# The search for an organizing physical framework for statistics

Colin LaMont

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

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Paul Wiggins, Chair

Mathias Drton

Jason Detwiler

Program Authorized to Offer Degree:
Physics

University of Washington

**Abstract**

The search for an organizing physical framework for statistics

Colin LaMont

Chair of the Supervisory Committee:
Paul Wiggins
Department of Biophysics

Theories of statistical analysis remain in conflict and contradiction. But nature reveals an elegant and coherent formulation of statistics in the thermal properties of physical systems. Demanding that a viable statistical theory share the properties of a viable physical theory—observer independence and coordinate invariance—resolves outstanding controversies in statistical model selection. Furthermore, by using constructions taken directly from thermodynamics, a predictive approach to model selection can be reconciled with a Bayesian approach to parameter uncertainty. This approach also solves the longstanding problem of the undetermined Bayesian prior.

# TABLE OF CONTENTS

Page

List of Figures . . . . . vi

Chapter 1: Introduction . . . . . 1
1.0.1 Method of maximum likelihood . . . . . 1
1.0.2 A unified framework for statistics? . . . . . 4
1.1 Model selection . . . . . 5
1.1.1 The change point problem . . . . . 6
1.1.2 Model selection and predictivity. . . . . 7
1.2 Uncertainty on the parameters . . . . . 8
1.3 AIC vs. BIC? . . . . . 9
1.4 Layout and contributions of the thesis . . . . . 11

Chapter 2: Change-point information criterion supplement . . . . . 13
2.1 Preliminaries . . . . . 15
2.2 Information criterion for change-point analysis . . . . . 20
2.3 The relation between frequentist and information-based approach . . . . . 27
2.4 Applications . . . . . 30
2.5 Discussion . . . . . 30

Chapter 3: Predictive model selection . . . . . 32
3.1 Introduction to Model selection . . . . . 33
3.2 Information-based inference . . . . . 34
3.2.1 Information criteria . . . . . 35
3.2.2 The Akaike Information Criterion . . . . . 37
3.3 Complexity Landscapes . . . . . 38
3.3.1 The finite-sample-size complexity of regular models . . . . . 38
3.3.2 Singular models . . . . . 42

3.3.3 Constrained models . . . 43
3.4 Frequentist Information Criterion (QIC) . . . 44
3.4.1 Measuring model selection performance . . . 45
3.5 Applications of QIC . . . 45
3.5.1 Small sample size and the step-size analysis . . . 46
3.5.2 Anomalously large complexity and the Fourier regression model . . . 49
3.5.3 Anomalously small complexity and the exponential mixture model . . . 55
3.6 Discussion . . . 57
3.6.1 QIC subsumes extends both AIC and AICc . . . 58
3.6.2 Asymptotic bias of the QIC complexity . . . 59
3.6.3 Advantages of QIC . . . 60
3.6.4 Conclusion . . . 62

Chapter 4: From postdiction to prediction . . . 64
4.1 Introduction . . . 65
4.1.1 A simple example of the Lindley paradox . . . 66
4.2 Data partition . . . 68
4.2.1 The definition of frequentism and Bayesian paradigms . . . 68
4.2.2 Prior information content . . . 69
4.2.3 The Bayesian cross entropy . . . 70
4.2.4 Pseudo-Bayes factors . . . 71
4.2.5 Information Criteria . . . 71
4.2.6 Decision rules and resolution . . . 73
4.3 The Lindley paradox . . . 73
4.3.1 Classical Bayes and the Bartlett paradox . . . 74
4.3.2 Minimal training set and Lindley Paradox . . . 75
4.3.3 Frequentist prescription and AIC . . . 76
4.3.4 Log evidence versus AIC . . . 79
4.3.5 When will a Lindley paradox occur? . . . 79
4.4 Discussion . . . 80
4.4.1 A canonical Bayesian perspective on the Lindley paradox . . . 80
4.4.2 Circumventing the Lindley paradox . . . 81
4.4.3 Loss of resolution . . . 82

4.4.4 The multiple comparisons problem . . . 83
4.4.5 Statistical significance does not imply scientific significance . . . 84
4.4.6 Conclusion . . . 84

Chapter 5: Thermodynamic correspondence . . . 87
5.1 Contributions . . . 87
5.2 Defining the correspondence . . . 89
5.2.1 Application of thermodynamic identities . . . 90
5.3 Results . . . 90
5.3.1 Models . . . 90
5.3.2 Thermodynamic potentials . . . 92
5.3.3 The principle of indifference . . . 94
5.3.4 A generalized principle of indifference . . . 96
5.4 Applications . . . 97
5.4.1 Learning Capacity . . . 98
5.4.2 Generalized principle of indifference . . . 108
5.4.3 Inference . . . 113
5.5 Discussion . . . 118
5.5.1 Learning capacity . . . 118
5.5.2 The generalized principle of indifference . . . 120
5.5.3 Comparison with existing approaches and novel features . . . 124
5.5.4 Conclusion . . . 126

Chapter 6: Conclusion . . . 127
6.1 The thermodynamics of inference . . . 128
6.2 Unanswered questions . . . 130

Appendix A: QIC for change-point algorithms and Brownian bridges . . . 147
A.1 Type I errors . . . 147
A.1.1 Asymptotic form of the complexity function . . . 147

Appendix B: QIC calculations . . . 154
B.1 Complexity Calculations . . . 154
B.1.1 Exponential families . . . 154

B.1.2 Modified-Centered-Gaussian distribution . . . 155
B.1.3 The component selection model . . . 155
B.1.4 n-cone . . . 156
B.1.5 Fourier Regression nesting complexity . . . 157
B.1.6 $L_1$ Constraint . . . 158
B.1.7 Curvature and QIC unbiasedness under non-realizability . . . 158
B.1.8 Approximations for marginal likelihood . . . 159
B.1.9 Seasonal dependence of the neutrino intensity . . . 160

Appendix C: Lindley Paradox . . . 162
C.1 Calculation of Complexity . . . 162
C.2 Definitions and Calculations . . . 163
C.2.1 Volume of a distribution . . . 163
C.2.2 Showing $I = I'$ to zeroth order . . . 164
C.2.3 Significance level implied by a data partition . . . 164
C.3 Other Methods . . . 164
C.3.1 Other Predictive Estimators . . . 165
C.3.2 Data-Validated Posterior and Double use of Data . . . 165
C.3.3 Fractional Bayes Factor . . . 166
C.4 Efficiency and correct models . . . 166

Appendix D: Thermodynamics Supplement . . . 169
D.1 Supplemental results . . . 169
D.1.1 Definitions of information, cross entropy, Fisher information matrix . 169
D.1.2 An alternate correspondence . . . 170
D.1.3 Finite difference is equivalent to cross validation . . . 170
D.1.4 Jeffreys prior is proportional to GPI prior in the large-sample-size limit 171
D.1.5 Reparametrization invariance of thermodynamic functions and the GPI prior . . . 172
D.1.6 Effective temperature of confinement . . . 173
D.1.7 A Bayesian re-interpretation . . . 173
D.2 Methods . . . 174
D.2.1 Computation of learning capacity . . . 174
D.2.2 Direct computation of GPI prior . . . 174

D.2.3 Computation of the free energy using a sufficient statistic . . . . . 175
D.2.4 Computation of the GPI prior using a recursive approximation . . . . 176
D.3 Details of applications . . . . . 177
D.3.1 Normal model with unknown mean and informative prior . . . . . 177
D.3.2 Normal model with unknown mean . . . . . 178
D.3.3 Normal model with unknown discrete mean . . . . . 179
D.3.4 Normal model unknown mean and variance . . . . . 181
D.3.5 Exponential model . . . . . 182
D.3.6 Uniform distribution . . . . . 183
D.3.7 Poisson stoichiometry problem . . . . . 184

# LIST OF FIGURES

Figure Number Page

1.1 **Schematic of a statistical model** A statistical model consists of a space of observed data, a space of parameters, and maps between the two. Candidate distributions describing the data are identified with points in a model or parameter space $\Theta$, the coordinates of which, $\theta_1, \theta_2, \ldots, \theta_K$, are the parameters. Data sets can be mapped to parameters through a learning or estimation method $\hat{\theta}(X^N) = \hat{\theta}_X$, such as the method of maximum likelihood. More generally the estimation process results in a *distribution of parameter values* such as in the Bayes procedure. We will consider both point estimates and distribution estimates. . . . . 3

1.2 **Panel A: Biophysical system which exhibits change points.** One potential application of Change-Point Analysis is to the characterization molecular-motor stepping along a cytoskeletal filament. **Panel B: Schematic of Change-Point Analysis.** A change-point model of motor stepping is shown for a series of position states. The blue dots represent measurements of motor position, corrupted by noise. The red line represents the change-point model for the true motor position. Each frame shows the optimal fit for $n = 1...8$ position states. From the figure, it is intuitively clear that $n = 4$ is the correct number of position states. Models with additional states improve the fit to the observed data but would result in information loss for an independent set of measurements of the same motor positions. . . . . 6

2.1 **Nesting complexity for AIC, QIC and BIC.** The nesting complexity is plotted for three state dimensions $d = \{1, 3, 6\}$ and $n = 2$. First note that the AIC penalty is much smaller than the other two nesting complexities. BIC is empirically known to produce acceptable results under some circumstances. For sufficiently large samples ($N$), the $k_{\mathrm{BIC}} > k_{\mathrm{QIC}}$, resulting in over penalization and the rejection of states that are supported statistically. This effect is more pronounced for large state dimension $d$ where the crossover occurs for small observation number $N$. $k_{\mathrm{BIC}}$ is too small for a wide range of sample sizes, resulting in over segmentation. . . . . 25

2.2 **Information-based model selection. Panel A: Nested models generated by a Change-Point Algorithm.** Simulated data (blue points) generated by a true model with four states is fit to a family of nested models (red lines) using a Change-Point Algorithm. Models fit with $1 \leq n \leq 8$ states are plotted. The fit change points are represented as vertical black lines. The true model has four states. **Panel B: Four changes points minimizes information loss.** Both the expectation of the information (red) and the cross entropy (green) are plotted as a function of the number of states $n$. The y-axis ($h$, information) is split to show the initial large changes in $h$ as well as the subsequent smaller changes for $4 \leq n \leq 8$. The cross entropy (green) is minimized by the the model that best approximates the truth ($n = 4$). The addition of parameters leads to an increase in cross entropy (green) for $n \geq 4$. The information loss estimator (red) is biased and continues to decreases with the addition of states as a consequence of overfitting. **Panel C: Complexity of Change-Point Analysis.** The true complexity is computed for the model shown in panel A via Monte Carlo simulation for $10^6$ realizations of the observations $X^N$ and compared with three models for the complexity AIC, QIC and BIC. For models with states numbering $1 \leq n \leq 4$, the true complexity (black) is correctly estimated by the AIC complexity (red dotted) and the QIC complexity (green). But for a larger number of states ($4 \leq n \leq 8$), only QIC accurately estimates the true complexity. . . . . 28

3.1 **Complexity at finite sample size.** Although AIC estimate accurately estimates the large-sample-size limit of the complexity of regular models, there can be significant finite-sample-size corrections. For instance, the modified-center-Gaussian model has a significantly larger complexity than the AIC limit for small $N$. In fact the complexity diverges for $N \leq \alpha$, implying that the model has insufficient data to make predictions. . . . . 39

3.2 **Complexity landscapes in singular models. AIC underestimates the complexity in the component selection model. Panel A-B:** Schematic sketches of the geometry of parameter space for two different multiplicity values: $n = 3$ and $n = 6$. **Panel C:** The AIC estimate $\mathcal{K} = 1$ (dashed line) matches the true complexity far from the singular point ($|\mu/\sigma| \gg 0$). Close to the singularity ($|\mu/\sigma| \approx 0$), the true complexity is much larger than the AIC estimate. The complexity grows with the number of means $n$ due to *multiplicity*. **AIC overestimates the complexity in the n-cone model. Panel D-E:** Schematic sketches of parameter space for a wide cone ($c = 1$) and a needle-like cone ($n = 0.1$). **Panel F:** For $n = 10$ dimensions, the AIC estimate $\mathcal{K} = n - 1$ (dashed line) matches the true complexity far from the singular point ($|\mu_1/\sigma| \gg 0$). Close to the singularity ($|\mu_1/\sigma| \approx 0$), the true complexity is much smaller than the AIC estimate. The complexity shrinks for small cone angles ($c \to 0$) since the cone geometry is needle-like with effectively a single degree for freedom ($\mu_1$). . . . . 40

3.3 **Complexity of $L_1$-constrained model. Panel A:** Schematic sketch of a slice of the seven-dimensional parameter space. Parameter values satisfying the $L_1$ constrain lie inside the simplex. **Panel B:** Complexity as a function of the true parameter value $\vec{\mu} = (\mu_1, ..., \mu_7)$. (Only a slice representing the $x$-$y$ plane is shown.) The black-hatched region represents parameter values satisfying the constraints. The $L_1$ constraint significantly reduces the complexity below the AIC estimate $\mathcal{K} = 7$. The complexity is lowest outside the boundaries of the simplex where the constraints trap MLE parameter estimates and reduce statistical fluctuations. . . . . 41

3.4 **Monte Carlo histogram of the LOOCV procedure:** A histogram of $10^6$ simulations of the effective cross-validation complexity $\mathcal{K}_{\mathrm{CV}}(X)$ for the normal model with unknown variance $\alpha = 2$ compared to the QIC result $\mathcal{K} = \frac{5}{5-2}$ for ($N = 5$). The lower variance of the QIC complexity often results in better model selection properties, especially at low sample sizes relative to cross-validation. . . . . 46

3.5 **Panel A: Truth, data and models.** (Simulated for $N = 100$.) The true mean intensity is plotted (solid green) as a function of season, along with the simulated observations (green points) and models fitted using two different algorithms, sequential (red) and greedy (blue). **Panel B: Failure of AIC for greedy algorithm.** (Simulated for $N = 100$.) For the greedy algorithm, the coefficients selected using AIC (red) are contrasted with the coefficients chosen using QIC. The QIC mean estimates (blue) track the true means very closely, unlike the AIC selected mean. **Panel C: Information as a function of model dimension.**(Simulated for $N = 100$.) The information is plotted as a function of the nesting index $n$. The dashed curves represent the information as a function of nesting index and both are monotonically decreasing. The solid curves (red and blue) represents the estimated average information (QIC), which is equivalent to estimated model predictivity. **Panel D: The true complexity matches** QIC **estimates.** (Simulated for $N = 1000$.) In the sequential-algorithm model, the true complexity (red dots) is AIC-like (solid red). In the greedy-algorithm model, the true complexity (blue dots) transitions from AIC-like (slope = 1) to BIC-like (slope $\propto \log N$) at $n = 4$. In both cases, the true complexity is correctly predicted by QIC (solid curve). 52

3.6 **Panel A: Performance of the sequential algorithm.** Simulated performance as measured by the KL Divergence $\overline{D}$ (Eqn. (3.23)) of sequential algorithm at different sample sizes using AIC, QIC and BIC (lower is better). AIC and QIC are identical in this case; they differ only because of the finite number of Monte Carlo samples. Larger fluctuations are arise from the structure of true modes at the resolvable scale of a given sample size. **Panel B: Performance of the greedy algorithm.** Simulated performance of greedy algorithm as measured by the KL Divergence $\overline{D}$ (Eqn. (3.23)) at different sample sizes using AIC, QIC and BIC (lower is better). QIC and BIC have very similar cutoff penalties. Because of the algorithmic sensitivity, QIC can have the appropriately complexity scaling with $N$ in both the greedy and sequential case. . . . . 54

4.1 **Loss of resolution in Bayesian inference. Panel A:** The resolution on detected bead displacement (the alternative hypothesis) is plotted as a function of sample size $N$. The increase in resolution is due to the decrease in the error in the mean $\sigma_\mu = \sigma/\sqrt{N}$. The resolution of both frequentist and Bayesian inference increase, but the frequentist resolution is higher. A dotted line represents the size of a putative displacement. The frequentist analysis detects this displacement at a smaller sample size than the Bayesian analysis. **Panel B:** To draw attention to the difference between Bayesian and frequentist resolution, we plot the resolution relative to the frequentist resolution $\mu_F$. To illustrate the prior dependence of the Bayesian analysis, we have drawn curves corresponding to various choices of prior volume $V_0$. . . . . 67

4.2 **Complexity as a function of data partition.** Complexity can be understood as a penalty for model dimension $K$. The data partition parameter controls the relative amount of information in the prior. In a predictive limit ($\nu \to 0$), the training set is large compared with the generalization set and the complexity is small. This is the AIC limit. At the other extreme ($\nu \to \infty$), all the data is partitioned into the generalization set and therefore the prior is uninformative and the complexity is large. This is the BIC limit. . . . . 72

4.3 **The geometry of the Occam factor.** The total volume of plausible parameter values for a model is $V_0$. The volume of acceptable parameter values after a single measurement is $V_1$. The volume of acceptable parameter values after $N$ measurements is $V_N$. The Occam factor is defined as the probability of randomly drawing a parameter from initial volume $V_0$ consistent with the $N$ measurements: $\Pr \approx \mathcal{N}^{-1}$ where $\mathcal{N} \equiv V_0/V_N$ is the number of distinguishable distributions after $N$ measurements. Lower dimension models are naturally favored by the Occam factor since the number of distinguishable models $\mathcal{N}$ is smaller. . . . . 74

4.4 **Visualizing the pre and postdictive decision rules.** The cross entropy difference for Bayesian inference (postdiction: $N|0$) and the predictive limit $(1|N-1)$ are plotted as a function of sample size $N$. $\Delta H > 0$ results in the selection of the alternative hypothesis. Both measures initially favor the null hypothesis. The use of a more (less) informative prior causes the postdictive curve to be shifted up (down). Since the predictive $H$ is the derivative of the postdictive $H$, the prior does not influence the inference of the predictive observer. The predictive curve crosses zero first, leading the predictive observer to support the alternative hypothesis. Since the predictive $H$ is the derivative of the postdictive $H$ with respect to $N$, the sample size at which the predictive observer switches to the alternative hypothesis corresponds to the sample size where the postdictive observer has the most evidence for the null hypothesis. The two measures are in conflict (grey region) until the postdictive $H$ crosses zero at a significantly larger sample size $N$. The Bayesian therefore requires significantly more evidence to reach the same conclusion as the predictive observer. . . . . 77

4.5 **Significance level as a function of data partition.** To make an explicit connection between the frequentist significance level and the data partition, it is useful to compute the significance level implied by the predictive decision rule. In the postdictive regime, corresponding to an uninformative prior, the significance level steeply decreases with increasing $\nu$, resulting in a strong Lindley paradox. . . . . 78

5.1 **Understanding Gibbs entropy.** The Gibbs entropy for the normal-model-with-prior is plotted as a function of sample size. The Gibbs entropy can be understood heuristically as the log ratio of the model consistent with the data to allowed models. At small sample size, the model structure determines the parametrization and therefore all models allowed are consistent with the data and there is zero Gibbs entropy. As the sample size grows beyond the critical sample size $N_0$, fewer and fewer of the allowed models are consistent with the data and the entropy decreases like $-\frac{1}{2}K \log N$. The non-positivity of the Gibbs entropy is a direct consequence of the normalization of the prior, which forces the Gibbs entropy to have a maximum value of zero. A prior determined by the generalized principle of indifference avoids this non-physical result. . 95

5.2 **The failure of equipartition.** The behaviors of the heat capacity and learning capacity are compared and related to the applicability or inapplicability of the Equipartition theorem in different regimes. **Panel A: Low-temperature freeze-out in a quantum system.** The heat capacity is plotted as a function of temperature. Equipartition predicts heat capacity should be constant, equal to half the degrees of freedom in the system. Plateaus can be observed at half-integer values, but the number of degrees of freedom is temperature dependent due to the discrete nature of quantum energy levels. At low temperature, some degrees of freedom are frozen out since the first excited state is thermally inaccessible. **Panel B: High-temperature freeze-out in the Learning capacity.** Analogous to the statistical mechanics system, the statistical learning capacity transitions between half integer plateaus, reflecting a temperature-dependent number of degrees of freedom. At low sample size $N$ (high temperature), the parameters are completely specified by model constraints (the prior) and therefore the parameters do not contribute to the learning capacity. At large sample size $N$, the parameters become data dominated and therefore the learning capacity is predicted by equipartition ($\frac{1}{2}K$). . . . . 99

5.3 **Panel A: Learning capacity at finite sample size.** At large sample size, equipartition predicts the learning capacity of all models. At sample size $N = 1$ the learning capacity diverges for models with unknown variance since the mean and variance cannot be simultaneously estimated from a single measurement. **Panel B: Learning capacity on a discrete manifold.** The learning capacity of a normal model with an unknown D-dimensional mean $\vec{\mu} \in \mathbb{Z}^D$ and variance $\sigma^2 = 15$. For statistical uncertainty $\delta\mu \gg 1$, the learning capacity is predicted by equipartition since the discrete nature of the parameter manifold cannot be statistically resolved. For $\delta\mu \ll 1$, there is no statistical uncertainty in the parameter value (due to the discreteness of $\mu$) and the degrees of freedom freeze out, giving a learning capacity of zero. . . 101

5.4 **Panel A: Posterior for low-temperature freeze-out.** Low-temperature freeze-out occurs when there is no statistical ambiguity in the parameter value. For long interval durations ($\lambda t = 500$) the posterior only weights a single parameter value ($m = 6$) whereas the manifold is effectively continuous for intermediate interval durations ($\lambda t = 10$) and multiple parameter values as weighted. **Panel B: Learning capacity for low-temperature freeze-out.** For long interval durations ($\lambda t = 500$), the stoichiometry $m$ is frozen-out and therefore the learning capacity is zero. For intermediate interval durations ($\lambda t = 10$), equipartition applies and the learning capacity is one-half. At short intervals, the large-sample-size limit assumption is violated and the learning capacity diverges from the limiting value. . . . . 105

5.5 **Effective complexity of models at finite sample size.** We computed the exact GPI prior for a series of models of different dimension. At large sample size, the dimension determines the effective complexity: $\mathscr{K} = \dim \Theta/2$. At finite sample size there are significant corrections. The effective complexity divergences for the normal model (dashed curves) with unknown mean and variance at $N = 1$. . . . . 111

5.6 **Complementary views on indifference:** A flat prior and the GPI prior offer two different ways to define the principle of indifference. **Panel A: The GPI prior depends on parameter and sample-size.** The flat prior is constant with respect to changes in source number $m$, and sample size, while the GPI prior (for Poisson stoichiometry problem) changes with the parameter $m$, and responds to sample size. **Panel B: The GPI prior has (nearly) constant entropy.** The average entropy under the GPI prior is almost flat and zero everywhere, but the entropy of the flat prior is *not constant*. Some models are entropically favored under the flat prior in violation of the generalized principle of indifference. . . . . 114

5.7 **Bayesian inference on model identity.** The posterior of model identity ($y$ axis) was computed for datasets generated by each model ($x$ axis). **Panel A: GPI prior.** For the simulated datasets, the generative model had the highest posterior probability as expected. **Panel B: Normalized objective prior.** The non-compactness of the parameter manifolds implies automatic rejection of all the higher-dimensional models. In this case, since the model $\mathcal{N}$ is parameter-free, it has posterior probability of 1 for all datasets, regardless of the fit. **Panel C: Revised informative prior.** To avoid this undesirable anomaly, we tune the prior parameter support to result in a reasonable posterior model probability. (See Tab. D.1.) Inference is no longer objective, as the posterior probabilities depend on how this tuning is performed. One representative plot of posterior probabilities to shown. In general, inference cannot be expect to identify the generative model unless the KL divergence is so large as to make the prior irrelevant. . . . . 115

5.8 **Panel A: Sloppiness is determined by parameter manifold geometry and posterior width.** Parameters are defined on a compact manifold $\boldsymbol{\Theta}$. In sloppy regions of parameter manifold, the parameters are model-structure dominated (red posterior) whereas in regular regions of parameter manifold parameters are data dominated (green posterior). From the perspective of the learning capacity, the model is effectively one dimensional in proximity to the red posterior and two dimensional in proximity to the green posterior. **Panel B: Generalized principle of indifference.** The posterior distribution $\varpi(\boldsymbol{\theta}_0|X^N)$ is shown schematically for two different sample sizes $N$. The resolution increases with sample size as the posterior shrinks. In the GPI prior (Eqn. 5.12), all parameter values consistent with the posterior are assigned unit prior weight collectively. . . . . 119

A.1 **Probability of a false positive change-point.** The probability of a false positive change-point is shown as a function of the number of observations in the interval length $N$ for three different model dimensions. . . . . 148

A.2 **Panel A: Brownian Walk and Brownian Bridge.** A visualization of a random walk $\boldsymbol{X}'_{[}1, n]$ (blue) and the corresponding Brownian bridge $\boldsymbol{B}'_n$ (red). **Panel B: Law of Iterated Logs.** A visualization of $S_n/\sqrt{n \log\log n}$ (blue) plotted as an orthographic projection as a function of $n$. $\sqrt{2}$ (red) is the limit of the supremum. . . . . 151

D.1 **Jeffreys prior and GPI prior at $t = 1$** The power law behavior of the GPI prior and the Jeffreys prior ($m^{-1/2}$) for the Poisson problem at $t = 1$ are compared on a log-log plot. At large fluorophore number, the discrete problem is very similar to the continuous problem, and the GPI prior converges to the same power law behavior as the Jeffreys prior. At small sample size, effects from the discretization deform the GPI prior away from the Jeffreys prior. The normalization of the Jeffreys prior has been chosen to make the two priors match at large $m$. . . . . 185

# ACKNOWLEDGMENTS

I wish to express sincere appreciation to Paul Wiggins, whose curiosity and insight have lead to many wonderful conversations and an amazingly fun project, and Chad Heilig, whose inspirational teaching got us into statistics. I also wishes to thank my committee members, especially my readers, who provided thoughtful conversation and met these ideas with active interest, and lab members Sarah Mangiameli and Julie Cass, for their helpful feedback during the revisions process (errors which remain are the authors alone). I also wish to thank those who have written about statistics with humanity and beauty: especially Gibbs, Fisher, Amari, Barndorf–Nielsen, Akaike, Gelfand, Dey, Tukey, Burnham and Anderson, Watanabe, and Mayo.

# DEDICATION

To my parents

# Chapter 1
# INTRODUCTION

Data analysis represents a significant portion of the time spent on scientific research for the average experimentalist. It is therefore surprising that the way data analysis is taught in statistics classes is almost laughably crude. Introductory lab or data analysis classes will suggest the following algorithm for measuring a quantity:

1. relate measurements to a line,

2. assume the data are normally distributed, and

3. do weighted-least-squares regression.

For most experiments, this procedure is going to be entirely hopeless. There is nothing "linear" to regress on in the measurements of a stationary stochastic process, or the RF resonance of nuclear spin. Generally, the experiment had to be artfully constructed specifically to allow a linear analysis. When it is possible to apply this algorithm, it will be with serious modifications: multiple layers of linear regression, kludges to account for uncertainty in both the x and y direction, non-linear transformations, etc.

### *1.0.1 Method of maximum likelihood*

Instead of a highly uncomfortable apparatus composed of nested linear regression problems, many of these elaborations can be elegantly treated using the method of maximum likelihood [43, 5]. Maximum likelihood is a clear, intuitive recipe which explains not only least-squares regression but also a host of other statistical procedures. It is one of those gems, like

the Fourier transform and diagonalization, which is both elementary and deep; simple and powerful. It is our first hint that it might be possible to replace increasingly ornate linear regression with a unified, solid foundation [120].

There are only two basic ingredients in Maximum Likelihood (ML): data $x^N = \{x_1 \dots x_N\}$ and a statistical model for the data $q(\cdot|\theta)$, which is specified by a set of parameters $\theta$. For instance, in an experiment designed to measure $\hbar$: the data $x^N$ are the observed detections in a quantum interference experiment, the parameters $\theta$ would include descriptors such as the decoherence time, the mass of the particles, detector error rates and the quantity of interest $\hbar$. The MLE would then give a best estimate of the quantity of interest $\hbar$, as well as estimates of the parameters for the stochastic processes responsible for the data.

ML can most immediately be understood in terms of the Bayes rule, [1].

$$\pi(\theta|x_N) = \frac{q(x^N|\theta)\pi(\theta)}{q(x^N)} \tag{1.1}$$

In the special case where the prior is constant, i.e. all values of the parameters are equally likely (do we really expect this to be the case?), the most likely value of the parameters is the parameters that maximize $q(x^N|\theta)$ (or equivalently, but more numerically stable $\log q(x^N|\theta)$) the so called Maximum Likelihood Estimators (MLEs)

$$\hat{\theta}_x = \arg\max_\theta \log q(x^N|\theta). \tag{1.2}$$

This estimation procedure recovers linear and non-linear least-squares fitting as the special case of normally distributed errors. In that case, $\theta$ define the parameters of a line and the log likelihood is given by.

$$\log q(x^N|\theta) = -\sum_i \frac{(y_i - \mu(x_i|\theta))^2}{2\sigma} - \frac{N}{2}\log 2\pi \tag{1.3}$$

so that maximizing the likelihood amounts to least-squares minimization. Maximum likelihood is reparametrization invariant, which makes it appealing from a physical and practi-

[1] More generally, the method of maximum likelihood has a theoretical justification which stands outside of a Bayesian analysis. For instance, it can be proven to asymptotically attain equality in the Cramer-Rao bound on the variance of an unbiased estimator.

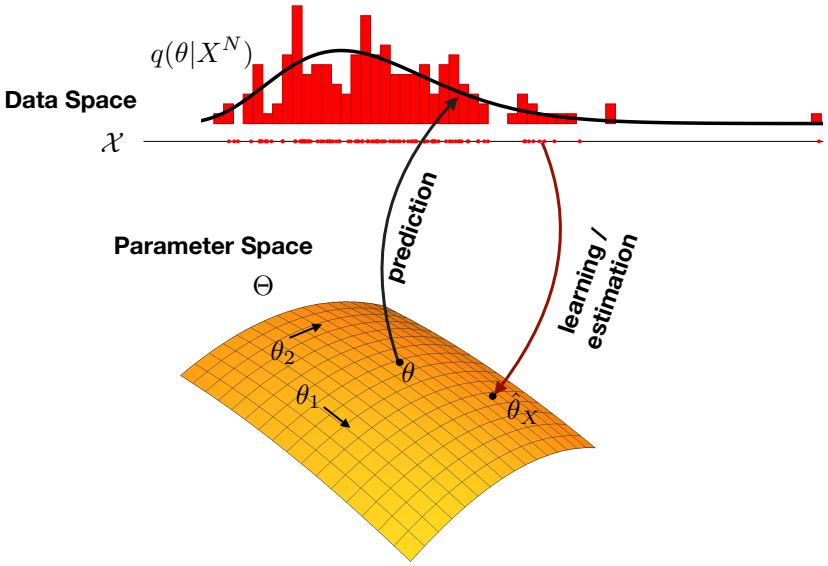


Figure 1.1: **Schematic of a statistical model** A statistical model consists of a space of observed data, a space of parameters, and maps between the two. Candidate distributions describing the data are identified with points in a model or parameter space $\Theta$, the coordinates of which, $\theta_1, \theta_2, \ldots, \theta_K$, are the parameters. Data sets can be mapped to parameters through a learning or estimation method $\hat{\theta}(X^N) = \hat{\theta}_X$, such as the method of maximum likelihood. More generally the estimation process results in a *distribution of parameter values* such as in the Bayes procedure. We will consider both point estimates and distribution estimates.

cal perspective. Everything from deep sequencing, to neural networks which return point-estimates, can be viewed as likelihood optimization on some underlying statistical model.

The picture of a statistical model which emerges, shown in Fig. 1.1, consists of data and parameter spaces, and maps between. In particular, to apply the maximum likelihood procedure, it is required that one provide a full statistical description of the data so that data can be simulated from the model. Not all statistical models are *generative*, that is, not all useful models are able to predict or simulate future data[2]. Nonetheless, this class of statistical models is very broad. We will assume that for models under consideration, we have access to the complete structure, data $X^N$, parameters $\theta$, and likelihood $q(X^N|\theta)$, to use as input for all the algorithms we propose.

### *1.0.2 A unified framework for statistics?*

Maximum likelihood brings together linear regression, parameter estimation and even primitive error analysis under a single principle. Can we arrange other statistical ideas in an even larger unified statistical machinery? Can we determine the fundamental quantities from which this hypothetical unified statistical machinery should be built?

In physics, given the dynamic equations of a physical system, and some boundary conditions, a solution is perfectly defined. The fully specified problem may be too computationally intractable to solve, it may be unnecessary or wasteful to solve it exactly, but the problem and the solution are well-specified. In this spirit, our aim to discover a physics of learning: a set of criteria which fully specifies a data analysis problem and a "solution" carrying the essential content of the observed data relative to the problem as specified. A full solution will be computationally intractable for all but the simplest problems, but we should aim to approximate this solution to whatever accuracy is required.

If we can understand existing methods as approximations to this to-be-discovered physics of learning, we can place them in relation to each other, and determine which is appropriate

[2]e.g. classifiers and most conventional neural nets are not generative.

given the parameters of data analysis context. This synthesis would alleviate controversy in statistics and point the way toward new methods. The scope of the maximum likelihood procedure is proof that such synthesis is, to some extent, possible. However, maximum likelihood alone is not enough.

Maximum likelihood, powerful and elegant a machine though it is, proves insufficient in at least two important contexts: Parameter uncertainty and model selection. These failures indicate a need to augment the maximum likelihood apparatus: An account of parameter uncertainty requires significantly more machinery in the form of a frequentist hypothesis with corresponding confidence intervals or, increasingly commonly, a fully Bayesian analysis[3]. A coherent account of model selection requires that we replace the likelihood with a predictive performance optimization principle. We will see that the Bayesian approach for uncertainty analysis and the predictive approach for model selection are fundamentally incompatible [150, 11, 133, 26, 1]. A unifying framework must first be able to explain and resolve this incompatibility.

### *1.1 Model selection*

Suppose we have several models $p_1(X^N|\theta) \dots p_m(X^N|\theta)$ We might reason that the *type* of statistical model is just another parameter to be optimized. Mathematically shouldn't we write:

$$q_i(X^N|\theta) = q(X^N|\theta, i)? \tag{1.4}$$

Then, based on the principle of maximum likelihood, we should maximize jointly over $\theta$ and $i$. This is particularly problematic if $i$ indexes the number of parameters (dimensions) in the statistical model, such as occurs in an analysis problem of interest to our lab: the Change-Point problem.

---

[3]The frequentist approach to parameter uncertainty, the confidence interval is a useful alternate formulation of parameter uncertainty, but it becomes increasingly cumbersome as the complexity of the model increases. We will not discuss it here.

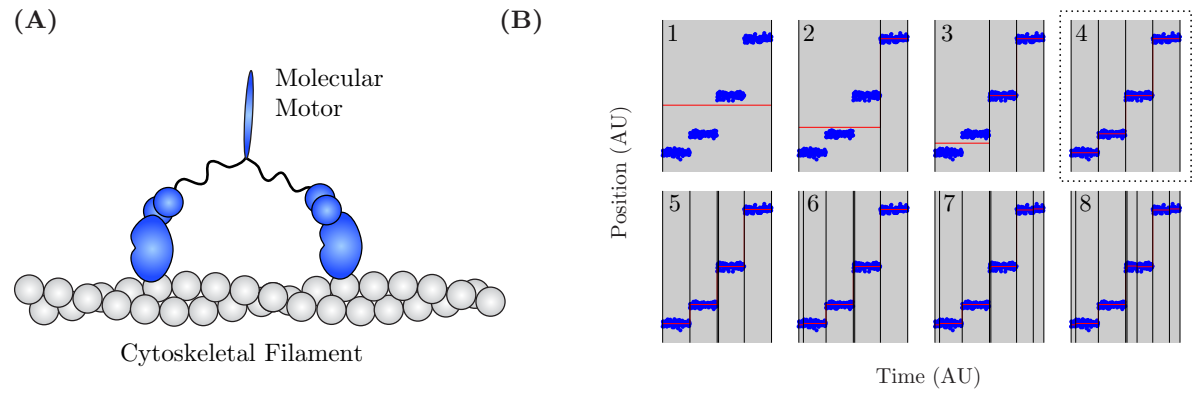


Figure 1.2: **Panel A: Biophysical system which exhibits change points.** One potential application of Change-Point Analysis is to the characterization molecular-motor stepping along a cytoskeletal filament. **Panel B: Schematic of Change-Point Analysis.** A change-point model of motor stepping is shown for a series of position states. The blue dots represent measurements of motor position, corrupted by noise. The red line represents the change-point model for the true motor position. Each frame shows the optimal fit for $n = 1...8$ position states. From the figure, it is intuitively clear that $n = 4$ is the correct number of position states. Models with additional states improve the fit to the observed data but would result in information loss for an independent set of measurements of the same motor positions.

### *1.1.1 The change point problem*

The change-point problem, the problem of determining the true state of a system that transitions between discrete states and whose observables are corrupted by noise is a canonical problem in statistics with a long history [99]. The approach we will discuss is called Change-Point Analysis and was first proposed by E. S. Page in the mid 1950s [115, 116]. Since its inception, Change-Point Analysis has been used in a great number of contexts and is regularly re-invented in fields ranging from geology to biophysics [32, 99, 100].

Change point analysis is applied to a signal consisting of a series of observations generated

by a stochastic process:

$$X^N \equiv (X_1, X_2, ..., X_N) \sim p(\cdot), \tag{1.5}$$

where the observation index is often but not exclusively temporal. We define a model for the signal corresponding to a system transitioning between a set of discrete states. For example, a molecular motor transitions between position states as it steps along the cytoskelletal fillament. Each state generates a distinct distribution of measurements as illustrated in Fig 1.2. We define the discrete time index corresponding to the start of the $I$th state $i_I$. This index is called a *change point*. The model parameters describing the signal distribution in the $I$th interval are $\theta_I$. Together these two sets of parameters, $i_I$ and $\theta_I$, parameterize the model. The model parameterization for the signal (including multiple states) can then be written explicitly:

$$\theta^n \;=\; \begin{pmatrix} 1 & i_2 & \dots & i_n \\ \theta_1 & \theta_2 & \dots & \theta_n \end{pmatrix}, \tag{1.6}$$

where $n$ is the number of states or change points. The problem of change-point analysis is then to determine the number and location of change points with the parameter values describing the underlying states.

The central difficulty in change-point analysis is the problem of the bias–variance tradeoff in selecting the dimension of the model: the determination of the number of states (or change points $n$). Adding states always improves the fit to the data. The maximum likelihood estimator over $\theta^n, n$ results in $\hat{n}_x = N$, i.e. where each $\mu_i$ only persists for a single observation and is equal to $X_i$. This model doesn't provide any new output, it simply replicates the initial data. But from the perspective of the maximum likelihood principle, it is the best of all possible models.

### *1.1.2 Model selection and predictivity.*

The paradoxically optimal likelihood of this useless model is due to the phenomenon of overfitting: it highly tuned to the "specific" features of the data set and not to the features of

data sets in general. New data points sampled from one of these states will not conform well to the fitted distribution: the model is therefore a failure from the perspective of prediction. If we select models based on likelihood alone then we will choose the largest (i.e. most parameterized) models: those models with the most overfitting. The fact that the vastly overfit model should be deprecated from a predictive rather than likelihood standpoint hints that predictive performance rather than likelihood should be our metric optimized by model selection. The predictive performance encodes a realization of Occam's Razor: "Prediction favors parsimony". Smaller models give us better predictive power.

A predictive solution is to penalize the likelihood of larger models: those for which parameter space is of higher dimension. This penalized likelihood is called an Information criterion (IC). The most important information criterion for our purposes is the Akaike Information Criterion (AIC),

$$AIC(X^N) \equiv \log q(X^N|\hat{\theta}_X) - \dim(\Theta) \tag{1.7}$$

If this quantity is optimized than we will find that the selected model dimension will not be the largest possible model, but rather one which balances goodness of fit $\log q(X^n|\theta_i)$, with the predictive loss due to the complexity of the model $\dim(\Theta)$. Unfortunately AIC isn't applicable in the change-point model where the data structured. In Chapter 2 , we will discuss a novel procedure that is applicable.

### *1.2 Uncertainty on the parameters*

Once we find $\hat{\theta}_x$ we would like to report uncertainty. A Laplace approximation which can be justified using the Cramer-Rao theorem (the central limit theorem for MLE's) is to look at the curvature of the log-likelihood:

$$J(\theta) = \frac{\partial}{\partial\theta} \otimes \frac{\partial}{\partial\theta} \log q(x^N|\theta) \tag{1.8}$$

This curvature tells us how well constrained our parameter estimates are: the larger the

curvature at the maximum, the smaller the uncertainty. Our uncertainty in the parameters can be represented as a Gaussian

$$\theta \sim \mathcal{N}\left\{\hat{\theta}_x, J^{-1}(\hat{\theta}_x)\right\}. \tag{1.9}$$

But our uncertainty may not be well represented by a Gaussian. Our variables may be constrained to have positive support, or it may have many local maxima in the likelihood. This later scenario is particularly common in a machine-learning context. Worse, the assumed distribution $\mathcal{N}\left\{\hat{\theta}_x, J^{-1}(\hat{\theta}_x)\right\}$ becomes inconsistent under a non-linear coordinate transformations[4].

Going back to the Bayes rule, it is obvious that we could get a distribution on our parameters directly, if we had access to a prior $\pi(\theta)$.

$$\pi(\theta|x_N) = \frac{q(x^N|\theta)\pi(\theta)}{q(x^N)} \tag{1.10}$$

Unfortunately, this makes the definition of parameter uncertainty dependent on the existence of a prior $\pi(\theta)$. Elicitation of an objective prior $\pi(\theta)$ is one of the oldest problems in statistics. A significant contribution of this thesis is a novel approach to the construction of a general objective prior.

### *1.3 AIC vs. BIC?*

The big picture of statistics that emerges is a discordant. While the maximum likelihood procedure has many virtues, it can only be part of some larger theory. When dealing with

[4]To see this, the density on transformed parameter space is $\pi(\theta|x^N)d\theta \to \pi(\phi(\theta)|x^N)\left|\frac{\partial\theta(\phi)}{\partial\phi}\right|$ which would imply *any* distribution depending on the choice of $\phi$ (e.g. uniform if $\phi$ is chosen to be the cumulative-density function (CDF) of the distribution of $\theta$). At the same time, in the new coordinate system, $J^{-1}$ simply undergoes a linear transformation, $J^{-1}(\hat{\theta}_x) \to J'^{-1}(\phi(\hat{\theta}_x))$ and our original procedure applied to the new coordinate system implies $\phi \sim \mathcal{N}\left\{\phi(\hat{\theta}_x), J'^{-1}(\phi(\hat{\theta}_x))\right\}$ i.e. *still* normally distributed. In short treating the $\mathcal{N}\left\{\hat{\theta}_x, J^{-1}(\hat{\theta}_x)\right\}$ as a real distribution for the parameters implies a contradiction when we perform a nonlinear transformation on parameter space.

parameter uncertainty, the most natural approach is to move to a fully Bayesian treatment with the elicitation of a prior. In the case of model selection, we would augment ML with a prediction-optimization principle, as exemplified by AIC.

There are questions on reproducibility and the predictive validity with wide implications. Debate on these issues is continuous and lively in biology, psychology, medicine, and economics [73, 13, 50, 107, 138, 39, 34, 119]. In our field of biophysics, Bayesian methods for model selection [18, 74, 125, 64] are extremely difficult to interpret as standard methods inadvertently create a systematic bias for models of smaller dimension. The bias present in "objective" Bayesian methods inferring model structure is not widely understood, and causes researchers to overstate the evidence for smaller, more compelling, more easily publishable models.

These two approaches are not mutually compatible [150]. The Bayesian approach has a different realization of Occam's razor. The mechanism for this realization of Occam's razor lies in the normalization of the prior. If the parameter space of a model $\Theta^m$ has a volume $\mathrm{vol}(\Theta^{\mathrm{m}})$, the posterior probability will be roughly proportional to $\propto 1/\mathrm{vol}(\Theta^{\mathrm{m}})$, the normalization of the prior. The slogan for this Occam razor can be stated succinctly "Probability favors parsimony." Smaller models tend to have greater posterior probability from a Bayesian standpoint because they didn't have to stretch their *a priori* probabilities over as large a volume. Roughly speaking the Bayesian formalism penalizes models based on parameter space volume while the predictive formulation penalizes models based on dimension.

An approximation for the log-Bayesian posterior weight is given by the penalized likelihood:

$$\mathrm{BIC}(X^N|\theta, i) \equiv \log q(X^n|\theta_i) - \frac{\dim \Theta^i}{2} \log(N) \tag{1.11}$$

Because this penalized likelihood has the same form as an information criterion, this is referred to as the Bayesian Information Criterion (BIC). Clearly these two forms of Occam's razor–one derived from prediction the other derived from posterior probability—are at odds.

*We found that predictive (AIC-type) approach is appropriate for most data analysis prob-*

*lems. The predictive approach could be reconciled with a Bayesian approach to parameter uncertainty using constructions taken directly from thermodynamics. This reconciliation simultaneously solves the problem of the undetermined prior.*

### *1.4 Layout and contributions of the thesis*

In Chapter 2 we discuss how our investigations into statistical foundations began in our attempt to determine the correct model penalty in an information criterion for the change-point problem [87] and how it leads to the development of an extension of AIC. We develop this novel information criterion from a more general theoretical standpoint in Chapter 3 [85]. To our surprise, we determined that BIC may "by luck" be numerically equally to the correct penalty under AIC-like arguments when the model has non-trivial geometry! This lead us to hypothesize that researchers who have found that BIC gives excellent *predictive* results may have been in a situation where the a BIC-like penalty follows from a predictive Occam razor in the presence of non-trivial model structure. Does this mean that prediction would make a viable unifying principle after all? Where does this leave Bayesian methods?

The difference between AIC and BIC is closely related to the disagreement between frequentist and Bayesian statistics in model selection: the Lindley paradox. We explore the relationship [89]between the predictive and Bayesian frameworks from the perspective of the Lindley paradox in Chapter 4. This exploration shows that the canonical (BIC-like) method typically leads to worse performance than AIC and related methods (including our novel information criterion developed in Chapter 3).

