QMUL-PH-22-13

On the Dynamics of Inference and Learning

David S. Berman[1*], Jonathan J. Heckman[2,3†] and Marc Klinger[4‡]

[1]Centre for Theoretical Physics, Queen Mary University of London, London E1 4NS, UK

[2]Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA

[3]Department of Mathematics, University of Pennsylvania, Philadelphia, PA 19104, USA

[4]Department of Physics, University of Illinois, Urbana IL 61801, USA

Abstract

Statistical Inference is the process of determining a probability distribution over the space of parameters of a model given a data set. As more data becomes available this probability distribution becomes updated via the application of Bayes' theorem. We present a treatment of this Bayesian updating process as a continuous dynamical system. Statistical inference is then governed by a first order differential equation describing a trajectory or flow in the information geometry determined by a parametric family of models. We solve this equation for some simple models and show that when the Cramér-Rao bound is saturated the learning rate is governed by a simple $1/T$ power-law, with T a time-like variable denoting the quantity of data. The presence of hidden variables can be incorporated in this setting, leading to an additional driving term in the resulting flow equation. We illustrate this with both analytic and numerical examples based on Gaussians and Gaussian Random Processes and inference of the coupling constant in the 1D Ising model. Finally we compare the qualitative behaviour exhibited by Bayesian flows to the training of various neural networks on benchmarked data sets such as MNIST and CIFAR10 and show how that for networks exhibiting small final losses the simple power-law is also satisfied.

April 2022

*e-mail: d.s.berman@qmul.ac.uk

†e-mail: jheckman@sas.upenn.edu

‡e-mail: marck3@illinois.edu

Contents

1 Introduction

It is difficult to overstate the importance of statistical inference. It forms the bedrock of how one weighs new scientific evidence, and is in some sense the basis for all of rational thought. Leaving philosophy aside, one can ask about the mechanics of inference: given new data, how quickly can we expect to adjust our understanding, and in what sense does this converge to the truth?

Bayes' rule [1] provides a concrete way to approach this question. Given events A and B, the conditional probabilities are related as:[1]

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{1.1}$$

As new evidence arrives, the posterior $P(A|B)$ can be treated as a new prior, and Bayes' rule thus provides a concrete way to continue updating (and hopefully improving) one's initial inference scheme.

In physical applications, one typically imposes a great deal of additional structure which allows one to weigh the various merits of new evidence. For example, in the context of quantum field theory, one is often interested in particle excitations where the structure of locality is built into the inference scheme. In this context, "new evidence" amounts to probing shorter distance scales with the help of a higher energy collider experiment or a more precise measurement of a coupling constant. In modern terms, this is organized with the help of the renormalization group [2–6], which provides a general way to organize new data as relevant, marginal or irrelevant (in terms of its impact on long distance observables).

More broadly, the issue of identifying relevant features as a function of scale is an important issue in a range of inference problems. For example, in machine learning applications, one might wish to classify an image according to "dog vs. cat", and then proceed to breed, and even finer distinguishing features. In this setting, however, the notion of a single quantity such as energy / wavenumber to define "locality" (as used in quantum field theory) is far less clear cut. This is also an issue in a wide range of systems with multiple scales and chaotic dynamics. In these settings it would seem important to seek out physically anchored organizational principles.

Our aim in this note will be to show that in many circumstances, there is an emergent notion of scaling which can be traced all the way back to incremental Bayesian updates. The essential idea is that as an inference scheme converges towards a best guess, the Bayesian update equation comes to resemble a diffusion equation. Much as in [7], the appropriate notion of energy in this context is the Kullback-Leibler divergence [8] between the model

[1] More symmetrically, one can write $P(A|B)P(B) = P(B|A)P(A)$.

distribution $m(x)$ and the true distribution $t(x)$:

$$D_{KL}(t||m) = \int dx\, t(x) \log \frac{t(x)}{m(x)}. \tag{1.2}$$

In many applications, the model depends on a set of fitting parameters θ^i, and the problem of inference amounts to performing an update $m(x|\theta_{k+1}) \leftarrow m(x|\theta_k)$. Indeed, in the infinitesimal limit where the θ_k's converge to an optimal θ^i_*, we can expand in the vicinity of this point. The second order expansion in θ^i of the KL divergence then yields the Fisher information metric which forms the basis for information geometry [9, 10]. (See [11–22] for contemporary appearances of the information metric in statistical inference and its relevance to quantum field theory and string theory.)

The processes of inference through Bayes' theorem then amounts to specifying the trajectory of a particle in the curved background described by the information geometry. In Bayesian inference, we treat the parameters θ as random variables as well, and these are dictated by a posterior distribution $\pi(\theta, T)$ which updates as a function of data steps T. For any observable $O(\theta)$ which depends on these parameters, its appearance in various averages results in an implicit time dependence and a corresponding flow equation:

$$\frac{\partial}{\partial T}\overline{O}(T) = -\mathrm{Var}(O, D), \tag{1.3}$$

where the line on top indicates an average with respect to $\pi(\theta, T)$, and $\mathrm{Var}(O, D)$ denotes the variance between the observable and D, a KL divergence between the "true" distribution and one which depends on θ at some intermediate stage of the inference scheme. There is a striking formal resemblance between the evolution of the parameters of the model (by taking $O = \theta$), and the evolution of parameters in renormalization group flow. We will explore this analogy further, especially with regards to "perturbations" — i.e. new data — which can alter the trajectory of a flow.

In favorable circumstances, we can obtain good approximate solutions to this flow equation for a wide class of observables. In fact, for observables where the Cramér-Rao bound [9] is saturated, we can solve the equation exactly. This then yields a simple $1/T$ power-law scaling. We examine the perturbed flow equation where the Cramér-Rao bound is not saturated and solve the equation numerically to give a power-law scaling with powers greater than -1. Finally, we examine the case where there are "hidden variables". These are non-updated parameters in the model and demonstrate an exponential behavior in this case. The interpolation from this simple $1/T$ to exponential falloff is well-approximated by a power-law decay of the form $1/T^{1+\nu}$ with $\nu > 0$.

We illustrate these general considerations with a number of examples. As one of the few cases we can treat analytically, we illustrate how the flow equations work in the case of inference on data drawn from a Gaussian distribution. This also includes the important case of a Gaussian Random Process, which is of relevance in the study of (untrained) neural

networks in the infinite width limit [23]. As a physically motivated numerical example, we ask how well an observer can learn the coupling constants of the 1D Ising model. In this setting, "data" amounts to sampling from the Boltzmann distribution of possible spin configurations, and inference corresponds to refining our prior estimates on the value of the nearest neighbor coupling. We indeed find that the trajectory of the coupling obeys the observable flow equation, and converges to a high level of accuracy.

It is also natural to ask whether we can apply these considerations even when we do not have a generative model for the probability distribution. A classic example of this sort is the "inference" performed by a neural network as it is learning. From the Bayesian perspective, neural networks are models that contain a large number of parameters given by the weights and biases of the network and training is a flow on those weights and biases induced by the training data set. See [24] for an introduction to neural networks aimed at physicists. Due to the large number of parameters a true quantitative analysis, as we did for the Ising model, is not possible. However, insofar as the neural network is engaging in rational inference, we should expect a flow equation to hold. To test this expectation we study the phenomenology of training a network as a function of the size of the data set. We consider simple dense feedforward networks (FF) and convolutional neural networks (CNNs) trained on the, by now classic, data sets of MNIST, Fashion-MNIST and CIFAR10.[2]

Quite remarkably, we find that the qualitative $1/T$ power-law behavior is emulated for the MNIST data set where the network final loss is very small and other power-laws occur for more complex data sets where the final loss is higher. The fact that our simple theoretical expectations match on to the rather opaque inference procedure of a neural network lends additional support to the formalism.

The rest of this paper is organized as follows. We begin in section 2 by reviewing Bayes' rule and then turn to the infinitesimal version defined by incremental updates and the induced flow for observables. After solving the flow equations exactly and in perturbation theory we then turn to some examples. First, in section 3 we present an analytic treatment in the context of inference for Gaussian data. In section 4 we study inference of coupling constants in the context of the 1D Ising model. In section 5 we turn to examples of neural network learning various data sets. We present our conclusions and potential avenues for future investigation in section 6. Appendix A presents a statistical mechanics interpretation of the Bayesian flow equations.

2 Bayes' Rule as a Dynamical System

In this section we present a physical interpretation of Bayes' rule as a dynamical system. By working in a limit where we have a large number of events $N \gg 1$ partitioned up into smaller number of events $N_k \gg 1$ such that $N_k/N \ll 1$, we show that this can be recast as

[2]These data sets are readily available with supporting notes in https://keras.io/api/datasets/

an integro-differential equation.

To begin, suppose we have observed some events $E = \{e_1, ..., e_N\}$, as drawn from some true distribution. In the Bayesian setting, we suppose that we have some model of the world specified by a posterior distribution conditional on observed events $f(\theta|e_1, ..., e_N) = \pi_{\text{post}}(\theta)$, which depends on "fitting parameters" $\theta = \{\theta^1, ..., \theta^m\}$ and our density for data conditional on θ specified as $f(e_1, ..., e_N|\theta)$. In what follows, we assume that there is a specific value $\theta = \alpha_*$ for which we realize the true distribution.[3] A general comment here is that we are framing our inference problem using Bayesian methods, which means that the parameters θ are themselves treated as values drawn from a random distribution. This is to be contrasted with how we would treat the inference problem as frequentists, where we would instead attempt to find a best estimate for these parameters (e.g. the mean and variance of a normal distribution). Rather, the notion of a Bayesian update means that additional "hyperparameters" for $\pi_{\text{post}}(\theta)$ are being updated as a function of increased data. For now, we keep this dependence on the hyperparameters implicit, but we illustrate later on how this works in some examples.

Now, assuming the observed events are conditionally independent of θ, we have:

$$\pi_{\text{post}}(\theta) = f(\theta|e_1, ..., e_N) \propto f(e_1, ..., e_N|\theta) \times \pi_{\text{prior}}(\theta), \tag{2.1}$$

where the constant of proportionality is fixed by the condition that $\pi_{\text{post}}(\theta)$ is properly normalized, i.e., we can introduce another distribution:

$$f(e_1, ..., e_N) = \int d\theta\, f(e_1, ..., e_N|\theta)\pi_{\text{prior}}(\theta), \tag{2.2}$$

and write:

$$\frac{\pi_{\text{post}}(\theta)}{\pi_{\text{prior}}(\theta)} = \frac{f(e_1, ..., e_N|\theta)}{f(e_1, ..., e_N)}. \tag{2.3}$$

Rather than perform one large update, we could instead consider partitioning up our events into separate sequences of events, which we can label as $E(k) = \{e_1(k), ..., e_{N_k}(k)\}$, where now we let $k = 1, .., , K$ such that $N_1 + ... + N_K = N$. Introducing the cumulative set of events:

$$S_k \equiv E(1) \cup ... \cup E(k), \tag{2.4}$$

we can speak of a sequential update, as obtained from incorporating our new data:

$$\frac{\pi_{k+1}(\theta)}{\pi_k(\theta)} = \frac{f(E(k+1)|S_k, \theta)}{f(E(k+1)|S_k)}. \tag{2.5}$$

This specifies a recursion relation and thus a discrete dynamical system. Indeed, writing

[3] A word on notation. We have chosen to write the fixed value of a given parameter as α_* instead of θ_*. We do this to emphasize that the θ's are to be treated as the values of a random variable, with α_* indicating what a frequentist might refer to as the estimator of this parameter.

