|W_1 - W_2\|_2$.

Observe that Assumption 7 implies Assumption 3(2) for $\kappa_2^2 = 2(L^2+\kappa^2)$ and that $\|b(z,W)\|_\infty \le C$ uniformly over all $z \in [-1,1]$ and $W \in \mathcal{W}$.

To argue about the existence of a unique solution of the system of SDEs (5.43), 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 5.4.3 below) and, $\Gamma^{(k)}$ is a deterministic process of kernels (Lemma 5.4.4). 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:

$$\begin{aligned} \mathrm{d}X^{(k+1)}_{i,j}(t) = b\Big(X^{(k)}_{i,j}(t), \Gamma^{(k)}(t)\Big)(U_i, U_j)\,\mathrm{d}t + \Sigma\left(\Gamma^{(k)}(t)\right)(U_i, U_j)\,\mathrm{d}B_{i,j}(t) \\ + \mathrm{d}L^{(k+1)-}_{i,j}(t) - \mathrm{d}L^{(k+1)+}_{i,j}(t), \end{aligned} \tag{5.45}$$

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 5.2.2) for every $(i,j) \in \mathbb{N}^{(2)}$. Since the drift and diffusion functions ϕ and Σ are deterministic and Lipschitz (Assumption 3), 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 5.4.3) as described in equation (5.39) in Lemma 5.4.1, and is therefore defined as

$$\Gamma^{(k+1)}(t)(x,y) \coloneqq \mathbb{E}\Big[X^{(k+1)}_{1,2}(t) \,\Big|\, U_1 = x, U_2 = y\Big], \qquad t \in \mathbb{R}_+. \tag{5.46}$$

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 5.4.3. *Suppose that, for some $k \in \mathbb{Z}_+$, there is a unique in law solution to the SDE (5.45) 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 $\big(X^{(k+1)}_{i,j}\big)_{(i,j)\in\mathbb{N}^{(2)}}$ has the same law as $\big(X^{(k+1)}_{\sigma_i,\sigma_j}\big)_{(i,j)\in\mathbb{N}^{(2)}}$ in the sense of equality of the two probability measures on $(C[0,\infty))^{\mathbb{N}^{(2)}}$.

Let $\widetilde{U}_i := U_{\sigma_i}$, 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 $Y^{(k+1)}_{i,j} \equiv X^{(k+1)}_{\sigma_i,\sigma_j}$ for every $(i,j) \in \mathbb{N}^{(2)}$. Since $Y^{(k+1)}_{i,j}(0) = W_0(U_{\sigma_i}, U_{\sigma_j}) =: W_0(\widetilde{U}_i, \widetilde{U}_j)$. It follows that $\big(Y^{(k+1)}_{i,j}(0)\big)_{(i,j)\in\mathbb{N}^{(2)}}$ has the same distribution as $\big(X^{(k+1)}_{i,j}(0)\big)_{(i,j)\in\mathbb{N}^{(2)}}$. Moreover for every $(i,j) \in \mathbb{N}^{(2)}$, the process $Y^{(k+1)}$ satisfies the SDEs

$$\begin{aligned}
\mathrm{d}Y^{(k+1)}_{i,j}(t) &= b\Big(X^{(k)}_{\sigma_i,\sigma_j}(t), \Gamma^{(k)}(t)\Big)(U_{\sigma_i}, U_{\sigma_j})\,\mathrm{d}t + \Sigma\Big(\Gamma^{(k)}(t))(U_{\sigma_i}, U_{\sigma_j}\Big)\,\mathrm{d}B_{\sigma_i,\sigma_j}(t) \\
&\qquad + \mathrm{d}L^{(k+1)-}_{\sigma_i,\sigma_j}(t) - \mathrm{d}L^{(k+1)+}_{\sigma_i,\sigma_j}(t) \\
&= b\Big(Y^{(k)}_{i,j}(t), \Gamma^{(k)}(t)\Big)(\widetilde{U}_i, \widetilde{U}_j)\,\mathrm{d}t + \Sigma\big(\Gamma^{(k)}(t))(\widetilde{U}_i, \widetilde{U}_j\big)\,\mathrm{d}B_{\sigma_i,\sigma_j}(t) \\
&\qquad + \mathrm{d}L^{(k+1)-}_{\sigma_i,\sigma_j}(t) - \mathrm{d}L^{(k+1)+}_{\sigma_i,\sigma_j}(t),
\end{aligned}$$

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 $\widetilde{B}_{i,j} := B_{\sigma_i,\sigma_j}$, $\widetilde{L}^{(k+1)-}_{i,j} := L^{(k+1)-}_{\sigma_i,\sigma_j}$ and $\widetilde{L}^{(k+1)+}_{i,j} := L^{(k+1)+}_{\sigma_i,\sigma_j}$ for every $(i,j) \in \mathbb{N}^{(2)}$, leaves their joint law unchanged, and we get

$$\begin{aligned}
\mathrm{d}Y^{(k+1)}_{i,j}(t) = b\Big(Y^{(k)}_{i,j}(t), \Gamma^{(k)}(t)\Big)(\widetilde{U}_i, \widetilde{U}_j)\,\mathrm{d}t + \Sigma\Big(\Gamma^{(k)}(t)\Big)(\widetilde{U}_i, \widetilde{U}_j)\,\mathrm{d}\widetilde{B}_{i,j}(t) \\
+ \mathrm{d}\widetilde{L}^{(k+1)-}_{i,j}(t) - \mathrm{d}\widetilde{L}^{(k+1)+}_{i,j}(t),
\end{aligned}$$

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 (5.45), their equivalence in law follows from the uniqueness in law of the SDE. □

Lemma 5.4.4. *Under the same assumption as in Lemma 5.4.3 and Assumption 7, the kernel-valued map $t \mapsto \Gamma^{(k)}(t)$, is deterministic and absolutely continuous. Moreover, for each $t \in \mathbb{R}_+$, we have*

$$\lim_{n\to\infty} \delta_{\square}\left(\left[K\left(\left(X^{(k)}_{i,j}(t)\right)_{(i,j)\in[n]^{(2)}}\right)\right], \left[\Gamma^{(k)}(t)\right]\right) = 0, \qquad a.s. \tag{5.47}$$

Proof. By definition, for $(x,y) \in [0,1]^{(2)}$, and $t \in \mathbb{R}_+$, $\Gamma^{(k)}(t)(x,y) := \mathbb{E}\Big[X_{1,2}^{(k)}(t) \,\Big|\, U_1 = x, U_2 = y\Big]$. This is a deterministic kernel for every $t \in \mathbb{R}_+$. To see (5.47), repeat the proof of Lemma 5.4.1. Notice that, there is no random variable U as in Lemma 5.4.1 (also see Remark 5.4.2). This is now a consequence of Kolmogorov's zero-one law [125, 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 $\mathcal{T} := \cap_{n\in\mathbb{N}} \vee_{\ell \geq n} \mathcal{G}_\ell$. 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 $\lim_{n\to\infty} h_F\big(K\big(\big(X_{i,j}^{(k)}(t)\big)_{(i,j)\in[n]^{(2)}}\big)\big)$, as in equation (5.41). 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 5.4.1, 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 $\big[\Gamma^{(k)}(t)\big]$ 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 5.4.5. *Assume that the drift functions $b\colon [-1,1]\times\mathcal{W} \to L^\infty\big([0,1]^{(2)}\big)$ satisfies Assumption 7, 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 5). Then the sequence of processes taking values in $C\left([0,\infty), [-1,1]\times\mathcal{W}\right)$ 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 $t \in \mathbb{R}_+$,*

$$\lim_{k\to\infty} \sup_{s\in[0,t]} \left[\left| X_{1,2}^{(k)}(s) - X_{1,2}(s) \right|^2 + \left\| \Gamma^{(k)}(s) - \Gamma(s) \right\|_2^2 \right] = 0, \qquad a.s. \tag{5.48}$$

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 [129, 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 [137] pathwise to the "noise before reflection" process $Y_{1,2}^{(k+1)}$ obtained as the unique strong solution to the SDE:

$$\mathrm{d}Y_{1,2}^{(k+1)}(t) = b\Big(X_{1,2}^{(k)}(t), \Gamma^{(k)}(t)\Big)(U_1,U_2)\,\mathrm{d}t + \Sigma\left(\Gamma^{(k)}(t)\right)(U_1,U_2)\,\mathrm{d}B_{1,2}(t), \tag{5.49}$$

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 $\sup_{s\in[0,t]}\left|X_{1,2}^{(k+1)}(s) - X_{1,2}^{(k)}(s)\right|$ for any $k \in \mathbb{N}$. Since the Skorokhod map is 4-Lipschitz in the local uniform norm (see Section 5.2.2), the above distance is bounded by $4\sup_{s\in[0,t]}\left|Y_{1,2}^{(k+1)}(s) - Y_{1,2}^{(k)}(s)\right|$. Now for every fixed $k \in \mathbb{N}$, from equation (5.49) we have

$$\begin{aligned}
&Y_{1,2}^{(k+1)}(t) - Y_{1,2}^{(k)}(t) \\
&= \int_0^t \Big(b\Big(X_{1,2}^{(k-1)}(t), \Gamma^{(k-1)}(t)\Big)(U_1,U_2) - b\Big(X_{1,2}^{(k)}(t), \Gamma^{(k)}(t)\Big)(U_1,U_2)\Big)\,\mathrm{d}s \\
&\quad - \int_0^t \Big(\Sigma\Big(\Gamma^{(k-1)}\Big)(U_1,U_2) - \Sigma\Big(\Gamma^{(k)}\Big)(U_1,U_2)\Big)\,\mathrm{d}B_{1,2}(s).
\end{aligned} \tag{5.50}$$

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 $\|A\|_2^2 = \mathbb{E}\big[A^2(U_1,U_2)\big]$, for U_1, U_2 i.i.d. as $\mathrm{Uni}[0,1]$. Using Jensen's inequality and interchanging expectation with integral and As-

sumption 7,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{s\in[0,t]}\Delta^2(s)\right]\\
&\leq t\mathbb{E}\left[\int_0^t\left|b\Big(X_{1,2}^{(k-1)}(t),\Gamma^{(k-1)}(t)\Big)(U_1,U_2)-b\Big(X_{1,2}^{(k)}(t),\Gamma^{(k)}(t)\Big)(U_1,U_2)\right|^2\mathrm{d}s\right]\\
&=t\int_0^t\left\|b\Big(X_{1,2}^{(k-1)}(t),\Gamma^{(k-1)}(t)\Big)-b\Big(X_{1,2}^{(k)}(t),\Gamma^{(k)}(t)\Big)\right\|_2^2\mathrm{d}s\\
&\leq 2\kappa^2t\int_0^t\left\|\Gamma^{(k-1)}(s)-\Gamma^{(k)}(s)\right\|_2^2\mathrm{d}s+2L^2t\int_0^t\mathbb{E}\left[\left|X^{(k-1)}(s)-X^{(k)}(s)\right|^2\right]\mathrm{d}s. \qquad (5.51)
\end{aligned}$$

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 [129, Theorem 3.8.iv, page 14], we get that,

$$\mathbb{E}\left[\sup_{s\in[0,t]}M^2(s)\right]\leq 4\int_0^t\mathbb{E}\left[\left|\Sigma\Big(\Gamma^{(k-1)}(s)\Big)(U_1,U_2)-\Sigma\Big(\Gamma^{(k)}(s)\Big)(U_1,U_2)\right|^2\right]\mathrm{d}s.$$

Using the assumption that Σ is κ_2-Lipschitz in $\|\cdot\|_2$ and the same argument as above,

$$\mathbb{E}\left[\sup_{s\in[0,t]}M^2(s)\right]\leq 4\kappa_2^2\int_0^t\left\|\Gamma^{(k-1)}(s)-\Gamma^{(k)}(s)\right\|_2^2\mathrm{d}s. \qquad (5.52)$$

Now, taking absolute values on both sides on (5.50), we immediately get,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{s\in[0,t]}\left|X_{1,2}^{(k+1)}(s)-X_{1,2}^{(k)}(s)\right|^2\right]\\
\leq&16\mathbb{E}\left[\sup_{s\in[0,t]}\left|Y_{1,2}^{(k+1)}(s)-Y_{1,2}^{(k)}(s)\right|^2\right]\leq 32\mathbb{E}\left[\sup_{s\in[0,t]}\Delta^2(s)+\sup_{s\in[0,t]}M^2(s)\right]\\
\leq&64(\kappa^2t+2\kappa_2^2)\int_0^t\left\|\Gamma^{(k-1)}(s)-\Gamma^{(k)}(s)\right\|_2^2\mathrm{d}s\\
&+64L^2t\int_0^t\mathbb{E}\left[\left|X^{(k-1)}(s)-X^{(k)}(s)\right|^2\right]\mathrm{d}s. \qquad (5.53)
\end{aligned}$$

Using the fact that the operator Γ, given by a conditional expectation (5.46), 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

$$\mathbb{E}\left[\sup_{s\in[0,t]}\left|X_{1,2}^{(k+1)}(s)-X_{1,2}^{(k)}(s)\right|^2+\sup_{s\in[0,t]}\left\|\Gamma^{(k+1)}(s)-\Gamma^{(k)}(s)\right\|_2^2\right]$$
$$\leq 128((\kappa^2+L^2)t+4\kappa_2^2)\int_0^t\mathbb{E}\left[\left|X^{(k-1)}(s)-X^{(k)}(s)\right|^2\right]\mathrm{d}s.$$

The rest of the argument follows exactly as in [129, page 290] by applications of Grönwall's lemma [100] and the Borel-Cantelli lemma [125, Theorem 4.18]. We skip the similar argument for pathwise uniqueness. See the proof of [129, Proposition 2.13, page 291]. □

Proposition 5.4.6. *Suppose the assumptions in Proposition 5.4.5 holds. Given any kernel $W_0 \in \mathcal{W}$, there exists a pathwise unique strong solution to the coupled system (5.43) and (5.44) 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 (5.43) and (5.44) 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 $\left(\left[K\left((X_{i,j}(t))_{(i,j)\in[n]^2}\right)\right]\right)_{n\in\mathbb{N}}$ locally uniformly in time. That is, for any $t\in\mathbb{R}_+$,

$$\lim_{n\to\infty}\sup_{s\in[0,t]}\delta_\square\left(\left[K\left((X_{i,j}(s))_{(i,j)\in[n]^{(2)}}\right)\right],[\Gamma(s)]\right)=0,\qquad a.s. \tag{5.54}$$

Proof. 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 5.4.5.

Given Γ and $(U_i)_{i\in\mathbb{N}}$ and following the system of SDEs (5.43), 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 5.4.5 and satisfies the constraint (5.43) 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 in the sense of Section 5.4 with $\mathcal{E}=C[0,\infty)$, the set of continuous functions from $[0,\infty)$ to $\mathbb{R}$.

For the limit (5.54) we will make use of the following result from [150, Proposition 8.12], which states that for any $V \in \mathcal{W}$,

$$\|V\|_{\square}^{4} \leq h_{C_4}(V) \leq 4\|V\|_{\square}. \tag{5.55}$$

Here C_4 is the cyclic graph with four vertices and $h_{C_4}(V)$ is the homomorphism density function of the simple graph C_4. We will apply this for the choice of $V_n(t) := K\left((X_{i,j}(t))_{(i,j)\in[n]^2}\right) - K\Big((\Gamma(t)(U_i,U_j))_{(i,j)\in[n]^2}\Big)$. Thus,

$$\begin{aligned} H_n(t) := h_{C_4}(V_n(t)) &= \frac{1}{n^{\downarrow 4}} \sum_{i_1,i_2,\ldots,i_4} \prod_{l=1}^{4} \big(X_{i_l,i_{l+1}}(t) - \Gamma(t)(U_{i_l},U_{i_{l+1}})\big) \\ &= \frac{1}{n^{\downarrow 4}} \sum_{i_1,i_2,\ldots,i_4} \prod_{l=1}^{4} \big(X_{i_l,i_{l+1}}(t) - \mathbb{E}\big[X_{i_l,i_{l+1}}(t) \,\big|\, \mathcal{F}_0\big]\big), \end{aligned}$$

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 5.4.1. 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 $s_1, s_2 \in [0,t]$,

$$\begin{aligned} &|H_n(s_2) - H_n(s_1)| \\ &\leq 16\left\|K\Big((X_{i,j}(s_2))_{(i,j)\in[n]^{(2)}}\Big) - K\Big((X_{i,j}(s_1))_{(i,j)\in[n]^{(2)}}\Big)\right\|_2 \\ &\qquad + 16\|\Gamma(s_2) - \Gamma(s_1)\|_2, \end{aligned} \tag{5.56}$$

where the inequality follows by an application of the counting lemma [150, 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 (5.14)), we therefore obtain

$$\begin{aligned}
&\left\| K\Big((X_{i,j}(s_2))_{(i,j)\in[n]^{(2)}}\Big) - K\Big((X_{i,j}(s_1))_{(i,j)\in[n]^{(2)}}\Big)\right\|_2^2 \\
&\leq \frac{2^4}{n^2}\sum_{(i,j)\in[n]^{(2)}} |Y_{i,j}(s_2) - Y_{i,j}(s_1)|^2 \\
&\leq \frac{2^5}{n^2}\sum_{(i,j)\in[n]^{(2)}} \left|\int_{s_1}^{s_2} b(X_{1,j}(u),\Gamma(u))(U_i,U_j)\,\mathrm{d}u\right|^2 \\
&\qquad + \frac{2^5}{n^2}\sum_{(i,j)\in[n]^{(2)}} \left|\int_{s_1}^{s_2} \Sigma(\Gamma(u))(U_i,U_j)\,\mathrm{d}B_{i,j}(u)\right|^2 \\
&\leq 2^5 M_\infty^2 |s_2 - s_1|^2 + \frac{2^5}{n^2}\sum_{(i,j)\in[n]^2} \left|\int_{s_1}^{s_2} \Sigma(\Gamma(u))(U_i,U_j)\,\mathrm{d}B_{i,j}(u)\right|^2. \qquad (5.57)
\end{aligned}$$

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 [199, Lemma A.4], there exist constants $C_{1,t}, C_{2,t} \in \mathbb{R}_+$ depending of t, such that for all $(i,j) \in [n]^{(2)}$,

$$\mathbb{E}[\eta_{i,j}] \leq M_\infty^2 C_{1,t}\delta\left|\log\frac{1}{\delta}\right|, \qquad \text{and} \qquad \mathbb{E}\big[\eta_{i,j}^2\big] \leq M_\infty^4 C_{2,t}^2\delta^2\log^2\frac{1}{\delta}. \qquad (5.58)$$

Since, $\eta_{i,j}$s are independent and have finite variance, it follows from the Chebyshev's inequality [125, 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 [125, Theorem 4.18], it follows that almost surely,

$$\frac{1}{n^2}\sum_{(i,j)\in[n]^{(2)}} \eta_{i,j} \leq M_\infty^2(C_{1,t}+C_{2,t})\delta\left|\log\frac{1}{\delta}\right|, \qquad (5.59)$$

for all $n \in \mathbb{N}$, sufficiently large. Combining equations (5.56) and (5.59), 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 [160, Theorem 47.1]. □

Proposition 5.4.7. *Suppose that $\Sigma \equiv \beta > 0$ and $b(z, W) = -\phi(W)$. Then, the limiting curve Γ in Proposition 5.4.6 has a velocity*

$$\dot{\Gamma}(t) = -\phi(\Gamma(t)) - \Big[p^{(+1)}_{\beta^2 t}(W_0, \phi\circ\Gamma, \beta) - p^{(-1)}_{\beta^2 t}(W_0, \phi\circ\Gamma, \beta)\Big], \tag{5.60}$$

where $p^{(\pm1)}_s(W_0, \phi\circ\Gamma, \beta)(x, y)$ 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 5.2.2).

Proof. 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 (5.44) and Itô's formula, we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\Gamma(t)(x, y) = &-\frac{\mathrm{d}}{\mathrm{d}t}\phi(\Gamma(t))(x, y)\\ &+ \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\Big[L^-_{1,2}(t)\,\Big|\,U_1 = x, U_2 = y\Big] - \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\Big[L^+_{1,2}(t)\,\Big|\,U_1 = x, U_2 = y\Big].\end{aligned} \tag{5.61}$$

Now consider the reflecting diffusion Z which solves the SDE

$$\mathrm{d}Z(s) = \Psi(s;\beta)\,\mathrm{d}s + \mathrm{d}B(s) + \mathrm{d}L^-(s) - \mathrm{d}L^+(s), \qquad s \in \mathbb{R}_+, \tag{5.62}$$

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 $\Psi(s;\beta) \coloneqq -\frac{1}{\beta^2}b\big(\Gamma(s/\beta^2)\big)(x, y)$ for all $s \in \mathbb{R}_+$ (see Section 5.2.2). 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

$$\begin{aligned}&\mathrm{d}Z(\beta^2 t) = -\frac{1}{\beta^2}\phi(\Gamma(t))(x, y)\,\mathrm{d}(\beta^2 t) + \mathrm{d}B(\beta^2 t) + \mathrm{d}L^-(\beta^2 t) - \mathrm{d}L^+(\beta^2 t),\\ \implies\ &X_{1,2}(t) = -\phi(\Gamma(t))\,\mathrm{d}t + \beta\,\mathrm{d}B(t) + \mathrm{d}L^-(\beta^2 t) - \mathrm{d}L^+(\beta^2 t), \qquad t \in \mathbb{R}_+,\end{aligned}$$

where the processes $(L^+(\beta^2 t))_{t\in\mathbb{R}_+}$ and $(L^-(\beta^2 t))_{t\in\mathbb{R}_+}$ constrain the process $X_{1,2}$ in the interval $[-1,1]$ (see Section 5.2.2). 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 $B(\beta^2 t)$ and $\beta B(t)$ coincide for all $\beta \in \mathbb{R}_+$. Let $p_s^{(\pm 1)}(W_0, \phi \circ \Gamma, \beta)(x,y)$ denote the transition density of the solution of SDE (5.62) 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 $p_{\beta^2 t}^{(\pm 1)}(W_0, \phi \circ \Gamma, \beta)(x,y)$.

Using [184, Exercise (1.12), page 407] and equation (5.61), we deduce that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\Big[L_{i,j}^{\pm}(t)\Big] = p_{\beta^2 t}^{(\pm 1)}(W_0, b \circ (X_{1,2}, \Gamma), \beta)(x,y), \tag{5.63}$$

which gives us the desired result. □

Remark 5.4.8. *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 (5.60), is qualitatively different from that when $\beta = 0$ (see Section 5.1.1).*

5.5 Convergence of the finite-dimensional processes

Consider now the finite dimensional SDE (RSDE):

$$\mathrm{d}X_n(t) = -n^2 \nabla R_n(X_n(t))\,\mathrm{d}t + \Sigma_n(X_n(t)) \circ \mathrm{d}B_n(t) + \mathrm{d}L_n^-(t) - \mathrm{d}L_n^+(t). \tag{5.64}$$

The Fréchet-like derivative of R is a symmetric kernel-valued map from $\mathcal{W} \to L^\infty\big([0,1]^{(2)}\big)$. Thus, for $(x,y) \in [0,1]^{(2)}$, there is a real-valued map $\phi_{x,y} \colon \mathcal{W} \to \mathbb{R}$ given by $\phi_{x,y}(V) = \phi(V)(x,y)$ for all $V \in \mathcal{W}$. This is the same map that we get when we replace (x,y) by (y,x). 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 8. *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*

$$\sup_{(x,y)\in[0,1]^2}\sup_{z\in[-1,1]}|b(z,W_1)(x,y)-b(z,W_2)(x,y)|\le\kappa_\square\|W_1-W_2\|_\square,\qquad \textit{and}$$

$$\sup_{(x,y)\in[0,1]^2}|\Sigma(W_1)(x,y)-\Sigma(W_2)(x,y)|\le\kappa_\square\|W_1-W_2\|_\square.$$

Proposition 5.5.1. *Suppose the assumptions in Proposition 5.4.5 and Assumption 8 hold. 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*

$$\lim_{n\to\infty}\Big\|W_0^{(n)}-W_0\Big\|_2=0, \tag{5.65}$$

the process of random kernels $(K(X_n(t)))_{t\in\mathbb{R}_+}$ *obtained from solutions of the SDEs (5.64), 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 Proposition 5.4.6.*

Proof. Consider a probability space satisfying the assumptions of Proposition 5.4.6 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 $X^{(k)}(\cdot):=\big(X_{(i,j)}\big)_{(i,j)\in[k]^{(2)}}$.

Now consider $K(X_n(t))$ from a solution of SDEs (5.64). 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 $E_k^c(n)$ 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 $\mathbb{P}\{E_k^c(n)\}\le\mathbb{P}\big\{\min_{i\in[k-1]}\big(U_{(i+1)}-U_{(i)}\big)\le\frac{1}{n}\big\}$. 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 (5.43). Since (5.64) 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

$$\begin{aligned}&48\int_0^t \Big|b(X_{1,2}(s), \Gamma(s))(u_1, u_2) - b\Big(\widetilde{X}_{n,1,2}(s), K(X_n(s))\Big)(u_1, u_2)\Big|^2\,\mathrm{d}s\\ &\quad + 48\sup_{s\in[0,t]}\Big|\int_0^s (\Sigma(\Gamma(r))(u_1, u_2) - \Sigma(K(X_n(r)))(u_1, u_2))\,\mathrm{d}B_{1,2}(r)\Big|^2\\ &\quad + 48\Big|\widetilde{X}_{n,1,2}(0) - X_{1,2}(0)\Big|^2.\end{aligned} \tag{5.66}$$

We can now use Assumption 7 and 8 on the first term in (5.66) to get

$$\begin{aligned}&\Big|b(X_{1,2}(s), \Gamma(s))(u_1, u_2) - b\Big(\widetilde{X}_{n,1,2}(s), K(X_n(t))\Big)(u_1, u_2)\Big|^2\\ &\le 2L^2\Big|X_{1,2}(s) - \widetilde{X}_{n,1,2}(s)\Big|^2 + 2\kappa_\square^2\|\Gamma(s) - K(X_n(s))\|_\square^2, \qquad s \in \mathbb{R}_+.\end{aligned} \tag{5.67}$$

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 (5.66) equal to $48 \sup_{s\in[0,t]} M^2(s)$. Using Markov's inequality followed by Doob's maximal inequality [129, page 14, Theorem 3.8.iv], we obtain

$$\begin{aligned}\mathbb{P}\left\{\sup_{s\in[0,t]} M^{(n)}(s)^2 \geq 2\lambda_k \mathbb{E}\left[M^{(n)}(t)^2\right]\right\} &\leq \left(2\lambda_k \mathbb{E}\left[M^{(n)}(t)^2\right]\right)^{-1} \mathbb{E}\left[\sup_{s\in[0,t]} M^{(n)}(s)^2\right] \\ &\leq \left(2\lambda_k \mathbb{E}\left[M^{(n)}(t)^2\right]\right)^{-1} \mathbb{E}\left[M^{(n)}(t)^2\right] = 2\lambda_k^{-1},\end{aligned} \tag{5.68}$$

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 $1 - 2\lambda_k^{-1}$,

$$\begin{aligned}\sup_{s\in[0,t]} M^{(n)}(s)^2 &\leq 2\lambda_k \mathbb{E}\left[M^{(n)}(t)^2\right] \\ &= 2\lambda_k \int_0^t \left|\Sigma(\Gamma(s))(u_1,u_2) - \Sigma(K(X_n(s)))(u_1,u_2)\right|^2 \mathrm{d}s \\ &\leq 2\lambda_k \kappa_\square^2 \int_0^t \|\Gamma(s) - K(X_n(s))\|_\square^2 \,\mathrm{d}s.\end{aligned} \tag{5.69}$$

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 (5.67) and (5.69) in equation (5.66) we get

$$\begin{aligned}\sup_{s\in[0,t]} \left|\widetilde{X}_{n,1,2}(s) - X_{1,2}(s)\right|^2 &\leq 48\left|W_0^{(n)}(U_1,U_2) - W_0(U_1,U_2)\right|^2 \\ &+ 96\kappa_\square^2(\lambda_k+1)\int_0^t \|\Gamma(s) - K(X_n(s))\|_\square^2 \,\mathrm{d}s \\ &+ 96L^2 \int_0^t \left|X_{1,2}(s) - \widetilde{X}_{n,1,2}(s)\right|^2 \mathrm{d}s.\end{aligned} \tag{5.70}$$

Replacing the role of $(1,2)$ by any other $(i,j) \in [k]^{(2)}$, and summing over, we get

$$\begin{aligned}&\sup_{s\in[0,t]} \frac{1}{k^2} \sum_{(i,j)\in[k]^{(2)}} \left|\widetilde{X}_{n,i,j}(s) - X_{i,j}(s)\right|^2 \\ &\leq \frac{48}{k^2} \sum_{(i,j)\in[k]^{(2)}} \left|W_0^{(n)}(U_i,U_j) - W_0(U_i,U_j)\right|^2 \\ &\quad + 96\kappa_\square^2(\lambda_k+1)\int_0^t \|\Gamma(s) - K(X_n(s))\|_\square^2 \,\mathrm{d}s \\ &\quad + 96L^2 \int_0^t \frac{1}{k^2} \sum_{(i,j)\in[k]^{(2)}} \left|X_{i,j}(s) - \widetilde{X}_{n,i,j}(s)\right|^2 \mathrm{d}s.\end{aligned} \tag{5.71}$$

By the triangle inequality,

$$\begin{aligned}\sup_{s\in[0,t]}\left\|K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right)-K\left((\Gamma(s)(U_i,U_j))_{(i,j)\in[k]^{(2)}}\right)\right\|_{\square}^2 \\ \leq 2\sup_{s\in[0,t]}\left\|K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right)-K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_{\square}^2 \\ +2\sup_{s\in[0,t]}\left\|K\left((\Gamma(s)(U_i,U_j))_{(i,j)\in[k]^{(2)}}\right)-K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_{\square}^2.\end{aligned} \tag{5.72}$$

Then notice that the kernel

$$\frac{1}{2}K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right)-\frac{1}{2}K\left((\Gamma(s)(U_i,U_j))_{(i,j)\in[k]^{(2)}}\right)$$

has entries in $[-1,1]$ and is sampled from the kernel $\frac{1}{2}K(X_n(s))-\frac{1}{2}\Gamma(s)$. By [150, Lemma 10.6], the difference

$$\left\|K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right)-K\left((\Gamma(s)(U_i,U_j))_{(i,j)\in[k]^{(2)}}\right)\right\|_{\square}^2-\|K(X_n(s))-\Gamma(s)\|_{\square}^2$$

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 $n\geq k$. Using this in (5.72) we get

$$\begin{aligned}\sup_{s\in[0,t]}\left\|K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right)-K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_{\square}^2 \\ \geq\frac{1}{2}\|K(X_n(s))-\Gamma(s)\|_{\square}^2-320k^{-1/4} \\ -\sup_{s\in[0,t]}\left\|K\left((\Gamma(s)(U_i,U_j))_{(i,j)\in[k]^{(2)}}\right)-K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_{\square}^2.\end{aligned} \tag{5.73}$$

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 (5.71) using equation (5.73) as

$$\begin{aligned}
&2 \sup_{s\in[0,t]} \left\| K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_2^2 \\
&\quad \geq \sup_{s\in[0,t]} \left\| K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_2^2 \\
&\qquad + \sup_{s\in[0,t]} \left\| K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_\square^2 \\
&\quad \geq \sup_{s\in[0,t]} \left\| K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_2^2 \\
&\qquad + \frac{1}{2}\|K(X_n(s)) - \Gamma(s)\|_\square^2 - 320k^{-1/4} \\
&\qquad - \sup_{s\in[0,t]} \left\| K\left((\Gamma(s)(U_i,U_j))_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_\square^2.
\end{aligned} \tag{5.74}$$

Here we used the fact that the L^2 norm is lower bounded by the cut norm. Using equation (5.74) back in equation (5.71) (multiplied by 2), and rearranging terms we get

$$\begin{aligned}
&\sup_{s\in[0,t]} \left\| K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_2^2 \\
&\qquad + \frac{1}{2}\sup_{s\in[0,t]} \|K(X_n(s)) - \Gamma(s)\|_\square^2 \\
&\quad \leq \sup_{s\in[0,t]} \left\| K\left((\Gamma(s)(U_i,U_j))_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_\square^2 \\
&\qquad + 320k^{-1/4} + \frac{96}{k^2}\sum_{(i,j)\in[k]^{(2)}} \left|W_0^{(n)}(U_i,U_j) - W_0(U_i,U_j)\right|^2 \\
&\qquad + 192L^2 \int_0^t \left\| K\left(\left(\widetilde{X}_{n,i,j}(s)\right)_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_2^2 \mathrm{d}s \\
&\qquad + 192\kappa_\square^2(\lambda_k+1)\int_0^t \|\Gamma(s) - K(X_n(s))\|_\square^2 \,\mathrm{d}s.
\end{aligned} \tag{5.75}$$

Now let

$$\begin{aligned}
A_k &:= \sup_{s\in[0,t]} \left\| K\left((\Gamma(s)(U_i,U_j))_{(i,j)\in[k]^{(2)}}\right) - K\left((X_{i,j}(s))_{(i,j)\in[k]^{(2)}}\right)\right\|_\square^2, \\
B_k(n) &:= \frac{96}{k^2}\sum_{(i,j)\in[k]^{(2)}} \left|W_0^{(n)}(U_i,U_j) - W_0(U_i,U_j)\right|^2 + 320k^{-1/4}.
\end{aligned}$$

Applying Grönwall's inequality [100] and noticing that the first term on the left of

equation (5.75) is always non-negative, gives us that on the event $E_k(n)$,

$$\begin{aligned}\sup_{s\in[0,t]}\Big\|K\Big(\Big(\widetilde{X}_{n,i,j}(s)\Big)_{(i,j)\in[k]^{(2)}}\Big)-K\Big(\big(X_{i,j}(s)\big)_{(i,j)\in[k]^{(2)}}\Big)\Big\|_2^2 \\ +\sup_{s\in[0,t]}\|K(X_n(s))-\Gamma(s)\|_\square^2\le 2\,(A_k+B_k(n))\exp\big(192(L^2+2\kappa_\square^2(\lambda_k+1))t\big),\end{aligned} \tag{5.76}$$

for every $n\ge k$. Note that

$$\mathbb{E}\Big[\Big|W_0^{(n)}(U_i,U_j)-W_0(U_i,U_j)\Big|^2\Big]=\Big\|W_0^{(n)}-W_0\Big\|_2^2\to 0,$$

as $n\to\infty$, by assumption (5.65). 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 5.4.6. Since $\lim_{k\to\infty}\lim_{n\to\infty}\mathbb{P}\{E_k(n)\}=1$,

$$\lim_{n\to\infty}\sup_{s\in[0,t]}\|K(X_n(s))-\Gamma(s)\|_\square=0,\qquad\text{and}$$

$$\lim_{k\to\infty}\lim_{n\to\infty}\sup_{s\in[0,t]}\frac{1}{k^2}\Big\|\big(K(X_n(s))(U_i,U_j)\big)_{(i,j)\in[k]^{(2)}}-\big(X_{i,j}(s)\big)_{(i,j)\in[k]^{(2)}}\Big\|_{\mathrm{F}}^2=0,$$

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 $k\to\infty$. This proves our claim. □

Remark 5.5.2. *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 5.4.6 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 $\lambda_k=\log(k)/\big(16\cdot 384t(L^2+2\kappa_\square^2)\big)$, we have the left hand side of (5.76) bounded by $M_tk^{-1/16}\log^{3/2}k$ with probability at least $1-\frac{k^2}{n}-4k^{-\frac{1}{\kappa^2 t}}-2t\mathrm{e}^{-\sqrt{k}/20}-2k^{-3/2}$, where $\kappa=32\sqrt{6}\big(L^2+2\kappa_\square^2\big)^{1/2}$. 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 5.3.2, we can get a non-asymptotic rate of convergence with finite n and $|\boldsymbol{\tau}_n|$.