Our developments in Chapter 4 rely heavily on the use of a derivative with respect to sample-size which mirrors a similar operation in thermodynamics. In Chapter 5 we expand on this newly discovered connection between statistical mechanics and thermodynamics, finding that it can be used to merge the predictive and Bayesian approaches [90]. One important result is a generalized principle of indifference. The generalized principle of indifference automatically generates a coherent statistical framework where predictive model selection is integrated with an objective Bayesian analysis of uncertainty. Another important result

from this analysis is a novel metric for model complexity, the Learning Capacity which offers significant insight into the nature of learning and the fundamental properties of a general statistical model. Finally in Chapter 6 we discuss the overall arc of this work, its expected impact and directions for future study.

# Chapter 2

# CHANGE-POINT INFORMATION CRITERION SUPPLEMENT

This thesis was not the result of some preexisting expectation that physics had a lot to say about statistics. We simply wanted to solve a statistics problem of interest to the biophysics experiments going on in the lab. In particular, we wanted to figure out how to solve the change-point problem, that is, how to get the computer to terminate at the correct number or change-points in an intensity trace for a fluorophore bleaching process. The results were to be used for Sarah Mangiameli's stoichiometry experiments on the replisome [104]. We believed that AIC, (or perhaps BIC? ) could be used to select the number of change points. In fact, neither AIC nor BIC worked! It was obvious that AIC found many spurious change-points. It was just as obvious that BIC missed resolvable change-points. Why was this happening, and what is the right IC (information criterion)? This chapter follows [87].

The change-point model is not *regular*; there exist singular points in parameter space for which the information matrix is not positive definite. As with non-analytic points in complex analysis, the Taylor expansion of the information poorly approximates its behavior in neighborhood of these singular points. The details of Akaike's derivation depend on the validity of this Taylor expansion, so AIC is not applicable to the change point problem [143]. Further complicating matters, the data in a change-point problem is potentially structured and therefore is not necessarily independent and identically distributed for all observations $X^N$. These properties make the application of tools like naïve cross validation and Watanabe's WAIC more difficult to apply [51].

**Proposed Approach.** Our approach can be seen as a direct extension of AIC. In regular models, the expected information is quadratic about its minimum in parameter space.

Realizations of the data generate maximum-likelihood estimators that fluctuate about this optimal value, in analogy with the thermal fluctuations of a particle confined to a harmonic potential. These fluctuations decrease the predictivity of models constructed using maximum likelihood procedure. AIC is derived through the consideration of these harmonic fluctuations. If a candidate change point $I$ is supported by the data, then the continuous parameters $\theta_I$ are subject to harmonic confinement and their contribution to the model complexity is equal to their dimensionality, as Akaike predicted, while the change point $i_I$, as a highly constrained discrete variable, does not contribute to the complexity at all.

If a candidate change point is unsupported, the maximum likelihood change point is not constrained; it can be realized anywhere over a candidate interval. We have recently proposed a Frequentist Information Criterion (QIC) applicable even in the context of singular models. Using QIC we find that the information as a function of change-point location can then be approximated with the squared norm of a Brownian bridge, and that expected predictive loss can be estimated with a modified measure of the model complexity derived from this description. Consideration of these two distinct behaviors gives a piecewise information criterion which does not depend on the detailed form of the model for the individual states but only on the number of model parameters, in close analogy with AIC. Therefore we expect this result to be widely applicable anywhere the change-point algorithm is applied.

**Relation to Frequentist Methods.** Frequentist statistical tests have been defined for a number of canonical change-point problems. It is interesting to examine the relation between this approach and our newly-derived information-based approach. We find the approaches are fundamentally related. The information-based approach can be understood to provide a predictively-optimal confidence level for a generalized ratio test. The Bayesian Information Criterion (BIC) has also been used in the context of Change-Point Analysis. We find very significant differences between our results and the BIC complexity that suggest that BIC is not suitable for application to change-point analysis.

## *2.1 Preliminaries*

The essential notation is summarized in Table 2.1. We shall represent the probability distribution for a change-point model $\Theta^n$ as:

$$q(X^N|\Theta^n) \tag{2.1}$$

**Information and cross entropy.** The information for signal $X^N$ given model $\Theta$ is:

$$h(X^N|\Theta^n) \equiv -\log q(X^N|\Theta^n), \tag{2.2}$$

and the cross entropy for the signal (average information) is:

$$H(\Theta^n) \equiv \mathop{\mathbb{E}_X}_{p} h(X^N|\Theta^n), \tag{2.3}$$

where the expectation over the signal $X^N$ is understood to be taken over the true distribution $p$.

The state parameters, $\theta_I$, and the change points, $i_I$, are fundamentally different parameters. We shall assume that the state model is regular: i.e. the parameters $\theta_I$ have non-zero Fisher information. By contrast, the change-point indices $i_I$ are discrete and typically non-harmonic parameters. For instance, consider a true model $p = q$ where $\theta_1 = \theta_2$. In this scenario the cross entropy will be independent of $i_2$ as long as $i_2 \in (i_1, i_3)$. The Fisher information corresponding to $i_2$ is therefore zero. These properties have important consequences for model selection.

**Determination of model parameters.** Fitting the change-point model is performed in two coupled steps. Given a set of change-point indices $\boldsymbol{i}^n \equiv (i_1, ..., i_n)$, we hold the change points fixed and find the maximum likelihood estimators (MLE) of the state parameters $\boldsymbol{\theta}^n \equiv (\theta_1, ..., \theta_n)$. These are defined:

$$\hat{\boldsymbol{\theta}}^n_X = \arg\min_{\boldsymbol{\theta}^n} h(X^N|\Theta^n). \tag{2.4}$$

| | **Data and observations** |
|---|---|
| $X^N,\ X^{[i,j]}$ | All $N$ observations / observations on interval $[i, j]$ |
| $p(\cdot)$ | True (unknown) distribution from which the data $X^N$ was generated |
| $\mathbb{E}_{\substack{X \\ q}}$ | Expectation over $X$ taken with respect to $q$ |
| | **Model parameterization** |
| $i_I$ | Change-point or first temporal index of state $I$ |
| $\theta_I$ | Parameters describing state $I$ |
| $\hat{\theta}_X$ | The maximum likelihood estimator (MLE) of $\theta$ |
| $\Theta^n$ | Vector of $\theta_I$ and $i_I$ describing $n$ states |
| $\theta_0$ | True parameter values |
| | **Measures of information and entropy** |
| $h(X^N \vert \Theta^n)$ | Information for $X^N$ (the negative of the log-likelihood) |
| $h_i$ | Information for the $i$th observation |
| $H^N(\Theta^n)$ | $N$-observation cross entropy (expected information) |
| $\mathscr{K}(n)$ | Complexity of a model with $n$ states |
| IC | Information Criterion or unbiased estimator of the cross entropy. |
| $\mathscr{k}(n)$ | Nesting complexity: $\mathscr{K}(n) - \mathscr{K}(n-1)$ |
| | **Derivatives of information** |
| $\mathbf{x}_i$ | Parameter gradient of information $h_i$ |
| $\mathbf{X}$ | Sum of the $\mathbf{x}_i$ (the negative of the score function) |
| $\mathbf{I}$ | Fisher information (Hessian matrix of the information $h_i$) |

Table 2.1: **Summary of essential notation:** The table contains a brief summary of the notation used in the paper.

The determination of the change-point indices $\boldsymbol{i}^n$ is a nontrivial problem since not only are the change-point indices unknown, but the number of transitions ($n$) is also unknown.

**Binary Segmentation Algorithm.** To determine the change-point indices, we will use a binary-segmentation algorithm that has been the subject of extensive study (e.g see the references in [32]). In the global algorithm, we initialize the algorithm with a single change point $i_1 = 1$. The data is sequentially divided into partitions by binary segmentation. Every segmentation is *greedy*: i.e. we choose the change point on the interval $(1, N)$ that minimizes the information in that given step, without any guarantee that this is the optimum choice over multiple segmentations. The family of models generated by successive rounds of segmentation are said to be *nested* since successive changes points are added without altering the time indices of existing change points. Therefore, the previous model is always a special case of the new model. In each step, after the optimum index for segmentation is identified, we statistically test the change in information (due to segmentation) to determine whether the new states are statistically supported. The $n$ change-points determined by binary segmentation with their MLE state parameters compose $\hat{\Theta}^n$. We later distinguish between local and global segmentation: the local binary-segmentation algorithm differs from the global algorithm only in that we consider binary segmentation of each partition of the data independently. The algorithms are described explicitly in the supplement.

**Information-based model selection.** The model that minimizes the cross entropy (Eqn. 2.3) is the most predictive model. Unfortunately, the cross entropy cannot be computed: the expectation cannot be taken with respect to the true but unknown probability distribution $p$ in Eqn. 2.3. The natural estimator of the cross entropy is the information (Eqn. 2.2), but this estimator is biased from below: Due to over-fitting, added model parameters always reduce the information, even as the predictivity of the model is reduced by the addition of superfluous parameters. To accurately estimate predictive performance, we construct an unbiased estimator of the cross entropy which we call the *information criterion*:

$$\mathrm{IC}(X^N, n) \equiv h(X^N|\hat{\Theta}^n_X) + \mathscr{K}(n), \tag{2.5}$$

where $\mathscr{K}$ is the complexity of the model which is defined as the bias in the information as an estimator of cross-entropy:

$$\mathscr{K}(n) \equiv \mathop{\mathbb{E}_{X,Y}}_{p} \left\{ h(Y^N|\hat{\Theta}_X^n) - h(X^N|\hat{\Theta}_X^n) \right\}, \tag{2.6}$$

where the expectations are taken with respect to the true distribution $p$ and $X^N$ and $Y^N$ are independent signals. Complexity is a measure of the flexibility of a family of models in fitting the observed data. A more complex model can be tuned to fit more features in the data, resulting in lower information than models with smaller complexity. However, the more complex model will be more prone to i.) artificially decreasing the information relative to its optimally predictive parameter values, and ii.) reducing the predictivity of the model by shifting the probability mass to accord with features not reproducible in different realizations of the data. The more flexible model is expected to be more predictive only if the decrease in observed information is greater than the expected magnitude of these detrimental effects as measured by the complexity.

For a regular model in the asymptotic limit, the complexity is equal to the number of model parameters and the information criterion is equal to AIC. In the context of singular models, a more generally applicable approach must be used to approximate the complexity.

**Frequentist Information Criterion.** The Frequentist Information Criterion (QIC) uses a more general approximation to estimate the model complexity. Since the true distribution $p$ is unknown, we make a frequentist approximation, computing the complexity for the model $\Theta^n$ as a function of the true parameterization:

$$\mathscr{K}_{\mathrm{QIC}}(\Theta^n, n) \equiv \mathop{\mathbb{E}_{X,Y}}_{\Theta^n)} \left\{ h(Y^N|\hat{\Theta}_X^n) - h(X^N|\hat{\Theta}_X^n \right\}, \tag{2.7}$$

and the corresponding information criterion is defined:

$$\mathrm{QIC}(X^N, n) \equiv h(X^N|\hat{\Theta}_X^n) + \mathscr{K}_{\mathrm{QIC}}(\hat{\Theta}_X^n, n), \tag{2.8}$$

where the complexity is evaluated at the MLE parameters $\hat{\Theta}_X^n$. The model that minimizes QIC has the smallest expected cross entropy.

**Approximating the QIC complexity.** The direct computation of the QIC complexity (Eqn. 2.7) appears daunting, but a tractable approximation allows the complexity to be estimated. The complexity difference between the models is:

$$\mathbb{k}(n) \equiv \mathcal{K}_{\mathrm{QIC}}(n) - \mathcal{K}_{\mathrm{QIC}}(n-1), \tag{2.9}$$

which is called the nesting complexity. An approximate piecewise expression can be computed as follows. Let the observed change in the MLE information for the addition of the $n$th change point be

$$\Delta h_n \equiv h(X^N|\hat{\Theta}_X^n) - h(X^N|\hat{\Theta}_X^{n-1}), \tag{2.10}$$

Consider two limiting cases: When the new parameters are identifiable, let the nesting complexity be given by $\mathbb{k}_+$ whereas when the new parameters are unidentifiable, let the nesting complexity be given by $\mathbb{k}_-$. When the new parameters are identifiable, the model is essentially regular therefore:

$$\mathbb{k}_+ = \mathcal{d}, \tag{2.11}$$

where $\mathcal{d}$ is the number of harmonic[1] parameters added to the model in the nesting procedure, as predicted by AIC.

To compute $\mathbb{k}_-$, we assume the unnested model is the true model and compute the complexity difference in Eqn. 2.9. We then apply a piecewise approximation for evaluating the nesting complexity [85]:

$$\mathbb{k}(n) \approx \begin{cases} \mathbb{k}_-(n), & -\Delta h_n < \mathbb{k}_-(n) \\ \mathbb{k}_+(n), & \text{otherwise} \end{cases}. \tag{2.12}$$

Since the nesting complexity represents complexity differences, the complexity can be summed:

$$\mathcal{K}_{\mathrm{QIC}}(n) \equiv \sum_{j=1}^{n} \mathbb{k}(j), \tag{2.13}$$

where the first term in the series, $\mathbb{k}(1)$ is computed using the AIC expression for the complexity. An exact analytic description of the complexity remains an open question.

[1] Harmonic parameters are parameter with sufficiently large Fisher information that they are not unidentifiable.

### *2.2 Information criterion for change-point analysis*

**Complexity of a state model.** As a first step towards computing the complexity for the change-point algorithm, we will compute the complexity for a signal with only a single state. It will be useful to break the information into the information per observation. Assuming the process is Markovian, the information associated with the $i$th observation is:

$$h_i(X^N|\theta) \equiv -\log q(X_i|X_{i-1};\theta). \tag{2.14}$$

For a stationary process, the average information per observation is constant $\overline{h} \equiv \mathbb{E}\ h$. The fluctuation in the information $\delta h_i \equiv h_i - \overline{h}$ has the property that it is independent for each observations:

$$\mathbb{E}\ \delta h_i\, \delta h_j = C_0 \delta_{ij}, \tag{2.15}$$

where $C_0$ is a constant and $\delta_{ij}$ is the Kronecker delta, due to the Markovian property. In close analogy to the derivation of AIC, we will Taylor expand the information in terms of the model parameterization $\theta$ around the true parameterization $\theta_0$. We make the following standard definitions:

$$\begin{aligned}
\delta\theta &\equiv \theta - \theta_0, && (2.16)\\
\hat{\boldsymbol{I}}_i &\equiv \nabla_\theta \nabla_\theta^T h_i(X^N|\theta_0), && (2.17)\\
\boldsymbol{I} &\equiv \mathbb{E}_X \nabla_\theta \nabla_\theta^T h_i(X^N|\theta_0), && (2.18)\\
\boldsymbol{x}_i &\equiv \nabla_\theta h_i(X^N|\theta_0), && (2.19)\\
\boldsymbol{X} &\equiv \sum_i \boldsymbol{x}_i. && (2.20)
\end{aligned}$$

where $\delta\theta$ is the perturbation in the parameters, $\boldsymbol{I}$ and $\hat{\boldsymbol{I}}_i$ are the Fisher information and its estimator respectively. We make the cannonical approximation that the estimator is well approximated by the true value: $\hat{\boldsymbol{I}}_i \to \boldsymbol{I}$. The subscript $i$ refers to the $i$th observation. Note that since the true parameterization minimizes the information by definition, $\mathbb{E}\ \boldsymbol{x}_i = 0$. Furthermore, Eqn. 2.15 implies that

$$\mathbb{E}\ \boldsymbol{x}_i\, \boldsymbol{x}_j^T = \boldsymbol{I}\delta_{ij} \tag{2.21}$$

where $\boldsymbol{I}$ is the Fisher information. The Taylor expansion of the information can then be written:

$$h(X^N|\theta) = h(X^N|\theta_0) + \delta\theta^T \boldsymbol{X} + \tfrac{1}{2}\delta\theta^T N \boldsymbol{I} \delta\theta + \mathcal{O}(\delta\theta^3), \tag{2.22}$$

to quadratic order in $\delta\theta$.

It is convenient to transform the random variables $\boldsymbol{x}_i$ to a new basis in which the Fisher information is the identity. This is accomplished by the transformation

$$\boldsymbol{x}'_i \equiv \boldsymbol{I}^{-1/2}\boldsymbol{x}_i, \tag{2.23}$$

$$\theta' \equiv \boldsymbol{I}^{1/2}\theta, \tag{2.24}$$

which results in the following expression for the information:

$$h(\theta|X_I) = h(X^N|\theta_0) + \delta\theta'^T \boldsymbol{X}' + \tfrac{1}{2} N \delta\theta'^T \delta\theta' + \mathcal{O}(\delta\theta^3). \tag{2.25}$$

In our rescaled coordinate system, $\boldsymbol{X}'$ can be interpreted as an unbiased random walk of $N$ steps with unit variance in each dimension.

We determine the MLE parameter values:

$$\delta\hat{\theta}'_X = -\tfrac{1}{N}\boldsymbol{X}'. \tag{2.26}$$

To compute the complexity we need the following expectations of the information:

$$\mathbb{E}_{X,Y}\ h(Y^N|\hat{\theta}_X) = \mathbb{E}_{X,Y}\ \left\{ h(Y^N|\theta_0) - \tfrac{1}{N}\boldsymbol{X}'^T\boldsymbol{Y}' + \tfrac{1}{2N}\boldsymbol{X}'^2 + \mathcal{O}(\delta\theta^3) \right\}, \tag{2.27}$$

$$\mathbb{E}_{X}\ h(X^N|\hat{\theta}_X) = \mathbb{E}_{X,Y}\ \left\{ h(X^N|\theta_0) - \tfrac{1}{2N}\boldsymbol{X}'^2 + \mathcal{O}(\delta\theta^3), \right\}. \tag{2.28}$$

Since the signals $X^N$ and $Y^N$ are independent, the second term on the RHS of Eqn. 2.27 is exactly zero. It is straightforward to demonstrate that

$$\mathbb{E}_X \boldsymbol{X}'^2_I = N\mathcal{d}, \tag{2.29}$$

where $\mathcal{d}$ is the dimension of the parameter $\theta$, which has an intuitive interpretation as the mean squared displacement ($\boldsymbol{X}'^2$) of a unbiased random walk of $N$ steps in $\mathcal{d}$ dimensions. The complexity is therefore:

$$\mathcal{K} \equiv \mathbb{E}_{X,Y}\ \left\{ h(Y^N|\hat{\theta}_X) - h(X^N|\hat{\theta}_X) \right\} = \mathcal{d}. \tag{2.30}$$

which is the AIC complexity.

This derivation of the AIC complexity through an expectation of a random walk in the score function $\boldsymbol{X}$ can now be extended to include the effects when the change point is not supported. When $i_I$ is not fixed by the data, it is another a free parameter that can be chosen to maximize the decrease in information. The nesting complexity will then be the *maximum* mean squared displacement of *many* (correlated) random walks.

The first unsupported changepoint in a single state system is the first segmentation. We compute the nesting complexity $\not{k}(2)$ of this first segmentation using Eqn. 2.12. We will therefore generate the observations $X^N$ and $Y^N$ using the unsegmented model $\Theta^1$. Remember that by convention we assign the first change-point index to the first observation $i_1 = 1$. The optimal but fictitious change-point index for binary segmentation is:

$$\hat{i}_2(X) \;=\; \arg\min_{1<i\leq N} \{\, h(X^{[1,i-1]}|\hat{\theta}_{X^{[1,i-1]}}) + h(X^{[i,N]}|\hat{\theta}_{X^{[i,N]}}) \,\}\,, \tag{2.31}$$

where the $X^{[j,k]}$ represent the respective partitions of the signal $X^N$ made by the change point $i$. (Note that in the case of an autoregressive process, it is possible to write overlapping partitions to account for the system memory.) The MLE model for two states is defined:

$$\hat{\Theta}^2_X \;\equiv\; \begin{pmatrix} 1 & \hat{i}_2 \\ \hat{\theta}_{X^{[1,\hat{i}_2-1]}} & \hat{\theta}_{X^{[\hat{i}_2,N]}} \end{pmatrix}. \tag{2.32}$$

To compute the nesting complexity, we compute the difference in the information between the two-state and one-state MLE models:

$$\begin{aligned} h(X^N|\hat{\Theta}^2_X) - h(X^N|\hat{\Theta}^1_X) \;=\; \min_{1<i\leq N} \{\, & \cancel{h(X^{[1,i-1]}|\theta_0)} + \cancel{h(X^{[i,N]}|\theta_0)} - \cancel{h(X^{[1,N]}|\theta_0)} \\ & - \tfrac{1}{2(i-1)}\boldsymbol{X}'^2_{[1,i-1]} - \tfrac{1}{2(N+1-i)}\boldsymbol{X}'^2_{[i,N]} + \tfrac{1}{2N}\boldsymbol{X}'^2_{[1,N]} \,\}\,, \end{aligned} \tag{2.33}$$

where $\boldsymbol{X}'_{[i,j]}$ are the $\boldsymbol{X}'$ computed in the two partitions of the data. The terms that are zeroth order in the perturbation cancel since the model is nested. (This equation is analogous to Eqn. 2.28.) It is straightforward to compute the analogous expression for information

difference for signal $Y^N$. The nesting penalty can then be written:

$$\begin{aligned} \cancel{k}_-(2) &\equiv \mathop{\mathbb{E}_{X,Y}}_{q(\cdot|\Theta_0^1)} \left\{ h(Y^N|\hat{\Theta}_X^2) - h(X^N|\hat{\Theta}_X^2) - h(Y^N|\hat{\Theta}_X^1) + h(X^N|\hat{\Theta}_X^1) \right\} \qquad (2.34) \\ &= \mathop{\mathbb{E}_{X}}_{q(\cdot|\Theta_0^1)} \max_{1<i\leq N} \left\{ \tfrac{1}{i-1} \boldsymbol{X'}^2_{[1,i-1]} + \tfrac{1}{N+1-i} \boldsymbol{X'}^2_{[i,N]} - \tfrac{1}{N} \boldsymbol{X'}^2_{[1,N]} \right\}, \qquad (2.35) \end{aligned}$$

where the cross terms between signals $X^N$ and $Y^N$ are zero since the signals are independent. It is now convenient to introduce a $\cancel{d}$-dimensional discrete Brownian bridge:

$$\boldsymbol{B'}_j \equiv \boldsymbol{X'}_{[1,j]} - \tfrac{j}{N} \boldsymbol{X'}_{[1,N]}, \qquad (2.36)$$

by using the well known relation between Brownian walks and bridges [122]. The Brownian bridge has the property that $\boldsymbol{B'}_0 = \boldsymbol{B'}_N = 0$, where each step has unit variance per dimension and mean zero. After some algebra, the nesting complexity can be written:

$$\cancel{k}_-(2) = \mathop{\mathbb{E}_{X}}_{q(\cdot|\Theta_0^1)} \max_{1\leq j<N} \left\{ \tfrac{N}{j(N-j)} \boldsymbol{B'}^2_j \right\}. \qquad (2.37)$$

It is not surprising that the nesting complexity should be well modeled by the square of a Brownian bridge. At the endpoints, the addition of a change point does nothing: it is indistinguishable from a change point already in place. The complexity almost certainly increases: the smaller model is nested in the larger model. These observations are captured in the facts that $\boldsymbol{B'}_0 = \boldsymbol{B'}_N = 0$ and $\boldsymbol{B}^2 \geq 0$ respectively.

The details of the state model will determine the distribution function for the discrete steps in the Brownian bridge, but the Central Limit Theorem implies that the distribution will approach the normal distribution. Therefore, it is convenient to approximate the discrete Brownian bridge $\boldsymbol{B'}_n$ as an idealized Brownian bridge with normally distributed steps:

$$\boldsymbol{B'}_j \to \boldsymbol{B}_j \equiv \sum_{i=1}^{n} \boldsymbol{b}_i, \quad \text{such that} \quad \boldsymbol{B}_N = 0, \qquad (2.38)$$

where the $\boldsymbol{b}_i$ are steps that are normally distributed with variance one per dimension $\cancel{d}$ and mean zero. We now introduce a new random variable $U(N, \cancel{d})$, the $\cancel{d}$-dimensional Change-Point Statistic [68, 69]:

$$U(N, \cancel{d}) \equiv \tfrac{1}{2} \max_{1\leq j<N} \tfrac{N}{j(N-j)} \boldsymbol{B}_j^2, \qquad (2.39)$$

which is a $d$-dimensional generalization of the change-point statistic computed by Hawkins [63]. In terms of the statistic $U$, the nesting penalty is

$$\kappa_-(2) = 2\,\mathbb{E}_U\; U(N, d) = 2\overline{U}(N, d). \tag{2.40}$$

We will discuss the connection to the frequentist likelihood-ratio test shortly.

**Nesting complexity for $n$ states.** The generalization of the analysis to $n$ states is intuitive and straightforward. In the local binary-segmentation algorithm, segmentation is tested locally. The relevant complexity is computed with respect to the length of the $J$th partition. It is convenient to work with the approximation that all partitions are of equal length since the complexity is slowly varying in $N$. We therefore define the local nesting complexity

$$\kappa_{L-}(n) = 2\,\mathbb{E}_U\; U(\tfrac{N}{n-1}, d) = 2\overline{U}(\tfrac{N}{n-1}, d), \tag{2.41}$$

where $\frac{N}{n-1}$ is the mean partition length. The nesting complexity for the binary segmentation of a single state is show in Fig. 2.1 for several different dimensions $d$, and compared with the complexity predicted by AIC and BIC.

In the global binary-segmentation algorithm, the next change-point is chosen by identifying the best position over all intervals. We therefore generalize all our expressions accordingly. We introduce a generalization of the Change-Point Statistic where we replace $N$ with a vector of the lengths of the constituent segment lengths $\boldsymbol{N}^n \equiv (N_1, ...N_n)$. We now define our new change-point statistic:

$$U_G(\boldsymbol{N}^n, d) \equiv \max_{1\le i\le n} U(N_i, d). \tag{2.42}$$

Because it is computationally intensive to compute $U_G$ for all possible segmentations $\boldsymbol{N}^n$, we assume that all the partitions are roughly the same size and consider $n$ segments length $N/(n-1)$. Since the complexity is slowly varying in $N$, this does not in general lead to significant information loss. [2] We therefore introduce another change-point statistic:

$$\kappa_{G-}(n) \;\equiv\; 2\,\mathbb{E}_U \max_{1\le i\le n} \{\, U_i(\tfrac{N}{n-1}, d)\,\} \quad (\approx 2\,\mathbb{E}_U\, U_G(\boldsymbol{N}^n, d)\,) \tag{2.43}$$

[2] We empirically invesigated this equal-interval approximation and it bounds the true complexity from above and is therefore conservative.

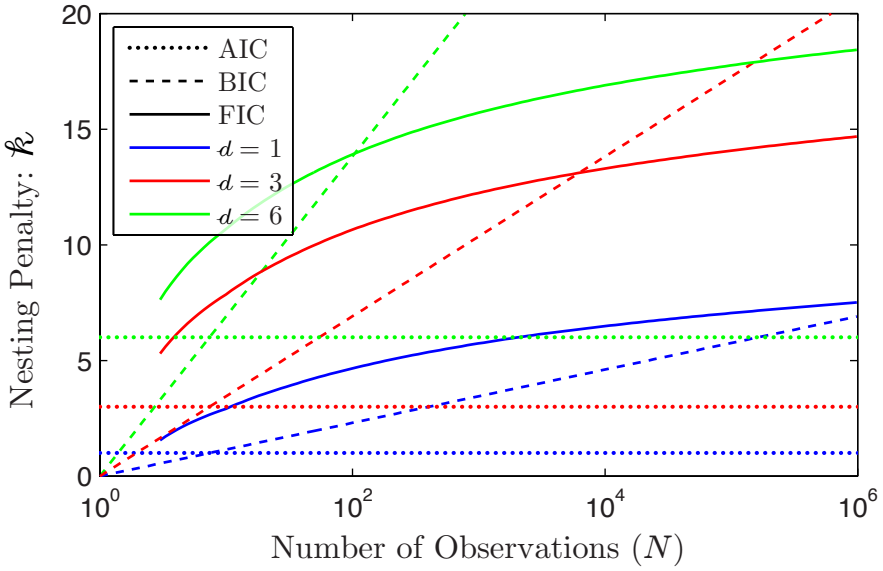


Figure 2.1: **Nesting complexity for AIC, QIC and BIC.** The nesting complexity is plotted for three state dimensions $d = \{1, 3, 6\}$ and $n = 2$. First note that the AIC penalty is much smaller than the other two nesting complexities. BIC is empirically known to produce acceptable results under some circumstances. For sufficiently large samples ($N$), the $\mathscr{k}_{\rm BIC} > \mathscr{k}_{\rm QIC}$, resulting in over penalization and the rejection of states that are supported statistically. This effect is more pronounced for large state dimension $d$ where the crossover occurs for small observation number $N$. $\mathscr{k}_{\rm BIC}$ is too small for a wide range of sample sizes, resulting in over segmentation.

that we will apply in the global binary-segmentation algorithm.

**Asymptotic expressions for the nesting complexity.** It is straightforward to compute the asymptotic dependence of the nesting penalty on the number of observations $N$ [68, 69]:

$$k_{G-}(n) \approx 2\log\log\tfrac{N}{n} + 2\log n + d\log\log\log\tfrac{N}{n} + ..., \tag{2.44}$$

$$k_{L-}(n) \approx 2\log\log\tfrac{N}{n} + d\log\log\log\tfrac{N}{n} + ... \tag{2.45}$$

These expression are slowly converging and in practice, we advocate using Monte Carlo

integration to determine the nesting penalty. If computationally cumbersome, Eqn. 2.44 and 2.45 are useful in placing our approach in relation to existing theory.

Both the local and the global encoding have the same leading-order $2 \log \log N$ dependence that has been advocated by Hannan and Quinn [60], although interestingly not in this context. In contrast, this $2 \log \log N$ dependence is in disagreement with the Bayesian Information Criterion, which has often been applied to change-point analysis. As illustrated by Fig. 2.1, the BIC complexity:

$$\mathscr{K}_{\mathrm{BIC}} = \tfrac{d}{2} \log N, \tag{2.46}$$

can be either too large or too small depending on the number of observations and the dimension of the model. It has long been appreciated that BIC can only be strictly justified in the large-observation-number limit. In this asymptotic limit, the BIC complexity is always larger than the QIC complexity due to the leading order $\log N$ dependence which will tend to lead to under fitting or under segmentation. It is clear from Fig. 2.1 that large ($N > 10^6$) may constitute much larger datasets than are produced in many applications.

**Global versus local complexity.** We proposed two possible parameter encoding algorithms above that give rise two distinct complexities: $k_{L-}$ and $k_{G-}$. Which complexity should be applied in the typical problem? For most applications, we expect the number of states $n$ to be proportional to the number of observations $N$. Doubling the length of the dataset will result in the observation of twice as many change points on average. The application of the local nesting complexity clearly has this desired property since it depends on the ratio of $N/n$. It is this complexity we advocate under most circumstances.

In contrast the global nesting complexity contains an extra contribution to the complexity $2 \log n$. The reason is subtle: In the global binary segmentation algorithm, one picks the best change point among $n$ segments and therefore complexity must reflect this added degree of choice. Consequently a larger feature must be observed to be above the expected background. The use of the global nesting complexity makes a statement of statistical significance against the entire signal, not just against a local region. In the context of discussing the significance

of the observation of a rare state that occurs just once in a dataset, the global nesting complexity is the most natural metric of significance.

**Computing the complexity from the nesting complexity.** To compute the QIC complexity, we sum the nesting complexities using Eqn. 2.13. For datasets with identifiable change points, the QIC complexity is initially identical to AIC:

$$\mathcal{K}_{\mathrm{QIC}}(n) = n\mathcal{d}, \tag{2.47}$$

until the change in the information on nesting $\Delta h < \mathcal{k}_{-}$, when QIC predicts that there is a change in slope of the penalty. The QIC, AIC, and BIC predicted complexities are compared with the true complexity for an explicit change-point analysis in Fig. 2.2, Panel C. It is immediately clear from this example that QIC quantitatively captures the true dependence of the penalty, including the change in slope at $n = 4$, exactly as predicted by the QIC complexity. As predicted, the AIC complexity is initially correct until the segmentation process must be terminated. At this point the complexity increases significantly with the result that the AIC complexity fails to terminate the segmentation process. In contrast, the BIC complexity is initially too large, but fails to grow at a sufficient pace to match the true complexity for $n > 4$.

When a change point is supported by the data (i.e. its location is reproducible in multiple realizations of the observations), the complexity is approximated by the expectation of a single chi-squared variable (i.e. the AIC complexity). When a change point is unidentifiable (the location is determined by the noise and is not reproducibly positioned), the complexity is effectively equivalent to the expectation of the maximum of a number of independent chi-squared random variables, and therefore is significantly larger than the AIC complexity. These two distinct complexity behaviors are captured by our piecewise approximation.

### *2.3 The relation between frequentist and information-based approach*

Consider the likelihood-ratio test for the following problem: We propose the binary segmentation of a single partition. In the null hypothesis ($H_0$) is the partition is described

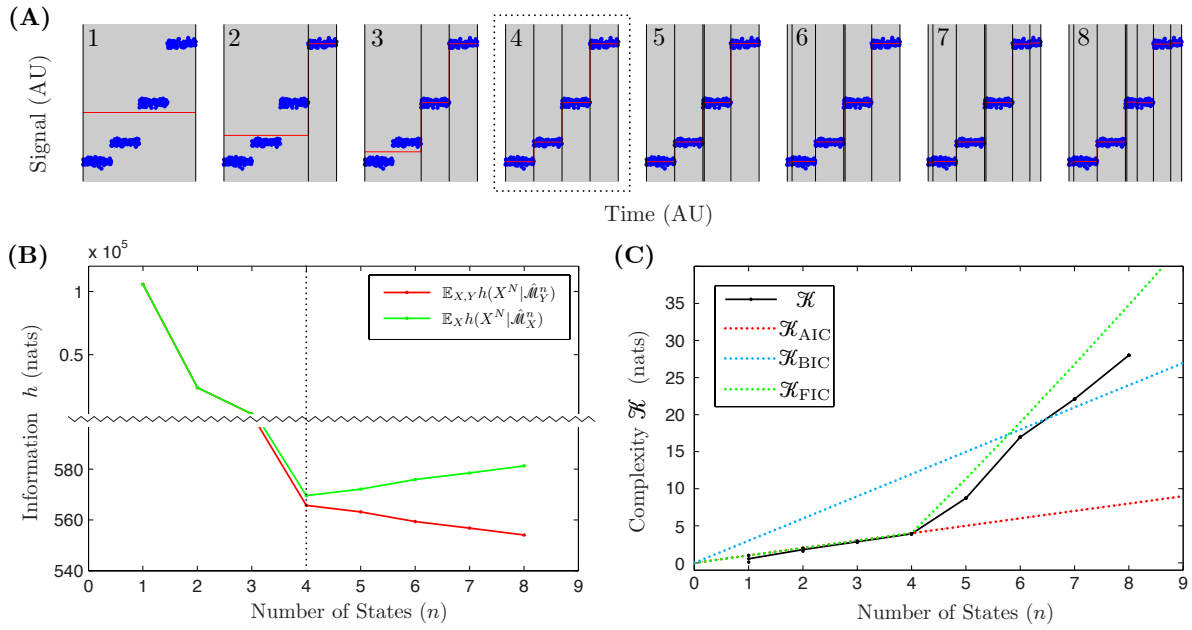


Figure 2.2: **Information-based model selection. Panel A: Nested models generated by a Change-Point Algorithm.** Simulated data (blue points) generated by a true model with four states is fit to a family of nested models (red lines) using a Change-Point Algorithm. Models fit with $1 \leq n \leq 8$ states are plotted. The fit change points are represented as vertical black lines. The true model has four states. **Panel B: Four changes points minimizes information loss.** Both the expectation of the information (red) and the cross entropy (green) are plotted as a function of the number of states $n$. The y-axis ($h$, information) is split to show the initial large changes in $h$ as well as the subsequent smaller changes for $4 \leq n \leq 8$. The cross entropy (green) is minimized by the the model that best approximates the truth ($n = 4$). The addition of parameters leads to an increase in cross entropy (green) for $n \geq 4$. The information loss estimator (red) is biased and continues to decreases with the addition of states as a consequence of overfitting. **Panel C: Complexity of Change-Point Analysis.** The true complexity is computed for the model shown in panel A via Monte Carlo simulation for $10^6$ realizations of the observations $X^N$ and compared with three models for the complexity AIC, QIC and BIC. For models with states numbering $1 \leq n \leq 4$, the true complexity (black) is correctly estimated by the AIC complexity (red dotted) and the QIC complexity (green). But for a larger number of states ($4 \leq n \leq 8$), only QIC accurately estimates the true complexity.

by a single state (unknown model parameters $\theta_0$) and the hypothesis to be tested ($H_1$) is that the partition is actually sub-divided into two states (unknown change point and model parameters $\theta_1$ and $\theta_2$). We use the log-likelihood ratio as the test statistic:

$$V(X^N) \equiv \log \frac{q(X^N|\hat{\Theta}^2_X)}{q(X^N|\hat{\Theta}^1_X)} = h(X^N|\hat{\Theta}^1_X) - h(X^N|\hat{\Theta}^2_X). \tag{2.48}$$

In the Neyman-Pearson approach to hypothesis testing, we assume the null hypothesis (1 state) and compute the distribution in the test statistic $V$. As before, we will expand the information around the true parameter values $\theta_0$. In exact analogy to Eqn. 2.33, we find that $V$ and our previously defined statistic $U$ identically distributed:

$$V \sim U, \tag{2.49}$$

up to the approximations discussed in the derivation. Therefore we will simply refer to $V$ as $U$.

In the canonical frequentist approach we specify a critical test statistic value $u_\gamma$ above which the alternative hypothesis is accepted. $u_\gamma$ is selected such that the alternative hypothesis $H_1$ is rejected given that the null hypothesis $H_0$ is true with a probability equal to the confidence level $\gamma$:

$$\gamma = F_U(u_\gamma), \tag{2.50}$$

where $F_U$ is the cumulative distribution of $U$.

Therefore we can interpret both the information-based approach and the frequentist approach as making use of the same statistic $U$. In the frequentist approach, a confidence level ($\gamma$) is specified to determine the critical value $u_\gamma$ with which to accept the two-state hypothesis. The information-based approach also uses the statistic $U$, but the critical value of the statistic ($\hbar_-$) is computed from the distribution of the statistic itself $\hbar_- = 2\overline{U}$. The information-based approach chooses the confidence level that optimizes predictivity.

### *2.4 Applications*

In the interest of brevity we have not included analysis of either experimental data or simulated data with a signal-model dimension larger than one, but we have tested the approach extensively. For instance, we have applied this technique to an experimental single-molecule biophysics application that is modeled by an Ornstein-Uhlenbeck process with state-model dimension of four [144]. We also applied the approach in other biophysical contexts including the analysis of bleaching curves, cell and molecular-motor motility [145].

### *2.5 Discussion*

In this paper, we present an information-based approach to change-point analysis using the Frequentist Information Criterion (QIC). The information-based approach to inference provides a powerful framework in which models with different parameterization, including different model dimension, can be compared to determine the most predictive model. The model with the smallest information criterion has the best expected predictive performance against a new dataset.

Our approach has two advantages over existing frequentist-based ratio tests for change-point analysis: (i) We derive an QIC complexity that depends only on the dimension of the state model ($d$), the number of states ($n$) and observations ($N$). Therefore it may be unnecessary to develop and compute custom statistics for specific applications. (ii) In the frequentist approach one must specify an *ad hoc* confidence level to perform the analysis. In the information-based approach, the confidence level is chosen automatically based upon the model complexity. The information-based approach is therefore parameter and prior free.

As the number of change-points increases, the model complexity is observed to transition between an AIC-like complexity $\mathcal{O}(N^0)$ and a Hannan-and-Quinn-like complexity $\mathcal{O}(\log\log N)$. We propose an approximate piecewise expression for this transition. The computation of this approximate model complexity can be interpreted as the expectation of the extremum of a $d$-dimensional Brownian bridge. We believe this information-based

approach to change-point analysis will be widely applicable.

# Chapter 3

# PREDICTIVE MODEL SELECTION

In Chapter 2 we developed an extension to AIC which allowed us to perform predictive model selection for the change point problem. In this chapter, our goal is to show that the method we developed in the context of the change point problem, QIC, is more generally applicable to singular models of all types, not just the change point problem. In the process we gain a deeper understanding of the effects of model singularity and it's relation to model geometry. We apply QIC to several realistic singular problems and show that it has excellent performance. We discuss the advantages of this form of information based inference over competing methods.

One of the most important results of this analysis is a deeper appreciation for the effects of a certain form of model singularity which we term *model multiplicity*. Model multiplicity is closely related to the frequentist multiple testing problem. QIC in the presence of model multiplicity displays an increased penalty on larger models which exactly mirrors the frequentist Bonferroni corrections for multiple testing. Surprisingly the effects of model multiplicity can make QIC equal to BIC! In the presence of model singularity BIC may be an appropriate *predictive* model selection criterion.

Mathematically this is an entirely uninteresting coincidence—there are no shortage of phenomena which scale logarithmically. But *sociologically* this is a very important piece of data. Multiplicity is a widespread feature of statistical models. At the same, statisticians have advocated for AIC in some contexts, and BIC in others where AIC obviously fails. Perhaps the continued preference for BIC by practicing statisticians stems from the catastrophic failure of AIC in the presence of multiplicity. It follows, then, that instead of abandoning predictive model selection in favor of Bayesian model selection, we need only correct from

multiplicity when it arises. QIC gives us one way to do this.

Where then, does this leave BIC? If predictive model selection works, does that mean Bayesian model selection is somehow problematic? We will return to these questions in Chapter 4.

### *3.1 Introduction to Model selection*

Model selection is a central problem in statistics. In the information-based paradigm of inference, models are selected to maximize the expected predictive performance. The canonical implementation of information-based inference is the minimization of the Akaike Information Criterion (AIC), an estimate for the (minus) predictive performance [3, 30]. Although it has enjoyed significant success, AIC is biased in many important applications. Model singularity, i.e. the absence of a one-to-one correspondence between model parameters and distribution functions, can make the bias extremely large and result in the catastrophic failure of model selection, as described below. There are three important and related mechanisms of failure: (i) finite-sample-size corrections, (ii) model singularity and (iii) model-training-algorithm dependence. In the course of our own analyses of biophysical and cell biology data, we frequently encounter all three phenomena. The goal of this paper is to propose a refinement to the information-based approach that overcomes these limitations.

We begin by studying the predictive complexity that plays a critical role in the mechanism of failure of AIC. We compute the exact predictive complexity of models to study its phenomenology and dependence on the parameters of the generative model. We discover that the AIC approximation for the complexity can significantly under or over-estimate the complexity, leading to pathological over-fitting or under-fitting in model selection problems. We find that parameter unidentifiability (i.e. model singularity), sample size, fitting algorithm and parameter manifold geometry can all play a critical role in determining the model complexity.

In real analyses, the true distribution is unknown and therefore the complexity must be approximated. Our exploration of the true complexity motivates a new approximation for

the complexity: the *frequentist complexity*. In this approximation, we assume the model of interest is the generative model at the estimated parameters. The frequentist complexity is not a universal function of model dimension and sample size. Instead it naturally adapts to the likelihood function, model training algorithm and sample size. We propose an improved information criterion based on this new frequentist complexity: the *Frequentist Information Criterion* (QIC).

For regular models in the large-sample-size limit, QIC is equal to AIC. Away from this limit, there can be large mismatches between the QIC and AIC. For instance, for models with large multiplicity, QIC can be much larger than AIC. For *sloppy* models [101], QIC can be much smaller than AIC. It is essential to note that QIC is still biased (since the true distribution is not know) but this bias is nearly always much smaller than the AIC estimate of the complexity and, as a result, QIC outperforms AIC (and other information criteria). QIC also outperforms other predictive methods in many contexts. To demonstrate this improved performance, we present three example analyses in Section 5.4 that highlight specific advantages QIC over other methods.

### *3.2 Information-based inference*

The goal of statistical modeling in this discussion is to approximate the unknown true distribution function $p$ which generated an observed dataset:

$$x \equiv (x_1 \dots x_N), \tag{3.1}$$

of sample size $N$. We will use $X$ (instead of $x$) when we interpret $X$ as random variables. The model $\boldsymbol{m}$ consists of a parameterized candidate probability distribution $q(x|\theta^m)$, called the likelihood, with parameters $\theta^m$ and an algorithm for training the model $\hat{\theta}^m$ [106]. The dependence of all quantities on the model $\boldsymbol{m}$ will be implicit, except where we make explicit comparisons between competing models. We will work predominantly in terms of Shannon information, defined:

$$h(x|\theta) \equiv -\log q(x|\theta), \tag{3.2}$$

where $h$ is the base-e message length (in nats) required to encode $x$ using distribution $q(\cdot|\theta)$. The output of a model-training algorithm, trained on measurements $x$, is a set of parameters $\hat{\theta}_x = \hat{\theta}(x)$. The methods that we explore apply to any model-training algorithm. For concreteness, we focus on models trained using maximum-likelihood parameter estimation. The Maximum-Likelihood-Estimate (MLE) of the parameters $\hat{\theta}_x$ is found by maximizing (minimizing) the likelihood $q(x|\theta)$ (information $h$) with respect to the parameters $\theta \in \Theta$.

It will be convenient to view both the true model and the candidate model parameter space $\boldsymbol{\Theta}$ as embedded in a higher dimensional space $\boldsymbol{\Phi}$, so $\boldsymbol{\Theta} \subseteq \boldsymbol{\Phi}$ with the true model parameterized by $\phi \in \boldsymbol{\Phi}$. We define the expected excess information loss, *i.e.* the KL-divergence:

$$D(\phi||\theta) \equiv \mathop{\mathbb{E}_X}_{\phi} h(X|\theta) - h(X|\phi). \tag{3.3}$$

The information loss, the empirical estimator for the KL-divergence is given by:

$$d_x(\phi||\theta) \equiv h(x|\theta) - h(x|\phi). \tag{3.4}$$

$D$ and $d$ act as directed distance functions and define a geometry for the parameter space of the model termed the *statistical manifold* [7, 15, 84]. For small perturbations around the true parameters, the KL-divergence can be computed using the Fisher information:

$$I(\theta) = \lim_{\theta' \to \theta} \partial_{\theta'} \otimes \partial_{\theta'} D(\theta||\theta'), \tag{3.5}$$

which can be reinterpreted as the Fisher-Rao metric and defines a local notion of distance on the manifold [14, 29].