$\pi_k(\theta) = \exp(\ell_k(\theta))$ and taking the logarithm of equation (2.5) yields the finite difference equation:

$$\ell_{k+1}(\theta) - \ell_k(\theta) = \log \frac{f(E(k+1)|S_k, \theta)}{f(E(k+1)|S_k)}. \tag{2.6}$$

To proceed further, we now make a few technical assumptions. First of all, we assume that each draw from the true distribution is independent so that we can write:

$$f(e_1, ..., e_N|\theta) = f\left(e_1|\theta\right) ... f\left(e_N|\theta\right). \tag{2.7}$$

Further, we assume that draws from the true distribution can always be viewed as part of the same parametric family as these densities:

$$f(E(k+1)|S_k) = \prod_{e \in E(k+1)} f(e|\alpha_k), \tag{2.8}$$

where α_k is the most likely estimate of θ given the data $E(k)$. Then, the finite difference equation of line (2.6) reduces to:

$$\ell_{k+1}(\theta) - \ell_k(\theta) = \sum_{e \in E(k+1)} \log \frac{f(e|\theta)}{f(e|\alpha_k)} \equiv N_{k+1} \left\langle \log \frac{f(y|\theta)}{f(y|\alpha_k)} \right\rangle_{E(k+1)}, \tag{2.9}$$

in the obvious notation.

We now show that in the limit $N \gg N_k \gg 1$, Bayesian updating is well-approximated by an integro-differential flow equation. The large N_k limit means that we can approximate the sum on the righthand side of equation (2.9) by an integral:

$$\ell_{k+1}(\theta) - \ell_k(\theta) = N_{k+1} \int dy \; f(y|\alpha_*) \log \frac{f(y|\theta)}{f(y|\alpha_k)} + ..., \tag{2.10}$$

where the correction terms are subleading in a $1/N_{k+1}$ expansion and we have switched from referring to events e_i by their continuous analogs y. The righthand side can be expressed in terms of a difference of two KL divergences, so we can write:

$$\ell_{k+1}(\theta) - \ell_k(\theta) = N_{k+1} \left(D_{KL}(\alpha_*||\alpha_k) - D_{KL}(\alpha_*||\theta)\right) + ... \tag{2.11}$$

We now approximate the lefthand side. The small N_k/N limit means we can replace the finite difference on the lefthand side by a derivative. More precisely, introduce a continuous parameter $\tau \in [0, 1]$, which we can partition up into discretized values τ_k with small timestep $\delta\tau_k = \tau_{k+1} - \tau_k$ between each step:

$$\tau_k \equiv \frac{1}{N}(N_1 + ... + N_k) \;\text{ and }\; \delta\tau_k \equiv \frac{N_k}{N}. \tag{2.12}$$

So, instead of writing $\pi_k(\theta)$, we can instead speak of a continuously evolving family of distributions $\pi(\tau;\theta)$. Similarly, we write $\alpha(\tau)$ to indicate the continuous evolution used in the parameter appearing in $f(e|\alpha_k) = f(e|\alpha(\tau_k))$. The finite difference can therefore be approximated as:

$$\ell_{k+1}(\theta) - \ell_k(\theta) = \ell'(\theta;\tau_k)\delta\tau_k + \frac{1}{2}\ell''(\theta;\tau_k)(\delta\tau_k)^2 + ..., \tag{2.13}$$

where the prime indicates a partial derivative with respect to the timestep, e.g. $\ell' = \partial\ell/\partial\tau$. Working in the approximation $N \gg N_k \gg 1$, equation (2.11) is then given to leading order by:

$$\frac{1}{N}\frac{\partial\ell(\theta;\tau)}{\partial\tau} = D_{KL}(\alpha_*||\alpha(\tau)) - D_{KL}(\alpha_*||\theta) + ... \tag{2.14}$$

To avoid overloading the notation, we write this as:

$$\frac{1}{N}\frac{\partial\ell(\theta;\tau)}{\partial\tau} = D(\alpha(\tau)) - D(\theta) + ... \tag{2.15}$$

Observe that the N_k dependence has actually dropped out from the righthand side; it only depends on the total number of events N. In terms of the posterior $\pi(\theta;\tau) = \exp\ell(\theta;\tau)$, we have:

$$\frac{\partial\pi(\theta;\tau)}{\partial\tau} = \pi(\theta;\tau)\frac{\partial\ell(\theta;\tau)}{\partial\tau} = N\pi(\theta;\tau)\left(D(\alpha(\tau)) - D(\theta)\right). \tag{2.16}$$

A formal solution to the posterior is then:

$$\pi(\theta;\tau) = \exp N\int^{\tau} d\tau' \ \left(D(\alpha(\tau)) - D(\theta)\right). \tag{2.17}$$

2.1 Observable Flows

Given an inference scheme over a random variable with parameters θ, we regard an observable as a function of the parameters; i.e. in the classical sense with the parameter space serving as a phase space for the theory. Given such an observable, $O(\theta)$, we now ask about the τ dependence, as obtained by evaluating the expectation value:

$$\overline{O}(\tau) = \int d\theta\ O(\theta)\pi(\theta;\tau). \tag{2.18}$$

This is subject to a differential equation, as obtained by differentiating both sides with respect to τ:

$$\frac{\partial\overline{O}(\tau)}{\partial\tau} = N\int d\theta\ O(\theta)\pi(\theta;\tau)\left(D(\alpha(\tau)) - D(\theta)\right), \tag{2.19}$$

and using equation (2.16). More compactly, we can write this as:

$$\left(\frac{\partial}{\partial\tau} - N(D(\alpha(\tau)) - \overline{D}(\tau))\right)\overline{O}(\tau) = -N\text{Var}(O,D), \tag{2.20}$$

where $\text{Var}(O,D)$ is just the variance between the operators O and D:

$$\text{Var}(O,D) = \left\langle (O-\overline{O})(D-\overline{D})\right\rangle_{\pi(\theta;\tau)}. \tag{2.21}$$

Before proceeding, let us consider the trivial observable $O(\theta) = 1$. This observable determines the normalization condition imposed on $\pi(\theta;\tau)$ as a formal probability density. Using equation (2.19), we find:

$$0 = N\int d\theta\,\pi(\theta;\tau)(D(\alpha(\tau)) - D(\theta)) = N(D(\alpha(\tau)) - \bar{D}(\tau)) \tag{2.22}$$

which implies that:

$$D(\alpha(\tau)) = \overline{D}(\tau)\,. \tag{2.23}$$

Thus, the equation obeyed by arbitrary observables is given by:

$$\frac{\partial}{\partial\tau}\overline{O}(\tau) = -N\text{Var}(O,D)\,. \tag{2.24}$$

To proceed further, it is helpful to work in terms of a rescaled time coordinate $T \equiv \tau N$. In terms of this variable, our equation becomes:

$$\frac{\partial}{\partial T}\overline{O}(T) = -\text{Var}(O,D)\,, \tag{2.25}$$

so that the N dependence has dropped out. By expanding the covariance we can also write this equation as:

$$\frac{\partial\overline{O}}{\partial T} = \overline{O}\;\overline{D} - \int d\theta\pi(\theta;T)O(\theta)D(\theta)\,. \tag{2.26}$$

We now turn to the interpretation of this equation in various regimes.

2.2 Generic Observables at Late T

Our approach to analyzing this equation will begin with expanding the observables in a power series to obtain manageable expressions that have interpretations as governing late T behavior. In particular, we will use the expansion of the divergence:

$$D(\theta) \approx \frac{1}{2}\mathcal{I}_{ij}\Big|_{\alpha_*}(\theta-\alpha_*)^i(\theta-\alpha_*)^j + \mathcal{O}(1/T^3)\,, \tag{2.27}$$

where $\mathcal{I}|_{\alpha_*} = \mathcal{I}_*$ is the Fisher information metric evaluated at the true underlying mean of the parameter distribution. It is necessary to expand around this parameter value if one wishes to represent the KL-Divergence as a quadratic form with no constant or linear contribution – that is, we have used the fact that α_* is the minimizing argument of $D(\theta)$ to set the constant and linear order terms in (2.27) to zero. Then, at late times any arbitrary observable satisfies the equation:

$$\begin{aligned}\frac{\partial \overline{O}}{\partial T} =& \frac{1}{2}\mathcal{I}^*_{kl}\,\overline{O}\int d\theta \pi(\theta;T)(\theta-\alpha_*)^k(\theta-\alpha_*)^l \\ &- \frac{1}{2}\mathcal{I}^*_{kl}\int d\theta\pi(\theta;T)O(\theta)(\theta-\alpha_*)^k(\theta-\alpha_*)^l + \mathcal{O}(1/T^3)\,. \end{aligned} \tag{2.28}$$

2.3 Centralized Moments at Late T

At this juncture, let us turn our attention to a special class of observables. Namely, those of the form:

$$C^{i_1\ldots i_{2l}}(\theta) = \prod_{j=1}^{2l}(\theta-\alpha_*)^{i_j}\,. \tag{2.29}$$

Such observables are precisely the pre-integrated centralized moments of the T-posterior. More precisely, this is true at sufficiently late times in which $\alpha(T) \sim \alpha_*$, meaning the parameter distribution has centralized around its true mean. The observable flow equation for these observables therefore governs the T-dependence of the centralized moments:

$$\overline{C}^{i_1\ldots i_{2l}}(T) = \int d\theta\pi(\theta;T)\prod_{j=1}^{2l}(\theta-\alpha_*)^{i_j} \tag{2.30}$$

which satisfy the differential equation:

$$\begin{aligned}\frac{\partial \overline{C}^{i_1\ldots i_{2l}}}{\partial T} =& \frac{1}{2}\mathcal{I}^*_{kl}\overline{C}^{i_1\ldots i_{2l}}\int d\theta\pi(\theta;T)(\theta-\alpha_*)^k(\theta-\alpha_*)^l \\ &- \frac{1}{2}\mathcal{I}^*_{kl}\int d\theta\pi(\theta;T)C^{i_1\ldots i_{2l}}(\theta)(\theta-\alpha_*)^k(\theta-\alpha_*)^l + \mathcal{O}(1/T^3)\,. \end{aligned} \tag{2.31}$$