5.6 Examples

In this section we will verify our assumptions for a class of functions introduced as linear functions in [167, Section 5.1]. 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) and let R admit a Fréchet-like derivative evaluation map $\phi\colon \mathcal{W} \to L^\infty\big([0,1]^{(2)}\big)$ (see [167, Section 5] for conditions). The map ϕ then satisfies

$$\phi(W)(x, y) = \sum\nolimits_{(i,j)\in[n]^2} \mathbb{E}\Big[\nabla R_n\Big((W(Z_p, Z_q))_{(p,q)\in[n]^{(2)}}\Big)\ \Big|\ (Z_i, Z_j) = (x, y)\Big], \tag{5.77}$$

and $D_{\mathcal{W}}\ell(\,\cdot\,; Z)$ for $Z \in [0,1]^n$ satisfies

$$(D_{\mathcal{W}}\ell(\,\cdot\,; Z))(W)(x, y) = \sum\nolimits_{(i,j)\in[n]^2} \nabla R_n\Big((W(Z_p, Z_q))_{(p,q)\in[n]^{(2)}}\big|_{(Z_i,Z_j)=(x,y)}\Big), \tag{5.78}$$

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

5.6.1 Scalar Entropy and Homomorphism density

Examples like the scalar entropy and the homomorphism density functions considered in the last chapter satisfy Assumption 3 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 $R = H_F$ 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 4 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 8 when R is the sum of scalar entropy and some homomorphism density H_F for a simple graph F with n vertices and m edges. Note that for this example, we have

$$b(z,W)(x,y) = \log\frac{z}{1-z} + \phi_{H_F}(W)(x,y), \qquad z = W(x,y) \in [\epsilon, 1-\epsilon], \tag{5.79}$$

for a.e. $(x,y) \in [0,1]^{(2)}$ where from [167, Equation 113],

$$\begin{aligned}\phi_{H_F}(W)(x,y) &= \sum_{l=1}^{m} \mathbb{E}\left[\prod_{r=1,r\neq l}^{m} W(Z_{e_r}) \,\middle|\, Z_{e_l} = (x,y)\right] \\ &=: \sum_{l=1}^{m} \mathbf{t}_{x,y}(F_{e_l}, W), \quad (x,y) \in [0,1],\end{aligned}$$

$Z_e = (Z_{e(1)}, Z_{e(2)})$ and F_{e_l} is the simple graph obtained from F by removing the edge e_l. It is shown in [167, Section 5.1.2] 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 $\phi_{H_F}(\,\cdot\,)(x,y)$ 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 $\{p,q\} \in E(F_e)$,

$$I_{p,q} \coloneqq \int_{[0,1]^{n-2}} (W_1(x_p,x_q) - W_2(x_p,x_q)) \prod_{(i,j)\in E(F_e)\setminus\{p,q\}} W_1(x_i,x_j) \prod_{v \in V(F_e)\setminus e} \mathrm{d}x_v. \tag{5.80}$$

Following the proof in [150, Lemma 10.24], we get $|I_{p,q}| \le \|W_1 - W_2\|_\square$, which yields

$$|\mathbf{t}_{x,y}(F_e, W_1) - \mathbf{t}_{x,y}(F_e, W_2)| \le (m-1)\|W_1 - W_2\|_\square, \tag{5.81}$$

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 (5.79), we have

$$\begin{aligned}|b(z,W_1)(x,y) - b(z,W_2)(x,y)| &= |\phi_{H_F}(W_1)(x,y) - \phi_{H_F}(W_1)(x,y)| \\ &\le m(m-1)\|W_1 - W_2\|_\square.\end{aligned} \tag{5.82}$$

Therefore b (as in equation (5.79)) satisfies Assumption 8 with $\kappa_\square = m(m-1)$.

5.6.2 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

$$R(W) \coloneqq \frac{1}{2}\sum_{\alpha=1}^{k}(H_{F^\alpha}(W) - c_\alpha)^2.$$

Note that a lower bound on R is achieved if $H_{F^\alpha} \equiv c_\alpha$ for all $\alpha \in [k]$. We note that R being a sum of squares of k many functions satisfies Assumption 3(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

$$\phi(W)(x,y) = \sum_{\alpha=1}^{k}(H_{F^\alpha}(W) - c_\alpha)\phi_{H_{F^\alpha}(W)}(W)(x,y)\ .$$

Note that $W \mapsto \phi_{H_{F^\alpha}}(W)$ satisfies Assumption 3(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 $|H_F(U) - H_F(V)| \le |E(F)|\|U - V\|_\square \le |E(F)|\|U - V\|_2$. A simple calculation using the fact that $|(H_{F^\alpha}(W) - c_\alpha)| \le 1$ for all W and that $\|\phi_{H_F}(W)\|_2 \le |E(F)|$, we obtain that ϕ satisfies Assumption 3(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 $W \mapsto H_F(W)$ is m-Lipschitz in cut norm where m is the number of edges in F. Using the fact that $\|\phi_{H_F}(W)\|_\infty \le m$ and $H_F(W) \in [0,1]$ 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 8.

5.6.3 Entropy minimization with edge-triangle constraints

We conclude with the discussion of the example mentioned in the Introduction. Recall the problem of minimizing the scalar entropy $\mathcal{E}$ over $\widehat{\mathcal{W}_0}$ with prescribed edge density $H_-(\,\cdot\,) =$

$e \in [0,1]$ and triangle density $H_{\triangle}(\,\cdot\,) = \tau \in [0,1]$ (see [167, Section 5.1-5.2]). As mentioned in [163], 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

$$W \mapsto R(W) \coloneqq \frac{1}{2}\left((H_{-}(W) - e)^2 + (H_{\triangle}(W) - \tau)^2\right) + \psi(\mathcal{E}(W)).$$

Since ψ is non-decreasing, minimizing $\mathcal{E}$ is equivalent to minimizing $\psi \circ \mathcal{E}$. On the other hand, the term $\frac{1}{2}\left((H_{-}(W) - e)^2 + (H_{\triangle}(W) - \tau)^2\right)$ penalizes any deviation from the marginal constraint on the edge and triangle densities.

It follows from the previous discussion that $W \mapsto \frac{1}{2}(H_{-}(W) - e)^2 + \frac{1}{2}(H_{\triangle}(W) - \tau)^2$ 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 $\psi = 4\mathrm{id}$ and consider the optimization algorithm described in Definition 5.1.2. 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 4 is satisfied. Theorem 5.1.3 and Theorem 5.1.7 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 [163]. Since R is not strictly convex, the gradient flow may not converge to the minimizer, however, it does converge to a stationary point with a polynomial rate.

5.7 Discussion

Before we move to the next chapter, let us summarize the key insights that we obtain from this chapter. The main goal of this chapter was to study the large n limit of $n \times n$ symmetric matrix-valued processes (see (RSDE))

$$\mathrm{d}X_n(t) = -n^2 \nabla R_n(X_n(t))\,\mathrm{d}t + \beta\,\mathrm{d}B_n(t) + \mathrm{d}L_n^-(t) - \mathrm{d}L_n^+(t), \tag{RSDE}$$

where R_n is an invariant function and B_n denotes the matrix-value process whose coordinates are i.i.d. Brownian motions (upto the symmetry). We obtain such processes as the scaling limit of Euclidean gradient descent (with noise) of appropriate functions. However, as (RSDE) only requires the drift to be an invariant function of the matrix, it is reasonable to expect that such processes will naturally arise in the study of evolutions of graphs. Somewhat similar evolutions have been investigated for instance in [10, 11, 20].

One of the main insights of this chapter (see Theorem 5.1.4) is that under appropriate scaling and invariant function as drift two coordinates of X_n become asymptotically uncorrelated. This significantly reduces the complexity of the limiting process that can be described by an IEA. Such propagation of chaos phenomenon for particle systems has proved to be a very useful tool [157, 170, 171, 52, 51] in probability. We hope that the propagation of chaos phenomenon in the matrix-valued processes will be similarly useful.

Chapter 6

PATH CONVERGENCE OF MARKOV CHAINS ON LARGE GRAPHS

6.1 Introduction

Let $(\Omega, \mathcal{F}, \mathbb{Q})$ be a probability space and let $\mathcal{H}\colon \Omega \to [0, \infty]$ be a measurable function (called the Hamiltonian). We are often interested in the set of minimizers of $\mathcal{H}$. Instead of finding the actual minimizers, which can be computationally expensive, one often considers a Gibbs measure on $(\Omega, \mathcal{F})$ whose density, with respect to $\mathbb{Q}$, is proportional to $\exp(-\beta\mathcal{H}(\cdot))$, for some $\beta > 0$. Here $\mathbb{Q}$ is usually taken to be some kind of a "uniform" probability distribution on Ω. The parameter β is often called the *inverse temperature*. It is well known that, in many circumstances of interest, as $\beta \to \infty$, the Gibbs measure concentrates around the minimizers of $\mathcal{H}$ (see [113]). Thus, one may replace the problem of optimization of $\mathcal{H}$ with 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 [209].

When Ω is continuous with a notion of differentiability of $\mathcal{H}$, such as $\mathbb{R}^d$, a natural stochastic process is the Langevin diffusion:

$$\mathrm{d}X(t) = -\nabla\mathcal{H}(X(t))\,\mathrm{d}t + \sqrt{\frac{2}{\beta}}\,\mathrm{d}B(t),$$

where B is standard d-dimensional Brownian motion and β is commonly called the inverse temperature parameter. In practice, stochastic gradient descent 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 $\mathcal{H}$, namely

$$\dot{x}(t) = -\nabla\mathcal{H}(x(t)),$$

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 an MCMC

algorithm [73], such as the celebrated Metropolis algorithm [185, Section 2.4], to sample from the Gibbs distribution.

For us, Ω is the space of dense edge-weighted unlabeled graphs and we consider natural stochastic processes used to optimize a Hamiltonian $\mathcal{H}$. Examples of such stochastic processes are graph-valued Metropolis Markov chains and processes arising out of optimization algorithms on edge-weights such as stochastic gradient descent (SGD) [187, 134, 24, 141, 41, 159, 142]. A theme that we pursue is that with certain modifications and in a certain limiting regime, as the size of the graph goes to infinity, both processes are related to a gradient flow. More specifically, we analyze a Metropolis Markov chain on a stochastic block model (SBM)(see [107, 210]). The base Markov chain runs on an SBM with r communities, with n members in each community, with an acceptance-rejection step specified by $\mathcal{H}$ 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. When we keep r fixed and let $n \to \infty$, with other parameters suitably scaled, the edge densities between communities converge to a stochastic differential equation on $r \times r$ symmetric matrices (see Section 6.3 for an overview and Proposition 6.3.3 for a precise statement of the result). Further, in Proposition 6.3.4 we prove that, as $r \to \infty$, the paths of the stochastic evolution of the edge density matrices converge to a deterministic curve on the space of *measure-valued graphons* (MVG) (see Chapter 2).

In Section 6.2, we define a general class of diffusions on symmetric $r \times r$ matrices and also show in Theorem 6.2.2 that under suitable assumptions these processes converge, as $r \to \infty$, to a deterministic curve on the space of MVGs specified by a McKean-Vlasov stochastic differential equation (SDE). This strengthens the main result in Chapter 5 where a similar convergence statement had been obtained on the space of graphons as opposed to MVGs. One example, within the purview of Theorem 6.2.2, is the $r \times r$ matrix of pairwise edge-densities from our Metropolis model above. We summarise this result in Proposition 6.3.4. The other kind of example covered by Theorem 6.2.2 is stochastic gradient descent of suitable functions on $r \times r$ symmetric matrices. In the last chapter, we argued that these stochastic gradient descents converge pathwise to the gradient flow on the space of graphons. We show in Proposition 6.3.6 that, as the relaxation parameter $\sigma \to 0+$,

the random trajectory of $r \times r$ random matrices of edge densities between communities generated by the Metropolis algorithm converges to a deterministic curve on graphons which the gradient flow of $\beta\mathcal{H}$ on the space of graphons. Therefore, under suitable convexity assumptions on $\mathcal{H}$, we obtain (see Proposition 6.3.8) an exponential rate of convergence of to the minimizer of $\mathcal{H}$. Combining this with Proposition 6.3.7, which gives non-asymptotic error bounds, we obtain that, in a certain limiting regime, the adjacency matrix from our Metropolis chain converges exponentially fast to the minimizer of $\mathcal{H}$. We introduce the Metropolis chain in Section 6.1.2. A reader familiar with these basics may directly go to Section 6.1.3 where we give a computational example.

6.1.1 Notation recall

We recall some frequently used notations in this chapter. For any set X, we use X^2 to denote the usual Cartesian product $X \times X$ while $X^{(2)}$ is used to denote the set $X^2/\sim$ where we identify $(a, b) \sim (b, a)$ for all $a, b \in X$. We use this notation for domains of symmetric functions.

Let us also denote the set of all $r \times r$ symmetric matrices with elements in $[0, 1]$ and $[-1, 1]$ by $\mathcal{M}_{r,+}$ and $\mathcal{M}_r$ respectively, i.e., $\mathcal{M}_{r,+} := [0, 1]^{[r]^{(2)}}$ and similarly for $\mathcal{M}_r$. Recall that the set $\mathcal{M}_r$ can be naturally identified with a subset of finite dimensional kernels, $\mathcal{W}_r \subset \mathcal{W}$. This identification/embedding is denoted by K (as in $K(A)$ is the kernel corresponding to the matrix A) and its inverse will be denoted by M_r (as in matrix). A simple unweighted graph G can be thought of as a weighted graph with edge-weights being the indicators that the edges exist.

6.1.2 Limits of Markov Processes on Weighted Graphs

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

$$b\colon [-1,1] \times \mathfrak{W} \to L^\infty\big([0,1]^{(2)}\big), \text{ and } \quad \Sigma\colon [-1,1] \times \mathfrak{W} \to L^\infty([0,1]^{(2)}), \tag{6.1}$$

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 $B(\cdot)$ 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

$$\begin{aligned} \mathrm{d}X(t) &= b\left(X(t), W(t)\right)(u, v)\,\mathrm{d}t + \Sigma(X(t), W(t))(u, v)\,\mathrm{d}B(t) \\ &\quad + \mathrm{d}L^-(t) - \mathrm{d}L^+(t), \\ W(t)(x, y) &\coloneqq \mathrm{Law}(X(t) \mid (U, V) = (x, y)), \quad \forall\, (x, y) \in [0,1]^{(2)}, \end{aligned} \tag{6.2}$$

where (X, L^+, L^-) solves the Skorokhod problem [137] with respect to $[-1, 1]$ (see Section 5.2.3 for details). The system described by equation (6.2) will be referred to as the MVG *McKean-Vlasov SDE* (MVSDE). Under appropriate assumptions on b and Σ, Proposition 6.2.1 shows that the MVSDE admits a pathwise unique solution. Notice that W is a deterministic curve on measure-valued kernels, and thus, on measure-valued graphons. A similar McKean-Vlasov SDE was introduced in [102] but the convergence was obtained 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 $w(t) = \mathbb{E}[W(t)]$ given by the natural projection map. It is useful to think of w as capturing the evolution of macroscopic properties while W describes the microscopic properties.

Where do such processes appear? In Section 6.2 we consider a general class of diffusion on symmetric $r \times r$ matrices whose coordinates are exchangeable and are evolving under a suitable mean-field interaction. In Theorem 6.2.2 we prove that processes in this general class have corresponding deterministic limits that are examples of (6.2). This is natural since one can spot that (6.2) 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 stochastic gradient descent of functions defined on the space of graphs. For another example, consider the problem of "soft" optimization described at the very beginning where, for $\beta > 0$, one may wish to consider a Gibbs measure on $\widehat{\mathcal{W}}$ with a "density" proportional to $\exp(-\beta\mathcal{H})$. 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 $r \times r$ graphons $\widehat{\mathcal{W}}_r$. By pulling back the natural map from kernels to graphons and identifying $r \times r$ kernels with $r \times r$ symmetric $[0,1]$-valued matrices, one can think of $\mathcal{H}$ as a function H_r on symmetric matrices, i.e., $H_r = \mathcal{H} \circ K$ on $\mathcal{M}_{r,+}$. One can define a natural Gibbs measure on $\mathcal{M}_{r,+}$ corresponding to H_r. A large class of commonly used models fall in this umbrella. See the thesis [60] for a historical development and some beautiful real-world applications. In particular, it appears in statistical physics models such as the Curie-Weiss models [60, Chapter 4], the exponential random graph models (ERM) [60, 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 nonparametric maximum likelihood estimator such as MCMLE [60, 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 [180, 161, 162, 181] 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 [162, Section 3.1] the authors construct an MCMC scheme to sample from an edge-triangle model. They justify convergence, not theoretically, but empirically. Given a target edge density e and a triangle density t, one may easily construct a Hamiltonian that gets minimized when the edge-density and the triangle densities are e and t, respectively. Then, sampling from this Gibbs measure will approximately sample from an edge-triangle model.

We will show that, for suitable $\mathcal{H}$, satisfying a semiconvexity condition (Assumption 10), the edge density matrix obtained from the Metropolis chain admits limits that are particular cases of (6.2). 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 [167, Section 5.1.2]. 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 $b_0 : [-1,1] \times \mathcal{W} \to\!\!\to L^\infty\big([0,1]^{(2)}\big)$ naturally gives a function $b : [-1,1] \times \mathfrak{W} \to L^\infty\big([0,1]^{(2)}\big)$ via the pullback $b(z,W) = b_0(z,w)$. Such drift functions will naturally arise in our examples. Let $\mathcal{H}$ be a Hamiltonian on $\mathfrak{W}$ that admits a Fréchet-

like derivative $D\mathcal{H}$ (see Definition 4.2.6), and fix a parameter $\beta > 0$. The solution to the McKean-Vlasov SDE (6.2) 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) [80].

Definition 6.1.1 (Empirical Stochastic Block Model (ESBM)). *For $r, n \in \mathbb{N}$ let $q \equiv (q_{i,j})_{1\le i,j\le r} \in \mathcal{M}_{r,+}$, and let $N = rn$. A random simple graph with N vertices is called* ESBM$[r, n, q]$ *if*

- *for $i \in [r]$, there are n many vertices having color i,*
- *for $i, j \in [r]$, $i \neq j$, $n^2 q_{i,j}$ 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 $i \in [r]$, $\binom{n}{2} q_{i,i}$ 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 $(i, j) \in [r]^{(2)}$.*

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

Recall that H_r is the restriction of $\mathcal{H}$ on the space of $r \times r$ symmetric matrices for each $r \in \mathbb{N}$. Fix a positive sequence $(\gamma_n)_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty} \gamma_n \log^2 n = 0, \quad \text{and} \quad \lim_{n\to\infty} \frac{\gamma_n n^2}{\log n} = \infty. \tag{6.3}$$

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

$$\widehat{Q}_{n,r,\beta}(\mathrm{d}q) = \frac{1}{Z_{n,r,\beta}} \mathrm{e}^{-\beta_{n,r} H_r(q)} U_{n,r}(\mathrm{d}q) = \frac{1}{Z_{n,r,\beta}} \mathrm{e}^{-\beta \gamma_n^{-1} r^{-2} H_r(q)} U_{n,r}(\mathrm{d}q).$$

where $Z_{n,r,\beta}$ is the normalizing constant. As each $q \in \mathcal{M}_{r,+}$ corresponds to a simple random graph in ESBM$[r, n, q]$, $\widehat{Q}_{n,r,\beta}$ can be thought of as a random probability distribution on simple graphs over rn vertices. We will denote the model specified by $\widehat{Q}_{n,r,\beta}$, as ESBM$[r, n, \beta, \mathcal{H}]$. It should be emphasized that the measure $\widehat{Q}_{n,r,\beta}$ depends on the choice of the parameter γ_n. Note that the above model closely resembles commonly used framework in exponential random graphs (see [56] and references therein).

The following Metropolis chain algorithm (see [146, Section 3.2]) can be used to sample from ESBM$[r, n, \beta, \mathcal{H}]$.

- *Base Markov Chain*: The state space of the chain is the set $\mathcal{S}_{n,r}$ of all simple graphs on rn vertices with r 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 $(i, j) \in [r]^{(2)}$, let $m_{i,j}(\ell)$ denote the number of edges between vertices of color $i \in [r]$ and color $j \in [r]$ in $G(\ell)$. The next step in the Markov chain is generated as follows.

  - For every $(i, j) \in [r]^{(2)}$, $i \neq j$, if $m_{i,j}(\ell) \notin \{0, n^2\}$, 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 n^2 by $\binom{n}{2}$ if $i = j$.
  - For every $(i, j) \in [r]^{(2)}$, 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 $m_{i,j}(\ell) = n^2$, $i \neq j$ (or $\binom{n}{2}$, 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 $(i, j) \in [r]^{(2)}$.

The resulting graph is $G(\ell+1)$ and $q(\ell+1) = (q_{i,j}(\ell+1))_{1\le i,j\le r}$ 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 $U_{n,r}$.

- *Metropolis Chain*: We run the base chain for $s_n \approx \gamma_n^2 n^4$ 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_n 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 $\exp\left(-\beta_{n,r}\left(\mathcal{H}(q') - \mathcal{H}(q)\right)^+\right)$, otherwise, remain at G. Here $x^+ := \max\{x, 0\}$.

It is standard to see the unique invariant distribution of this Metropolis Markov chain is the Gibbs measure $\widehat{Q}_{n,r,\beta}$. We will explore scaling limits of the chain as $n, r \to \infty$, $\gamma_n, \beta_{n,r}$ as specified above and when $s_n = O(\gamma_n^2 n^4)$. 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 $\ell_{n,r}(\sigma) = O(\sigma^2 r^{-4}\gamma_n n^4)$ many steps, for some $\sigma > 0$, and always accept the last state.

Thus, our final Markov chain repeatedly runs the base chain for s_n many steps, performs an accept-reject step and then runs another $\ell_{n,r}(\sigma)$ many steps of the base chain. We call this the *relaxed Metropolis chain*. Since $\ell_{n,r}(0) = 0$, when $\sigma = 0$, we recover the true Metropolis chain. However, note that the relaxed chain has a different invariant distribution for any positive σ.

Consider the resulting process of $r \times r$ matrix of pairwise edge-densities $q(\cdot)$. One may think of this exchangeable process as a Markov chain on symmetric matrices. In Proposition 6.3.3 and Proposition 6.3.4 we show that, as $n \to \infty$ and $r \to \infty$, and the other parameters are scaled as above, the paths of this process converge to a deterministic curve

on MVG, describe by a McKean-Vlasov SDE (6.2) 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 6.3.5) to a deterministic curve on the space of graphons in the same scaling limit. For a fixed r, as $n \to \infty$ the adjanceny matrix of ESBM$[r, n, q]$ converges to the edge density matrix q in the cut metric. 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 $\sigma > 0$.

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 6.3.6 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 6.3.8 establishes the exponential convergence rate of this flow under appropriate convexity conditions on the Hamiltonian.

6.1.3 A computational example from extremal graph theory: Revisited

To illustrate our results, we give a concrete example with numerical simulations. To motivate our example, we first recall the celebrated Mantel's theorem [155] 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 [32]. One may attempt to computationally verify this theorem by considering a "softer" version of the problem. That is, consider the Hamiltonian $\mathcal{H}(\,\cdot\,) \coloneqq t(\triangle, \cdot\,) - \alpha t(-, \cdot\,)$ for sufficiently small $\alpha > 0$. Here $\triangle$ and $-$ are the triangle and the edge graphs respectively. Recall that the homomorphism density function $t(F, \cdot\,)$ 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 [58]. 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 $\alpha = 1/4$, $n = 16$, $r = 16$, $\sigma = 1$, $\gamma_n = 1/4n$ and $\beta = 1/4$. In particular, $\mathcal{H}(\,\cdot\,) \coloneqq t(\triangle, \cdot\,) - \frac{1}{4}t(-, \cdot\,)$. The Fréchet-like derivative of the Hamiltonian is given by $D\mathcal{H}(w)(x,y) = 3\int_{[0,1]} w(x,z)w(z,y)\,\mathrm{d}z - 1/4$, 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 as $n \to \infty$, followed by $r \to \infty$, and finally $\sigma \to 0$, is given by the a curve $w\colon \mathbb{R}_+ \to \mathcal{W}_{[0,1]}$ given by

$$w(t)(x,y) = w(0)(x,y) - \beta \int_0^t D\mathcal{H}(w(s))(x,y)\mathbb{1}_{G_{w(s)}}\,\mathrm{d}s, \qquad (x,y) \in [0,1]^{(2)}, \tag{6.4}$$

where the starting point $w(0) \in \mathcal{W}_{[0,1]}$ is the L^2-limit of the community edge density kernel of the initialization of the Metropolis chain as $r \to \infty$. The set function G_u 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 w is a time reparametrization of the gradient flow of $\mathcal{H}$. In Figure 6.1 we see that the iteration sequence of the MCMC chain has a close resemblance with the curves shown in [167, Figure 1, Section 1.2] which is a forward Euler discretization 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 $G^{(n)}_{r,3.7\times 10^5}$ is close to the graph the one corresponding to a complete bipartite graph as one would expect from Mantel's theorem.

In Proposition 6.3.8 we show that if the Hamiltonian is strongly convex, the curve w 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 Hamiltonian a multiple of the scalar entropy function [167, Section 5.1.3] that makes it strongly convex enough to guarantee an exponential rate of convergence. That is, for $\gamma > 0$ large enough, the following new Hamiltonian $\mathcal{H}_\gamma$ defined as $\mathcal{H}_\gamma(w) \coloneqq t(\triangle, w) - \frac{1}{4}t(-, w) + \gamma\int_{[0,1]^2} h(w(x,y))\,\mathrm{d}x\,\mathrm{d}y$, where $h(p) = p\log p + (1-p)\log(1-p)$ for $p \in (0,1)$ and

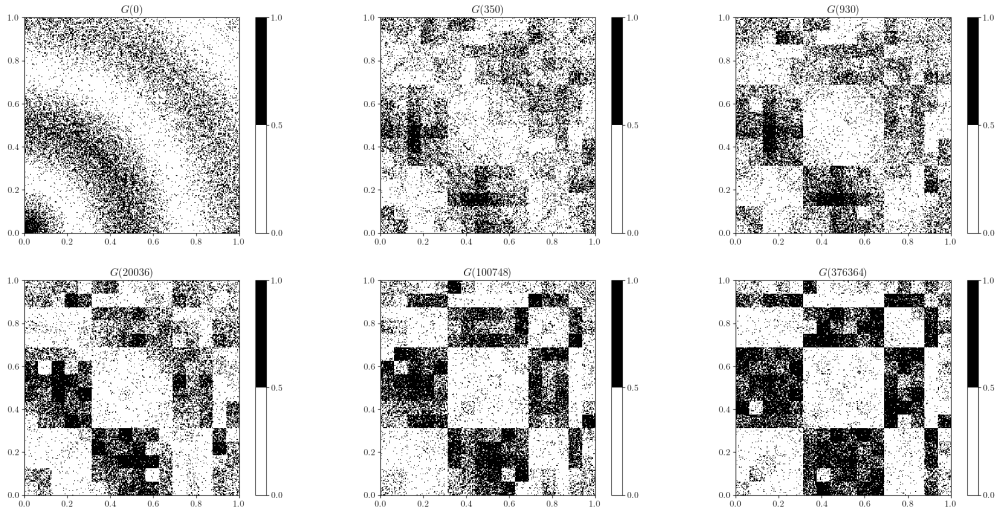


Figure 6.1: A relaxed Metropolis chain algorithm simulation for $\mathcal{H} = t(\triangle, \cdot) - \frac{1}{4}t(-, \cdot)$ at initialization and after 3.5×10^2, 9.3×10^2, 2.0×10^4, 1.0×10^5 and 3.7×10^5 iterations respectively (order: from left to right).

zero if $p \in \{0, 1\}$, is strongly convex and the corresponding gradient flow curve converges exponentially fast. In fact, in this particular example, $\mathcal{H}_\gamma$ as defined above is strongly convex for any $\gamma > 9/2$. In particular, if we set $\gamma = 5$, following Proposition 6.3.8, 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 $\lambda > \gamma - 9/2 = 1/2$. However, to compare our current simulation with those in [167] we do not add the entropy regularization here.

6.2 Dynamics

In this section we study the limit of exchangeable processes on symmetric matrices as the dimension grows to infinity. In Theorem 6.2.2, we show that a general class of processes on symmetric matrices converges to a deterministic curve on $\widehat{\mathfrak{W}}$ that is described by (6.2) as the dimension grows to infinity. In Section 6.3, we study the relaxed Metropolis chain algorithm on SBMs described in Section 6.1.2. We begin with preliminaries on Skorokhod map.

6.2.1 McKean-Vlasov SDE

Recall MVG McKean-Vlasov SDE (6.2) described in Section 6.1. Following a standard Picard's iteration argument, as done in [102, Proposition 4.5], it can be shown that MVG McKean-Vlasov SDE (6.2) admits a pathwise unique solution under appropriate assumptions on b and Σ. For completeness, we record this as Proposition 6.2.1 but skip the proof.

Assumption 9. *Recall the definition of the generalized cut norm, $\|\cdot\|_{\blacksquare}$, from Definition 2.4.3. Let b, Σ be as in (6.1) and satisfy global Lipschitz conditions, that is, there exists $L, \kappa_{\blacksquare} \in \mathbb{R}_+$ such that*

$$\sup_{W\in\mathfrak{W}} \|b(z_1, W) - b(z_2, W)\|_\infty, \sup_{W\in\mathfrak{W}} \|\Sigma(z_1, W) - \Sigma(z_2, W)\|_\infty \le L|z_1 - z_2|,$$
$$\sup_{z\in[-1,1]} \|b(z, W_1) - b(z, W_2)\|_\infty, \sup_{z\in[-1,1]} \|\Sigma(z, W_1) - \Sigma(z, W_2)\|_\infty \le \kappa_{\blacksquare}\|W_1 - W_2\|_{\blacksquare},$$

for all $W_1, W_2 \in \mathfrak{W}$ and $z_1, z_2 \in [-1, 1]$.

Proposition 6.2.1. *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, W) to the MVG McKean-Vlasov SDE (6.2).*

To motivate the study of McKean-Vlasov SDE (6.2), consider the problem of minimizing triangle density while maximizing the edge density as in Section 6.1.3. As explained in Section 6.1.3, a close proxy would be minimize the function $\mathcal{H}(\,\cdot\,) := t(\triangle, \cdot\,) - \alpha t(-, \cdot\,)$ for sufficiently small $\alpha > 0$ over the space of $\widehat{\mathcal{W}}$. Notice that $\mathcal{H}$ naturally restricts to a well-defined function H_r, on $\mathcal{M}_{r,+}$, the space of symmetric matrices with entries in $[0, 1]$. Consider the following SDE on $\mathcal{M}_{r,+}$

$$\mathrm{d}X_{r,(i,j)}(t) = -r^2\nabla H_r(X_r) + \sigma\,\mathrm{d}B_{r,(i,j)}(t) + \mathrm{d}L^-_{r,(i,j)}(t) - \mathrm{d}L^+_{r,(i,j)}(t),$$

with where $\left\{B_{r,(i,j)}\right\}_{(i,j)\in\mathbb{N}^{(2)}}$ is a collection of i.i.d. Brownian motion. Note that the Euclidean gradient ∇H_r is scaled by a factor of r^2. The above SDE can be thought of as a time-scaling of noisy Euclidean gradient flow of H_r. These processes are considered in [102]

and they are indeed obtained as the continuous time limit of *projected noisy stochastic gradient descent* (PNSGD) of H_r [102, Definition 1.2, Theorem 3.2]. It is clear that X_r is a symmetric matrix valued process whose coordinates are exchangeable. It is shown in [102, Theorem 1.4] that, under appropriate assumptions on the initial condition $X_r(0)$, the matrix valued process $(X_r(t))_{t\in\mathbb{R}_+}$ converges to a deterministic curve $(w_\sigma(t))_{t\in\mathbb{R}_+}$ on the space of graphons, uniformly on compact time intervals, as $r\to\infty$. Moreover, the curve $(w_\sigma(t))_{t\in\mathbb{R}_+}$ is described by a McKean-Vlasov SDE similar to (6.2).