### *3.2.1 Information criteria*

The true distribution is approximated in two steps: (i) the parameters $\hat{\theta}$ are selected in each model $\boldsymbol{m}$ as described above and (ii) a model $\hat{\boldsymbol{m}}$ is then selected among a small number of competing models. In information-based inference, models are selected to maximize the

estimated *predictive performance*. Predictive performance of the fitted model parameterized by $\hat{\theta}_x$ is measured by the cross entropy:

$$H(\phi||\hat{\theta}_x) \equiv \mathop{\mathbb{E}_X}_{\phi} h(X|\hat{\theta}_x), \tag{3.6}$$

where $X$ has identical structure to the observed data $x$. The model with the smallest cross entropy is the most predictive model, but $H$ is unknown since $\phi$ is unknown.

In information-based inference, $H$ is approximated by an *information criterion*. We will use the information $h$ as an empirical estimator. Although $h(x|\theta)$ is an unbiased estimator of $H(\phi||\theta)$, $h(x|\hat{\theta}_x)$ is biased from below:

$$\mathop{\mathbb{E}_X}_{\phi} h(X|\hat{\theta}_X) \leq \mathop{\mathbb{E}_X}_{\phi} H(\phi||\hat{\theta}_X). \tag{3.7}$$

$h(x|\hat{\theta}_x)$ describes in-sample performance, but in-sample and out-of-sample performance are distinct due to the phenomenon of overfitting. In the context of nested models[1], this bias in $h(x|\hat{\theta}_x)$ cannot be ignored since $h(x|\hat{\theta}^m_x)$ typically monotonically decreases with model dimension even as the cross entropy $H(\phi||\hat{\theta}^m_x)$ increases. Minimizing $h(x|\hat{\theta}^m_x)$ with respect to $\boldsymbol{m}$ would lead to the selection of the most complex model.

To select the model with optimum predictive performance, we must correct the bias of the cross-entropy estimator $h(x|\hat{\theta}_x)$. This bias is defined:

$$\mathscr{K} \equiv \mathop{\mathbb{E}_X}_{\phi} \left\{ H(\phi||\hat{\theta}_X) - h(X|\hat{\theta}_X) \right\}, \tag{3.8}$$

but for the purposes of computation, it is often convenient and more computationally efficient to re-write the bias in terms of the KL divergence:

$$\mathscr{K} = \mathop{\mathbb{E}_X}_{\phi} \left\{ D(\phi||\hat{\theta}_X) - d_X(\phi||\hat{\theta}_X) \right\}. \tag{3.9}$$

$\mathscr{K}$ is called the predictive complexity, or *complexity* in the interest of brevity. The complexity can be understood intuitively as the *flexibility* of the model in fitting data $x$.

---

[1] An important class of models is referred to as *nested* [106]. Lower-dimension model $\boldsymbol{m}$ is nested in higher-dimensional model $\boldsymbol{n}$ if all candidate distributions in $\Theta^m$ are realizable in $\Theta^n$.

By construction, an unbiased estimator of the cross entropy $H(\hat{\theta}_x)$ is:

$$\mathrm{IC}(x) = h(x|\hat{\theta}_x) + \mathscr{K}, \tag{3.10}$$

which is called an *information criterion*. The first term, the minimum information $h$, measures the *goodness-of-fit* of the model and typically decreases with model complexity (or dimension). The second term, the *complexity*, is a penalty that represents expected information loss due to the statistical variation of the parameter values fit to the training set $x$. As the model dimension increases, so does the complexity, while the information $h$ decreases with model dimension. As a consequence of these competing imperatives (improving the fit while minimizing the model complexity), the information criterion has a minimum with respect to model dimension corresponding to the estimated optimally predictive model.

### *3.2.2 The Akaike Information Criterion*

Although neither the complexity (Eqn. (3.9)) nor the information criterion (Eqn. (3.10)) can be computed if the true parameters $\phi$ are unknown, in practice the $\phi$ dependence vanishes asymptotically. In the large-sample-size limit of a regular model, a surprisingly simple expression is derived for the complexity:

$$\mathscr{K} = K + \mathcal{O}(N^{-1}), \tag{3.11}$$

where the model dimension is $K \equiv \dim \Theta$. This complexity approximation can be understood as the leading-order contribution to a perturbative expansion of the complexity in inverse powers of the sample size $N$. Using Eqn. (3.11), we can write the well-know Akaike Information Criterion (AIC)[2]:

$$\mathrm{AIC}(x) = h(x|\hat{\theta}_x) + K, \tag{3.12}$$

which does not depend on (i) the true distribution $\phi$, (ii) the detailed functional form of the candidate models $q(x|\theta)$, (iii) the data structure or (iv) the sample size $N$.

[2]Historically, AIC was defined as twice Eqn. (3.12) for consistency with the deviance [30]. There is no significance to this multiplicative factor.

Although the AIC information criterion has been successfully applied in many problems, the AIC approximation for the complexity can fail in many unexceptional contexts. For instance, the $\mathcal{O}(N^{-1})$ correction may not be small at finite sample size. Alternatively, the structure of the model can cause AIC to fail. For instance, a parameter $\theta$ is called *unidentifiable* if $q(\cdot|\theta) = q(\cdot|\theta')$ for $\theta \neq \theta'$. If a model includes unidentifiable parameters, the model is called *singular*, as opposed to a *regular* statistical model [143]. For singular models, AIC fails at all samples sizes. As our examples in Sec. 5.4 will illustrate, both these mechanisms of failure naturally arise in many analyses.

## *3.3 Complexity Landscapes*

To study the phenomenology and investigate novel approximations for the complexity, we compute it for realizable models as a function of the true parameter $\theta$. We will find that although the AIC complexity is correct in the large-sample-size limit of a regular model, there can be significant deviation from this approximation at finite sample size, in singular models, and as a result of parameter-space constraints.

### *3.3.1 The finite-sample-size complexity of regular models*

In general, the complexity will depend on both the sample size $N$ and the true parameter $\theta$. However, statistical models with symmetries can lead to a complexity independent of the true parameter. For instance, consider a family of distributions:

$$q(x|\theta) = C_\alpha \, \lambda^{1/\alpha} \, e^{-\lambda|x|^\alpha}, \tag{3.13}$$

with parameters $\theta = (\lambda, \alpha)$ and support $\lambda, \alpha \in \mathbb{R}_+$ and normalization:

$$C_\alpha^{-1} \equiv \Gamma(1 + \alpha^{-1}) \times \begin{cases} 2, & x \in \mathbb{R} \\ 1, & x \in \mathbb{R}_+. \end{cases} \tag{3.14}$$

This family includes the exponential ($\alpha = 1$, $x \in \mathbb{R}_+$), the centered-Gaussian ($\alpha = 2$, $x \in \mathbb{R}$), the Laplace ($\alpha = 1$, $x \in \mathbb{R}$) and the uniform ($\alpha \to \infty$) distributions. The transformation

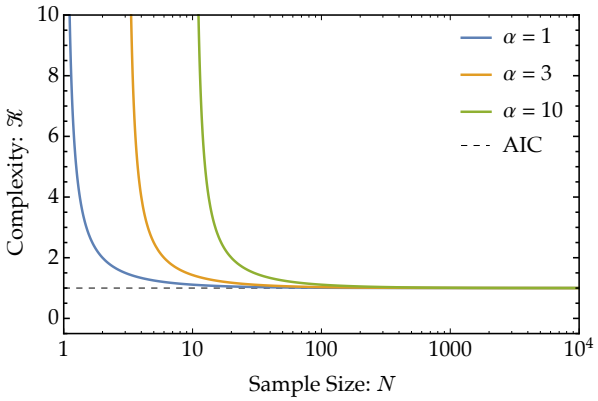


Figure 3.1: **Complexity at finite sample size.** Although AIC estimate accurately estimates the large-sample-size limit of the complexity of regular models, there can be significant finite-sample-size corrections. For instance, the modified-center-Gaussian model has a significantly larger complexity than the AIC limit for small $N$. In fact the complexity diverges for $N \leq \alpha$, implying that the model has insufficient data to make predictions.

of this distribution under dilations on $x$ implies the complexity must be independent of $\lambda$. In the Appendix Sec. B.1.1, we derive a general result for exponential-family models. This problem is a special case of that expression. The complexity for unknown $\lambda$ and known $\alpha$ is:

$$\mathscr{K} = \tfrac{N}{N-\alpha} \quad \text{for} \quad N > \alpha, \tag{3.15}$$

as shown in Appendix Sec. B.1.2.

The complexity of this centered-modified Gaussian family is equal to one asymptotically ($N \to \infty$) but can significantly diverge from this AIC limit at finite sample size $N$, as shown in Fig. 3.1. The finite-sample-size correction is particularly large for large values of the exponent $\alpha$. In this regime, the MLE algorithm tends to strongly overestimate the fit to the data. In fact, the complexity is infinite in the uniform distribution limit ($\alpha \to \infty$) where a Bayesian approach, which hedges over parameter $\lambda$, is required to give acceptable predictive performance at any sample size $N$. (See Ref. [90] for more information.)

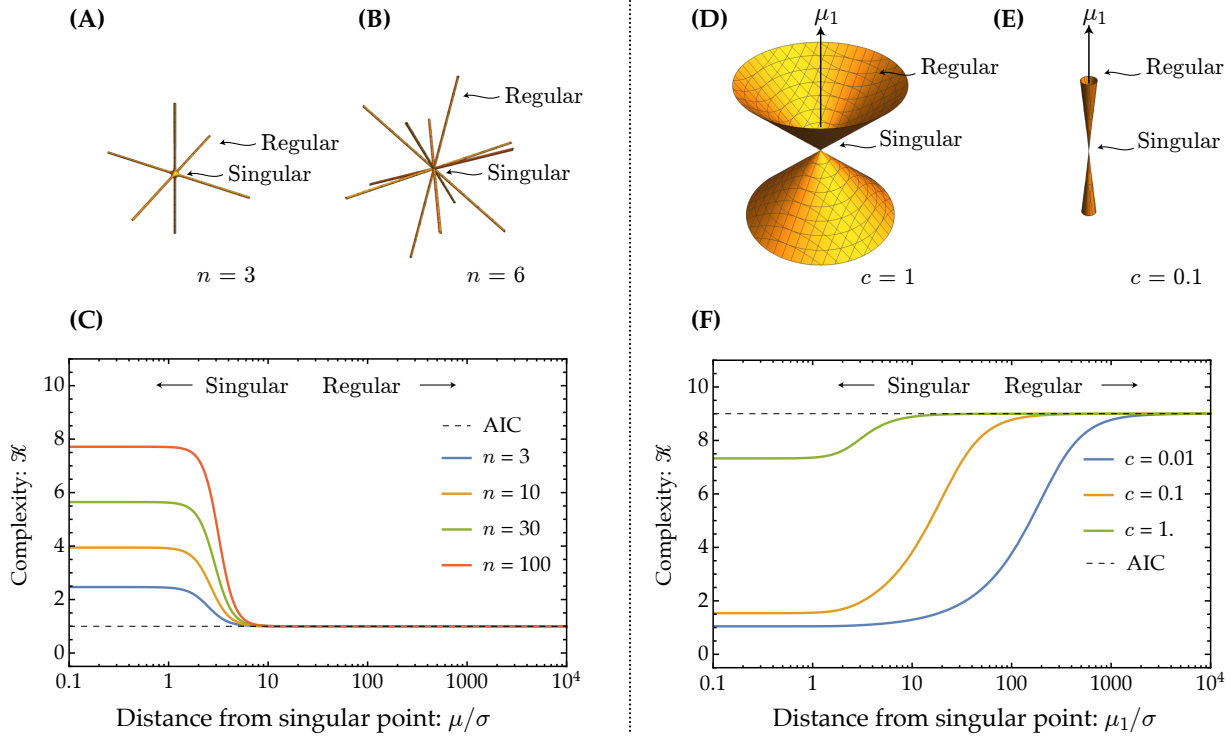


Figure 3.2: **Complexity landscapes in singular models. AIC underestimates the complexity in the component selection model. Panel A-B:** Schematic sketches of the geometry of parameter space for two different multiplicity values: $n = 3$ and $n = 6$. **Panel C:** The AIC estimate $\mathscr{K} = 1$ (dashed line) matches the true complexity far from the singular point ($|\mu/\sigma| \gg 0$). Close to the singularity ($|\mu/\sigma| \approx 0$), the true complexity is much larger than the AIC estimate. The complexity grows with the number of means $n$ due to *multiplicity*. **AIC overestimates the complexity in the n-cone model. Panel D-E:** Schematic sketches of parameter space for a wide cone ($c = 1$) and a needle-like cone ($n = 0.1$). **Panel F:** For $n = 10$ dimensions, the AIC estimate $\mathscr{K} = n - 1$ (dashed line) matches the true complexity far from the singular point ($|\mu_1/\sigma| \gg 0$). Close to the singularity ($|\mu_1/\sigma| \approx 0$), the true complexity is much smaller than the AIC estimate. The complexity shrinks for small cone angles ($c \to 0$) since the cone geometry is needle-like with effectively a single degree for freedom ($\mu_1$).

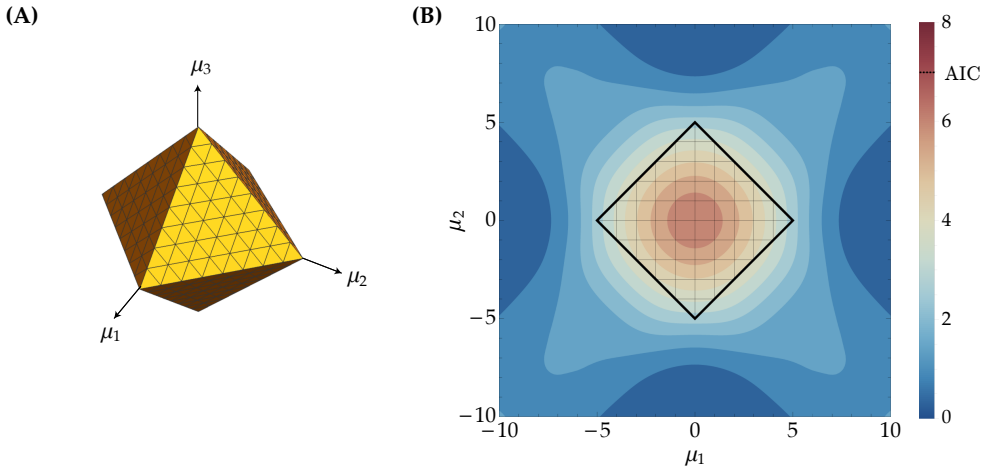


Figure 3.3: **Complexity of $L_1$-constrained model. Panel A:** Schematic sketch of a slice of the seven-dimensional parameter space. Parameter values satisfying the $L_1$ constrain lie inside the simplex. **Panel B:** Complexity as a function of the true parameter value $\vec{\mu} = (\mu_1, ..., \mu_7)$. (Only a slice representing the $x$-$y$ plane is shown.) The black-hatched region represents parameter values satisfying the constraints. The $L_1$ constraint significantly reduces the complexity below the AIC estimate $\mathcal{K} = 7$. The complexity is lowest outside the boundaries of the simplex where the constraints trap MLE parameter estimates and reduce statistical fluctuations.

### *3.3.2 Singular models*

Singular models have parameter unidentifiability that cannot be removed by coordinate transformation. These models can show very large deviations from the AIC complexity at all sample sizes in the vicinity of the singular point in parameter space. The deviation can either significantly increase or decrease the complexity as we will illustrate with two closely related examples. This singular class of models are common place in many analyses, especially in the context of nested models, and therefore they pose a significant limitation to the more general use of AIC.

To explore the properties of a singular model, consider the following simple example: the component selection model. An $n$-dimensional vector of observations $\vec{x} \in \mathbb{R}^n$ is normally distributed about an $n$-dimensional vector of means $\vec{\mu} \in \mathbb{R}^n$ with variance one. The likelihood is:

$$q(\vec{x}|\boldsymbol{\theta}) = (2\pi)^{-n/2} \exp[-\tfrac{1}{2}(\vec{x} - \vec{\mu})^2]. \tag{3.16}$$

We consider the model where all but one of the components of the vector mean are zero:

$$\vec{\mu} = (0, ..., \mu_i, ..., 0), \tag{3.17}$$

but the identity, $i$, of the non-zero component is unknown as well as the mean $\mu_i \equiv \mu$. The parameters are defined $\theta \equiv (i, \mu)$ with support $\mu \in \mathbb{R}$ and where the index $i$ is an integer on the interval $[1, n]$. This model is singular when $\mu = 0$ since the likelihood is independent of $i$. The complexity must be computed numerically (Appendix Sec. B.1.3) and depends on $\mu$ but is independent of $i$ (permutation symmetry) and is plotted in Fig. 3.2A. As shown in the figure, there is a large deviation from the AIC complexity in the singular region $\mu = 0$ and the complexity is large compared with the model dimension, irrespective of sample size $N$. Far from the singular point, the complexity is $\mathcal{K} = 1$ which matches the AIC complexity for a single continuous parameter ($\mu$) and the discrete parameter $i$ does not contribute to the complexity in this limit.

To demonstrate that singular models can have reduced complexity relative to AIC, consider the same likelihood function (Eqn. (3.16)), but a different parameter manifold. We constrain the mean $\vec{\mu}$ to lie on the surface of a $n$-cone, defined by the equation:

$$\mu_1^2 c^2 = \sum_{i=2}^{n} \mu_i^2, \tag{3.18}$$

where $\alpha = \tan^{-1} c$ is the angle of the cone. This cone geometry has been previously suggested to represent the fundamental geometry mixture models [59, 9]. The model is singular at the vertex of the cone $\mu_1 = 0$. The complexity can be computed analytically (Appendix Sec. B.1.4) and is shown in Fig. 3.2B. Like the previous singular model, there is a large deviation from the AIC complexity at the singular point $\mu_1 = 0$ where the complexity is small ($\mathscr{K} \approx 1$) compared with the model dimension ($n - 1$), irrespective of sample size $N$. Far from the singular point, the complexity is $\mathscr{K} \approx n-1$, which matches the AIC complexity for $n - 1$ dimensional parameter manifold. In general, we expect a strong failure of the AIC approximation in the vicinity of the singularity, but far from the singular point, the AIC approximation applies in the large-sample-size limit.

### *3.3.3 Constrained models*

A canonical approach to regularizing high-dimensional models are convex constraints, including $L_1$ constrained optimization. Consider the same likelihood function (Eqn. (3.16)), but with convex constraint:

$$\sum_{i=1}^{n} |\mu_i| \leq \lambda, \tag{3.19}$$

where $\lambda$ is a constraint chosen by the analyst. The complexity landscape can be computed numerically (Appendix Sec. B.1.6) and is shown in Fig. 3.3. As expected, the constraint works to significantly reduce the complexity far below the AIC estimate at finite sample size, especially when the true parameter lies somewhere close or outside the subspace of parameter space that satisfies the constraint (Eqn. (3.19)).

### *3.4 Frequentist Information Criterion (QIC)*

In each example discussed in the previous section, we demonstrated a significant mismatch between the AIC complexity and the true complexity. In practice, these corrections are often important since (i) singular model are widespread and (ii) all real analyses occur at finite sample size. A significant bias in the complexity can lead to failures in model selection and, in the context of recursively-nested singular models, it can lead a catastrophic breakdown in model selection where the selected model dimension grows with sample size indefinitely, irrespective of the generative distribution (*e.g.* Sec. 3.5.2). Our goal is therefore to develop an improved approximation for the complexity.

Clearly the ideal situation would be to use the true complexity $\mathscr{K}(\phi)$, but $\mathscr{K}(\phi)$ depends upon the unknown generative distribution, *i.e.* the unknown parameter $\phi$. To circumvent this difficulty, we propose using a natural approximation in the current context: We approximate $\theta$ with the point estimate $\hat{\theta}_x$ and define the frequentist approximation of the complexity:

$$\mathscr{K}_{\mathrm{QIC}}(x) \equiv \mathscr{K}(\hat{\theta}_x). \tag{3.20}$$

where $\mathscr{K}(\theta)$ is the true complexity for data generated by a realizable distribution with parameter $\theta$. We call this a *frequentist approximation* since the *complexity is computed with respect to hypothesized data distributions* in close analogy to the computation of the distribution of a frequentist test statistic. Unlike a frequentist test, no *ad hoc* confidence level must be supplied by the analyst. We generically expect the frequentist complexity to depend on (i) the data $x$, (ii) the functional form of candidate models $q$, (iii) the training algorithm and (iv) the sample size $N$. In general, the complexity must be computed numerically, although analytic results or approximations can be used in many models.

Now that we have defined a novel approximation for the complexity, we can define the corresponding information criterion:

$$\mathrm{QIC}(x) = h(x|\hat{\theta}_x) + \mathscr{K}_{\mathrm{QIC}}, \tag{3.21}$$

which we call the Frequentist Information Criterion (QIC). In analogy to the AIC analysis,

the model that minimizes QIC is estimated to have the best predictive performance.

### *3.4.1 Measuring model selection performance*

The QIC approach is to compute an approximate complexity (Eqn. (3.20)) in order to construct the information criterion (Eqn. (3.10)), an estimator of the cross entropy $H(\phi||\hat{\theta}_x)$. In simulations, $\phi$ is known. Therefore the estimated complexity can be compared with the true complexity and the information criterion can be compared with the cross entropy $H(\phi||\hat{\theta}_x)$.

A more direct metric for the performance of information criteria is the information loss of the selected model $\hat{m}$. The selected model $\hat{m}$ is that which minimizes the information criterion

$$\hat{m}(x) = \arg\min_m \mathrm{QIC}^m(x). \tag{3.22}$$

The expected performance of a selection criterion is the KL divergence averaged over training sets $X$,

$$\overline{D} \equiv \mathop{\mathbb{E}}_{\phi}{}_X \, D(\phi||\hat{\theta}_X^{\hat{m}(X)}), \tag{3.23}$$

where $q_X(\cdot)$ is the estimate of $p$, which is the result of the model selection procedure (3.22). The better the performance of the model selection criterion, the smaller the information loss $\overline{D}$.

## *3.5 Applications of QIC*

In Sec. 3.3, we described two important contexts in which AIC fails: (i) finite sample size and (ii) in singular models. Before considering a formal analysis of the performance of QIC, we explore this criterion in the context of a number of sample problems. First, we analyze a problem of modeling the motion of large complexes in the cell in Sec. 3.5.1. In this problem, finite sample size plays a central role in the choice of models when we compare two models with the same dimension. As expected, QIC outperforms AIC. In the next analysis in Sec. 3.5.2, we analyze a Fourier Regression problem. In this analysis, we fit the data using two different algorithms, one of which is singular. In the analysis of the singular model, there

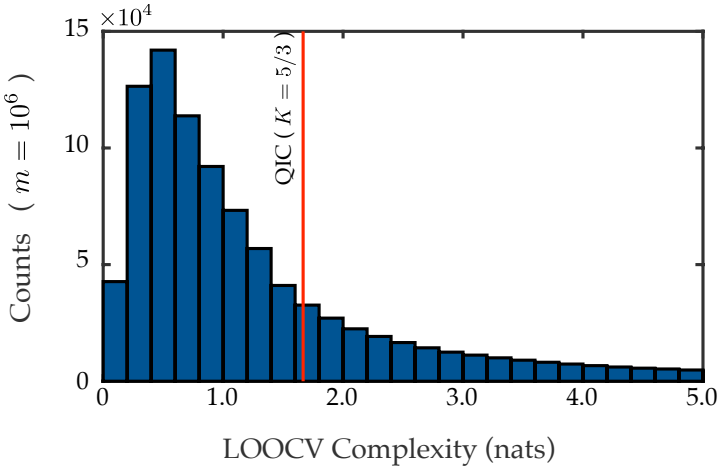


Figure 3.4: **Monte Carlo histogram of the LOOCV procedure:** A histogram of $10^6$ simulations of the effective cross-validation complexity $\mathscr{K}_{\mathrm{CV}}(X)$ for the normal model with unknown variance $\alpha = 2$ compared to the QIC result $\mathscr{K} = \frac{5}{5-2}$ for $(N = 5)$. The lower variance of the QIC complexity often results in better model selection properties, especially at low sample sizes relative to cross-validation.

is a catastrophic failure of AIC where the dimension of the AIC-selected model is much larger than the optimally predictive model due to the large size of the true complexity relative to the AIC estimate. Again, we demonstrate that QIC gives a good approximation for the complexity, both in the context of the singular and regular models. In the final example in Sec. 3.5.3, we analyze a singular model in which the complexity is significantly smaller than the model dimension. As expect, QIC outperforms AIC in this context as well.

### *3.5.1 Small sample size and the step-size analysis*

In this section, we explore the small-sample-size limit in the context of a problem with two competing models of the same dimension but different true complexities. Inspired by our recent experimental work [92, 135], we model the step-size distribution of large protein complexes in the cytoplasm undergoing stochastic motion. In this problem, individual complexes

can only be tracked over a short interval of $N_t$ steps. Although many trajectories can be captured ($N_T$), complex-to-complex and cell-to-cell variation implies that different parameters describe each short trajectory (length $N_t$) and therefore the complexity is in the finite-sample-size limit. In the current context, we simulate two experiments where the generative distributions are the (i) centered-Gaussian and the (ii) Laplace distributions, respectively.

*Analysis*

The likelihoods are defined in Eqn. (3.13) for $\alpha = 1$ for the Laplace and $\alpha = 2$ for the centered-Gaussian model. We define differences in the information criterion as the Gaussian minus the Laplace model, $\Delta(\cdot) \equiv (\cdot)_2 - (\cdot)_1$, where negative values of $\Delta(\cdot)$ select the Gaussian model and positive values select the Laplace model. The AIC complexity for both models is $\mathscr{K}_{\rm AIC} = 1$ per trace. The QIC complexity is given by Eqn. (3.15) per trace. Therefore the overall complexities for all traces are:

$$\mathscr{K}_{\rm AIC} = N_T, \tag{3.24}$$

$$\mathscr{K}_{\rm QIC} = N_T \tfrac{N_t}{N_t - \alpha}. \tag{3.25}$$

In this particular applications, QIC subsumes AICc, a previously proposed corrected AIC [71, 30]. The average IC differences are:

| Generative Model | Sample size $N_t$ | Sample size $N_T$ | $\Delta\overline{\rm AIC}$ (nats) | AIC Model Selection | $\Delta\overline{\rm QIC}$ (nats) | QIC Model Selection |
|---|---|---|---|---|---|---|
| Gaussian | 5 | 100 | $-46.7$ | Gaussian | $-5.1$ | Gaussian |
| **Laplace** | 5 | 100 | $-17.1$ | **Gaussian** | $+24.5$ | **Laplacian** |
| Gaussian | 100 | 5 | $-25.3$ | Gaussian | $-25.3$ | Gaussian |
| Laplace | 100 | 5 | $+32.6$ | Laplacian | $+32.6$ | Laplacian |

where we have highlighted the discrepancy in the analysis in bold. At large sample size ($N_t = 100$), AIC and QIC both correctly select the generative distribution. But at small sample

size, AIC incorrectly selects the Gaussian model when the generative model is Laplace. Qualitatively, the larger complexity of the Gaussian relative to the Laplace model implies that the model has a greater propensity to overfit by underestimating the information at the MLE parameters. As a result, AIC model selection prefers the Gaussian over the Laplace model, even when the Laplace model is both (i) the generative distribution and (ii) more predictive. Furthermore, since $|\Delta\text{AIC}| \gg 1$, the AIC analysis incorrectly indicates that there is extremely strong support for the Gaussian model. In contrast to AIC, QIC selects the optimal model in both experiments and at both sample sizes.

*Comparison of QIC and cross-validation*

In the current example, the data is assumed to be *unstructured* meaning that each observation $x_i$ is independent and identically distributed (in each trace). In these cases, there is a powerful alternative approach to estimating the predictive performance: Leave-One-Out-Cross-Validation (LOOCV). In the LOOCV estimate, each data point is predicted with parameters fit to the remaining $N - 1$:

$$\text{LOOCV}(x) = \sum_{i=1}^{N} h(x_i|\hat{\theta}_{x_{\neq i}}), \tag{3.26}$$

where $x_{\neq i}$ is shorthand for the dataset excluding $x_i$. To examine the relative performance LOOCV, AIC and QIC, we now consider performing model selection trajectory-by-trajectory ($N_T = 1$) for five-step trajectories ($N_t = 5$). For simplicity, consider data generated by the Laplace model where the complexity plays a central role in model selection due to the propensity of the Gaussian model to overfit. We then simulate the probability of the selection of the Laplace model by each criterion:

| Criterion | AIC | QIC | LOOCV |
|---|---|---|---|
| Probability of selecting Laplace | 34% | 61% | 53% |

which demonstrates that QIC outperforms both AIC and LOOCV, at least in the current context.

Why does LOOCV perform poorly? Although LOOCV is only weakly biased, it typically has a larger variance than QIC. To understand qualitatively why this is the case, we define an effective LOOCV complexity:

$$\mathscr{K}_{\mathrm{CV}}(x) \equiv \mathrm{LOOCV}(x) - h(x|\hat{\theta}_x), \tag{3.27}$$

which reinterprets LOOCV as a information criterion with a data-dependent complexity. The complexity $\mathscr{K}_{\mathrm{CV}}(x)$ acts like a weakly-biased estimator of the true complexity, but is subject to statistical variation, as shown in Fig. 3.4. It is this variance that can lead to a loss in performance, even when the bias of the estimator is small. In contrast, the QIC complexity is constant in the current example.

LOOCV and QIC each have respective advantages. The advantage of LOOCV is that the data used to compute the estimated predictive performance were all generated by the true distribution. The frequentist complexity depends upon an assumed distribution, which can lead to a bias in QIC. LOOCV is also biased since it estimates the performance of predicting 1 measurement given $N-1$ rather than 1 measurement given $N$. Our own unpublished experiments indicate that whether LOOCV or QIC is more biased is model and sample-size dependent. However, QIC does have two important and generic advantages: (i) it typically has less variance than LOOCV and (ii) it can also be applied to analyses of structured data where LOOCV cannot be applied, as illustrated in the next example.

### *3.5.2 Anomalously large complexity and the Fourier regression model*

In this example we have two principal aims: (i) to explore the behavior of QIC in the context of a singular model with large complexity and (ii) to demonstrate the dependence of the QIC complexity on the model fitting algorithm. We present a model of simulated data inspired by the measurements of the seasonal dependence of the neutrino intensity detected at *Super-Kamiokande* [45].

*Problem setup*

We simulate normally distributed intensities with arbitrary units (AU) with unit variance: $X_j \sim \mathcal{N}(\mu_j, 1)$, where the true mean intensity $\mu_j$ depends on the discrete-time index $j$:

$$\mu_j \quad = \quad \sqrt{120 + 100 \sin(2\pi j/N + \pi/6)} \text{ AU}, \tag{3.28}$$

and the sample size is equal to the number of bins: $N = 100$. This true distribution is therefore *unrealizable* for a finite number of Fourier modes. The generating model, simulated data and two model fits are shown in Figure 3.5, Panel A.

We expand the model mean ($\mu_i$) and observed intensity ($X_i$) in Fourier coefficients $\tilde{\mu}_i$ and $\tilde{X}_i$ respectively. A detailed description is provided in the Appendix Sec. B.1.9. The MLE that minimizes the information is $\hat{\tilde{\mu}}_i = \tilde{X}_i$. We now introduce two different approaches to encoding our low-level model parameters $\{\tilde{\mu}_i\}_{i=-N/2 \ldots N/2}$: the *sequential* and *greedy algorithms*. In both cases, the models will be built by selecting a subset of the same underlying model parameters, the Fourier coefficients ($\tilde{\mu}_i$).

*Sequential-algorithm analysis*

In the *sequential algorithm* we will represent our nested-parameter vector as follows:

$$\theta_{(n)} = \begin{pmatrix} & \tilde{\mu}_{-1} & \ldots & \tilde{\mu}_{-n} \\ \tilde{\mu}_0 & \tilde{\mu}_1 & \ldots & \tilde{\mu}_n \end{pmatrix}, \tag{3.29}$$

where all selected $\tilde{\mu}_i$ are set to their respective maximum likelihood values and all other $\tilde{\mu}_i$ are identically zero. We initialize the algorithm by encoding the data with parameters $\theta_{(0)}$. We then execute a sequential nesting procedure, increasing temporal resolution by adding the Fourier coefficients $\tilde{\mu}_{\pm i}$ corresponding to the next smallest integer frequency index $i$. (Recall there are two Fourier coefficients at every frequency, labeled $\pm i$, except at $i = 0$.) The cutoff frequency is indexed by $n$ and is determined by the model selection criterion.

From the AIC perspective, the complexity is simply a matter of counting the parameters fit for each model as a function of the nesting index. Counting the parameters in Eqn. (3.29)

gives the expression for the complexity $\mathscr{K}_{\rm AIC} = 2n+1$, since both an $\tilde{\mu}_i$ and an $\tilde{\mu}_{-i}$ are added at every level. Since this is a normal model with known variance, QIC estimates the same complexity as AIC. In the Bayesian analysis, the complexity is: $\mathscr{K}_{\rm BIC} = \frac{1}{2}(2n+1)\log N$, where $N = 100$, which is significantly larger than the AIC and QIC. (See Sec. B.1.8 for a discussion of the BIC analysis.) Panel B of Figure 3.5 shows QIC model selection for the sequential algorithm. The $n = 2$ nesting level minimizes QIC and this model ($n = 2$) is shown in Panel A. The true and QIC complexity are compared in Panel D for a sample size of $N = 1000$. Both AIC and QIC are excellent approximations of the true complexity.

*Greedy-algorithm analysis*

Instead of starting with the lowest frequency and sequentially adding terms, an alternative approach would be to consider all the Fourier coefficients and select the largest magnitude coefficients to construct the model. In the *greedy algorithm* we will represent the Fourier coefficients as

$$\theta_{(n)} = \begin{pmatrix} 0 & i_1 & ... & i_n \\ \tilde{\mu}_0 & \tilde{\mu}_{i_1} & ... & \tilde{\mu}_{i_n} \end{pmatrix}, \tag{3.30}$$

where the first row represents the Fourier index and the second row is the corresponding Fourier coefficient. As before, all unspecified coefficients are set to zero. We initialize the algorithm by encoding the data with parameters $\theta_{(0)}$ and then we execute a sequential nesting procedure: At each step in the nesting process, we choose the Fourier coefficient with the largest magnitude (not already included in $\theta_{(n-1)}$). The optimal nesting cutoff will be determined by model selection.

If one counts the parameters, the AIC and BIC complexities are unchanged. There are still two parameters in Eqn. (3.30) at every nesting level $n$. For the QIC complexity, the distinction between the sequential and greedy algorithms has profound consequences. The greedy-algorithm model is singular since the Fourier mode number $i_n$ becomes unidentifiable after the last resolvable Fourier mode is incorporated into $\theta_{(n)}$. There are two approaches to computing the QIC complexity: (i) Monte Carlo and (ii) an analytical piecewise approxima-

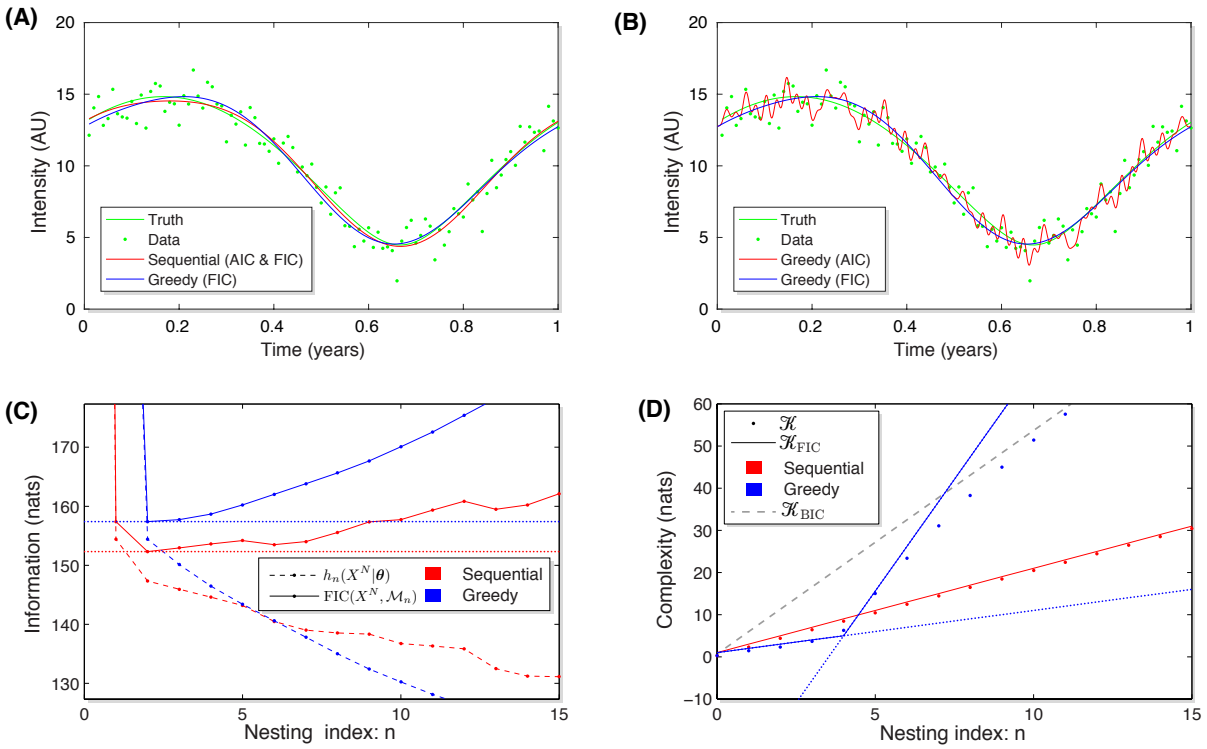


Figure 3.5: **Panel A: Truth, data and models.** (Simulated for $N = 100$.) The true mean intensity is plotted (solid green) as a function of season, along with the simulated observations (green points) and models fitted using two different algorithms, sequential (red) and greedy (blue). **Panel B: Failure of AIC for greedy algorithm.** (Simulated for $N = 100$.) For the greedy algorithm, the coefficients selected using AIC (red) are contrasted with the coefficients chosen using QIC. The QIC mean estimates (blue) track the true means very closely, unlike the AIC selected mean. **Panel C: Information as a function of model dimension.**(Simulated for $N = 100$.) The information is plotted as a function of the nesting index $n$. The dashed curves represent the information as a function of nesting index and both are monotonically decreasing. The solid curves (red and blue) represents the estimated average information (QIC), which is equivalent to estimated model predictivity. **Panel D: The true complexity matches** QIC **estimates.** (Simulated for $N = 1000$.) In the sequential-algorithm model, the true complexity (red dots) is AIC-like (solid red). In the greedy-algorithm model, the true complexity (blue dots) transitions from AIC-like (slope $= 1$) to BIC-like (slope $\propto \log N$) at $n = 4$. In both cases, the true complexity is correctly predicted by QIC (solid curve).

tion that we developed for computing an analogous complexity in change point analysis [88]. We will use the analytical approach, which gives a change in complexity on nesting of:

$$\mathscr{K}_n - \mathscr{K}_{n-1} \approx \begin{cases} 1, & -\Delta h > k \\ k, & \text{otherwise} \end{cases}, \tag{3.31}$$

where the change in information is defined $\Delta h \equiv h_n(x|\hat{\theta}_x) - h_{n-1}(x|\hat{\theta}_x)$ and the singular complexity is $k \approx 2 \log N$ (*i.e.* BIC scaling). The singular complexity $k$ arises due to picking the largest remaining Fourier mode. The approximation is given by computing the expectation of the largest of $N$ chi-squares, which is discussed in more detail in the supplement (Sec. B.1.5). If $-\Delta h > k$ the model is in a regular part of parameter spaces whereas if $-\Delta h < k$, the model is essentially singular [88]. The complexity is computed by re-summing Eqn. (3.31).

Panel B of Figure 3.5 shows QIC model selection for the greedy algorithm. The $n = 2$ nesting level minimizes QIC and this model ($n = 2$) is shown in Panel A. The true and QIC complexity are compared in Panel D for a sample size of $N = 1000$. This large sample size emphasizes the difference between the slopes. In the greedy algorithm, only QIC provides an accurate approximation of the true complexity. For large nesting index, the piecewise approximation made to compute the QIC complexity fails due to order statistics. (The largest of $m$ $\chi^2$ random variables is larger than the second largest.) This is of little consequence since the complexity in this regime is not relevant to model selection. The use of AIC model selection in this context leads to significant over fitting by the erroneous inclusion of noise-dominated Fourier modes, as shown in Panel B of Fig. 3.5.

The predictive performance of the average selected model has been determined by Monte-Carlo simulations and is plotted in Fig. 3.6, for the greedy and sequential algorithms. QIC shows correct scaling behavior for both fitting algorithms, which allows it to achieve good performance in both cases, whereas AIC (and not BIC) performs well in the Sequential case and BIC (and not AIC) performs well in the Greedy case.

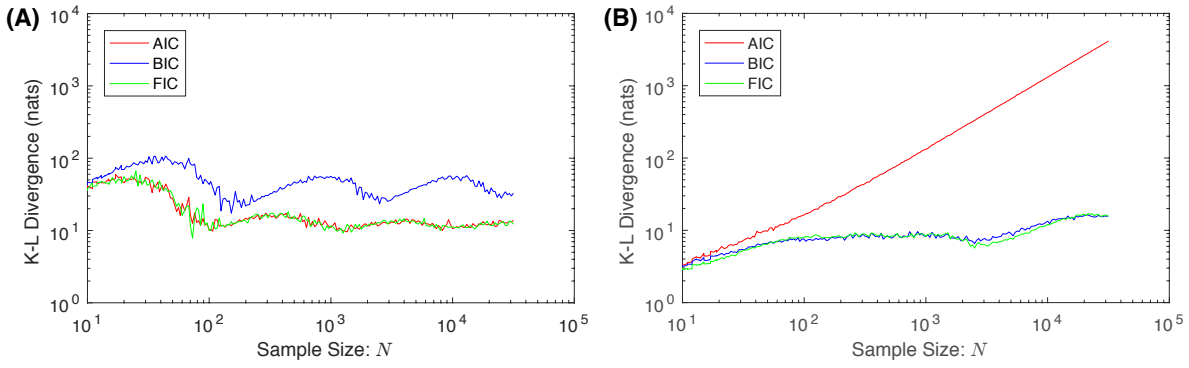


Figure 3.6: **Panel A: Performance of the sequential algorithm.** Simulated performance as measured by the KL Divergence $\overline{D}$ (Eqn. (3.23)) of sequential algorithm at different sample sizes using AIC, QIC and BIC (lower is better). AIC and QIC are identical in this case; they differ only because of the finite number of Monte Carlo samples. Larger fluctuations are arise from the structure of true modes at the resolvable scale of a given sample size. **Panel B: Performance of the greedy algorithm.** Simulated performance of greedy algorithm as measured by the KL Divergence $\overline{D}$ (Eqn. (3.23)) at different sample sizes using AIC, QIC and BIC (lower is better). QIC and BIC have very similar cutoff penalties. Because of the algorithmic sensitivity, QIC can have the appropriately complexity scaling with $N$ in both the greedy and sequential case.

### *3.5.3 Anomalously small complexity and the exponential mixture model*

In the greedy algorithm implementation of Fourier regression, both AIC and BIC *underestimated* the true complexity. But, the true complexity is not always underestimated by AIC. In sloppy models [101, 90], we find that the AIC (and BIC) approximation for the complexity typically *overestimate* the true complexity at finite sample size. To explore this phenomenon, we analyze an exponential mixture model.

In an exponential mixture model, $m$ different components decay at rate $\lambda_i$. The rates ($\lambda_i$), the relative weighting of each component in the mixture ($\omega_i$) and even the number of components ($m$) are all unknown. We represent the model parameters $\boldsymbol{\theta} = (\boldsymbol{\lambda},\ \boldsymbol{\omega})$ and the candidate distribution function for the probability density of observing an event at time $t$ is:

$$q(t|\boldsymbol{\theta}) = \sum_{i=1}^{m} \omega_i\ \lambda_i e^{-\lambda_i t}, \tag{3.32}$$

with support $\omega_i, \lambda_i \in \mathbb{R}_+$ and constraint $\sum_i \omega_i = 1$. For $m > 1$, this model is singular where $\omega_i = 0$ or $\lambda_i = \lambda_j$ for $i \neq j$. Exponential mixture models are frequently applied in biological and medical contexts where the different rates might correspond to independent signaling pathways, or sub-populations in a collection of organisms, *etc.*

#### *Problem setup*

To explore the properties of the model, we simulate data from a realizable model with $m = 4$ components and parameters:

$$\boldsymbol{\theta} \equiv \begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\omega} \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 5 \\ 0.3 & 0.2 & 0.2 & 0.3 \end{pmatrix}. \tag{3.33}$$

For a large enough sample size, $N$, AIC could be expected to accurately estimate the complexity for an $m = 4$ model. In practice, the sample size is always finite and therefore it is important to investigate the finite-sample size properties of the complexity. We simulated $N = 100$ samples from the true distribution.

*Analysis*

In our statistical analysis, we consider just two competing models, $m = 1$ and 4 component models, for simplicity. For the AIC and BIC complexities, we used a model dimension of $K = 2m - 1$ due to the normalization constraint on component weights $\omega_i$. The QIC complexity for $m = 1$ has an analytic form given by Eqn. 3.35 while the complexity for $m = 4$ was computed by Monte Carlo. The true complexity and the AIC, BIC and QIC approximations are compared for the two models below:

| | Complexity $\mathscr{K}$ (nats) | | | |
|---|---|---|---|---|
| Model | True | QIC | AIC | BIC |
| $m = 1$ | 1.77 | 1.01 | 1 | 4.61 |
| $m = 4$ | 3.33 | 3.45 | 7 | 16.1 |

QIC shows excellent agreement with the true complexity for $m = 4$. The discrepancy when $m = 1$ occurs because QIC has approximated the true distribution ($m = 4$) with the fitted model ($m = 1$). The true distribution in this case is *not realizable*, but nonetheless this approximation still provides the best estimate of the true complexity. For the one component model ($m = 1$), AIC makes nearly the same estimate for the complexity as QIC, but it significantly overestimates the complexity of the larger 4 component model ($m = 4$). At finite sample size, this model is therefore more predictive than estimated by AIC. The BIC complexity never accurately approximates the true complexity.