Using our notation this can be written more briefly as:

$$\frac{\partial}{\partial T}\overline{C}^{i_1\ldots i_{2l}} = \frac{1}{2}\mathcal{I}^*_{kl}\overline{C}^{i_1\ldots i_{2l}}\overline{C}^{kl} - \frac{1}{2}\mathcal{I}^*_{kl}\overline{C}^{i_1\ldots i_{2l}\,kl} + \mathcal{O}(1/T^3)\,. \tag{2.32}$$

2.4 Centralized Moments for Gaussian Distributions

The analysis we have done up to this point is valid for the centralized moments of any arbitrary posterior distribution. Now, we will specialize to the case that the posterior dis-

tribution is Gaussian at late T. This is quite generic, and will always be the case when the parameters being inferred over are truly non-stochastic. The aspect of the Gaussian model which is especially useful is that we can implement Isserlis', a.k.a, Wick's theorem, to reduce $2l$-point functions into sums of products of 2-point functions (i.e. the covariance). In particular, we have the formula:

$$\overline{C}^{i_1 \dots i_{2l}} = \sum_{p \in \mathcal{P}^2_{2l}} \prod_{(r,s) \in p} \overline{C}^{i_r i_s} \,. \tag{2.33}$$

Here $\mathcal{P}^2_{2l}$ is the set of all partitions of $2l$ elements into pairs, and a generic element $p \in \mathcal{P}^2_{2l}$ has the form $p = \{(r_1, s_1), ..., (r_l, s_l)\}$, hence the notation in the product. Isserlis' / Wick's theorem implies that for a Gaussian model it is sufficient to understand the T-dependent behavior of the 2-point function, and the behavior of the other $2l$-point functions immediately follow suit. Therefore, let us consider the equation satisfied by $\overline{C}^{ij}$:

$$\frac{\partial}{\partial T}\overline{C}^{ij} = \frac{1}{2}\mathcal{I}^*_{kl}\overline{C}^{ij}\overline{C}^{kl} - \frac{1}{2}\mathcal{I}^*_{kl}\overline{C}^{ijkl} + \mathcal{O}(1/T^3) \,. \tag{2.34}$$

Using (2.33) we can write:

$$\overline{C}^{ijkl} = \overline{C}^{ij}\overline{C}^{kl} + \overline{C}^{ik}\overline{C}^{jl} + \overline{C}^{il}\overline{C}^{jk} \tag{2.35}$$

Plugging this back into equation (2.34) we get:

$$\begin{aligned} \frac{\partial}{\partial T}\overline{C}^{ij} =& \frac{1}{2}\mathcal{I}^*_{kl}\overline{C}^{ij}\overline{C}^{kl} - \frac{1}{2}\mathcal{I}^*_{kl}\left(\overline{C}^{ij}\overline{C}^{kl} + \overline{C}^{ik}\overline{C}^{jl} + \overline{C}^{il}\overline{C}^{jk}\right) + \mathcal{O}(1/T^3) && (2.36) \\ =& -\mathcal{I}^*_{kl}\overline{C}^{ik}\overline{C}^{jl} + \mathcal{O}(1/T^3) && (2.37) \end{aligned}$$

where we have used the fact that $\mathcal{I}^*_{ij}$ is symmetric. The equation satisfied by the covariance of a Gaussian distribution at late T is thus:

$$\frac{\partial}{\partial T}\overline{C}^{ij} + \mathcal{I}^*_{kl}\overline{C}^{ik}\overline{C}^{jl} = 0 + \mathcal{O}(1/T^3) \,. \tag{2.38}$$

At this stage, we can recognize our equation as predicting familiar behavior for the $2l$-point functions of a Gaussian posterior.

2.5 Cramér-Rao Solution

Cramér and Rao [9] demonstrated that there is a lower bound on the variance of any unbased estimator which is given by the inverse of the Fisher information. An estimator that saturates this bound is as efficient as possible and reaches the lowest possible mean squard error.

We now show that the evolution equation (2.38) is satisfied when the model saturates the Cramér-Rao bound. That is, at sufficiently late T, we take the bound to be saturated

such that:

$$\overline{C}^{ij}_{CR} = \frac{\mathcal{I}^{ij}_*}{T} + \mathcal{O}(1/T^2)\,. \tag{2.39}$$

One can see that (2.39) is then a solution to (2.38) by straightforward computation:

$$\frac{\partial}{\partial T}\overline{C}^{ij}_{CR} = \frac{\partial}{\partial T}(\frac{\mathcal{I}^{ij}_*}{T}) = -\frac{\mathcal{I}^{ij}_*}{T^2} = -\mathcal{I}^*_{kl}\frac{\mathcal{I}^{ik}_*}{T}\frac{\mathcal{I}^{jl}_*}{T} = -\mathcal{I}^*_{kl}\overline{C}^{ik}_{CR}\overline{C}^{jl}_{CR}\,. \tag{2.40}$$

We leave it implicit that there could be correction terms of higher order in the expansion $1/T$. The T-dependent behavior of any arbitrary $2l$-point function in the theory is subsequently given by:

$$\overline{C}^{i_1\ldots i_{2l}}_{CR} = \frac{1}{T^l}\sum_{p\in\mathcal{P}^2_{2l}}\prod_{(r,s)\in p}\mathcal{I}^{i_r i_s}_*\,. \tag{2.41}$$

2.6 Higher Order Effects

The assumptions that led to the saturation of the Cramér-Rao bound are based on two leading order approximations: Firstly, that the maximum likelihood parameter is near the "true" value, and secondly, that the posterior distribution is approximately Gaussian. Both of these assumptions become increasingly valid at late update times ie. with more data, hence we have expressed our equations as a power series expansion in the small quantity, $1/T$.

One can look for corrections to these assumptions by systematically reintroducing higher order effects via a perturbation series. To be precise, as one moves into earlier update times, there will be contributions to the KL-Divergence which are higher than quadratic order in θ. Similarly, as one moves away from a Gaussian posterior, either by moving back in time or by including some additional implicit randomness to the parameters, one finds new terms in the posterior distribution away from Gaussianity which also lead to new terms in the Observable Flow equations.

We provide an outline of the perturbative analysis including effects both from the additional higher order expansion of the KL-Divergence and from the deviation of the Posterior from Gaussianity. To begin, the KL-Divergence can be Taylor expanded into a power series in the un-integrated n-point functions as follows:

$$D(\theta) = \sum_{n=2}^{\infty}\frac{1}{n!}\mathcal{I}^{(n)}_{i_1\ldots i_n}C^{i_1\ldots i_n}(\theta) \tag{2.42}$$

where

$$\mathcal{I}^{(n)}_{i_1\ldots i_n} = \prod_{j=1}^{n}\frac{\partial}{\partial\theta^{i_j}}D(\theta)\Big|_{\theta=\alpha_*}\,. \tag{2.43}$$

We will now perturb the posterior distribution away from Gaussiantiy as follows:

$$\pi(\theta; T) = \text{Gaussian} \cdot \ e^{-\lambda f(\theta)}, \tag{2.44}$$

where in the above, $f(\theta)$ is treated as an arbitrary bounded function, and the size of the small parameter λ governing the perturbation is, in general, dependent on the update time T. We take the unperturbed Gaussian Distribution to be centered around the maximum likelihood estimate (MLE) with covariance given by the two-point correlator. The expectation value of an observable $O(\theta)$ can therefore be expanded in a series with respect to λ:

$$\langle O \rangle_\pi = \langle e^{-\lambda f(\theta)} O(\theta) \rangle_{\text{Gauss}} = \sum_{n=0}^{\infty} \frac{(-1)^n \lambda^n}{n!} \langle f(\theta)^n O(\theta) \rangle_{\text{Gauss}} \,. \tag{2.45}$$

Now, one can input these series expansions back into (2.26) to obtain higher order corrections to the scaling behavior of any arbitrary observable expectation value. Doing so explicitly and terminating the perturbation series at order N in powers of λ and order M in powers of $1/T$ we find:

$$\begin{aligned}
&\sum_{n=0}^{N} \frac{(-1)^n \lambda^n}{n!} \frac{\partial}{\partial T} \langle f(\theta)^n O(\theta) \rangle_{\text{Gauss}} \\
&\quad = \sum_{n=0}^{N} \frac{(-1)^n \lambda^n}{n!} \langle f(\theta)^n O(\theta) \rangle_{\text{Gauss}} \sum_{n'=0}^{N} \sum_{m=2}^{M} \frac{(-1)^{n'} \lambda^{n'}}{n'!} \frac{1}{m!} \mathcal{I}^{(m)}_{i_1 \ldots i_m} \langle f(\theta)^{n'} C^{i_1 \ldots i_m}(\theta) \rangle_{\text{Gauss}} \\
&\qquad - \sum_{n=0}^{N} \sum_{m=2}^{M} \frac{(-1)^n \lambda^n}{n!} \frac{1}{m!} \mathcal{I}^{(m)}_{i_1 \ldots i_m} \langle f(\theta)^n O(\theta) C^{i_1 \ldots i_m}(\theta) \rangle_{\text{Gauss}} + \mathcal{O}(T^{-M}, \lambda^N) \,.
\end{aligned} \tag{2.46}$$

We will work just to the next to leading order. To see the impact of these higher order effects, we performed two simple numerical experiments in which the Observable Flow for the two-point function can be solved exactly. In each case we consider only a single parameter being inferred upon during the Bayesian update.

1. In the first numerical experiment, we consider a perturbation in which we accept terms in the expansion of the KL-Divergence up to fourth order, but in which the posterior is assumed to remain approximately Gaussian. In this case, the Observable Flow equation for the second centralized moment, $\overline{C}^{(2)}$, becomes:

$$\frac{\partial \overline{C}^{(2)}}{\partial T} = -\mathcal{I}^{(2)} (\overline{C}^{(2)})^2 - \frac{1}{2} \mathcal{I}^{(4)} (\overline{C}^{(2)})^3 \,. \tag{2.47}$$

   We fix the $\mathcal{I}^{(2)}$ and $\mathcal{I}^{(4)}$ by hand and then use the above equation to solve for the time dependence of $C^{(2)}$. This is done using numerical methods, and subsequently fit to a

$\mathcal{I}^{(4)}/\mathcal{I}^{(2)}$	a	b	c
0.1	0.96	0.99	-4.7×10^{-6}
0.2	0.92	0.98	-2.5×10^{-4}
0.3	0.88	0.96	-6.9×10^{-4}
0.4	0.83	0.94	-1.3×10^{-3}
0.5	0.78	0.92	-2.1×10^{-3}
0.6	0.74	0.89	-2.9×10^{-3}
0.7	0.70	0.87	-3.9×10^{-3}
0.8	0.66	0.84	-4.9×10^{-3}
0.9	0.62	0.82	-6.0×10^{-3}
1	0.58	0.79	-7.1×10^{-3}
1.1	0.55	0.77	-8.2×10^{-3}
1.2	0.52	0.74	-9.3×10^{-3}
1.3	0.49	0.72	-1.0×10^{-2}
1.4	0.47	0.69	-1.2×10^{-2}
1.5	0.44	0.67	-1.3×10^{-2}

Table 1: Result of the first numerical experiment involving perturbations to the KL divergence. This entails numerically solving equation (2.47) by fitting to a power-law $aT^{-b} + c$, as in equation (2.48). In all cases, the R^2 value is $\sim 0.99 + O(10^{-3})$. As the ratio $\mathcal{I}^{(4)}/\mathcal{I}^{(2)}$ increases, we observe that the size of the constant offset increases in magnitude and the exponent b in T^{-b} decreases.

power-law of the form

$$C^{(2)} = \frac{a}{T^b} + c\,. \tag{2.48}$$

The resulting curve has a power in which $b < 1$, and typically in the range between 0.65 and 1 depending on the ratio between $\mathcal{I}^{(2)}$ and $\mathcal{I}^{(4)}$ (only the ratio matters). The results of this numerical experiment are given in table (1).