Note that the graphon convergence of $X_r(\cdot)$ comes with a loss of microscopic information as discussed earlier in Section 6.1.2. It may be reasonable to expect that the matrix valued process $X_r(\cdot)$ converges to an IEA as $r\to\infty$ and one should rather consider the convergence of $X_r(\cdot)$ to a deterministic curve on MVG space. In this section we accomplish this convergence and do so for a larger class of $\mathcal{H}$ than those allowed in [102, Theorem 1.4]. We begin with an illustrative example.

Example 13. 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 $\mathcal{H}\coloneqq t_{\mathrm{d}}(F,\,\cdot\,)$. Note that $\mathcal{H}$ restricts naturally to a function $H_r\colon \mathcal{M}_r\to\mathbb{R}$ as defined in equation (2.24). One can easily see that $\partial_{(i,j)}t_{\mathrm{d}}(F,\,\cdot\,)(X_r)=\frac{4}{r^3}\sum_{k=1}^r X_{r,(i,k)}^2X_{r,(k,j)}$ for every $(i,j)\in[r]^{(2)}$. Let $B_{r,(i,j)}$ be a collection of i.i.d. Brownian motions. Consider the following SDE on $\mathcal{M}_r$:

$$\mathrm{d}X_{r,(i,j)}(t)=-\frac{4}{r}\sum_{k=1}^{r}X_{r,(i,k)}^2(t)X_{r,(k,j)}(t)\,\mathrm{d}t+\mathrm{d}B_{r,(i,j)}(t)+\mathrm{d}L^-_{r,(i,j)}(t)-\mathrm{d}L^+_{r,(i,j)}(t),$$

where $(i,j)\in[r]^{(2)}$ and (X,L^+,L^-) solves the Skorokhod problem. The above SDE can be recovered as a continuous time limit of the PNSGD algorithm [102, Definition 1.2, Theorem 3.2] when we consider $t_{\mathrm{d}}(F,\,\cdot\,)$ to be the optimization objective. We remark that the function $\mathcal{H}$ does not satisfy the assumptions of [102, Theorem 1.4] as it is not continuous in the cut-metric (see [117, Example C.3]). However, the function $\mathcal{H}$ does satisfy Assumption 9 for b defined as $b(z,W)(x,y)\coloneqq$

$$\begin{aligned}&\sum_{\ell=1}^{m}\mathbb{E}\left[\prod_{s=1}^{\ell-1}\Gamma(F_{e_s},W)(Z_{e_s})\cdot\Gamma(F'_{e_\ell},W)(Z_{e_\ell})\cdot\prod_{s=\ell+1}^{m}\Gamma(F_{e_s},W)(Z_{e_s})\,\middle|\,Z_{e_\ell}=(x,y)\right],\\&=:\sum_{\ell=1}^{m}\mathbf{t}_{x,y}(\partial_{e_\ell}F,W),\qquad z\in[-1,1],W\in\mathfrak{W},(x,y)\in[0,1]^{(2)},\end{aligned}\tag{6.5}$$

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 [102, Example 5] and Lemma 2.4.14.

As a consequence of our main result (Theorem 6.2.2), we will see that the solution of the SDE on $\mathcal{M}_r$ as defined above, converges to the solution X to the MVG McKean-Vlasov SDE (6.2), 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 3. This example, naturally extends to all decorated homomorphism density functions, thereby expanding the scope of [102, Theorem 1.4].

More generally we consider the following family of diffusions on symmetric matrices. Let b and Σ be as defined in Section 6.1. For $r \in \mathbb{N}$, let $\Sigma_r \colon [-1,1] \times \mathcal{M}_r$ be the restrictions of Σ, i.e., $\Sigma_r(x,X) = \Sigma(x, \mathcal{K}(X))$ and similarly define $b_r \colon [-1,1] \times \mathcal{M}_r$. Consider the diffusions X_r defined on $\mathcal{M}_r$ as follows

$$\begin{aligned} \mathrm{d}X_{r,(i,j)}(t) = b_{r,(i,j)}(X_{r,(i,j)}(t), X_r(t))\,\mathrm{d}t + \Sigma_{r,(i,j)}(X_{r,(i,j)}(t), X_r(t))\,\mathrm{d}B_{r,(i,j)}(t) \\ + \mathrm{d}L^-_{r,(i,j)}(t) - \mathrm{d}L^+_{r,(i,j)}(t), \end{aligned} \tag{6.6}$$

for each $(i,j) \in [r]^{(2)}$ and $t \in \mathbb{R}_+$, with the initial condition $X_r(0) \in \mathcal{M}_r$.

In Theorem 6.2.2 we show that under appropriate assumption on $(X_r(0))_{r\in\mathbb{N}}$, the process $X_r(\cdot)$ converges, uniformly on compact intervals of time, to a deterministic curve $W(\cdot)$ on MVGs as $r \to \infty$. And, this curve $W(\cdot)$ is described by the McKean-Vlasov system (6.2).

Theorem 6.2.2. *Suppose Assumption 9 holds. Let $W_0 \in \mathfrak{W}$ and let W be described by the MVG McKean-Vlasov SDE (6.2) with initial condition $W(0) = W_0$. Let X_n be the solution of equation (6.6) for $r = n$ 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,*

$$\lim_{n\to\infty} \sup_{t\in[0,T]} \Delta_{\blacksquare}(\mathcal{K}(X_n(t)), W(t)) = 0. \tag{6.7}$$

The proof of Theorem 6.2.2 closely parallels the proof of [102, Proposition 4.9]. Therefore, we only give a sketch of the proof highlighting only the crucial differences.

Proof. Consider the probability space satisfying Assumption of Proposition 6.2.1 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 $W(t)[k]$ defined as $W(t)[k](i,j) = W(t)(U_i, U_j)$ for $(i,j) \in [k]^{(2)}$. Consider also the corresponding $k \times k$ matrix of diffusions $X[k](\cdot) \coloneqq (X_{i,j})_{(i,j)\in[k]^{(2)}}$. Now consider $\mathcal{K}(X_n(t))$, the measure-valued finite dimensional kernel from a solution of SDE (6.6). 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 $W(t)[k]$ 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

$$M^{(n)}(s) \coloneqq \int_0^s \Big(\Sigma(X_{1,2}(s), W(r))(u_1,u_2) - \Sigma\Big(\widetilde{X}_{n,1,2}(s), \mathcal{K}(X_n(r))\Big)(u_1,u_2)\Big)\,\mathrm{d}B_{1,2}(r),$$

for $s \in [0,t]$. Note that

$$\mathbb{P}\Big\{\sup_{s\in[0,t]} M^{(n)}(s) \geq \sqrt{\lambda_k \mathbb{E}\big[M^{(n)}(t)^2\big]}\Big\} = \mathbb{P}\Big\{\sup_{s\in[0,t]} \exp\Big(uM^{(n)}(s)\Big) \geq \exp(\lambda_k)\Big\},$$

where $u = \sqrt{\lambda_k / \mathbb{E}\big[M^{(n)}(t)^2\big]}$. Using Markov's inequality followed by Doob's maximal inequality [129, page 14, Thoerem 3.8.iv], we obtain that with probability at least $1-4\mathrm{e}^{-\lambda_k/2}$,

$$\sup_{s\in[0,t]} M^{(n)}(s)^2 \le 2\lambda_k \left[\kappa_{\blacksquare}^2 \int_0^t \|W(s) - \mathcal{K}(X_n(s))\|_{\blacksquare}^2 + L^2 \Big|X_{1,2}(s) - \widetilde{X}_{n,1,2}(s)\Big|^2 \,\mathrm{d}s\right], \tag{6.8}$$

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 (6.8) and the Lipschitz property of the Skorokhod map, triangle inequality and Assumption 9 (replacing $(1,2)$ by any other $(i,j) \in [k]^{(2)}$ and summing over),

$$\begin{aligned}\sup_{s\in[0,t]} \Big\|\mathcal{K}\Big(\widetilde{X}_n[k](s)\Big) - \mathcal{K}(X[k](s))\Big\|_{\blacksquare}^2 &\le \frac{48}{k^2} \sum_{(i,j)\in[k]^{(2)}} \Big|\widetilde{X}_{n,i,j}(0) - X_{i,j}(0)\Big|^2 \\ &+ 96(\lambda_k+1)\kappa_{\blacksquare}^2 \int_0^t \|W(s) - \mathcal{K}(X_n(s))\|_{\blacksquare}^2 \,\mathrm{d}s \\ &+ 96(\lambda_k+1)L^2 \int_0^t \sup_{s\in[0,t]} \Big\|\mathcal{K}\Big(\widetilde{X}_n[k](s)\Big) - \mathcal{K}(X[k](s))\Big\|_{\blacksquare}^2 \,\mathrm{d}s.\end{aligned} \tag{6.9}$$

We now want to replace $\Big\|\mathcal{K}\Big(\widetilde{X}_n[k](s)\Big) - \mathcal{K}(W(s)[k])\Big\|_{\blacksquare}^2$ by $\|\mathcal{K}(X_n(s)) - W(s)\|_{\blacksquare}^2$ up to some error that goes to zero as $k \to \infty$. This is achieved by exploiting the first sampling lemma [150, Lemma 10.6] for cut norm in [102]. The first sampling lemma is not available directly to us for the $\|\cdot\|_{\blacksquare}$. However, we notice that using the first sampling lemma [150, Lemma 10.6] and equation (2.31) we obtain that for every $\epsilon > 0$ there exists a constant $F_\epsilon < \infty$ such that

$$\left|\Big\|\mathcal{K}\Big(\widetilde{X}_n[k](s)\Big) - \mathcal{K}(W(s)[k])\Big\|_{\blacksquare}^2 - \|\mathcal{K}(X_n(s)) - W(s)\|_{\blacksquare}^2\right| \le \frac{1}{k^{1/4}} + \epsilon,$$

with probability at least $F_\epsilon \mathrm{e}^{-\sqrt{k}/10}$. Moreover, from Lemma 2.4.15, we can choose $\epsilon_k = \frac{64}{k^{1/4}}$ so that $F_{\epsilon_k} \le \mathrm{e}^{\sqrt{k}/40}$. In particular, setting $C_k = \frac{65}{k^{1/4}}$ and $c_k = \sqrt{k}/20$ we can repeat the same proof as in [102] to obtain

$$\begin{aligned}\sup_{s\in[0,t]} \Big\|\mathcal{K}\Big(\widetilde{X}_n[k](s)\Big) - \mathcal{K}(X[k](s))\Big\|_{\blacksquare}^2 &\ge \frac{1}{2}\|\mathcal{K}(X_n(s)) - W(s)\|_{\blacksquare}^2 - C_k \\ &- \sup_{s\in[0,t]} \|\mathcal{K}(W(s)[k]) - \mathcal{K}(X[k](s))\|_{\blacksquare}^2.\end{aligned} \tag{6.10}$$

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 $\mathbb{P}\{E_k(n)\} \ge 1 - 4\mathrm{e}^{-\lambda_k/2} - \frac{k^2}{n} - \mathrm{e}^{-c_k}$.

We can now repeat the same argument as in [102]. After doing some rearrangement and applying Grönwall's inequality [100] we obtain that on the event $E_k(n)$,

$$\begin{aligned}\sup_{s\in[0,t]} D_2^2\Big(\mathcal{K}\Big(\widetilde{X}_n[k](s)\Big),\mathcal{K}(X[k](s))\Big) + \sup_{s\in[0,t]}\|\mathcal{K}(X_n(s)) - W(s)\|_{\blacksquare}^2 \\ \leq 2\left(A_k + B_k(n)\right)\exp\big(192(L^2+2\kappa_{\blacksquare}^2)(\lambda_k+1)t\big),\end{aligned} \tag{6.11}$$

where $A_k = \sup_{s\in[0,t]}\left\|\widetilde{A}_k(s)\right\|_{\blacksquare}^2$ and

$$\begin{aligned}\widetilde{A}_k(s) &:= \mathcal{K}(W(s)[k]) - \mathcal{K}(X[k](s)),\\ B_k(n) &:= C_k + \frac{96}{k^2}\sum_{(i,j)\in[k]^{(2)}}\left|\widetilde{X}_{n,i,j}(0) - X_{i,j}(0)\right|^2.\end{aligned} \tag{6.12}$$

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 2.3.6 and Lemma 2.3.7 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 [102, Proposition 4.5]) 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 $\lim_{k\to\infty}\lim_{n\to\infty}\mathbb{P}\{E_k(n)\} = 1$,

$$\lim_{n\to\infty}\sup_{s\in[0,t]}\|\mathcal{K}(X_n(s)) - W(s)\|_{\blacksquare} = 0,\ \lim_{k\to\infty}\lim_{n\to\infty}\sup_{s\in[0,t]} D_2^2\Big(\mathcal{K}\Big(\widetilde{X}_n[k](s)\Big),\mathcal{K}(X[k](s))\Big) = 0,$$

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. □

Remark 6.2.3. *Note that the proof is stable with respect to small perturbations of drift. More precisely, suppose b_r in (6.6) is replaced by $\widetilde{b}_r$ such that $\left\|\widetilde{b}_r - b_r\right\|_{\infty} \leq \alpha_r$ for some $\alpha_r \to 0$ as $r\to\infty$. Then, the proof continues to hold and we still obtain the limiting McKean-Vlasov SDE with the same drift b as in Theorem 6.2.2.*

Remark 6.2.4. *Note that the McKean-Vlasov equation (6.2) depends only on $w(t) = \mathbb{E}[W(t)]$. One can, therefore, say that $w(t)$ satisfies a graphon McKean-Vlasov SDE that*

satisfies on $(U, V) = (u, v)$

$$\begin{aligned} \mathrm{d}X(t) &= b_0(X(t), w(t)) + \Sigma_0(X(t), w(t)) + \mathrm{d}L^-(t) - \mathrm{d}L^+(t), \\ w(t)(x, y) &= \mathbb{E}[X(t) \mid (U, V) = (x, y)], \end{aligned} \qquad t \in \mathbb{R}_+. \tag{6.13}$$

Thus we recover [102, Theorem 1.4].

We should emphasize the point made in Remark 2.3.5 once again here. The same IEA gives rise to both McKean-Vlasov SDEs (6.2) and (6.13). The crucial difference is $X_r(\cdot)$ *converging to this IEA in cut-metric is equivalent to only checking the convergence of homomorphism densities with respect to simple graphs, while* $X_r(\cdot)$ *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.*

6.3 Analysis of the relaxed Metropolis chain

Recall that our goal is to minimize some function $\mathcal{H}$ defined on large networks. One class of functions that is of interest is linear combination of homomorphism densities. The key takeaway of the above discussion is that one can either solve for the McKean-Vlasov system described above or perform (stochastic) gradient flow of $\mathcal{H}$ restricted to $\mathcal{M}_r$ for large r. Both of these techniques will yield approximate minimizers of $\mathcal{H}$. However, notice that both these techniques give a graphon/MVG or symmetric matrices with entries in $[-1, 1]$. In other words, these methods do not directly yield a curve on the space of graphs. Naturally, given a function $\mathcal{H}$ defined on all graphs, one would want to run a Markov chain directly on the graphs that mimics the behaviour of McKean-Vlasov SDE (6.2). This is what is achieved by our relaxed Metropolis chain that we study in this section. Another crucial feature of the relaxed Metropolis chain is that it does not need the access to the gradient of $\mathcal{H}$ but it can still mimic the gradient flow.

Assumption 10. *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 4.2.6) and satisfy*

$$\frac{\lambda}{2}\|u - v\|_2^2 \le \mathcal{H}(v) - \mathcal{H}(u) - \langle D\mathcal{H}(u), v - u\rangle \le \frac{L}{2}\|u - v\|_2^2, \tag{6.14}$$

for every $u, v \in \mathcal{W}_{[0,1]}$, for some constants $\lambda \in \mathbb{R}$ and $L > 0$. Furthermore, assume that $D\mathcal{H}$ is Lipschitz, that is, there exists $\kappa_{\blacksquare} > 0$ such that for all $u, v \in \mathcal{W}$

$$\|D\mathcal{H}(u) - D\mathcal{H}(v)\|_\infty \le \kappa_{\blacksquare} \|u - v\|_{\blacksquare}, \tag{6.15}$$

where $\|u - v\|_{\blacksquare}$ is defined as in Remark 2.4.4.

Recall that $W \mapsto \mathbb{E}[W]$ is $(\|\cdot\|_{\blacksquare} \to \|\cdot\|_{\square})$-Lipschitz. In particular, if $w \mapsto D\mathcal{H}(w)$ is $(\|\cdot\|_{\square} \to \|\cdot\|_\infty)$-Lipschitz, then (6.15) holds. All decorated homomorphism density functions satisfy Assumption 10. Using a similar argument as in [167, Section 5.1.2], it can be shown that $\max\{|\lambda|, L\}$ for the decorated homomorphism density function of a decorated graph H is bounded by $|E(H)||V(H)|(|V(H)|-1)$, and $\kappa_{\blacksquare} = |E(H)|(|E(H)|-1)$ following [102, Section 5, equation (84)].

Definition 6.3.1. *Let $\mathcal{H}$ satisfy Assumptions 10. Let $\beta > 0$. Define $b_0 \colon \mathcal{W}_{[0,1]} \to L^\infty\big([0,1]^{(2)}\big)$ as*

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

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

Recall that b_0 defined as above naturally defines a function on $\mathfrak{W}$ via pullback. 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 6.4.1 $r^{-4} b_r(w) = \mathbb{E}\big[Z \exp\big(-\beta_{n,r}\gamma_n \langle \nabla H_r, Z\rangle_F^+\big)\big]$ where Z is $r \times r$ symmetric matrix with i.i.d. Gaussian entries, and $\beta_{n,r} := \beta r^{-2}/\gamma_n$ as defined in Section 6.1.2. Thus $\|b_r\|_\infty < \infty$.

Remark 6.3.2. *It follows from (6.15) that $\|D\mathcal{H}\|_\infty \le C$ for some C 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 $\|b_0(w) + \beta D\mathcal{H}\|_\infty \to 0$ as $r \to \infty$.*

We now recall, from Section 6.1.2, the Metropolis algorithm to sample from the ESBM$[r, n, \beta, \mathcal{H}]$. Given $G(k) \in \mathcal{S}_{n,r}$ and the matrix $q_{r,k}^{(n)} \in \mathcal{M}_{r,+}$ 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 $s_n := \lceil \gamma_n^2 n^4 \rceil$ many steps. Let $\widetilde{G}(k+1) \in \mathcal{S}_{n,r}$ be the graph obtained after such s_n many steps. Let $\widetilde{q}^{(n)}_{r,k+1}$ 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

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

where $a^+ = \max\{0, a\}$ for $a \in \mathbb{R}$ and $\beta_{n,r} := \beta r^{-2}/\gamma_n$ as defined in Section 6.1.2. Let $p^{(n)}_{r,k+1}$ be the matrix of edge-densities for $Y(k+1)$. Observe that

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

3. After the accept-reject step, we again run the base chain starting from $Y(k+1) \in \mathcal{S}_{n,r}$ for $\ell_{n,r} := \lceil r^{-4}\sigma^2\gamma_n n^4 \rceil$ many steps for some $\sigma > 0$. Let the graph obtained thereafter be $G(k+1)$, and let $q^{(n)}_{r,k+1}$ 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 $\mathcal{S}_{n,r}$ with corresponding process of edge-density matrix $\left(q^{(n)}_{r,k}\right)_{k\in\mathbb{N}}$. In the following we show that, as $n \to \infty$ and $r \to \infty$, the latter process converges to a MVG McKean-Vlasov SDE with drift b as defined in Definition 6.3.1. Taking a natural projection of this MVG curve to the space of graphons, we recover a deterministic curve on the space of graphons. For fixed r, the adjacency matrix of $G(k)$ converges to the corresponding edge-density matrix $q^{(n)}_{r,k}$ in the cut metric as $n \to \infty$ uniformly. We can thus 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 $n \to \infty$ followed by $r \to \infty$.

A heuristic analysis of the edge-density process

Before we state our result we analyse heuristically the process of edge-density matrix $\left(q^{(n)}_{r,k}\right)_{k\in\mathbb{Z}_+}$ as defined above. To this end, for any $k \in \mathbb{Z}_+$, define $\Delta q^{(n)}_{r,k} := q^{(n)}_{r,k+1} - q^{(n)}_{r,k}$ and

let $\mathcal{F}_k$ be the sigma algebra generated by $\left\{q_{r,i}^{(n)} \;\middle|\; i = 0, \ldots, k\right\}$. Given $k \in \mathbb{Z}_+$, let us analyze the $\mathbb{E}\left[\Delta q_{r,k}^{(n)} \;\middle|\; \mathcal{F}_k\right]$. Notice that

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

where $\widetilde{\Delta} q_{r,k}^{(n)} \coloneqq p_{r,k+1}^{(n)} - q_{r,k}^{(n)}$. Notice that given $p_{r,k+1}^{(n)}$, the increment of the (i,j)-th coordinate, $q_{r,k+1,(i,j)}^{(n)} - p_{r,k+1,(i,j)}^{(n)}$, has the same distribution as the reflected random walk of step-size $\frac{1}{n^2}$ run for $\ell_{n,r}$ 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 $\mathbb{E}\left[\Delta q_{r,k}^{(n)} \;\middle|\; \mathcal{F}_k\right] \approx \mathbb{E}\left[\widetilde{\Delta} q_{r,k}^{(n)} \;\middle|\; \mathcal{F}_k\right]$

$$= \mathbb{E}\left[\left(\widetilde{q}_{r,k+1}^{(n)} - q_{r,k}^{(n)}\right) \exp\left(-\beta_{n,r}\left[\mathcal{H}\left(K\left(\widetilde{q}_{r,k+1}^{(n)}\right)\right) - \mathcal{H}\left(K\left(q_{r,k}^{(n)}\right)\right)\right]^+\right)\right]. \tag{6.16}$$

By Assumption 10,

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

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

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

2. $\widetilde{q}_{r,k+1}^{(n)} - q_{r,k}^{(n)}$ has approximately the same distribution as $\gamma_n Y_r$ where Y_r is an $r \times r$ symmetric matrix of independent standard Gaussians. Notice that if $q_{r,k,(i,j)}^{(n)} \in \{0,1\}$ for any $(i,j) \in [r]^{(2)}$, this is not true. This approximation is valid only when all the coordinates of $q_{r,k}^{(n)}$ 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 (6.16) we obtain that with high probability, $\mathbb{E}\Big[\Delta q_{r,k}^{(n)} \,\Big|\, \mathcal{F}_k\Big]$

$$\begin{aligned}
&\approx \mathbb{E}\Big[\Big(\widetilde{q}_{r,k+1}^{(n)} - q_{r,k}^{(n)}\Big) \exp\Big(-\beta_{n,r}\Big\langle D\mathcal{H}\Big(K\Big(q_{r,k}^{(n)}\Big)\Big), K\Big(\widetilde{q}_{r,k+1}^{(n)}\Big) - K\Big(q_{r,k}^{(n)}\Big)\Big\rangle^{+}\Big)\Big] \\
&= \gamma_n \mathbb{E}\Big[Y_r \exp\Big(-\beta_{n,r}\gamma_n\Big\langle \nabla H_r\Big(q_{r,k}^{(n)}\Big), Y_r\Big\rangle_{\mathrm{F}}^{+}\Big)\Big], \qquad (6.17)
\end{aligned}$$

where we used the fact that for any two $r \times r$ symmetric matrices $A, B \in \mathcal{M}_r$ we have $\langle A, B\rangle_F = r^2\langle K(A), K(B)\rangle$ and that the fact that $D\mathcal{H} = r^{-2}\nabla H$ (see [167, Lemma 4.10]). The expectation in the last expression above is very amenable to analysis. It follows from Lemma 6.4.1 that $\mathbb{E}\Big[Y_r \exp\Big(-\beta_{n,r}\gamma_n\Big\langle \nabla H_r\Big(q_{r,k}^{(n)}\Big), Y_r\Big\rangle_{\mathrm{F}}^{+}\Big)\Big] = r^{-4} b_r\Big(q_{r,k}^{(n)}\Big)$ where b_r is defined in Definition 6.3.1.

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

$$\mathbb{E}\Big[\Delta q_{r,k}^{(n)} \,\Big|\, \mathcal{F}_k\Big] \approx \gamma_n r^{-4} b_r\Big(q_{r,k}^{(n)}\Big), \qquad (6.18)$$

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

$$\mathrm{Cov}\Big(\Delta q_{r,k}^{(n)} \,\Big|\, \mathcal{F}_k\Big) = r^{-4}\gamma_n\sigma^2 I + O(\gamma_n^2), \qquad (6.19)$$

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

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

$$q_r^{(n)}(t) := q_{r,t_{n,r}}^{(n)}, \qquad t \in \mathbb{R}_+, \qquad (6.20)$$

In particular, we obtain

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

Using the heuristic derived in (6.18) and (6.19), one expects that

$$q_r^{(n)}(t) - q_r^{(n)}(0) \approx \sum_{k=0}^{t_{n,r}-1} \gamma_n r^{-4} b_r\Big(q_{r,k}^{(n)}\Big) + \sum_{k=0}^{t_{n,r}-1} \Delta M_{r,k}^{(n)}, \qquad t \in \mathbb{R}_+, \tag{6.21}$$

where $\Big(\Delta M_{r,k}^{(n)}\Big)_{k\in\mathbb{Z}_+}$ is a $\mathcal{M}_r$-valued martingale difference sequence with uniform coordinatewise variance $\gamma_n r^{-4}\sigma^2$. We must caution that the approximation in (6.21) is not valid if $q_{r,k}^{(n)}$ is close to boundary $\{0,1\}$. The heuristic calculations have been derived under the assumption that all coordinates of $q_{r,k}^{(n)}$ are away from $\{0,1\}$. Ignoring this boundary contribution, it is reasonable to conclude that

$$q_r^{(n)}(t) - q_r^{(n)}(0) \approx \int_0^t b_r\Big(q_r^{(n)}(s)\Big)\,\mathrm{d}s + \sigma B_r(t), \qquad t \in \mathbb{R}_+,$$

where B_r is an $r \times r$ 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 $\Big(q_{r,k}^{(n)}\Big)_{k\in\mathbb{Z}_+}$ spends negligible proportion of time at the boundary, then

$$q_r^{(n)}(t) - q_r^{(n)}(0) \approx \int_0^t b_r\Big(q_r^{(n)}(s)\Big)\,\mathrm{d}s + \sigma B_r(t) + L_r^{(0)}(t) - L_r^{(1)}(t), \qquad t \in \mathbb{R}_+,$$

where $\Big(q_r^{(n)}, L_r^{(0)}, L_r^{(1)}\Big)$ solves the Skorokhod problem on the cube $\mathcal{M}_{r,+}$. That is, each coordinate process of $q_r^{(n)}$ satisfies the above SDE with reflection at the boundary $\{0,1\}$. This heuristic argument can be made precise (see Proposition 6.3.3) 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 Proposition 6.3.3 and its proof.

Proposition 6.3.3. *Let $\mathcal{H}$ satisfy Assumption 10 and let $(\gamma_n)_{n\in\mathbb{N}}$ satisfy condition (6.3). Let $D([0,\infty), \mathcal{M}_{r,+})$ be the space of right continuous functions with left limits equipped with the topology of uniform convergence over compact subsets. Let $q_r^{(n)}\colon \mathbb{R}_+ \to \mathcal{M}_{r,+}$ be a piecewise interpolation of $\Big(q_{r,k}^{(n)}\Big)_{k\in\mathbb{Z}_+}$ (see equation (6.20)). Then, $q_r^{(n)}$ converges weakly*

in $D([0,\infty),\mathcal{M}_{r,+})$ *to a process* X_r *over compact time intervals, with continuous path that satisfies the SDE*

$$\mathrm{d}X_r(t) = b_r(X_r(t))\,\mathrm{d}t + \sigma\,\mathrm{d}B_r(t) + \mathrm{d}L_r^{(0)}(t) - \mathrm{d}L_r^{(1)}(t), \qquad t \in \mathbb{R}_+, \tag{6.22}$$

with initial condition $X_r(0) = q_{r,0}^{(n)}$*, where* B_r *is a symmetric* $r \times r$ *matrix with whose coordinates are i.i.d. Brownian motions (up to matrix symmetry) and* $\left(X_r, L_r^{(0)}, L_r^{(1)}\right)$ *solves the Skorokhod problem w.r.t. the finite dimensional cube* $\mathcal{M}_{r,+}$ *(see Section 5.2.3).*

Proof of Proposition 6.3.3

The proof is long and requires several lemmas. Therefore, we first give an outline of the proof before presenting the details. Fix $r \in \mathbb{N}$. For every $k \in \mathbb{Z}_+$, let $\mathcal{F}_k^n$ be the sigma algebra generated by $\left\{q_{r,\ell}^{(n)} \,\middle|\, \ell \in \{0\} \cup [k]\right\}$. Let $t_{n,r} = \lfloor tr^4/\gamma_n \rfloor$ as defined earlier. For $i,j \in [r]$, notice that

$$\begin{aligned} q_{r,(i,j)}^{(n)}(t) - q_{r,(i,j)}^{(n)}(0) = &\sum_{\ell=0}^{t_{n,r}-1} \mathbb{E}\left[\Delta q_{r,\ell,(i,j)}^{(n)} \,\middle|\, \mathcal{F}_\ell^n\right] \mathbb{1}_{(0,1)}\left\{q_{r,\ell,(i,j)}^{(n)}\right\} \\ &+ \sum_{\ell=0}^{t_{n,r}-1} \Delta M_{r,\ell,(i,j)}^{(n)} + L_{r,(i,j)}^{(n,0)}(t) - L_{r,(i,j)}^{(n,1)}(t), \end{aligned}$$

for every $t \in \mathbb{R}_+$, where $\Delta M_{r,\ell}^{(n)} = \Delta q_{r,\ell}^{(n)} - \mathbb{E}\left[\Delta q_{r,\ell}^{(n)} \,\middle|\, \mathcal{F}_\ell^n\right]$ for all $\ell \in \mathbb{Z}_+$ and

$$\begin{aligned} L_{r,(i,j)}^{(n,0)}(t) &= \sum_{\ell=0}^{t_{n,r}-1} \mathbb{E}\left[\Delta q_{r,\ell,(i,j)}^{(n)} \,\middle|\, \mathcal{F}_\ell^n\right] \mathbb{1}_{\{0\}}\left\{q_{r,\ell,(i,j)}^{(n)}\right\}, \\ L_{r,(i,j)}^{(n,1)}(t) &= \sum_{\ell=0}^{t_{n,r}-1} \mathbb{E}\left[\Delta q_{r,\ell,(i,j)}^{(n)} \,\middle|\, \mathcal{F}_\ell^n\right] \mathbb{1}_{\{1\}}\left\{q_{r,\ell,(i,j)}^{(n)}\right\}, \end{aligned}$$

for $t \in \mathbb{R}_+$, where $L_{r,(i,j)}^{(n,0)}(t)$ is $\frac{1}{n^2}$ times the number of times the process $q_{r,(i,j)}^{(n)}$ visits $\{0\}$ before time t and similarly for $L_{r,(i,j)}^{(n,1)}(t)$. Note that $\left(M_{r,k}^{(n)} := \sum_{\ell=0}^{k-1} \Delta M_{r,\ell}^{(n)}\right)_{k\in\mathbb{Z}_+}$ is a $\mathcal{M}_r$-valued martingale and we define a piecewise constant interpolation of this martingale process $M_r^{(n)}$ defined as $M_r^{(n)}(t) = M_{r,t_{n,r}}^{(n)}$ for $t \in \mathbb{R}_+$. Let Sko be the Skorokhod map (see Section 5.2.3), then for any $t \in \mathbb{R}_+$, and any $(i,j) \in [r]^{(2)}$,

$$q_{r,(i,j)}^{(n)}(t) = \mathrm{Sko}\left(q_{r,(i,j)}^{(n)}(0) + \sum_{\ell=0}^{t_{n,r}-1} \mathbb{E}\left[\Delta q_{r,\ell,(i,j)}^{(n)} \,\middle|\, \mathcal{F}_\ell^n\right] \mathbb{1}_{(0,1)}\left\{q_{r,\ell,(i,j)}^{(n)}\right\} + M_{r,(i,j)}^{(n)}(t)\right). \tag{6.23}$$

1. Since the Skorokhod map Sko is a 4-Lipschitz map [137], to show that $q_r^{(n)}(t)$ converges uniformly to $X_r(t)$ as $n \to \infty$, it is sufficient to show that

$$\sum_{\ell=0}^{t_{n,r}-1} \mathbb{E}\Big[\Delta q^{(n)}_{r,\ell,(i,j)} \;\Big|\; \mathcal{F}^n_\ell\Big] \mathbb{1}_{(0,1)}\Big\{q^{(n)}_{r,\ell,(i,j)}\Big\} \to \int_0^t b_{r,(i,j)}(X_r(s)) \mathbb{1}_{(0,1)}\big\{X_{r,(i,j)}(s)\big\}\,\mathrm{d}s,$$

$$\text{and} \quad M_r^{(n)}(t) \to \sigma B_r(t),$$

uniformly over compact time intervals as $n \to \infty$.

2. In Lemma 6.4.3, we show that the quadratic variation of the martingale $M_r^{(n)}$ in the time interval $[0,t]$ converges to $t\sigma^2$ for every $t \in \mathbb{R}_+$ as $n \to \infty$. The key ingredient is the fact that a simple symmetric reflected random walk spends negligible amount of time at the boundary.

3. Using Lemma 6.4.3 and [82, Theorem 1.4, Chapter 7] we conclude that the process $M_r^{(n)}$ converges to the process (weakly) σB_r where B_r is an $r \times r$ symmetric matrix with i.i.d. Brownian motions. Using Skorokhod representation theorem both $M_r^{(n)}$ and B_r 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_r and $q_r^{(n)}$ using (6.22) and (6.23) respectively.