The difference in estimated complexity has important consequences for model selection. We will define the difference $\Delta(\cdot) \equiv (\cdot)_1 - (\cdot)_4$, where $\Delta(\cdot) > 0$ implies the $m = 4$ model is expected to be more predictive. Consider the training-sample average differences between the MLE information, the information criteria and the cross entropy difference:

| | Average information difference (nats) | | | | |
|---|---|---|---|---|---|
| *12/14 | $\Delta\overline{H}$ | $\Delta\overline{\text{QIC}}$ | $\Delta\overline{h(X\|\hat{\boldsymbol{\theta}}_X)}$ | $\Delta\overline{\text{AIC}}$ | $\Delta\overline{\text{BIC}}$ |
| | 3.73 | 2.84 | 5.29 | -0.72 | -8.53 |

In a nested model, the larger model is always favored by $\Delta h$ due to overfitting. The average cross entropy is also positive, which implies that the trained $m = 4$ model is more predictive

than the $m = 1$ trained model on average. The QIC complexity most-closely estimates the true complexity and there is the best agreement between the average cross entropy difference and average QIC. QIC also favors the $m = 4$ model. Due to the overestimate of the complexity for $m = 4$, both AIC and BIC tend to favor the smaller model.

Although QIC better estimates the true complexity on average, unlike the AIC and BIC estimates, it depends on the MLE parameter estimate and so there are statistical fluctuations in the estimated complexity. A large variance might still lead to a degradation in model selection performance, even if the mean were unbiased. We therefore compute the model selection probabilities and the expected predictive performance of model the selection criteria for AIC, BIC and QIC by computing KL Divergence, averaged over the training set:

| Performance metric | Model selection criterion AIC | BIC | QIC |
|---|---|---|---|
| $\Pr_1$ | 0.64 | 0.98 | 0.19 |
| $\Pr_4$ | 0.36 | 0.02 | 0.81 |
| $\overline{D}$ (nats) (Eqn. (3.23)) | 4.02 | 5.24 | 2.17 |

where $\Pr_m$ is the probability of selecting model $\boldsymbol{m}$, *Choose* $\boldsymbol{m}$ is a criterion where model $\boldsymbol{m}$ is always chosen. As expected, QIC has superior performance to AIC and BIC since it picks the $m = 4$ model with higher probability. Both AIC and BIC underestimate the performance of the larger model and therefore need a significantly larger dataset to justify the selection of the model family that contains the true distribution. We believe this example is representative of many systems biology problems where the complexity is significantly smaller than predicted by the model dimension alone.

### *3.6 Discussion*

Although the AIC and BIC complexities depend only on the number of parameters, the true and QIC complexities depend on the likelihood and the fitting algorithm itself. In general, the QIC complexity will not be exactly equal to the true complexity and therefore QIC

remains a biased estimator of cross entropy. In this section, we shall outline the known properties of QIC.

### *3.6.1* QIC *subsumes extends both AIC and AICc*

In comparing QIC to existing information criteria, it is first important to note that, for an important class of analyses, QIC is expected to be exactly equivalent to AIC or corrected AIC. In the large sample size limit of regular models, the frequentist complexity is equal to the AIC complexity and therefore AIC and QIC are identical. Furthermore, QIC subsumes an important class of previously proposed refinements to AIC. These complexities follow from the assumption of realizability, and the special case of parameter-invariant frequentist complexity discussed in 3.3.1. The AIC complexity is itself exact for the normal model with unknown mean(s) and known variance at any sample size. Another example is $\mathrm{AIC_C}$, derived in the context of linear least-squares regression with unknown variance. In this case the complexity is [70]:

$$\mathscr{K} = K\tfrac{N}{N-K-1}, \tag{3.34}$$

which is equal to $K$ in the large-sample-size limit ($N \to \infty$), but deviates significantly for small $N$ corrected AIC [71, 30]. Another exact result is found for the exponential model, $q(x|\theta) = \theta\, e^{-\theta x}$, where the complexity is [30]

$$\mathscr{K} = \tfrac{N}{N-1}. \tag{3.35}$$

The appealing property of these complexities is that, like AIC, they do not require knowledge of the true distribution and therefore maintain all the advantages of AIC while potentially correcting for finite-sample-size effects.

Burnham and Anderson have previously advocated the use of Eqn. (3.34) even outside the case for linear regression, on the grounds that some finite-sample-size correction is better than none [30]. The QIC complexity is a more principled approach, using the assumption of realizability without presupposing a complexity of the model. When the frequentist com-

plexity of a particular model *is* constant, QIC recovers a form of $\mathrm{AIC_C}$. When it is not, the generative parameters must be estimated using the frequentist complexity, Eqn. (3.20).

### *3.6.2 Asymptotic bias of the QIC complexity*

A canonical approach to analyzing the performance of an estimator is to study the bias of that estimator in the large-sample-size limit. An asymptotic unbiased estimator of the cross-entropy will be an asymptotically efficient model selection criteria under standard conditions (See **(author?)** [10, 127] for details). Efficiency is an important goal for predictive model selection[127, 149, 26].

QIC is not a significant improvement over AIC in terms of asymptotic bias. First, just as with AIC, we must assume that the true model is realizable (although this condition can be relaxed, see B.1.7) . If the true model is realizable $\phi \in \Theta$, then we can Taylor expand the frequentist complexity around the true parameter value:

$$\overline{\mathscr{K}(\phi + \delta\theta_X)} = \mathscr{K}(\phi) + \overline{\delta\theta}_X \cdot \nabla\mathscr{K}(\phi) + \tfrac{1}{2}\overline{\delta\theta_X \otimes \delta\theta_X} \cdot \nabla \otimes \nabla\mathscr{K}(\phi) + \ldots, \tag{3.36}$$

where the over line represents expectations with respect to $X \sim q(\cdot|\phi)$. If the estimated parameters are unbiased, the second term is zero. For nonsingular points the third term is asymptotically zero—but at non-singular points QIC is asymptotically equal to AIC. At singular points the bias due to the third term is expected to be greater than $\mathcal{O}(N^{-1})$ and QIC *will* be asymptotically biased.

However, in practice the QIC estimate of the complexity often appears to be *good enough*, and certainly superior to the alternatives. For example, the Greedy algorithm of the Fourier analysis is a useful test case. This problem is singular. The use of the AIC complexity in this problem leads to a catastrophic breakdown in model selection: The number of overfit parameters added is very large and grows with the sample size $N$. In contrast, the QIC estimate of the complexity, though biased, has the correct $\log N$ scaling behavior near the singular point: the QIC method shows excellent model selection performance in this context.

We measured the relative performance of QIC using three metrics: we compared (i)

QIC complexity to the true complexity and (ii) QIC to the cross entropy, and (iii) directly computing the KL Divergence of the trained-selected model. By all three metrics, we demonstrate that QIC outperforms AIC and BIC. We therefore conclude that, while QIC does not generically offer asymptotic efficiency when AIC does not, QIC is often vastly superior to AIC at a finite sample size, where all real analyses occur.

### *3.6.3 Advantages of QIC*

QIC has several advantages compared with existing methods. Although QIC is not universally unbiased, a good estimator should balance bias and variance—in a bias-variance tradeoff [52, 118]. QIC tends to have both relatively low bias (compared to AIC, $C_p$ and similar penalized methods) and low variance, compared to CV, bootstrap, and the Takeuchi information criterion (TIC) [28].

#### *QIC has smaller biases than AIC and similar methods*

Although QIC and AIC have similar asymptotic behavior and performance, at finite sample size, AIC will have greater bias in a cross entropy estimator, and will typically have greater predictive loss. This performance loss due to the bias of AIC can be significant [16, 26, 44], especially for small $N/\mathcal{K}$. For regular, realizable models with constant or slowly varying $\mathcal{K}(\theta)$, QIC will have negligible bias even at small sample size.

#### *QIC has smaller variance than empirical methods*

One practical method to circumvent the QIC assumption of realizability is the use of estimators depend only on empirical expectations taken with respect to the observed data (i.e. LOOCV, bootstrap, etc). Empirical estimates for the complexity such as the bootstrap methods are guaranteed to be asymptotically unbiased in a very wide range of model selection scenarios. If the sample size is large, cross-validation has highly desirable properties. However empirical methods are inferior to both AIC and AICc in the regular limit because they

suffer from a large variance resulting from the subsampling procedure [129, 30, 128, 44, 26]. This increased variance leads to degraded performance when unbiased estimators of the complexity are available. QIC therefore has provably superior performance in many situations [44].

*QIC is applicable to models of structured data such as time series*

Both LOOCV and bootstrap rely on an assumption that the data are unstructured, i.e. they take the form of independent and identically distributed random vectors. QIC can be applied, without modification, to structured data such as time series, where correlations exist between measurements. We originally developed a version of QIC in one such structured context: the change-point problem [88]. If calculations of QIC requires a Monte-Carlo calculation, data are sampled from the joint distribution, which therefore preserves the relevant dependencies in the data. In contrast, it is not as straightforward to *leave out* or *subsample* a data-point when doing Fourier analysis or DNA sequencing, although workarounds exist in specialized situations (e.g. generalized CV [36]).

*QIC responds to the effects of manifold geometry*

QIC is non-perturbative, unlike AIC and TIC, and other methods that rely on Taylor expansion. The putative distribution of $\hat{\theta}_x$ in the frequentist expectation will explore parameter space in the vicinity of the optimal value, and meet constraints and nearby singularities. Although these features usually result in QIC being biased, these biases are often small when compared with the complete failure of other methods. Two of our example applications are in singular spaces, where empirical evidence suggests that QIC is robust with complexity estimates that are accurate enough to achieve good performance.

*QIC can account for the multiplicity.*

Assuming a generative model gives QIC the ability to simulate the behavior of the entire procedure including stopping rules, outlier removal, thresholding and the fitting algorithm itself. In particular, the order in which a model family is traversed can have a profound effect on the complexity due to the multiplicity of competing models [53, 40]. These multiplicity effects are ubiquitous, and in frequentist tests they lead to Bonferonni corrections [27, 66] to the significance level. QIC automatically generates an information-based realization of the Bonferonni corrections—models with large multiplicity have substantially increased complexities. This increase in complexity lead to a much stronger preference for smaller models in the presence of multiplicity than in sequential model selection. We studied these effects in Sec. 3.5.2.

*QIC accounts for the learning algorithm*

Algorithmic dependence plays an interesting and important role in determining the complexity in some simple applications we discuss. The two approaches to the neutrino problem illustrate this point: Although both the sequential and greedy algorithms represent the intensity signal as Fourier modes, the complexities are fundamentally different as a result of the fitting algorithms. This algorithmic dependence is typical. For example, the greedy addition of regressors in linear regression problems is a common realization of a singular model that results in significant increases in complexity. QIC facilitates an information-based approach to these problems for the first time and reinforces the notion that the fitting algorithm can be of equal importance to the number of model parameters.

### *3.6.4 Conclusion*

We have proposed a new information criterion: the Frequentist Information Criterion (QIC). QIC is a significantly better approximation for the true complexity and results in better model selection performance than AIC in many typical analyses. Although, QIC is equal to AIC

in the large-sample-size limit of regular models, QIC is a superior approximation in regular models at finite sample size as well as singular model at all sample sizes and can account changes in the complexity due to algorithmic dependence. The QIC approach to model selection is objective and free from *ad hoc* prior probability distributions, regularizations, and the choice of a null hypothesis or confidence level. It therefore offers a promising alternative to other model selection approaches, especially when existing information-based approaches fail.

# Chapter 4

# FROM POSTDICTION TO PREDICTION

When we began exploring how to solve the change-point problem in Chapter 2 we saw that there were two typical options for doing model selection: AIC justified from a predictivist perspective, and BIC, which is nothing more than a Laplace (saddle-point) approximation for a Bayesian posterior weight. We found that neither method was applicable to the change point problem because of model singularity. Using a new approximation, developed further in Chapter 3, which rescued the predictive approach and gained a viable information criterion for the Change-point problem and many other singular models. The reappearance of the $\log N$ BIC penalty in a *predictive* information criteria exactly analogous to AIC suggested to us that the favoritism enjoyed by BIC in the literature was really only a crude way to correct for multiplicity.

Does this mean that we should abandon the Bayesian approach to model selection altogether? Is there some issue with the Bayesian evidence? In fact these questions have ancient history and go back to Laplace, and Bayes himself. It was the problems with Bayesianism that led to the construction of mainstream frequentist statistics in the first place.

In this chapter we able to answer these questions in new ways because i.) We had built some skepticism for the Bayesian machinery born from our understanding of the multiplicity issue and ii.) We had developed some facility in manipulating information criteria to compare different penalties. Finally iii.) we knew, to some extent, the punchline "AIC is the derivative of BIC", because we had already discovered some of the thermodynamic relations discussed in Chapter 5. A key contribution of this chapter is an effective Bayesian complexity which shows the significance of partitioning the data between prior construction and prior updating.

### *4.1 Introduction*

With advances in computing, Bayesian methods have experienced a strong resurgence. Proponents of Bayesian inference cite numerous practical and philosophical advantages of the paradigm over classical (frequentist) statistics [24]. The most compelling argument in favor of Bayesian methods is the natural hedging between competing hypotheses and parameter values. This hedging mechanism (i.e. model averaging) protects against over-fitting in singular models and has led to excellent performance in machine learning applications and many other contexts, especially those which require the synthesis of many forms of evidence [143, 24]. But the practical and philosophical problems that motivated the development of frequentist methods remain unresolved: (i) There is no commonly agreed upon procedure for specifying the Bayesian prior and (ii) statistical inference can depend strongly upon the prior. This dependence creates a discrepancy between Bayesian and frequentist methods: the Lindley paradox.

We analyze Bayesian model selection with respect to the relative partition of information between the data and the prior. This analysis leads to novel connections between Bayesian, information-based, and frequentist methods. We demonstrate that a large prior information partition results in model selection consistent with the Akaike Information Criterion (AIC) [3], while the opposite limit of the information partition results in model selection consistent with the Bayesian Information Criterion (BIC) [126]. Intermediate partitions interpolate between these well known limits. Although the AIC limit is well defined and robust, the BIC limit depends sensitively on the *ad hoc* definition of a *single measurement*. Furthermore, the BIC limit corresponds to a loss of resolution. This loss of resolution might result in the unnecessary purchase of more sensitive equipment or the collection of unreasonable sample sizes.

As a result, we question the suitability of BIC model selection (or Bayesian inference with an uninformative prior) at finite sample size. The large-prior-information regime of Bayesian inference can be achieved in almost any practical Bayesian implementation by the

use of pseudo-Bayes factors[47, 49]. This approach circumvents the Lindley paradox while maintaining many advantages of Bayesian inference.

### *4.1.1 A simple example of the Lindley paradox*

A simple example emphasizes the difference between Bayesian and frequentist forms of statistical support for models of unknown dimension. Suppose an observer measures the position of a bead in the course of a biophysics experiment. The position is first determined with negligible uncertainty. After a perturbation is applied, $N$ measurements are made of the bead position: $x^N \equiv (x_1, x_2, ..., x_N)$. The $N$ measurements are assumed to be independent and identically distributed (iid) in a normal distribution centered on the unknown true displacement $\mu$ with known variance $\sigma^2$ where $\mu = 0$ if the bead is unmoved and $\mu \neq 0$ otherwise.

In the Bayesian paradigm, we must specify priors over the parameters for the two models $\pi_i(\boldsymbol{\theta})$. Model zero (null hypothesis) is parameter free since $\mu = 0$, but model one (alternative hypothesis) is parameterized by unknown mean $\mu$. The true value $\mu_0$ is unknown and to represent this ignorance, we use a *vague* conjugate prior, choosing a normal prior centered on zero with a large variance $\tau^2$. A canonical objective Bayesian approach to model selection is to assume the competing models have equal prior probability. The model with the largest posterior probability is selected. The experimental resolution for detecting a change in the bead position is then:

$$|\hat{\mu}| > \sigma_\mu \cdot \sqrt{2 \log \tau / \sigma_\mu}, \tag{4.1}$$

while the frequentist *rule of thumb* ($\approx 95\%$ confidence level) for rejecting the null hypothesis is:

$$|\hat{\mu}| > \sigma_\mu \cdot 2, \tag{4.2}$$

where $\sigma_\mu \equiv \sigma/\sqrt{N}$ is the uncertainty in $\mu$. The difference between the conditions defined by Eqns. 4.1 and 4.2 reveals that the paradigms may come to conflicting conclusions about model selection, as illustrated in Fig. 4.1. D. Lindley emphasized this conflict by describing the following scenario: If the alternative hypothesis is true, for a suitable choice of $\tau$ and

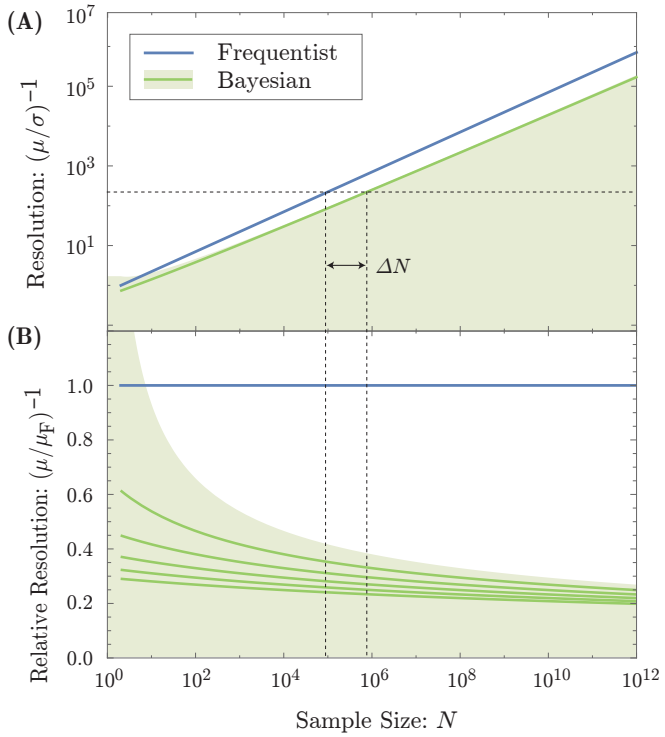


Figure 4.1: **Loss of resolution in Bayesian inference. Panel A:** The resolution on detected bead displacement (the alternative hypothesis) is plotted as a function of sample size $N$. The increase in resolution is due to the decrease in the error in the mean $\sigma_\mu = \sigma/\sqrt{N}$. The resolution of both frequentist and Bayesian inference increase, but the frequentist resolution is higher. A dotted line represents the size of a putative displacement. The frequentist analysis detects this displacement at a smaller sample size than the Bayesian analysis. **Panel B:** To draw attention to the difference between Bayesian and frequentist resolution, we plot the resolution relative to the frequentist resolution $\mu_F$. To illustrate the prior dependence of the Bayesian analysis, we have drawn curves corresponding to various choices of prior volume $V_0$.

sample size $N$, the null hypothesis could be simultaneously (i) rejected at a 95% confidence level *and* (ii) have 95% posterior probability [97]! This conflict between statistical paradigms

has been called the Lindley paradox.

Many practitioners of Bayesian inference believe that priors may be a formal necessity but have minimal influence on inference. For instance, the posterior probability for $\mu$ is independent of the prior in the uninformative limit $\tau \to \infty$. However, as we see in Eqn. 4.1, inference on model identity remains critically dependent on the prior (value of $\tau$). In the limit that $\tau \to \infty$, no finite observed displacement $\hat{\mu}$ is sufficient to support the alternative hypothesis that the bead has moved! This paradoxical condition is called the Bartlett paradox [17].

## *4.2 Data partition*

### *4.2.1 The definition of frequentism and Bayesian paradigms*

We wish to study a generalized class of decision rules that include methods from all three paradigms of inference. In the current context, we will use the log likelihood ratio:

$$\lambda(x^N) \equiv h_0(x^N|\hat{\boldsymbol{\theta}}_x) - h_1(x^N|\hat{\boldsymbol{\theta}}_x), \tag{4.3}$$

as a *frequentist test statistic* where $h$ is the Shannon information $h \equiv -\log q$ and $\hat{\boldsymbol{\theta}}_x$ is the maximum likelihood estimate of the parameters of the respective model. We shall define a decision rule:

$$\lambda(x^N) < \lambda_*, \tag{4.4}$$

to select model zero where $\lambda_*$ is the critical value of the test statistic. We will refer to the decision rule as *frequentist* if $\lambda_*$ is sample-size independent in the large-sample-size limit of a regular model. This definition includes both the frequentist Neyman-Pearson likelihood ratio test as well as the information-based paradigm (AIC). In the Bayesian paradigm, we will define the decision rule in terms of the log-Bayes factor:

$$\lambda_B(x^N) \equiv h_0(x^N) - h_1(x^N), \tag{4.5}$$

where $q(x^N)$ is the marginal likelihood and $h(x^N)$ is the respective Shannon information. We define the decision rule: $\lambda_B(x^N) < 0$ to select model zero. Although the Bayes factor

is not a test statistic—an orthodox Bayesian approach is to compute a posterior on model identity—the decision rule captures how the Bayes factor is typically used in practice.

In the large-sample-size limit, the Bayesian decision rule is equivalent to Eqn. 4.4 with $\lambda_*$ proportional to $\log N$ to leading order in $N$. Therefore, we will define a decision rule Eqn. 4.4 as *Bayesian* if the critical test statistic $\lambda_*$ is sample-size dependent. This definition includes standard Bayesian model selection as well as the Bayesian information criterion (BIC).

### *4.2.2 Prior information content*

The paradoxically-large displacement needed to select the alternative hypothesis is a consequence of the uninformative prior ($\tau \to \infty$). To be more precise about the descriptors *informative* and *uninformative*, we can compute expected-parameter-information content of the data set $x^N$ [98]:

$$I(x^N) \equiv \mathop{\mathbb{E}_{\boldsymbol{\theta}}}_{\pi(\cdot|x^N)} \log \pi(\boldsymbol{\theta}|x^N)/\pi(\boldsymbol{\theta}), \tag{4.6}$$

which is equal to the KL Divergence of the posterior and prior. $I \geq 0$ and will increase with sample size. Given $N$ new measurements, we call the prior $\pi$ *uninformative* if $I$ is large.

A standard approach to specify an informative prior is the elicitation of a prior from an expert [24]. It is convenient to make the concrete assumption that the expert knowledge is the result of previous measurements, which we can write explicitly $x^{N_T}$. Our posterior on these measurements $\pi(\boldsymbol{\theta}|x^{N_T})$ is computed from some suitably flat prior $\pi(\boldsymbol{\theta})$. The $x^{N_T}$ is then used to construct a new informative prior:

$$\pi'(\boldsymbol{\theta}) \equiv \pi(\boldsymbol{\theta}|x^{N_T}), \tag{4.7}$$

where the primed distributions are computed with respect to the informative prior (Eqn. 4.7). This Bayesian update rule was concisely summarized by D. Lindley: *Today's posterior is tomorrow's prior.*

Let the new measurements made be re-labeled $x^{N_G}$. We can re-compute the marginal likelihood $q'$ using the new prior $\pi'$. $q'$ has a second interpretation, the Bayesian predictive

distribution computed from the original prior $\pi$:

$$q'(x^{N_G}) = q(x^{N_G}|x^{N_T}) \equiv q(x^N)/q(x^{N_T}), \tag{4.8}$$

where $x^N$ represents the entire data set of $N = N_G + N_T$ measurements. This distribution is predictive since it predicts or *generalizes* on data set $x^{N_G}$ given a *training* data set $x^{N_T}$. Adjustment of the data partition between the training set (size $N_T$) and the generalization set (size $N_G$) can be understood as adjusting the information content of the prior. If $N_G \gg N_T$, the prior is uninformative relative to the data.

### *4.2.3 The Bayesian cross entropy*

The general problem of predicting observations $x^{N_G}$, conditioned on $x^{N_T}$ where $N = N_G + N_T$ is closely related to a natural metric of performance: a predictive *cross entropy* [30]

$$H^{N_G|N_T} \equiv \tfrac{N}{N_G} \underset{p(\cdot)}{\mathbb{E}_X} h(X^{N_G}|X^{N_T}), \tag{4.9}$$

where $p(x^N)$ is the true distribution of observations $x^N$. The cross entropy is rescaled to correspond to the total sample size $N$. We can view model inference using the evidence Eq. 4.8 as choosing the model which is estimated to have the optimal performance under this metric. Since $H$ can only be computed if the true distribution $p$ is known, it will be useful to empirically estimate it. A natural estimator is the leave-k-out estimator [48]

$$\hat{H}^{N_G|N_T}(x^N) \equiv \tfrac{N}{N_G} \underset{P\{x^N\}}{\mathbb{E}_X} h(X^{N_G}|X^{N_T}), \tag{4.10}$$

where $\hat{H}$ estimates $H$ and the empirical expectation $\mathbb{E}$ is taken over all unique permutations of the observed data between the training and generalization sets.

This estimator uses *cross validation*: there is no double use of data since the same observations never appear in both the generalization and training sets. Methods like empirical Bayes [102, 103, 2], where the prior is fit to the data to maximize the evidence, implicitly use the data twice and are therefore subject to the same over-fitting phenomenon as maximum likelihood estimation.

### *4.2.4 Pseudo-Bayes factors*

The natural strategy would be to compute the model posterior probability (or Bayes factor) using the evidence $q'(x^{N_G})$. But, for small $N_G$, $h(x^{N_G}|x^{N_T})$ typically exhibits large statistical fluctuations since only a small fraction of the data $x^{N_G}$ is used for inference on the model identity even though there is more non-redundant information encoded in $x^{N_T}$. To re-capture this missing information, we replace $h(x^{N_G}|x^{N_T})$ with $\hat{H}^{N_G|N_T}$. Therefore, in analogy with the log Bayes factor, the log-*pseudo-Bayes factor* is defined [48]:

$$\lambda_{\mathrm{PB}}^{N_G|N_T}(x^N) \equiv \hat{H}_0^{N_G|N_T} - \hat{H}_1^{N_G|N_T}, \tag{4.11}$$

which depends on the data partition $N_G$ and $N_T$. We define the decision rule: $\lambda_{PB}(x^N) < 0$ to select model zero.

Two data partitions have been discussed in some detail. A maximal-traning-set limit, where $N_T = N - 1$ and $N_G = 1$, corresponds to Leave-one-out cross validation (LOOCV) and has been studied extensively [48, 47, 49, 140, 132]. A mininimal-training-set limit has also been explored in which $N_T$ is as small as possible such that $\pi'$ is proper [12, 131].

We focus on the example of a pairwise model selection to compare with canonical frequentist inference, but a selection among any number of models can be performed by selecting the model with the smallest cross-entropy estimator.

### *4.2.5 Information Criteria*

To systematically investigate the dependence of inference on the data partition in the pseudo-Bayes Factor, we propose a novel estimator of the cross entropy $H$ whose dependence on the data partition is explicit. The data partition will be parameterized by $\nu \equiv N_G/N_T$. We define a generalized Information Criterion:

$$\mathrm{IC}^{\nu}(x^N) \equiv h(x^N|\hat{\boldsymbol{\theta}}_x) + \mathscr{K}_{\nu}, \tag{4.12}$$

where the complexity $\mathscr{K}_{\nu}$ is the bias, chosen to make $\mathrm{IC}^{\nu}$ an unbiased estimator of $H^{N_G|N_T}$. The log-pseudo-Bayes Factor can be constructed using the information criterion. The infor-

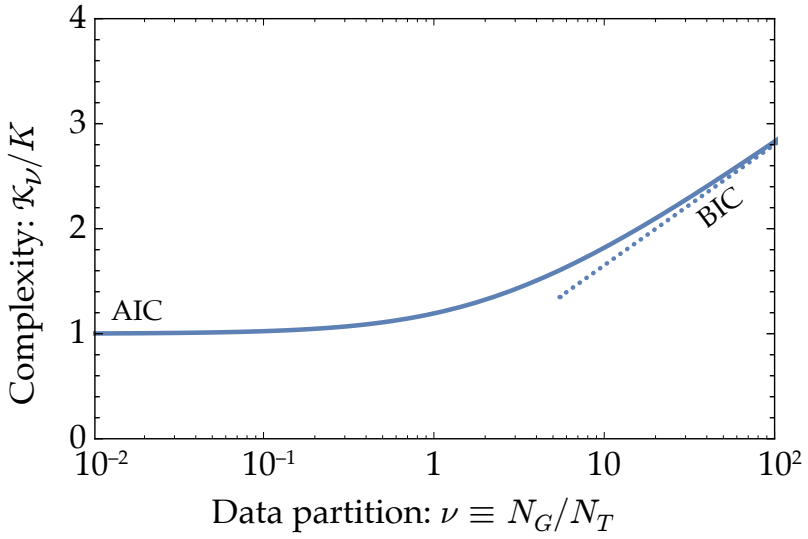


Figure 4.2: **Complexity as a function of data partition.** Complexity can be understood as a penalty for model dimension $K$. The data partition parameter controls the relative amount of information in the prior. In a predictive limit ($\nu \to 0$), the training set is large compared with the generalization set and the complexity is small. This is the AIC limit. At the other extreme ($\nu \to \infty$), all the data is partitioned into the generalization set and therefore the prior is uninformative and the complexity is large. This is the BIC limit.

mation criterion is typically much easier to evaluate than the leave-k-out formulation. Since the first term in the definition of $\mathrm{IC}^\nu$ is independent of $\nu$, the data-partition dependence is completely characterized by the complexity $\mathscr{K}_\nu$.

Assuming $\pi$ is uninformative and $q(x^N|\boldsymbol{\theta})$ is a regular model in the large sample size limit, the Laplace approximation holds and the complexity has a simple form:

$$\mathscr{K}_\nu = \tfrac{1}{2} K \left[1 + (1+\nu^{-1})\log(1+\nu)\right], \tag{4.13}$$

which is only a function of the parameter-space dimension $K$ and the data partition $\nu$. The complexity is plotted as a function of the data partition $\nu$ in Fig. 4.2.

### *4.2.6 Decision rules and resolution*

With the information criterion above, we can connect a (pseudo-)Bayes factor with a particular data partition $\nu$ to an effective decision rule. We choose model one if

$$h_0(x^N|\hat{\boldsymbol{\theta}}_x) - h_1(x^N|\hat{\boldsymbol{\theta}}_x) > \Delta\mathscr{K}_\nu, \tag{4.14}$$

where $\Delta\mathscr{K}_\nu$ is the difference in the complexity of the models. We can also connect these decision rules to choices of a frequentist significance level as described in the supplement. A plot of this function for two different values of $\Delta K$ is shown in Fig. 4.5. Of particular practical experimental importance is the minimal signal to noise ratio at which our decision rule will choose a larger model. Returning to the biophysical problem described in the introduction, the minimal resolvable bead displacement is

$$|\hat{\mu}| > \sigma_\mu \cdot \sqrt{1 + (1+\nu^{-1})\log(1+\nu)} \tag{4.15}$$

where the RHS is the inverse resolution. The resolution is monotonically decreasing in $\nu$. The smallest $\nu$ gives us the highest resolution.

## *4.3 The Lindley paradox*

The Lindley paradox can be understood from the perspective of the relative partition of information between the prior and the data. By defining the complexity (Eqn. 4.13), we

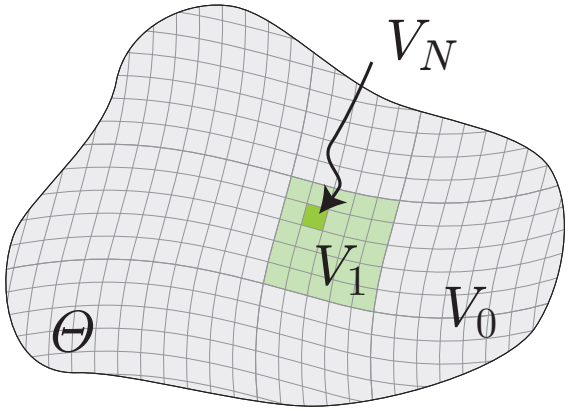


Figure 4.3: **The geometry of the Occam factor.** The total volume of plausible parameter values for a model is $V_0$. The volume of acceptable parameter values after a single measurement is $V_1$. The volume of acceptable parameter values after $N$ measurements is $V_N$. The Occam factor is defined as the probability of randomly drawing a parameter from initial volume $V_0$ consistent with the $N$ measurements: $\Pr \approx \mathcal{N}^{-1}$ where $\mathcal{N} \equiv V_0/V_N$ is the number of distinguishable distributions after $N$ measurements. Lower dimension models are naturally favored by the Occam factor since the number of distinguishable models $\mathcal{N}$ is smaller.

can explore the partitioning of this data by studying the decision rule and resolution as a function of the partition $\nu$.

### *4.3.1 Classical Bayes and the Bartlett paradox*

For the classical Bayes factor, $N_T = 0$ or $\nu \to \infty$. If the prior is flat on an infinite-volume parameter manifold, the complexity $\mathcal{K}_\nu$ becomes infinite. This scenario always favors the smaller model, regardless of the goodness of fit, resulting in the Bartlett paradox.

If the parameter-manifold volume $V_0$ is finite, so is the complexity. In the large sample size limit, the marginal likelihood can be written in an intuitive form:

$$q(x^N) = \tfrac{V_N}{V_0} \times q(x^N|\hat{\boldsymbol{\theta}}_x), \tag{4.16}$$

where $V_N$ is the volume of parameter manifold consistent with the data $x^N$ and $\hat{\boldsymbol{\theta}}_x$ is the maximum likelihood estimate of the parameters. (We define this volume more precisely in the supplement.) The first factor on the RHS is the Occam factor or the probability of randomly drawing a parameter (consistent with $x^N$) from the prior distribution $\pi$. Complex models (large $K$) with uninformative priors have *small* Occam Factors, due to the large volume of plausible parameters ($V_0$), relative to the volume of the parameter manifold consistent with the observations ($V_N$). Large Occam factors give rise to a natural mathematical realization of the Occam Razor: *Among competing hypotheses, the one with the fewest assumptions [parameters] should be selected* [102]. This effect is illustrated schematically in Fig. 4.3. Both infinite and finite-but-large-volume parameter manifolds can give rise to strong Lindley paradoxes.

### *4.3.2 Minimal training set and Lindley Paradox*

We might use a *minimal training set* to remove the dependence on the potentially divergent volume $V_0$ [22] which corresponds to the large-data-partition limit: $\nu \gg 1$. It is difficult to define this minimal training set in a satisfactory way [114]. The most natural option is to set $N_T = 1$ and $N_G = N - 1$, which results in the Bayesian information criterion (BIC) [126]:

$$\mathrm{BIC}(x^N) \equiv h(x^N|\hat{\boldsymbol{\theta}}_x) + \tfrac{K}{2}\log N, \tag{4.17}$$

in the large-sample-size limit. We can compute a limit on the smallest resolvable change in position:

$$|\hat{\mu}| > \sigma_\mu \cdot \sqrt{\log N}, \tag{4.18}$$

which is free from the *ad hoc* volume $V_0$ of the uninformative prior. This approach resolves the Bartlett paradox, but leads to a strong Lindley paradox—conflict with Frequentist methods in some critical range of sample sizes.

The $\log N$ dependence of BIC results in some troubling properties. If we now bin pairs of data points, the empirical mean $\hat{\mu}$ and the standard error $\sigma_\mu$ are unchanged, but $N \to N/2$, changing the complexity and therefore the decision rule and resolution. Therefore, although

BIC does not depend on the choice of prior support, it does depend on an *ad hoc* choice as to what constitutes a single sample.

### *4.3.3 Frequentist prescription and AIC*

The complementary limit describes a maximal training set ($\nu \to 0$). In this limit, $\mathrm{IC}^{\nu}$ corresponds to the Akaike Information Criterion (AIC):

$$\mathrm{AIC}(x^N) \equiv h(x^N|\hat{\boldsymbol{\theta}}_x) + K, \tag{4.19}$$

where the complexity is equal to the dimension of the parameter manifold $K$. This leads to a sample-size-independent critical value of the test statistic in Eqn. 4.4, and is therefore *frequentist*. Like the log-Occam factor, $K$ can be reinterpreted as a penalty for model complexity that gives rise to a distinct information-based realization of the Occam Razor: *Parsimony implies predictivity.*

The smallest resolvable change in position using AIC is

$$|\hat{\mu}| > \sigma_{\mu} \cdot \sqrt{2}. \tag{4.20}$$

AIC can also be viewed as the performance of the model against the next observation $x_{N+1}$ [30]. For this reason, $\nu \to 0$ is called the *predictive limit*. The canonical Bayesian approach estimates the marginal likelihood of the observed data $x^N$ from the prior. It is therefore postdictive. We therefore call $\nu \to \infty$ the *postdictive limit*. The AIC and BIC penalties are often used as complimentary heuristics in model selection [136]. Our cross entropy description shows that they can be interpreted as Bayesian objects which differ only in the choice of data partition $\nu$.

The predictive limit is expected to result in maximum resolution and consistent inference (*i.e.* independent of the prior and data partition). Unlike BIC, it is essentially independent of data binning in the large sample size limit. (See Fig. 4.2.) Although AIC is computed using a point estimate, a pseudo-Bayes factor for $N_G = 1$ and $N_T = N - 1$ (*i.e.* LOOCV) corresponds to the same predictive limit.

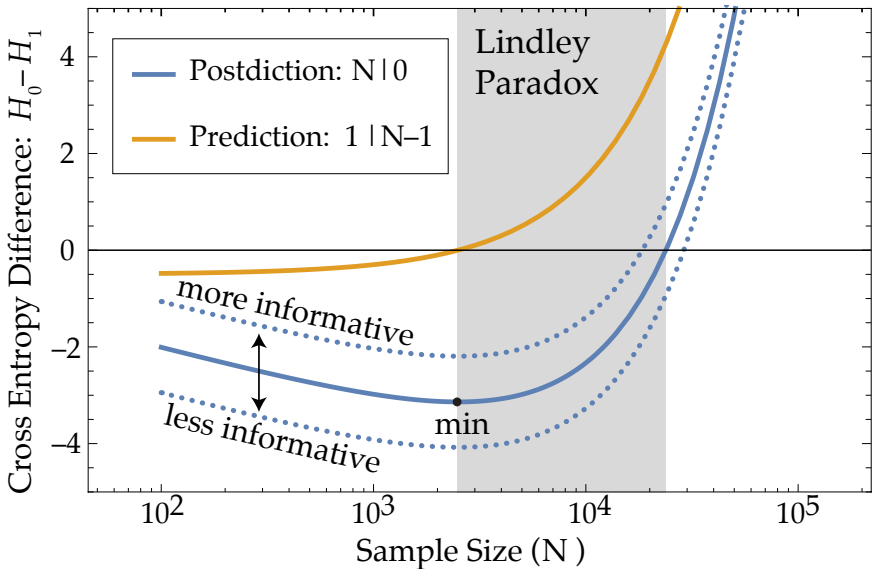


Figure 4.4: **Visualizing the pre and postdictive decision rules.** The cross entropy difference for Bayesian inference (postdiction: $N|0$) and the predictive limit ($1|N-1$) are plotted as a function of sample size $N$. $\Delta H > 0$ results in the selection of the alternative hypothesis. Both measures initially favor the null hypothesis. The use of a more (less) informative prior causes the postdictive curve to be shifted up (down). Since the predictive $H$ is the derivative of the postdictive $H$, the prior does not influence the inference of the predictive observer. The predictive curve crosses zero first, leading the predictive observer to support the alternative hypothesis. Since the predictive $H$ is the derivative of the postdictive $H$ with respect to $N$, the sample size at which the predictive observer switches to the alternative hypothesis corresponds to the sample size where the postdictive observer has the most evidence for the null hypothesis. The two measures are in conflict (grey region) until the postdictive $H$ crosses zero at a significantly larger sample size $N$. The Bayesian therefore requires significantly more evidence to reach the same conclusion as the predictive observer.

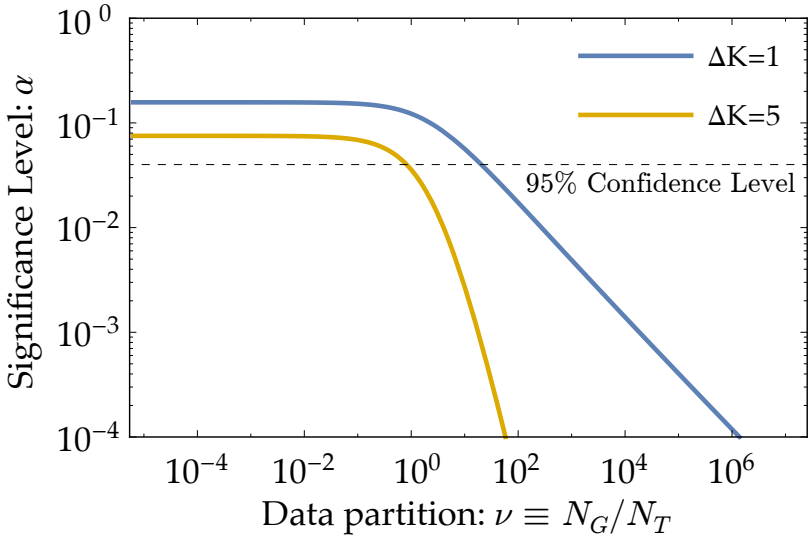


Figure 4.5: **Significance level as a function of data partition.** To make an explicit connection between the frequentist significance level and the data partition, it is useful to compute the significance level implied by the predictive decision rule. In the postdictive regime, corresponding to an uninformative prior, the significance level steeply decreases with increasing $\nu$, resulting in a strong Lindley paradox.

### *4.3.4 Log evidence versus AIC*

A convenient heuristic for understanding the relation between AIC and the log evidence can be understood from the relation between the cross entropy using $H^{1|N-1}$ and $H^{N|0}$ which are estimated by AIC and $h(x^N)$ respectively. If we approximate the finite difference in $H^{1|N-1}$ as a derivative in the large $N$ limit, $H^{1|N-1}$ can be approximated:

$$H^{1|N-1} = N\partial_N H^{N|0} + \mathcal{O}(N^{-1}). \tag{4.21}$$

Therefore, we can understand the relation between the conflicting information criteria AIC and BIC in the following way: AIC is the derivative of BIC.

This heuristic can naturally explain why AIC is free from the strong prior dependence which leads to the Lindley paradox. In the context of an uninformative prior, the expected log evidence $H^{N|0}$ has an ambiguous offset corresponding to the prior choice, leading different individuals to make different inference on the model identity. $H^{1|N-1}$, estimated by AIC, is independent of the unknown constant since the slope of $H^{N|0}$ is independent of its offset. This relationship is illustrated schematically in Fig. 4.4.

A second interesting feature of the heuristic relates to the sample size dependence of the predictive and postdictive decision rules. The sample size at which the predictive statistician begins to favor the alternative hypothesis corresponds to the same sample size at which the postdictive statistician has maximum confidence in the null hypothesis! (See Fig. 4.4.) The difference between a function and its derivative explains both the connection and inconsistency of the predictive and postdictive decision rules.

### *4.3.5 When will a Lindley paradox occur?*

We stated that the Lindley paradox is a consequence of an insufficiently informative prior, but we have studied differences in the performance of predictive and postdictive decision rules. We now discuss the connection between these equivalent formulations. Let us define the difference between the predictive and postdictive cross entropy:

$$I'(x^N) \equiv \hat{H}^{1|N-1} - \hat{H}^{N|0}. \tag{4.22}$$

In the large-sample-size limit, we can express $I'$ in terms of the expected-parameter-information content of the data $I$:

$$I' = I(x^N) + \mathcal{O}(N^0), \tag{4.23}$$

as shown in the supplement. $I'$, the mismatch between pre and postdictive measures of performance, can be interpreted as the *missing information* from an uninformative prior $I$. The missing information is only *missing* before sample $x^N$ is observed. A model may be extremely predictive, even if the missing information was infinite, once $x^N$ has been observed.

## *4.4 Discussion*

By defining a novel information criterion that estimates the cross-entropy, we established a continuous bridge between canonical Bayesian and information-based model selection defined in terms of a data-partition between training and generalization data sets. The strength of the Lindley paradox, the mismatch between Bayesian and frequentist inference on hypotheses, can be re-interpreted in terms of prior information content (*i.e.* the data partition). We studied the properties of model selection with respect to the data partition. Two solutions to the Lindley paradox have been widely discussed: (i) adapt the frequentist paradigm by making the significance level sample-size dependent [20] or (ii) adapt the Bayesian paradigm by making the prior sample-size dependent. We advocate for taking the second approach.

### *4.4.1 A canonical Bayesian perspective on the Lindley paradox*

It is important to acknowledge that the Bayesian perspective on the Lindley paradox is valid. Returning to the biophysical example, if we interpret the alternative hypothesis *precisely*, we define a uniform prior probability density over an infinite-volume manifold in the uninformative limit ($\tau$ and $V_0 \to \infty$). Therefore, the *a priori* probability of picking a displacement consistent with the data ($\approx V_N/V_0$) is vanishingly small in the alternative hypothesis. *Fine tuning* would be required to make $\hat{\mu}$ finite and therefore the null hypothesis is strongly favored, whatever $\hat{\mu}$.

In this context, the Bayesian perspective is correct and intuitive. This approach is useful in many contexts where we have a precisely defined alternative hypothesis. However, this interpretation of the alternative hypothesis is *not* what the authors intended. Although we wished (i) *to allow a large range of putative parameter values*, we also unintentionally specified a corollary: (ii) *a vanishingly small prior density on the parameter manifold*. In our conception of the statistical problem, we are *not* interested in testing any *precise model* for the distribution of $\mu$ (*e.g.* diffusion, stage drift, etc) as a requisite for determining whether the bead movement can be detected. If possible, we wish to achieve condition (i) without the corollary (ii). The predictive formulation of inference can achieve this goal. The vanishingly small prior density subtracts out of the predictive cross entropy as illustrated in Sec. 4.3.4.