2. In the second numerical experiment, we consider perturbations away from Gaussianity in which we accept terms of order λ with $f(\theta) = \theta^4$, but regard the KL-Divergence as sufficiently well approximated at quadratic order. In this case, the two-point function has the form:

$$\overline{C}^{(2)} = \overline{C}^{(2)}_G - 15\lambda(\overline{C}^{(2)}_G)^3 \tag{2.49}$$

where $\overline{C}^{(2)}_G$ is the expectation of the second centralized moment with respect to the Gaussian distribution. The Gaussian two-point function subsequently satisfies the ODE:

$$\frac{\partial}{\partial T}\left(\overline{C}^{(2)}_G - 15\lambda(\overline{C}^{(2)}_G)^3\right) = \frac{1}{2}\mathcal{I}^{(2)}(\overline{C}^{(2)}_G)^2 - 15\lambda\mathcal{I}^{(2)}(\overline{C}^{(2)}_G)^3 - \frac{1}{2}\mathcal{I}^{(2)}\left(3(\overline{C}^{(2)}_G)^2 - 105\lambda(\overline{C}^{(2)}_G)^4\right). \tag{2.50}$$

Again, this equation can be solved using numerical methods, and fit to a power-law of

$\lambda/\mathcal{I}^{(2)}$	a	b	c
0.01	0.78	0.89	-3.9×10^{-3}
0.02	0.63	0.80	-7.7×10^{-3}
0.03	0.51	0.71	-1.1×10^{-2}
0.04	0.42	0.63	-1.7×10^{-2}
0.05	0.36	0.55	-2.1×10^{-2}
0.06	0.31	0.49	-2.6×10^{-2}
0.07	0.28	0.44	-3.0×10^{-2}
0.08	0.25	0.40	-3.5×10^{-2}
0.09	0.23	0.36	-3.9×10^{-2}
0.1	0.22	0.32	-4.3×10^{-2}
0.11	0.21	0.29	-4.8×10^{-2}
0.12	0.20	0.26	-5.3×10^{-2}
0.13	0.19	0.24	-5.9×10^{-2}
0.14	0.19	0.21	-6.5×10^{-2}
0.15	0.19	0.19	-7.2×10^{-2}

Table 2: Result of the second numerical experiment involving perturbation away from Gaussianity. This involves numerical solutions to equation (2.50) obtained by fitting to a power-law $aT^{-b} + c$, as in equation (2.48). In all cases, the R^2 value is $\sim 0.99 + O(10^{-3})$. As the ratio $\lambda/\mathcal{I}^{(2)}$ increases, we observe that the size of the constant offset increases in magnitude and the exponent b in T^{-b} decreases.

the form (2.48). The resulting curves exhibit scaling in which $b < 1$. The size of b is governed by the ratio of $\mathcal{I}^{(2)}$ and λ. The results of this experiment for various values of this ratio can be found in (2).

The upshot of these experiments, and of this section, is that higher order corrections to the Cramér-Rao solution of the Observable Flow equation can be implemented systematically by considering a bi-perturbation series which takes into account both changes to the KL-Divergence due to the proximity of the MLE and the data generating parameter, and deviation of the posterior distribution from a Gaussian. The impact of these corrections are to decrease the steepness of the learning curve, in accord with expectations from the Cramér-Rao Bound.

2.7 Hidden Variables and the Breaking of the Cramér-Rao Bound

In the last section we showed that the Dynamical Bayesian Inference scheme respects the Cramér-Rao Bound as an upper limit on the rate at which the two-point function can scale with respect to the update time. An analogous phenomenon occurs in conformal field theories, and is often referred to as the "unitarity bound," (see [25]) which controls the

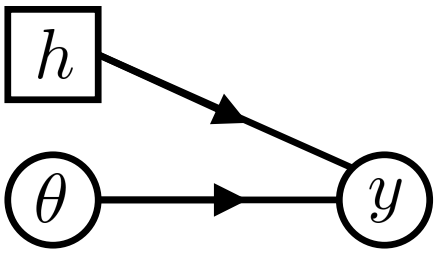


Figure 1: Graphical model depiction of a conditional distribution $p(y|\theta, h)$ we wish to infer, where now we explicitly account for both visible training parameters θ and hidden parameters h. The presence of such hidden variables can impact the inference scheme.

strength of correlations as a function of distance in the underlying spacetime.[4]

Now, in the physical setting, a simple way to violate constraints from unitarity is to treat the system under consideration as "open", i.e. degrees of freedom can flow in or out. In the context of Bayesian inference, we have a direct analogy in terms of hidden variables h which may also impact the distribution $f(y|\theta, h)$, but which we may not be able to access or even parameterize (see figure 1). This can lead to dissipative phenomena, as well as driving phenomena.[5]

The basic setup is to consider a data generating distribution which belongs to a parametric family, $p(y \mid \theta, h)$ which depends on two sets of variables, θ (visible) and h (hidden). An experimenter who is observing the data generated by this distribution may, either due to ignorance or by choice,[6] train a model that depends only on the variables, θ, i.e. $f(y \mid \theta)$. Observables involving the hidden variables will evolve indirectly over the course of the update due to the changing of the total joint probability density over trained and hidden variables, however this evolution is not necessarily governed by a Bayesian updating scheme. Insofar as we have a reliable inference scheme at all, we can neglect the explicit time variation in the

[4]For example, in a relativistic conformal field theory (CFT) in $D \geq 2$ spacetime dimensions, a scalar primary operator $O(x)$ of scaling dimension Δ will have two-point function: $\langle O^\dagger(x)O(x)\rangle \sim 1/|x|^{2\Delta}$, and $\Delta \geq \Delta_0$ specifies a unitarity bound which is saturated by a free scalar field, i.e. a Gaussian random field. In the context of the Cramér-Rao bound, the limiting situation is again specified by the case of a Gaussian. The analogy is not perfect, however, because we do not have the same notion of spacetime locality in Bayesian inference, and the referencing to the Gaussian case is different ($\Delta > 1$ for a CFT, but powers $1 \ / \ T^b$ for $b < 1$ in the case of Cramér-Rao. It is, nevertheless, extremely suggestive, and the physical intuition about how to violate various unitarity bounds will indeed have a direct analogy in the statistical inference setting as well.

[5]Returning to the case of a CFT in D dimensions, observe that a 4D free scalar can be modeled in terms of a collection of 3D scalars coupled along a discretized dimension. The unitarity bound for a scalar primary operator in 3D is $\Delta_{3D} \geq 1/2$, while in 4D it is $\Delta_{4D} \geq 1$.

[6]For example, a model builder may choose to fix the "hidden" variables if they are close to their maximum likelihood values, or do not vary greatly across samples. In this regard the hidden variables may more aptly be identified as non-dynamical rather than hidden, but their impact is the same either way.

hidden variables, i.e., we can treat them as non-dynamical. Summarizing, only observables in the visible parameters θ will satisfy Observable Flow equations.

Now, although we are treating the hidden parameters as non-dynamical, they still enter in the KL-Divergence between the likelihood and the data generating model and therefore impact all observable flow equations. To be precise, at leading order

$$D_{KL}(\Phi_* \parallel \Phi) = \frac{1}{2}\mathcal{I}_{AB}(\Phi - \Phi_*)^A(\phi - \Phi_*)^B \,. \tag{2.51}$$

Here we are using notation in which $\Phi = (\theta, h)$ is the complete set of parameters, and $\Phi_* = (\alpha_*, h_*)$ (by abuse of notation) denotes the actual parameters from which we draw the distribution. The index $A = (i, I)$ spans all parameters, with the index $i = 1, ..., n$ corresponding to the trained parameters, and the index $I = 1, ...m$ corresponding to the hidden parameters. In this more explicit notation, the information metric appearing in (2.51) takes the form

$$\mathcal{I} = \mathcal{I}_{ij}d\theta^i \otimes d\theta^j + \mathcal{I}_{IJ}dh^I \otimes dh^J + \mathcal{I}_{iI}d\theta^i \otimes dh^I + \mathcal{I}_{Ii}dh^I \otimes d\theta^i \tag{2.52}$$

where $\mathcal{I}_{iI} = \mathcal{I}_{Ii}$.

Consider next the scaling of the two-point function between trained parameters. The scaling of the two-point function over the course of the Dynamical Bayesian inference scheme is dictated by the differential equation governing $C^{ij}(\theta)$. Using the observable flow equation, and remembering to include the complete KL-Divergence including contributions from both hidden and trained variables, one finds the equation:[7]

$$\frac{\partial}{\partial T}\overline{C}^{ij} = -\mathcal{I}_{kl}\overline{C}^{ik}\overline{C}^{jl} - \mathcal{I}_{KL}\overline{C}^{iK}\overline{C}^{jL} - 2\mathcal{I}_{lL}\overline{C}^{il}\overline{C}^{jL} \,. \tag{2.54}$$

The new contributions are the final two terms on the righthand side which depend on the covariance between trained and hidden variables. As noted above, such observables are to be considered as slowly varying in comparison with observables involving only trained parameters. Hence, for the purposes of this exercise we can regard these covariances as approximately constant in time.