5. It remains to show that the first condition in Step 1 holds. To this end, we first show that for every fixed $\epsilon > 0$, $\mathbb{E}\Big[\Delta q^{(n)}_{r,\ell,(i,j)} \;\Big|\; \mathcal{F}^n_\ell\Big] \mathbb{1}_{(\epsilon,1-\epsilon)}\Big\{q^{(n)}_{r,\ell,(i,j)}\Big\} - b_{r,(i,j)}\Big(q^{(n)}_{r,\ell}\Big) \mathbb{1}_{(\epsilon,1-\epsilon)}\Big\{q^{(n)}_{r,\ell,(i,j)}\Big\} \to 0$ as $n \to \infty$. This is achieved by a sequence of reductions in Lemma 6.4.7.

6. Finally, we show that $\sum_{\ell=0}^{t_{n,r}-1} \mathbb{1}_{(\delta,1-\delta)^c}\Big\{q^{(n)}_{r,\ell,(i,j)}\Big\} \to 0$ as $n \to \infty$. Since $\frac{1}{\gamma_n}\mathbb{E}\Big[\Delta q^{(n)}_{r,\ell} \;\Big|\; \mathcal{F}_\ell\Big]$ is uniformly bounded. We conclude that the first condition in Step 1 holds. This completes the proof.

Proof of Proposition 6.3.3. Let $q_r^{(n)}$ be defined by (6.23). In the following we will keep r 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 $D([0,t], \mathbb{R}^{[r]^{(2)}})$, the space of right continuous paths with left limits from $[0,t]$ to $\mathbb{R}^{[r]^{(2)}}$. Define

$$b^{(n)}_{r,(i,j)}\left(q^{(n)}_{r,\ell}\right) = \frac{1}{\gamma_n r^{-4}} \mathbb{E}\left[\Delta q^{(n)}_{r,\ell,(i,j)} \,\middle|\, \mathcal{F}^n_\ell\right] \mathbb{1}_{(0,1)}\left\{q^{(n)}_{r,\ell,(i,j)}\right\},$$

and let b_r be in Definition in 6.3.1. Recall that both $b_r^{(n)}$ and b_r are uniformly bounded by some constant C. Also, recall that B_r is an $r \times r$ symmetric matrix with i.i.d. Brownian coordinates. Define the stochastic processes $Y_r^{(n)}, \widetilde{Y}_r^{(n)}, \widetilde{Z}_r^{(n)}, Z_r^{(n)} \colon \mathbb{R}_+ \to \mathbb{R}^{[r]^{(2)}}$ as:

$$\begin{aligned}
Y^{(n)}_{r,(i,j)}(t) &= q^{(n)}_{r,(i,j)}(0) + \sum_{\ell=0}^{t_{n,r}-1} \gamma_n r^{-4} b^{(n)}_{r,(i,j)}\left(q^{(n)}_{r,\ell}\right) + M^{(n)}_{r,(i,j)}(t),\\
\widetilde{Y}^{(n)}_{r,(i,j)}(t) &= q^{(n)}_{r,(i,j)}(0) + \sum_{\ell=0}^{t_{n,r}-1} \gamma_n r^{-4} b^{(n)}_{r,(i,j)}\left(q^{(n)}_{r,\ell}\right) + \sigma B_{r,(i,j)}(t), \qquad t \in \mathbb{R}_+, (i,j) \in [r]^{(2)}.\\
\widetilde{Z}^{(n)}_{r,(i,j)}(t) &= q^{(n)}_{r,(i,j)}(0) + \int_0^t b^{(n)}_{r,(i,j)}\left(q^{(n)}_r(s)\right) \mathrm{d}s + \sigma B_{r,(i,j)}(t),\\
Z^{(n)}_{r,(i,j)}(t) &= q^{(n)}_{r,(i,j)}(0) + \int_0^t b_{r,(i,j)}\left(q^{(n)}_r(s)\right) \mathrm{d}s + \sigma B_{r,(i,j)}(t),
\end{aligned}$$

Notice that $Y_r^{(n)}$ is the "unconstrained version" of the process $q_r^{(n)}$ in the sense that $\mathrm{Sko}\left(Y_r^{(n)}\right)(s) = q_r^{(n)}(s)$ for every $s \in \mathbb{R}_+$. Finally, let $\left(X_r, L_r^{(0)}, L_r^{(1)}\right)$ be the process that satisfies the Skorokhod SDE

$$\mathrm{d}X_r(t) = b_r(X(t))\,\mathrm{d}t + \sigma\,\mathrm{d}B_r(t) + \mathrm{d}L_r^{(0)}(t) - \mathrm{d}L_r^{(1)}(t), \qquad X_{r,(i,j)}(0) = q^{(n)}_{r,(i,j)}(0).$$

Denote the corresponding unconstrained process

$$\widetilde{X}_r(t) = q_r^{(n)}(0) + \int_0^t b_r(X_r(s))\,\mathrm{d}s + \sigma\,\mathrm{d}B_r(t).$$

Note that $\mathrm{Sko}\left(\widetilde{X}_r\right) = X_r$. Recall that the goal is to show that $q_r^{(n)}$ converges to the process X_r as $n \to \infty$. Using the fact that the Skorokhod map is Lipschitz (see Section 5.2.3), it is

sufficient to show that $Y_r^{(n)}$ converges to $\widetilde{X}_r$. To this end, set

$$\Delta^{(n)}(t) \coloneqq \mathbb{E}\left[\sup_{s\in[0,t]}\left\|Y_r^{(n)}(s)-\widetilde{X}_r(s)\right\|_{\mathrm{F}}^2\right], \qquad \Delta_1^{(n)}(t) \coloneqq \mathbb{E}\left[\sup_{s\in[0,t]}\left\|Y_r^{(n)}(s)-\widetilde{Y}_r^{(n)}(s)\right\|_{\mathrm{F}}^2\right],$$
$$\Delta_2^{(n)}(t) \coloneqq \mathbb{E}\left[\sup_{s\in[0,t]}\left\|\widetilde{Z}_r^{(n)}(s)-\widetilde{Y}_r^{(n)}(s)\right\|_{\mathrm{F}}^2\right], \qquad \Delta_3^{(n)}(t) \coloneqq \mathbb{E}\left[\sup_{s\in[0,t]}\left\|\widetilde{Z}_r^{(n)}(s)-Z_r^{(n)}(s)\right\|_{\mathrm{F}}^2\right],$$
$$\Delta_4^{(n)}(t) \coloneqq \mathbb{E}\left[\sup_{s\in[0,t]}\left\|\widetilde{X}_r^{(n)}(s)-Z_r^{(n)}(s)\right\|_{\mathrm{F}}^2\right], \qquad \text{for all } t\in\mathbb{R}_+.$$

Since $(a+b+c+d)^2 \le 4(a^2+b^2+c^2+d^2)$ for all $(a,b,c,d)\in\mathbb{R}^4$, $\Delta^{(n)}(t) \le 4\sum_{i=1}^4 \Delta_i^{(n)}(t) \le 4\sum_{i=1}^3 \Delta_i^{(n)}(t) + 64t\int_0^t \Delta^{(n)}(s)\,\mathrm{d}s$, where the final inequality is due to the 4-Lipschtizness of the Skorokhod map. In particular, for $t\in[0,T]$, we have $\Delta^{(n)}(t) \le 4\sum_{i=1}^3 \Delta_i^{(n)}(t) + 64T\int_0^t \Delta^{(n)}(s)\,\mathrm{d}s$. Note that $t\mapsto\Delta_i^{(n)}(t)$ is increasing. Therefore, using Grönwall's inequality [100], we obtain

$$\Delta^{(n)}(T) \le 4\left(\sum_{i=1}^3 \Delta_i^{(n)}(T)\right)\exp\bigl(64T^2\bigr). \tag{6.24}$$

It is therefore sufficient to show that $\Delta_i^{(n)}(t)\to 0$ as $n\to\infty$ for $i\in[3]$. This is done in following steps. Using Lemma 6.4.3 below and Theorem [82, Theorem 1.4, Chapter 7], we know that the process $M_r^{(n)}$ converges to σB_r uniformly on compact subsets of time. In particular, for fixed $t>0$ we have that $\Delta_1^{(n)}(t)\to 0$ as $n\to\infty$. For $\Delta_2^{(n)}$, we notice that the error is actually the error from the Riemann sum approximation. Hence, $\Delta_2^{(n)}(t)\le C_r\gamma_n^2\to 0$ as $n\to\infty$. To see this, first recall that $q^{(n)}(s)$ is piecewise constant on the interval of length $\gamma_n r^{-4}$, that is, $q_r^{(n)}(s) = q_{r,\lfloor s/(\gamma_n r^{-4})\rfloor}^{(n)}$. Now observe that

$$\begin{aligned}\left|\widetilde{Y}_{r,(i,j)}^{(n)}(t)-\widetilde{Z}_{r,(i,j)}^{(n)}(t)\right| &= \left|\sum_{\ell=0}^{t_{n,r}-1}\gamma_n r^{-4} b_{(i,j)}^{(n)}\Bigl(q_{r,\ell}^{(n)}\Bigr) - \int_0^t b_{(i,j)}^{(n)}\Bigl(q_r^{(n)}(s)\Bigr)\right| \\ &= \left|\bigl(t-\gamma_n r^{-4}(t_{n,r}-1)\bigr) b_{(i,j)}^{(n)}\Bigl(q_{r,t_{n,r}-1}^{(n)}\Bigr)\right| \le C\gamma_n r^{-4},\end{aligned}$$

where the inequality in the last line follows from the fact that $b^{(n)}$ is uniformly bounded. Squaring both sides and summing over all $(i,j)\in[r]^{(2)}$ we conclude that $\Delta_2^{(n)}(t)\le C_r\gamma_n^2$.

We now show that $\Delta_3^{(n)}(t)\to 0$ as $n\to\infty$. To do this, we fix $\epsilon>0$, $\delta>0$ (we assume that $\delta\ll\epsilon$ and $\delta+\epsilon\ll 1$). Define

$$A_\epsilon \coloneqq \left\{M\in\mathcal{M}_r \;\middle|\; \epsilon\le M_{(i,j)}\le 1-\epsilon \quad \forall\ (i,j)\in[r]^{(2)}\right\}, \quad B_\epsilon \coloneqq \mathcal{M}_r\setminus A_\epsilon. \tag{6.25}$$

Start observing that $\sup_{s\in[0,t]}\left\|\widetilde{Z}_r^{(n)}(s)-Z_r^{(n)}(s)\right\|_{\mathrm{F}}^2$ is at most

$$
\begin{aligned}
&t\int_0^t\left\|b_r^{(n)}\Big(q_r^{(n)}(s)\Big)-b_r\Big(q_r^{(n)}(s)\Big)\right\|_{\mathrm{F}}^2\mathrm{d}s\\
&\quad\le t\int_0^t\left\|b_r^{(n)}\Big(q_r^{(n)}(s)\Big)-b_r\Big(q_r^{(n)}(s)\Big)\right\|_{\mathrm{F}}^2\mathbb{1}_{A_\epsilon}\Big\{q_r^{(n)}(s)\Big\}\,\mathrm{d}s\\
&\qquad+t\int_0^t\left\|b_r^{(n)}\Big(q_r^{(n)}(s)\Big)-b_r\Big(q_r^{(n)}(s)\Big)\right\|_{\mathrm{F}}^2\mathbb{1}_{B_\epsilon}\Big\{q_r^{(n)}(s)\Big\}\,\mathrm{d}s. \qquad (6.26)
\end{aligned}
$$

From Lemma 6.4.7 below $b_r^{(n)}\Big(q_r^{(n)}(s)\Big)$ and $b_r\Big(q_r^{(n)}(s)\Big)$ are close in $\|\cdot\|_{\mathrm{F}}^2$ by $\frac{Cr^2}{n^4}+4r^2\beta_{n,r}^2e_n^3\max\{|\lambda|,L\}$ with probability at least $1-p_{n,\epsilon}$, when $q_r^{(n)}(s)\in A_\epsilon$ where $p_{n,\epsilon}=\frac{4r^2}{n^4}+2r^2\mathrm{erf}\Big(\frac{r^2\epsilon}{4\sigma\sqrt{2\gamma_n}}\Big)$ and $e_n=O(\gamma_n^2\log n)$. On the other hand, we notice that $b_r^{(n)}$ and b_r are both uniformly bounded by in $\|\cdot\|_\infty$. Using these two facts we conclude that

$$
\begin{aligned}
&\mathbb{E}\left[\int_0^t\left\|b_r^{(n)}\Big(q_r^{(n)}(s)\Big)-b_r\Big(q_r^{(n)}(s)\Big)\right\|_{\mathrm{F}}^2\mathbb{1}_{A_\epsilon}\Big\{q_r^{(n)}(s)\Big\}\,\mathrm{d}s\right]\\
&\quad\le C\frac{t}{\gamma_n}p_{n,\epsilon}+\frac{Cr^2}{n^4}+4r^2\beta_{n,r}^2e_n^3\max\{|\lambda|,L\}. \qquad (6.27)
\end{aligned}
$$

Since $b_r^{(n)}$ and b_r are uniformly bounded, the second term in (6.26) is bounded as

$$
\int_0^t\left\|b_r^{(n)}\Big(q_r^{(n)}(s)\Big)-b_r\Big(q_r^{(n)}(s)\Big)\right\|_{\mathrm{F}}^2\mathbb{1}_{B_\epsilon}\Big\{q_r^{(n)}(s)\Big\}\,\mathrm{d}s\le CD^{(n)}(t),
$$

for some constant $C>0$, where $D^{(n)}(t):=\int_0^t\mathbb{1}_{B_\epsilon}\Big\{q_r^{(n)}(s)\Big\}\,\mathrm{d}s$.

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 $I_\epsilon:=[0,\epsilon)\cup(1-\epsilon,1]$ and $0\le\psi_{\epsilon,\delta}\le 1$ and $\mathrm{supp}(\psi)\subset[0,\epsilon+\delta)\cup(1-\epsilon-\delta,1]$. Recall that by our assumption $\epsilon+\delta\ll 1$. Then, $D^{(n)}(t)\le\int_0^t\psi_{\epsilon,\delta}\Big(q_r^{(n)}(s)\Big)\,\mathrm{d}s$.

Recall that $q_r^{(n)}(s)=\mathrm{Sko}\big(Y^{(n)}\big)(s)$ 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 $\Psi_{\epsilon,\delta}(s):=\psi_{\epsilon,\delta}\Big(\mathrm{Sko}\Big(Y_r^{(n)}\Big)(s)\Big)$ and $\widetilde{\Psi}_{\epsilon,\delta}(s):=\psi_{\epsilon,\delta}\Big(\mathrm{Sko}\Big(\widetilde{Z}_r^{(n)}\Big)(s)\Big)$. Now observe that

$$
\begin{aligned}
\Psi_{\epsilon,\delta}(s)&\le\widetilde{\Psi}_{\epsilon,\delta}(s)+L_{\epsilon,\delta}\left\|\widetilde{Z}_r^{(n)}(s)-Y_r^{(n)}(s)\right\|_{\mathrm{F}}^2\\
&\le\widetilde{\Psi}_{\epsilon,\delta}(s)+2L_{\epsilon,\delta}\left(\left\|\widetilde{Z}_r^{(n)}(s)-\widetilde{Y}_r^{(n)}(s)\right\|_{\mathrm{F}}^2+\left\|Y_r^{(n)}(s)-\widetilde{Y}_r^{(n)}(s)\right\|_{\mathrm{F}}^2\right).
\end{aligned}
$$

Therefore, we obtain

$$\mathbb{E}\Big[D^{(n)}(t)\Big] \le \mathbb{E}\bigg[\int_0^t \widetilde{\Psi}_{\epsilon,\delta}(s)\,\mathrm{d}s\bigg] + 2L_{\epsilon,\delta}t\Big(\Delta_1^n(t) + \Delta_2^{(n)}(t)\Big). \tag{6.28}$$

Note that $\mathbb{E}\Big[\int_0^t \widetilde{\Psi}_{\epsilon,\delta}(s)\,\mathrm{d}s\Big] = \mathbb{E}\Big[F_{\epsilon,\delta}\Big(\widetilde{Z}^{(n)}\Big)\Big]$ for some bounded continuous function $F_{\epsilon,\delta}\colon C([0,t],\mathcal{M}_{r,+}) \to \mathbb{R}$. Also, recall that $\widetilde{Z}^{(n)}$ satisfies the SDE $\widetilde{Z}_r^{(n)}(t) = q_r(0) + \int_0^t f(s)\,\mathrm{d}s + \sigma B_r(t)$, where $f(s) = b_r^{(n)}\Big(q_r^{(n)}(s)\Big)$ is a bounded function. Set

$$\mathcal{E} = \exp\bigg(\frac{1}{\sigma}\int_0^t b_r^{(n)}\Big(q_r^{(n)}(s)\Big)\,\mathrm{d}B_r(s) - \frac{1}{2\sigma^2}\int_0^t b_r^{(n)}\Big(q_r^{(n)}(s)\Big)\,\mathrm{d}s\bigg).$$

Using Girsanov's theorem and the Cauchy–Schwarz inequality we obtain

$$\mathbb{E}\Big[F_{\epsilon,\delta}\Big(\widetilde{Z}_r^{(n)}\Big)\Big]^2 = \mathbb{E}[F_{\epsilon,\delta}(B)\mathcal{E}]^2 \le \mathbb{E}\big[F_{\epsilon,\delta}^2(B)\big]\mathbb{E}\big[\mathcal{E}^2\big].$$

Finally using the fact that $b_r^{(n)}$ is uniformly bounded, we obtain that $\mathbb{E}\big[\mathcal{E}^2\big] \le C_{r,t,\sigma}$. On the other hand, we notice that by definition

$$\mathbb{E}\big[F_{\epsilon,\delta}^2(B)\big] = \mathbb{E}\bigg[\bigg(\int_0^t \psi_{\epsilon,\delta}(\mathrm{RBM}(s))\,\mathrm{d}s\bigg)^2\bigg] \le t\int_0^t \mathbb{P}\{\mathrm{RBM}(s) \in I_\epsilon\}\,\mathrm{d}s =: C(\epsilon,\delta,r,t)^2, \tag{6.29}$$

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 (6.27), (6.28) and (6.29) we obtain that $\Delta_3^{(n)}(t) \le \frac{t}{\gamma_n}p_n + C(\epsilon,\delta,r,t) + 2L_{\epsilon,\delta}t\Big(\Delta_1^n(t) + \Delta_2^{(n)}(t)\Big)$.

Putting this back in equation (6.24) we conclude that

$$\Delta^{(n)}(t) \le \bigg(\frac{t}{\gamma_n}p_n + C(\epsilon,\delta,r,t) + (2L_{\epsilon,\delta}t + C)\Big(\Delta_1^n(t) + \Delta_2^{(n)}(t)\Big)\bigg)\mathrm{e}^{Ct^2}.$$

Recall that $\frac{p_n}{\gamma_n} \to 0$ and $\Delta_1^{(n)}(t) \to 0$ and $\Delta_2^{(n)}(t) \le \gamma_n^2 \to 0$ as $n \to \infty$. Therefore, we conclude that $\limsup_{n\to\infty} \Delta^{(n)}(t) \le C(\epsilon,\delta,r,t)\mathrm{e}^{Ct^2}$. Since $C(\epsilon,\delta,r,t) \to 0$ as $\epsilon,\delta \to 0$, this concludes the proof. □

Let b_0 be as in Definition 6.3.1. Recall from Remark 6.3.2 that $\|b_0 + \beta D\mathcal{H}\|_\infty \to 0$ as $r \to \infty$. As an immediate consequence of Theorem 6.2.2 (see Remark 6.2.3) we obtain that the process X_r converges to the McKean-Vlasov SDE defined in (6.2) with drift given by $-\beta D\mathcal{H}(w)$ as $r \to \infty$. 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 $\{(U,V) = (u,v)\}$,

$$\begin{aligned} \mathrm{d}X(t) &= -\beta D\mathcal{H}(\mathbb{E}[W^\sigma(t)])(u,v)\,\mathrm{d}t + \sigma\,\mathrm{d}B(t) + \mathrm{d}L^{(0)}(t) - \mathrm{d}L^{(1)}(t), \\ W^\sigma(t)(x,y) &= \mathrm{Law}(X(t)\,|\,(U,V) = (x,y)), \qquad (x,y) \in [0,1]^{(2)}, \end{aligned} \tag{6.30}$$

for $t \in \mathbb{R}_+$, where $\left(X, L^{(0)}, L^{(1)}\right)$ solves the Skorokhod problem with respect to $[0,1]$.

Proposition 6.3.4. *Let X_r be a solution of (6.22) with initial condition $X_r(0) \in \mathcal{M}_{r,+}$. If $\lim_{r\to\infty} \|\mathcal{K}(X_r(0)) - W_0\|_\blacksquare = 0$, then X_r converges in MVG sense, in probability, uniformly over compact time intervals, to a deterministic curve W^σ in the space of MVGs as $r \to \infty$. Moreover, W^σ is described by (6.30) with initial condition W_0.*

Remark 6.3.5. *Consider the curve w^σ in the space of graphons defined as $w^\sigma(t) \coloneqq \mathbb{E}[W^\sigma(t)]$ for all $t \in \mathbb{R}_+$. It follows from Remark 6.2.4 that the random curves $(X_r)_{r\in\mathbb{N}}$ converge in cut-metric, uniformly on compact intervals of time, to w^σ in probability. And, w^σ can be recovered as a solution of a MKV SDE satisfying on $\{(U,V) = (u,v)\}$*

$$\begin{aligned} \mathrm{d}X(t) &= -\beta D\mathcal{H}(w^\sigma)(u,v)\,\mathrm{d}t + \sigma\,\mathrm{d}B(t) + \mathrm{d}L^{(0)}(t) - \mathrm{d}L^{(1)}(t), \\ w^\sigma(t)(x,y) &= \mathbb{E}[X(t)\,|\,(U,V) = (x,y)], \qquad (x,y) \in [0,1]^{(2)}, \quad t \in \mathbb{R}_+, \end{aligned} \tag{6.31}$$

where $\left(X, L^{(0)}, L^{(1)}\right)$ solves the Skorokhod problem with respect to $[0,1]$.

When $\sigma = 0$, the graphon McKean-Vlasov (6.31) reduces to a deterministic evolution w of kernels given by

$$w(t)(x,y) = w(0)(x,y) - \beta \int_0^t D\mathcal{H}(w(s))(x,y)\mathbb{1}_{G_{w(s)}}\,\mathrm{d}s, \tag{6.32}$$

for $(x,y) \in [0,1]^{(2)}$ and $t \in \mathbb{R}_+$.

Here $G_u \subseteq [0,1]^{(2)}$ for any $u \in \mathcal{W}_{[0,1]}$ is defined as

$$\begin{aligned} G_u \coloneqq \left\{(x,y) \in [0,1]^{(2)} \,\middle|\, u(x,y) = 1, b(u)(x,y) < 0\right\} \cup \left\{(x,y) \in [0,1]^{(2)} \,\middle|\, u(x,y) = 0, b(u)(x,y) > 0\right\}, \\ \cup \left\{(x,y) \in [0,1]^{(2)} \,\middle|\, 0 < u(x,y) < 1\right\}. \end{aligned} \tag{6.33}$$

This can be seen by defining $L^{(0)}$ and $L^{(1)}$ as

$$\begin{aligned} L^{(1)}(t) &\coloneqq + \int_0^t b(w(s))(u,v)\mathbb{1}\{w(s)(u,v) = 1, b(w(s)) > 0\}\,\mathrm{d}s, \\ L^{(0)}(t) &\coloneqq - \int_0^t b(w(s))(u,v)\mathbb{1}\{w(s)(u,v) = 0, b(w(s)) < 0\}\,\mathrm{d}s, \end{aligned}$$

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 5.2.3). It is clear that that w 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 w^σ converges to w 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 $\left\{B_{(i,j)}\right\}_{(i,j)\in\mathbb{N}^{(2)}}$. We can therefore define an IEA X^σ on this probability space such that

$$\begin{aligned}
\mathrm{d}X^\sigma_{(i,j)}(t) &= -\beta D\mathcal{H}(\mathbb{E}[W^\sigma(t)])(U_i,U_j)\,\mathrm{d}t + \sigma\,\mathrm{d}B_{(i,j)}(t) + \mathrm{d}L^{(0)}_{(i,j)}(t) - \mathrm{d}L^{(1)}_{(i,j)}(t),\\
W^\sigma(t)(x,y) &= \mathrm{Law}(X^\sigma_{(i,j)}(t)\,|\,(U_i,U_j)=(x,y)), \qquad (x,y)\in[0,1]^{(2)},
\end{aligned}$$

for $t\in\mathbb{R}_+$, where $\left(X^\sigma, L^{(0)}, L^{(1)}\right)$ solves the Skorokhod problem with respect to $[0,1]$.

Let w be as defined in (6.32). Recall that $w(t)$ can be naturally identified with an MVG $W(t)$ defined as $W(t)(x,y) \coloneqq \delta_{w(t)(x,y)}$ 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

$$\begin{aligned}
\mathrm{d}X_{(i,j)}(t) &= -\beta D\mathcal{H}(\mathbb{E}[W(t)])(U_i,U_j)\,\mathrm{d}t + \mathrm{d}L^{(0)}_{(i,j)}(t) - \mathrm{d}L^{(1)}_{i,j}(t),\\
W(t)(x,y) &= \mathrm{Law}(X_{(i,j)}(t)\,|\,(U_i,U_j)=(x,y)), \qquad (x,y)\in[0,1]^{(2)},
\end{aligned} \qquad t\in\mathbb{R}_+, \tag{6.34}$$

where $\left(X, L^{(0)}, L^{(1)}\right)$ solves the Skorokhod problem with respect to $[0,1]$. Note that given $\{U_i\}_{i\in\mathbb{N}}$ the IEA X is deterministic. In particular, $W(t)(x,y) = \delta_{w(t)(x,y)}$.

Proposition 6.3.6. *Let $w_0\in\mathcal{W}_{[0,1]}$ be a kernel. Let w^σ and w be defined in equation (6.31) and (6.32). Then, for every finite $t>0$, $\sup_{s\in[0,t]}\|w^\sigma(s)-w(s)\|_\square \le 2C\sigma^2 t\exp(Ct^2)$ for some universal constant $C>0$.*

Proof of Proposition 6.3.6. We first prove a slightly stronger result, that is, we show that W^σ converges to W in the MVG sense. The desired result therefore follows immediately. The proof closely resembles the proof of Theorem 6.2.2. Let $(\Omega,\mathcal{F},\mathbb{P})$ be as above and X^σ, X, W^σ, W be as above with the initial condition $W^\sigma(0)(x,y) = W(0)(x,y) = \delta_{w_0(x,y)}$. Using the Lipschitzness of Skorokhod map as in the proof of Theorem 6.2.2, we observe that

for any (i,j) we have

$$\left|X^{\sigma}_{(i,j)}(t) - X_{(i,j)}(t)\right|^2 \leq Ct\int_0^t |b(w^{\sigma}(s))(U_i,U_j) - b(w(s))(U_i,U_j)|^2 + C\sigma^2|B_{i,j}(t)|^2.$$

Summing over $(i,j)\in[k]^2$ and diving by $\frac{1}{k^2}$ we obtain that for each $k\in\mathbb{N}$ we have

$$\begin{aligned}
&\|\mathcal{K}(X^{\sigma}[k](t)) - \mathcal{K}(X[k](t))\|_2^2 \leq I_k(t) + J_k(t),\\
\text{where}\quad &I_k(t) = Ct\int_0^t \frac{1}{k^2}\sum_{(i,j)\in[k]^2} |b(w^{\sigma}(s))(U_i,U_j) - b(w(s))(U_i,U_j)|^2\,\mathrm{d}s,\\
\text{and}\quad &J_k(t) \coloneqq C\sigma^2\frac{1}{k^2}\sum_{(i,j)\in[k]^2}\left|B_{(i,j)}(t)\right|^2.
\end{aligned}$$

By Doob's maximal inequality [129, page 14, Theorem 3.8.iv] and Markov's inequality we get $\mathbb{P}\Big\{\sup_{s\in[0,t]} J_k(s) \geq 2C\sigma^2 t\Big\} \leq \frac{C}{4k^2}$. Using our assumption on b, we conclude that (compare with (6.9))

$$\sup_{s\in[0,t]}\|\mathcal{K}(X^{\sigma}[k](s)) - \mathcal{K}(X[k](s))\|_{\blacksquare}^2 \leq C\beta^2\kappa_{\blacksquare}^2 t\int_0^t \|W(s) - W^{\sigma}(s)\|_{\blacksquare}^2\,\mathrm{d}s + 2C\sigma^2 t, \quad (6.35)$$

with probability at least $1-\frac{C}{4k^2}$. Note that compared to equation (6.9) in the proof of Theorem 6.2.2, 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 6.2.2 to replace $\|\mathcal{K}(X^{\sigma}[k](s)) - \mathcal{K}(X[k](s))\|_{\blacksquare}^2$ with $\|W(s) - W^{\sigma}(s)\|_{\blacksquare}^2$ 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 [100] as in the proof of Theorem 6.2.2 we conclude that

$$\begin{aligned}
&\sup_{s\in[0,t]} D_2^2(\mathcal{K}(X^{\sigma}[k](s)),\mathcal{K}(X[k](s))) + \sup_{s\in[0,t]}\|W^{\sigma}(s) - W(s)\|_{\blacksquare}^2\\
&\qquad\leq (C_k + 2C\sigma^2 t)\mathrm{e}^{C\beta^2\kappa_{\blacksquare}^2 t^2},\ \text{with probability at least } 1-\mathrm{e}^{-c_k} - \frac{C}{4k^2}.
\end{aligned}$$

Letting $k\to\infty$, we conclude that $\sup_{s\in[0,t]}\|W^{\sigma}(s) - W(s)\|_{\blacksquare}^2 \leq 2C\sigma^2 t\mathrm{e}^{C\beta^2\kappa_{\blacksquare}^2 t^2}$. The desired claim now follows from the fact that $W\to\mathbb{E}[W]$ is a contraction. □

Proposition 6.3.7 (Non-asymptotic high probability error bound)**.** *Let X_n^{σ} be a solution to equation (6.6) for $\Sigma_n\equiv\sigma$. Let the assumptions of Theorem 6.2.2 and Theorem 6.3.6*

hold. If the initial condition is i.i.d., then

$$\sup_{s\in[0,t]} \|\mathcal{K}(X_n^\sigma(s)) - W(s)\|_{\blacksquare}^2 \le C_t n^{-1/56}\log^{3/2} n + (64n^{-1/14} + 2C\sigma^2 t)\mathrm{e}^{C\beta^2\kappa_{\blacksquare}^2 t^2}, \tag{6.36}$$

with probability at least $1 - 5n^{-3/7} - tn^{-\frac{2}{7\kappa^2 t}}$. *Here* $W(s)(x,y) = \delta_{w(s)(x,y)}$ *for a.e.* $(x,y) \in [0,1]^{(2)}$, *and* $s \in \mathbb{R}_+$ *following equation* (6.31); *and* $\kappa = 32\sqrt{6}\big(L^2 + 2\kappa_{\blacksquare}^2\big)^{1/2}$.

Proof. To get a non-asymptotic error rate, we need to control on A_k and $B_k(n)$ in equation (6.11). 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, combining the arguments in [102, Proposition 4.5] and [150, 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 $\mathbb{P}\big\{A_k \ge M_t(\delta\log(1/\delta))^{1/4}\big\} \le k^{-2} + t\delta^{-1}\mathrm{e}^{\frac{128}{\delta\log(1/\delta)}}\mathrm{e}^{-k\delta\log(1/\delta)/2}$.

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 $\mathbb{P}\Big\{t(C_4, \Gamma(\psi, \widetilde{A}_k(s)) \ge \sqrt{\delta\log(1/\delta)}\Big\} \le \mathrm{e}^{-\frac{k\delta\log(1/\delta)}{2}}$. Note that the right side is independent of ψ. For any subset $F \subseteq \mathcal{L}$ define $\Delta_{k,F}(s) := \sup_{\psi\in F}\Big|t(C_4, \Gamma(\psi, \widetilde{A}_k(s)))\Big|$. Fix $\epsilon > 0$ and note that by Lemma 2.4.15 there exists a finite set $F \subseteq \mathcal{L}$ such that $|F| \le \mathrm{e}^{\frac{32}{\epsilon^2}}$ and $|\Delta_{k,\mathcal{L}}(s) - \Delta_{k,F}(s)| \le \epsilon$. Taking $\epsilon = \frac{1}{2}\sqrt{\delta\log(1/\delta)}$, we get $\mathbb{P}\Big\{\Delta_{k,\mathcal{L}}(s) \ge \sqrt{\delta\log(1/\delta)}\Big\} \le \mathbb{P}\Big\{\Delta_{k,F}(s) \ge 2^{-1}\sqrt{\delta\log(1/\delta)}\Big\} \le \mathrm{e}^{\frac{128}{\delta\log(1/\delta)}}\mathrm{e}^{-k\delta\log(1/\delta)/2}$.

Repeating the proof of [102, Proposition 4.5], 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 [150, Proposition 8.12], we have $\Big\|\widetilde{A}_k(s)\Big\|_{\blacksquare}^4 \le \Delta_{k,\mathcal{L}}(s)$. This yields the desired conclusion. In particular, choosing $\delta = 64\sqrt{k^{-1}\log k}$ and $\lambda_k = \log(k)/\big(16\cdot 384t(L^2 + 2\kappa_{\blacksquare}^2)\big)$, we have the left hand side of (6.11) 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 $\kappa = 32\sqrt{6}\big(L^2 + 2\kappa_{\blacksquare}^2\big)^{1/2}$.