### *4.4.2 Circumventing the Lindley paradox*

By partitioning the data into a training and generalization partition in the pseudo-Bayes factor, we are able to circumvent the most severe forms of the Lindley paradox by generating inference that is prior independent (for sufficiently large sample sizes). The postdictive limit depends sensitively on the data partition, but the predictive limit does not. Fig. 4.2 shows that for sufficiently large sample size $N$, the complexity rapidly converges to its limit as $\nu \to 0$ for $N_G < N_T$. Due to this convergence, two researchers will report the same predictive pseudo-Bayes factor, even if they make different decisions about the prior and the data partition.

Our discussion of the Lindley paradox focusses mainly on critiques of a *Bayesian* perspective and the defense of a *frequentist* perspective on hypothesis testing or model selection. In fact the *frequentist* perspective we discuss includes methods from all three paradigms of inference. Our criticism of the Bayesian paradigm is confined strictly to a criticisms of the use of Bayes factors and their undesirable consequences on model selection, as described above. However, the Bayesian paradigm offers many strengths. The posterior is an elegant and intuitive framework for representing parameter uncertainty. Furthermore, hedging between parameter values (and models) typically leads to superior frequentist performance relative

to point estimates. Finally, the Bayesian paradigm offers a coherent framework for combining different types of data. Therefore we advocate retaining as many of these advantages as possible while eliminating paradoxical behavior in the context of model selection. The pseudo-Bayes factor has these desired properties.

Predictive methods (the information-based paradigm and predictive pseudo-Bayes factor) also circumvent many criticisms of the classical frequentist procedure: (i) Observed data that is unlikely in both the alternative and null hypothesis results in the rejection of the null hypothesis. (ii) An *ad hoc* confidence level must be supplied. (iii) Only pairwise comparisons between models can be made. (iv) A null hypothesis must be defined. The predictive approach circumvents each of these criticisms. Predictive methods also have provable asymptotic efficiency in terms of cross entropy loss in typical modeling situations[127], a feature which we discuss in the supplement.

### *4.4.3 Loss of resolution*

To place the discussion of experimental resolution in context, it is useful to remember that biophysicists will routinely pay thousands of dollars more for a 1.49 NA versus a 1.4 NA objective with nominally a 6% increase in the signal-to-noise ratio. This obsession with signal-to-noise ratio might suggest that a similar effort would be expended to optimize the resolution of the experimental analysis to exploit the data as efficiently as possible, especially in the context of single-molecule experiments where the sample size is often extremely limited. The Bayesian formulation of inference can imply a prohibitively stringent significance level for the discovery of new phenomena. The frequentist formulation of inference is tuned for discovery in the sense that it explicitly controls for the largest acceptable false positive probability. The Bayesian may require a much larger sample size to detect the same phenomena, as illustrated in Figs. 4.1, 4.4 and 4.5.

### *4.4.4 The multiple comparisons problem*

We have demonstrated that predictive inference has a lower threshold for discovery, but proponents have argued that the loss of resolution is in fact a feature rather than a flaw of the Bayesian paradigm. There is a perception that the canonical frequentist significance test is too weak and leads to spurious claims of discovery. An important and subtle problem with Frequentist significance testing is the multiple comparisons problem (multiplicity). For instance, if 20 independent false hypotheses for tumor genesis were independently tested at a 95% confidence level, one would expect spurious support for one of these hypotheses. Multiplicity can arise in more subtle contexts: Hypotheses (or priors) are modified after some results are known in the course of research, often unconsciously. In singular statistical models, there is often implicit multiplicity in the maximum likelihood estimation procedure [143, 88, 86]. The peer-review process itself may favor the most extreme results among multiple competing articles. In exact analogy to the tumor genesis example, multiplicity can result in the spurious selection of the alternative hypothesis in each of these cases.

These false discoveries are a consequence of using an incorrect frequentist significance test [42]. For instance, we have described how the complexity in information-based inference must be modified in the context of a singular model [88, 86]. (*e.g.* [73]). From a frequentist perspective, the significance test must reflect the presence of multiple alternative hypotheses which leads to corrections (*e.g.* Bonferonni correction [42]). These corrections increase the critical test statistic value to reflect the true confidence level of the test in the context of multiple alternative hypotheses. In summary, the failure of frequentist methods due to un-corrected multiplicity is not a flaw in the frequentist paradigm but rather a flaw in its application. Bayesian inference can naturally circumvent some of these problems in a principled way, but in many applications there are parameters for which one must supply an uninformative prior. As a result, the effective confidence level is *ad hoc*. If multiplicity is the source of spurious false discoveries, a principled approach is to correct for this problem explicitly.

### *4.4.5 Statistical significance does not imply scientific significance*

Simpler models are often of greater scientific significance. Therefore there is a perception that frequentism is flawed because it typically assigns higher statistical significance to larger models, relative to the Bayesian paradigm. This perception conflates statistical and scientific significance. Almost all natural systems appear to be described by models with a clear hierarchy of effect sizes [101]. Scientific progress is achieved by studying the largest effects first, irrespective of the statistical significance of smaller effects. The selection of effects to include in a model is a matter of judgment and scientific insight. There are important non-statistical systematic sources of error that must be considered. If sample size is large enough, these systematic effects will suggest the use of a larger model from a predictive standpoint, even if the larger model is not scientifically relevant [105]. Statistics supplies only a lower bound on scientific significance by determining whether a hypothetical effect can be explained by chance.

### *4.4.6 Conclusion*

Bayesian inference can be powerful in many contexts, especially in singular models and in the small-sample-size limit where point estimates are unsuitable [143]. But Bayesian inference can result in strong conflict with frequentist inference when uninformative priors are used. When a Bayesian analysis is desired, we advocate the pseudo-Bayes factor method [48] for inference on model identity with a small ratio $\nu$ of generalization to training set sample size. We demonstrate that only in this predictive limit can inference be expected to be consistent between independent analyses. This approach is fully Bayesian for parameter inference, but free from the Lindley paradox. Therefore it preserves all of the advantages of Bayesian methods without the risk of paradoxical inference on model identity and optimizes experimental resolution.

# INTERLUDE

The continued disagreement between "Bayesians" and "frequentists" is something of a religious war. To stretch the simile, the battlefront for this religious war is model selection and the battle line is the Lindley-Bartlett paradox. However a careful examination of the issue shows that neither paradigm is strictly wedded to one horn of the paradoxes or the other. In fact there are two distinct situations for doing model selection:

1. when we wish to test a hypothesis about how the parameters might have been distributed, and

2. when we wish to choose the most predictive model given the data we have.

The first is certainly more naturally treated using the Bayesian framework, but neither question is strictly Bayesian or Frequentist! The frequentist is allowed to use the Bayes rule to calculate test thresholds in the first situation, the Bayesian is allowed to partition the data in such a way that the posterior measures prediction in the second situation. Our conclusion can be stated simply: in most modeling situations, a true prior of interest is rare and only the predictive perspective, AIC, not BIC, makes sense.

Still there is something weird about the pseudo-Bayes factors. There is this "average over permutations" which is inherently non-Bayesian and yet absolutely necessary. A clear explanation for why are we allowed to replace the Bayes factor with the pseudo-Bayes factor remains missing. We attempted to close this gap by using the freedom inherent to the Bayes approach, the prior, to correct the classical Bayes procedure. That is, make the classical Bayesian evidence *predictive*. This was our initial motivation for the the thermodynamic approach which follows in Chapter 5: unification of the post and predictive methods. It soon became clear that the implications of this thermodynamic approach were perhaps even

more interesting in their own right, providing new insight into the learning process.

# Chapter 5

# THERMODYNAMIC CORRESPONDENCE

Despite significant advances in learning algorithms, fundamental questions remain about the mechanisms of learning and the relationships between learning algorithms. How does model complexity affect learning? Why do some models have anomalously good learning performance? We explore the phenomenology of learning by exploiting a correspondence between Bayesian inference and statistical mechanics. This correspondence has been previously described by Jaynes, Balasubramanian, and many others [76, 75, 15, 14, 113, 109] and many methods from statistical physics have been adapted to statistics [38, 67, 33, 110, 148, 111, 112, 96, 72, 108, 151]. We extend this correspondence by using the canonical bridge between statistical mechanics and thermodynamics to compute the standard thermodynamic potentials and properties of a learning system. The correspondence identifies two novel statistical quantities, a learning capacity and the Gibbs entropy. These quantities generate new insights into the mechanism of learning and new learning algorithms.

## *5.1 Contributions*

The analysis of a novel *learning capacity* (corresponding to the heat capacity) reveals an interesting connection between the Akaike Information Criterion (AIC) of information-based inference and the equipartition theorem in statistical mechanics [117]. In addition, the learning capacity also provides new insights into the mechanism of learning. It has long been known that some high-dimensional models learn anomalously well. These models have been termed *sloppy* [101]. We demonstrate that the learning capacity both provides a natural definition for the sloppiness phenomenon as well as providing a mechanism: a statistical analogue of the well known freeze-out mechanism of statistical mechanics. We hypothesize

| **Thermodynamics** | | | **Statistics** | |
|---|---|---|---|---|
| **Quantity:** | **Interpretation:** | | **Quantity:** | **Interpretation:** |
| $\beta = T^{-1}$ | Inverse temperature | $\leftrightarrow$ | $N$ | Sample size |
| $\boldsymbol{\theta}$ | State variables/vector | $\leftrightarrow$ | $\boldsymbol{\theta}$ | Model parameters |
| $X^N$ | Quenched disorder | $\leftrightarrow$ | $X^N$ | Observations |
| $E_X(\boldsymbol{\theta})$ | State energy | $\leftrightarrow$ | $\hat{H}_X(\boldsymbol{\theta})$ | Cross entropy estiamtor |
| $E_0$ | Disorder-averaged ground state energy | $\leftrightarrow$ | $H_0$ | Shannon entropy |
| $\rho(\boldsymbol{\theta})$ | Density of states | $\leftrightarrow$ | $\varpi(\boldsymbol{\theta})$ | Prior |
| $Z$ | Partition function | $\leftrightarrow$ | $Z$ | Evidence |
| $Z^{-1}\rho\ \exp -\beta E_X$ | Normalized Boltzmann weight | $\leftrightarrow$ | $\varpi(\boldsymbol{\theta}\|X^N)$ | Posterior |
| $F = -\beta^{-1}\log Z$ | Free energy | $\leftrightarrow$ | $F = -N^{-1}\log Z$ | Minus-log-evidence |
| $U = \partial_\beta \beta F$ | Average energy | $\leftrightarrow$ | $U = \partial_N NF$ | Minus-log-prediction |
| $C = -\beta^2\partial_\beta^2 \beta F$ | Heat capacity | $\leftrightarrow$ | $C = -N^2\partial_N^2 NF$ | *Learning capacity* |
| $S = \beta^2\partial_\beta F$ | Gibbs entropy | $\leftrightarrow$ | $S = N^2\partial_N F$ | *Statistical Gibbs entropy* |

Table 5.1: **Thermodynamic-Bayesian correspondence.** The top half of the table lists the correspondences that can be determined directly from the definition of the marginal likelihood as the partition function. The lower half of the table lists the implied thermodynamic expressions and their existing or proposed statistical interpretation.

that this mechanism is responsible for the anomalously high predictive performance of many high-dimensional models.

We also propose that the Gibbs entropy provides a natural device for determining model multiplicity, *i.e.* counting indistinguishable distributions in the context of statistical inference. This interpretation allows us to define a *generalized principle of indifference* (GPI) for selecting a prior in absence of *a priori* information about parameters or models. The GPI unifies a number of known, but seemingly unconnected objective Bayesian methods, while also providing an algorithm applicable to small sample size and singular models where existing approaches fail. The GPI also resolves a number of troubling anomalies in objective Bayesian analysis, providing a natural resolution to the Lindley-Bartlett paradox in which larger models are automatically rejected.

The paper is organized as follows: In Sec. 5.2, we define the correspondence. In Sec. 5.3,

we compute the thermodynamics potentials and properties of inference in the analytically tractable large-sample-size limit and use these results to deduce the statistical meaning of each quantity. In Sec. 5.4, we compute the learning capacity and GPI prior for a number of example analyses to demonstrate that the results in Sec. 5.3 generalize beyond the normal model.

### *5.2 Defining the correspondence*

We assume that a true parameter value $\boldsymbol{\theta}_0$ is drawn from a known prior distribution $\varpi(\boldsymbol{\theta})$. We observe $N$ samples $x^N \equiv \{x_1, ..., x_N\}$ which are distributed like $q(x|\boldsymbol{\theta}_0)$:

$$X_i \sim q(\cdot|\boldsymbol{\theta}_0), \tag{5.1}$$

where we use capital $X$ to denote random variables and the symbol $\sim$ to denote *distributed like*. For simplicity, we will assume that the observations are independent and identically distributed, but the approach can be generalized.

The correspondence between statistical physics and Bayesian inference is clearest when expressions are written in terms of the empirical estimator of the cross-entropy (Eqn. D.2):

$$\hat{H}(\boldsymbol{\theta}) \equiv -\left\langle \log q(X|\boldsymbol{\theta}) \right\rangle_{X \in x^N}, \tag{5.2}$$

where the angle brackets represent the empirical expectation (Eqn. D.4). The marginal likelihood (*i.e.* evidence) can be written [14]:

$$Z(x^N) \equiv \int_{\boldsymbol{\Theta}} \mathrm{d}\boldsymbol{\theta}\ \varpi(\boldsymbol{\theta})\, e^{-N\hat{H}}, \tag{5.3}$$

which can be directly compared to the partition function in the canonical ensemble [76, 75, 15, 14, 113, 109]. The model parameters $\boldsymbol{\theta}$ are the variables that define the physical state vector, the cross entropy $\hat{H}(\boldsymbol{\theta})$ is the energy $E(\boldsymbol{\theta})$, the prior $\varpi(\boldsymbol{\theta})$ is the density of states $\rho(\boldsymbol{\theta})$.The data $x^N$ is quenched disorder in the physical system. The sample size $N$ is identified with the inverse temperature $\beta \equiv T^{-1}$. (Choosing $\beta \leftrightarrow N$ is only one of at least two proposals for the identification of the temperature. See App. D.1.2.) This assignment is

natural in the following sense: At small-sample-size $N$, many parameter values are consistent with the data, in analogy with the large range of states $\boldsymbol{\theta}$ occupied at high temperature $T$ in the canonical ensemble. The analogy between different quantities is summarized in Tab. 5.1.

### *5.2.1 Application of thermodynamic identities*

To extend the previously proposed correspondence, we follow the standard prescriptions from statistical mechanics to compute thermodynamic potentials, properties, and variables for the system [55, 117]. These are shown in the lower half of Tab. 5.1. The thermodynamic quantities depend on the particular realization of the data $X^N$. In the current context we are interested in the expectation over this *quenched disorder* (*i.e.* data). We define the disorder average with an overbar:

$$\overline{f}(N) \equiv \langle\, f(X^N, \boldsymbol{\theta}_0) \,\rangle_{X,\boldsymbol{\theta}_0}\,, \tag{5.4}$$

where $X \sim q(\cdot|\boldsymbol{\theta}_0)$ and $\boldsymbol{\theta}_0 \sim \varpi$.

## *5.3 Results*

We motivate interpretations of the thermodynamic quantities by developing the thermodynamic potentials in the normal model and the large-sample-size limit of singular models. The similarity between these large-sample-size results and familiar results in statistical mechanics show that interpretations of the thermodynamic quantities of a statistical model can be deduced from the meaning of their physical counterparts.

### *5.3.1 Models*

*Free particle and the normal model*

We compare a free particle in $K$-dimensions confined to a $K$-cube in statistical mechanics to a $K$-dimensional normal model with $K$ unknown means $\vec{\mu}$ and known variance $\sigma^2$ in Bayesian inference. The true parameter $\vec{\mu}_0$ is drawn from a $K$-dimensional normal distribution (the

| Model | K-D-Free-particle | K-D-Normal-prior | K-D-singular |
|---|---|---|---|
| $\overline{F}$ | $E_0 + \frac{K}{2\beta}\log\frac{\beta}{\beta}_0$ | $H_0 + \frac{K}{2N}\log\frac{N}{N_0}$ | $H_0 + \frac{\gamma}{2N}\log N + ...$ |
| $\overline{U}$ | $E_0 + \frac{K}{2\beta}$ | $H_0 + \frac{K}{2N}$ | $H_0 + \frac{\gamma}{2N} + ...$ |
| $\overline{C}$ | $\frac{K}{2}$ | $\frac{K}{2}$ | $\frac{\gamma}{2} + ...$ |
| $\overline{S}$ | $\frac{K}{2}\left(1 - \log\frac{\beta}{\beta_0}\right)$ | $\frac{K}{2}\left(1 - \log\frac{N}{N_0}\right)$ | $-\frac{\gamma}{2}\log N + ...$ |

Table 5.2: **Thermodynamic-Bayesian correspondence.** The thermodynamic quantities of a $K$-dimensional free particle with ground-state energy $E_0$ and thermal de-Broglie inverse temperature $\beta_0$, defined in the App. D.1.6 are compared to a $K$-dimensional normal model with a conjugate prior. Inspection reveals that the free particle is exactly equivalent to the normal model, identifying the parameters as described in Tab. 5.1. For the singular model, we supply only the leading order contributions in the large $N$ limit. The learning coefficient $\gamma \leq K$. The special case of $\gamma = K$ is a regular model.

prior $\varpi$) with mean $\vec{\mu}_\varpi$ and variance $\sigma^2_\varpi$. It will be convenient to define a critical sample size:

$$N_0 \equiv \sigma^2/\sigma^2_\varpi, \tag{5.5}$$

where the information content of the observations $x^N$ is equal to the information content of the prior. (See Appendix Sec. D.3.1.) In the current context, we will be interested in the uninformative limit: $N_0 \to 0$.

*Singular models*

The normal model is representative of the large sample-size-limit of a regular Bayesian model of dimension $K$ (i.e. the Bernstein-von Mises theorem [25, 94, 54]). For generality, we also study the large-sample-size limit of a singular model of dimension $K$. Models are *singular* when parameters are *structurally unidentifiable* [143]:

$$q(x|\boldsymbol{\theta}_1) = q(x|\boldsymbol{\theta}_2) \qquad \text{for} \qquad \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2, \tag{5.6}$$

where the unidentifiability cannot be removed by coordinate transformation. A regular model is the special case where all parameters are identifiable, the parameter manifold is continuous, and the Fisher information matrix (defined in Eqn. D.6) is positive definite. Using exact asymptotic results for singular models [143], the thermodynamic quantities for each model and limit are shown in Tab. 5.2.

### *5.3.2 Thermodynamic potentials*

*Free energy*

A relation between the partition function and Bayesian evidence has long been discussed [76, 75, 15, 14, 113, 109]. The free energy $F$ represents the Bayesian model preference: the minus-log-evidence per observation, or message length per observation. In Tab 5.2 it is seen that $F$ breaks up into two parts. The first term is the code length per observation using the optimal encoding, $H_0$, and the second term is the length of the code required to encode the model parameters using the prior (per observation) [123, 57]. The model that maximizes the evidence and therefore minimizes $F$ is selected in the canonical approach to Bayesian model selection.

*Average energy*

The thermodynamic prescription for computing the average energy involves a derivative with respect to temperature (Tab. 5.1). In the context of statistics, we will formally interpret this derivative using a finite difference definition, such that

$$U(x^N) \equiv -\left\langle \log q(x_i|x^{\neq i}) \right\rangle_{i=1..N}, \tag{5.7}$$

where $q(x_i|x^{\neq i}) \equiv Z(x^N)/Z(x^{\neq i})$ is the Bayes-predictive distribution. The RHS is a well-known statistical object: the Leave-One-Out-Cross-Validation (LOOCV) estimator of model performance (See App. D.1.3). The statistical interpretation of average energy $U$ is therefore the minus-expected-predictive-performance of the model (*e.g.* [62]).

As shown in the examples in Tab. 5.2, the averaged energy can be written as the sum of two contributions: The first term $H_0$ is the performance of the model if the true parametrization was known and corresponds to a ground state energy. The second term represents the loss associated with predicting a new observation $X$ using estimated model parameters and corresponds to the thermal energy. The loss term follows the typical behavior predicted by the equipartition theorem: *there is a half $k_BT$ of thermal energy per harmonic degree of freedom* [117]. This is an important universal property of regular models in the large-sample-size limit: Independent of the detailed structure of the model, there is a universal predictive loss $\frac{1}{2N}$ per degree of freedom in the model. This universal predictive loss can be interpreted as the mechanism by which the Akaike Information Criterion (AIC) estimates the predictive performance [3, 30].

*Learning capacity*

To study the predictive loss, it is natural to study the statistical quantity corresponding to the heat capacity. The heat capacity measures the rate of increase in thermal energy with temperature ($\overline{C}$ in Tab. 5.1). The statistical analogue of the heat capacity, a *learning capacity*, is a measure of the rate of increase in predictive performance with sample size. For the normal model:

$$\overline{C} = \tfrac{1}{2}K, \tag{5.8}$$

as implied by the equipartition theorem. (See Tab. 5.2.)

To consider how this analogy generalizes to a generic statistical model, we use the large-sample-size limit asymptotic expression for the Bayesian evidence from Ref. [143] to compute the learning capacity. (See Tab. 5.2.) Like the normal model, the learning capacity for a singular model has the equipartition form but with an effective dimension:

$$\overline{C} = \tfrac{1}{2}K_{\text{eff}}, \tag{5.9}$$

where $K_{\text{eff}} = \gamma$ is the learning coefficient defined by Watanabe [143]. A regular model is a special case of this expression where $K_{\text{eff}} = K$, the dimension of the parameter manifold. We

therefore conclude that equipartition theorem describes the universal properties of regular statistical models in the large-sample-size limit: The learning capacity is half the number of degrees of freedom.

.

*The Gibbs entropy*

In physics, the Gibbs entropy generalizes the Boltzmann formula: $S = \log \Omega$ where $\Omega$ is the number of accessible states. We propose that the Gibbs entropy has the analogous meaning in the context of Bayesian statistics: The Gibbs entropy is the log-number of models consistent with the data. The finite-sample-size expression for the entropy is

$$S(x^N) \equiv N(U - F), \tag{5.10}$$

where Eqn. 5.7 provides an explicit expression for $U$. The Gibbs entropy of a normal model is shown in Fig. 5.1. Above the critical sample size $N_0$, the data is informative to the parameter values and therefore the number of models consistent with the data is reduced. As a result the Gibbs entropy becomes increasingly negative as sample size $N$ grows.

### *5.3.3 The principle of indifference*

In statistical physics, the density of states is known (*i.e.* measured) but in inference the selection of a prior is often subjective. The construction of an objective or uninformative prior is a long-standing problem in Bayesian statistics. What insight does the proposed correspondence provide for prior choice?

Prior construction since Bayes and Laplace has often attempted to apply a *Principle of Indifference*: All *mutually exclusive* and *exhaustive* possibilities should be assigned equal prior probability [93, 81]. One interpretation of this prescription is that it maximizes entropy [76, 130]. However, the principle of indifference is difficult to interpret in the context of continuous parameters, or across models of different dimension. For example, are normal models with means $\mu$ and $\mu + \mathrm{d}\mu$ mutually exclusive (distinguishable)? Even if the mean were

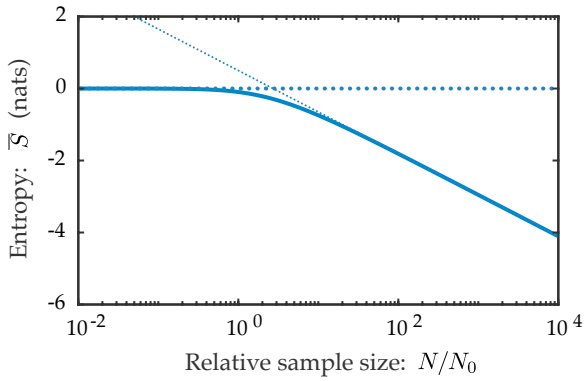


Figure 5.1: **Understanding Gibbs entropy.** The Gibbs entropy for the normal-model-with-prior is plotted as a function of sample size. The Gibbs entropy can be understood heuristically as the log ratio of the model consistent with the data to allowed models. At small sample size, the model structure determines the parametrization and therefore all models allowed are consistent with the data and there is zero Gibbs entropy. As the sample size grows beyond the critical sample size $N_0$, fewer and fewer of the allowed models are consistent with the data and the entropy decreases like $-\frac{1}{2}K \log N$. The non-positivity of the Gibbs entropy is a direct consequence of the normalization of the prior, which forces the Gibbs entropy to have a maximum value of zero. A prior determined by the generalized principle of indifference avoids this non-physical result.

constrained to be an integer ($\mu \in \mathbb{Z}$) to define *mutually exclusive*, the exhaustive condition is also problematic. Exhaustive would correspond to a uniform weighting over all integers. This vanishing prior weight ($1/\infty$) on the non-compact set $\mathbb{Z}$ results in a paradoxical value for the evidence $Z \to 0$ and the rejection of the model irrespective of the data, as described in Sec. 5.5.2 (Lindley-Bartlett paradox).

### *5.3.4 A generalized principle of indifference*

To define *mutually exclusive* in a statistical context, we look for natural analogues to this problem in statistical physics. A surprising result from the perspective of classical physics is that Nature makes no distinction between states with identical particles exchanged (*e.g.* electrons) and counts only distinguishable states (the Gibbs paradox). Following V. Balasubramanian [14], we proposed that the concept of indistinguishability must be applied to objective Bayesian inference. We take the *mutually exclusive* criteria in the principle of indifference to refer to distributions which are mutually distinguishable at the experimental resolution available. We propose a *generalized principle of indifference*: sets of indistinguishable models are each collectively assigned the weight of a single distinguishable model.

To study the weighting of each model, *we must prepare the data using a different procedure*. We distribute $X^N$ according to an assumed true parameter $\boldsymbol{\theta}$: $X^N \sim q(\cdot|\boldsymbol{\theta})$, omitting the expectation over $\boldsymbol{\theta}$ [1]. A generalized principle of indifference states that the prior $\varpi$ should be chosen such that:

$$\overline{S}(\boldsymbol{\theta}; N, \varpi) \approx \text{const} \quad \forall \quad \boldsymbol{\theta} \in \Theta, \tag{5.11}$$

at sample size $N$, where the Gibbs entropy is now a function of $\boldsymbol{\theta}$. Eqn. 5.11 realizes a statistically principled definition for the condition of equal model weighting on mutually exclusive models.

---

[1] We must make the important distinction between the *inference prior* $\varpi$, used to perform inference, and the *true prior*, used to generate the true parameters $\boldsymbol{\theta}_0$. This approach has a long and important precedent: *e.g.* [143, 79].

The correspondence also offers a natural mechanism for resolving statistical anomalies arising from the *exhaustive* condition in the principle of indifference that gives rise to the Lindley-Bartlett paradox. In statistical mechanics, the partition function $Z$ is not normalized by construction since the density of states $\rho$ is a density but not a probability density. Therefore, a natural solution to statistical anomalies arising from the exhaustive condition is to re-interpret the objective inference prior as a *density of models*. To specify a consistent density of models between different parameter values and model families, we replace the prior $\varpi(\boldsymbol{\theta})$ with a model density $w(\boldsymbol{\theta})$ such that:

$$\overline{S}(\boldsymbol{\theta}; N, w) \approx 0, \tag{5.12}$$

assigning unit multiplicity to all parameters $\boldsymbol{\theta}$ and model families $I$. (*Technical note:* We avoid specifying Eqns. 5.11 and 5.12 as equalities since the condition is typically not exactly realizable for all $\boldsymbol{\theta}_0$ at finite sample size $N$. A precise formulation will be described elsewhere, but is analogous to the mini-max approach of Kashyap where the largest violation of the GPI condition is minimized [79].) Eqn. 5.12, and the resulting prior are reparametrization invariant (see App. D.1.5). The prior $w$ will be improper, but none-the-less the normalization is well defined. We shall refer to Eqn. 5.12 as the *Generalized Principle of Indifference* which realizes both the mutually-exclusive and exhaustive conditions using a principled statistical approach, regardless of the nature of the parameter manifold. We will call the prior $w$ that satisfies Eqn. 5.12 the *GPI prior*.

### *5.4 Applications*

We have used the correspondence between thermodynamics and statistics to motivate the definition of the learning capacity and the Gibbs entropy (and resultant generalized principle of indifference). We now wish to investigate the statistical properties of these definitions. We will find that both these novel statistical objects provide new insight into statistics and learning.

### *5.4.1 Learning Capacity*

The applicability and failure of the equipartition theorem are well understood phenomena in physics. At high or low temperature, degrees of freedom can become anharmonic, altering their contribution to the heat capacity [117]. For instance, due to the discrete structure of the quantum energy levels, degrees of freedom can *freeze out* at low temperature. (See Fig. 5.2A.) Degrees of freedom can also become irrelevant at high temperature. For instance, the position degrees of freedom of a gas do not contribute to the heat capacity [121]. We see an analogous high-temperature freeze-out mechanism in the context of inference.

In this section, we investigate the phenomenology of the learning capacity in a series of simple examples. In each case, we first compute $\overline{F}$, then we compute the learning capacity (as defined in Tab. 5.1). In most of our examples, both computations can be performed analytically.

- In Sec. 5.4.1, we analyze the finite-sample-size behavior of the learning capacity in the context of the normal model with unknown mean and variance. The dependence of the log-likelihood on the variance is anharmonic and therefore there are interesting finite-sample-size corrections to equipartition.

- In Sec. 5.4.1, we analyze a quantum-like freeze-out phenomenon for a model on a discrete parameter manifold.

- In Sec. 5.4.1, we analyze a problem where a discrete parameter manifold arise naturally: the stoichiometry of a Poisson processes.

- In Sec. 5.4.1, we analyze a singular model, the exponential mixture model, which has previously been identified as *sloppy*. We show that the learning capacity is smaller than predicted by equipartition for parameter values in the vicinity of the singularity.

- In Sec. 5.4.1, we analyze the learning capacity in a non-regular but analytically-tractable model. In this example the learning capacity is larger than predicted by

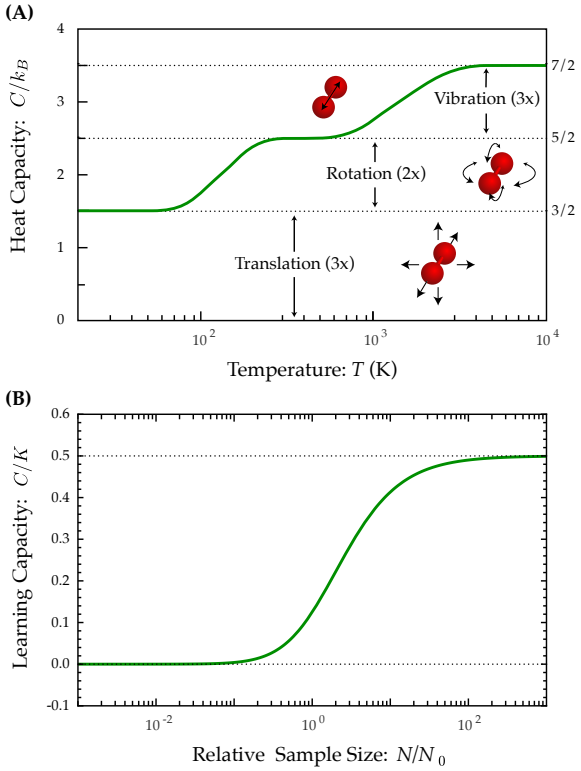


Figure 5.2: **The failure of equipartition.** The behaviors of the heat capacity and learning capacity are compared and related to the applicability or inapplicability of the Equipartition theorem in different regimes. **Panel A: Low-temperature freeze-out in a quantum system.** The heat capacity is plotted as a function of temperature. Equipartition predicts heat capacity should be constant, equal to half the degrees of freedom in the system. Plateaus can be observed at half-integer values, but the number of degrees of freedom is temperature dependent due to the discrete nature of quantum energy levels. At low temperature, some degrees of freedom are frozen out since the first excited state is thermally inaccessible. **Panel B: High-temperature freeze-out in the Learning capacity.** Analogous to the statistical mechanics system, the statistical learning capacity transitions between half integer plateaus, reflecting a temperature-dependent number of degrees of freedom. At low sample size $N$ (high temperature), the parameters are completely specified by model constraints (the prior) and therefore the parameters do not contribute to the learning capacity. At large sample size $N$, the parameters become data dominated and therefore the learning capacity is predicted by equipartition ($\frac{1}{2}K$).

equipartition.

*Normal model with an unknown mean*

In many regular statistical models at finite sample size, it is the model structure and not the data that constrain the parameter values. In these cases, structurally constrained parameters will not contribute to the learning capacity. A simple and exactly tractable example of this phenomenon has already been discussed: the normal model with unknown mean and an informative prior.

*Model:* We define a normal model on a $D$-dimensional observation space with unknown mean and unknown variance $\sigma^2$. The likelihood function is:

$$q(\vec{x}|\boldsymbol{\theta}) \equiv (2\pi\sigma^2)^{-D/2} \ \exp[-\tfrac{1}{2\sigma^2}(\vec{x}-\vec{\mu}\,)^2], \tag{5.13}$$

with $\boldsymbol{\theta} \equiv (\vec{\mu})$ and informative prior:

$$\varpi(\boldsymbol{\theta}) \equiv (2\pi\sigma_\varpi^2)^{-D/2} \ \exp[-\tfrac{1}{2\sigma_\varpi^2}(\vec{\mu}-\vec{\mu}_\varpi\,)^2], \tag{5.14}$$

We now consider the informative limit where the critical sample size $N_0$ (Eqn. 5.5) is finite.

*Analysis:* The learning capacity can be computed analytically:

$$\overline{C} = \tfrac{K}{2(1+N_0/N)^2}. \tag{5.15}$$

At large sample size, the learning capacity is equal to the equipartition expression (Eqn. 5.9). At small sample size (high temperature), the prior determines the parametrization, and therefore the parameter does not contribute to the learning capacity and $\overline{C} \to 0$. (See Fig. 5.2B.) This situation is roughly analogous to the heat capacity of a gas. In the solid phase, the position degrees of freedom contribute to the heat capacity in the canonical way whereas in the gas phase, the walls of the box confine the atoms, the energy does not depend on the position degrees of freedom, and these variables no longer contribute to the heat capacity.

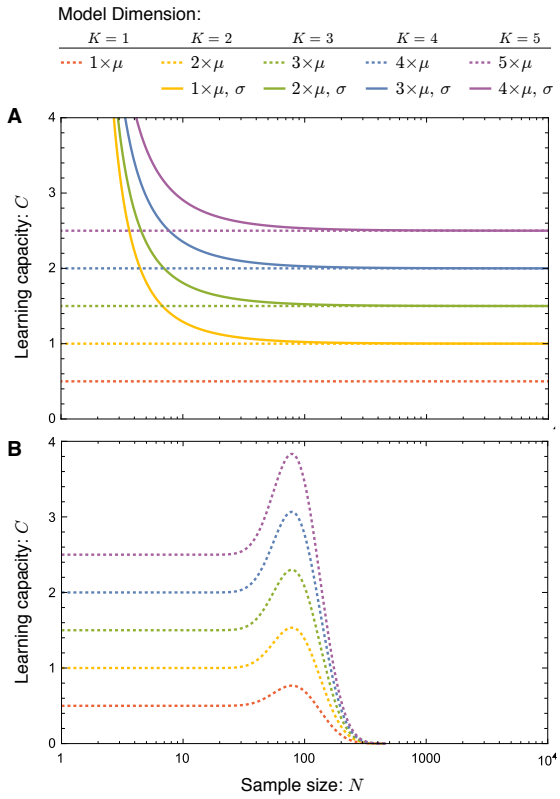


Figure 5.3: **Panel A: Learning capacity at finite sample size.** At large sample size, equipartition predicts the learning capacity of all models. At sample size $N = 1$ the learning capacity diverges for models with unknown variance since the mean and variance cannot be simultaneously estimated from a single measurement. **Panel B: Learning capacity on a discrete manifold.** The learning capacity of a normal model with an unknown D-dimensional mean $\vec{\mu} \in \mathbb{Z}^D$ and variance $\sigma^2 = 15$. For statistical uncertainty $\delta\mu \gg 1$, the learning capacity is predicted by equipartition since the discrete nature of the parameter manifold cannot be statistically resolved. For $\delta\mu \ll 1$, there is no statistical uncertainty in the parameter value (due to the discreteness of $\mu$) and the degrees of freedom freeze out, giving a learning capacity of zero.

*Normal model with an unknown mean and variance*

The learning capacity $\overline{C}$ typically deviates from equipartition-like behavior in the small sample-size (high temperature) limit in analogy to anharmonic effects in the heat capacity of metals near the melting point [142, 46]. To demonstrate this phenomenon, we analyze the normal model with unknown mean and variance.

*Model:* We define a normal model on a $D$ dimensional observations space with unknown mean and unknown variance $\sigma^2$. The likelihood function is given by Eqn. 5.14 with $\boldsymbol{\theta} \equiv (\vec{\mu}, \sigma)$ and impose an improper Jeffreys prior $\varpi = \sigma^{-D-1}$. Although, the dependence of the information is harmonic in $\mu$, it is non-harmonic in variance $\sigma^2$.

*Analysis:* The learning capacity can be computed analytically:

$$\begin{aligned} \overline{C} &= \frac{D}{2}\Big[1 - Dn^2\psi^{(1)}(\tfrac{D(n-1)}{2}) + ... \\ &\quad + \tfrac{Dn^2}{2}\psi^{(1)}(\tfrac{Dn}{2}) - \tfrac{D^2n^3}{4}\psi^{(2)}(\tfrac{D(n-1)}{2})\Big], \end{aligned} \tag{5.16}$$

in terms of the polygamma functions $\psi$. The learning capacity is plotted as a function of sample size in Fig. 5.3A. The learning capacity diverges at sample size $N = 1$ since the mean and variance cannot be estimated from a single measurement and the divergence of the learning capacity signals an infinite predictive loss. The learning capacity of the normal model with known mean and variance is representative of the behavior of many models: Typically, the learning capacity can show significant differences with the equipartition limit at very small sample size but rapidly converges to the equipartition value as the sample size grows.

*Normal model with discrete mean.*

An important exception to the generic behavior described in the previous paragraph occurs when the *parameter manifold is discrete rather than continuous.* In exact analogy to the freeze-out phenomenon in quantum statistical mechanics, as the sample size increases (and the temperature decreases) the discrete nature of the parameter manifold (energy levels)

becomes the dominant structure in the analysis and the system condenses into a single distribution (ground state). We will consider a contrived but analytically tractable example: a normal model with unknown discrete mean.

*Model:* The likelihood for the $D$-dimensional normal model is defined in Eqn. 5.14. The parameters now include only the the mean: $\boldsymbol{\theta} = (\vec{\mu})$, with the mean constrained to have an integer values: $\vec{\mu} \in \mathbb{Z}^D$. We assume a flat improper Jeffreys prior: $\varpi = \sigma^{-D}$.

*Analysis:* The learning capacity can be computed analytically and the expression is given in the Appendix Sec. D.3.3. The learning capacity is plotted in Fig. 5.3B. To discuss the phenomenology, it is useful to define a frequentist *statistical resolution* with respect to parameter coordinate $\theta^i$:

$$\delta\theta^i(N) \equiv N^{-\frac{1}{2}} \sqrt{[\boldsymbol{I}^{-1}]^{ii}}, \tag{5.17}$$

in terms of the Fisher information matrix $\boldsymbol{I}$ (Eqn. D.6), which is a naturally covariant symmetric tensor on the continuous parameter manifold. $\delta\theta^i(N)$ is the width of the posterior in the large-sample-size limit. For the normal model, $\delta\mu = \sigma/\sqrt{N}$. For a regular model with discrete parameters in the large sample size limit, the learning capacity is:

$$\overline{C} = \begin{cases} \frac{1}{2}K, & \Delta\theta^i \ll \delta\theta^i \text{ for all } i \\ 0, & \Delta\theta^i \gg \delta\theta^i \text{ for all } i \end{cases} \tag{5.18}$$

where $\Delta\theta^i$ is the lattice spacing for parameter coordinate $\theta^i$. The physical interpretation is clear: At large sample size, the system condenses into a single state. Therefore the corresponding degrees of freedom freeze out, and no longer contribute to the learning capacity. At small sample size, the discrete nature of the parameter manifold cannot be resolved, and the parameter manifold is effectively continuous. The learning capacity therefore assumes the equipartition value (provided that the sample size is large enough such that the information is effectively harmonic, as discussed in Sec. 5.4.1).

*Stoichiometry of a Poissn process*

In the previous example, we considered a somewhat contrived example where the mean of a normal process was discrete, but many important statistical problems are naturally defined on discrete parameter manifolds (*e.g.* the dimension of a chi-square distribution, the shape of gamma distribution, *etc*). For an explicit example, we will analyze a problem that arises in the context of our experimental work, the analysis of protein stoichiometry. The molecular complex is known to be a multimer: a complex consisting of a well-defined number of identical protein subunits. In our experiments, we measure stoichiometry by the fluorescence intensity of the protein labels.

*Model:* The number of photons emitted per fluorophore is well modeled by a Poisson process. We will assume that the average intensity per fluorophore is known (i.e. the rate $\lambda$ is known). The stoichiometry, or number of fluorophores, is $m \in \mathbb{N}$, an unknown natural number. Let $x_i$ be the binary observation of a photon ($x = 1$) or no photon ($x = 0$) in a short interval length $\delta t$:

$$q(x|\boldsymbol{\theta}) \equiv (m\,\lambda\,\delta t)^x (1 - m\,\lambda\,\delta t)^{1-x}, \tag{5.19}$$

where we will work in the limit as $\delta t \to 0$ where the parameter is $\boldsymbol{\theta} = (m)$. We apply a flat prior $\varpi_m = 1$.

*Analysis:* The learning capacity can be computed analytically, as shown in Appendix Sec. D.3.7. The limiting cases can be understood intuitively in terms of the results in the previous section (Eqn. 5.20). The lattice spacing is $\Delta m = 1$ and the statistical resolution is $\delta m = \sqrt{m/\lambda t}$ and the limiting learning capacity is:

$$\overline{C} = \begin{cases} \frac{1}{2}, & \Delta m \ll \delta m \\ 0, & \Delta m \gg \delta m \end{cases} \tag{5.20}$$

where the sample size dependence is represented as an interval duration: $t = N\,\delta t$. For $\Delta m \ll \delta m$, the stoichiometry $m$ appears continuous since the posterior on $m$ spans multiple values of $m$ (Fig. 5.4A) and therefore the learning capacity is predicted by equipartition

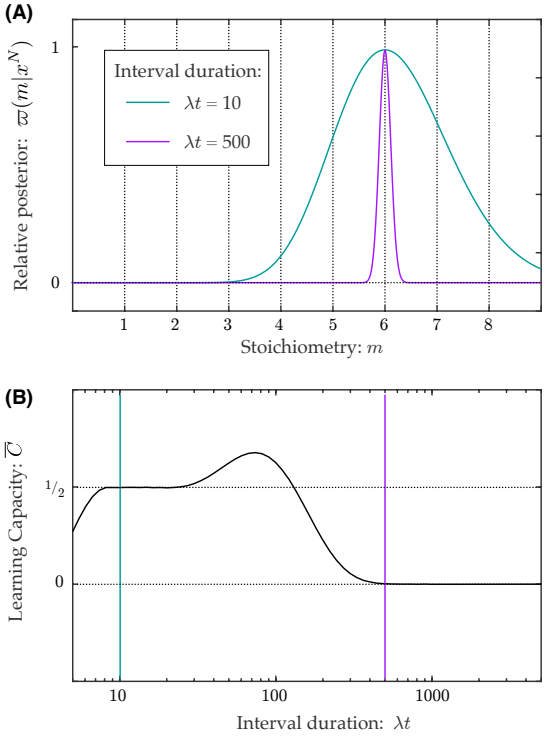


Figure 5.4: **Panel A: Posterior for low-temperature freeze-out.** Low-temperature freeze-out occurs when there is no statistical ambiguity in the parameter value. For long interval durations ($\lambda t = 500$) the posterior only weights a single parameter value ($m = 6$) whereas the manifold is effectively continuous for intermediate interval durations ($\lambda t = 10$) and multiple parameter values as weighted. **Panel B: Learning capacity for low-temperature freeze-out.** For long interval durations ($\lambda t = 500$), the stoichiometry $m$ is frozen-out and therefore the learning capacity is zero. For intermediate interval durations ($\lambda t = 10$), equipartition applies and the learning capacity is one-half. At short intervals, the large-sample-size limit assumption is violated and the learning capacity diverges from the limiting value.

(Fig. 5.4B). For $\Delta m \gg \delta m$, there is no statistical ambiguity in the parameter value $m$ (Fig. 5.4A), the parameter freezes out, leading to zero learning capacity (Fig. 5.4B). The plot of the learning capacity (Fig. 5.4B) shows one additional important feature that has already been discussed: At the smallest interval lengths $t$, the large sample-size limit assumed in equipartition fails, leading to a second deviation from equipartition.

*Exponential mixture models.*

The previous two examples demonstrated the quantum-like freeze-out that results from discrete parameter manifolds. Our next example illustrates another small sample size (*higher-temperature*) freeze-out phenomenon analogous to the loss of heat capacity when a solid sublimes to form a gas. Here, this phenomenon arises as the result of model singularity: A zero mode appears in the Fisher information matrix corresponding to one (or more) parameter coordinates becoming unidentifiable (Eqn. 5.6). To explore this phenomenon, we analyze the exponential mixture model which has previously been identified as *sloppy model* by Transtrum, Machta and coworkers using a criterion defined by the distribution of the eigenvalues of the Fisher information matrix [101].