We now make the well-motivated assumption that the joint probability model between the trained and hidden parameters is such that the covariance amongst all pairs of trained

[7]Here we have extended the notation $\overline{C}^{A_1 \ldots A_n}$ to refer to the expectation value of n-point functions including arbitrary combinations of trained and hidden parameters:

$$\overline{C}^{A_1 \ldots A_n} = \int d\theta \; dh \; \rho(h \mid \theta)\pi(\theta; T) \prod_{i=1}^{n} (\Phi - \Phi_*)^{A_i} \tag{2.53}$$

$\rho(h \mid \theta)$ is a fixed conditional distribution encoding the probability density for hidden parameters given trained parameters.

and hidden parameters satisfies the series of inequalities:

$$\overline{C}^{ij} \gg \overline{C}^{iI} \gg \overline{C}^{IJ} \tag{2.55}$$

for all values of i, j, I, J. This assumption (along with the assumption that hidden variable observables are slowly varying) basically serve to justify the distinction between hidden and trained variables. If the hidden variables were rapidly varying and/or highly correlated with observed data it would not be reasonable to exclude them from the model. Alternatively, we may use these conditions as a criterion for *defining* hidden variables as those variables which vary slowly and have limited covariance with relevant parameters.[8] Under these assumptions, we can then write the ODE governing the scaling of the two-point function as:

$$\frac{\partial}{\partial T}\overline{C}^{ij} = -\mathcal{I}_{kl}\overline{C}^{ik}\overline{C}^{jl} - 2\mathcal{I}_{lL}\overline{C}^{il}\overline{C}^{jL}. \tag{2.56}$$

From equation (2.56) we can recognize that the evolution of the two-point function, (as well as the n-point functions) depends directly on the covariance between trained and hidden parameters. Note also that coupling to the hidden variables involves the off-diagonal terms of the information metric, $\mathcal{I}_{lL}$, which can in principle be either positive or negative, provided the whole metric is still positive definite. This can lead to a flow of information into the visible system (driving), or leakage out (dissipation).

The presence of this additional coupling to the hidden variables can produce an apparent violation of the Cramér-Rao bound on just the visible sector. To see why, it is already enough to consider the simplest case where we have a single visible parameter θ, with the rest viewed as hidden. In this case, the observable flow equation is:

$$\frac{\partial}{\partial T}\overline{C}^{11} = -\mathcal{I}_{11}(\overline{C}^{11})^2 - \beta\overline{C}^{11} + \mathcal{O}(\beta^2) \tag{2.57}$$

where here

$$\beta = 2\mathcal{I}_{1I}\overline{C}^{1I} \tag{2.58}$$

is twice the sum of the covariances of the trained parameter with the hidden parameters, and we have made it explicit that this is a leading order result in the size of these correlations. Dropping the order β^2 terms, the differential equation (2.57) can be solved exactly:

$$\overline{C}^{11}(T) = \frac{\kappa\beta}{e^{\beta T} - \kappa\mathcal{I}_{11}}, \tag{2.59}$$

where the constant κ depends on the initial conditions. Observe that for $\beta > 0$, $\overline{C}^{11}(T)$ decays exponentially at large T, i.e., faster than $1/T$. We interpret this as driving information into the visible sector. Conversely, for $\beta < 0$, we observe that the solution asymptotes to

[8] This is again reminiscent of the splitting between fast and slow modes which one uses in the analysis of renormalization group flows.

$-\beta/\mathcal{I}_{11} > 0$, i.e., we are well above the Cramér-Rao bound (no falloff at large T at all). We interpret this as dissipation: we are continually losing information.

In the above, we made several simplifying assumptions in order to analytically approximate the solution to the observable flow equations. At a phenomenological level, the interpolation from a simple $1/T$ behavior to an exponential decay law can be accomplished by a more general power-law of the form $1/T^{1+\nu}$ with $\nu > 0$, dependent on the particular inference scheme. This will be borne out by our numerical experiments, especially the ones in section 5 involving inference in a neural network, where we study the loss function and its dependence on T.

It is interesting to note that the crossover between a power-law and exponential decay is also implicitly tied to the accuracy of the underlying model. This suggests that at lower accuracy there is more information left for the algorithm to draw into its estimates. As the accuracy improves, the available information decreases and hence the driven behavior is slowly deactivated, resulting in more approximately power-law type behavior. Stated in this way, it is interesting to ponder what the precise nature of this crossover is, and whether it may be regarded as a kind of phase transition. We leave a more fundamental explanation of this crossover behavior to future work.

3 Dynamical Bayes for Gaussian Data

To give an analytic example of dynamical Bayesian updating, we now consider the illustrative case of sampling from a Gaussian distribution. An interesting special case is that of the Gaussian random process which can also be used to gain insight into the inference of neural networks (see, e.g., [23]).

3.1 Analysis for Multivariate Gaussian Data

A d-dimensional Gaussian random variable can be regarded as a random variable distributed according to a family of distributions governed by two parameters – a mean vector μ, and a symmetric, positive semi-definite covariance matrix Σ. Explicitly:

$$f(y \mid \mu, \Sigma) = ((2\pi)^d \det(\Sigma))^{-1/2} \ \exp(-\frac{1}{2}(y-\mu)^\intercal \Sigma^{-1}(y-\mu)) \,. \tag{3.1}$$

Here Σ^{-1} is the matrix inverse of the covariance; $\Sigma\Sigma^{-1} = \mathbb{I}$. By a simple counting argument, the number of free parameters governing the distribution of a d-dimensional Gaussian random variable is $d + \frac{d(d+1)}{2}$.

Bayesian inference over Gaussian data consists in determining a posterior distribution in the space of parameters $\Theta = (\mu, \Sigma)$. We can be slightly more general by allowing for

reparameterizations of the space of parameters in terms of some $\theta \in \mathcal{S} \subset \mathbb{R}^{(d+\frac{d(d+1)}{2})}$, that is:

$$\Theta = \Theta(\theta) = (\mu(\theta), \Sigma(\theta)) \,. \tag{3.2}$$

Hence, the result of a Dynamical Bayesian inference procedure on Gaussian data is to determine a flow in the parameters, $\theta = \alpha(T)$, giving rise to a flow in the posterior distribution $\pi(\theta; T)$.

In the case of the Gaussian distribution, and many other standard distributions for that matter, we can say slightly more than what we could when the family governing data remains unspecified. In particular, we have an explicit form for the KL-Divergence between multivariate Gaussian distributions:

$$D_{KL}((\mu_0, \Sigma_0) \parallel (\mu_1, \Sigma_1)) = \frac{1}{2}\left(\mathrm{tr}(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^\intercal \Sigma_1^{-1}(\mu_1 - \mu_0) + \ln(\frac{\det(\Sigma_1)}{\det(\Sigma_0)}) - d\right) . \tag{3.3}$$

This can be expressed in terms of θ_0 and θ_1 by composition with the reparameterization 3.2 provided $(\mu_a, \Sigma_a) = (\mu(\theta_a), \Sigma(\theta_a))$ for $a = 0, 1$:

$$D_{KL}(\theta_0 \parallel \theta_1) = \frac{1}{2}\begin{pmatrix} \mathrm{tr}(\Sigma(\theta_1)^{-1}\Sigma(\theta_0)) + \ln(\frac{\det(\Sigma(\theta_1))}{\det(\Sigma(\theta_0))}) - d) \\ +(\mu(\theta_1) - \mu(\theta_0))^\intercal \Sigma(\theta_1)^{-1}(\mu(\theta_1) - \mu(\theta_0)) \end{pmatrix} . \tag{3.4}$$

Given the flowing of the parameters, $\alpha(T)$, and the true underlying parameters, α_*, the posterior distribution is given by the solution to the Dynamical Bayesian updating equation:

$$\pi(\theta; T) = \exp\left(\int_0^T dT' (D_{KL}(\alpha_* \parallel \alpha(T')) - D_{KL}(\alpha_* \parallel \theta))\right) . \tag{3.5}$$

This solution can be written in the form:

$$\pi(\theta; T) = \exp\left(-T D_{KL}(\alpha_* \parallel \theta)\right) \exp\left(N \int_0^T dT' \, D_{KL}(\alpha_* \parallel \alpha(T'))\right) . \tag{3.6}$$

Note that the posterior distribution is proportional to the exponentiated KL-Divergence evaluated against the true underlying model parameter – a standard result from the theory of large deviations:

$$\pi(\theta; T) \propto \exp\left(-T D_{KL}(\alpha_* \parallel \theta)\right) . \tag{3.7}$$

Using the explicit form of the KL-Divergence for the normal distribution we find:

$$\pi(\theta; T) \propto \det(\Sigma(\alpha_*)) \det(\Sigma(\theta))^{-1} \exp\left(T\left\{-\frac{1}{2}\mathrm{tr}(\Sigma(\alpha_*)\Sigma(\theta)^{-1}) - \frac{1}{2}(\mu(\theta) - \mu(\alpha_*))\Sigma(\theta)^{-1}(\mu(\theta) - \mu(\alpha_*))\right\}\right) . \tag{3.8}$$

This distribution is of the form of a Normal-Inverse-Wishart with location parameter $\mu(\alpha_*)$

and inverse scale parameter $\Sigma(\alpha_*)$. This is precisely the expected result for the posterior of a normal data model whose conjugate prior distribution is Normal-Inverse-Wishart.

3.2 Gaussian Random Processes

Having addressed the Dynamical Bayesian inference of multivariate Gaussian data it becomes natural to discuss the Dynamical Bayesian inference of data which is distributed according to a Gaussian Random Process (GRP).[9] A GRP may be interpreted as the functional analog of a Gaussian distribution. That is, instead of considering random vectors, one considers random *functions*, and instead of specifying a mean vector and a covariance matrix one specifies a mean *function* and a covariance *kernel*. Let us be more precise:

Suppose the data we are interested in consists of the space of random functions, $\phi : D \to \mathbb{R}$.[10] To specify a GRP on such a sample space one must specify a mean function:

$$\mu : D \to \mathbb{R} \tag{3.9}$$

and a covariance kernel:

$$\Sigma : D \times D \to \mathbb{R} \tag{3.10}$$

Then, the distribution over functions takes the symbolic form:

$$f(\phi \mid \mu, \Sigma) = \mathcal{N} \exp\left(-\frac{1}{2}\int_{D\times D} dxdy\ (\phi(x) - \mu(x))\Sigma^{-1}(x,y)(\phi(y) - \mu(y))\right). \tag{3.11}$$

Here $\Sigma^{-1}(x,y)$ is the inverse of $\Sigma(x,y)$ in the functional sense:

$$\int_D dy\ \Sigma^{-1}(x,y)\Sigma(y,z) = \delta(x-z) \tag{3.12}$$

and the prefactor $\mathcal{N}$ is formally infinite, and can be identified with the partition function (path integral) of the unnormalized GRP.

Taken literally, the distribution (3.11) is difficult to use. It should rather be viewed as a set of instructions for how to interpret the GRP. Formally, a GRP is defined by restricting our attention to a finite partition of the domain D: $P = \{x_1, ..., x_n\} \subset D$. A functional random variable $f : D \to \mathbb{R}$ follows a Gaussian Process with mean $\mu(x)$ and covariance $\Sigma(x,y)$ if, for any such partition, the n-vector, $f_P = (f(x_1), ..., f(x_n))$ in a multivariate Gaussian random variable with mean $\mu = (\mu(x_1), ..., \mu(x_n))$ and covariance $\Sigma = \Sigma(x_i, x_j)$.