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 6.3.6 and a triangle inequality. □

Proposition 6.3.8 (Convergence McKean-Vlasov (6.32) to equilibrium when $\sigma = 0$)**.** *Let $\mathcal{H}$ be $\delta_\square$-lower semicontinuous and satisfy Assumption 10 with $\lambda \geq 0$ and $L \in [\lambda, \infty) \cup \{\infty\}$. Let w be the graphon valued curve as defined in (6.32). Let $w_* \in \widehat{\mathcal{W}}_{[0,1]}$ be a minimizer of $\mathcal{H}$, then*

$$\mathcal{H}(w(t)) - \mathcal{H}(w_*) \leq \frac{\delta_2^2(w(0), w_*)}{2\beta t}, \qquad t \in \mathbb{R}_+. \tag{6.37}$$

Moreover, if $\lambda > 0$ and $L < \infty$ and $w_* \in \widehat{\mathcal{W}}_{[0,1]}$ is the unique minimizer of the strongly convex function $\mathcal{H}$, then for $t \in \mathbb{R}_+$,

$$\delta_2(w(t), w_*) \leq \mathrm{e}^{-\beta\lambda t}\delta_2(w(0), w_*), \quad \mathcal{H}(w(t)) - \mathcal{H}(w_*) \leq \frac{L}{2}\mathrm{e}^{-2\beta\lambda t}\delta_2^2(w(0), w_*). \tag{6.38}$$

Proof. Notice that $\beta D\mathcal{H}$ is the Fréchet-like derivative evaluation map of $\beta\mathcal{H}$. The proof immediately follows from [167, Remark 4.16] following [5, Remark 4.0.5, part (d)], [5, Corollaray 4.0.6] and Assumption 10. □

The notion of (geodesic) convexity of functions on graphons can be found in [167, Definition 2.15]. When $\mathcal{H}$ is a linear combinations of homomorphism densities, it is only semi-convex. However, one may regularize $\mathcal{H}$ by adding a large enough multiple of scalar entropy to make it strictly convex [167, Section 5.1.1, Section 5.1.3] and satisfy the conditions for exponential convergence in Proposition 6.3.8.

6.4 Remaining Proofs

6.4.1 Scaling limit of Metropolis

Properties of drift function

Lemma 6.4.1. *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*

$$\mathbb{E}_Y\left[Y \exp\left(-t\langle v, Y\rangle_{\mathrm{F}}^+\right)\right] = -2tv \exp\left(t^2\|v\|_{\mathrm{F}}^2\right)\overline{\Phi}(\sqrt{2}t\|v\|_{\mathrm{F}}).$$

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

$$\begin{aligned}\mathbb{E}\big[Y\exp\big(-t\langle v,Y\rangle_{\mathrm{F}}^{+}\big)\big] &= \mathbb{E}_X\big[\exp\big(-tX^+\big)\mathbb{E}_Y[Y \mid \langle v,Y\rangle = X]\big]\\ &= \frac{v}{\|v\|_{\mathrm{F}}^2}\mathbb{E}\big[X\exp\big(-tX^+\big)\big] = \frac{\sqrt{2}v}{\|v\|_{\mathrm{F}}}\mathbb{E}\Big[Z\exp\Big(-\sqrt{2}t\|v\|_{\mathrm{F}}Z^+\Big)\Big],\end{aligned}$$

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 $\alpha = \sqrt{2}t\|v\|_{\mathrm{F}}$. □

Martingale quadratic variation

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 6.4.2 (Time at boundary of reflected RW)**.** *Let $\ell_n = \lceil r^{-4}\sigma^2\gamma_n n^4\rceil$. Fix $x \in \big\{i/n^2 \,\big|\, i = 0,\dots,n^2\big\}$. Let S denote the symmetric random walk with step size $\frac{1}{n^2}$ reflected at $\{0,1\}$ starting at x. Then,*

$$\lim_{n\to\infty}\frac{1}{\gamma_n n^4}\sum_{k=1}^{\ell_n}\mathbb{1}_{\{0,1\}}\{S_k\} = 0, \quad \textit{in probability.}$$

We now compute the quadratic variation of the martingale $M_r^{(n)}(t)$ defined in Section 6.3.

Lemma 6.4.3 (Martingale Quadratic Variation)**.** *For $r \in \mathbb{N}$, $n \in \mathbb{N}$ and $t \in \mathbb{R}_+$, let $M_r^{(n)}(t) \coloneqq \sum_{\ell=0}^{t_{n,r}-1}\Delta M_{r,\ell}^{(n)}$ where $t_{n,r} = \lfloor tr^4/\gamma_n\rfloor$. Then, the quadratic variation of M_n in the time interval $[0,t]$ converges to $t\sigma^2 I$ for all $t \in \mathbb{R}_+$. That is, the following convergence holds in probability:*

$$\lim_{n\to\infty}\sum_{\ell=0}^{t_{n,r}-1}\mathbb{E}\Big[\Big(\Delta M_{r,\ell,(i,j)}^{(n)}\Big)\Big(\Delta M_{r,\ell,(i',j')}^{(n)}\Big)\,\Big|\,\mathcal{F}_\ell\Big] = t\sigma^2\mathbb{1}\big\{i=i', j=j'\big\},$$

for all $(i,j),(i',j') \in [r]^{(2)}$.

Proof. We first notice that for each $k \in \mathbb{N}$ we have $\mathbb{E}\left[\left\|\widetilde{q}_{r,k+1}^{(n)} - q_{r,k}^{(n)}\right\|_2^2 \,\middle|\, \mathcal{F}_k\right] \le r^2\gamma_n^2$. Let $\mathcal{G}_k$ be the sigma algebra generated by $\mathcal{F}_k \vee \left\{p_{r,k+1}^{(n)}\right\}$. Recall that given $p_{r,k+1}^{(n)}$, the iterate $q_{r,k+1}^{(n)}$ is obtained by running ℓ_n steps of independent symmetric random walk with step size $\frac{1}{n^2}$

(with reflection at $\{0,1\}$) starting at $p^{(n)}_{r,k+1}$. Fix $(i,j) \in [r]^{(2)}$. Let S_k denote the symmetric random walk with step size $1/n^2$ run for m steps starting at $p^{(n)}_{r,k+1,(i,j)}$. Now observe that

$$\mathbb{E}\left[\left(q^{(n)}_{r,k+1,(i,j)} - p^{(n)}_{r,k+1,(i,j)}\right)^2 \,\middle|\, \mathcal{G}_k\right] = \frac{1}{n^4}\sum_{m=1}^{\ell_n} \mathbb{1}_{(0,1)}\{S_{k,m}\} + \frac{1}{2n^4}\sum_{m=1}^{\ell_n} \mathbb{1}_{\{0,1\}}\{S_{k,m}\}$$
$$= \frac{\ell_n}{n^4} - \frac{1}{2n^4}\sum_{m=1}^{\ell_n} \mathbb{1}_{\{0,1\}}\{S_{k,m}\}.$$

Set $h^{(n)}_k = \frac{1}{2n^4}\sum_{m=1}^{\ell_n} \mathbb{1}_{\{0,1\}}\{S_m\}$. Note that $\lim_{n\to\infty} \sum_{m=0}^{t_{n,r}-1} \frac{\ell_n}{n^4} = t\sigma^2$. It follows that

$$\lim_{n\to\infty}\left|\sum_{\ell=0}^{t_{n,r}-1} \mathbb{E}\left[\left(\Delta M^{(n)}_{k,\ell,(i,j)}\right)^2\right] - t\sigma^2\right| \le \lim_{n\to\infty} r^2\gamma_n^2 t_{n,r} + \lim_{n\to\infty}\sum_{\ell=0}^{t_{n,r}-1} h^{(n)}_k.$$

It is clear that $r^2\gamma_n^2 t_{n,r} \to 0$ as $n \to \infty$, and $\lim_{n\to\infty}\sum_{k=0}^{t_{n,r}-1} h^{(n)}_k = 0$ by Lemma 6.4.2.

For simplicity define $\widehat{\Delta} q^{(n)}_{r,k,(i,j)} := q^{(n)}_{r,k+1,(i,j)} - p^{(n)}_{r,k+1,(i,j)}$. If $\{i,j\} \neq \{i',j'\}$ then $\widehat{\Delta} q^{(n)}_{r,k,(i,j)}$ and $\widehat{\Delta} q^{(n)}_{r,k,(i',j')}$ are independent given $\mathcal{G}_k$. In particular,

$$\mathbb{E}\left[\widehat{\Delta} q^{(n)}_{r,k,(i,j)}\widehat{\Delta} q^{(n)}_{r,k,(i',j')} \,\middle|\, \mathcal{G}_k\right] = \mathbb{E}\left[\widehat{\Delta} q^{(n)}_{r,k,(i,j)} \,\middle|\, \mathcal{G}_k\right]\mathbb{E}\left[\widehat{\Delta} q^{(n)}_{r,k,(i',j')} \,\middle|\, \mathcal{G}_k\right]$$
$$\le \frac{1}{n^4}\sum_{m=1}^{\ell_n} \mathbb{1}_{\{0,1\}}\{S_{k,m}\},$$

where $S_{k,m}$ is as above. Using Lemma 6.4.2 we conclude that

$$\lim_{n\to\infty}\left|\sum_{\ell=0}^{t_{n,r}-1} \mathbb{E}\left[\Delta M^{(n)}_{k,\ell,(i,j)}\Delta M^{(n)}_{k,\ell,(i',j')}\right]\right| \le \lim_{n\to\infty}\sum_{\ell=0}^{t_{n,r}-1} h^{(n)}_k = 0.$$

This completes the proof. □

Away from 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 [136] which states that one can couple S with some Brownian motion B such that

$$\mathbb{P}\left\{\max_{0\le k\le T}\frac{|S_k - B(k)|}{n^2} \ge C\frac{\log T + x}{n^2}\right\} \le \mathrm{e}^{-x},$$

for any $T \in \mathbb{N}$. Taking $T = s_n$ (and ℓ_n respectively), we obtain that for n sufficiently large we have

$$\mathbb{P}\left\{\max_{0\le k\le s_n} \frac{|S_k - B(k)|}{n^2} \ge \frac{C\log n}{n^2}\right\} \le \mathbb{P}\left\{\max_{0\le k\le \ell_n} \frac{|S_k - B(k)|}{n^2} \ge \frac{C\log n}{n^2}\right\} \le \frac{1}{n^4}.$$

Further observe that for a fixed $\delta > 0$, we have that

$$\mathbb{P}\left\{\max_{0\le t\le s_n/n^4} |B(t)| \ge \delta\right\} \le \mathbb{P}\left\{\max_{0\le t\le \ell_n/n^4} |B(t)| \ge \delta\right\} \le 2\overline{\Phi}\left(\frac{\delta}{r^{-2}\sigma\sqrt{\gamma_n}}\right).$$

We combine these observations to obtain the following lemma.

Lemma 6.4.4. *Let $\widetilde{S}_k = \frac{1}{n^2} S_k$ for every $k \in \mathbb{Z}_+$. Let $\epsilon > 0$ be fixed. Then, for all $n \in \mathbb{N}$ sufficiently large, we have*

$$\mathbb{P}\left\{\max_{k\le s_n} \left|\widetilde{S}_k\right| \ge \epsilon/2\right\} \le \mathbb{P}\left\{\max_{k\le \ell_n} \left|\widetilde{S}_k\right| \ge \epsilon/2\right\} \le \frac{1}{n^4} + 2\overline{\Phi}\left(\frac{\epsilon}{4r^{-2}\sigma\sqrt{\gamma_n}}\right).$$

Lemma 6.4.5. *Let $\epsilon > 0$ fixed. Let $\ell \in \mathbb{Z}_+$ be such that $q^{(n)}_{r,\ell} \in A_\epsilon$. Then, for n sufficiently large, we have*

$$\left\|\mathbb{E}\left[\Delta q^{(n)}_{r,\ell} \;\middle|\; \mathcal{F}_\ell\right] - \mathbb{E}\left[\widetilde{\Delta} q^{(n)}_{r,\ell} \;\middle|\; \mathcal{F}_\ell\right]\right\|_{\mathrm{F}}^2 \le 2\left(\frac{r^2}{n^4} + 2r^2\overline{\Phi}\left(\frac{\epsilon}{4r^{-2}\sigma\sqrt{\gamma_n}}\right)\right),$$

with probability at least $1 - \frac{r^2}{n^4} - 2r^2\overline{\Phi}\left(\frac{\epsilon}{4r^{-2}\sigma\sqrt{\gamma_n}}\right)$.

Proof. Let $\epsilon > 0$, r, ℓ be fixed. Let $\widehat{\Delta} q^{(n)}_{r,\ell} \coloneqq q^{(n)}_{r,\ell+1} - \widetilde{q}^{(n)}_{r,\ell+1}$. Begin by observing that

$$\left\|\mathbb{E}\left[\Delta q^{(n)}_{r,\ell} \;\middle|\; \mathcal{F}_\ell\right] - \mathbb{E}\left[\widetilde{\Delta} q^{(n)}_{r,\ell} \;\middle|\; \mathcal{F}_\ell\right]\right\|_{\mathrm{F}}^2 = \left\|\mathbb{E}\left[\widehat{\Delta} q^{(n)}_{r,\ell} \;\middle|\; \mathcal{F}_\ell\right]\right\|_2^2.$$

Let $E_{n,\ell}$ be the event that $\widetilde{q}^{(n)}_{r,\ell+1} \in A_{\epsilon/2}$. Using Lemma 6.4.4 and union bound we conclude that

$$\mathbb{P}\left\{\widetilde{E}_{n,\ell}\right\} \ge 1 - \frac{r^2}{n^4} - 2r^2\overline{\Phi}\left(\frac{\epsilon}{4r^{-2}\sigma\sqrt{\gamma_n}}\right).$$

Given $p^{(n)}_{r,\ell+1}$, we observe that $\widehat{\Delta} q^{(n)}_{r,\ell}$ has the same distribution as symmetric random walk with step-size $\frac{1}{n^2}$ (reflected at boundary $\{0,1\}$) run for $\ell_{n,r}$ steps. Let us denote this j-th step of this walk by $S_{k,j}$. Also define a simple random walk with step-size $\frac{1}{n^2}$ (without reflection) $\widetilde{S}_k$ starting with the same initial condition as S_k. Given $\widetilde{q}^{(n)}_{r,\ell+1} \in A_{\epsilon/2}$, we can couple the walk S_k and $\widetilde{S}_k$ so that $S_{k,j} = \widetilde{S}_{k,j}$ for all $j \le T$ where $T = \min\left\{i \in \mathbb{Z}_+ \;\middle|\; \left|\widetilde{S}_{k,i} - \widetilde{S}_{k,0}\right| \ge \epsilon/2\right\}$.

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

$$\mathbb{E}\Big[\widehat{\Delta} q_{r,\ell}^{(n)} \,\Big|\, \mathcal{G}_\ell\Big] = \mathbb{E}\Big[\widehat{\Delta} q_{r,\ell}^{(n)} \mathbb{1}_{T\le \ell_n} \,\Big|\, \mathcal{G}_\ell\Big] + \mathbb{E}\Big[\widehat{\Delta} q_{r,\ell}^{(n)} \mathbb{1}_{T> \ell_n} \,\Big|\, \mathcal{G}_\ell\Big].$$

using the bound $\widehat{\Delta} q_{r,\ell}^{(n)} \le 1$ and Lemma 6.4.4 we have that

$$\Big\|\mathbb{E}\Big[\widehat{\Delta} q_{r,\ell}^{(n)} \mathbb{1}_{T\le \ell_n} \,\Big|\, \mathcal{G}_\ell\Big]\Big\|_{\mathrm{F}}^2 \le \frac{r^2}{n^4} + 2r^2\overline{\Phi}\bigg(\frac{\epsilon}{4r^{-2}\sigma\sqrt{\gamma_n}}\bigg).$$

On the other hand, using the fact that $\mathbb{E}\Big[\widetilde{S}_{k,\ell_n} \,\Big|\, \mathcal{G}_\ell\Big] = 0$, we obtain

$$\begin{aligned}
\Big\|\mathbb{E}\Big[\widehat{\Delta} q_{r,\ell}^{(n)} \mathbb{1}_{T> \ell_n} \,\Big|\, \mathcal{G}_\ell\Big]\Big\|_{\mathrm{F}}^2 &= \Big\|\mathbb{E}\Big[\widetilde{S}_{k,\ell_n} \mathbb{1}_{T> \ell_n} \,\Big|\, \mathcal{G}_\ell\Big]\Big\|_{\mathrm{F}}^2 \\
&= \Big\|\mathbb{E}\Big[\widetilde{S}_{k,\ell_n} \mathbb{1}_{T> \ell_n} \,\Big|\, \mathcal{G}_\ell\Big] - \mathbb{E}\Big[\widetilde{S}_{k,\ell_n} \,\Big|\, \mathcal{G}_\ell\Big]\Big\|_{\mathrm{F}}^2 \\
&= \Big\|\mathbb{E}\Big[\widetilde{S}_{k,\ell_n} \mathbb{1}_{T\le \ell_n} \,\Big|\, \mathcal{G}_\ell\Big]\Big\|_{\mathrm{F}}^2 \le r^{-2}\sigma^2\gamma_n\bigg(\frac{r^2}{n^4} + 2r^2\overline{\Phi}\bigg(\frac{\epsilon}{4r^{-2}\sigma\sqrt{\gamma_n}}\bigg)\bigg).
\end{aligned}$$

Thus, we conclude that for n sufficiently large we have

$$\Big\|\mathbb{E}\Big[\widehat{\Delta} q_{r,\ell}^{(n)} \,\Big|\, \mathcal{F}_\ell\Big]\Big\|_{\mathrm{F}}^2 \le 2\bigg(\frac{r^2}{n^4} + 2r^2\overline{\Phi}\bigg(\frac{\epsilon}{4r^{-2}\sigma\sqrt{2\gamma_n}}\bigg)\bigg).$$

completing the proof. □

Lemma 6.4.6. *Let $r \in \mathbb{N}$, for any $k \in \mathbb{Z}_+$, let $q_{r,k}^{(n)}$ and $\widetilde{q}_{r,k+1}^{(n)}$ be as defined in Step 1 of our algorithm in Section 6.3. Then, there exists a universal constant $c > 0$ such that for all $n \in \mathbb{N} \setminus \big[\big[\mathrm{e}^{cr/4}\big]\big]$, we have*

$$\Big\|K\Big(\widetilde{q}_{r,k+1}^{(n)}\Big) - K\Big(q_{r,k}^{(n)}\Big)\Big\|_2^2 \le 4\gamma_n^2 + (16/c)\gamma_n^2 \log n =: e_n \le (32/c)\gamma_n^2 \log n,$$

with probability at least $1 - 2n^{-4}$.

Proof. For every $i, j \in [r]$, let $(S_{i,j,k})_{k\in\mathbb{Z}_+}$ denote the 1-dimensional symmetric random walk with step-size $\frac{1}{n^2}$ starting at 0. Let these random walks be independent up to the double index symmetry for indices $(i, j) \in [r]^{(2)}$. Recall that $s_n = \big\lceil \gamma_n^2 n^4 \big\rceil$, and note that for any $i, j \in [r]$, given $q_{r,k,(i,j)}^{(n)}$ we have

$$\widetilde{q}_{r,k+1,(i,j)}^{(n)} - q_{r,k,(i,j)}^{(n)} \overset{\mathrm{d}}{=} \mathrm{Sko}(S_{s_n}).$$

Since the Skorokhod map is 4-Lipschitz, we conclude that

$$\mathbb{P}\left\{\left\|K\left(\widetilde{q}_{r,k+1}^{(n)}\right)-K\left(q_{r,k}^{(n)}\right)\right\|_2^2\geq e_n\right\}\leq\mathbb{P}\left\{\frac{1}{r^2}\sum_{(i,j)\in[r]^{(2)}}(S_{i,j,s_n})^2\geq e_n/4\right\}. \tag{6.39}$$

We will now show that the quantity $\frac{1}{r^2}\sum_{i,j\in[r]^{(2)}}(S_{i,j,s_n})^2$ is concentrated near its expectation, that is $s_n\cdot\left(\frac{1}{n^2}\right)^2$. From the Hanson-Wright concentration inequality [189],

$$\begin{aligned}\mathbb{P}\left\{\left|\frac{1}{r^2}\sum_{i,j\in[r]^{(2)}}(S_{i,j,s_n})^2-\gamma_n^2\right|>t_n\right\}&\leq\mathbb{P}\left\{\left|\frac{1}{r^2}\sum_{i,j\in[r]^{(2)}}(S_{i,j,s_n})^2-s_nn^{-4}\right|>t_n\right\}\\&\leq 2\exp\left(-c\min\left\{\frac{t_n^2}{\left(\frac{1}{n^2}\right)^4rs_n^2},\frac{t_n}{\left(\frac{1}{n^2}\right)^2s_n}\right\}\right)\leq 2\exp\left(-c\min\left\{\frac{t_n^2}{r\gamma_n^4},\frac{t_n}{\gamma_n^2}\right\}\right),\end{aligned} \tag{6.40}$$

for every $t_n\geq 0$, for some universal constant $c>0$. Let us consider $t_n\geq r\gamma_n^2$. Then, the above probability becomes $2\exp\left(-ct_n/\gamma_n^2\right)$. Moreover, for $n\geq \mathrm{e}^{cr/4}$ if we choose $t_n=(4/c)\gamma_n^2\log n$, we have that for all $(i,j)\in[r]^{(2)}$,

$$\frac{1}{r^2}\sum_{i,j\in[r]^{(2)}}(S_{i,j,s_n})^2\leq\gamma_n^2+(4/c)\gamma_n^2\log n,$$

with probability at least $1-2n^{-4}$, for $e_n\coloneqq 4\gamma_n^2+(16/c)\gamma_n^2\log n\leq(32/c)\gamma_n^2\log n$. □

Lemma 6.4.7. *Let $\epsilon>0$ be fixed, e_n be as defined in Lemma 6.4.6, and let $q_{r,\ell}^{(n)}\in A_\epsilon$ where A_ϵ is defined in (6.25). Then,*

$$\left\|\mathbb{E}\left[\Delta q_{r,\ell}^{(n)}\,\middle|\,\mathcal{F}_\ell\right]-\gamma_nr^{-4}b_r\left(q_{r,\ell}^{(n)}\right)\right\|_{\mathrm{F}}^2\leq\frac{Cr^2}{n^4}+4r^2\beta_{n,r}^2e_n^3\max\{\lambda,L\}+2r^2\overline{\Phi}\left(\frac{\epsilon}{4r^{-2}\sigma\sqrt{\gamma_n}}\right),$$

with probability at least $1-\frac{2}{n^4}-\frac{2r^2}{n^4}-4r^2\overline{\Phi}\left(\frac{\epsilon}{4r^{-2}\sigma\sqrt{\gamma_n}}\right)$.

Proof. Let $I\coloneqq\left\|K\left(\widetilde{q}_{r,\ell+1}^{(n)}\right)-K\left(q_{r,\ell}^{(n)}\right)\right\|_2^2$ and let $A_{n,\ell}$ be the event that $\{I\leq e_n\}$. In the following we will work on this event. Set

$$J\coloneqq\frac{\exp\left(-\beta_{n,r}\left[\mathcal{H}\left(K\left(\widetilde{q}_{r,\ell}^{(n)}\right)\right)-\mathcal{H}\left(K\left(q_{r,\ell}^{(n)}\right)\right)\right]^+\right)}{\exp\left(-\beta_{n,r}\left\langle D\mathcal{H}\left(K\left(q_{r,\ell}^{(n)}\right)\right),K\left(\widetilde{q}_{r,\ell+1}^{(n)}\right)-K\left(q_{r,\ell}^{(n)}\right)\right\rangle^+\right)},$$

From our assumption on $(\gamma_n)_{n\in\mathbb{N}}$, we have that for sufficiently large n, $\beta\gamma_n\log^2 n\leq 1$. Notice that by Assumption 10, we have

$$1-2\beta_{n,r}\lambda I\leq\exp(-\beta_{n,r}\lambda I)\leq J\leq\exp(\beta_{n,r}LI)\leq 1+2\beta_{n,r}LI,$$

if $\beta_{n,r}\lambda I, \beta_{n,r}LI \leq 1$, i.e., when n is sufficiently large. Define $b_r^{(n)}$ at $q_{r,\ell}^{(n)}$ as

$$\mathbb{E}\Big[\Big(\widetilde{q}_{r,\ell+1}^{(n)} - q_{r,\ell}^{(n)}\Big)\exp\Big(-\beta_{n,r}\Big\langle D\mathcal{H}\Big(K\Big(q_{r,\ell}^{(n)}\Big)\Big), K\Big(\widetilde{q}_{r,\ell+1}^{(n)}\Big) - K\Big(q_{r,\ell}^{(n)}\Big)\Big\rangle^{+}\Big)\,\Big|\,\mathcal{F}_k\Big].$$

Then, on the event $A_{n,\ell}$ we have $\Big\|\mathbb{E}\Big[\widetilde{\Delta}q_{r,\ell}^{(n)}\,\Big|\,\mathcal{F}_k\Big] - b_r^{(n)}\Big(q_{r,\ell}^{(n)}\Big)\Big\|_{\mathrm{F}}^2 \leq 4r^2\beta_{n,r}^2 e_n^3\max\{\lambda, L\}$.

Let $E_{n,\ell}$ be the event as in the proof of Lemma 6.4.4. On this event, we have

$$\Big\|\mathbb{E}\Big[\Delta q_{r,\ell}^{(n)}\,\Big|\,\mathcal{F}_\ell\Big] - \mathbb{E}\Big[\widetilde{\Delta}q_{r,\ell}^{(n)}\,\Big|\,\mathcal{F}_\ell\Big]\Big\|_{\mathrm{F}}^2 \leq C\bigg(\frac{r^2}{n^4} + 2r^2\overline{\Phi}\bigg(\frac{\epsilon}{4r^{-2}\sigma\sqrt{\gamma_n}}\bigg)\bigg). \tag{6.41}$$

Moreover, on the event $E_{n,\ell}$ we also have that given $\mathcal{F}_k$, the coordinates of the $r\times r$ symmetric matrix $\Big(\widetilde{q}_{r,\ell+1}^{(n)} - q_{r,\ell}^{(n)}\Big)$ are i.i.d. and have the same distribution as $\widetilde{S}_{s_n}$.

Let $\widetilde{Y}_n$ be $r\times r$ matrix with independent entries such that $\widetilde{Y}_{n,(i,j)}$ be increment of the symmetric random walk (without reflection) of step-size n^{-2} starting from $q_{r,\ell,(i,j)}^{(n)}$ run for $s_n = \lceil\gamma_n^2 n^4\rceil$ steps. Let B_r be an $r\times r$ symmetric matrix of standard Brownian motions. On the event $E_{n,\ell}\cap A_{n,\ell}$, we use the Berry-Esseen lemma (see [177, Theorem 16]) with a union bound to obtain $\mathbb{W}_2^2\Big(\widetilde{Y}_n, B_r(\gamma_n^2)\Big) \leq \frac{Cr^2}{n^4}$, for some universal constant $C>0$.

Let $\nabla H_r\Big(q_{r,\ell}^{(n)}\Big) = V$. Define a function $G(Y) \coloneqq Y\exp\big(-\beta_{n,r}\langle V, Y\rangle_{\mathrm{F}}^{+}\big)$. Note that G is a bounded Lipschitz function of Y. Observe that $b_r^{(n)}\Big(q_{r,\ell}^{(n)}\Big) = \mathbb{E}\Big[G\Big(\widetilde{Y}_n\Big)\Big]$. On the other hand, we know that $\gamma_n r^{-4}b_r\Big(q_{r,\ell}^{(n)}\Big) = \mathbb{E}[G(B_r(\gamma_n))]$. We conclude that on the event $E_{n,\ell}\cap A_{n,\ell}$ we have

$$\Big\|\mathbb{E}\Big[\widetilde{\Delta}q_{r,\ell}^{(n)}\,\Big|\,\mathcal{F}_\ell\Big] - \gamma_n r^{-4}b_r\Big(q_{r,\ell}^{(n)}\Big)\Big\|_{\mathrm{F}}^2 \leq \frac{Cr^2}{n^4} + 4r^2\beta_{n,r}^2 e_n^3\max\{\lambda, L\}.$$

The conclusion follows by using (6.41) and noticing that the

$$\mathbb{P}\{E_{n,\ell}\cap A_{n,\ell}\} \geq 1 - \frac{2}{n^4} - \frac{2r^2}{n^4} - 4r^2\overline{\Phi}\bigg(\frac{\epsilon}{4r^{-2}\sigma\sqrt{\gamma_n}}\bigg).$$

□

Chapter 7

SCALING LIMIT OF ITERATED MATRIX PRODUCTS

7.1 Introduction

Beginning with the seminal work of Bellman [23], Furstenberg and Kesten [90] and Berger [26] there is a significant work on the law of large number and CLT type results for (the entries of the) product of random matrices and operators [89, 207, 213, 121, 206, 78]. Concentration inequalities for the product of random matrices have also been extensively studied [81, 17, 130, 105, 59]. Iterated products of matrices have been studied in the context of random walks on groups [145, 88, 25]. In this paper, we study the scaling limit for the iterated products of triangular arrays of matrices in large dimensions.

To state the problem, let us consider a triangular array of $n \times n$ (possibly random) matrices $\left(\left(A_{n,k}^{(m)}\right)_{k\in[m]}\right)_{m\in\mathbb{N}}$ and define the following iterated product of matrices

$$P_n^{(m)}(k) := \left(I_n + \frac{\mu_n}{m} A_{n,k}^{(m)}\right) \dots \left(I_n + \frac{\mu_n}{m} A_{n,2}^{(m)}\right) \left(I_n + \frac{\mu_n}{m} A_{n,1}^{(m)}\right), \qquad k \in [m], \tag{7.1}$$

where μ_n is a dimension-dependent scaling factor. We set $P_n^{(m)}(0) = I_n$. Our goal is to establish the scaling limit for $P_n^{(m)}$ in equation (7.1) as $m, n \to \infty$. In the following, we first explain the scaling limit as $m \to \infty$ and then consider the limit as $n \to \infty$. The role of μ_n becomes important only when we consider the limit as $n \to \infty$. Therefore, in the following discussion, we fix $\mu_n = 1$.

A particularly important situation where such matrix products arise in more generality is deep residual Neural Networks (NNs). Particularly, a deep residual NN with linear activation and $m \in \mathbb{N}$ layers consists of a sequence of m matrices $(A_{n,k})_{k\in[m]}$. More generally, instead of a scaling of $\frac{1}{nm}$ as described in equation (7.1), one can consider $\tau_{n,m} \in \left\{\frac{1}{nm}, \frac{1}{n\sqrt{m}}, \frac{1}{\sqrt{nm}}\right\}$ and consider the following product. Given an input $H_0 \in \mathbb{R}^n$, the k-th layer of the NN computes $H_k \in \mathbb{R}^n$ as $H_k = H_{k-1} + \tau_{n,m} A_{n,k} H_{k-1} = (I_n + \tau_{n,m} A_{n,k}) H_{k-1}$, for $k \in [m]$. The output of the NN is simply the average of H_m. Different choices of scaling

of $\tau_{n,m}$ with respect to n and m lead to different parameterizations of the NN. We refer the reader to [216, 217] and references within for the importance and implications of some of the choices. We also refer the readers to [104, 64, 191, 101, 94, 6, 148, 218, 194, 33, 120] for recent literature on deep neural network where the evolution of neurons are thought of as appropriate Gaussian processes.

First, let us consider the scaling limit of $P_n^{(m)}$, fixing the dimension $n \in \mathbb{N}$, as $m \to \infty$. As n is fixed, we will drop μ_n for simplicity in the following discussion. Note that $P_n^{(m)}$ satisfies following difference equation

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

It is reasonable to expect that $P_n^{(m)}$ admits a scaling limit. That is, under appropriate conditions on the curve A_n defined as $A_n(t) := \lim_{m\to\infty} A_{n,\lfloor mt \rfloor}^{(m)}$ for $t \in [0,1]$, we should expect that the curve $P_{n,m}$ defined as $P_{n,m}(t) := P_n^{(m)}(\lfloor mt \rfloor)$ for $t \in [0,1]$, converges to an absolutely continuous curve, say P_n, satisfying $\frac{\mathrm{d}}{\mathrm{d}t} P_n(t) = A_n(t) P_n(t)$ as $m \to \infty$. If $A_n(t) \equiv A_n$ is a constant curve, then the solution to this differential equation is $P_n(t) = \mathrm{e}^{tA_n}$. For a more general curve A_n, one may guess the solution of the above differential equation to be $P_n(t) = \mathrm{e}^{\int_0^t A_n(s)\,\mathrm{d}s}$. 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 analog of exponential that we define later.

Now 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 (7.1), we consider the iterated product of matrices such that

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

Following a similar heuristic as above, one may expect that $P_{n,m}$ converges to a matrix valued SDE satisfying $\mathrm{d}P_n(t) = \mathrm{d}B_n(t) P_n(t)$ 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 7.2.4 we consider a general class of triangular sequence of random matrices that encompasses both the cases above. That is, we consider

$$P_n^{(m)}(k) := \left(I_n + X_{n,k}^{(m)}\right) \ldots \left(I_n + X_{n,2}^{(m)}\right)\left(I_n + X_{n,1}^{(m)}\right), \quad k \in [m], \qquad P_n^{(m)}(0) := I_n,$$

where

$$X_{n,k}^{(m)} := \frac{\mu_n}{m} M_{n,k}^{(m)} + \frac{\sigma_n}{\sqrt{m}} G_{n,k}^{(m)}, \qquad k \in [m], \tag{7.2}$$

such that $\mathbb{E}\left[M_{n,k}^{(m)}\right] = A_{n,k}^{(m)}$ for every $k \in [m]$. 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 7.2.1). Let $P_{n,m}(t) := P_n^{(m)}(\lfloor mt \rfloor)$, and $A_n(t) := \lim_{m\to\infty} A_{n,\lfloor mt \rfloor}^{(m)}$, for $t \in [0,1]$, as also defined earlier. We state below an informal version of Theorem 7.2.4.