*Model:* Consider a model for the lifetime of a mixed population consisting of several different chemical species $I$ with different transition rates. Both the transition rates ($k_I$) and the relative abundance of the species ($p_I$) are unknown. For an $m$ species model, the likelihood function for the lifetime $t$ is:

$$q(t|\boldsymbol{\theta}) \equiv \sum_{I=1}^{m} p_I \, k_I \, \mathrm{e}^{-k_I t}, \tag{5.21}$$

with parameters:

$$\boldsymbol{\theta} \equiv \begin{pmatrix} p_1 & \ldots & p_m \\ k_1 & \ldots & k_m \end{pmatrix}, \tag{5.22}$$

subject to the constraint: $\sum_I p_I = 1$ and we apply improper prior $\varpi(\boldsymbol{\theta}) = 1$. The exponential mixture model is singular since parameter $k_I$ is unidentifiable for $p_I = 0$ and $p_1$ is uniden-

tifiable for $k_1 = k_2$. (See Eqn. 5.6.) For simplicity, we analyze the smallest model with a singularity ($m = 2$) to facilitate the numerical Bayesian marginalization.

*Analysis:* We compute the learning capacity at two locations in parameter manifold, at the singularity ($\boldsymbol{\theta}_S$) and far from it ($\boldsymbol{\theta}_R$):

$$\boldsymbol{\theta}_S = \begin{pmatrix} 1 & 0 \\ 1 & 10 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\theta}_R = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 10 \end{pmatrix}. \tag{5.23}$$

The learning capacity is computed numerically for $N = 100$ observations with distribution $T^N \sim q(\cdot|\boldsymbol{\theta})$:

$$\overline{C}(\boldsymbol{\theta}) = \begin{cases} 0.61, & \boldsymbol{\theta} = \boldsymbol{\theta}_S \\ 1.5, & \boldsymbol{\theta} = \boldsymbol{\theta}_R \end{cases}. \tag{5.24}$$

Far from the singularity ($\boldsymbol{\theta}_R$), the equipartition theorem predicts the learning capacity ($\dim/2$) whereas close to the singularity ($\boldsymbol{\theta}_S$), where the model is effectively described by only a single parameter ($k_1$), the learning capacity reflects this smaller effective model dimension. As expected, the exponential mixture model is predictively sloppy in the vicinity of the singular point, but not elsewhere.

We expect that the behavior of the exponential mixture model is representative of many machine learning and systems biology problems. In these problems, the effective dimension $K_{\text{eff}}$ may be very much smaller than the true dimension of model. In practice, these models do not require exact structural unidentifiability (Eqn. 5.6) to show a reduced learning capacity. For instance, the reduced complexity is not only at the singularity but in the vicinity of the singularity as well. For models with structural unidentifiability, this vicinity-of-singularity region will shrink with sample size. For models without structural unidentifiability (Eqn. 5.6) that are regular everywhere, equipartition will hold at sufficiently large sample size.

*Uniform distribution with unknown upper limit.*

In the previous example, the non-regular model showed reduced learning capacity at the singularity but non-regular models can also have increased learning capacity as well. To

illustrate this phenomenon, consider a continuous version of the German Tank problem, estimation of the support of a uniform distribution [56].

*Model:* Consider a uniform distribution $X \sim \mathcal{U}(0, L)$ with unknown end point $L \geq 0$. The likelihood function is:

$$q(x|\boldsymbol{\theta}) = \begin{cases} L^{-1}, & 0 \leq x \leq L \\ 0, & \text{otherwise} \end{cases}, \tag{5.25}$$

with parameter $\boldsymbol{\theta} \equiv (L)$ and improper prior $\varpi(L) = L^{-1}$.

*Analysis:* In this model, neither the first nor second derivative of the cross entropy $H(L; L_0)$ exist at the true parameter $L_0$ and therefore the model is not regular. It is straightforward to compute the learning capacity:

$$\overline{C} = 1, \tag{5.26}$$

corresponding to an effective dimension of two, even though the parameter manifold is one dimensional. This result is exact and independent of sample size $N$. (See Appendix Sec. D.3.6.)

### *5.4.2 Generalized principle of indifference*

The proposed generalized principle of indifference has properties that rectify significant shortcomings with other approaches to prior selection. We showcase these properties with the analysis of a series of examples.

- In Sec. 5.4.2, we compute the GPI prior for regular models in the large-sample-size limit. This analysis reveals a connection between the GPI prior and the Jeffreys prior.

- In Sec. 5.4.2, we demonstrate an exact computation of the GPI prior for a number of non-harmonic models.

- In Sec. 5.4.2, we demonstrate an exact computation of the GPI prior for a non-regular model.

- In Sec. 5.4.2, we analyze the limiting behavior of the GPI prior on a discrete parameter manifold. GPI unifies two conflicting approaches, and interpolates between respective limits as a function of sample size.

- In Sec. 5.4.2, we analyze a problem where the GPI prior cannot be computed exactly: the analysis of stoichiometry in a Poisson process.

*Approximate GPI prior for regular models*

We will first explore the properties of the generalized principle of indifference by computing the GPI prior in the large-sample-size limit of a regular model. To define the GPI prior, it is first useful to define the scaled-Jeffreys prior:

$$\rho(\boldsymbol{\theta}; N) \equiv \left(\tfrac{N}{2\pi}\right)^{K/2} I^{1/2}, \tag{5.27}$$

where $I$ is the determinant of the Fisher information matrix defined for a single sample (Eqn. D.6), $K$ is the dimension of the continuous parameter manifold $\Theta$. The prior $\rho$ is a density on parameter manifold with the qualitative meaning of the inverse volume of indistinguishable models at sample size $N$. The GPI prior is

$$w(\boldsymbol{\theta}; N) \approx \rho\ e^{-K}, \tag{5.28}$$

where $K = \dim \Theta$, as shown in Appendix Sec. D.1.4.

In the large-sample-size limit, the parameter dependence of the GPI prior is identical to the Jeffreys prior, which has enjoyed a long and successful history [80]. The Jeffreys prior was initially proposed because it was reparametrization invariant [78]. More recently the same prior has been motivated by numerous other arguments (*e.g.* [23, 14]). From the perspective of parameter inference, in the large-sample-size limit of a regular model, the GPI approach simply recapitulates a widely-applied method rather than generating novel algorithm.

*Exact GPI prior for symmetric models*

For simple models, symmetry and dimensional analysis often imply that $w$ must still be proportional to the Jeffreys prior even at small sample size. We compute the exact GPI prior analytically for the normal model with unknown mean and variance and the exponential model in the Appendix Sec. D.2.2. Both these models have a log-likelihood that is anharmonic in the parameters and therefore are expected to have non-trivial high-temperature behavior. The calculation reveals that the asymptotic form of the GPI prior (Eqn. 5.28) closely approximates the exact prior. In many models it is convenient to define the finite sample-size correction as an effective complexity $\mathscr{K}$ that replaces the model dimension $K$ in Eqn. 5.28:

$$w(\boldsymbol{\theta}; N) \approx \rho \; e^{-\mathscr{K}}, \tag{5.29}$$

$\mathscr{K}$ is plotted as a function of sample size ($N$) for a number of different models in Fig. 5.5. On an empirical basis, it is clear that Eqn. 5.28 is typically an excellent approximation for $w$ even small to intermediate sample sizes for many models.

*Exact GPI prior for a non-regular model*

In this section we explore another key motivation to the GPI approach: the analysis of non-regular models. We return to the example of the uniform distribution $X \sim \mathcal{U}(0, L)$ with unknown end point $L \geq 0$. In this case, the Fisher information matrix (Eqn. D.6) is not defined and so the Jeffreys prior approximation (Eqn. 5.28) cannot be applied. The definition of $w$ does not depend on assuming a regular model and it is still straight forward to compute $w$ (Appendix Sec. D.3.6):

$$w(L; N) = \tfrac{N}{L} \exp\left[-1 - N \log(1 + N^{-1})\right]. \tag{5.30}$$

which has a scaling of sample size $N$ corresponding to an effective dimension of two, even though the parameter manifold is one dimensional.

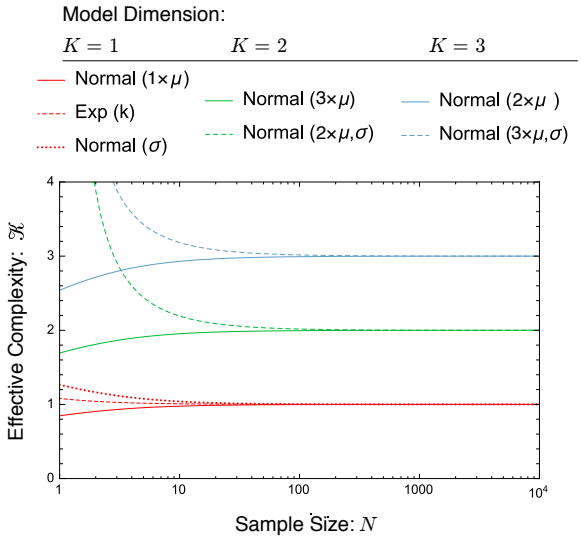


Figure 5.5: **Effective complexity of models at finite sample size.** We computed the exact GPI prior for a series of models of different dimension. At large sample size, the dimension determines the effective complexity: $\mathcal{K} = \dim \Theta/2$. At finite sample size there are significant corrections. The effective complexity divergences for the normal model (dashed curves) with unknown mean and variance at $N = 1$.

*GPI prior for discrete parameter manifolds*

For discrete parameter manifolds two competing and well-established methods exist for choosing a prior: (i) A literal interpretation of the principle of indifference would seem to imply that all parameter values are given equal weight. (ii) Alternatively, we can consider the continuous parameter limit where the prior can be chosen to give consistent results with the Jeffreys prior. Both approaches have desirable properties in different analytical contexts [21]. GPI provides an elegant resolution to this conflict: When the discrete nature of the parameter manifold can be statistically resolved, the GPI prior prior assigns equal weight to all discrete parameter values (i), whereas, if the discreteness of the space cannot be statistically resolved, the large $N$ limit gives rise to a Jeffreys prior (ii):

$$w = \begin{cases} \rho \; e^{-K} \prod_i \Delta\theta^i, & \Delta\theta^i \ll \delta\theta^i \\ 1, & \Delta\theta^i \gg \delta\theta^i \end{cases} \tag{5.31}$$

where $\Delta\theta^i$ is the lattice spacing and the statistical resolution $\delta\theta^i$ is defined in Eqn. 5.17. The GPI prior for a normal model with a discrete mean can be computed exactly and is described in Appendix Sec. D.3.3.

*Stoichiometry of a Poissn process (revisited)*

In each example discussed so far, it is possible to compute the GPI prior exactly. In most applications this approach is *not* tractable. Our analysis of regular models suggests that Eqn. 5.28 is often an excellent approximation. If this does not suffice, a recursive algorithm (Appendix, Sec. D.2.4) is a practical refinement to Eqn. 5.28. As an example of this approach, we return the analysis of the stoichiometry of a Poissn process. (See Sec. 5.4.1.) In this case it is clear from the learning capacity (Eqn. 5.20) that $w$ must interpolate between the discrete and continuous limits of Eqn. 5.31 as a function of the stoichiometry $m$. We compute $w$ numerically using the recursive algorithm (Appendix Sec. D.3.7).

We plot the $w$ (GPI) and flat (PI) priors as a function of the stoichiometry $m$ in Fig. 5.6A and the Gibbs entropy of the $w$ and flat priors in Fig. 5.6B. The traditional interpretation

| Model | Parameter support $\boldsymbol{\theta}$ | Generative parameters $\boldsymbol{\theta}_0$ |
|---|---|---|
| $\mathcal{N}$ | | $\mu_0 = 5$, $\sigma_0 = 1$ |
| $\mathcal{N}(\mu)$ | $\mu \in \mathbb{R}$ | $\mu_0 = 6$, $\sigma_0 = 1$ |
| $\mathcal{N}(\mu, \sigma)$ | $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+$ | $\mu_0 = 5$, $\sigma_0 = 0.75$ |
| $\mathrm{Exp}(\lambda)$ | $\lambda \in \mathbb{R}_+$ | $\lambda_0 = 2$ |
| $\mathcal{U}(L)$ | $L \in \mathbb{R}_+$ | $L_0 = 10$ |

Table 5.3: **Models for inference on simulated data.** Five data sets were generated, one for each model, using the generative parameters: $X^N \sim q(\cdot|\boldsymbol{\theta}_0)$. Inference was performed on the simulated data using the GPI prior $w$.

of the principle of indifference leads to a constant prior and non-constant Gibbs entropy. In contrast, GPI results in a non-constant prior and constant Gibbs entropy. As is seen in Fig. 5.6B, the Gibbs entropy of the flat prior rises for large stoichiometry because models with stoichiometry $m$ and $m+1$ cannot be distinguished for the dimensionless interval length $\lambda t$ and therefore the flat prior *over weights* these indistinguishable models by counting them independently. The GPI prior compensates by weighting these large stoichiometry distributions less, resulting in constant Gibbs entropy. As the dimensionless interval (*i.e.* sample size) increases, these distributions become increasingly distinguishable and the GPI increases the weight of these models. For small stoichiometry where the models can be statistically distinguished, the improper-flat prior and $w$ are identical (Fig. 5.6A).

### *5.4.3 Inference*

To demonstrate that the GPI prior automatically leads to non-anomalous inference (*i.e.* free from infinite normalization factors) and is also free from *ad hoc* parameters, we analyze simulated datasets for parameters defined on non-compact manifolds. We consider five competing models: three realizations of the normal model: known mean and variance, unknown mean

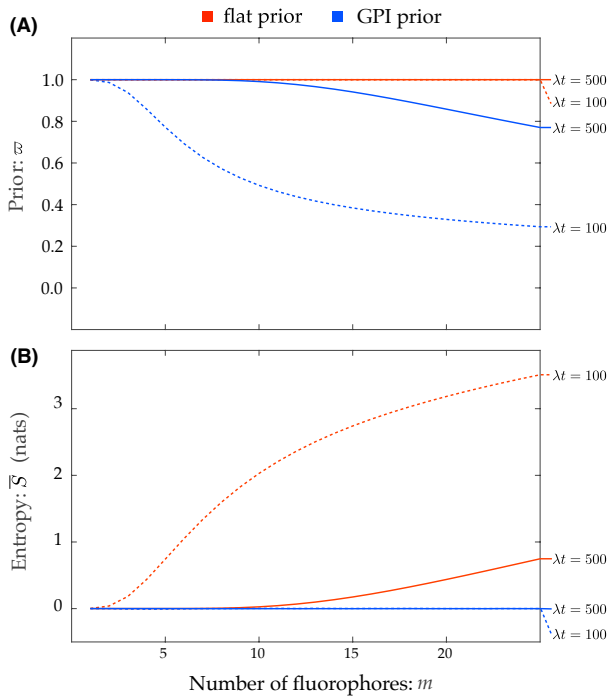


Figure 5.6: **Complementary views on indifference:** A flat prior and the GPI prior offer two different ways to define the principle of indifference. **Panel A: The GPI prior depends on parameter and sample-size.** The flat prior is constant with respect to changes in source number $m$, and sample size, while the GPI prior (for Poisson stoichiometry problem) changes with the parameter $m$, and responds to sample size. **Panel B: The GPI prior has (nearly) constant entropy.** The average entropy under the GPI prior is almost flat and zero everywhere, but the entropy of the flat prior is *not constant*. Some models are entropically favored under the flat prior in violation of the generalized principle of indifference.

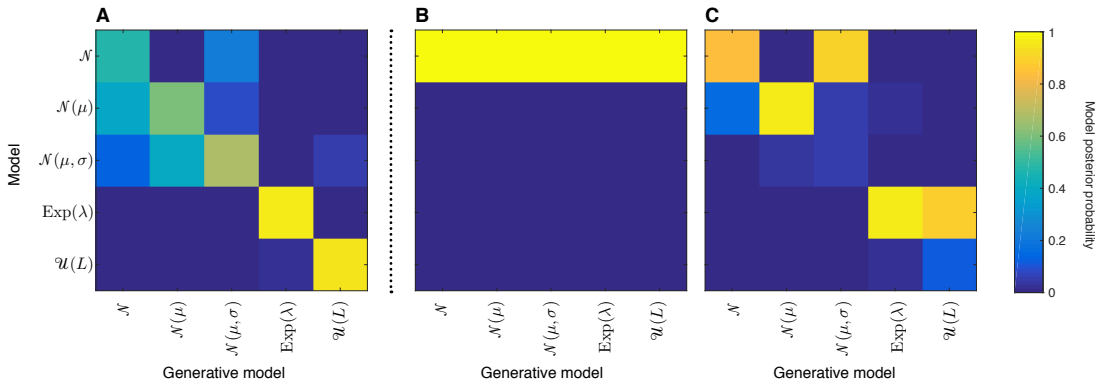


Figure 5.7: **Bayesian inference on model identity.** The posterior of model identity ($y$ axis) was computed for datasets generated by each model ($x$ axis). **Panel A: GPI prior.** For the simulated datasets, the generative model had the highest posterior probability as expected. **Panel B: Normalized objective prior.** The non-compactness of the parameter manifolds implies automatic rejection of all the higher-dimensional models. In this case, since the model $\mathcal{N}$ is parameter-free, it has posterior probability of 1 for all datasets, regardless of the fit. **Panel C: Revised informative prior.** To avoid this undesirable anomaly, we tune the prior parameter support to result in a reasonable posterior model probability. (See Tab. D.1.) Inference is no longer objective, as the posterior probabilities depend on how this tuning is performed. One representative plot of posterior probabilities to shown. In general, inference cannot be expect to identify the generative model unless the KL divergence is so large as to make the prior irrelevant.

and known variance, and unknown mean and variance and we also consider and exponential model with unknown rate and the uniform distribution model with unknown end-point. (See Tab. 5.3.) We will generate datasets from all fives models and then perform inference on the model parameters and identity for each dataset. We purposefully choose the true parameter values and sample size ($N = 20$) such that cross entropies are not so large as to make prior choice irrelevant. (See Tab. 5.3.)

*GPI approach*

We have computed the GPI prior for each of the proposed models. Inference on parameter values follows the standard Bayesian framework using the GPI prior $w$. The GPI prior includes the model prior (Appendix Sec. D.1.7) and therefore the posterior probability of model $I$ is:

$$\varpi(I|x^N) = Z_I / \sum_J Z_J, \tag{5.32}$$

where the model index $J$ runs over the five competing models. The model posteriors for the five sets of simulated data for a sample size of $N = 20$ are shown in Fig. 5.7A. The results show a number of important characteristics of the GPI prior: (i) There is an unambiguous Bayesian procedure for computing inference on both parameters and models. (ii) Inference on both parameters and models leads to non-anomalous results in which the generative distribution has non-zero posterior probability. (iii) For the normal models, the higher-dimensional models have lower posterior probability for the data generated by model $\mathcal{N}$, even though the generative distribution is realizable in $\mathcal{N}(\mu)$ and $\mathcal{N}(\mu, \sigma)$. This shows that the GPI prior contains an endogenous model selection mechanism favoring model parsimony, and we will discuss this in detail in Sec. 5.5.2.

*A canonical uninformative Bayesian approach*

For contrast, we briefly describe a canonical Bayesian approach to this analysis. We attempt to use the Jeffreys prior for each model. A problem immediately presents itself in the context

of the Uniform model where the Jeffreys prior is undefined. We must therefore deviate from our protocol and apply some other prior. We set a flat prior on the parameter $L$ motivated by the principle of indifference. The priors for the four models with parameters cannot be normalized due to their non-compact parameter manifolds. Parameter posteriors for each model can still be computed using an improper prior and the results are identical to the GPI approach, except for the uniform model where no Jeffreys prior exists. If the Bayesian approach is interpreted literally, the model posterior for the parameter-free normal model is one, regardless of which distribution was used to generate the data, due to the prior impropriety of the other models. (See Fig. 5.7B.) This is an undesirable outcome and this phenomenon is discussed in more detail in Sec. 5.5.2.

A number of *ad hoc* modifications to the proposed procedure are now possible to avoid this outcome. (i) After having seen the data, a Bayesian will often reconsider the prior and localize it around the values favored by the data. (See Fig. 5.7C.) This approach is formalized in variational or empirical Bayesian methods. In this case, the prior is no longer determined *a priori* and this double-use of data can lead to difficulties due to the potential for overfitting. (ii) A more rigorous approach is to sub-divide the data: One subset is used to train the prior to make it informative and the second set is used to do inference using the canonical procedure. There are two important disadvantages to this technique: An *ad hoc* decisions must then be made about the size of the data partitions. (This approach is essentially equivalent to eliciting a prior for the parameters from an expert.) A second disadvantage is that some information is lost from inference on the model identity from the data in the prior training subset. (iii) Alternatively, one could use a non-cannocical approach for inference on the model identity, like the use of the pseudo-Bayes factor [51, 48, 132, 141, 31]. As we shall discuss below in Sec. 5.5.2, this approach is consistent with the GPI approach in the large-sample-size limit.

In summary, the GPI approach results in an unambiguous protocol for selecting the prior and then performing inference whereas the canonical Bayesian approach requires *ad hoc* modifications to lead to acceptable results.

## 5.5 Discussion

### 5.5.1 Learning capacity

One valuable feature of the proposed correspondence is the potential to gain new insights into statistical phenomenology using physical insights into the thermodynamic properties of physical systems. Artificial Neural Networks (ANN) and systems-biology models are two examples of systems with a large number of poorly-specified parameters that none-the-less prove qualitatively predictive. This phenomena has been discovered empirically and has been termed *model sloppiness* [101, 137]. These models often have a logarithmic distribution of Fisher information matrix eigenvalues and this characteristic has been used as a definition of sloppiness [101]. But, this definition is unsatisfactory since it is not reparametrization invariant. It is easy to construct counterexamples for this definition: For instance, in a $K$-dimensional normal model where the variance for each dimension is logarithmically distributed, the Fisher information eigenvalues are likewise logarithmically distributed, but the model none-the-less behaves like a normal regular model from the standpoint of prediction and statistical analyses.

The correspondence suggests a definition directly written in terms of the predictive performance of the model and the equipartition theorem. We propose that *predictive sloppiness* be defined as models that have a smaller learning capacity than estimated from the model dimension:

$$\overline{C} < \tfrac{1}{2} \dim \Theta. \tag{5.33}$$

This definition (i) would exclude all regular models in the large sample-size limit, (ii) is reparametrization invariant and (iii) can be generalized to other non-Bayesian frameworks by expressing the learning capacity in terms of the predictive performance.

The sloppiness phenomenon, or freeze out, is the result of the model parameters being determined by model structure rather than the data. To understand the role of model structure in the freeze-out phenomenon, it is useful to write a qualitative expression for the free energy $F$. We project the parameters into a regular sector $\boldsymbol{\theta}_R$ dimension $K_R$ and a

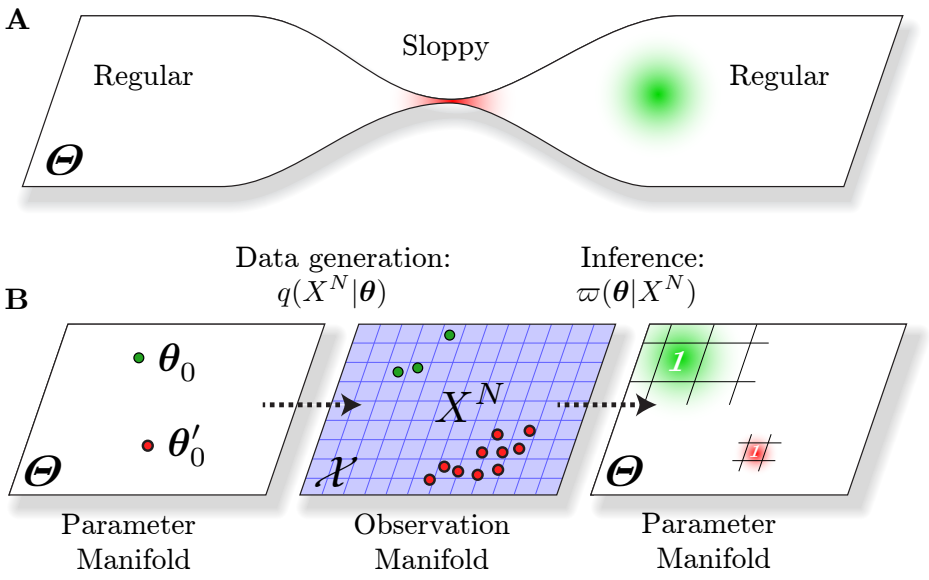


Figure 5.8: **Panel A: Sloppiness is determined by parameter manifold geometry and posterior width.** Parameters are defined on a compact manifold $\boldsymbol{\Theta}$. In sloppy regions of parameter manifold, the parameters are model-structure dominated (red posterior) whereas in regular regions of parameter manifold parameters are data dominated (green posterior). From the perspective of the learning capacity, the model is effectively one dimensional in proximity to the red posterior and two dimensional in proximity to the green posterior. **Panel B: Generalized principle of indifference.** The posterior distribution $\varpi(\boldsymbol{\theta}_0|X^N)$ is shown schematically for two different sample sizes $N$. The resolution increases with sample size as the posterior shrinks. In the GPI prior (Eqn. 5.12), all parameter values consistent with the posterior are assigned unit prior weight collectively.

singular sector $\boldsymbol{\theta}_S$ dimension $K_S$ and assume the improper prior $\varpi = 1$. We can then write:

$$NF \approx NH_0 - K_R \log (NI_R)^{-\frac{1}{2}} - K_S \log v_S + ..., \tag{5.34}$$

where $I_R$ is the $K_R$-root of the determinant Fisher information matrix (Eqn. D.6) projected onto the regular sector. The regular sector give rises to an $N$-dependent $K_R$-dimensional parameter volume $V_R \approx (NI_R)^{-K_R/2}$ whereas the singular sector gives rise to a $N$-independent $K_S$-dimensional parameter volume $V_S = v_S^{K_S}$. The resultant learning capacity is sloppy:

$$\overline{C} = \tfrac{1}{2} K_R \ < \dim \Theta, \tag{5.35}$$

with only the regular parameter coordinates contributing. This scenario is drawn schematically in Fig. 5.8A. The posteriors for two parameter values are shown on the parameter manifold: At a regular point (green), $K_S = 0$ and $K_R = 2$ and all parameter coordinates are regular and data dominated. At the sloppy point (red), the manifold is not rigorously singular but $K_S$ is effectively 1 since the manifold constraints determine the parameter value in the vertical coordinate direction.

In summary, it is parameters inference dominated by the model structure rather than data that in each case give rise to the anomalously small learning capacity and the qualitative phenomenon of anomalously predictive models. The learning capacity has the potential to offer new insights into the mechanism of learning in more complex systems, including ANNs. Our preliminary investigations suggest some training algorithms may map to physical systems with well-understood thermodynamic properties. The detailed physical understanding of the complex phenomenology of physical systems, including phase transitions, renormalization, *etc.*, have great promise for increasing our understanding of the fundamental mechanisms of learning.

### *5.5.2 The generalized principle of indifference*

We argue that a natural approach to objective Bayesian inference is to choose a prior such that the number of indistinguishable distributions is one for all parameter values. This

generalized principle of indifference is simply written $S = 0$. (See Eqn. 5.12.) Schematically, this procedure assigns equal prior weighting to all models that can be distinguished at finite sample size $N$. As the sample size increases, the prior must be modified to accommodate the increased resolution, *i.e.* $(\delta\theta^i)^{-1}$, due to shrinking of the posterior support.

It is important to stress that GPI gives rise to a *sample-size-dependent prior* and therefore this inference is *not* Bayesian in a classical sense: (i) It violates Lindley's dictum: *today's posterior is tomorrow's prior.* (ii) Furthermore, the evidence and prior are no longer interpretable as probabilities but rather statistical weightings. On-the-other-hand, the method codes parameter uncertainty in terms of a posterior probability distribution and facilitates Bayesian parameter and model averaging. Therefore, we would argue the approach maintains all of the attractive features of the Bayesian framework while avoiding problematic aspects. At small sample size or in singular models, the GPI prior must be computed explicitly. (See Fig. 5.8B.) For a regular model in the large-sample-size limit, no calculation is required and GPI prior is equal to the scaled-Jeffreys prior (Eqns. 5.27.)

*Model selection*

The normalization of the GPI prior has significant consequences for inference on model identity (*i.e.* model selection). Returning to the regular model, it is straight forward to apply the Laplace approximation to compute the minus-log evidence using the GPI prior:

$$-\log Z(x^N; w) \approx -\log q(x^N|\hat{\boldsymbol{\theta}}) + K, \tag{5.36}$$

where $K = \dim \Theta$. The scaled-Jeffreys prior cancels the Occam factor from the integration. The two remaining contributions each have clear qualitative interpretations: the MLE estimate of the information $(-\log q)$ and a penalty for model complexity $(K)$. Eqn. 5.36 is already well known as the Akaike Information Criterion (AIC)[2]:

$$-\log Z \approx \mathrm{AIC}(x^N). \tag{5.37}$$

[2]where AIC is defined in nats (rather than the more common demi-nat expression which is twice Eqn. 5.36)

Information-based inference is performed by selecting the model which minimizes AIC, maximizing the estimated predictive performance. The reason why AIC and GPI Bayesian inference are equivalent is most easily understood by rewriting Eqn. 5.12:

$$-N\overline{F} \approx -N\overline{U}, \tag{5.38}$$

which in statistical language corresponds to using a prior that makes the log partition function (LHS) an unbiased estimator of the log predictive performance (RHS). Since the Akaike Information Criterion (AIC) is an unbiased estimator of RHS at large sample size $N$, the generalized principle of indifference encodes an AIC-like model selection [134] and an information-based (AIC) realization of Occam's razor: *parsimony increases predictivity* [30]. The log-predictive performance (RHS Eqn. 5.38) has been advocated in the context of Bayesian model selection through the use of pseudo-Bayes factors by Gelman and coworkers [51, 48, 132, 141, 31].

*Lindley-Bartlett paradox*

In contrast to the unambiguous inference generated by the GPI approach, the canonical uninformative Bayesian approach leads to a number of difficulties and ambiguities, as we discovered in Sec. 5.4.3. These difficulties arise due to the normalization of the prior. We can compute the total number of distinguishable distributions defined by the GPI prior in model $I$ at sample size $N$ by integrating the GPI prior over the parameter manifold:

$$M_I(N) \equiv \int_\Theta \mathrm{d}\boldsymbol{\theta}\ w(\boldsymbol{\theta}; N). \tag{5.39}$$

For non-compact manifolds, $M_I(N)$ may diverge, but this is of no significance. A more detailed discussion of the Bayesian meaning of this procedure is provided in the Appendix Sec. D.1.7.

To explore the significance of this normalization, we will compute the minus-log evidence of a regular model in the large-sample-size limit with a normalized Jeffreys prior. As described above, this prior is equivalent to $w$ normalized by the distribution number

(Eqn. 5.39):

$$\varpi_J(\boldsymbol{\theta}) = M^{-1} w(\boldsymbol{\theta}). \tag{5.40}$$

Using the Laplace approximation, it is straightforward to compute the minus-log-evidence:

$$-\log Z(x^N; \varpi_J) = -\log q(x^N|\hat{\boldsymbol{\theta}}) + \log M + \mathcal{O}(N^0), \tag{5.41}$$

where where the first term (order $N$) has an interpretation of goodness-of-fit, the second term (order $\log N$) is the minus-log Occam factor or log-number of distinguishable distributions (Eqn. 5.39). The corresponding GPI expression is Eqn. 5.36.

The difficulty with using evidence in Eqn. 5.41 as a measure of statistical support is clearest in the context of a simple example: the normal model. Consider the analysis of a normal model with $N$ observations $x^N$ and known variance $\sigma^2$. Consider two competing models: mean is zero (the null hypothesis) and $\mu \neq 0$ (the alternative hypothesis). We must define an acceptable range of values for $\mu$. Assume a flat proper prior on an interval length $L$. The log number of distinguishable distributions is approximately:

$$\log M = \log \tfrac{L}{\delta\mu} \left(\tfrac{1}{2\pi}\right)^{\frac{1}{2}} + ... \tag{5.42}$$

where the error in the mean is $\delta\mu \equiv \sigma/N^{\frac{1}{2}}$. The condition that the evidence is greater for the $\mu \neq 0$ model than for the $\mu = 0$ model is:

$$\hat{\mu} \geq \delta\mu \begin{cases} 2^{\frac{1}{2}}, & \text{GPI} \\ (2\log M)^{\frac{1}{2}}, & \text{normalized} \end{cases}. \tag{5.43}$$

The expression for GPI is independent of the interval length $L$ whereas the normalized expression still retains a dependence on $L$. In fact, in the limit that the prior is uninformative ($L \to \infty$), no finite observed mean is large enough to support the $\mu \neq 0$ model for model selection with a normalized prior. We have described just such an example in Sec. 5.4.3. This automatic rejection of high-dimensional models in the context of uninformative priors is called the Lindley-Bartlett paradox [97, 17]). The use of the GPI prior circumvents this anomalous result.

*Posterior impropriety*

The use of GPI prior often, but not always, gives non-zero evidence for all models under consideration. One such exception is shown in Fig. 5.5 which reveals that the normal model with unknown mean and variance has a divergent effective complexity at a sample size of $N = 1$. The effect of this divergence is to give these models zero statistical weight. Although this may initially appear problematic, it is an important feature of the generalized principle of indifference. A mean and variance cannot be estimated from a single observation and as a result the model parameter posterior would be improper and the predictive loss would be infinite. Therefore the generalized principle of indifference automatically gives this model zero statistical weight ($Z = 0$). The inability of other approaches to automatically handle posterior impropriety is recognized as a significant shortcoming [80].

### *5.5.3 Comparison with existing approaches and novel features*

The generalized principle of indifference subsumes a patchwork of conflicting methods for prior and model selection, resolving many conflicting approaches and generating a single, generally-applicable and self-consistent framework. The GPI approach subsumes the following approaches: (i) For discrete parameter manifolds in the large sample size limit, the GPI gives equal weight to all mutually exclusive models, consistent with the original formulation of the principle of indifference by Bayes and Laplace [93, 81]. (See Eqn. 5.31.) (ii) In the large-sample-size limit, GPI generates a GPI prior proportional to the well-known Jeffreys prior. In this sense, the approach is closely related to the reference prior approach of Bernardo and Berger [19, 23]. (iii) With respect to model selection (inference on model identity), the GPI evidence behaves like pseudo-Bayes factors (or AIC) and therefore circumvents the Lindley-Bartlett paradox. (See Sec. 5.5.2.) To date, the pseudo-Bayes approach has always been un-Bayesian in the sense that the pseudo-Bayes method consists of the *ad hoc* combination of a canonical Bayesian prior for inference on parameters but a cross-validation-based weighting for inference on models. The GPI provides a self-consistent equivalent approach

to inference on both parameters and models.

The GPI addresses a number of problems with existing approaches to objective Bayesian inference. (iv) *Lindley-Bartlett paradox:* As already discussed above, an important shortcoming with existing objective Bayesian approaches relates to the compactness of the parameter manifold and the automatic rejection of higher-dimensional models in model selection (the Bartlett-Lindley paradox [97, 17]). More generally, the evidence of the canonical objective Bayesian approach depends on *ad hoc* modeling decisions, like the range of allowed parameter values. The GPI-Bayes evidence circumvents these anomalies by generating a consistent distribution density $w$ over competing models. As a result the GPI evidence is independent of *ad hoc* modeling decisions. (See Sec. 5.5.2.) (v) *Unification of statistical paradigms:* The absence of the Lindley-Bartlett paradox implies coherent inference between paradigms [35, 91] and therefore the generalized principle of indifference naturally unifies objective Bayesian inference with information-based inference. (vi) *Prior and posterior impropriety:* Another important flaw identified in other objective Bayesian approaches is the inability to handle impropriety. In many cases where parameters are defined on non-compact manifolds, the prior (and sometimes the posterior) cannot be normalized. The redefinition of the prior as a density of models introduces a well-defined and consistent method for defining prior normalization, regardless of the global structure of the manifold. Furthermore, the approach automatically assigns zero statistical weight to models that suffer from posterior impropriety. (See Sec. 5.5.2.) (vii) *Discrete parameter manifolds:* The GPI-Bayes approach also unifies two well-established approaches to defining objective prior on discrete manifolds: equal weight versus Jeffreys prior. GPI Bayes interpolates between these two limits as a function of sample size. (See Eqn. 5.31.) (viii) *Singularity and sloppiness:* Finally, the GPI-Bayes approach does not assume model regularity. It treats singularity and the sloppiness phenomenon in a natural way. (See Sec. 5.4.2.)

### *5.5.4 Conclusion*

Nature reveals an elegant formulation of statistics in the thermal properties of physical systems. Measurements of the heat capacity, compressibility or susceptibility reveal unambiguously how Nature enumerates states and defines entropy. These physical insights provide clues to the definition of novel statistical quantities and the resolutions of ambiguities in the formulation of objective Bayesian statistics. We have refined a previously proposed correspondence between the Bayesian marginal likelihood and the partition function of statistical physics. We demonstrate a novel and substantive mapping between the average energy, heat capacity, entropy and other statistical quantities. The newly-defined learning capacity is a natural quantity for characterizing and understanding learning algorithms and generates new insight into the Akaike Information Criterion (AIC) and model sloppiness through a correspondence with the equipartition theorem and the freeze-out phenomenon, respectively. Finally, we use the Gibbs entropy to define a generalized principle of indifference and an objective Bayesian weighting prior with the property that all distributions have equal prior weight. This approach subsumes many seemingly inconsistent and disparate methods into a single, coherent statistical approach.

# Chapter 6

# CONCLUSION

We wished at the outset to address two major tasks in statistics which were of experimental interest to the Wiggins lab:

- Performing parametric model selection to identify signal against noisy backgrounds
- Represent the uncertainty in the parameters of the selected models.

We wished to perform these tasks with minimal or no subjective input so as to make our analyses as clear and robust to criticism as possible. These aims could not be achieved coherently using existing statistical theory.

Although the Bayesian framework is ideal for expressing parameter uncertainty, it usually leads to unacceptable model performance in the context of model selection when using objective priors. The dependence of the posterior model probabilities on the details of the objective prior and the chosen volume of support lead to subjective and inefficient results.

Nonetheless it is known that AIC can drastically overfit models, and this had been seen as a failure of the predictive model selection paradigm. The much more severe Bayesian penalty on each added dimension ( *infinite* in the Bartlett-paradox limit, and $\frac{1}{2}\log N$ in BIC) is typically seen a reason to embrace the Bayesian paradigm in those circumstances where AIC fails.

But with the development of QIC we had learned that the true *predictive* complexity may in fact be must greater than the model dimension due to the effects of multiplicity. The QIC predictive penalty might need to be as large as $\log N$, twice the BIC penalty! When these multiplicity effects are taken into account, predictive model selection can be even more intolerant of large models as the Bayesian model selection in the regular limit. This lead

us to believe that the current ambiguity in model selection (AIC or BIC?) is unwarranted. Prediction is (almost always) the right thing so long as we account for model singularity. [1]

The discovery of QIC warranted a deeper explanation of the relationship between Bayesian and frequentist penalties and the Lindley-Bartlett paradox. Using our new understanding of the predictive complexity we were able to analyze both Bayes factors and predictive (pseudo-)Bayes factors using the framework of information-based inference. We found several surprising results. First there is a natural indeterminacy in how data is divided in a Bayesian context: information can either be used to construct the prior, or update the prior, (this indeterminacy only effects the model evidence and not the posterior density). Second this indeterminacy is closely related to the choice of $k$ in $k$-fold cross-validation. Third, varying $k$ continuously dials between the BIC and AIC results. This lead to a surprising realization: AIC is the derivative of BIC with respect to sample size.

These results showed that AIC and BIC are fully Bayesian in the fundamental sense, that predictive model selection need not be in conflict a Bayesian treatment of uncertainty and finally that the AIC-type limit (predictive model selection) is optimal from the perspectives of stability, invariance, and resolution for discovery. Although we are not the first to argue for predictive model selection, the literature in the last 20 years heavily favors BIC as a solution. Our discussion will serve as a much needed development in arguing for a predictive (AIC-type) resolution.

### *6.1 The thermodynamics of inference*

As important as these results are from Chapters 3 and 4 —we feel they argue definitively for predictive model selection—the results discussed in Chapter 5 are the most exciting. Once

[1] This picture is complicated slightly in that we would make an exception for the case where we really do have priors for each of the models under our consideration. Fundamentally Bayesian model selection tests the model and the prior *jointly* and therefore is only appropriate if the prior is an essential part of the model — if there is assumed to be a particular generative process for the quenched disorder represented by the distribution parameters $\theta$.

we understood that AIC was the derivative of BIC, or in other words:

$$\text{Prediction} = N\frac{\partial}{\partial N}\log Z(N) \tag{6.1}$$

the connection with the thermodynamic internal energy was clear.

The full thermodynamic treatment of the Bayesian system was an important missing link in the known connections between statistical mechanics and Bayesian inference. It immediately yielded two new concepts: the *learning capacity*, which gives us a generalized method for measuring model dimension, and the thermodynamic Gibbs entropy which we could use to define a generalized principle of indifference. These objects are extremely rich (sometimes prohibitively so, making calculations difficult!) Nearly every problem we have calculated them has brought some unexpected insight into the statistical features of the system.

The generalized principle of indifference (GPOI), which is written in terms of the Gibbs entropy, allowed us to define an objective prior which makes the classical Bayesian model selection procedure consistent with predictive model selection, and extends objective priors to models with discrete or otherwise non-trivial topology. It is a unique addition to the existing family of objective priors and it fully resolves the Lindley-Bartlett paradox.

The learning capacity is a universal tool for measuring model dimension. The mysterious property of systems biology models is that they work at all. We can understand this mystery as being explained by an anomalously small effective dimension, which we can define unambiguously for the first time using the learning capacity.

The learning capacity has already proven itself to be more than of entirely theoretical useful: Jon Craig in the Gundlach lab has been attempting to use it to measure the effective number of rate limiting processes at work in an enzymatic pathway. Nourmohommad, who will be leading the group where I am taking a post-doc at the Max Planck Institute has been thinking about how to use it understand the fitness landscape of a virus coevolving with the human immune system. I believe these ideas will bear tremendous fruit in the coming years, and I have been honored to be a part of their development.

### *6.2 Unanswered questions*

"Every thesis will feel unfinished" Jason Detwiler one of my readers has told me. I feel like I'm just begging to understand the contributions we have made. However I just would like to take the time here to list some of the most intriguing questions which have arisen:

- **Role of Curvature in Complexity** We see the curvature plays an important role in the complexity. We have preliminary results that this role can be quantified in the nearly flat case, for MLE's using perturbation theory. What is the corresponding effect of curvature for a Bayesian analysis? We expect the model response to statistical curvature to be (unlike the predictive performance) highly dependent on whether we use point estimates or bet-hedging strategy. This line of inquiry would be of great interest to the field of information geometry.

- **What is the true nature of $N$?** Although in Chapter 5 we argue that $N$ corresponds to temperature, there is a compelling argument to be made that the temperature should be another free parameter (power scaling the likelihood so that $e^{-\beta N H_x}$), while the sample size actually corresponds to the Grand ensemble particle number and the predictive performance the chemical potential. The interaction between temperature and particle number would have a number of non-trivial models and suggests an entropic treatment for mixture models.

- **Leave-None-Out cross validation.** At the center of the thermodynamic interpretation is a derivative with respect to sample size. We can interpret the derivative with respect to sample size unambiguously using a finite difference approximation. But for many models we have been able to analytically continue to take an true (infinitesimal) derivative! What is the nature of this analytic treatment of an integer-valued parameter? Preliminary investigations are quite promising and connect our results to the sample-moments and unbiased estimation theory.

- **Monte Carlo calculation of $\overline{S}$** For the development of the GPI prior it is necessary to calculate the Gibbs entropy over the entirety of parameter space. For complicated models this remains computationally infeasible. However, certain partial expressions are amenable to Monte-Carlo methods. Is there a way to use these partial expressions to quickly develop expressions for C and S? Again we have several promising preliminary results that suggest there may be empirical estimates for these quantities analogous to the pseudo-Bayes factor result for $U$.

These are only the most developed directions that I have explored. I am excited to see other people take up these ideas and see where they can go. I hope the reader shares my excitement. Thank you.

# BIBLIOGRAPHY

[1] Ken Aho, DeWayne Derryberry, and Teri Peterson. Model selection for ecologists: the worldviews of aic and bic. *Ecology*, 95(3):631–636, 2014.

[2] M. Aitkin. Posterior Bayes factors (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 53:111–142, 1991.

[3] H. Akaike. Information theory and an extension of the maximum likelihood principle. In Petrov and E. Csaki, editors, *2nd International Symposium of Information Theory.*, pages 267–281. Akademiai Kiado, Budapest., 1973.

[4] Hirotugu Akaike. A new look at the bayes procedure. *Biometrika*, 65(1):53–59, 1978.

[5] John Aldrich. Ra fisher and the making of maximum likelihood 1912-1922. *Statistical science*, pages 162–176, 1997.

[6] S. I. Amari. *Differential Geometrical Methods in Statistics.* Springer-Verlag, Berlin., 1985.

[7] Shun-ichi Amari. *Information geometry and its applications.* Springer, 2016.

[8] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

[9] Shun-ichi Amari, Hyeyoung Park, and Tomoko Ozeki. Geometrical singularities in the neuromanifold of multilayer perceptrons. In *Advances in neural information processing systems*, pages 343–350, 2002.

[10] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4(40–79), 2010.

[11] Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

[12] Anthony C Atkinson. Posterior probabilities for choosing a regression model. *Biometrika*, 65(1):39–48, 1978.

[13] Monya Baker. Is there a reproducibility crisis? a nature survey lifts the lid on how researchers view the'crisis rocking science and what they think will help. *Nature*, 533(7604):452–455, 2016.

[14] V. Balasubramanian. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation*, 9:349–368, 1997.

[15] Ole E Barndorff-Nielsen and Peter E Jupp. Statistics, yokes and symplectic geometry. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 6, pages 389–427, 1997.

[16] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

[17] M. S. Bartlett. A comment on D.˜V.˜Lindley's statistical paradox. *Biometrika*, 44(3/4):533–534, 1957.

[18] Matthew J Beal, Francesco Falciani, Zoubin Ghahramani, Claudia Rangel, and David L Wild. A bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21(3):349–356, 2004.

[19] J. O. Berger and J. M. Bernardo. On the development of the reference prior method. Techincal Report 91-15C, Purdue University, 1991.