In this respect, the study of a GRP is precisely the same as the study of the multivariate Gaussian – we need only restrict our attention to some finite partition of the domain of

[9] For an introduction to GRPs in machine learning, see reference [26].

[10] Notice, this construction can be straightforwardly generalized to functions with values in arbitrary spaces, we consider maps into $\mathbb{R}$ for the sake of clarity.

the functional random variable and then perform Dynamical Bayesian inference over the resulting multivariate normal random variable.

4 Inference in the Ising Model

We now turn to some numerical experiments to test the general framework of dynamical Bayesian updating. Along these lines, we consider the basic physical question: Given a collection of experimental data, how well can an observer reconstruct the underlying model?[11] To make this tractable, we assume that the particular physical model is known, but the couplings are unknown. A tractable example of this sort is the statistical mechanics of the Ising model, as specified by a collection of spins $\sigma = \pm 1$ arranged on a graph. In this setting, the statistical mechanics provides us with a probability distribution over spin configurations $\{\sigma\}$ as specified by the Boltzmann factor:

$$P[\{\sigma\}|J] = \frac{1}{\mathcal{Z}(J)} \exp(-H_{\text{Ising}}[\{\sigma\}|J]), \tag{4.1}$$

where $\mathcal{Z}(J)$ is a normalization constant (i.e., the partition function) introduced to ensure a normalized distribution and H_{Ising} is the Ising model Hamiltonian with coupling constant J:

$$H_{\text{Ising}}[\{\sigma\}|J] = -J \sum_{n.n.} \sigma\sigma'. \tag{4.2}$$

In the above, the sum is over nearest neighbors on the graph. One can generalize this model in various ways, by changing the strength of any given bond in the graph, but for ease of analysis we focus on the simplest non-trivial case as stated here. In this case, each draw from the distribution $P[\{\sigma\}|J]$ is specified by a collection of spins $\{\sigma\}$. We can bin all of these events, as we already explained in section 2, and this specifies a posterior distribution $\pi_{\text{post}}(J;T)$. Using this, we can extract the T dependence of various observables, for example:

$$\langle J^m \rangle = \int dJ\, \pi_{\text{post}}(J;T) J^m. \tag{4.3}$$

We can also introduce the centralized moments:

$$\overline{C}^m = \langle (J - \langle J \rangle)^m \rangle. \tag{4.4}$$

[11] See also [14, 27, 28, 22] for related discussions.

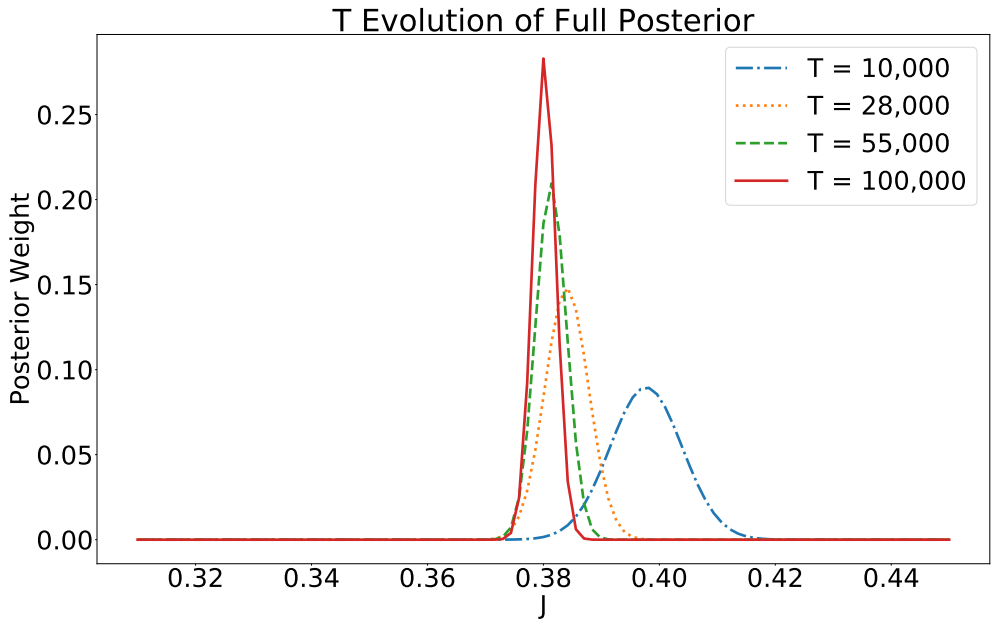


Figure 2: Example of a trial in which the the posterior distribution over couplings is inferred at different update "times" incremented in steps of 900 starting from an initial training at $T = 10,000$. We observe that the central value of the distribution converges to $J_* = 0.38$, and the width of the distribution narrows sharply. The match on higher order moments is displayed in table 3.

4.1 Numerical Experiment: 1D Ising Model

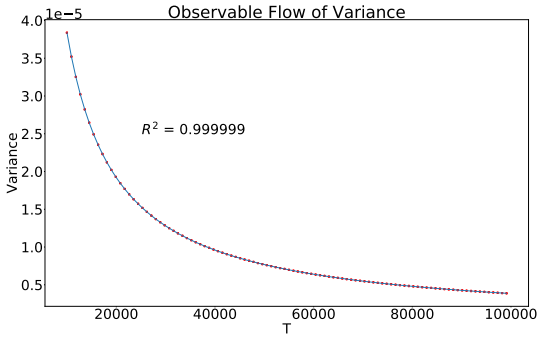
As an explicit example, we now turn to the specific case of the 1D Ising model, i.e., a one-dimensional periodic lattice of evenly spaced spin. The Hamiltonian in this case is:

$$H_{\text{Ising}} = -J \sum_{1 \leq i \leq L} \sigma_i \sigma_{i+1}, \tag{4.5}$$

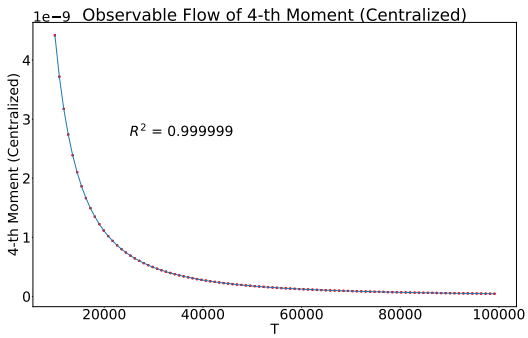
with $\sigma_{L+1} \equiv \sigma_1$. We have an analytic expression for the partition function (see, e.g., [29]), and can also explicitly extract the Fisher information metric:

$$\mathcal{I}(J) = (L-1)\,\text{sech}^2(J). \tag{4.6}$$

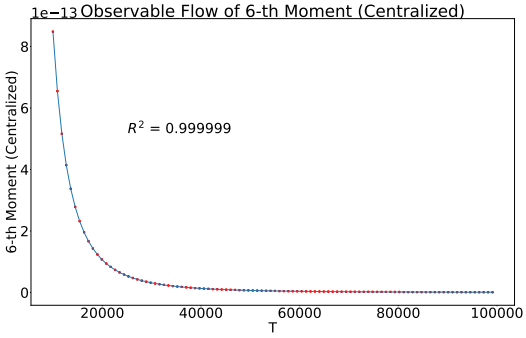
We would like to understand the convergence of the model to the true value of the parameter. Since the main element of our analysis involves adjusting the posterior distribution, it is enough to work with a small number of spins, i.e., $L = 4$. We take a benchmark value of $J_* = 0.38$ (so the Fisher information metric is $\mathcal{I}(J_*) = 2.60533$) and track the dynamical Bayesian updating on the inference of this coupling. For a given trial, we performed



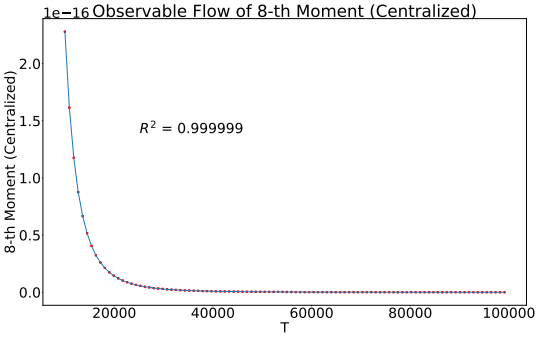
(a) **Variance of Posterior Distribution**



(b) **Fourth Centralized Moment**



(c) **Sixth Centralized Moment**



(d) **Eighth Centralized Moment**

Figure 3: Observable flows for the first four centralized moments $\langle (J - \langle J \rangle)^{2l} \rangle$ for $l = 1, 2, 3, 4$ of the posterior distribution for the Ising Model Experiment. In all cases, we observe a power-law decay which is in close accord with the behavior saturated by the Cramér-Rao bound (see equation 4.7).

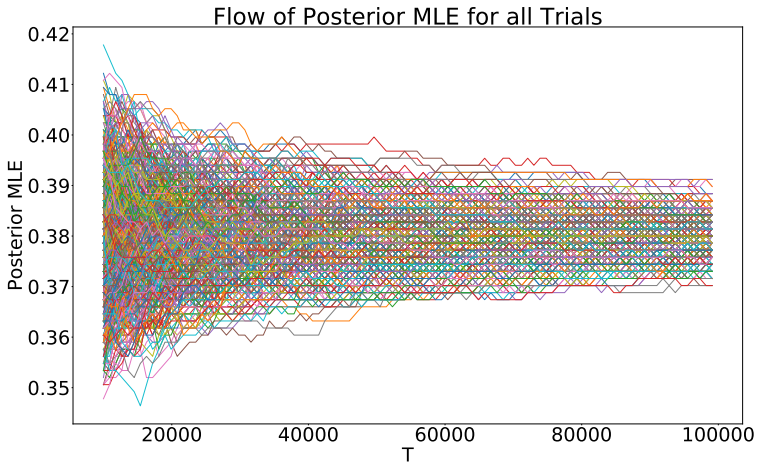


Figure 4: Dynamical Bayesian Trajectories for 1000 Ising Trials.

Moment	C-R Limit	Experiment
$\left\langle (J - \langle J \rangle)^2 \right\rangle$	$0.38/T$	$0.38/T^{0.9997}$
$\left\langle (J - \langle J \rangle)^4 \right\rangle$	$0.44/T^2$	$0.44/T^{1.9996}$
$\left\langle (J - \langle J \rangle)^6 \right\rangle$	$0.84/T^3$	$0.84/T^{2.9993}$
$\left\langle (J - \langle J \rangle)^8 \right\rangle$	$2.28/T^4$	$2.26/T^{3.999}$

Table 3: Comparison of predicted scaling for n-point functions from Dynamical Bayesian Inference in the limit where the Cramér-Rao bound is saturated (see equation 4.7), and the observed scaling from the Ising Model Experiment. We have displayed additional significant figures to exhibit the extent of this match. Observe that in all cases, the experimentally determined power-law is of the form $1/T^{1-\nu}$ for $\nu > 0$, i.e., it respects the lower limit expected from the Cramér-Rao bound.

a Bayesian update to track how well we could infer the value of the coupling constant. In each trial, we sampled from the Boltzmann distribution $100,000$ distinct spin configurations. Starting from the initial prior $J = 0$ (uniform distribution), we performed an initial update using $10,000$ events to get the first estimate for J. We then used the remaining $90,000$ events to obtain a series of sequential updates. The posterior was updated after the inclusion of every additional set of 900 events. This then ran for a total of 1000 time steps.