Theorem 7.1.1 (Informal Theorem 7.2.4)**.** *The curve $P_{n,m}$ uniformly converges to $\mathrm{Texp}[Y_n]$, where $Y_n(t) = \mu_n \int_0^t A_n(s)\,\mathrm{d}s + \sigma_n B_n(t)$, for $t \in [0,1]$.*

Various authors have studied similar problems – in fixed dimension $n \in \mathbb{N}$. For instance, it is shown in [81] that if

$$Q_{n,k}^{(m)} = \left(I_n + \frac{1}{m} A_{n,k}\right) \ldots \left(I_n + \frac{1}{m} A_{n,2}\right) \left(I_n + \frac{1}{m} A_{n,1}\right), \quad k \in [m], \tag{7.3}$$

then $Q_{n,m}^{(m)}$ converges to e^{A_n} where $A_n := \lim_{m\to\infty} \frac{1}{m} \sum_{k=1}^m A_{n,k}$ (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 [105, 130]. 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 $A_{n,k}^{(m)} := A_{n,k}$ for all $k \in [m]$ and all $m \in \mathbb{N}$, then our result (Theorem 7.2.4) recovers this result (see Example 16). 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 $A_n := \lim_{m\to\infty} \frac{1}{m} \sum_{k=1}^m A_{n,k}^{(m)}$ exists and also the limit $P_n(t) := \lim_{m\to\infty} P_n^{(m)}(\lfloor mt \rfloor)$ exists, it is not necessarily true that $P_n(t) = \mathrm{e}^{tA_n}$.

Our second problem concerns the iterated matrix product like equation (7.3) in very large dimension (that is, as $n \to \infty$). In the view of Theorem 7.2.4, 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 (7.1) becomes important here.

Let $\mathcal{M}_n$ denote the set of all $n \times n$ matrices. Notice in the simple case when $A_{n,k}^{(m)} = A_n$ for all $k \in [m]$ and $m \in \mathbb{N}$ for some fixed $A_n \in \mathcal{M}_n$. Here, as $m \to \infty$, the product $Q_{n,m}$ in equation (7.3) 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 $\mu_n = \frac{1}{n}$ in dimension as in equation (7.1), then $\mathrm{e}^{A_n/n} = I_n + O\big(n^{-1}\big)$ 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. It is also instructive to consider the case when A_n is a matrix with, say, i.i.d. standard Gaussian entries. In this case, one can see that $n\mathcal{E}_n = A_n + O\big(\frac{1}{n}\big)$. Therefore, it may have a non-trivial limit.

Now, we come to meaning of limit. Consider the previous case, that is, where A_n is a matrix of i.i.d. Gaussian entries and consider the matrix $n\mathcal{E}_n = n\big(\mathrm{e}^{A_n/n} - I_n\big) = A_n + O(\frac{1}{n})$. 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. More generally, we define an IEA as a random variable X taking values in $\mathbb{R}^{\mathbb{N}\times\mathbb{N}}$, that is, $X = (X_{i,j})_{(i,j)\in\mathbb{N}^2}$ such that $\mathrm{Law}(X) = \mathrm{Law}(X^\sigma)$ where $X^\sigma_{i,j} = X_{\sigma(i),\sigma(j)}$ for any finite permutation σ of $\mathbb{N}$. Naturally, we say a matrix $A_n \in \mathcal{M}_n$ is exchangeable if $\mathrm{Law}(A_n) = \mathrm{Law}(A_n^\sigma)$ for all permutations σ of $[n]$. Note that given any matrix A_n, one can obtain an exchangeable random matrix $\widetilde{A}_n = A_n^\sigma$ where σ is a permutation of $[n]$ chosen uniformly at random. We can now define a notion of limit of a sequence of matrices $(A_n \in \mathcal{M}_n)_{n\in\mathbb{N}}$ as $n \to \infty$. We say that A_n converges to an IEA X if for every $r \in \mathbb{N}$, the $r \times r$ exchangeable matrix $A_n\{r\} \coloneqq (A_n(x_i, x_j))_{(i,j)\in[r]^2}$ where $x_1, \ldots, x_r$ is chosen uniformly at random from all r-subsets of $[n]$ converges weakly to $X[r] \coloneqq (X_{i,j})_{(i,j)\in[r]^2}$.

Before, we state our second main result, we consider another example. Let G_n be a matrix with i.i.d. Gaussian coordinates, we can also 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 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 $\left(n^{1/2}\mathcal{E}_n\right)_{n\in\mathbb{N}}$ 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 7.2.4, we consider the limits of $\mathrm{Texp}[Y_n]$ – with suitable scaling and centering – for a semimartingale $Y_n(t) = \mu_n \int_0^t A_n(s)\,\mathrm{d}s + \sigma_n B_n(t)$, $t \in [0,1]$. Now we state an informal version of Theorem 7.2.6 – but before that we some definitions and notations. A *kernel* W is 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. 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 A_n is a $n \times n$ matrix, we will write $A_n\{r\}$ for $K(A_n)\{r\}$.

Theorem 7.1.2 (Informal Theorem 7.2.6)**.** *Let $Y_n(t) = \mu_n \int_0^t A_n(s)\,\mathrm{d}s + \sigma_n B_n(t)$ be a semimartingale. Let W be a continuous curve of kernel such that $\sup_{s\in[0,t]} \|K(A_n)(s) - W(s)\|_2 \to 0$ as $n \to \infty$ for every $t \in [0,1]$. Let $\mathcal{E}_n = \mathrm{Texp}[Y_n] - I_n$. Then, there exists a curve of kernels $t \mapsto \Gamma(t)$ such that*

1. *When $\mu_n = \sigma_n = n^{-1}$, then $(n\mathcal{E}_n)_{n\in\mathbb{N}}$ converges to an IEA $\Gamma(t)\{\infty\} + B_\infty$.*

2. *When $\mu_n = \sigma_n^2 = n^{-1}$, then $n^{1/2}\mathcal{E}_n = n^{1/2}(\mathrm{Texp}[\sigma_n B_n] - I_n) + O\big(n^{-1/2}\big)$ converges, as $n \to \infty$, to an IEA X satisfying $X(t) = B_\infty(\mathrm{e}^t - 1)$ for $t \in [0,1]$.*

3. *When $\mu_n = \sigma_n^2 = n^{-1}$, then $n^{1/2}\big(n^{1/2}\mathcal{E}_n - n^{1/2}(\mathrm{Texp}[\sigma_n B_n] - I_n)\big)$ converges to an IEA with Gaussian entries with mean as in the case 1, i.e., $\Gamma(t)\{\infty\}$ (albeit with a non-trivial covariance).*

7.2 Setup and Main Results

7.2.1 Iterated matrix product in fixed dimension

We now explore the scaling limit of iterated product of matrices as defined in (7.1) as $m \to \infty$ while the dimension n is kept fixed. As throughout this section we keep the dimension n fixed, we will drop the scaling depending on n whenever convenient. We begin with some definitions and notations.

Definition 7.2.1 (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 $\Delta_k(t) := \{(s_1, \ldots, s_k) \in [0,t]^k \mid t \geq s_k \geq s_{k-1} \geq \ldots s_1 \geq 0\}$ and define*

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

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

$$\mathrm{Texp}[Y_n](t) := \sum_{k=0}^{\infty} J_k(Y_n)(t), \qquad t \in \mathbb{R}_+. \tag{7.4}$$

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 14.

Definition 7.2.2 (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 [43] for more detail on Poisson point processes.*

Example 14. 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

$$\begin{aligned} J_k(Y_n)(t) &= \int_{\Delta_k(t)} A_n(s_k)A_n(s_{k-1}) \ldots A_n(s_1) \prod_{j=1}^{k} \mathrm{d}s_j \\ &= |\Delta_k(t)| \int_{\Delta_k(t)} A_n(s_k)A_n(s_{k-1}) \ldots A_n(s_1)\, \mathrm{d}\sigma_{k,t}(s_k, \ldots, s_1) \\ &= \mathrm{e}^t \mathbb{E}\left[\prod_{\alpha \in N_t} A_n(\alpha) \,\middle|\, |N_t| = k \right] \mathbb{P}\{|N_t| = k\}, \end{aligned}$$

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 7.2.3. *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 $t \in \mathbb{R}_+$.

We consider the product defined in equation (7.2), 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

$$P_{n,m}(t) \coloneqq \prod_{k=1}^{\lfloor mt \rfloor} \left(I_n + X_{n,k}^{(m)}\right), \qquad m \in \mathbb{N}, \quad t \in [0,1]. \tag{7.5}$$

Assumption 11. *There exists $C, D \geq 0$ such that for every $n \in \mathbb{N}$,*

1. *For every $m \in \mathbb{N}$, and $k \in [m]$, $\mathrm{Cov}\left(M_{n,k}^{(m)}, \preceq\right) nDI_n$.*
2. *For every $m \in \mathbb{N}$, and $k \in [m]$, the absolute value of elements in $A_{n,k}^{(m)}$ is at most C.*
3. *The piecewise constant interpolation $A_{n,m}$ 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 converges to a continuous curve A_n as $m \to \infty$.*
4. *For every $m \in \mathbb{N}$, $\left(G_{n,k}^{(m)}\right)_{k\in[m]}$ is a sequence of i.i.d. Wigner matrices with i.i.d. standard Gaussian entries.*

With the required assumptions stated above, we now state our first main result.

Theorem 7.2.4 (Convergence to SDE for fixed dimension)**.** *Let* $\left(\left(X_{n,k}^{(m)}\right)_{k\in[m]}\right)_{m\in\mathbb{N}}$ *be the triangular array defined in equation* (7.2)*. Under Assumption 11, the curve* $P_{n,m}$ *(as defined in equation* (7.5)*) 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.*

Example 15. 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 7.2.4 that $P_m(t)$ converges to $\mathrm{e}^{B_t-t/2}$ where B_t is a standard BM (compare with [79]).

Example 16. 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 7.2.4 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 [81].

7.2.2 Iterated matrix product in large dimension

We now turn towards the limit of (7.1) as the dimension of the matrix n grows to infinity. In the view of Theorem 7.2.4, it makes sense to consider the limit of $\mathrm{Texp}[Y_n]$ as $n \to \infty$ for semimartingale $t \mapsto Y_n(t) := \mu_n \int_0^t A_n(s)\,\mathrm{d}s + \sigma_n B_n(t)$. As we explain in Remark 7.2.5, the two interesting choices of the scaling are: $\mu_n = \sigma_n = n^{-1}$ and $\mu_n = \sigma_n^2 = n^{-1}$. For the following discussion, we will work with the case $\mu_n = \sigma_n = \frac{1}{n}$ for concreteness. As $\mathrm{Texp}[Y_n]$ is a curve in $\mathcal{M}_n$, we first need to define a notion for the limits of matrices as their dimension $n \to \infty$. There are at least two natural ways to go about this.

I Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of $n \times n$ matrices. Let $\sigma \in S_n$ be a permutation chosen uniformly at random and let $\widetilde{A}_n = A_n^\sigma$ where $A^\sigma(i,j) = A(\sigma(i),\sigma(j))$ for $(i,j) \in [n]^2$. Thus, $\widetilde{A}_n$ is an exchangeable matrix, that is, $\mathrm{Law}\Big(\widetilde{A}_n\Big) = \mathrm{Law}\Big(\widetilde{A}_n^\sigma\Big)$ for any $\sigma \in S_n$. Let Ω be a Polish space. Equip $\Omega^{\mathbb{N}^2}$ with the usual product sigma algebra and define an Ω-valued *infinite exchangeable array* (IEA) X as random variable taking values in $\Omega^{\mathbb{N}^2}$ such that $\mathrm{Law}(X^\sigma) = \mathrm{Law}(X)$ for every finite permutation σ of $\mathbb{N}$. For our purposes $\Omega = \mathbb{R}$ or $\Omega = C(\mathbb{R})$. And, we will drop the mention of Ω wherever it is convenient. It is natural to consider the limit of $(A_n)_{n\in\mathbb{N}}$ to be an IEA. We refer to Section 2.1 for details, but roughly, we say that $A_n \in \mathcal{M}_n$ converges to X as $n \to \infty$ if $A_n[r] := \Big(\widetilde{A}_n(i,j)\Big)_{(i,j)\in[r]^2}$ converges weakly to $X[r] := (X(i,j))_{(i,j)\in[r]^2}$ for every $r \in \mathbb{N}$.

Example 17. 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 [184, 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 $\alpha = (i_1, \dots, i_{k-1}) \in [n]^{k-1}$, then

$$U_{\alpha,k}(t) = \int_0^t \mathrm{d}B(s_k)(i, i_{k-1}) \int_0^{s_k} \mathrm{d}B(s_{k-1})(i_{k-1}, i_{k-2}) \dots \int_0^{s_2} \mathrm{d}B(s_1)(i_1, j).$$

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.

An IEA is also intimately related to kernels and graphons as we explain briefly. We refer the reader to Section 2.1 for the detailed discussion. A *kernel* is a measurable map $W \colon [0,1]^2 \to \mathbb{R}$. For most of our discussion, we consider the kernels that are bounded, say $|W(x,y)| \leq 1$ for a.e. $(x,y) \in [0,1]^2$. Given a kernel W, one can construct an infinite exchangeable array as follows. Let $\{U_i\}_{i \in \mathbb{N}}$ be a collection of i.i.d. $\mathrm{Uni}([0,1])$ random variables. Define $X_{i,j} \coloneqq W(U_i, U_j)$ for $(i,j) \in \mathbb{N}^2$. Then, $X = (X_{i,j})_{i,j \in \mathbb{N}}$ is an IEA.

Finally, for two kernels U, V, we define the kernel $U \cdot V$ as $(U \cdot V)(x,y) \coloneqq \int_0^1 U(x,z)V(z,y)\,\mathrm{d}z$ 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

$$J_k(U)(t) \coloneqq \int_{\Delta_k(t)} W(s_k) \cdot W(s_{k-1}) \cdot W(s_1)\,\mathrm{d}s_k \cdots \mathrm{d}s_1.$$

Also observe that $n \times n$ matrix A_n can be naturally identified with a kernel $K(A_n)$ defined as $K(A_n)(x,y) = A_n(i,j)$ if $(x,y) \in (i-1/n, i/n] \times (j-1/n, j/n]$ for some $(i,j) \in [n]^2$.

Example 18. 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 $\mathcal{E}_n \coloneqq \mathrm{Texp}[Y_n](t) - I_n$. 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 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 $\Gamma(t)(U_1,U_2)$ where $\Gamma(t)$ is the kernel $\Gamma(t) \coloneqq \sum_{k=1}^{\infty} J_k(U)(t)$.

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 $\Gamma(t)\{r\}$. In other words, $(n\mathcal{E}_n)_{n\in\mathbb{N}}$ converges to IEA X defined as $X(t) = \Gamma(t)\{\infty\}$ as $n \to \infty$.

It would be important for the later use to define a curve $\Gamma(U)$ for a curve of kernels $U\colon t \mapsto \int_0^t W(s)\,\mathrm{d}s$ as

$$\Gamma(U)(t) \coloneqq \sum_{k=1}^{\infty} J_k(U)(t) = (\mathrm{e}^t - 1)\mathbb{E}\left[\prod_{s\in N_t} W(s) \,\middle|\, |N_t| \geq 1\right], \tag{7.6}$$

and similarly $\Gamma(Y_n)$ for a curve of matrices $Y_n\colon t \mapsto \int_0^t \mu_n A_n(s)\,\mathrm{d}s$ as

$$\Gamma(Y_n)(t) \coloneqq \sum_{k=1}^{\infty} J_k(Y_n)(t), \tag{7.7}$$

where N_t is a Poisson point process with unit intensity on interval $[0,t]$ for $t \in [0,1]$.

Remark 7.2.5. *Notice the difference in the scaling of $n \times n$ matrices in Example 17 and Example 18. In Example 18 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 17 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* $\mathbf{0}$ *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.*

We now state our second main result.

Theorem 7.2.6 (IEA convergence)**.** *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 $\mathcal{E}_n(t) \coloneqq \mathrm{Texp}[Y_n](t) - I_n$ *for* $t \in [0,1]$*. Let* $W \in C\big([0,1], L^2([0,1]^2)\big)$ *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* $(n\mathcal{E}_n(t))_{n\in\mathbb{N}}$ *converges, as* $n \to \infty$*, to an IEA* $X(t) = \Gamma(U)(t)\{\infty\} + B_\infty(t)$*, where* $B_\infty(t)$ *is an IEA with i.i.d. BM coordinates and is independent of* $\Gamma(U)(t)\{\infty\}$*.*

2. *If* $\mu_n = \sigma_n^2 = n^{-1}$*, then* $n^{1/2}\mathcal{E}_n = n^{1/2}(\mathrm{Texp}[\sigma_n B_n] - I_n) + O\big(n^{-1/2}\big)$ *and it converges, as* $n \to \infty$*, to an IEA* X *satisfying* $X(t) = B_\infty(\mathrm{e}^t - 1)$ *for* $t \in [0,1]$*.*

3. *If* $\mu_n = \sigma_n^2 = n^{-1}$*, then* $n^{1/2}\big(n^{1/2}\mathcal{E}_n - n^{1/2}(\mathrm{Texp}[\sigma_n B_n] - I_n)\big)$ *converges, as* $n \to \infty$*, to an IEA* X *satisfying* $X(t) = \Gamma(U)(t)\{\infty\} + Z(t)$*, where* $Z(t)$ *is a zero mean IEA with explicit covariance described in the proof.*

Notice that limiting IEA in case 2 in the 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.

II 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 $L^2_n([0,1]) \coloneqq \left\{ f \in L^2([0,1]) \,\middle|\, f \text{ is constant a.e. on } (i-1/n, i/n] \right\}$. 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 $\left(T_{K(A_n)}\right)_{n\in\mathbb{N}}$ converges to T in strong sense. Note that even if $\left(T_{K(A_n)}\right)_{n\in\mathbb{N}}$ 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 (see Section 2.1 for more detail).

Theorem 7.2.7 (Operator convergence)**.** *For every $n \in \mathbb{N}$, let Y_n be a semimartingale 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.*

If $K(A_n)(t)$ converges to a kernel $W(t)$ in the cut metric, uniformly over compact subsets of time t as $n \to \infty$, then $A_n(t)$ converges to operators $T(t) \coloneqq T_{W(t)}$ as $n \to \infty$ [150, Lemma 8.12], i.e., $\big\|T_{K(A_n)}(t) - T_W(t)\big\|_{\mathrm{op}} \leq \|K(A_n)(t) - W(t)\|_{\square}^{1/4}$. 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 7.2.7.

Corollary 7.2.8. *Let* Y_n *be as in Theorem 7.2.7. 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* $L^2([0,1])$.

Before we begin the proofs, we end this section with the following remarks.

Remark 7.2.9 (IEA approximation)**.** *Following Theorem 7.2.6, 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 7.2.10. *Following Remark 7.2.9, 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 Section 7.3.2) 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 $O\big(n^{-1}\|h_0\|_2\big)$ 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 a 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.*

7.3 Proofs

Proof of Proposition 7.2.3. The existence and uniqueness of the solution follows from the standard arguments for vector valued SDEs [128, 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

$$\begin{aligned}
\mathrm{Texp}[Y_n](t) &= I_n + \sum_{k=1}^{\infty} J_k(Y_n)(t) \\
&= I_n + \sum_{k=1}^{\infty} \int_0^t \mathrm{d}Y_n(s)J_{k-1}(s) \\
&= I_n + \int_0^t \mathrm{d}Y_n(s)\left(\sum_{k=1}^{\infty} J_{k-1}(Y_n)(s)\right) \\
&= I_n + \int_0^t \mathrm{d}Y_n(s)\,\mathrm{Texp}[Y_n](s), \qquad t \in \mathbb{R}_+.
\end{aligned}$$

This completes the proof. □

Proof of Theorem 7.2.4. 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 $H_p = \prod_{k=1}^{p}\Big(I_n + X_{n,k}^{(m)}\Big)$ 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 $M_j = \frac{\mu_n}{m}\Big(X_{n,j}^{(m)} - \mathbb{E}\Big[X_{n,j}^{(m)}\Big]\Big) + \frac{\sigma_n}{\sqrt{m}} G_{n,j}^{(m)}$ 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 $\Big(G_{n,k}^{(m)}\Big)_{k\in[m]}$ 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

$$\begin{aligned} \|P_{n,m}(t) - P_n(t)\|_{\mathrm{F}}^2 \le 3t\mu_n^2 \int_0^t \Big\|\widetilde{A}_n^{(m)}(s) P_{n,m}(s) - A_n(s) P_n(s)\Big\|_{\mathrm{F}}^2 \,\mathrm{d}s \\ + 3\sigma_n^2 \Big\| \int_0^t \mathrm{d}B_n(s)(P_{n,m}(s) - P_n(s))\Big\|_{\mathrm{F}}^2 + \frac{3\mu_n^2}{m^2}\Big\|\sum_{j=1}^{\lfloor mt \rfloor} Z_{n,j}^{(m)} H_{j-1}\Big\|_{\mathrm{F}}^2, \end{aligned}$$

where $Z_{n,j}^{(m)} = M_{n,j}^{(m)} - \mathbb{E}\Big[M_{n,j}^{(m)}\Big]$ 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

$$\begin{aligned} \Delta_m(t) \le 3t\mu_n^2 \int_0^t \Big\|\widetilde{A}_n^{(m)}(s)\Big\|_{\mathrm{F}}^2 \Delta_m(s)\,\mathrm{d}s + 3t\mu_n^2 \int_0^t \zeta_{n,m}(s)\|P_n(s)\|_{\mathrm{F}}^2\,\mathrm{d}s \\ + 3\sigma_n^2 \sup_{s\in[0,t]} \Big\| \int_0^s \mathrm{d}B_n(r)(P_{n,m}(r) - P_n(r))\Big\|_{\mathrm{F}}^2 \\ + \sup_{s\in[0,t]} \frac{3\mu_n^2}{m^2} \sum_{j,j'=1}^{\lfloor ms \rfloor} \mathrm{Tr}\Big[H_{j-1}^\top Z_{n,j}^{(m)\top} Z_{n,j'}^{(m)} H_{j'-1}\Big], \end{aligned}$$

where $\zeta_{n,m}(t) := \sup_{s\in[0,t]}\left\|\widetilde{A}_n^{(m)}(s) - A_n(s)\right\|_{\mathrm{F}}^2$. Finally, since $\left\|\widetilde{A}_n^{(m)}(s)\right\|_{\mathrm{F}}^2 \le Cn^2$ for all $s \in [0,1]$, for some constant $C > 0$. Since $\left(Z_{n,j}^{(m)}\right)_{j\in[m]}$ are all independent for every $m \in \mathbb{N}$, and $\mathbb{E}\left[Z_{n,j}^{(m)\top} Z_{n,j}^{(m)}\right] \preccurlyeq nDI_n$ for all $j \in [m]$ and every $m \in \mathbb{N}$, taking expectations and using Doob's maximal inequality, we get

$$\begin{aligned}
\mathbb{E}[\Delta_m(t)] &\le 3(Ctn^2\mu_n^2 + 4\sigma_n^2)\int_0^t \mathbb{E}[\Delta_m(s)]\,\mathrm{d}s + 3t\mu_n^2\zeta_{n,m}(t)\int_0^t \mathbb{E}\left[\|P_n(s)\|_{\mathrm{F}}^2\right]\mathrm{d}s \\
&\quad + 24nD\mu_n^2 m^{-1}\int_0^t \mathbb{E}\left[\|P_n(s)\|_{\mathrm{F}}^2\right]\mathrm{d}s + 24nD\mu_n^2 m^{-1}\int_0^t \mathbb{E}[\Delta_m(s)]\,\mathrm{d}s \\
&= 3(Ctn^2\mu_n^2 + 4\sigma_n^2 + 8nD\mu_n^2 m^{-1})\int_0^t \mathbb{E}[\Delta_m(s)]\,\mathrm{d}s \\
&\quad + 3(t\mu_n^2\zeta_{n,m}(t) + 8nD\mu_n^2 m^{-1})\int_0^t \mathbb{E}\left[\|P_n(s)\|_{\mathrm{F}}^2\right]\mathrm{d}s.
\end{aligned}$$

Now we apply Grönwall inequality [100] to get

$$\mathbb{E}[\Delta_m(t)] \le 3(t\mu_n^2\zeta_{n,m}(t) + 8nD\mu_n^2 m^{-1})\int_0^t \mathbb{E}\left[\|P_n(s)\|_{\mathrm{F}}^2\right]\mathrm{d}s \cdot \mathrm{e}^{3t(Ctn^2\mu_n^2 + 4\sigma_n^2 + 8nD\mu_n^2 m^{-1})}.$$

The claim now follows from the assumption that $\zeta_{n,m}(t) \to 0$ as $m \to \infty$. □

7.3.1 Dimension going to infinity

In this section, we prove Theorem 7.2.6. 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

$$\mathrm{Texp}[Y_n] = \mathrm{Texp}\left[\int_0^{\cdot} \mu_n A_n(s)\,\mathrm{d}s\right] + \sigma_n B_n + \sum_{k=2}^{\infty} \widetilde{J}_k,$$

where $\widetilde{J}_k$ is the sum of all k-fold integrals that contain at least one (scaled) BM. Note that

$$n(\mathrm{Texp}[Y_n] - I_n) = n\left(\mathrm{Texp}\left[\int_0^{\cdot} \mu_n A_n(s)\,\mathrm{d}s\right] - I_n\right) + B_n + n\sum_{k=2}^{\infty} \widetilde{J_k}. \tag{7.8}$$

On the other hand,

$$nK\left(\mathrm{Texp}\left[\int_0^{t} \mu_n A_n(s)\,\mathrm{d}s\right] - I_n\right) = (\mathrm{e}^t - 1)\mathbb{E}\left[\prod_{s\in N_t} K(A_n(s)) \,\middle|\, N_t \geq 1\right].$$

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 $\Gamma(U)(t) := (\mathrm{e}^t - 1)\mathbb{E}\big[\prod_{s\in N_t} W(s) \,\big|\, N_t \geq 1\big]$. A randomly chosen $r \times r$ submatrix of $\left(\mathrm{Texp}\left[\int_0^t \mu_n A_n(s)\,\mathrm{d}s\right] - I_n\right)$ therefore converges to $\Gamma(U)(t)\{r\}$ for i.i.d. Uni$([0,1])$ random variables $\{U_i\}_{i\in\mathbb{N}}$.

It is reasonable to believe that, as $n \to \infty$, the sum $n\big(\mathrm{Texp}\big[\int_0^{\cdot} \mu_n A_n(s)\,\mathrm{d}s\big] - I_n\big) + B_n$ converges to the appropriate IEA $X(t) = \Gamma(U)(t)\{\infty\} + B(t)$ where B is an IEA with all BMs. 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

$$\mathrm{Texp}[Y_n] = \mathrm{Texp}\left[\int_0^{\cdot} \mu_n A_n(s)\,\mathrm{d}s\right] + \sum_{k=1}^{\infty} \widetilde{J}_k.$$

Now notice that the same heuristic as above shows that $\sqrt{n}\big(\mathrm{Texp}\big[\int_0^{\cdot} \mu_n A_n(s)\,\mathrm{d}s\big] - I_n\big) = O\Big(\frac{1}{\sqrt{n}}\Big)$. On the other hand, rewrite $\sum_{k=1}^{\infty} \widetilde{J}_k = (\mathrm{Texp}[\sigma_n B_n] - I_n) + \sum_{k=2}^{\infty} \widehat{J}_k$, where $\widehat{J}_k$ is the sum of all k-fold integrals which contain at least one BM but not all are BMs. Following [184, page 151], we now notice that $\sqrt{n}(\mathrm{Texp}[\sigma_n B_n](t) - I_n)$ 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 BM.

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 $\sqrt{n}(\mathrm{Texp}[\sigma_n B_n] - I_n)$. 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

$$\begin{aligned} n(\mathcal{E}_n - (\mathrm{Texp}[\sigma_n B_n] - I_n)) &= n(\mathrm{Texp}[Y_n] - \mathrm{Texp}[\sigma_n B_n]) \\ &= n\left(\mathrm{Texp}\left[\int_0^{\cdot} \mu_n A_n(s)\,\mathrm{d}s\right] - I_n\right) + \sum_{k=2}^{\infty} n\widehat{J}_k. \end{aligned}$$

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 $n(\mathrm{Texp}[Y_n] - \mathrm{Texp}[\sigma_n B_n])$ 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 $p \leq k$, define

$$I_{k,p,\pi}(t) := \sum_{\alpha \in \binom{k}{p}} \int_{\Delta_k(t)} \mathrm{d}U_{\alpha,\pi}(\mathbf{s}), \qquad t \in [0,1].$$

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 de-

fine $I_{k,p}(t)$ 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 7.3.1. *For $k_1, k_2 \in \mathbb{N}$, $p \leq k_1 \wedge k_1$, $\pi_1 \in [n]^{k_1+1}$, $\pi_2 \in [n]^{k_2+1}$, $t \in \mathbb{R}_+$, 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*

$$\left| \int_{\Delta_{k_1}(t)} \int_{\Delta_{k_2}(t)} \mathbb{E}[\mathrm{d}U_{\alpha,\pi_1}(\mathbf{s})\, \mathrm{d}U_{\beta,\pi_2}(\boldsymbol{\tau})] \right| \leq C^{k_1-p} C^{k_2-p} \cdot |\Delta(p; k_1, k_2, \alpha, \beta; t)|, \tag{7.9}$$

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. 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 $C \geq 0$, we get

$$\left| \int_{\Delta_{k_1}(t)} \int_{\Delta_{k_2}(t)} \mathbb{E}[\mathrm{d}U_{\alpha,\pi_1}(\mathbf{s})\, \mathrm{d}U_{\beta,\pi_2}(\boldsymbol{\tau})] \right| \leq C^{k_1-p} C^{k_2-p} \cdot |\Delta(p; k_1, k_2, \alpha, \beta; t)|.$$

□

Claim 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,

$$|\Delta(p; k_1, k_2, \alpha, \beta; t)| \leq \frac{t^p}{p!} \frac{t^{k_1-p}}{\delta_\alpha(1)! \dots \delta_\alpha(p+1)!} \frac{t^{k_2-p}}{\delta_\beta(1)! \dots \delta_\beta(p+1)!}.$$

Proof of Claim 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 $b \geq v_n \geq v_{n-1} \geq v_1 \geq a$. Given a vector $u = (u_1, \ldots, u_p)$ 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

$$\begin{aligned} |\Delta(p; k_1, k_2, \alpha, \beta, t)| &= \mathbb{P}\{E_1(U) \cap E_2(U)\} \\ &\leq \frac{t^{k_1-p}}{\delta_{\alpha(1)}! \ldots \delta_{\alpha(p+1)}!} \frac{t^{k_2-p}}{\delta_{\beta(1)}! \ldots \delta_{\beta(p+1)}!} \int_{\Delta_p(t)} \mathrm{d}u_1 \ldots \mathrm{d}u_p. \end{aligned}$$ □

We use the Lemma 7.3.1 to compute the variances of the error terms in Case 1 to 3 above.