[20] James O. Berger. Could Fisher, Jeffreys and Neyman have agreed on testing? *Statist. Sci.*, 18(1):1–32, 2003.

[21] James O. Berger, Jose M. Bernardo, and Dongchu Sun. Objective priors for discrete parameter spaces. *Journal Of The American Statistical Association*, 107(498):636–648, 2012.

[22] James O. Berger and Luis R. Pericchi. The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122, 1996.

[23] J. M. Bernardo. *Bayesian statistics 6: Nested Hypothesis Testing: The Bayesian Reference Criterion*. Oxford University Press, 1999.

[24] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory.* Chichester: Wiley., 1994.

[25] S Bernstein. Theory of probability, 1927.

[26] Lucien Birgé and Pascal Massart. Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138(1-2):33–73, 2007.

[27] J Martin Bland and Douglas G Altman. Multiple significance tests: the bonferroni method. *Bmj*, 310(6973):170, 1995.

[28] Hamparsum Bozdogan. Akaike's information criterion and recent developments in information complexity. *Journal of mathematical psychology*, 44(1):62–91, 2000.

[29] Jacob Burbea and C Radhakrishna Rao. Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *Journal of Multivariate Analysis*, 12(4):575–596, 1982.

[30] K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference.* Springer-Verlag New York, Inc., 2nd. edition, 1998.

[31] Kenneth P. Burnham and David R. Anderson. Multimodel inference: Understanding AIC and BIC in model selection. *Socialogical Methods & Research*, 33(2):261–304, 2004.

[32] J. Chen and A. K. Gupta. On change point detection and estimation. *Communications in Statistics–Simulation and Computation*, 30(3):665–697, 2007.

[33] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International Conference on Machine Learning*, pages 1683–1691, 2014.

[34] Garret S Christensen and Edward Miguel. Transparency, reproducibility, and the credibility of economics research. Technical report, National Bureau of Economic Research, 2016.

[35] Robert D. Cousins. The Jeffreys–Lindley paradox and discovery criteria in high energy physics. *Synthese*, pages 1–38, 2014.

[36] Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. *Numerische mathematik*, 31(4):377–403, 1978.

[37] D. A. Darling and P. Erdös. A limit theorem for the maximum of normalized sums of independent random variables. *Duke Math J.*, 23:143–155, 1956.

[38] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3):197–208, 2000.

[39] Sorin Draghici, Purvesh Khatri, Aron C Eklund, and Zoltan Szallasi. Reliability and reproducibility issues in dna microarray measurements. *TRENDS in Genetics*, 22(2):101–109, 2006.

[40] VP Draglia, Alexander G Tartakovsky, and Venugopal V Veeravalli. Multihypothesis sequential probability ratio tests. i. asymptotic optimality. *IEEE Transactions on Information Theory*, 45(7):2448–2461, 1999.

[41] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279. ACM, 2008.

[42] Charles W Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955.

[43] AWF Edwards. Likelihood: an account of the statistical concept of likelihood and its application to scientific inference. 1972.

[44] Bradley Efron. The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632, 2004.

[45] S Fukuda, Y Fukuda, M Ishitsuka, Y Itow, T Kajita, J Kameda, K Kaneyuki, K Kobayashi, Y Koshio, M Miura, et al. Determination of solar neutrino oscillation parameters using 1496 days of super-kamiokande-i data. *Physics Letters B*, 539(3-4):179–187, 2002.

[46] Brent Fultz. Vibrational thermodynamics of materials. *Progress in Materials Science*, 55(4):247–352, 2010.

[47] Seymour Geisser and William F. Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160, 1979.

[48] Alan E. Gelfand and Dipak K. Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 501–514, 1994.

[49] Alan E. Gelfand, Dipak K. Dey, and Hong Chang. Model determination using predictive distributions with implementation via sampling-based methods. Technical report, DTIC Document, 1992.

[50] Andrew Gelman. The connection between varying treatment effects and the crisis of unreplicable research: A bayesian perspective, 2015.

[51] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

[52] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

[53] Christopher R Genovese, Kathryn Roeder, and Larry Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

[54] Jayanta K Ghosh, Subhashis Ghosal, and Tapas Samanta. Stability and convergence of the posterior in non-regular problems. *Statistical Decision Theory and Related Topics V*, page 183, 2012.

[55] Josiah Willard Gibbs. Elementary principles of statistical mechanics. *Compare*, 289:314, 1902.

[56] Leo A Goodman. Some practical techniques in serial number analysis. *Journal of the American Statistical Association*, 49(265):97–112, 1954.

[57] Peter D. Grünwald. *The Minimum Description Length Principle.* MIT, Cambridge, MA, 2007.

[58] Emil Julius Gumbel. Les valeurs extrêmes des distributions statistiques. *Ann. Inst. Henri Poincaré*, 5(2):115–158, 1935.

[59] Katsuyuki Hagiwara, Taichi Hayasaka, Naohiro Toda, Shiro Usui, and Kazuhiro Kuno. Upper bound of the expected training error of neural network regression for a gaussian noise sequence. *Neural Networks*, 14(10):1419–1429, 2001.

[60] E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B.*, 41, 1979.

[61] Enkelejd Hashorva, Zakhar Kabluchko, and Achim Wübker. Extremes of independent chi-square random vectors. *Extremes*, 15(1):35–42, March 2012.

[62] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[63] D. M. Hawkins. Testing a sequence of observations for a shift in location. *Journals of the american statistical association.*, 72(357):180–186, 1977.

[64] David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

[65] Keegan E Hines. Inferring subunit stoichiometry from single molecule photobleaching. *The Journal of general physiology*, 141(6):737–746, 2013.

[66] Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[67] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[68] Lajos Horváth. The maximum likelihood method for testing chnages in the parameters of normal observations. *The Annals of Statistics*, 21(2):671–680, 1993.

[69] Lajos Horváth, Piotr Kokoszka, and Josef Steinebach. Testing for changes in multivariate dependent observations with an application to temperature changes. *Jounral of Multivariate Analysis*, 68:96–119, 1999.

[70] C. M. Hurvich and C. L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76, 1989.

[71] C. M. Hurvich and C. L. Tsai. Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika*, 78:499–509, 1991.

[72] Jun-ichi Inoue. Application of the quantum spin glass theory to image restoration. *Physical Review E*, 63(4):046114, 2001.

[73] John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[74] Ronald Jansen, Haiyuan Yu, Dov Greenbaum, Yuval Kluger, Nevan J Krogan, Sambath Chung, Andrew Emili, Michael Snyder, Jack F Greenblatt, and Mark Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *science*, 302(5644):449–453, 2003.

[75] E. T. Jaynes. *Probability Theory: The Logic of Science.* Cambridge University Press., 2003.

[76] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[77] Edwin T Jaynes. Prior probabilities. *IEEE Transactions on systems science and cybernetics*, 4(3):227–241, 1968.

[78] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[79] R. L. Kashyap. Prior probability and uncertainty. *IEEE Transactions on information theory*, IT-17(6):641–650, 1971.

[80] Robert E. Kass and Larry Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 1996.

[81] J. M. Keynes. *A Treatise on Probability.* Macmillan Limited, London, 1921.

[82] A. Khinchine. Über einen satz der wahrscheinlichkeitsrechnung. *Fundamenta Mathematica*, 6:9–20, 1924.

[83] A. Kolmogoroff. Über das gesetz des iterierten logarithmus. *Mathematische Annalen*, 101:126–135, 1929.

[84] Fumiyasu Komaki. On asymptotic properties of predictive distributions. *Biometrika*, 83(2):299–313, 1996.

[85] Colin H LaMont and Paul A Wiggins. The frequentist information criterion (fic): The unification of information-based and frequentist inference. *stat*, 1050:19, 2015.

[86] Colin H. LaMont and Paul A. Wiggins. The frequentist information criterion (FIC): The unification of information-based and frequentist inference. *Under revision for PNAS. (https://arxiv.org/abs/1506.05855arXiv:1506.05855)*, 2015.

[87] Colin H LaMont and Paul A Wiggins. The development of an information criterion for change-point analysis. *Neural computation*, 28(3):594–612, 2016.

[88] Colin H. LaMont and Paul A. Wiggins. The development of an information criterion for Change-Point analysis with applications to biophysics and cell biology. *Neural Computation*, 28(3):594–612, 2016.

[89] Colin H LaMont and Paul A Wiggins. The lindley paradox: The loss of resolution in bayesian inference. *arXiv preprint arXiv:1610.09433*, 2016.

[90] Colin H LaMont and Paul A Wiggins. A correspondence between thermodynamics and inference. *arXiv preprint arXiv:1706.01428*, 2017.

[91] Colin H. LaMont and Paul A. Wiggins. The lindley paradox: The loss of resolution in bayesian inference. *Under review. (arXiv:1610.09433)*, 2017.

[92] Thomas J Lampo, Stella Stylianidou, Mikael P Backlund, Paul A Wiggins, and Andrew J Spakowitz. Cytoplasmic rna-protein particles exhibit non-gaussian subdiffusive behavior. *Biophysical journal*, 112(3):532–542, 2017.

[93] Pierre Simon Laplace. *Théorie analytique des probabilités*. Courcier, 1820.

[94] Lucien LeCam. On some asymptotic properties of maximum likelihood estimates and related bayes estimates. *Univ. California Pub. Statist.*, 1:277–330, 1953.

[95] D. Leung and Mathias Drton. Order-invariant prior specification in Bayesian factor analysis (http://arxiv.org/abs/1409.7672arXiv:1409.7672).
*it Under review*, 2014.

[96] Faming Liang and Wing Hung Wong. Real-parameter evolutionary monte carlo with applications to bayesian mixture models. *Journal of the American Statistical Association*, 96(454):653–666, 2001.

[97] D. V. Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.

[98] Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

[99] Max A. Little and Nick S. Jones. Generalized methods and solvers for noise removal from piecewise constant signals. i. background theory. *Proc Math Phys Eng Sci*, 467(2135):3088–3114, November 2011.

[100] Max A. Little and Nick S. Jones. Generalized methods and solvers for noise removal from piecewise constant signals. II. new methods. *Proc Math Phys Eng Sci*, 467(2135):3115–3140, November 2011.

[101] Benjamin B. Machta, Ricky Chachra, Mark K. Transtrum, and James P. Sethna. Parameter space compression underlies emergent theories and predictive models. *Science*, 342(6158):604–7, November 2013.

[102] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[103] D. J. C. MacKay. A practical framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

[104] Sarah M Mangiameli, Julie A Cass, Houra Merrikh, and Paul A Wiggins. The bacterial replisome has factory-like localization. *Current genetics*, pages 1–8, 2018.

[105] John I Marden. Hypothesis testing: from p values to bayes factors. *Journal of the American Statistical Association*, 95(452):1316–1320, 2000.

[106] Peter McCullagh. What is a statistical model? *Ann. Statist.*, 30(5):1225–1310, October 2002.

[107] Richard McElreath and Paul E Smaldino. Replication, communication, and the population dynamics of scientific discovery. *PLoS One*, 10(8):e0136088, 2015.

[108] Marc Mézard and Andrea Montanari. Reconstruction on trees and spin glass transition. *Journal of statistical physics*, 124(6):1317–1350, 2006.

[109] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[110] Thomas P Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

[111] Tom Minka et al. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005.

[112] Teppo Mikael Niinimäki and Mikko Koivisto. Annealed importance sampling for structure learning in bayesian networks. In *IJCAI*, pages 1579–1585, 2013.

[113] Hidetoshi Nishimori. *Statistical physics of spin glasses and information processing: an introduction*, volume 111. Clarendon Press, 2001.

[114] Anthony O'Hagan. Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 99–138, 1995.

[115] E. S. Page. A test for a change in a parameter occuring at an unknown point. *Biometrika*, 42:523–527, 1955.

[116] E. S. Page. On problems in which a change in a parameter occurs at an unknown point. *Biometrika*, 44:248–252, 1957.

[117] R. K. Pathria and P. D. Beale. *Statistical Mechanics*. Elsevier Science, 1996.

[118] Juho Piironen and Aki Vehtari. Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735, 2017.

[119] Florian Prinz, Thomas Schlange, and Khusru Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*, 10(9):712–712, 2011.

[120] MA RA Fisher. On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A*, 222(594-604):309–368, 1922.

[121] F. Reif. *Statistical Physics*. McGraw-Hill (New York), 1967.

[122] Daniel Revuz and Marc Yor. *Continuous Martingales and Brownian Motion*. Springer-Verlag, New York, 2nd edition, 1999.

[123] J. Rissanen. Modeling by the shortest data description. *Automatica*, 14:465–471, 1978.

[124] Timothy D Ross. Accurate confidence intervals for binomial proportion and poisson rate estimation. *Computers in biology and medicine*, 33(6):509–531, 2003.

[125] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[126] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–4, 1978.

[127] Jun Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

[128] Ritei Shibata. An optimal selection of regression variables. *Biometrika*, 68(1):45–54, 1981.

[129] Ritei Shibata. Bootstrap estimate of Kullback-Leibler information for model selection. *Statistica Sinica*, 7:375–394, 1997.

[130] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *Information Theory, IEEE Transactions on*, 26(1):26–37, January 1980.

[131] Adrian FM Smith and David J Spiegelhalter. Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 213–220, 1980.

[132] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, B64:583–639, 2002.

[133] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[134] M. Stone. An asymptotic equivalence of choice of model by cross-validation and Akaike's Criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1977.

[135] Stella Stylianidou, Thomas J Lampo, Andrew J Spakowitz, and Paul A Wiggins. Strong disorder leads to scale invariance in complex biological systems.

[136] Meysam Tavakoli, J. Nicholas Taylor, Chun-Biu Li, Tamiki Komatsuzaki, and Steve Pressé. Single molecule data analysis: An introduction. *arXiv preprint arXiv:1606.00403*, 2016.

[137] Mark K. Transtrum, Benjamin B. Machta, Kevin S. Brown, Bryan C. Daniels, Christopher R. Myers, and James P. Sethna. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *J Chem Phys*, 143(1):010901, July 2015.

[138] Jay J Van Bavel, Peter Mende-Siedlecki, William J Brady, and Diego A Reinero. Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, 113(23):6454–6459, 2016.

[139] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Efficient implementation of leave-one-out cross-validation and WAIC for evaluating fitted bayesian models. *arXiv preprint arXiv:1507.04544*, 2015.

[140] Aki Vehtari and Jouko Lampinen. Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, 14(10):2439–2468, 2002.

[141] Aki Vehtari and Janne Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

[142] Duane C Wallace. *Statistical physics of crystals and liquids: a guide to highly accurate equations of state*. World Scientific, 2002.

[143] S. Watanabe. *Algerbraic geometry and statistical learning theory*. Cambridge Univeristy Press, 2009.

[144] Paul A. Wiggins. An information-based approach to change-point analysis with applications to biophysics and cell biology. *Submitted to Biophys J.*, 2015.

[145] Paul A. Wiggins. An information-based approach to change-point analysis with applications to biophysics and cell biology. *In preparation*, 2015.

[146] Wikipedia. Gumbel distribution — wikipedia, the free encyclopedia, 2015. [Online; accessed 19-May-2015].

[147] Wikipedia. Law of the iterated logarithm — wikipedia, the free encyclopedia, 2015. [Online; accessed 19-May-2015].

[148] Eric P Xing, Michael I Jordan, and Stuart Russell. A generalized mean field algorithm for variational inference in exponential families. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 583–591. Morgan Kaufmann Publishers Inc., 2002.

[149] Y. Yang. Can the strengths of AIC and BIC be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

[150] Yuhong Yang. Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

[151] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

# Appendix A

# QIC FOR CHANGE-POINT ALGORITHMS AND BROWNIAN BRIDGES

## *A.1 Type I errors*

In terms of the Cummulative Probability Distribution (CDF), the probability of a false positive change-point is:

$$\alpha = 1 - F_U(2\overline{U}), \tag{A.1}$$

where $U$ is the relevant change-point statistic and $\overline{U}$ is its expectation. Using the local binary-segmentation algorithm, $\alpha$ corresponds to the probability of a false positive per data partition and the change-point statistic is defined by Eqn. 41 (in the main text) evaluated at the average partition length $N_p \equiv \frac{N}{n}$. The false positive change-point acceptance probability is plotted in Figure A.1.

The analogous false positive rate for the global binary-segmentation algorithm describes the probability of a false positive in the entire data set, including all partitions. In this cases, we use the change-point statistic defined by Eqn. 45 (in the main text).

### *A.1.1 Asymptotic form of the complexity function*

In order to discuss the scaling of the complexity relative to the BIC complexity, we need to derive an asymptotic form for the complexity in the large $N$ limit. We do not recommend explicitly using this asymptotic expression for the complexity for Change-Point Analysis since it converges to the true complexity very slowly, especially for large $d$.

First let us consider related results for and Brownian walk rather than a Brownian bridge.

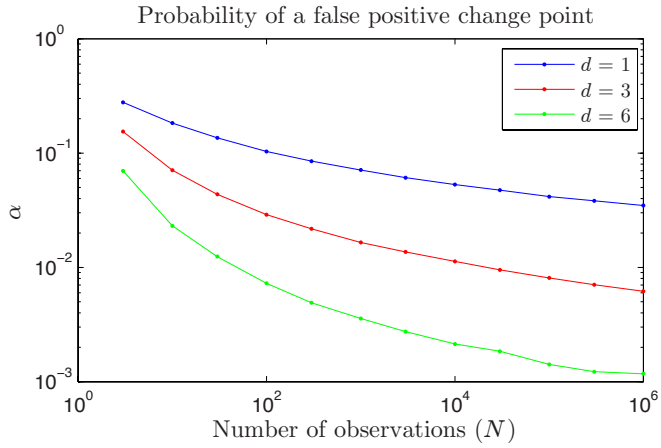


Figure A.1: **Probability of a false positive change-point.** The probability of a false positive change-point is shown as a function of the number of observations in the interval length $N$ for three different model dimensions.

Let us define $S_n$ as follows:

$$S_n \equiv |\boldsymbol{Z}_n| \tag{A.2}$$

$$\boldsymbol{Z}_n \equiv \sum_{i=1}^{n} \boldsymbol{z}_i \tag{A.3}$$

where the $\boldsymbol{z}_i$ are independent normally-distributed random variables with mean zero variance one per dimension $\boldsymbol{d}$. The Law of Iterated Logs states that [82, 83, 147]:

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{n \log\log n}} = \sqrt{2} \quad \text{a.s.}, \tag{A.4}$$

where a.s. is the acronym for almost surely. (See Figure A.2.) This behavior of $S_n$ is described in more detail by the Darling-Erdös Theorem [37]. Let us define a new random variable

$$U'(N_p) \equiv \max_{1\le n\le N_p} \frac{S_n}{\sqrt{n}}, \tag{A.5}$$

in $\boldsymbol{d} = 1$ dimensions, the asymptotic cumulative distribution of $U'$ approaches the cumulative

distribution for a Gumbel Distribution [37]:

$$\lim_{N_p\to\infty} \Pr\left[U' < \beta t + u\right] = \exp\left[-e^{-t}\right], \tag{A.6}$$

$$\beta(N_p) \equiv (2\log\log N_p)^{-1/2}, \tag{A.7}$$

$$u(N_p) \equiv \beta\left[\beta^{-2} + \tfrac{1}{2}\log\log\log N_p - \log 2\pi^{1/2}\right], \tag{A.8}$$

where Pr denotes probability and the distribution parameters $u$ and $\beta$ are called the location and scale respectively and the average partition length is $N_p \equiv \frac{N}{n}$. Let us introduce the cumulative distribution function for $U$:

$$F_U(U) \equiv \Pr\left[U'_{(n)} < U\right]. \tag{A.9}$$

This expression can be reordered to put it in the canonical form of the Gumbel Distribution [58, 146]:

$$F_U(U) = \exp\left[-\exp\left(-\tfrac{U-u}{\beta}\right)\right], \tag{A.10}$$

We can then use the well known expression in terms the cdf to compute the cdf of the maximum of $n$ random variables $U'$:

$$\Pr\left[U'_{(n)} < U_{(n)}\right] = F_U^n(U), \tag{A.11}$$

$$= \left(\exp\left[-\exp\left(-\tfrac{U-u}{\beta}\right)\right]\right)^n, \tag{A.12}$$

$$= \exp\left[-\exp\left(-\tfrac{U-u_n}{\beta}\right)\right], \tag{A.13}$$

where

$$u_n \equiv u + \beta\log n. \tag{A.14}$$

The mean and variance of the Gumbel Distribution are well known, allowing us to compute the expectation of ${U'}^2_{(n)}$:

$$\mathbb{E}_x\, {U'}^2_{(n)} \approx (u_n + \cancel{\beta\gamma})^2 + \cancel{\frac{\pi^2}{6}\beta^2}, \tag{A.15}$$

$$\approx 2\log\log N_p + 2\log n + ... \tag{A.16}$$

where $\gamma$ is the Euler-Mascheroni constant and we have used the cancel notation to show which terms have been dropped to lowest order. In the second line, we have written the expression to lowest order in $N$ and $n$.

Horváth has generalized the Darling-Erdös Theorem for a Brownian bridge in $\cancel{d}$ dimensions for the application to Change-Point Analysis in the context of the LPT test [68, 69]. The generalized expression for the cumulative distribution leads to a change in the expression for $u$ only:

$$u_d(N_p) \equiv \beta \left[\beta^{-2} + \tfrac{d}{2} \log\log\log N_p - \cancel{\log \Gamma\left(\tfrac{d}{2}\right)}\right] \tag{A.17}$$

where $\Gamma$ is the Gamma Function. We drop the last term since it is not leading order for large $N_p$. We now follow the same steps to generate the distribution for the maximum of $n$ random variables $U'$, leading to a new Gumbel Distribution with location $\mu_{n,d}$:

$$u_{n,d}(N_p) = \beta \left[\beta^{-2} + \tfrac{d}{2} \log\log\log N_p + \log n\right] \tag{A.18}$$

We now recompute the expectation for $\cancel{d}$ dimensions:

$$\begin{aligned}
\cancel{k}_{\mathrm{G}-}(N_p, n, \cancel{d}) &\equiv \mathbb{E}_x \; U'^2_{(n)}(N_p, n, \cancel{d}), && \text{(A.19)}\\
&\approx (u_{n,d} + \beta\gamma)^2 + \frac{\pi^2}{6}\beta^2 && \text{(A.20)}\\
&\approx 2\log\log N_p + 2\log n + \cancel{d}\log\log\log N_p + ... && \text{(A.21)}
\end{aligned}$$

where we have kept terms only to highest order in $n$ and $N_p$.

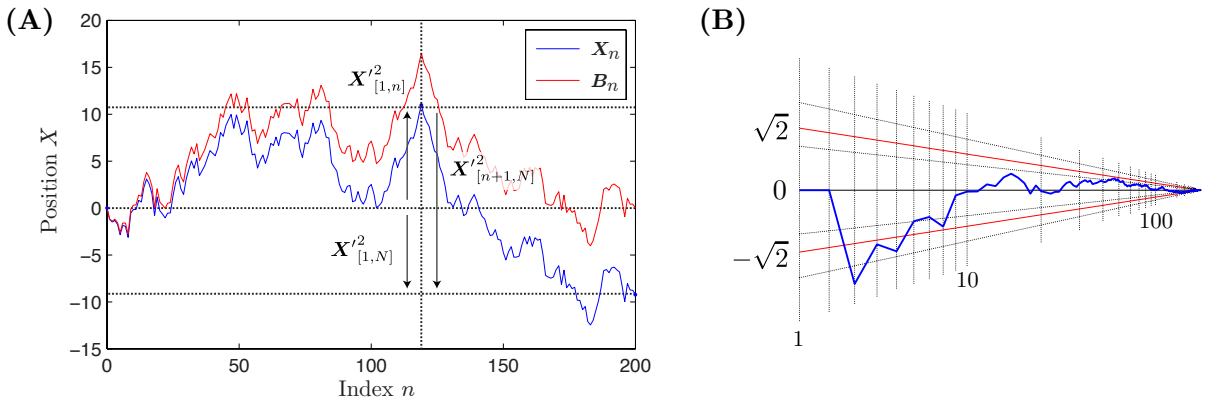


Figure A.2: **Panel A: Brownian Walk and Brownian Bridge.** A visualization of a random walk $\boldsymbol{X}'_[1, n]$ (blue) and the corresponding Brownian bridge $\boldsymbol{B}'_n$ (red). **Panel B: Law of Iterated Logs.** A visualization of $S_n/\sqrt{n \log\log n}$ (blue) plotted as an orthographic projection as a function of $n$. $\sqrt{2}$ (red) is the limit of the supremum.

**Global Binary-Segmentation Algorithm**

A global algorithm for binary segmentation. The information $\hat{h}$ is implicitly evaluated at the MLE state-model parameters $\hat{\boldsymbol{\Theta}}$.

1. Initialize the change-point vector: $\boldsymbol{i} \leftarrow \{1\}$

2. Segment model $\hat{\Theta}^n \rightarrow \hat{\Theta}^{n+1}$:

   (a) Compute the entropy change that results from all possible new change-point indices $j$:

   $$\Delta h_j \leftarrow \hat{h}(X|\{i_1, ..., j, ..., i_n\}) - \hat{h}(X|\boldsymbol{i}), \tag{A.22}$$

   (b) Find the minimum information change $\Delta h_{\min}$, and the corresponding index $j_{\min}$.

   (c) **If** the information change plus the nesting complexity is less than zero:

   $$\Delta h_{\min} + \not{k}_{G-} < 0 \tag{A.23}$$

   **then** accept the change-point $j_{\min}$

      i. Add the new change-point to the change-point vector.

      $$\boldsymbol{i} \leftarrow \{i_1, ..., j_{\min}, ...i_{n+1}\} \tag{A.24}$$

      ii. Segment model $\hat{\Theta}^{n+1}$

   (d) **Else** terminate the segmentation process.

**Local Binary-Segmentation Algorithm**

A local algorithm for binary segmentation. The information $\hat{h}$ is implicitly evaluated at the MLE state model parameters $\hat{\boldsymbol{\Theta}}$.

1. Initialize the change-point vector: $\boldsymbol{i} \leftarrow \{1\}$, $I \leftarrow 1$.

2. Segment model $\hat{\Theta}^n$ on state $I$:

   (a) Compute the entropy change that results from all possible new change-point indices $j$ on the interval $[i_I, i_{I+1})$:

   $$\Delta h_j \leftarrow \hat{h}(X|\{...i_I, j, i_{I+1}, ...\}) - \hat{h}(X|\boldsymbol{i}), \tag{A.25}$$

   (b) Find the minimum information change $\Delta h_{\min}$, and the corresponding index $j_{\min}$.

   (c) **If** the information change plus the nesting complexity is less than zero:

   $$\Delta h_{\min} + \not{k}_{\mathrm{L}-} < 0 \tag{A.26}$$

   **then** accept the change-point $j_{\min}$

      i. Add the new change-point to the change-point vector.

      $$\boldsymbol{i} \leftarrow \{..., i_I, j_{\min}, i_{I+1}, ...\} \tag{A.27}$$

      ii. Segment model $\Theta^{n+1}$ on states $I$ and $I+1$.

      iii. Merge the resulting index lists.

   (d) **Else** terminate the segmentation process.

# Appendix B

# QIC CALCULATIONS

## *B.1 Complexity Calculations*

Here are some of the descriptions of the calculations we performed.

### *B.1.1 Exponential families*

An important case is the exponential-family, where the likelihood can be written:

$$q(x|\theta) = \exp[t(x) \cdot \theta - N\psi(\theta) + r(x)], \tag{B.1}$$

the sufficient statistics $t(x)$ and function $r(x)$ are functions of the dataset $x$ only and $\psi(\theta)$ is a function of the parameters only and $N$ is the sample size. In this case, the complexity can be computed from Eqn. (3.8) and can be written:

$$\mathcal{K}(\theta) = \mathop{\mathbb{E}}_{X,Y|\theta} [t(X) - t(Y)] \cdot \hat{\theta}_X, \tag{B.2}$$

where $X$ and $Y$ are two independent datasets of sample size $N$ generated from distribution $q(\cdot|\theta)$ and

$$\hat{\theta}_X \equiv (\nabla\psi)^{-1}[t(X)/N], \tag{B.3}$$

where $(\nabla\psi)^{-1}$ is the functional inverse of the gradient of function $\psi$.

### *B.1.2 Modified-Centered-Gaussian distribution*

The likelihood for the Modified-Centered-Gaussian model is given by Eqn. (3.13). The MLE parameters and sufficient statistic are:

$$\hat{\theta}_x = -\frac{N}{\alpha\, t(X)} \tag{B.4}$$

$$t(X) = -\sum_{i=1}^{N} |x|^{\alpha}, \tag{B.5}$$

respectively. The sufficient statistic $t$ is distributed like a Gamma distribution:

$$-t \sim \Gamma(N/\alpha, \lambda), \tag{B.6}$$

which has well-known moments:

$$\overline{(-t)^m} = \tfrac{\Gamma(m+N/\alpha)}{(-\lambda)^m \Gamma(N/\alpha)}. \tag{B.7}$$

Using the last results in combination with expression for the complexity of an exponential model, Eqn. (B.2), we find:

$$\mathscr{K} = \tfrac{N}{N-\alpha} \quad \text{for} \quad N > \alpha, \tag{B.8}$$

which is always larger than the AIC complexity $K = 1$ for $\alpha > 0$.

### *B.1.3 The component selection model*

For convenience, consider a true model where $j = n$, which is general due to permutation symmetry. Let the observations be defined as:

$$X_j = \xi_j + [j = n]\mu, \tag{B.9}$$

where we have used the Iverson bracket and the $\xi_j$ are iid random variables centered around zero with unit variance. The MLE parameters for the model are:

$$\hat{\imath} = \arg\max_j X_j^2, \tag{B.10}$$

$$\hat{\mu} = X_{\hat{\imath}}. \tag{B.11}$$

The complexity can then be written:

$$\mathcal{K}(\theta) = \mathop{\mathbb{E}}_{\xi} \left\{ \max_j X_j^2 - \mu^2[\hat{\imath} = n] \right\}, \tag{B.12}$$

which can be computed using one-dimensional integrals of the CDFs.

It is useful to consider the large and small multiplicity limit. For large multiplicity ($n$), the complexity is

$$\mathcal{K}(\theta) = 2\log n - \log\log n - 2\log\Gamma(\tfrac{1}{2}) + 2\gamma + ..., \tag{B.13}$$

where $\gamma$ is the EulerMascheroni constant [61]. For $n = 1$ or sufficiently large $\mu$, there is no multiplicity and we recover the AIC result:

$$\mathcal{K}(\theta) = 1. \tag{B.14}$$

### *B.1.4 n-cone*

Following notation used in special relativity, we denote the *space-like* component of a vector $\vec{A} = \{A_2, \dots A_n\}$ and the *time-like* component $A_1$. The implicit function of constraint is

$$\rho(\theta) = \vec{\mu}^2 - (c\mu_1)^2 = 0, \tag{B.15}$$

which is to say that the mean must lie on the *light cone*. The observations $X = (X_1, \vec{X})$ can be represented as:

$$X = \mu + \xi, \tag{B.16}$$

where $\xi$ is an $n$-vector of iid random variables normally around zero with unit variance. The MLE parameters satisfying the constraints are:

$$\hat{\vec{\mu}} = \frac{\left(\frac{c|X_1|}{|\vec{X}|} + c^2\right)}{c^2 + 1}\vec{X}, \tag{B.17}$$

$$\hat{\mu}_1 = \frac{X_1 + c\,\mathrm{sgn}(X_1)\left|\vec{X}\right|}{c^2 + 1}. \tag{B.18}$$

We can take the expectation using known properties of the non-central $\chi$ distribution. The result can be expressed in terms of the generalized Laguerre polynomials:

$$\mathscr{K}(\theta) = \frac{c^2(k-1) - c\vec{\mu}^2\left(\sqrt{\frac{\pi}{2}}\mu_1 \operatorname{erf}\left(\frac{\mu_1}{\sqrt{2}}\right) + e^{-\frac{\mu_1^2}{2}}\right) L_{-\frac{1}{2}}^{\frac{k-1}{2}}\left(-\frac{\vec{\mu}^2}{2}\right)}{c^2+1} + \frac{c\left(\sqrt{\frac{\pi}{2}}\mu_1 \operatorname{erf}\left(\frac{\mu_1}{\sqrt{2}}\right) + 2e^{-\frac{\mu_1^2}{2}}\right) L_{\frac{1}{2}}^{\frac{k-3}{2}}\left(-\frac{\vec{\mu}^2}{2}\right) + 1}{c^2+1}. \tag{B.19}$$

This result recovers the known results of AIC on the realizable surface far from the singularity, and $\mathscr{K} = 1$ when c is very large, corresponding to a needle-like geometry where the surface of constraint is essentially one-dimensional compared to the scale of the Fisher information.

### *B.1.5 Fourier Regression nesting complexity*

A literal treatment of the QIC algorithm requires a Monte Carlo simulation. However, as can be seen in Fig. 1, this complexity interpolates between two limiting behaviors that can be treated analytically. To treat the nesting complexity analytically, we will make two assumptions: (i) All previously included models are unambiguously resolved and (ii) the number of modes included is small compared to the total $n$. Under these two assumptions, the nesting complexity is equivalent to selecting the largest magnitude coefficient of the remaining unselected Fourier components. Since each is independent and normally distributed, this problem is exactly equivalent to a problem that we have already analyzed: the component selection model. In this case, we can simply reuse the complexity derived in Eqn. (B.13) as the nesting complexity, with limiting behavior:

$$k_i(\theta) = \begin{cases} 2\log N & \text{when } \mu^2 \ll 2\log N \\ 1 & \text{when } \mu^2 \gg 2\log N \end{cases}, \tag{B.20}$$

in exact analogy to Eqn. (B.13) where the number of components $n = N$. The total complexity can be summed,

$$\mathscr{K}(\theta) = \sum_i k_i(\theta). \tag{B.21}$$

We have previously used this approximation in the context of change-point analysis [88, 144].

### *B.1.6 $L_1$ Constraint*

We use the simplex projection algorithm described in **(author?)** [41] with the MATLAB code to project onto an $L_1$ ball provided by John Duchi at `https://stanford.edu/~jduchi/projects/DuchiShSiCh08.html`. We computed the complexity using $10^5$ samples on a $10^{-1}$ grid, with the resulting complexity linearly filtered in Fourier space.

### *B.1.7 Curvature and QIC unbiasedness under non-realizability*

If $|\theta_X - \theta_0|$ is small (on average) relative to the inverse-mean-curvatures of the manifold $\boldsymbol{\Theta}$, then we have that the true complexity is given by

$$\mathcal{K}(\phi) \approx \mathop{\mathbb{E}}_{X|\phi} \left\{ D(\theta_0||\theta_X) - d_X(\theta_0||\theta_X) \right\} \tag{B.22}$$

This follows from Amari's "generalized pythagorean theorem" [8, 6] where $D(\phi||\theta)$ is analogous to the half-squared-distance between $\phi$ and $\theta$. If $\hat{\theta}_X$ is a MLE then $d_X(\theta_0||\theta_X)$ is equivalent to another K-L divergence [8]. We can finally write this as

$$\approx \mathop{\mathbb{E}}_{\theta_X|\phi} \left\{ D(\theta_0||\theta_X) + D(\theta_X||\theta_0) \right\}. \tag{B.23}$$

For (nearly) flat manifolds, such as the unconstrained exponential family, with $\hat{\theta}_X$ being the MLE, we do not need the distribution of the *data* $X|\phi$ to be well approximated by $X|\theta_0$, we only need the distribution of the *fitted parameters* to match $\mathop{\mathbb{E}}_{\theta_X|\phi} \approx \mathop{\mathbb{E}}_{\theta_X|\theta_0}$.

$$\mathcal{K}(\phi) \approx \mathop{\mathbb{E}}_{\theta_X|\theta_0} \left\{ D(\theta_0||\theta_X) + D(\theta_X||\theta_0) \right\} \tag{B.24}$$

$$= \mathcal{K}(\theta_0) \tag{B.25}$$

In which case the model is *effectively* realizable for our purposes in the sense that $\mathcal{K}(\theta_0)$ is unbiased, even though $D(\phi||\theta_0)$ may be large. Eqn. (3.36) and the subsequent considerations then apply.

We would expect QIC to be biased if $\theta_0$ poorly describes the variance of $\theta_X$. For instance, if we assume a fixed, incorrect, variance $\sigma'^2$, instead of the true value of $\sigma$, this will bias the QIC complexity by a scale factor of $\sigma^2/\sigma'^2$. Although we'd expect Eq. B.22 to be very generally asymptotically true, our complexity landscapes show that the presence or absence of extrinsic curvature of $\Theta$ is an important factor in whether or not the variance of $\theta_X$ will be well estimated by $\theta_0$. When the true distribution is not realizable, the variance of $\theta_X$ will depend on the curvature, and QIC may have significant bias.

### *B.1.8 Approximations for marginal likelihood*

A second canonical information criterion (BIC) is motivated by Bayesian statistics. In Bayesian model selection, the canonical approach is to select the model with the largest marginal likelihood:

$$q(x) \equiv \int_\Theta \mathrm{d}\boldsymbol{\theta}\ \varpi(\boldsymbol{\theta})\ q(x|\boldsymbol{\theta}), \tag{B.26}$$

where $\varpi$ is the prior probability density of parameters $\boldsymbol{\theta}$. If we assume (i) the large $N$ limit, (ii) that the model is regular, (iii) the model dimension is constant as $N$ increases and (iv) the prior is uninformative, the negative log of the marginal likelihood can be computed using the Laplace approximation [126, 30]:

$$-\log q(x) = h(x|\hat{\boldsymbol{\theta}}_x) + \tfrac{1}{2}K\log N + \log\frac{\sqrt{(2\pi)^K \det \boldsymbol{I}}}{\varpi(\hat{\boldsymbol{\theta}}_x)} + ... \tag{B.27}$$

where $K$ is the dimension of the model and $\boldsymbol{I}$ is Fisher Information Matrix. The first three terms have $N^1$, $\log N$ and $N^0$ scaling with sample size $N$, respectively. A canonical approach is to keep only the first two terms of the negative log of the marginal likelihood, which define the Bayesian Information Criterion (BIC):

$$\mathrm{BIC}(x) = h(x|\hat{\boldsymbol{\theta}}_x) + \tfrac{1}{2}K\log N, \tag{B.28}$$

which has the convenient property of dropping the prior dependence since it is constant order in $N$ [126, 30]. The BIC complexity grows with sample size and is therefore larger than the AIC complexity in the large $N$ limit. This tends to lead to the selection of smaller models

than AIC. Since the prior typically depends on *ad hoc* assumptions about the system, the absence of prior dependence is an attractive feature of BIC. On-the-other-hand, in many practical analyses $\log N$ is not large, which makes the canonical interpretation of BIC dubious. A more palatable interpretation of BIC is to imagine withholding a minimal subset of the data (*i.e.* $N \approx 1$) to generate an informative prior, then computing marginal likelihood. This sensible Bayesian procedure is well approximated by BIC [89].

### *B.1.9 Seasonal dependence of the neutrino intensity*

*Analysis of the data*

We expand the model mean ($\mu_i$) and observed intensity ($X_i$) in Fourier coefficients $\tilde{\mu}_i$ and $\tilde{X}_i$ respectively. The MLE parameters that minimize the information are $\hat{\tilde{\mu}}_i = \tilde{X}_i$. We now introduce two different approaches to encoding our low-level model parameters $\{\tilde{\mu}_i\}_{i=-N/2\ldots N/2}$: The *Sequential* and *Greedy Algorithms*. Note that in both cases, the models will be

*Analysis of the data*

We expand the model mean ($\mu_i$) and observed intensity ($x_i$) into Fourier coefficients $\tilde{\mu}_i$ and $\tilde{X}_i$ respectively:

$$\mu_j = \sum_{i=-N/2}^{N/2} \tilde{\mu}_i \psi_i(j) \quad \text{where} \quad \tilde{\mu}_i = \sum_{j=1}^{N} \mu_j \psi_i(j), \tag{B.29}$$

$$x_j = \sum_{i=-N/2}^{N/2} \tilde{X}_i \psi_i(j) \quad \text{where} \quad \tilde{X}_i = \sum_{j=1}^{N} x_j \psi_i(j), \tag{B.30}$$

where the orthonormal Fourier basis functions are defined:

$$\psi_i(j) \equiv N^{-1/2} \begin{cases} \sqrt{2}\,\cos(2\pi ij/N), & i < 0 \\ 1, & i = 0 \\ \sqrt{2}\,\sin(2\pi ij/N), & i > 0. \end{cases} \tag{B.31}$$

Substituting these expressions into the expression of the data-encoding information gives

$$h(X^N|\boldsymbol{\theta}) = \frac{N}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=-N/2}^{N/2}(\tilde{X}_i - \tilde{\mu}_i)^2, \tag{B.32}$$

where we have used the orthagonality in the large $N$ limit for all terms. We chose the eigenfunction normalization in order to give this expression its concise form.

Note that there is no need to (re)compute the information *etc* since the structure of the problem is identical to the resonance problem discussed above.

# Appendix C

# LINDLEY PARADOX

## *C.1 Calculation of Complexity*

In the large sample-size limit of a regular model, we can view a model as a flat prior on some subspace $\Theta$ of dimension $K$ embedded in a larger parameter space $J + K$. The marginal likelihood of $N$ measurements with dimension $J + K$ can then be written as;

$$q(X^N) = \int \mathrm{d}^J\theta_\perp \, \delta^J(\theta_\perp - \theta_\perp^0) \int \mathrm{d}^K\theta_\parallel \, \frac{\exp \frac{-\sum_i (X^i-\theta)^2}{2\sigma^2}}{(2\pi\sigma^2)^{N(K+J)/2}} \tag{C.1}$$

Which gives for the predictive distribution,

$$h(X^{N_G}|X^T) = h(X^N) - h(X^{N_T}) \tag{C.2}$$

$$= \frac{\sum_i (X_\perp^i - \theta_\perp^0)^2}{2\sigma^2} + \frac{N_G(J+K)}{2} \log 2\pi\sigma^2 \tag{C.3}$$

$$+ \frac{K}{2} \log \frac{N}{N_T} + \frac{S_N - S_{N_T}}{2\sigma^2}$$

Where $S_{N_T}$, is the sum of (projected) squared deviations from the (projected) mean $\mu_{N_T} = N_T^{-1} \sum_{i \in N_T} X_\parallel^i$. A straightforward calculation shows that

$$S_N - S_{N_T} - S_{N_G} = \frac{N_T N_G}{N} \left(\hat{\mu}_{N_T} - \hat{\mu}_{N_G}\right)^2 \tag{C.4}$$

Expanding around $\theta_0$, where $\theta_0$ is the parameter in the manifold $\Theta$ minimizing the KL divergence from the true distribution $p(\cdot)$,

$$h(X^{N_G}|X^T) = h(X^{N_G}|\theta_0) + \frac{K}{2} \log \frac{N}{N_T} - \frac{N_G^2}{N} \frac{\left(\hat{\mu}_{N_G} - \theta_0\right)^2}{2\sigma^2} + \tag{C.5}$$

$$\frac{N_G N_T}{N} \frac{\left(\hat{\mu}_{N_T} - \theta_0\right)^2 - 2\left(\hat{\mu}_{N_G} - \theta_0\right)\left(\hat{\mu}_{N_T} - \theta_0\right)}{2\sigma^2}.$$

The deviance terms cancel under expectation. After rescaling, we can write,

$$\frac{N}{N_G}\overline{h(X^{N_G}|X^T)} = \overline{h(X^N|\theta_0)} + \frac{K}{2}\frac{N}{N_G}\log\frac{N}{N_T}. \tag{C.6}$$

Defining $\nu = N_G/N_T$ emphasizes the limit behavior

$$\frac{N}{N_G}\overline{h(X^{N_G}|X^T)} = \overline{h(X^N|\theta_0)} + \frac{K}{2}(1+\nu^{-1})\log(1+\nu). \tag{C.7}$$

The term $\overline{h(X^N|\theta_0)}$ is estimated by the observed information at MLE $h(x^N|\hat{\theta}_x)$. The error in this estimator (training error) is again $\frac{1}{2}\chi^2(K)$ distributed [3], making the following estimator unbiased

$$\overline{h(X^N|\theta_0)} \hat{=} h(x^N|\hat{\theta}_x) + \frac{K}{2}. \tag{C.8}$$

When Eqn. C.8 is used with Eqn. C.7, it gives us an information criterion corresponding to the pseudo-Bayes Factor for each partition choice $\nu$.

## *C.2 Definitions and Calculations*

### *C.2.1 Volume of a distribution*

Our intuitions about the volume of a distribution can be made mathematically precise using the self-entropy $S$. The self entropy functional is defined as

$$S[q(\cdot)] \equiv -\int \mathrm{d}\theta\, q(\theta)\log q(\theta), \tag{C.9}$$

and the volume is defined in turn as

$$V_q \equiv e^{-S[q(\cdot)]}. \tag{C.10}$$

For uniform distributions, this entropic definition reduces to the volume of the support. For a normal distribution of dimension K the volume is

$$V_\Sigma = (2\pi e)^{\frac{K}{2}}|\Sigma|^{\frac{1}{2}} \approx \left(2\pi e\sigma^2\right)^{\frac{K}{2}}. \tag{C.11}$$

where the second equality only holds if $\Sigma$ is proportional to the identity.

### *C.2.2 Showing $I = I'$ to zeroth order*

The first term in the information difference

$$I'(x^N) = \hat{H}^{1|N-1} - \hat{H}^{N|0} \tag{C.12}$$

$$= \left( \mathop{\mathbb{E}_\theta}_{\varpi(\cdot|x^N)} h(x^N|\theta) \right) - h(X^N) + \mathcal{O}(N^0) \tag{C.13}$$

$$= \mathop{\mathbb{E}_\theta}_{\varpi(\cdot|x^N)} \log \frac{q(x^N|\theta)}{q(x^N)} + \mathcal{O}(N^0). \tag{C.14}$$

By multiplying the numerator and denominator by $\varpi(\boldsymbol{\theta})$, we can identify this first term as the KL divergence that we used to define $I(x^N)$

$$= I(x^N) + \mathcal{O}(N^0) \tag{C.15}$$

### *C.2.3 Significance level implied by a data partition*

Under the assumption of the smaller model, the information difference is expected to be distributed like a $\frac{1}{2}\chi^2$ with $\Delta K$ degrees of freedom, where $\Delta K = K_2 - K_1$. The effective significance level $\alpha_\nu$ is therefore

$$\alpha_\nu = 1 - \mathrm{CDF}\left[\chi^2_{\Delta K}\right]\left(\Delta K\left[1 + (1+\nu^{-1})\log(1+\nu)\right]\right). \tag{C.16}$$

This function is plotted in Fig. 4.5 for two choices of the dimensional difference. An interesting corollary is that for large $\Delta K$, typical confidence levels may actually be less than equivalent predictive methods such as AIC. In other words, we can reject the null hypothesis before it become predictively optimal to use the larger model.