For each trial we observe some amount of random fluctuation, but after averaging over 1000 trials, we observe strikingly regular behavior, especially in the moments of the coupling J as computed by the posterior distribution (see equation (4.3)). The late T posterior distribution is Gaussian, and can be seen for a sample run at progressively later times in figure (2). The observable flow of the even centralized moments for the update dependent posterior distribution can be seen below. Assuming we saturate the Cramér-Rao bound, we find:

$$\langle (J - \langle J \rangle)^{2l} \rangle = \overline{C}^{2l} = \frac{(2l-1)!!}{(\mathcal{I}_*)^l} T^{-l} \tag{4.7}$$

Where $n!! = \prod_{k=0}^{[\frac{n}{2}]-1} (n - 2k)$. This agrees very well with the numerical experiment, as can be seen in figure (3) and summarized in table (3).

Finally, we note there is some statistical variation present on the space of trajectories for the maximum likelihood estimate (MLE) (see figure 4). This makes manifest that there is statistical variation in any individual inference scheme, but that on aggregate, the paths converge to the maximum likelihood estimate. This observation inspires a path integral interpretation of dynamical Bayesian updating that we leave for future work.

5 Neural Networks and Learning

The Bayesian approach to neural networks was pioneered by Neal in [23]. In what follows we will examine whether the dynamical inference model described in the present work can be applied to neural networks. We will take the viewpoint that a neural network is simply a model whose parameters are given by its weights and biases. Training a neural network using data infers the most likely set of weights given the training set (at least one hopes that this is true). As such one may adopt the view that the training of neural networks is a Bayesian problem of inferring a posterior distribution over the weights given the data available and then one chooses a net with the most likely weights from the posterior distribution. Note that training a network is a stochastic process where the outcome depends on the initialization of weights and the path taken through training.

To apply the reasoning in the paper we will examine how the trained neural network is dependent on the quantity of data used in its training. In particular we will measure how a trained neural network changes as we increment the amount of data used in the training process. We will certainly not be able to follow in a fully quantitative way the calculations in the previous sections because a neural network has far too many parameters (its weights) to carry out the Bayesian analysis explicitly. Instead we will empirically investigate whether the neural network follows a similar qualitative dependence on data as indicated by dynamical Bayesian updating. Insofar as the loss function can be approximated near the final inference in terms of quantities which are quadratic in the underlying θ parameters, we expect a simple power-law behavior as we approach a high level of accuracy. We expect the loss function to exhibit an exponential decaying profile when the inference is only moderately successful. The fact that we empirically observe precisely this sort of behavior provides support for the general picture developed in section 2.

Let us outline the experiment. For a helpful glossary of terms and additional background, see e.g. reference [30]. The basic idea is that we will consider training a neural network using differing sample sizes from the same data set and see how loss depends on the amount of data. (For comparison we will repeat the whole experiment using the MNIST, Fashion-MNIST and CIFAR10 data sets.) The first neural network we use will have a very simple feedforward (FF) architecture. The input layer is a 28×28 layer, corresponding to the MNIST input data. Next is a simple 128 node dense layer followed by the final 10 node output layer with softmax activation. The cost function is taken to be the categorical cross-entropy.

We also consider some experiments involving more sophisticated convolutional neural networks, training on the MNIST data set and the CIFAR10 data set. In the case of the MNIST data set, we consider a convolutional layer with kernel size 2 and filter size 64, followed by max pooling (with pool size 2), followed by a drop out layer (with drop out parameter 0.3) and then another convolution layer, kernel size 2 and filter size 32, then max pooling (with pool size 2), a dropout layer (parameter 0.3), followed by a dense layer with 256 neurons with rectified linear unit (ReLU) activation and a final dropout layer (parameter

0.5) and a final dense layer with 10 outputs and softmax activation.

For the CIFAR10 data we used a convolutional neural network with 3 convolutional layers with respective filter sizes 32, 64, and 128, with kernel size 3 for each layer, a max pooling layer with 3×3 poolsize was included after each convolutional layer. This set of convolution/pooling layers are then followed by a 128 node dense layer with ReLU activation followed by a dropout layer with dropout parameter 0.4 leading on to the final dense layer of 10 outputs with softmax activation.

The main difference between the convolutional neural networks used in the MNIST and CIFAR10 experiments, apart from having the larger input layer for CIFAR10 is the kernel size of the convolutions. In all cases the hyperparameters such as for dropout were untuned. Given that such hyperparameter tuning tends to depend on the specifics of the data being learnt for the purposes of the questions in this paper we did not consider hyperparameter tuning as necessary.

Crucially, we wish to investigate the dependence of the loss on the amount of data and not the amount of training of the network. Usually in training a neural network these two become connected since in any given epoch the amount of training depends on the amont of data. But crucially, neural networks often learn by repeated training using the same data set over many epochs. We are interested in the final state of the neural network after we have completed training.

We wish to keep the amount of training fixed and only compare the loss with different amounts of data used to do the training. (By training, we really mean the attempt to minimize the cost through some form of repeated gradient flow.) To do this we link the number of epochs to the size of the training set we use. We have chosen to train over 4 epochs if the data set is maximal, i.e., 60,000 samples. This is a reasonable choice that produces good accuracy without overfitting. To demonstrate the reasoning behind this, consider training one neural net with N data samples and another with $2N$. One training epoch for the network trained with $2N$ samples will have effectively twice the amount of training as the network with just N samples. Thus to compare the effect of the larger data set as opposed to the amount of training we should train the network that uses the $2N$ data half the number of epochs as the one using the N data set.

We train the networks using the `Adam` optimizer [31] with learning rate set to a standard 0.001. (For the full 60,000 samples and 4 epochs, this gives a healthy sparse categorical accuracy of around 0.97 for MNIST with the simple neural net.) After the network has been trained using the training set of N samples it is tested on the full test set of 10,000 samples.

In what follows, we begin with a large sample size (e.g., 3,000) and then examine the loss after the training is complete as a function of the size of the training data set. We will then increase the training set size N by some increment δN where typically we take δN to be around 500 and then repeat this until we reach a final data set an order of magnitude bigger e.g., 30,000 data points. We then fit the resulting curve to the power-law behavior

as expected from the dynamical Bayesian updating analysis. We find that for MNIST with the convolutional net the power-law is close to one but for Fashion-MNIST where the loss is higher, the power-law is of the form $1/T^{1+\nu}$ for $\nu > 0$. This is compatible with the contribution from hidden variables for Bayesian flows given in section 2.

We then repeat this with the CIFAR10 data set and the even more involved convolutional network where we find the exponential decay is a better fit than power-law indicating that the network has untrained parameters as in the hidden variable example discussed before.

The reader familiar with Stochastic and Batch gradient descent may feel that we are just doing the same thing in this experiment and these are just the traditional learning curves. This is *not* the case since we train for multiple epochs and the curves measure only the loss as a function of *total* data used in the training.

All the code is available to view in a Google Colab:

https://colab.research.google.com/drive/1zNxHj7qCoE1-WzawqRTbaa9QhFWpQZqr?usp=sharing

5.1 Results

Training neural nets is notoriously stochastic. To take this into account we actually perform multiple trials of each experiment (with different initial conditions in each case). We plot loss against T and then fit to a power-law in each case. Performing multiple trials, we also extract the mean and variance for these fitting parameters, in particular the exponent appearing in the power-law fit. We also quote the root mean variance as an indicator of how robust the results are. For 10 trials, the root variance of the power-law was between 8% and 10% depending on the data set in question.

We display here some representative examples of this analysis, as in figure 5 for the experiments with a feedforward neural network trained on the MNIST and Fashion-MNIST data sets, as well as figure 6 for the convolutional neural network experiments trained on the MNIST and CIFAR10 data sets. In these plots we display the loss function (i.e., the categorical cross-entropy) on the vertical axis and the number of data samples used for training on the horizontal axis. In each case, we also display the corresponding fit for these particular examples, and the results are collected in table 4. As discussed above, an important aspect of these individual fits is that the actual parameters deviate from trial to trial; and so we also give the central values of the fitting parameters and their 1σ deviations. The mean values of the fitting parameters are displayed in table 5.

5.1.1 A Simple Feedforward Network

The first curve is with 3000 initial samples used as training data and then incremented in steps of 500. The fit to a power-law has an R^2 value of 0.98, showing a very strong fit to the data with a power-law behavior $\sim 1/T^{0.74}$. We then repeated the experiment with the

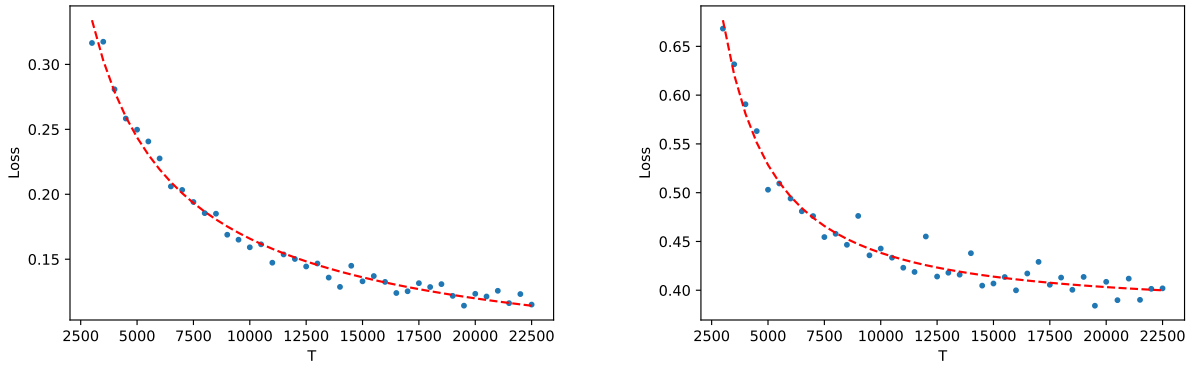
Fashion-MNIST data set, which had an R^2 fit to power-law of 0.96 with power-law behavior $\sim 1/T^{1.36}$. See figure 5 for the plots of the loss function and the fitting curves, and table 4 for a summary of the fitting functions for these particular examples. Table 5 also reports the mean and 1σ uncertainties for the power law fitting parameters.