Lemma 7.3.2. *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

$$\sum_{k=2}^{\infty} \widetilde{J}_k(t) = \sum_{p=1}^{\infty} H_{n,p}(t), \qquad H_{n,p}(t) \coloneqq \sum_{k=p\vee 2}^{\infty} \mu_n^{k-p} \sigma_n^p I_{k,p,(x,y)}(t). \tag{7.10}$$

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,

$$\begin{aligned}
\operatorname{Var}\left[n\sum_{k=2}^{\infty}\widetilde{J}_{k,(x,y)}(t)\right] &\leq \operatorname{Var}\left[n\sum_{p=1}^{\infty}H_{n,p}(t)\right] \\
&= n^2\sum_{p=1}^{\infty}\operatorname{Var}[H_{n,p}(t)] \\
&= n^2\sum_{p=1}^{\infty}\mathbb{E}\left[\left(\sum_{k=p}^{\infty}\mu_n^{k-p}\sigma_n^p I_{k,p,(x,y)}(t)\right)^2\right] \\
&= n^2\sum_{p=1}^{\infty}\mathbb{E}\left[\sum_{k_1,k_2=p}^{\infty}\mu_n^{k_1-p}\mu_n^{k_2-p}\sigma_n^{2p}I_{k_1,p,(x,y)}(t)I_{k_2,p,(x,y)}(t)\right] \\
&= n^2\sum_{p=1}^{\infty}\sum_{k_1,k_2=p}^{\infty}\mu_n^{k_1-p}\mu_n^{k_2-p}\sigma_n^{2p}\sum_{\pi_1}\sum_{\pi_2}\mathbb{E}[I_{k_1,p,\pi_1}(t)I_{k_2,p,\pi_2}(t)]
\end{aligned}$$

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 $\pi_1 = (z_{k_1}, \ldots, z_0)$, and $\pi_2 = (\tilde{z}_{k_2}, \ldots, \tilde{z}_0)$. 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 7.3.1 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

$$\begin{aligned}
&\mathrm{Var}\left[n\sum_{k=2}^{\infty}\widetilde{J}_{k,(x,y)}(t)\right]\\
&\leq n^2\sum_{p=1}^{\infty}\sum_{k_1,k_2=p}^{\infty}\mu_n^{k_1-p}\mu_n^{k_2-p}\sigma_n^{2p}n^{k_1-1}n^{k_2-1}n^{-p}t^{k_1+k_2-p}\frac{1}{p!}\frac{(C(p+1))^{k_1-p}}{(k_1-p)!}\frac{(C(p+1))^{k_2-p}}{(k_2-p)!}\\
&= n^2\sum_{p=1}^{\infty}\frac{1}{p!}(n\sigma_n)^{2p}n^{-(p+2)}t^p\mathrm{e}^{2Ct(p+1)}\\
&\leq \mathrm{e}^{2Ct}\sum_{p=1}^{\infty}\frac{1}{p!}\big(n\sigma_n^2t\mathrm{e}^{2Ct}\big)^p\\
&= \mathrm{e}^{2Ct}\big(\exp\big(n\sigma_n^2t\mathrm{e}^{2Ct}\big)-1\big)=O\bigg(\frac{1}{n}\bigg).
\end{aligned}$$

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 7.3.3. *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*

$$\begin{aligned}
&C_n(((i_1,j_1),t_1),((i_2,j_2),t_2))\\
&\quad\coloneqq\mathbb{1}\{i_1=i_2\}\int_0^{\min\{t_1,t_2\}}\frac{1}{n}\Big(\Gamma_{n,1}(s)^\top\Gamma_{n,1}(s)\Big)(j_1,j_2)\,\mathrm{d}s\\
&\qquad+\mathbb{1}\{j_1=j_2\}\int_0^{t_1}\int_0^{t_2}\min\{s_1,s_2\}\frac{1}{n}\Big(\Gamma_{n,2}(s_1;t_1)\Gamma_{n,2}(s_2;t_2)^\top\Big)(i_1,i_2)\,\mathrm{d}s_2\,\mathrm{d}s_1,
\end{aligned}\tag{7.11}$$

where

$$\begin{aligned}\Gamma_{n,1}(s) &\coloneqq n\bigg(\operatorname{Texp}\bigg[\int_0^{\cdot}\mu_n A_n(r)\,\mathrm{d}r\bigg](s) - I_n\bigg),\\ \Gamma_{n,2}(s;t) &\coloneqq n\operatorname{Texp}\bigg[\int_0^{\cdot}\mu_n\tau_s(A_n)(r)\,\mathrm{d}r\bigg](t-s)\mu_n A_n(s),\end{aligned} \tag{7.12}$$

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. 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

$$\int_0^t\sigma_n\,\mathrm{d}B_n(s_k)\cdot n\bigg(\operatorname{Texp}\bigg[\int_0^{\cdot}\mu_n A_n(r)\,\mathrm{d}r\bigg](s_k) - I_n\bigg).$$

The sum of the two kinds of term finally is

$$\sigma_n\int_0^t\mathrm{d}B_n(s)\Gamma_{n,1}(s) + \sigma_n\int_0^t\Gamma_{n,2}(s)B_n(s)\,\mathrm{d}s. \tag{7.13}$$

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

$$
\begin{aligned}
&C_n(((i_1, j_1), t_1), ((i_2, j_2), t_2)) \\
&= \mathbb{E}\left[\int_0^{t_1}\int_0^{t_2} \frac{1}{n}\sum_{k_1,k_2=1}^{n} \mathrm{d}B_{n,(i_1,k_1)}(s_1)\,\mathrm{d}B_{n,(i_2,k_2)}(s_2)\Gamma_{n,1,(k_1,j_1)}(s_1)\Gamma_{n,1,(k_2,j_2)}(s_2)\right] \\
&+ \mathbb{E}\left[\int_0^{t_1}\int_0^{t_2} \frac{1}{n}\sum_{k_1,k_2=1}^{n} B_{n,(k_1,j_1)}(s_1)B_{n,(k_2,j_2)}(s_2)\Gamma_{n,2,(i_1,k_1)}(s_1;t_1)\Gamma_{n,2,(i_2,k_2)}(s_2;t_2)\,\mathrm{d}s_2\,\mathrm{d}s_1\right] \\
&= \mathbb{1}\{i_1 = i_2\}\int_0^{\min\{t_1,t_2\}} \left(\frac{1}{n}\sum_{k=1}^{n}\Gamma_{n,1,(k,j_1)}(s)\Gamma_{n,1,(k,j_2)}(s)\right)\mathrm{d}s \\
&\quad + \mathbb{1}\{j_1 = j_2\}\int_0^{t_1}\int_0^{t_2} \min\{s_1, s_2\}\left(\frac{1}{n}\sum_{k=1}^{n}\Gamma_{n,2,(i_1,k)}(s_1;t_1)\Gamma_{n,2,(i_2,k)}(s_2;t_2)\right)\mathrm{d}s_2\,\mathrm{d}s_1 \\
&= \mathbb{1}\{i_1 = i_2\}\int_0^{\min\{t_1,t_2\}} \frac{1}{n}\Big(\Gamma_{n,1}(s)^\top\Gamma_{n,1}(s)\Big)(j_1, j_2)\,\mathrm{d}s \\
&\quad + \mathbb{1}\{j_1 = j_2\}\int_0^{t_1}\int_0^{t_2} \min\{s_1, s_2\}\frac{1}{n}\Big(\Gamma_{n,2}(s_1;t_1)\Gamma_{n,2}(s_2;t_2)^\top\Big)(i_1, i_2)\,\mathrm{d}s_2\,\mathrm{d}s_1.
\end{aligned}
$$

This completes the proof. □

We are now ready to prove Theorem 7.2.6. 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

$$
\begin{aligned}
&C_\infty(((x_1,y_1),t_1),((x_2,y_2),t_2)) \\
&\quad \coloneqq \mathbb{1}\{x_1 = x_2\} \int_0^{\min\{t_1,t_2\}} \Big(\Gamma_1(s)^\top \odot \Gamma_1(s)\Big)(y_1,y_2)\,\mathrm{d}s \\
&\qquad + \mathbb{1}\{y_1 = y_2\} \int_0^{t_1}\int_0^{t_2} \min\{s_1,s_2\}\Big(\Gamma_2(s_1;t_1) \odot \Gamma_2(s_2;t_2)^\top\Big)(x_1,x_2)\,\mathrm{d}s_2\,\mathrm{d}s_1.
\end{aligned} \tag{7.14}
$$

Lemma 7.3.4. *Let W_1 and W_2 be two curves of kernels and let $U_i = \int_0^t W_i(s)\,\mathrm{d}s$. Assume that $\sup_{s\in[0,t]}\|W_i(s)\|_\infty \le C_t$ for some $C_t > 0$ and for $i = 1,2$. Define,*

$$
\eta(t) = \sup_{s\in[0,t]} \|W_1(s) - W_2(s)\|_2 .
$$

Then, for every fixed $0 \le s \le t \in \mathbb{R}_+$ we have

$$
\begin{aligned}
\|\Gamma_1(U_1)(t) - \Gamma_1(U_2)(t)\|_2 &\le tC_t \mathrm{e}^{tC_t}\eta(t), \\
\|\Gamma_2(U_1)(s;t) - \Gamma_2(U_2)(s;t)\|_2 &\le \big(tC_t(\mathrm{e}^{tC_t}-1) + \mathrm{e}^{tC_t}\big)\eta(t).
\end{aligned}
$$

Proof. The proof for the continuity of Γ_1 follows exactly the same argument as in Lemma 7.3.6, 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

$$
\begin{aligned}
&\|\Gamma_2(U_1)(s;t) - \Gamma_2(U_2)(s;t)\|_2 \\
&\le \|(\mathrm{Texp}[U_1](t-s) - \mathrm{Texp}[U_2](t-s)) \odot W_1(s)\|_2 + \|\mathrm{Texp}[U_2](t-s) \odot (W_1(s) - W_2(s))\|_2 \\
&\le \|\mathrm{Texp}[U_1](t-s) - \mathrm{Texp}[U_2](t-s)\|_{\mathrm{op}}\|W_1(s)\|_2 + \|\mathrm{Texp}[U_2](t-s)\|_{\mathrm{op}}\|W_1(s) - W_2(s)\|_2 \\
&\le tC_t(\mathrm{e}^{tC_t}-1)\eta(t) + \mathrm{e}^{tC_t}\eta(t) = \big(tC_t(\mathrm{e}^{tC_t}-1) + \mathrm{e}^{tC_t}\big)\eta(t),
\end{aligned}
$$

where the last line uses Lemma 7.3.6. □

The proof of Theorem 7.2.6 in Case 1 and Case 2 now follows easily.

Proof of Theorem 7.2.6 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. $\mathrm{Uni}([0,1])$ 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 7.3.2 we obtain

$$\mathbb{W}_2^2((n\mathcal{E}_n)\{r\}(t), X[r](t)) \leq 2e_{n,r}(t) + 2\mathrm{Var}\left[n\sum_{k=2}^{\infty} \widetilde{J}_k(t)\right] \leq 2e_{n,r}(t) + 2\left(\frac{r^2}{n}\right),$$

where

$$e_{n,r}(t) \coloneqq \|\Gamma_n(t)\{r\} - \Gamma(U)(t)\{r\}\|_{\mathrm{F}}^2.$$

Now notice that

$$\mathbb{E}\left[\sup_{s\in[0,t]} e_{n,r}(s)\right] \leq 2r^2 \sup_{s\in[0,t]} \|m_n(s) - m(s)\|_2^2.$$

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 $\mathrm{Var}\left[n^{1/2}\sum_{k=2}^{\infty} \widehat{J}_{k,(i,j)}(t)\right] \leq \frac{C_t}{n}$. By our assumption, we have that $\left\|\sqrt{n}\left(\mathrm{Texp}\left[\int_0^{\cdot} \mu_n A_n(s)\,\mathrm{d}s\right](t) - I_n\right)\right\|_{\max} \leq \frac{C_t}{\sqrt{n}}$. Now observe that $\sqrt{n}\Gamma\left(\frac{B_n}{\sqrt{n}}\right)$ 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 $\sqrt{n}\Gamma\left(\frac{B_n}{\sqrt{n}}\right)(t) = B_\infty[n](\mathrm{e}^t - 1)$. With this coupling, it is immediate that

$$\mathbb{W}_2^2\left((\sqrt{n}\mathcal{E}_n)\{r\}(t), B_\infty[r](t)\right) \leq 8\left(\frac{r^2 C_t}{n}\right),$$

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 $\left\|\sqrt{n}\big(\mathrm{Texp}\big[\int_0^\cdot \mu_n A_n(s)\,\mathrm{d}s\big](t) - I_n\big)\right\|_{\max} \leq \frac{C_t}{\sqrt{n}}$ 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.

Proof of Theorem 7.2.6 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

$$\mathcal{E}_{n,\mathrm{det}} = n\left(\mathrm{Texp}\left[\int_0^\cdot \mu_n A_n(s)\,\mathrm{d}s\right] - I_n\right)$$
$$\mathcal{E}_{n,0} = \sum_{k=2}^{\infty} n\widehat{J}_{k,0}, \qquad \mathcal{E}_{n,1} = \sum_{k=2}^{\infty} n\widehat{J}_{k,1}.$$

The proof strategy is similar to the first two cases with some technical differences that we explain first. From Lemma 7.3.3, we have that entries of $\mathcal{E}_{n,1}$ have variance $O(1/n)$. Now, notice that $\mathcal{E}_{n,\mathrm{det}} = n\Gamma\big(\int_0^\cdot \mu_n A_n(s)\,\mathrm{d}s)\big)$. We know from our assumption that $K(\mathcal{E}_{n,\mathrm{det}})$ converges in L^2 to $\Gamma(t)$. In particular, a randomly chosen $r \times r$ submatrix $\mathcal{E}_{n,\mathrm{det}}\{r\}$ of $\mathcal{E}_{n,\mathrm{det}}$ 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 7.3.3 that the covariance kernel of $\mathcal{E}_{n,0}$ is given by C_n.

Roughly, the idea of the proof is as follows. Let $\widetilde{\mathcal{E}}_n = \mathcal{E}_{n,\mathrm{det}} + \mathcal{E}_{n,0}$. Ignoring the $O(1/n)$ term $\mathcal{E}_{n,1}$, we notice that

$$\widetilde{\mathcal{E}}_n\{r\} = \mathcal{E}_{n,\mathrm{det}}\{r\} + G_{n,r},$$

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 $\mathcal{E}_{n,\mathrm{det}}\{r\}$ 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 $G_\infty := (G_{i,j})_{(i,j)\in\mathbb{N}^2}$ 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 $\mathcal{E}_{n,\mathrm{det}}(t)\{r\} \approx \Gamma(t)\{r\}$. 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

$$C(((x_1,y_1),t_1),((x_2,y_2),t_2)) := \delta_{x_1=y_1,x_2=y_2} \min\{t_1,t_2\},$$

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 C. Let $G(x,y;\,\cdot\,)$ be a process defined as

$$G(x,y;t) = \int_0^t \int_0^1 \mathrm{d}B(x,z;s)\Gamma_1(s)(z,y)\,\mathrm{d}z\,\mathrm{d}s + \int_0^t \int_0^1 \Gamma_2(s;t)(x,z)B(z,y;s)\,\mathrm{d}z\,\mathrm{d}s, \tag{7.15}$$

for every $(x,y;t) \in \mathcal{I}$. Notice that G has a covariance kernel given by C_∞ defined in Lemma 7.3.3.

Possibly after extending the probability space $(\Omega,\mathcal{F},\mathbb{P})$ we assume that it supports a collection of i.i.d. $\mathrm{Uni}([0,1])$ random variables $\{U_i\}_{i\in\mathbb{N}}$ independent of B. Define an IEA Y by setting

$$Y_{i,j} := \Gamma(t)(U_i,U_j) + G(U_i,U_j;t), \tag{7.16}$$

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

$$G_n(x,y;t) := \int_0^t \int_0^1 \mathrm{d}B(x,z;s)K(\Gamma_{n,1}(s))(z,y)\,\mathrm{d}z \tag{7.17}$$

$$+ \int_0^t \int_0^1 K(\Gamma_{n,2}(s;t))(x,z)B(z,y;s)\,\mathrm{d}z\,\mathrm{d}s. \tag{7.18}$$

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 7.3.4, 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.

Lemma 7.3.5. *Let C_n and C_∞ be as defined in Lemma 7.3.3 and equation (7.14) 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*

$H_{n,0} = h(U_i)$ for every $i \in [n]$ and $n \in \mathbb{N}$. Let $Z_n(t) = (G_n(t)(U_i, U_j))_{(i,j)\in[n]^2}$ for every $n \in \mathbb{N}$ where G_n is as defined in equation (7.18). Then for any $i_1, i_2 \in [n]$,

$$\lim_{n\to\infty} \mathbb{E}\left[\left(\frac{1}{n} Z_n(t) H_{n,0}\right)_{i_1} \left(\frac{1}{n} Z_n(t) H_{n,0}\right)_{i_2}\right] = \mathbb{1}\{i_1 = i_2\} \int_0^t \|\Gamma(s)h\|_2^2 \,\mathrm{d}s. \tag{7.19}$$

Proof. Following the definition of C_n,

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{n} Z_n(t) H_{n,0}\right)_{i_1} \left(\frac{1}{n} Z_n(t) H_{n,0}\right)_{i_2}\right] \\
&= \frac{1}{n^2} \sum_{j_1,j_2=1}^{n} H_{n,j_1} \mathbb{E}\big[Z_{n,(i_1,j_1)}(t) Z_{n,(i_2,j_2)}(t)\big] H_{n,j_2} \\
&= \frac{1}{n^2} \sum_{j_1,j_2=1}^{n} H_{n,j_1} C_n(((U_{i_1}, U_{j_1}), t), ((U_{i_1}, U_{j_1}), t)) H_{n,j_2}
\end{aligned}$$

Separating the terms when $j_1 = j_2$ and otherwise, the above simplifies to

$$\begin{aligned}
&\frac{1}{n^2} \sum_{j_1=j_2=1}^{n} H_{n,j_1} C_n(((U_{i_1}, U_{j_1}), t), ((U_{i_1}, U_{j_1}), t)) H_{n,j_2} \\
&\qquad + \frac{1}{n^2} \sum_{j_1\neq j_2=1}^{n} H_{n,j_1} C_n(((U_{i_1}, U_{j_1}), t), ((U_{i_1}, U_{j_1}), t)) H_{n,j_2}
\end{aligned}$$

Let us first consider the case when $i_1 \neq i_2$. Then 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, but the second term survives. Plugging in the expression for C_n in this case, we get that the above converges to

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n^2} \sum_{j_1,j_2=1}^{n} h(U_{j_1}) h(U_{j_2}) \int_0^t \Big(\Gamma_1(s)^\top \odot \Gamma_2(s)\Big)(U_{j_1}, U_{j_2}) \,\mathrm{d}s \\
&= \lim_{n\to\infty} \int_0^t \int_0^1 \frac{1}{n^2} \sum_{j_1,j_2=1}^{n} h(U_{j_1}) h(U_{j_2}) \Gamma_1(s)(z, U_{j_1}) \Gamma_1(s)(z, U_{j_2}) \,\mathrm{d}z \,\mathrm{d}s \\
&= \int_0^t \int_0^1 \left(\lim_{n\to\infty} \frac{1}{n} \sum_{j_1=1}^{n} \Gamma_1(s)(z, U_{j_1}) h(U_{j_1})\right) \left(\lim_{n\to\infty} \frac{1}{n} \sum_{j_2=1}^{n} \Gamma_1(s)(z, U_{j_2}) h(U_{j_2})\right) \\
&= \int_0^t \int_0^1 (\Gamma(s)h)^2(z) \,\mathrm{d}z \,\mathrm{d}s = \int_0^t \|\Gamma(s)h\|_2^2 \,\mathrm{d}s.
\end{aligned}$$

□

7.3.2 Proof of Theorem 7.2.7

We will prove Theorem 7.2.7 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 $f_{n,i} = n\int_{(i-1)/n}^{i/n} f(x)\,\mathrm{d}x$ for $i \in [n]$. Note that $\frac{1}{n}\|f_n\|_2^2 \le \|f\|_2^2$. Let X_n be an $n \times n$ matrix. Notice that

$$\|K(X_n)f\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{j=1}^{n} X_{n,(i,j)} f_{n,j}\right)^2 \le \|f\|_2^2\left(\frac{1}{n^2}\sum_{i,j=1}^{n} X_{n,(i,j)}^2\right). \tag{7.20}$$

In particular, if $X_{n,(i,j)}$s are mean 0 random variables with variance bounded by ς_n^2, then

$$\mathbb{E}\Big[\|K(X_n)f\|_2^2\Big] \le \varsigma_n^2\|f\|_2^2.$$

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 0 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 0 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.

Before we start the proof we prove the following lemma. 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 7.3.6. *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

$$\begin{aligned}
\|J_k(S_1)(t) - J_k(S_2)(t)\|_{\mathrm{op}} &\le \int_{\Delta_k(t)} \|T_1(s_k)\dots T_1(s_1) - T_2(s_k)\dots T_2(s_1)\|_{\mathrm{op}}\,\mathrm{d}s_k\dots\mathrm{d}s_1 \\
&\le \int_{\Delta_k(t)} \sum_{j=1}^{k} \left\| \prod_{i=j+1}^{k} T_1(s_i)\cdot(T_1(s_j) - T_2(s_j))\cdot \prod_{i=1}^{j-1} T_2(s_i) \right\|_{\mathrm{op}} \prod_{i=1}^{k} \mathrm{d}s_i \\
&\le k C_t^{k-1} \eta(t) \int_{\Delta_k(t)} \mathrm{d}s_k \cdots \mathrm{d}s_1 \\
&\le \eta(t) C_t^{k-1} k \frac{t^k}{k!}.
\end{aligned}$$

Finally observe that

$$\|\mathrm{Texp}[S_1](s) - \mathrm{Texp}[S_2](s)\|_{\mathrm{op}} \le \sum_{k\ge1} \|J_k(S_1)(s) - J_k(S_2)(s)\|_{\mathrm{op}} \le \eta(s) t(\mathrm{e}^{tC_t} - 1).$$

Taking supremum over $s \in [0,t]$ we get the final result. □

Proof of Theorem 7.2.7. 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 (7.8), 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

$$\begin{aligned}
&\mathbb{E}\left[\sup_{s\in[0,t]} \|(K(n\,\mathrm{Texp}[Y_n](s)) - \mathcal{E}_{n,\mathrm{det}}(s))f\|_2^2\right] \\
&\le 2\mathbb{E}\left[\sup_{s\in[0,t]} \left(\|K(\mathcal{E}^0_{n,\mathrm{err}}(s))f\|_2^2 + \|K(\mathcal{E}^1_{n,\mathrm{err}}(s))f\|_2^2\right)\right] \le \frac{D_t}{n}\|f\|_2^2,
\end{aligned}$$

where $D_t \geq 1$ is a constant that depends only on t.

Let T_n be the integral operator corresponding to A_n and set $\eta_n(t) := \sup_{s\in[0,t]} \|T_n(s) - T(s)\|_{\mathrm{op}}$. Similarly define the integral curve S_n of T_n. It follows that

$$\sup_{s\in[0,t]} \|\mathrm{Texp}[S_n](s) - \mathrm{Texp}[S](s)\|_{\mathrm{op}} \leq \eta_n(t) t \mathrm{e}^{tC_t} \to 0,$$

as $n \to \infty$.

The proof in the Case 2, follows exactly from the same argument, by noting that $\|\mathcal{E}_{n,\mathrm{det}}\|_\infty \leq \frac{1}{\sqrt{n}}$ in this case. We skip the details. □

7.4 Discussion

7.4.1 Oja's Algorithm

In this section, we analyze the Oja's algorithm [168] 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 [147, 110].

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 $P_{n,m}(t) := \prod_{k=1}^{\lfloor mt \rfloor} \left(I_n + \frac{1}{mn} \cdot n x_k x_k^\top\right)$. By assumption, we have $\mathbb{E}\left[x_{n,k} x_{n,k}^\top\right] = \Sigma_n$ 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 7.2.4, we find that $(P_{n,m})_{m\in\mathbb{N}}$ uniformly converges to $t \mapsto \mathrm{Texp}\left[\int_0^\cdot \Sigma_n\right](t) = \mathrm{e}^{t\Sigma_n}$.

Since 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 eigendecomposition $V_n\Lambda_n V_n^\top$ for $\Lambda_n = \operatorname{diag}\big((\lambda_{n,i})_{i=1}^n\big)$ 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

$$\|q_n(t) - v_{n,1}\|_2^2 = \|z_n(t) - e_{n,1}\|_2^2, \tag{7.21}$$

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

$$\begin{aligned} z_{n,1}^2(t) &= \frac{z_{n,1}^2(0)}{z_{n,1}^2(0) + \sum_{j=2}^n \mathrm{e}^{-2t(\lambda_{n,1}-\lambda_{n,j})} z_{n,j}^2(0)}, \\ z_{n,i}^2(t) &= \frac{\mathrm{e}^{-t(\lambda_{n,1}-\lambda_{n,i})} z_{n,i}^2(0)}{z_{n,1}^2(0) + \sum_{j=2}^n \mathrm{e}^{-2t(\lambda_{n,1}-\lambda_{n,j})} z_{n,j}^2(0)}, \qquad i \in [n] \setminus \{1\}. \end{aligned} \tag{7.22}$$

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

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

7.4.2 Gossip Algorithms

Gossip algorithms are distributed algorithms that are used to average values over the nodes of a graph. Simple applications arise when we have certain sensors capturing values over a small region or space. And, in order 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 7.2.4,

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

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

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

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 $t(1-\lambda_{n,2})$. Therefore we get that the relative error in the limit 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 [86], yielding that as $n \to \infty$, the rate of convergence is $\mathrm{e}^{-T(1-\Theta(d^{-1/2}))}$ with probability 1.

7.4.3 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 [61] 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 [164], 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 $t \in [0,1]$,

$$x_{\lfloor mt\rfloor} = \prod_{k=1}^{\lfloor mt\rfloor}\left(I_n - \frac{na_ka_k^\top}{\lambda_{k,n}mn}\right)\cdot x_1 + \frac{n}{m}\sum_{k=1}^{\lfloor mt\rfloor}\prod_{j=k+1}^{\lfloor mt\rfloor}\left(I_n - \frac{na_ja_j^\top}{\lambda_{j,n}mn}\right)\cdot\frac{a_k}{\lambda_{k,n}}(n^{-1}a_k^\top x^* + \epsilon_k). \tag{7.23}$$

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 7.2.4, 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

$$\mathbb{E}[x(t)] - x^* = \mathrm{e}^{-t\Sigma_n/\lambda_n}(x(0) - x^*). \tag{7.24}$$

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 (7.24) 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.

7.4.4 Infinitely deep and infinitely wide Residual Neural Network

Authors in [216] introduce the maximal update parameterization (μP) to initialize a deep and wide neural network that enjoys various scale-free properties that allow hyperparameter transfers across networks of different sizes. Consider a deep residual neural network defined

as follows. Let $x \in \mathbb{R}^d$ be a fixed input, where $d \in \mathbb{N}$ is fixed. A m-layer network $\hat{y}$ under μP scaling takes the following form. For every $i \in [n]$, compute

$$\begin{aligned} H_{n,0} &= Jx, \\ H_{n,k} &= H_{n,k-1} + \frac{1}{\sqrt{nm}} \theta_{n,k}^{(m)} H_{n,k-1}, \qquad k \in [m], \\ \hat{y}(x) &= \frac{1}{n} \sum_{i=1}^{n} H_{n,m,i}. \end{aligned}$$

The matrix $J \in \mathbb{R}^{n \times d}$ is a fixed sampling matrix with (possibly random) entries of the order $\Theta(1)$. An example of such a matrix is when all elements are i.i.d. and distributed as $N(0,1)$. The trainable matrices $\left(\theta_{n,k}^{(m)}\right)_{k \in [m]}$ for every $m \in \mathbb{N}$ are the weight matrices corresponding to every layer. At the time of initialization, these matrices are set to be independent with 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 $\left(\theta_{n,k}^{(m)}\right)_{k \in [m]}$ change to have non-zero mean that allows $\hat{y}$ to change from zero and learns the desired function. A reasonable way through which these weight matrices $\left(\theta_{n,k}^{(m)}\right)_{k \in [m]}$ can be modeled is

$$\theta_{n,k}^{(m)} = \frac{1}{\sqrt{nm}} M_{n,k}^{(m)} + G_{n,k}^{(m)}, \qquad k \in [m],\ m \in \mathbb{N}, \tag{7.25}$$

where $\left(M_{n,k}^{(m)}\right)_{k \in [m]}$ are random matrices with mean $\left(A_{n,k}^{(m)}\right)_{k \in [m]}$ respectively. Using the above model, the layer $\lfloor mt \rfloor$ of the network for any $t \in [0,1]$ computes

$$H_{n,\lfloor mt \rfloor} = \prod_{k=1}^{\lfloor mt \rfloor} \left(I_n + \frac{1}{nm} M_{n,k}^{(m)} + \frac{1}{\sqrt{nm}} G_{n,k}^{(m)} \right) \cdot H_{n,0}.$$

The iterative product of matrices in the above equation is nothing but $P_{n,m}(t)$ as discussed in Section 5.1 for $\sigma^2 = n^{-1}$. If we now assume that the piecewise constant interpolations $\left(A_n^{(m)} : t \mapsto A_{n,\lfloor mt \rfloor}^{(m)}\right)_{m \in \mathbb{N}}$ uniformly converge to a curve A_n in L^2 as $m \to \infty$, such that $\sup_{t \in [0,1]} \|A_n(s)\|_{\max} \le C$ for some $C \ge 0$, and $\mathrm{Cov}\left(M_{n,k}^{(m)}, \preceq\right) nDI_n$ for all $k \in [m]$, $m \in \mathbb{N}$ for some $D \ge 0$; then Theorem 7.2.4 tells us that $P_{n,m}$ uniformly converges to $\mathrm{Texp}[Y_n]$ as $m \to \infty$, where

$$Y_n(t) = \frac{1}{n} \int_0^t A_n(s)\, \mathrm{d}s + \frac{1}{\sqrt{n}} B_n(t), \qquad t \in [0,1],$$

where B_n is a $n \times n$ matrix with i.i.d. Brownian motions.

7.4.5 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:

$$H_{n,k-1,i} + \frac{1}{m} \cdot \left(\frac{1}{n} \sum_{j=1}^{n} M^{(m)}_{n,k,(i,j)} H_{n,k-1,j} \right) + \frac{1}{\sqrt{m}} \cdot \left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} G^{(m)}_{n,k,(i,j)} H_{n,k-1,j} \right). \quad (7.26)$$

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 [**?**].

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

Chapter 8

SOME REMARKS AND FUTURE DIRECTIONS

We now give a summary of this thesis: we developed a notion of gradient flow on the space of graphons. This is done following the general theory of gradient flows in metric spaces as developed in [5]. Our main contribution, in this regard, is to specialize this theory to the space of graphons. Restricting ourselves to a particular space, we prove some useful results that may have practical significance. For instance, we show that the Euclidean gradient flows on symmetric matrices well-approximate the gradient flows on graphons.

We then turn our attention to studying the scaling limits of the evolution of large graphs that possess some symmetry. In particular, we consider a general class of SDEs (with reflections) on symmetric matrices where the drift is a permutation invariant function of matrices. Under appropriate assumptions, we establish a propagation of chaos phenomenon for these processes. As a result, in the limit, these processes can be described by an infinite exchangeable array whose coordinates satisfy a novel McKean-Vlasov type SDE. This is particularly important for studying the evolution of networks in large dimensions. This can be seen as a generalization of mean-field interacting particle systems. In this thesis, we show that such SDEs naturally arise in the context of stochastic gradient descent with noise for functions of symmetric matrices that satisfy permutation invariance property. It is reasonable to expect that similar phenomenon in many natural models of graph evolution.

In Chapter 6, we consider a variant of the Metropolis chain on the stochastic block models with r communities and n individuals in each community. The chain is designed to minimize certain Hamiltonian that is permutation invariant. We take the number of individuals n in each community to infinity and study the evolution of the $r \times r$ connection probability matrix between communities. With a novel relaxation step in the Metropolis chain, and with an appropriate scaling limit, we show that the $r \times r$ connection probability matrix between communities converges to a matrix-valued diffusion. This diffusion has

permutation invariant drift. Following a similar analysis as in Chapter 5 we show that, as $n \to \infty$, the trajectories of these $r \times r$ matrix-valued SDEs converge, in probability, to a deterministic curve on the space of graphons. In other words, the relaxed Metropolis closely approximates a deterministic curve on the space of graphons. En route, we extended some of the formalism of so-called decorated graphons introduced in [153]. We refer to these decorated graphons as measure-valued graphons. We introduced a metric analogous to the cut metric on the space of measure-valued graphons and established its equivalence to the topology given in [153]. We expect that this will be of independent interest. We further established the equivalence of convergence of measure-valued graphons with the convergence of infinite exchangeable arrays thus extending the results in [117, 118].

Finally, in Chapter 7, we consider the iterated product of matrices where at each step a small perturbation of the identity matrix is multiplied from the left. We study the scaling limit of such matrix products as the dimension goes to infinity. The iterated product of matrices has a rich and long history as we refer in 7. Our contribution in this regard is two-fold. Firstly, in the fixed dimension, we study the product of a triangular array of matrices instead of the product of a sequence of matrices. This generalizes some of the previous works. In fixed dimensions, the scaling limit of the iterated product of matrices is given by a non-commutative exponential of a semimartingale. This description is explicit but may not always be very easy. We provide a neat combinatorial interpretation of the mean of this process. Our second contribution concerns the study of the limit of this non-commutative exponential as the dimension goes to infinity. In this limit, we obtain an infinite exchangeable array where the coordinates are Gaussian and the mean and covariance of these coordinates can be explicitly described. This can also be seen as a matrix-valued generalization of classical symmetric statistics [79].

We now describe some potential applications of the current work and point out some immediate directions that naturally emerge out of the current work that needs further research.

8.1 A computational tool for extremal graph theory

Extremal graph theory is replete with problems that involve maximizing or minimizing certain graph functions. For instance, we mentioned the Mantel-Turán problem in Chapter 1. As already explained, our theory does not yield any information about the structure of the minimizer. However, as we exhibited in the example in Section 6.1.3, running a gradient flow for a function can provide insight into the structure of minimizers. Thus, it may serve as a useful tool for obtaining candidate minimizer(s). For instance, as mentioned in Chapter 3 the minimizers for the rate function of ERGM are only known in the so-called replica symmetry regime where we know that the minimizers are constant graphons. This effectively reduces the problem of finding minimizers to a calculus problem. Outside the replica symmetry regime, even a reasonable conjecture for minimizers is often out of reach. For concreteness, consider the problem of minimizing $\mathcal{F}(W) := \beta_1 t(K_2, W) + \beta_2 t(K_3, W)$ for some (β_1, β_2) that is not in the replica-symmetry regime. One may run the gradient flow of $\mathcal{F}(W)$ for sufficiently long time and the resulting graphon would be a good candidate for a minimizer of $\mathcal{F}(W)$. On the other hand, our technique may also be of use in generating counterexamples in some cases. For instance consider the problem of testing whether a connected finite graph H with $E(H)$ number of edges has Siderenko property (see [195]) or not. This is equivalent to testing if the inequality

$$t(H, W) \geq t(K_2, W)^{|E(H)|} ,$$

is true or not. One can again run a gradient flow $(W_t)_{t\geq 0}$ of $\mathcal{F}(W) := t(H, W) - t(K_2, W)^{E(H)}$ over the space of graphons. If H does not have Siderenko's property, then one would expect that $\mathcal{F}(W_t) < 0$ for sufficiently large t. It is worth emphasizing that this is a heuristic idea. The homomorphism densities are generally non-convex functions (see 4.5.1) and therefore the gradient flow may get stuck at a local minima or at a stationary point. Deveoping a systematic method to handle this issue is an important problem that needs significant future work.

8.2 Constrained optimization on the space of graphons

We now describe a related theme that is not considered fully in this thesis and is the next natural problem to study. As explained in Section 3.4, there is a significant interest in studying the optimization of some convex function, say, the entropy function $\mathcal{E}(W) = \int\int h(W(x,y))\,\mathrm{d}\mathrm{d}y$ where $h(p) = p\log(p) + (1-p)\log(1-p)$ over the subset of graphons where the edge density $t(K_2, W) = \epsilon$ and the triangle density $t(K_3, W) = \tau$ are fixed. One can consider more number of such constraints. While one can consider a relaxation of this problem, that is, consider minimizing the functions like $F(W) := A(t(K_2,W) - \epsilon)^2 + B(t(K_3,W) - \tau)^2 + \mathcal{E}(W)$ with some large positive constants A and B. This is not completely satisfying. The functions like $A(t(K_2,W) - \epsilon)^2 + B(t(K_3,W) - \tau)^2$ are generally only semi-convex. In particular, for large A and B, the function F becomes non-convex (see 5.6.3). In this case, while the gradient flow curve still exists, there need not be a unique minimizer and even the convergence to a stationary point is slow (see 5.6.3). We again point out that the structure of the minimizer or stationary point is not within the scope of current work but it is an active area of research.

Naturally, one may envision developing a Lagrange multiplier theory for optimization with constraints on the space of graphons. This is an interesting and important direction to pursue. However, as we have pointed out (see 4.5.1, 5.6.3), the homomorphism densities are generally non-convex. Even the feasible region of parameters, where the intersection of finitely many constraints is non-empty, has been the subject of intense research and is known only in a few cases (see, for example, [181, 182, 180, 161, 162]).