## *C.3 Other Methods*

There are several methods that deviate more drastically from the standard Bayesian probability calculus. We mention here just a few of the interesting ideas which have been proposed.

### *C.3.1 Other Predictive Estimators*

Once a data division strategy is chosen and we can agree on what we are trying to estimate, there are many information criteria which can be used. For instance, the predictive limit can be estimated using AIC, DIC [132] and WAIC [143]. When the posterior is known to be approximately normal, AIC can perform with minimal variance [129]. Far from normality and the large sample-size limit, WAIC has a uniquely well developed justification in terms of algebraic geometry, but the standard LOOCV seems to have better properties in numerical experiments [139]. Similar alternatives to BIC exist for postdictive performance estimation [95, 143].

### *C.3.2 Data-Validated Posterior and Double use of Data*

Aitkin [2] attempted to address the Lindley paradox by proposing training and validating the data using the entire dataset $X^N$. The resulting posterior Bayes factor comes from the observed posterior information:

$$H_{\text{POBF}}(X^N) = h(X^N|X^N) \tag{C.17}$$

This has a complexity $\mathscr{K}_{\text{Aitkin}} = \frac{K}{2}\log 2 \approx 0.35K$. This is far too weak to realize Occam's razor. This weakness results from two effects: i.) We use here the a generalization sample size of $N_G = N$ instead of the predictive limit where the generalization sample size is zero. ii.) The double use of data means that the posterior is over-fit to the particular dataset. This posterior appears to performs better than even knowledge of the true parameter $\mathscr{K}_0 = \frac{K}{2}$. Overfitting can also occur when data are double used implicitly through posterior training, as in empirical Bayes methods where prior hyperparameters are optimized with respect to the model information.

We do not believe that the double use of data is completely ruled out of a principled statistical analysis [4]. But because double use of data is antithetical to the interpretation of conditional probability, and because it very often leads to overfitting, double use of data requires careful justification.

### *C.3.3 Fractional Bayes Factor*

O'Hagan [114] sought to define a minimal training set mathematically by taking some small power of the likelihood. The fractional model information is then

$$H_{\mathrm{FBF}}(b) = \log \mathop{\mathbb{E}_{\boldsymbol{\theta}}}_{\pi} q^b(X^N|\boldsymbol{\theta}) - \log \mathop{\mathbb{E}_{\boldsymbol{\theta}}}_{\pi} q(X^N|\boldsymbol{\theta}) \tag{C.18}$$

where $b$ is chosen to be very small. If epsilon goes to zero, this expression is obviously identical to the original model information. As O'Hagan notes "The key question remaining in the use of FBFs is the choice of $b$. It may seem that the only achievement of this paper is to replace an arbitrary ratio [*i.e.* $N_G/N_T$] with an arbitrary choice of $b$." The same issues with defining a minimal experiment for minimal training also arise for this approach.

## *C.4 Efficiency and correct models*

The landmark treatment by J. Shao[127] and its discussion by Yang [149] are sometimes viewed as supporting BIC and Bayes factors in certain situations. We therefore wish to discuss this important work in more detail. We suppress many of the technical details for the purposes of our discussion, and refer to [127, 149] for more precision.

Let $\alpha_0^N$ be the identifier for the most predictive model at sample size $N$ (which may not be the true model!), and let $\hat{\alpha}_\nu(x^N)$ identify the model chosen by selecting the largest pseudo-Bayes factor parameterized by $\nu$. We can define the loss ratio in terms of the predictive cross-entropy of the trained model,

$$\epsilon_\nu(x^N) \equiv \frac{\mathbb{E}_Y\, h(Y|x^N, \hat{\alpha}_\nu(x^N))}{\mathbb{E}_Y\, h(Y|x^N, \alpha_0^N)}, \tag{C.19}$$

where the expectations are taken with respect to the true distribution. Shao identifies a reasonable criteria for the performance of a model estimator $\hat{\alpha}$: *asymptotic-loss-efficiency* which is equivalent to the condition that

$$\epsilon_\nu(X^N) \to_p 1, \tag{C.20}$$

That is, the loss ratio converges in probability to unity as the sample size goes to infinity.

Shao found that the *context* in which model selection is performed is incredibly important to whether or not asymptotic efficiency is achieved. Specifically, there are two very different situations:

1. There is no correct model, or there is exactly one correct model which is not nested inside a more complicated model.

2. There is more than one correct model. The smallest correct model is nested inside potentially an infinite set of increasingly complicated models, which are all capable of realizing the smaller model.

If condition (1) holds predictive methods ($\nu \to 0$) are guaranteed to be asymptotically efficient, and (pseudo-)Bayes factors for which $\nu > 0$ are *not* guaranteed to be asymptotically efficient. But if condition (2) holds, then statistical fluctuations will cause AIC and pseudo-Bayesian methods to choose larger models than $\alpha_0$ with a probability that never goes to zero. It is necessary for the penalty that is, $\nu$, to diverge to ensure that the probability of choosing a larger correct model will converge to zero, and that asymptotic efficiency can be achieved.

If the possibility of condition (2), and the true model is realizable at finite dimension, many would suggest that we are justified in using Bayesian methods which have a divergent $\log N$ penalty and thus hope for asymptotic efficiency. We criticize this position on several points.

First, condition (2) is unlikely to ever hold. The Boxian proverb, "All models are wrong," expresses the general truth that nature is too complicated to ever yield the exact truth to a finite dimensional model. Condition (1) is far more likely in any typical modeling context.

Second, whereas predictive methods occupy a unique place in relation to condition (1), the rate at which penalties must go to infinity to satisfy efficiency under condition (2) is not uniquely determined. All methods whose complexities go to infinity slower than $N$, will

(with some technical caveats) satisfy asymptotic efficiency. A complexity of $\log \log N$ would be no less favored under this argument than the $\log N$ complexity of BIC.

Finally, the asymptotic argument which prefers BIC under condition (2) seems to have little bearing on the conditions we would observe at finite sample size. At finite sample size, we do not know if we are in the regime where we are selecting the true model, or if the true model cannot yet be resolved with the available data. If the true model cannot be resolved, we'd still expect AIC to typically outperform BIC for the same reasons that hold in condition (1). BIC is unjustified unless we know a priori the scale at which a true effect will be observed. This is exactly the situation which holds when we have a precise distribution for the parameter of interest, and the Bayesian approach is indistinguishable from the way the frequentist would use *a priori* information in accordance with the Bayes law.

# Appendix D

# THERMODYNAMICS SUPPLEMENT

## *D.1 Supplemental results*

### *D.1.1 Definitions of information, cross entropy, Fisher information matrix*

The Shannon Information is defined:

$$h(x|\boldsymbol{\theta}) \equiv -\log q(x|\boldsymbol{\theta}). \tag{D.1}$$

Let $X$ be distributed with a true distribution with parameter $\boldsymbol{\theta}_0$: $X \sim q(\cdot|\boldsymbol{\theta}_0)$. The cross entropy is defined:

$$H(\boldsymbol{\theta};\boldsymbol{\theta}_0) \equiv \overline{h(X|\boldsymbol{\theta})}, \tag{D.2}$$

and which has a minimum at the true entropy:

$$H_0(\boldsymbol{\theta}_0) \equiv H(\boldsymbol{\theta}_0;\boldsymbol{\theta}_0). \tag{D.3}$$

The empirical estimator of the cross entropy is defined:

$$\hat{H}(\boldsymbol{\theta}) \equiv N^{-1}\sum_{i=1}^{N} h(x_i|\boldsymbol{\theta}), \tag{D.4}$$

which scales like $N^0$ in spite of the prefactor. The KL-Divergence:

$$D_{\mathrm{KL}}(\boldsymbol{\theta}_0||\boldsymbol{\theta}) = H(\boldsymbol{\theta};\boldsymbol{\theta}_0) - H_0(\boldsymbol{\theta}_0), \tag{D.5}$$

is the natural distance-like measure on the parameter manifold. The Fisher information matrix is defined:

$$I_{ij} = \left[\tfrac{\partial}{\partial\theta^i}\tfrac{\partial}{\partial\theta^j}H(\boldsymbol{\theta};\boldsymbol{\theta}_0)\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \tag{D.6}$$

which is a rank two covariant tensor known as the Fisher-Rao metric.

### *D.1.2 An alternate correspondence*

In establishing the correspondence between inference and statistical mechanics, we identify the partition function $Z$ as the marginal likelihood and $N \leftrightarrow \beta$ in agreement with V. Balasubramanian [14]. This is not the only choice. For instance Watanabe [143] instead chooses to define the inverse-temperature $\beta$ so that the likelihood is given by $q^{\beta}(X^N|\boldsymbol{\theta})$, that is raised to an arbitrary power $\beta$. This identification has two advantages: i.) It seems to be more closely related to the physical temperature, which can be varied independently with the strength of the quenched disorder ii.) It allows one to interpolate between a Bayesian posterior (given by $\beta = 1$) and the point estimates of the MLE's (given by $\beta = \infty$). This temperature has also been applied in tempering schemes in MCMC methods, and simulated annealing—increasing the temperature promotes a better exploration of the sample space (chain-mixing) that can be used to better sample multimodal distributions, or find the minima in a rough function.

On the other hand, there are two disadvantages of a power $\beta$ relative to $N \leftrightarrow \beta$ which we believe outweigh the advantages: i.) First, it is not a preexisting statistical parameter within the Bayesian framework. ii.) Second, the internal energy under this other choice of $\beta$ is not the predictive performance $U$. Consequently, the principle of indifference which results from a likelihood-power $\beta$ does not induce the Akaike weights as the model averaging procedure. Instead $\overline{U} = H_0$, which does not encode a realization of Occam's razor.

Thermodynamic expressions using both definitions may give somewhat complementary information. Which is useful will depend on the context. We do not believe that statistical mechanics prescribes a uniquely-correct procedure for objective Bayesian inference. It is the reproduction of a principled model selection criteria, AIC with its proven asymptotic efficiency [127] that justifies the proposed correspondence in the context of model selection.

### *D.1.3 Finite difference is equivalent to cross validation*

The log-predictive distribution can be written as a finite difference

$$\log q(X_i|X^{\neq i}) = \log Z(X^N) - \log Z(X^{\neq i}), \tag{D.7}$$

We can interpret the $\log q(X_i|X^{\neq i})$ as a finite difference estimate of the the sample size derivative of the free energy. We take the mean over all permutations of the data so that this estimate is symmetric with respect to all data points. Under expectation, analytically continuing sample size, the LOOCV relationship to the internal energy is clear:

$$\langle \log q(X_i|X^{\neq i}) \rangle \approx \frac{\partial}{\partial N} \langle \log Z(X^N) \rangle + O(N^{-1}). \tag{D.8}$$

This identity is crucial in establishing the thermodynamic interpretation in terms of predictive performance

### *D.1.4 Jeffreys prior is proportional to GPI prior in the large-sample-size limit*

In the large-sample-size limit, the partition function can be evaluated using the Lapalce (saddle-point) approximation and the resulting prior is proportional to the Jeffreys prior. The integral is evaluated by expanding around the minimum of $\hat{H}_X(\boldsymbol{\theta})$, the maximum likelihood estimator: $\hat{\boldsymbol{\theta}}_X$. The partitition function $Z(X^N) = \int_{\boldsymbol{\Theta}} \mathrm{d}\boldsymbol{\theta}\ \varpi(\boldsymbol{\theta})\ \exp[-N\hat{H}_X(\boldsymbol{\theta})]$, becomes

$$Z(X^N) \approx e^{-N\hat{H}_X(\hat{\boldsymbol{\theta}}_X)} \left( \frac{2\pi}{N(\det I)^{1/K}} \right)^{K/2} \varpi(\boldsymbol{\theta}_X) \tag{D.9}$$

By the standard $\chi^2_K$ representation of the overfitting error, $\langle \hat{H}(\hat{\theta}_X) \rangle_X = H_0 - \frac{K}{2N}$. Therefore the disorder average becomes

$$\overline{F}(\boldsymbol{\theta}_0, \varpi, N) = H_0 - \frac{K}{2N} - \frac{K}{2N} \log \frac{2\pi}{N(\det I)^{1/K}} - \frac{1}{N} \log \varpi(\boldsymbol{\theta}_0) + O(N^{-2}) \tag{D.10}$$

We can then calculate the Gibbs entropy $N^2 \partial_N F$,

$$\overline{S}(\boldsymbol{\theta}_0, \varpi, N) = \frac{K}{2} \log \frac{2\pi}{N(\det I)^{1/K}} + K + \log \varpi(\boldsymbol{\theta}_0) + O(N^{-1}) \tag{D.11}$$

If we enforce the generalized principle of indifference, ignoring higher powers of $N^{-1}$,

$$0 = S(\boldsymbol{\theta}_0, w, N) \tag{D.12}$$

and substituting the $w$ for $\varpi$ in the entropy expression then gives us the conditon

$$w(\boldsymbol{\theta}_0) = (\det I)^{1/2} \left(\frac{N}{2\pi}\right)^{K/2} e^{-K}. \tag{D.13}$$

Thus the generalized principle of indifference is satisfied by the Jeffries prior in the large-sample-size limit. The constant weighting factor is important in model selection as $e^{-K}$ expresses the Akaike weighting.

This constant factor shows another important characteristic of the GPI prior: it has sample-size dependence. This sample size dependence will in general break the de-Finetti likelihood principle: that the prior should not depend on the nature of the data-generating procedure (including the sample size). The departure from the likelihood principle is the origin of the departure from the conventional Bayesian model selection behavior.

### *D.1.5 Reparametrization invariance of thermodynamic functions and the GPI prior*

Reparametrization invariance of the thermodynamic quantities follows from there being derived from the partition function which is also reparametrization invariant. The partition function under an invertible coordinate transformation becomes

$$Z(X^N) = \int \mathrm{d}\boldsymbol{\phi}\, q(X^N|\boldsymbol{\phi}) \varpi\left(\boldsymbol{\theta}(\boldsymbol{\phi})\right) J(\boldsymbol{\phi}). \tag{D.14}$$

At the same time the density transforms so that $\varpi\left(\boldsymbol{\theta}(\boldsymbol{\phi})\right) J(\boldsymbol{\phi}) \to \varpi'\left(\boldsymbol{\phi}\right)$ where $J$ is the determinant of the Jacobian. Notably, if $\varpi\left(\boldsymbol{\theta}\right)$ satisfies the GPI, then the transformed density $\varpi\left(\boldsymbol{\theta}(\boldsymbol{\phi})\right) J(\boldsymbol{\phi}) \to \varpi'\left(\boldsymbol{\phi}\right)$ results in the same partition function and Gibbs entropy and therefore still satisfies the GPI in this new coordinate system.

### *D.1.6 Effective temperature of confinement*

To calculate the free energy $\mathscr{F}$ of a free particle confined to a volume $V = L^3$, we calculate the partition function by integrating over available phase space:

$$Z(\beta) = \int \frac{d^K \boldsymbol{p}\, d^K \boldsymbol{x}}{(2\pi\hbar)^K} e^{-\beta H(\boldsymbol{p},\boldsymbol{x})} \tag{D.15}$$

$$= \frac{e^{-\beta E_0} L^K}{(2\pi\hbar)^K} \left( \int dp\, e^{\frac{-\beta p^2}{2m}} \right)^K = \left( \frac{mL^2}{2\pi\hbar^2\beta} \right)^{K/2} e^{-\beta E_0}. \tag{D.16}$$

The Free energy is then

$$\mathscr{F}(\beta) = E_0 + \frac{K}{2\beta} \log \frac{mL^2}{2\pi\hbar^2\beta} = E_0 + \frac{K}{2\beta} \log \frac{\beta_0}{\beta} \tag{D.17}$$

where we have made the identifications

$$\beta_0 = \frac{mL^2}{2\pi\hbar^2} \quad \text{and} \quad K = 3. \tag{D.18}$$

$\beta_0$ can be interpreted as the inverse of the (typically negligibly small) temperature at which the thermal de Broglie wavelength of the confined particle is on the order of the width of the confining box.

### *D.1.7 A Bayesian re-interpretation*

The replacement of the prior (a probability density) with an unnormalized density of states may make a Bayesian reader uncomfortable since the evidence ($Z$) no longer has the meaning of a probability. But there is a natural Bayesian interpretation in terms of the *a priori* model probability. Typically, when models are compared in a Bayesian context, all mutually exclusive models are assigned equal *a priori* probabilities (*i.e.* the principle of indifference). But, we have now proposed a new concept of model enumeration by introducing a density of models. We can compute the total number of distinguishable distributions in model $I$ at sample size $N$ by integrating the GPI prior (density of states) over the parameter manifold:

$$\mathcal{N}_I(N) \equiv \int_\Theta \mathrm{d}\boldsymbol{\theta}\ w_I(\boldsymbol{\theta}; N). \tag{D.19}$$

Since models $I$ and $J$ contain different numbers of distinguishable distributions, we reason that the principle of indifference should be interpreted to apply at the distinguishable distribution level rather than the model level. Therefore the *a prior* model probabilities should be:

$$\varpi_I \equiv \mathcal{N}_I / \sum_I \mathcal{N}_I. \tag{D.20}$$

and the proper parameter prior is

$$\varpi(\boldsymbol{\theta}|I) \equiv w_I(\boldsymbol{\theta}; N)/\mathcal{N}_I. \tag{D.21}$$

Inference with the improper GPI prior is equivalent to assuming proper prior $\varpi_I$ on models and proper prior $\varpi(\boldsymbol{\theta}|I)$ on parameters. The numerator in RHS of Eqn. D.20 will cancel the denominator in the RHS of Eqn. D.21 when the model posterior is computed and the normalization $\mathcal{N}_I$ divides out of parameter posterior distributions.

## *D.2 Methods*

### *D.2.1 Computation of learning capacity*

To compute the learning capacity, we will use the definition from Tab. 5.1:

$$\overline{C}(\boldsymbol{\theta}; N, \varpi) = N^2 \partial_N^2 \overline{\log Z}(\boldsymbol{\theta}; N, \varpi), \tag{D.22}$$

where $X \sim q(\cdot|\boldsymbol{\theta})$.

### *D.2.2 Direct computation of GPI prior*

We will use the discrete difference definition of the entropy (Eqns. 5.7 and 5.10) to enforce the generalized principle of indifference (Eqn. 5.12). The relation for the GPI prior can be written:

$$(N+1)\,\overline{\log Z}(\boldsymbol{\theta}; N, w) = N\,\overline{\log Z}(\boldsymbol{\theta}; N+1, w), \tag{D.23}$$

in terms of the partition function. We will use Eqn. D.23 explicitly to solve for the GPI prior $w$. For the models we work analytically, we will be be able to use the asymptotic form of $w$

(Eqn. 5.28) to define an effective model dimension $\mathscr{K}$ (Eqn. 5.29). The general strategy will be:

1. Use symmetry and dimensional analysis to deduce the scaling of $w$ with respect to the parameters $\boldsymbol{\theta}$.

2. Compute $\log Z(X^N; w)$ and re-express in terms of canonical random variables.

3. Compute $\overline{\log Z(X^N; w)}$.

4. Solve for the unknown normalization $c$ of $w$ using GPI (Eqn. D.23).

### *D.2.3 Computation of the free energy using a sufficient statistic*

It is often convenient to work in terms of sufficient statistics because (i) all the data dependence of the posterior enters through the sufficient statistic and (ii) the statistics have well known statistical distributions that significantly simplify many calculations. We define a sufficient statistic $\boldsymbol{t} = \boldsymbol{T}(X^N)$ such that

$$\Pr(\boldsymbol{\theta}|X^N) = \Pr(\boldsymbol{\theta}|\boldsymbol{t}), \tag{D.24}$$

or all the information about the parameters is encoded in $\boldsymbol{t}$. We can therefore write:

$$q(X^N|\boldsymbol{\theta}) = q(X^N|\boldsymbol{t})\ q(\boldsymbol{t}|\boldsymbol{\theta}), \tag{D.25}$$

and we can define a Statistic Shanon entropy:

$$H_{\boldsymbol{t}}(\boldsymbol{\theta}) = -\overline{\log q(\boldsymbol{t}|\boldsymbol{\theta})}. \tag{D.26}$$

In terms of the sufficient statistic, the partition function factors:

$$Z(X^N; \varpi) \;=\; q(X^N|\boldsymbol{t})\ z(\boldsymbol{t}; N, \varpi), \tag{D.27}$$

where the statistic partition function is

$$z(\boldsymbol{t}; N, \varpi) \;\equiv\; \int_\Theta \mathrm{d}\boldsymbol{\theta}\ \varpi(\boldsymbol{\theta})\ q(\boldsymbol{t}|\boldsymbol{\theta}). \tag{D.28}$$

The expected free energy can be written:

$$\overline{F}(\boldsymbol{\theta}; N, \varpi) = -N^{-1}\overline{\log z(\boldsymbol{t}; N, \varpi)} + H_0(\boldsymbol{\theta}) - N^{-1}H_{\boldsymbol{t}}(\boldsymbol{\theta}), \tag{D.29}$$

where $H_0$ is the entropy.

### *D.2.4 Computation of the GPI prior using a recursive approximation*

The Gibbs entropy has the property that it is linear in the prior so that the following holds:

$$\overline{S}(\theta_0, N, e^{\alpha}\varpi) = \alpha + \overline{S}(\theta_0, N, \varpi) \tag{D.30}$$

If the prior and entropy are flat, then setting $\alpha = -\overline{S}(\theta_0, N, \varpi)$ will result in $\overline{S}(\theta_0, N, e^{\alpha}\varpi) = 0$; the w-prior condition. This suggests the following simple recursive scheme for a successive approximation for the w-prior:

1: **procedure** RecursiveW($\varpi$)
2: **repeat**
3: $\varpi(\theta) \leftarrow \varpi(\theta)e^{-\overline{S}(\theta,\varpi,N)}$
4: **until** $\overline{S}(\theta; \varpi, N) \approx 0$
5: **end procedure**

To the extent the entropy is slowly varying and only locally dependent on the prior, this algorithm will very quickly converge to an exact w-prior. However, effects due to manifold boundaries and model singularities may create artifacts that lead to unstable updates. Empirical evidence suggests that the algorithm should be terminated before the exact GPI prior condition is met. Typically very few iterations are required. In the Poisson stoichiometry problem, $w$ for $t = 100$ and $t = 500$ were calculated with only a single iteration. At smaller sample-sizes, more iterations are required.

## D.3 Details of applications

### D.3.1 Normal model with unknown mean and informative prior

The likelihood for the normal model is defined by Eqn. 5.14 with parameters $\boldsymbol{\theta} \equiv (\vec{\mu})$ for support $\mu \in \mathbb{R}^K$ for a normal model with unknown mean and known variance $\sigma^2$. In this example, we assume a conjugate prior:

$$\varpi(\boldsymbol{\theta}) = (2\pi\sigma_\varpi^2)^{-K/2} \exp[-\tfrac{1}{2\sigma_\varpi^2}(\vec{\mu} - \vec{\mu}_\varpi)^2], \tag{D.31}$$

where we introduce the critical sample size $N_0 \equiv \sigma^2/\sigma_\varpi^2$. The partition function is computed by completing the square in the exponential. If $X^N \sim q(\cdot|\boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim \varpi$, the log partition function can be expressed in terms of three independent chi-squared random variables:

$$\sigma^{-2} \sum_{i=1}^{N} (\vec{X}_i - \hat{\vec{\mu}}_X)^2 \sim \chi^2_{K(N-1)}, \tag{D.32}$$

$$\sigma^{-2} N (\vec{\mu} - \hat{\vec{\mu}}_X)^2 \sim \chi^2_K, \tag{D.33}$$

$$\sigma^{-2} N_0 (\vec{\mu} - \vec{\mu}_\varpi)^2 \sim \chi^2_K. \tag{D.34}$$

The log partition function is therefore distributed:

$$\log Z(X^N; \varpi) \sim -\tfrac{KN}{2} \log 2\pi\sigma^2 - \tfrac{K}{2} \log \tfrac{N+N_0}{N_0} - \tfrac{1}{2}\chi^2_{K(N-1)} - \tfrac{1}{2}\tfrac{N_0 N}{N+N_0}(N^{-1}\chi^2_K + N_0^{-1}\chi^{2\prime}_K), \tag{D.35}$$

where $\chi^2_j$ is a chi-squared random variable dimension $j$ and the expect-log partition function is

$$\overline{\log Z}(N, \varpi) = -NH_0 - \tfrac{K}{2} \log \tfrac{N+N_0}{N_0}, \tag{D.36}$$

where $H_0$ is the entropy and the free energy is:

$$\overline{F}(N, \varpi) = H_0 + \tfrac{K}{2N} \log \tfrac{N+N_0}{N_0}. \tag{D.37}$$

The other results in the Tab. 5.2 are generated by apply the definitions of the correspondence in Tab. 5.1.

### *D.3.2 Normal model with unknown mean*

The likelihood for the normal model is defined by Eqn. 5.14 with parameters $\boldsymbol{\theta} \equiv (\vec{\mu})$ for support $\vec{\mu} \in \mathbb{R}^D$ for a normal model with unknown mean and known variance $\sigma^2$. The cross entropy is

$$H(\boldsymbol{\theta}; \boldsymbol{\theta}_0) \equiv \tfrac{D}{2} \left[\log 2\pi\sigma^2 + 1\right] + \tfrac{1}{2\sigma^2}(\vec{\mu} - \vec{\mu}_0)^2, \tag{D.38}$$

where the true distribution is $X \sim q(\vec{x}|\boldsymbol{\theta}_0)$ and the determinant of the Fisher information matrix is:

$$\det \boldsymbol{I} = \sigma^{-2D}. \tag{D.39}$$

The scaled Jeffreys prior (Eqn. 5.27) is therefore:

$$\rho = (\tfrac{N}{2\pi\sigma^2})^{D/2}. \tag{D.40}$$

We will assume $w$ matches the asymptotic form:

$$w = c\sigma^{-D}, \tag{D.41}$$

and solve for the unknown constant $c(N, D)$. The partition function is

$$\log Z(X^N; w) \sim \log c - \tfrac{DN}{2} \log 2\pi\sigma^2 + \tfrac{D}{2} \log \tfrac{2\pi}{N} - \tfrac{1}{2}\chi^2_{D(N-1)} \tag{D.42}$$

where $\chi^2_{D(N-1)}$ is a $D(N-1)$-dimensional chi-squared random variable. The expected log partition function is:

$$\overline{\log Z}(\boldsymbol{\theta}; N, w) = -NH_0(\boldsymbol{\theta}) + \log c + \tfrac{D}{2}\left[\log \tfrac{2\pi}{N} + 1\right], \tag{D.43}$$

where $H_0$ is the true entropy. The learning capacity is:

$$\overline{C}(\boldsymbol{\theta}; N) = \tfrac{D}{2}, \tag{D.44}$$

where $D$ is both the dimension of mean parameter and the model. The unknown normalization of $w$ is:

$$\log c = \tfrac{D}{2} \log \tfrac{N}{2\pi} - \tfrac{D}{2}\left[1 + N \log(1 + N^{-1})\right] \tag{D.45}$$

which can be re-written as an effective dimension:

$$\mathscr{K} = \tfrac{D}{2}\left[1 + N\log(1+N^{-1})\right], \tag{D.46}$$

to define the GPI prior $w$ using Eqn. 5.29.

### *D.3.3 Normal model with unknown discrete mean*

The likelihood for the normal model is defined by Eqn. 5.14 with parameters $\boldsymbol{\theta} \equiv (\vec{\mu})$ for support $\vec{\mu} \in \mathbb{Z}^D$ for a normal model with unknown mean and known variance $\sigma^2$. We use Eqn. D.29 to treat the problem in terms of sufficient statitics. The statistic partition function breaks up by dimension: For each dimension with flat prior $\varpi(\mu) = \sum_m \delta(\mu - m)$, and sufficient statistic $t = N^{-1}\sum_i^N X_i$ the statistic partition function becomes the sum over discrete prior values:

$$z(t; N, \varpi) = \sum_{m=-\infty}^{\infty} q(t|m) = \left(\frac{N}{2\pi}\right)^{1/2} \sum_{m=-\infty}^{\infty} e^{-N(t-m)^2} \tag{D.47}$$

$$= \vartheta\left(t; r = e^{\frac{-2\pi^2}{N}}\right) = \sum_{m=-\infty}^{\infty} r^{m^2} e^{2\pi t} \tag{D.48}$$

Where $\vartheta$ is the Jacobi theta function with nome $r$. We can use the Jacobi triple product formula to write down a log partition function

$$\log z(t; N, \varpi) = \sum_{m=1}^{\infty} \log(1 - r^{2m}) + \log\left(1 + r^{2m-1}e^{-i2\pi t}\right) + \log\left(1 + r^{2m-1}e^{i2\pi t}\right) \tag{D.49}$$

We are now able to take the disorder average analytically. Assume (without loss of generality) that $m_0 = 0$, then, because $|r| < 1$, we can safely expand the logarithm, and

$$\mathop{\mathbb{E}}_{t|m_0} \sum_{m=1}^{\infty} \log\left(1 + r^{2m-1}e^{-i2\pi t}\right) = -\sum_{m=1}^{\infty}\sum_{k=1}^{\infty} \frac{(-1)^k}{k} r^{(2m-1)k} \mathop{\mathbb{E}}_{t|m_0} e^{-i2\pi k t} \tag{D.50}$$

The expectation is just $\phi_N(2\pi k)$, the characteristic function of the central normal distribution with variance $\sigma^2 = 1/N$. Specifically $\phi_N(2\pi k) = e^{-\frac{2\pi^2 k^2}{n}} = q^{k^2}$

$$-\sum_{m=1}^{\infty}\sum_{k=1}^{\infty} \frac{(-1)^k}{k} r^{(2m-1)k+k^2} = \sum_{k=1}^{\infty} \frac{(-1)^k}{k} \frac{r^{k^2+k}}{r^{2k}-1} \tag{D.51}$$

This series is convergent, but converges slowly for very large $N$. We therefore must also develop a series for when $N \gg 1$. We can use the Poisson resummation formula to convert the partition function into a sum over reciprocal space. First we have to subtract off the singularity at zero, by adding a piece to the summand that can be explicitly summed. Then we can extend this function to both positive and negative integers:

$$\mathop{\mathbb{E}}_{t|m_0} \sum_{m=1}^{\infty} \log\left(1 + r^{2m-1} e^{-i2\pi t}\right) = \frac{1}{2} \sum_{k=1}^{\infty} \frac{\cos(\pi k)}{k^2} \frac{k}{\sinh\left(\frac{2\pi^2 k}{N}\right)} e^{-\frac{2\pi^2 k^2}{N}} \tag{D.52}$$

$$= \frac{n}{48} + \frac{n}{4\pi^2} \sum_{k=1}^{\infty} \frac{\cos(\pi k)}{k^2} \left( \frac{2\pi^2 k}{n \sinh\left(\frac{2\pi^2 k}{N}\right)} e^{-\frac{2\pi^2 k^2}{N}} - 1 \right) \tag{D.53}$$

$$= \frac{n}{48} + \frac{1}{4} - \frac{\pi^2}{12n} + \frac{n}{8\pi^2} \sum_{k=-\infty}^{\infty} \frac{\cos(\pi k)}{k^2} \left( \frac{2\pi^2 k}{n \sinh\left(\frac{2\pi^2 k}{N}\right)} e^{-\frac{2\pi^2 k^2}{N}} - 1 \right) \tag{D.54}$$

The sum can now be represented as the sum of the Fourier transform of the summand. At large n, even the first term is exponentially small, and the whole sum can be ignored, leaving:

$$\mathop{\mathbb{E}}_{t|m_0} \sum_{m=1}^{\infty} \log\left(1 + r^{2m-1} e^{-i2\pi t}\right) = \begin{cases} \sum_{k=1}^{\infty} \frac{(-1)^k}{k} \frac{r^{k^2+k}}{r^{2k}-1} & \text{for all n} \\ \frac{n}{48} + \frac{1}{4} - \frac{\pi^2}{12n} & n > 10^2 \end{cases} \tag{D.55}$$

Similarly we have for the Euler function piece of the Jacobi-theta triple product

$$\sum_{m=1}^{\infty} \log(1 - r^{2m}) = \begin{cases} \sum_{k=1}^{\infty} \frac{1}{k} \frac{r^{2k}}{1-r^{2k}} & \text{for all n} \\ -\frac{n}{24} + \frac{\pi^2}{6n} + \frac{1}{2} \log\left(\frac{n}{2\pi}\right) & n > 50 \end{cases} \tag{D.56}$$

The rational tail from these two contributions in the asymptotic expansion cancel exactly in the free-energy so only the logarithmic term in the Euler function remains. This term exactly cancels the 1/2 from the $H_t$ contribution to the full free-energy and shows that the learning capacity reaches zero at large sample size.

### *D.3.4 Normal model unknown mean and variance*

The likelihood for the normal model is defined by Eqn. 5.14 with parameters $\boldsymbol{\theta} = (\vec{\mu}, \sigma)$ with support $\vec{\mu} \in \mathbb{R}^D$ and $\sigma \in \mathbb{R}_{>0}$. The cross entropy is:

$$H(\boldsymbol{\theta}; \boldsymbol{\theta}_0) \equiv \tfrac{D}{2}\left[\log 2\pi\sigma^2 + \tfrac{\sigma_0^2}{\sigma^2}\right] + \tfrac{1}{2\sigma^2}(\vec{\mu} - \vec{\mu}_0)^2, \tag{D.57}$$

where the true distribution is $X \sim q(\vec{x}|\boldsymbol{\theta}_0)$ and the determinant of the Fisher information matrix is:

$$\det \boldsymbol{I} = 2\sigma^{-2(D+1)}. \tag{D.58}$$

The scaled Jeffreys prior (Eqn. 5.27) is therefore:

$$\rho = \sqrt{2}(\tfrac{N}{2\pi\sigma^2})^{(D+1)/2}. \tag{D.59}$$

We will assume $w$ matches the asymptotic form:

$$w = c\,\sigma^{-D-1}. \tag{D.60}$$

Note that $w$ must have units of inverse length to the $D+1$ power in order to give the evidence the correct units. Due to translation symmetry in $\mu$, $w$ must be a function of $\sigma$ only. The partition function is

$$\log Z(X^N) \sim \log c - \tfrac{DN}{2}\log 2\pi\sigma^2 + \tfrac{D}{2}\log\tfrac{2\pi}{N} - \log 2 + \log\Gamma(\tfrac{DN}{2}) - \tfrac{DN}{2}\log\tfrac{\chi^2_{D(N-1)}}{2} \tag{D.61}$$

where $\chi^2_{D(N-1)}$ is a $D(N-1)$-dimensional chi-squared random variable. The expected log partition function is:

$$\overline{\log Z}(\boldsymbol{\theta}; N) = -NH_0(\boldsymbol{\theta}) + \log c + \tfrac{DN}{2} + \tfrac{D}{2}\log\tfrac{2\pi}{N} - \log 2 + \log\Gamma(\tfrac{DN}{2}) - \tfrac{DN}{2}\psi(\tfrac{D(N-1)}{2}) \tag{D.62}$$

where $H_0$ is the true entropy and $\psi$ is the polygamma function. The learning capacity is:

$$\overline{C}(\boldsymbol{\theta}; N, w) = \tfrac{D}{2} + N^2(\tfrac{D}{2})^2\psi^{(1)}(\tfrac{DN}{2}) - 2N^2(\tfrac{D}{2})^2\psi^{(1)}(\tfrac{D(N-1)}{2}) - N^3(\tfrac{D}{2})^3\psi^{(2)}(\tfrac{D(N-1)}{2}), \tag{D.63}$$

where $D$ is both the dimension of mean parameter and the model. The unknown normalization of $w$ is:

$$\log c = \tfrac{D+1}{2}\log\tfrac{N}{2\pi} + \tfrac{1}{2}\log 2 - \mathscr{K} \tag{D.64}$$

written in terms of the effective dimension:

$$\mathscr{K} = \tfrac{1}{2}\log\tfrac{N}{2\pi} - \tfrac{1}{2}\log 2 - \tfrac{DN}{2}\log\tfrac{N}{N+1} - N\log\Gamma[\tfrac{D(N+1)}{2}] + (N+1)\log\Gamma[\tfrac{DN}{2}] + \tfrac{D(N+1)N}{2}\left[\psi(\tfrac{DN}{2} - \psi(\tfrac{D(N-1)}{2}))\right], \tag{D.65}$$

which is used to define the GPI prior $w$ using Eqn. 5.29.

### *D.3.5 Exponential model*

The likelihood for the normal model is defined:

$$q(x|\boldsymbol{\theta}) \equiv \lambda e^{-\lambda x} \tag{D.66}$$

which parameters $\boldsymbol{\theta} \equiv (\lambda)$ with support $\lambda \in \mathbb{R}_{>0}$. The cross entropy is:

$$H(\boldsymbol{\theta}; \boldsymbol{\theta}_0) = -\log\lambda + \tfrac{\lambda}{\lambda_0}, \tag{D.67}$$

where the true distribution is $X \sim q(x|\boldsymbol{\theta}_0)$ and the determinant of the Fisher information matrix is:

$$\det \boldsymbol{I} = \lambda^{-2} \tag{D.68}$$

The scaled Jeffreys prior (Eqn. 5.27) is therefore:

$$\rho = (\tfrac{N}{2\pi\lambda^2})^{1/2} \tag{D.69}$$

We will assume $w$ matches the asymptotic form:

$$w = c\,\lambda^{-1}. \tag{D.70}$$

The partition function is

$$\log Z(X^N) \sim -N\log Y + \log\Gamma(N) + N\log\lambda, \tag{D.71}$$

where $Y$ is a Gamma-distributed random variable with unit scale and shape $N$. The expected log partition function is:

$$\overline{\log Z}(\boldsymbol{\theta}; N) = -NH_0(\boldsymbol{\theta}) + N(1 - \psi[N]) + \log\Gamma(N) \tag{D.72}$$

where $H_0$ is the true entropy and $\psi$ is the polygamma function. The learning capacity is:

$$\overline{C}(\boldsymbol{\theta}; N) = N^2[\psi^{(1)}(N) - 2\psi^{(1)}(N) - N\psi^{(2)}(N)], \tag{D.73}$$

where $\psi$ is the polygamma function. The unknown normalization of $w$ is:

$$\log c = \tfrac{1}{2}\log\tfrac{N}{2\pi} - \mathscr{K} \tag{D.74}$$

which can be re-written as an effective dimension:

$$\mathscr{K} = \tfrac{1}{2}\log\tfrac{N}{2\pi} - N\log\Gamma(N+1) + (N+1)\log\Gamma(N) + N(N+1)[\psi(N+1) - \psi(N)], \tag{D.75}$$

to define the GPI prior $w$ using Eqn. 5.29.

### *D.3.6 Uniform distribution*

The likelihood for the normal model is defined:

$$q(x|\boldsymbol{\theta}) \equiv \begin{cases} L^{-1}, & 0 \leq x \leq L \\ 0, & \text{otherwise} \end{cases}, \tag{D.76}$$

where the parameter $\boldsymbol{\theta} \equiv (L)$ with support $L \in \mathbb{R}_{>0}$. The cross entropy is:

$$H(\boldsymbol{\theta}; \boldsymbol{\theta}_0) = \begin{cases} \log L, & L_0 \leq L \\ \infty, & \text{otherwise} \end{cases}, \tag{D.77}$$

which is minimized at $L = L_0$ but neither the first nor second derivative is defined at this point and therefore the Fisher information matrix cannot be defined. We can still infer the dependence of the $w$ by symmetry and dimensional analysis:

$$w = c\, L^{-1}. \tag{D.78}$$

The partition function is

$$\log Z(X^N) \sim \log c - N\log L - \log N - N\log Y, \tag{D.79}$$

where $Y$ is the maximum of $N$ uniformly-distributed random variables on the interval $[0, 1]$. The CDF for $Y$ is the $N$th power of the CDF for a single uniformly-distributed random variable. The expected log partition function is:

$$\overline{\log Z}(\boldsymbol{\theta}; N) = -NH_0(\boldsymbol{\theta}) + \log c - \log N + 1. \tag{D.80}$$

The learning capacity is:

$$\overline{C}(\boldsymbol{\theta}; N) = 1. \tag{D.81}$$

The unknown normalization of $w$ is:

$$\log c = \log N - N \log(1 + N^{-1}) - 1, \tag{D.82}$$

which can be plugged into Eqn. D.78 to calculation the GPI prior $w$.

### *D.3.7 Poisson stoichiometry problem*

Choosing temporal units so that $\lambda = 1$, we subdivide our observation time into short times $t_0 \ll 1$, for which we can approximate our Poisson distribution as the joint distribution of Bernoulli trials.

$$q(X^N|m) = (mt_0)^k \, (1 - mt_0)^{N-k} \tag{D.83}$$

The domain of $m$ is discrete and takes on values $\{1, 2, \ldots\}$ We use this Bernoulli representation of the Poisson process where afterwards we must take the limit where $t_0 \to 0$. It is only in this "extensive" representation $X^N$, that the derivative with respect to sample size is a predictive density. However, we can to perform thermodynamic calculations without this limiting procedure using the sufficient statistic likelihood $p(k|\theta)$, where $k = \sum_i x_i$ following the reasoning laid out in D.2.3. The sufficient statistic partition function as a function of $k$ is

$$z(k; t) = \sum_{m=1}^{\infty} e^{-mt} \frac{(mt)^k}{k!} \varpi(m). \tag{D.84}$$

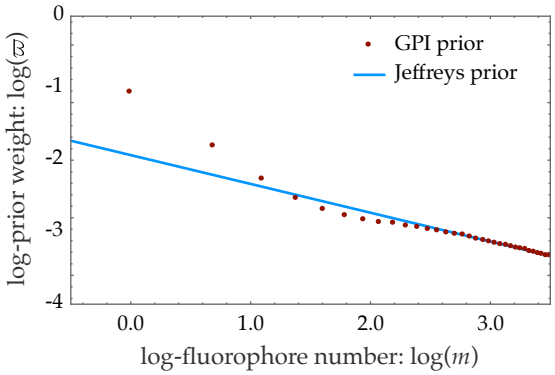


Figure D.1: **Jeffreys prior and GPI prior at** $t = 1$ The power law behavior of the GPI prior and the Jeffreys prior ($m^{-1/2}$) for the Poisson problem at $t = 1$ are compared on a log-log plot. At large fluorophore number, the discrete problem is very similar to the continuous problem, and the GPI prior converges to the same power law behavior as the Jeffreys prior. At small sample size, effects from the discretization deform the GPI prior away from the Jeffreys prior. The normalization of the Jeffreys prior has been chosen to make the two priors match at large $m$.

The likelihood with the one parameter prior $\varpi(m) = e^{-bm}$. This can be recast using Poisson resummation into an equivalent sum

$$z(k;t) = \frac{t^k}{(b+t)^{k+1}} \left( 1 + 2 \sum_{\nu=1}^{\infty} \left( \frac{4\pi^2\nu^2}{(b+t)^2} + 1 \right)^{-\frac{k+1}{2}} \cos\left[ (k+1) \arctan\left( \frac{2\pi\nu}{b+t} \right) \right] \right) \quad \text{(D.85)}$$

The first expression is useful for large k, when the width of the likelihood is much smaller than the unit spacing. The second one can be viewed as an expansion from the approximation when the sum is replaced by an integral, it is therefore useful when the posterior is of large or moderate width ($k/t^2 > .1$).

Finally, we also have the closed form recursion relation which is useful for small $k$

$$z(k;t) = \frac{-t\partial_t}{k} z(k-1;t) \quad \text{with} \quad z(0;t) = (e^{b+t} - 1)^{-1} \quad \text{(D.86)}$$

when $\varpi(m) = e^{-bm}$. This is useful when $k$ is small. When k is large, this closed expression grows into sums with combinatorial coefficients. It is then from the disorder averaged free energy that the partition functions can be calculated after the sample size derivatives are corrected for the changing point entropy which we describe in App. D.2.3

$$\overline{F}(m_0) - H_0 = -t \mathop{\mathbb{E}}_{k|m_0} \log z_t(k) + s_t(m_0) \quad \text{(D.87)}$$

where $s_t$ is the entropy of the count distributions.

Using the expressions for the entropy, and the recursive scheme described in App. D.2.4, we can construct the w-prior. The Jeffreys prior obeys power law scaling for the continuous case [77, 65, 124]. In the small-sample size limit we recover the standard Jeffreys prior at moderate sized fluorophore number as shown inf Fig. D.3.7, while in the large sample-size limit we converge to the Laplace principle of indifference: each discrete value of the parameter assigned equal, in fact unit, weight.

| Model | Parameter support $\boldsymbol{\theta}$ | Generative parameters $\boldsymbol{\theta}_0$ |
|---|---|---|
| $\mathcal{N}$ | | $\mu_0 = 5$, $\sigma_0 = 1$ |
| $\mathcal{N}(\mu)$ | $\mu \in [0, 10]$ | $\mu_0 = 6$, $\sigma_0 = 1$ |
| $\mathcal{N}(\mu, \sigma)$ | $\mu \in [0, 10]$, $\sigma \in [0.1, 10]$ | $\mu_0 = 5$, $\sigma_0 = 0.75$ |
| $\text{Exp}(\lambda)$ | $\lambda \in [0.1, 10]$ | $\lambda_0 = 2$ |
| $\mathcal{U}(L)$ | $L \in [0, 10]$ | $L_0 = 10$ |

Table D.1: **Models for inference on simulated data with revised support.** Five data sets were generated, one for each model, using the generative parameters: $X^N \sim q(\cdot|\boldsymbol{\theta}_0)$. Inference was performed on the simulated data using the GPI prior $w$.