We should mention that one of the most important problems in this area is Siderenko's conjecture. We explain the problem below. Let F be a bipartite graph. Siderenko's conjecture states that minimizer of $t(F,W)$ over all graphons $W : [0,1]^2 \to [0,1]$ such that $\int\int W\,\mathrm{d}x\,\mathrm{d}y = p$ is given by the constant graphon $W \equiv p$. It is known [149] that $W \equiv p$ is a local minimizer. One can replace F with some other class of graphs. The conjecture is known to be true for the class of graphs F called *norming graphs* [103]. However, a full characterization of norming graphs is unknown.

8.3 Dynamic graphon-based interacting particle systems

As mentioned in Chapter 3, there is a significant interest in understanding the evolution of interacting particle systems where the interaction is determined by an underlying graph. We discussed this in detail in Chapter 3.

Our work concerns the evolution of the graph itself where the edges of the graph symmetrically interact with each other. We show that despite the complicated evolution of such networks, for large n (the number of vertices), the evolution of the graph is close to a deterministic curve of graphons. It is natural to study the evolution of graph-based particle systems where the graph itself is evolving with time according to the dynamics considered in our work. This problem is also important from a practical point of view. Most of the literature on graph-based interacting particle systems is inspired by models of the behavior of agents (e.g. financial markets) in an environment where not everyone interacts with everyone else. It is reasonable to assume in this case that the environment itself evolves with time. There is some work in this direction as we discussed in Chapter 3, however, the evolution of the networks in those works depends only on the position of the agents and not on the strength (edge-weights) of the network itself [19, 99, 21]. Such systems arise in the study of gossip algorithms, epidemiology, SIR Model, and so on. We particularly refer the reader to the introduction of [21] and the references therein for more details. For instance [21] describes as a toy example a system of n children labeled $1, \dots, n$ and their location at times t is denoted by $X_i^{(n)}(t)$ for $i = 1, \dots, n$. While there is an underlying network $\xi^{(n)}$ where the edge $\xi_{i,j}^{(n)}(t)$ denotes the type of friendship between i and j. The authors in [21] consider the case where the dynamics of the location $X_i^{(n)}$ of the child i depends on the joint empirical distribution of all the other children's position and their friendship with i. In their model the evolution of the friendship type $\xi_{i,j}^{(n)}$ between i and j only depends only on $X_i^{(n)}$ and $X_j^{(n)}$. It is natural to imagine a more general model of a graph-based interacting particle system that encompasses the two worlds. That is, one may allow the network evolution to depend on the position of the agents (vertex weights) as well as the strength of the connection (edge weights). For instance, in the previous example of children and the friendship network, it is imaginable that the evolution of friendship also

depends on the current status of the friendship and not just on the status of two individuals (at least in a world we would all like to live).

8.4 Fluctuations of subgraph densities

In Chapter 5 and Chapter 6, we establish the convergence of matrix valued process, say $X_n \in C([0,1], \mathcal{M}_n)$, with certain symmetries to some deterministic curve, say $\Gamma \in C([0,1], \mathcal{W})$, on the space of graphons. This convergence entails that for any fixed finite graph F, the homomorphism density $(t(F, X_n(s)))_{s\in[0,1]}$ converges to $(t(F, \Gamma(s)))_{s\in[0,1]}$. This is analogous to the law of large numbers. It is the natural next step to study the fluctuations of the homomorphism density for these processes.

This should be compared to our discussion in Chapter 3. Recall that the fluctuations of homomorphism densities of the random graph $\mathbb{G}(n, W)$ generated from a graphon W is subject to intensive research. However, the available results in the literature do not directly apply to our case. For instance, fix $t > 0$, we show that the random matrix $X_n(t) \in \mathcal{M}_n$ converges to some kernel $\Gamma(t)$. In this case, however, $X_n(t)$ is not the adjacency matrix of $\mathbb{G}(n, \Gamma(t))$. Therefore, while the convergence of homomorphism density follows from our results, the fluctuations of such processes need to be studied in the future. The theory developed for the generalized U-statistics and incomplete U-statistics have proved to be useful tools to study the fluctuations of subgraph counts [119, 166] and it may be useful in studying this problem.

8.5 Analysis of deep neural networks

We now discuss an important example that was one of our primary motivations for this series of work. A neural network (NN) (see Figure 8.1) consists of $b \in \mathbb{N}$ hidden layers, an initial input $x_0 \in \mathbb{R}^{d_0}$, and a terminal output $\hat{y}(x_0) \in \mathbb{R}$ (say), computed by a sequence of transformations

$$x_0 \mapsto x_1 \in \mathbb{R}^{d_1} \mapsto x_2 \in \mathbb{R}^{d_2} \mapsto \cdots \mapsto x_b \in \mathbb{R}^{d_b} \mapsto \hat{y}.$$

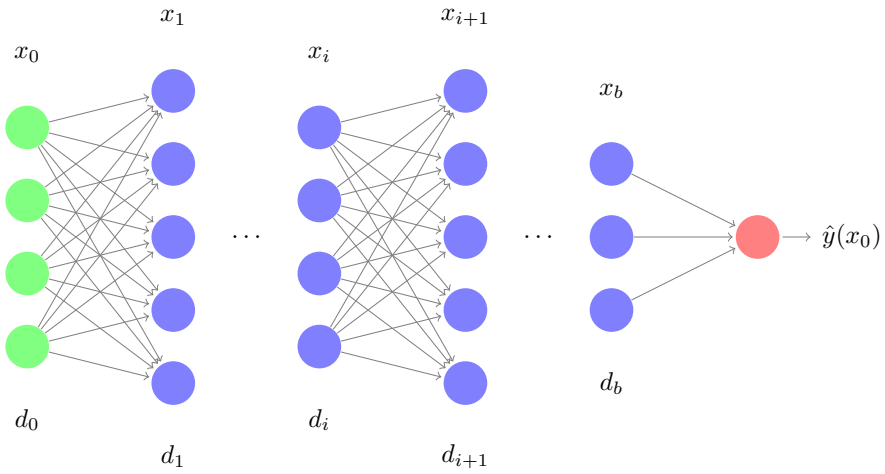


Figure 8.1: Finite width Neural Network with multiple hidden layers

Each transformation involves a matrix $A_{i+1} \in \mathbb{R}^{d_{i+1} \times d_i}$, a vector $\beta_{i+1} \in \mathbb{R}^{d_{i+1}}$, and the transformation is defined as

$$x_{i+1} = \sigma(A_{i+1}x_i + \beta_{i+1}), \tag{8.1}$$

for all $i \in \{0\} \cup [b-1]$, 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 x_b.

The distribution of training data is given by a probability measure μ on $\mathbb{R}^{d_0} \times \mathbb{R}$. The goal of a NN is to minimize a risk function R, often a quadratic loss,

$$R\left(A_i, \beta_i,\ i \in [b-1]\right) \coloneqq \mathbb{E}_{(X,Y)\sim\mu}\Big[(Y - \hat{y}(X))^2\Big], \tag{8.2}$$

where the minimization is over all choices of the sequence of matrices A_i and vectors β_i, for $i \in [b-1]$.

Let us ignore the vectors β_i from our discussion. The entries of the matrix A_{i+1} can be thought of as associated with the edges of the bipartite graph connecting the nodes in layers i and $i+1$. The output R in equation (8.2) 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 A_{i+1} the output R remains the same. Therefore the risk function R can be thought of as a function of edge weights of a sequence of bipartite graphs that is

invariant under vertex relabeling. A gradient descent algorithm on the NN tries to compute the Euclidean gradient flow with respect to the edge weights to reach the minimum value of R. One can again 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? This is a multivariate generalization of our set-up of the gradient flow of a function on graphons. Instead of a single graph, we have a sequence of b graphs, all bipartite, and successive graphs share vertices. Many significant problems are open in this area. For an NN with a single hidden layer, a similar analysis has been successfully performed in [200, 201] where it is shown that the training of NN can be modeled by a gradient flow on Wasserstein space. For deep neural networks, the theory is still unsatisfying despite a large number of works. We believe that a multivariate generalization of our work can address this challenge.

Another important insight and direction that our work (especially Chapter 7) opens up is the following. Let us consider the model where the function σ is the identity $\sigma(x) = x$ and the vectors $\beta_i \equiv 0$. In this case, we see that $x_d = A_d A_{d-1} \cdot A_1 x_0$, that is, the output of the network is obtained via a sequence of products of matrices. In Chapter 7, we study a particular model of iterated product of matrices where the dimensions are all the same. This allows us to interpret the output of a deep NN (in infinite width and infinite depth limit). Given the weight matrices, in this simple setting, we can describe the output. However, it suggests (at least at the initialization) that the evolution of a fixed neuron can be described as a Gaussian process whose covariance kernel can be described by a function of some curve in $L^2([0,1]^2)$. More generally, one can think of the evolution of the full network as a Gaussian process indexed by $[0,1] \times \mathbb{R}_{\geq 0}$ where the index $[0,1]$ essentially models the location of a neuron and the index $\mathbb{R}_{\geq 0}$ represents the time. While this requires a significant amount of future work, this is an important and promising direction. We should again point out that in the current setup, we do not consider any training in the network. It would be an interesting direction to analyze the training dynamics in deep neural networks. One key issue that arises here is that at any positive time during the training, the weight matrices are highly correlated. In recent times, many authors have attempted similar analysis [104, 64, 191, 101, 94, 6, 148, 218, 194, 33, 120].

BIBLIOGRAPHY

[1] Edo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. *Advances in Neural Information Processing Systems*, 26, 2013.

[2] David J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

[3] David J. Aldous. On exchangeability and conditional independence. *Exchangeability in probability and statistics (Rome, 1981)*, pages 165–170, 1982.

[4] 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.

[5] 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, 2008.

[6] Joseph M Antognini. Finite size corrections for neural network gaussian processes. *arXiv preprint arXiv:1908.10030*, 2019.

[7] Dyego Araújo, Roberto I Oliveira, and Daniel Yukimura. A mean-field limit for certain deep neural networks. arXiv preprint arXiv:1906.00193, 2019.

[8] Siva Athreya, Frank den Hollander, and Adrian Röllin. Graphon-valued stochastic processes from population genetics. *The Annals of Applied Probability*, 31(4):1724 – 1745, 2021.

[9] Siva Athreya, Soumik Pal, Raghav Somani, and Raghavendra Tripathi. Path convergence of markov chains on large graphs. *arXiv preprint arXiv:2308.09214*, 2023.

[10] Siva Athreya and Adrian Röllin. Dense graph limits under respondent-driven sampling. *The Annals of Applied Probability*, 26(4):2193–2210, 2016.

[11] Siva Athreya and Adrian Röllin. Respondent-driven sampling and sparse graph convergence. *Electronic Communications in Probability*, 23:1–12, 2018.

[12] Alexander Aurell, René Carmona, and Mathieu Lauriere. Stochastic graphon games: Ii. the linear-quadratic case. *Applied Mathematics & Optimization*, 85(3):1–33, 2022.

[13] Tim Austin. On exchangeable random variables and the statistics of large graphs and hypergraphs. *Probability Surveys*, 5:80–145, 2008.

[14] Tim Austin. Exchangeable random arrays. In *Notes for IAS workshop*, 2012.

[15] Tim Austin. Exchangeable random measures. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 51(3):842 – 861, 2015.

[16] Francis Bach and Lenaïc Chizat. Gradient descent on infinitely wide neural networks: Global convergence and generalization. arXiv preprint arXiv:2110.08084, 2021.

[17] Sina Baghal. A matrix concentration inequality for products. *arXiv preprint arXiv:2008.05104*, 2020.

[18] Andrew D Barbour, Michal Karoński, and Andrzej Ruciński. A central limit theorem for decomposable random variables with applications to random graphs. *Journal of Combinatorial Theory, Series B*, 47(2):125–145, 1989.

[19] Julien Barré, Paul Dobson, Michela Ottobre, and Ewelina Zatorska. Fast non-mean-field networks: Uniform in time averaging. *SIAM Journal on Mathematical Analysis*, 53(1):937–972, 2021.

[20] Erhan Bayraktar, Suman Chakraborty, and Ruoyu Wu. Graphon mean field systems. *arXiv preprint arXiv:2003.13180*, 2020.

[21] Erhan Bayraktar and Ruoyu Wu. Mean field interaction on random graphs with dynamically changing multi-color edges. *Stochastic Processes and their Applications*, 141:197–244, 2021.

[22] Erhan Bayraktar and Ruoyu Wu. Graphon particle system: Uniform-in-time concentration bounds. *Stochastic Processes and their Applications*, 156:196–225, 2023.

[23] Richard Bellman. Limit theorems for non-commutative operations. I. *Duke Journal of Mathematics*, 1954.

[24] Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pages 1–68. Springer, 1999.

[25] Yves Benoist and Jean-François Quint. *Random walks on reductive groups*. Springer, 2016.

[26] Marc A Berger. Central limit theorem for products of random matrices. *Transactions of the American Mathematical Society*, 285(2):777–803, 1984.

[27] Gianmarco Bet, Fabio Coppini, and Francesca R Nardi. Weakly interacting oscillators on dense random graphs. *arXiv preprint arXiv:2006.07670*, 2020.

[28] 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.

[29] Bhaswar B Bhattacharya, Anirban Chatterjee, and Svante Janson. Fluctuations of subgraph counts in graphon based random graphs. *Combinatorics, Probability and Computing*, 32(3):428–464, 2023.

[30] Alessandra Bianchi, Francesca Collet, and Elena Magnanini. Limit theorems for exponential random graphs. *arXiv preprint arXiv:2105.06312*, 2021.

[31] Béla Bollobás and Oliver Riordan. Metrics for sparse graphs. *arXiv preprint arXiv:0708.1919*, 2007.

[32] J Adrian Bondy. Pancyclic graphs i. *Journal of Combinatorial Theory, Series B*, 11(1):80–84, 1971.

[33] 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.

[34] Christian Borgs, Jennifer Chayes, Henry Cohn, and Yufei Zhao. An l^p theory of sparse graph convergence i: Limits, sparse random graph models, and power law distributions. *Transactions of the American Mathematical Society*, 372(5):3019–3062, 2019.

[35] Christian Borgs, Jennifer Chayes, Julia Gaudio, Samantha Petti, and Subhabrata Sen. A large deviation principle for block models. *arXiv preprint arXiv:2007.14508*, 2020.

[36] Christian Borgs, Jennifer Chayes, Jeff Kahn, and László Lovász. Left and right convergence of graphs with bounded degree. *Random Structures & Algorithms*, 42(1):1–28, 2013.

[37] 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.

[38] Christian Borgs, Jennifer T Chayes, Henry Cohn, and Yufei Zhao. An l^p theory of sparse graph convergence ii: Ld convergence, quotients and right convergence. *The Annals of Probability*, 46(1):337–396, 2018.

[39] 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.

[40] 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.

[41] Vivek S. Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

[42] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[43] P. Brémaud. *Point Process Calculus in Time and Space: An Introduction with Applications*. Probability Theory and Stochastic Modelling. Springer International Publishing, 2020.

[44] Sébastien Bubeck. Convex Optimization: Algorithms and Complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[45] John Charles Butcher. *Numerical methods for ordinary differential equations*. John Wiley & Sons, 2016.

[46] Diana Cai, Nathanael Ackerman, and Cameron Freer. An iterative step-function estimator for graphons. *arXiv preprint arXiv:1412.2129*, 2014.

[47] René Carmona, Daniel B Cooney, Christy V Graves, and Mathieu Lauriere. Stochastic graphon games: I. the static case. *Mathematics of Operations Research*, 47(1):750–778, 2022.

[48] 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.

[49] Patrick Cattiaux, Arnaud Guillin, and Florent Malrieu. Probabilistic approach for granular media equations in the non uniformly convex case. *Probability Theory and Related Fields*, 140(1-2):19–40, 2008.

[50] Augustin-Louis Cauchy. Méthode générale pour la résolution des systemes d'équations simultanées. *Comptes Rendus de l'Académie des Science*, 25:536–538, 1847.

[51] Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: A review of models, methods and applications. I. Models and methods. *Kinetic and Related Models*, 15(6):895–1015, 2022.

[52] Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: A review of models, methods and applications. . Applications. *Kinetic and Related Models*, 15(6):1017–1173, 2022.

[53] Anirban Chatterjee, Soham Dan, and Bhaswar B Bhattacharya. Higher-order graphon theory: Fluctuations, degeneracies, and inference. *arXiv preprint arXiv:2404.13822*, 2024.

[54] Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

[55] Sourav Chatterjee. An introduction to large deviations for random graphs. *Bulletin of the American Mathematical Society*, 53(4):617–642, 2016.

[56] Sourav Chatterjee. *Large deviations for random graphs: École d'Été de Probabilités de Saint-Flour XLV-2015*, volume 2197. Springer, 2017.

[57] Sourav Chatterjee and Persi Diaconis. Estimating and understanding exponential random graph models. *The Annals of Statistics*, 41(5):2428–2461, 2013.

[58] Sourav Chatterjee and SR Srinivasa Varadhan. The large deviation principle for the Erdős-Rényi random graph. *European Journal of Combinatorics*, 32(7):1000–1017, 2011.

[59] Chi-Fang Chen, Hsin-Yuan Huang, Richard Kueng, and Joel A Tropp. Concentration for random product formulas. *PRX Quantum*, 2(4):040305, 2021.

[60] Bobbie G. Chern. *Large deviations approximation to normalizing constants in exponential models*. PhD thesis, Stanford University, 2016.

[61] Lénaïc Chizat. Mean-field langevin dynamics : Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022.

[62] 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*, NIPS'18, page 3040–3050, Red Hook, NY, USA, 2018. Curran Associates Inc.

[63] David Choi and Patrick J. Wolfe. Co-clustering separately exchangeable network data. *The Annals of Statistics*, 42(1), 2014.

[64] 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.

[65] Nicholas Cook and Amir Dembo. Large deviations of subgraph counts for sparse Erdős–Rényi graphs. *Advances in Mathematics*, 373:107289, 2020.

[66] Fabio Coppini. A note on Fokker–Planck equations and graphons. *Journal of Statistical Physics*, 187(2):1–12, 2022.

[67] Fabio Coppini, Anna De Crescenzo, and Huyen Pham. Nonlinear graphon mean-field systems. *arXiv preprint arXiv:2402.08628*, 2024.

[68] Fabio Coppini, Helge Dietert, and Giambattista Giacomin. A law of large numbers and large deviations for interacting diffusions on erdős–rényi graphs. *Stochastics and Dynamics*, 20(02):2050010, 2020.

[69] Anthony Cousien, Jean-Stéphane Dhersin, Viet Chi Tran, and Thi Phuong Thuy Vo. Respondent driven sampling on sparse erd\" os-r\'enyi graphs. *Acta Math Vietnam*, 48:479–513, 2023.

[70] Harry Crane. Dynamic random networks and their graph limits. *The Annals of Applied Probability*, 26(2):691 – 721, 2016.

[71] 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.

[72] 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.

[73] Persi Diaconis. The Markov chain Monte Carlo revolution. *Bulletin of the AMS*, 46(2):179–205, 2009.

[74] Persi Diaconis and Svante Janson. Graph limits and exchangeable random graphs. *Rendiconti di Matematica e delle sue Applicazioni*, 28(1):33–61, 2008.

[75] Peter Diao, Dominique Guillot, Apoorva Khare, and Bala Rajaratnam. Differential calculus on graphon space. *Journal of Combinatorial Theory, Series A*, 133:183–227, 2015.

[76] L. Dobrushin, R. Vlasov equations. *Functional Analysis and its applications*, 13:115–123, 1979.

[77] 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.

[78] 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.

[79] Eugene B Dynkin and Avishai Mandelbaum. Symmetric statistics, poisson point processes, and multiple wiener integrals. *The Annals of Statistics*, pages 739–745, 1983.

[80] 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.

[81] Jordan Emme and Pascal Hubert. Limit laws for random matrix products. *Mathematical Research Letters*, 25(4):1205–1212, 2018.

[82] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*. John Wiley & Sons, USA, 2009.

[83] Valentin Féray, Pierre-Loïc Méliot, and Ashkan Nikeghbali. Graphons, permutons and the thoma simplex: three mod-gaussian moduli spaces. *Proceedings of the London Mathematical Society*, 121(4):876–926, 2020.

[84] Stephen E Fienberg. Introduction to papers on the modeling and analysis of network data. *The Annals of Applied Statistics*, pages 1–4, 2010.

[85] Christoph Fretter, Matthias Müller-Hannemann, and Marc-Thorsten Hütt. Subgraph fluctuations in random graphs. *Physical Review E*, 85(5):056119, 2012.

[86] 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.

[87] Alan Frieze and Ravi Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

[88] Alex Furman. Random walks on groups and random transformations. In *Handbook of dynamical systems*, volume 1, pages 931–1014. Elsevier, 2002.

[89] Harry Furstenberg. Noncommuting random products. *Transactions of the American Mathematical Society*, 108(3):377–428, 1963.

[90] Harry Furstenberg and Harry Kesten. Products of random matrices. *The Annals of Mathematical Statistics*, 31(2):457–469, 1960.

[91] Wilfrid Gangbo and Robert J McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

[92] Wilfrid Gangbo and Adrian Tudorascu. On differentiability in the Wasserstein space and well-posedness for Hamilton-Jacobi equations. *Journal de Mathématiques Pures et Appliqués*, 125:119–174, 2019.

[93] Chao Gao, Yu Lu, and Harrison H. Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6), 2015.

[94] Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow gaussian processes. *arXiv preprint arXiv:1808.05587*, 2018.

[95] J. Gärtner. On the McKean-Vlasov limit for interacting diffusions. *Math. Nachr.*, 137:197–248, 1988.

[96] Saeid Ghafouri and Seyed Hossein Khasteh. A survey on exponential random graph models: an application perspective. *PeerJ Computer Science*, 6:e269, 2020.

[97] Daniel Glasscock. What is... a graphon. *Notices of the AMS*, 62(1):46–48, 2015.

[98] Jan Grebík and Oleg Pikhurko. Large deviation principles for block and step graphon random graph models. *arXiv preprint arXiv:2101.07025*, 2021.

[99] Pablo Groisman, Ruojun Huang, and Hernán Vivas. The kuramoto model on dynamic random graphs. *Nonlinearity*, 36(11):6177, 2023.

[100] 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.

[101] 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.

[102] Z. Harchaoui, S. Oh, S. Pal, R. Somani, and R. Tripathi. Stochastic optimization on matrices and a graphon McKean-Vlasov limit. arXiv preprint arXiv:2210.00422, 2022.

[103] Hamed Hatami. Graph norms and sidorenko's conjecture. *Israel Journal of Mathematics*, 175:125–150, 2010.

[104] Soufiane Hayou and Greg Yang. Width and depth limits commute in residual networks. In *International Conference on Machine Learning*, pages 12700–12723. PMLR, 2023.

[105] Amelia Henriksen and Rachel Ward. Concentration inequalities for random matrix products. *Linear Algebra and its Applications*, 594:81–94, 2020.

[106] Jan Hladkỳ, Christos Pelekis, and Matas Šileikis. A limit theorem for small cliques in inhomogeneous random graphs. *Journal of Graph Theory*, 97(4):578–599, 2021.

[107] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[108] D. N. Hoover. Relations on probability spaces and arrays of random variables, 1979. Preprint. Institute for Advanced Studies.

[109] D. N. Hoover. Row-column exchangeability and a generalized model for probability. *Exchangeability in probability and statistics (Rome, 1981)*, pages 281–291, 1982.

[110] 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.

[111] R.E. Huff. The Radon-Nikodým 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, 1977.

[112] 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.

[113] Chii-Ruey Hwang. Laplace's method revisited: Weak convergence of probability measures. *The Annals of Probability*, 8(6):1177–1182, 1980.

[114] Pierre-Emmanuel Jabin. A review of the mean field limits for vlasov equations. *Kinetic and Related Models*, 7(4):661–711, 2014.

[115] Svante Janson. A functional limit theorem for random graphs with applications to subgraph count statistics. *Random Structures & Algorithms*, 1(1):15–37, 1990.

[116] Svante Janson. *Orthogonal decompositions and functional limit theorems for random graph statistics*, volume 534. American Mathematical Soc., 1994.

[117] Svante Janson. Graphons, cut norm and distance, couplings and rearrangements. *NYJM Monographs*, 4, 2013.

[118] Svante Janson. Graphons and cut metric on sigma-finite measure spaces. arXiv preprint arXiv:1608.01833, 2016.

[119] Svante Janson and Krzysztof Nowicki. The asymptotic distributions of generalized u-statistics with applications to random graphs. *Probability theory and related fields*, 90:341–375, 1991.

[120] 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.

[121] Dudley Paul Johnson. Central limit theorems for random evolutions. *Stochastic Processes and their Applications*, 53(2):221–232, 1994.

[122] 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.

[123] Mark Kac. Foundations of kinetic theory. In *Proceedings of The third Berkeley symposium on mathematical statistics and probability*, volume 3, pages 171–197, 1956.

[124] O. Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Probability and Its Applications. Springer, New York, NY, 2005.

[125] O. Kallenberg. *Foundations of Modern Probability*. Probability Theory and Stochastic Modelling. Springer International Publishing, 2021.

[126] Olav Kallenberg. On the representation theorem for exchangeable arrays. *Journal of Multivariate Analysis*, 30(1):137–154, 1989.

[127] Olav Kallenberg. Multivariate sampling and the estimation problem for exchangeable arrays. *Journal of Theoretical Probability*, 12:859–883, 1999.

[128] Rajeeva L Karandikar and Bhamidi V Rao. *Introduction to stochastic calculus*. Springer, 2018.

[129] I. Karatzas and S. E. Shreve. *Brownian motion and stochastic calculus.*, volume 113 of *Graduate Texts in Mathematics*. Springer, second edition, 1991.

[130] Tarun Kathuria, Satyaki Mukherjee, and Nikhil Srivastava. On concentration inequalities for random matrix products. *arXiv preprint arXiv:2003.06319*, 2020.

[131] Gursharn Kaur and Adrian Röllin. Higher-order fluctuations in dense random graph models. *Electronic Journal of Probability*, 26:1–36, 2021.

[132] 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.

[133] Richard Kenyon and Mei Yin. On the asymptotics of constrained exponential random graphs. *Journal of Applied Probability*, 54(1):165–180, 2017.

[134] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[135] Nicolás Kim. *Local Structure and Inference in Large Random Graphs*. PhD thesis, Illinois Institute of Technology, 2018.

[136] 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.

[137] Lukasz Kruk, John Lehoczky, Kavita Ramanan, and Steven Shreve. An explicit formula for the Skorokhod map on [0, a]. *The Annals of Probability*, pages 1740–1768, 2007.

[138] Christian Kuehn. Network dynamics on graphops. *New Journal of Physics*, 22(5):053030, 2020.

[139] Dávid Kunszenti-Kovács. Uniqueness of banach space valued graphons. *Journal of Mathematical Analysis and Applications*, 474(1):413–440, 2019.

[140] Dávid Kunszenti-Kovács, László Lovász, and Balázs Szegedy. Multigraph limits, unbounded kernels, and banach space decorated graphs. *arXiv preprint arXiv:1406.7846*, 2014.

[141] Harold Kushner and G. George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[142] Harold Joseph Kushner and Dean S. Clark. *Stochastic approximation methods for constrained and unconstrained systems*, volume 26. Springer Science & Business Media, 2012.

[143] Daniel Lacker, Kavita Ramanan, and Ruoyu Wu. Local weak convergence for sparse networks of interacting processes. *arXiv preprint arXiv:1904.02585*, 2019.

[144] Daniel Lacker and Agathe Soret. A label-state formulation of stochastic graphon games and approximate equilibria on large networks. *Mathematics of Operations Research*, 48(4):1987–2018, 2023.

[145] François Ledrappier. Some asymptotic properties of random walks on free groups. In *CRM Proceedings and Lecture Notes*, pages 117–152. American Mathematical Society, 2001.

[146] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Society, Providence, RI, 2017.

[147] 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.

[148] Qianyi Li and Haim Sompolinsky. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Physical Review X*, 11(3):031059, 2021.

[149] László Lovász. Subgraph densities in signed graphons and the local sidorenko conjecture. *arXiv preprint arXiv:1004.3026*, 2010.

[150] László Lovász. *Large Networks and Graph Limits*, volume 60 of *Colloquium publications*. American Mathematical Society, 2012.

[151] László Lovász and Balázs Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957, 2006.

[152] László Lovász and Balázs Szegedy. Szemerédi's lemma for the analyst. *Geometric And Functional Analysis*, 17:252–270, 2007.

[153] László Lovász and Balázs Szegedy. Limits of compact decorated graphs. arXiv preprint arXiv:1010.5155, 2010.

[154] Eyal Lubetzky and Yufei Zhao. On replica symmetry of large deviations in random graphs. *Random Structures & Algorithms*, 47(1):109–146, 2015.

[155] Willem Mantel. Problem 28. *Wiskundige Opgaven*, 10(2):60–61, 1907.

[156] Robert J McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.

[157] H. P. McKean. Fluctuations in the kinetic theory of gases. *Communications on Pure and Applied Mathematics*, 28(4):435–455, 1975.

[158] Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Found. Comput. Math*, 11:417–487, 2011.

[159] 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.

[160] James R Munkres. *Topology*. Prentice Hall Upper Saddle River, 2000.

[161] J. Neeman, C. Radin, and L. Sadun. Phase transitions in finite random networks. *Journal of Statistical Physics*, 181:305–328, 2020.

[162] J. Neeman, C. Radin, and L. Sadun. Typical large graphs with given edge and triangle densities. *PTRF*, 186:1167–1223, 2023.

[163] 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.

[164] Praneeth Netrapalli. Stochastic gradient descent and its variants in machine learning. *Journal of the Indian Institute of Science*, 99(2):201–213, 2019.

[165] 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.

[166] Krzysztof Nowicki and John C Wierman. Subgraph counts in random graphs using incomplete u-statistics methods. *Discrete Mathematics*, 72(1-3):299–310, 1988.

[167] Sewoong Oh, Soumik Pal, Raghav Somani, and Raghav Tripathi. Gradient flows on graphons: existence, convergence, continuity equations. arXiv preprint arXiv:2111.09459, 2021.

[168] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15:267–273, 1982.

[169] Sofia C Olhede and Patrick J Wolfe. Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences*, 111(41):14722–14727, 2014.

[170] 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.

[171] 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.

[172] Peter Orbanz and Daniel M Roy. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):437–461, 2014.

[173] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.

[174] Felix Otto and Cédric Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

[175] Francesca Parise and Asuman Ozdaglar. Graphon games. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 457–458, 2019.

[176] Juyong Park and Mark EJ Newman. Statistical mechanics of networks. *Physical Review E*, 70(6):066117, 2004.

[177] Valentin V Petrov. *Sums of independent random variables*, volume 82. Springer, Berlin, Heidelberg, 2011.

[178] Oleg Pikhurko et al. Large deviation principles for graphon sampling. *arXiv preprint arXiv:2311.06531*, 2023.

[179] Oleg Pikhurko and Alexander Razborov. Asymptotic structure of graphs with the minimum number of triangles. *Combinatorics, Probability and Computing*, 26(1):138–160, 2017.

[180] C. Radin, K. Ren, and L. Sadun. The asymptotics of large constrained graphs. *Journal of Physics A; Mathematical and Theoretical*, 47(17), 2014.

[181] Charles Radin and Lorenzo Sadun. Optimal graphons in the edge-2star model. arXiv preprint arXiv:2305.00333, 2023.

[182] Charles Radin and Mei Yin. Phase transitions in exponential random graphs. *The Annals of Applied Probability*, pages 2458–2471, 2013.

[183] Alexander A Razborov. On the minimal density of triangles in graphs. *Combinatorics, Probability and Computing*, 17(4):603–618, 2008.

[184] D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2004.

[185] Matthew Richey. The evolution of markov chain monte carlo methods. *The American Mathematical Monthly*, 117(5):pp. 383–413, 2010.

[186] Hannes Risken and Hannes Risken. *Fokker-planck equation*. Springer, 1996.

[187] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951.

[188] 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.

[189] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. arXiv preprint arXiv:1306.2872, 2013.

[190] Luana Ruiz, Luiz Chamon, and Alejandro Ribeiro. Graphon neural networks and the transferability of graph neural networks. *Advances in Neural Information Processing Systems*, 33:1702–1712, 2020.

[191] 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.

[192] 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, 2015.

[193] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

[194] 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.

[195] Alexander Sidorenko. A correlation inequality for bipartite graphs. *Graphs and Combinatorics*, 9:201–204, 1993.

[196] 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.

[197] 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.

[198] Leszek Słomiński. On approximation of solutions of multidimensional SDE's with reflecting boundary conditions. *Stochastic processes and their Applications*, 50(2):197–219, 1994.

[199] Leszek Słomiński. Euler's approximations of solutions of SDEs with reflecting boundary. *Stochastic processes and their applications*, 94(2):317–337, 2001.

[200] Mei Song, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.

[201] 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.

[202] Karl-Theodor Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. Available at arXiv:1208.0434v1, 2012.

[203] Alain-Sol Sznitman. Nonlinear reflecting diffusion process, and the propagation of chaos and fluctuations associated. *Journal of Functional Analysis*, 56(3):311–336, 1984.

[204] 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.

[205] H. Tanaka. Probabilistic treatment of the Boltzmann equation of Maxwellian molecules. *Z. Wahrsch. Verw. Gebiete*, 46(1):67–105, 1978/79.

[206] Behrouz Touri and Angelia Nedić. Product of random stochastic matrices. *IEEE Transactions on Automatic Control*, 59(2):437–448, 2013.

[207] VN Tutubalin. On limit theorems for the product of random matrices. *Theory of Probability & Its Applications*, 10(1):15–27, 1965.

[208] 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.

[209] 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.

[210] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018.

[211] 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.

[212] Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.

[213] Joseph C Watkins. A central limit problem in random evolutions. *The Annals of Probability*, pages 480–513, 1984.

[214] Patrick J Wolfe and Sofia C Olhede. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.

[215] Ruoyu Wu. Open problem—weakly interacting particle systems on dense random graphs. *Stochastic Systems*, 9(3):315–317, 2019.

[216] Greg Yang and Edward J. Hu. Tensor Programs IV: Feature Learning in Infinite-Width Neural 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 11727–11737. PMLR, 18–24 Jul 2021.

[217] 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.

[218] 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.