Dataset	Network	Function Type	Loss(T)	R^2
MNIST	FF	Power Law	$103T^{-0.74} + 0.05$	0.98
Fashion-MNIST	FF	Power Law	$16033T^{-1.36} + 0.41$	0.96
MNIST	CNN	Power Law	$241T^{-1.03} + 0.03$	0.99
CIFAR10	CNN	Exponential	$4.1e^{-0.000113T} + 0.60$	0.96

Table 4: Fitting functions categorical cross-entropy loss as a function of T for the example trial runs displayed in figures 5 and 6 for various data sets and neural network architectures (FF refers to feedforward and CNN refers to convolutional neural network). In most cases, we observe a rather good fit to a power-law behavior when the accuracy of inference is also high. For situations where there is a degraded performance as in the CIFAR10 data set, we instead observe a better fit to an exponential decay function. Note also that in some cases, we obtain a power-law with exponent above or below -1. Including hidden variables in the Bayesian flow equations can accommodate both phenomena. Comparing over multiple trial runs, we observe some variance in individual fits. We collect the central values and variance of the decay law parameters for the different data sets in table 5.

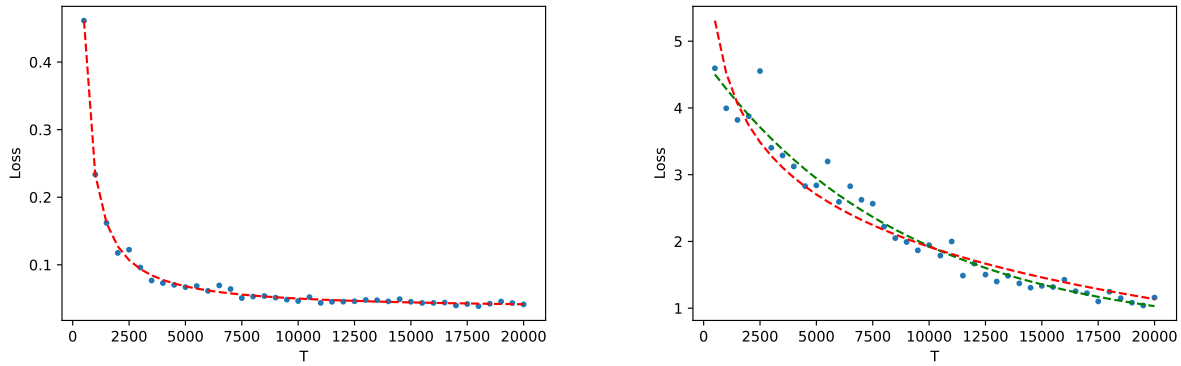
Dataset	Network	Loss(T)	b
MNIST	FF	$aT^{-b} + c$	0.74 ± 0.06
Fashion-MNIST	FF	$aT^{-b} + c$	1.32 ± 0.12
MNIST	CNN	$aT^{-b} + c$	1.01 ± 0.06
CIFAR10	CNN	$ae^{-bT} + c$	$1.6 \times 10^{-4} \pm 2.4 \times 10^{-5}$

Table 5: Central values of the fitting parameters averaged over 10 different trials. Uncertainties are quoted at the 1σ level. For the MNIST and Fashion-MNIST data sets, these fit well to power-law behavior of the form $aT^{-b} + c$. For the CIFAR10 where the overall accuracy was lower, we instead find a better fit to an exponential decay law $ae^{-bT} + c$. While there is some variance in the overall value of these fitting parameters, each individual trial fits well to the expectations of the dynamical Bayesian evolution equations. The experiments thus reveal the sensitivity to initial conditions in the training of the neural networks.



(a) **MNIST Trial** (b) **Fashion-MNIST Trial**

Figure 5: Categorical cross-entropy loss as a function of T in a simple feedforward neural network with varying amounts of trial data. Here, we display the results for a single complete run in the case of the MNIST and Fashion-MNIST data sets. In nearly all examples, we observe a highly accurate fit to a power-law behavior, with respective power-laws $103T^{-0.74}+0.05$ (R^2 of 0.98) and $16033T^{-1.36}+0.38$ (R^2 of 0.96) for the MNIST and Fashion-MNIST and examples. See also table 4. We collect the central values and variance of the decay law parameters for the different data sets in table 5.



(a) **MNIST Trial** (b) **CIFAR10 Trial**

Figure 6: Categorical cross-entropy loss as a function of T in a convolutional neural network with varying amounts of trial data. Here, we display the results for a single complete run in the case of the MNIST and CIFAR10 data sets. In this case, we obtain a good fit to a power-law decay in the case of the MNIST data set, while in the case of the CIFAR10 data set, the lower accuracy is better fit by an exponential function (red curve) as opposed to a power-law (green curve). See also table 4. We collect the central values and variance of the decay law parameters for the different data sets in table 5.

5.1.2 Convolutional Neural Networks

We also performed a similar set of experiments using the convolutional neural networks as described above. We again repeated the experiments 10 times so as to take into account the stochastic nature of the training process and take mean values. We took the initial data size to be 500 and increment size 500 as before.

In the case of the MNIST data set, we find the mean power-law fit has $R^2 = 0.99$ and mean decay coefficient 1.01. (The root of the variance of the decay constant was 0.06). This network had a very low final loss 0.99 and captured well the properties of the full data set. It is interesting that when this happened, the exponent of the power-law approached the value for the Cramér-Rao bounded flow. Figure 6a displays one such trial. Averaging over all the trials, we also determined the exponent for the power-law decay, the results are displayed in table 5.

Finally, for the CIFAR10 data set with the three layer convolutional network, we took an initial data size of 500 and increment size of 500. We repeated the experiment 10 times, and in each trial we performed a best fit to the loss function, and in general we observed the data was better fit by an exponential rather than a power-law. In figure 6b we present the data from one such trial, where the power-law fit (green curve) gave an R^2 of 0.92, while the exponential fit (red curve) gave an R^2 of 0.96. Averaging over all the trials, we also determined the decay constant for the exponential fit, the results are displayed in table 5. Note that although it is better fit by an exponential decay, the actual decay constant is quite small.

6 Conclusions and Discussion

In this note we have presented an interpretation of Bayesian updating in terms of a dynamical system. In a given model of the world, each new piece of evidence provides us with an improved understanding of the underlying system, thus generating an effective flow in the space of parameters which is saturated by a simple $1/T$ power-law, the analog of a "unitarity bound" in conformal field theory. This can be exceeded when additional information flows in via hidden variables. We have shown how this works in practice both in an analytic treatment of Gaussian distributions and Gaussian Random Processes, and have also performed a number of numerical experiments, including inference on the value of the coupling constants in the 1D Ising Model, and in training of neural networks. We find it remarkable that simple Bayesian considerations accurately capture the asymptotic behavior of so many phenomena.

The appearance of a $1/T^b$ power-law scaling for learning in neural networks is of course quite suggestive. In the context of statistical field theory, the onset of such a scaling law behavior is usually a clear indication of a phase transition. We have also seen that inference in the presence of hidden variables provides a simple qualitative explanation for some of this behavior. It would be very interesting to develop a more fundamental explanation.

A unifying thread of this work has centered on giving a physical interpretation of Bayesian updating. This equation shares a number of common features with the related question of renormalization group flow in a quantum field theory.[12] But whereas renormalization is usually interpreted as a flow from the ultraviolet to the infrared wherein we *lose* information about microscopic physics, the Bayesian updating procedure does precisely the opposite: we are *gaining* information as we evolve along a flow. We have also seen that new evidence in Bayesian updating can either perturb a trajectory, or not impact it very much, and this again parallels similar notions of relevant and irrelevant perturbations. We have also taken some preliminary steps in developing a path integral interpretation of Bayesian flows in Appendix A. This in turn suggests that there should be a direct analog of Polchinski's exact renormalization group equation which would be exciting to develop.

One of the original motivations of this work was to better understand the sense in which the structure of quantum gravity might emerge from an observer performing local measurements in their immediate vicinity (see, e.g., [14, 27, 39–41, 28] for related discussions). From this perspective, each new piece of data corresponds to this local observer making larger excursions in the spacetime, as well as the parameters of the theory. This is particularly well-motivated in the specific context of the AdS/CFT correspondence [42], where the radial direction of the bulk anti-de Sitter space serves as a renormalization scale in the CFT with a cutoff. Given that we have a flow equation, and that it shares many formal similarities to an RG equation, this suggests a natural starting point for directly visualizing radial evolution in terms of such an inference procedure.

At a more practical level, it would also be interesting to test how well an observer can infer such "spacetime locality". Along these lines, there is a natural class of numerical experiments involving a mild generalization of our Ising model analysis in which we continue to draw from the same Ising model with only nearest neighbor interactions, but in which the model involves additional contributions coupling neighbors which might be very far away.

Acknowledgments

We thank J.G Bernstein, R. Fowler and R.A. Yang for helpful discussions and many of the members of the "Physics meets ML" group. DSB thanks Pierre Andurand for his generous donation supporting this work. The work of JJH is supported in part by the DOE (HEP) Award DE-SC0013528, and a generous donation by P. Kumar, as well as a generous donation by R.A. Yang and Google.

[12] The notion of "renormalization" has been discussed in [32, 33, 30], though we should point out that in a quantum field theory, the utility of organizing by scale has a great deal to do with the fact that there is a clear notion of locality, something which is definitely *not* present in many inference problems! For additional discussion on connections between statistical / quantum field theory and machine learning, see, e.g., [34–38].

A Interpreting Dynamical Bayesian Updating

In the main text of our paper we implemented an approach to dynamical Bayesian Inference in which the posterior distribution is probed by observing the scaling of its various centralized moments as a function of update "time". In this appendix we would like to draw attention to an alternative strategy for studying Dynamical Bayesian inference in which one solves the flow equation for the complete posterior, (2.17), directly. As was the case in the main text, we will find it more natural to consider our update in terms of a "time" parameter $T = N\tau$. One can think of T as corresponding to the number of data point utilized in the Bayesian Inference model up to a given iteration. In these terms we can write the T-dependent posterior distribution which solves the flow equation as:

$$\pi(\theta; T) = \exp(-TD_{KL}(\alpha_* \parallel \theta)) \exp\left(\int_0^T dT' D_{KL}(\alpha_* \parallel \alpha(T'))\right) \tag{A.1}$$

We will see that the structure of this solutions calls to mind many of the common approaches utilized in the analysis of physical systems, especially statistical ensembles.

To begin, observe that $\pi(\theta; T)$ is a *normalized* probability density function for each value of T:

$$1 = \int d\theta \pi(\theta; T) \ \ \forall\, T \tag{A.2}$$

Performing the integration explicitly, we notice that only the first factor in (A.1) depends on θ. Thus, we find:

$$1 = \exp(\int_0^T dT'\, D_{KL}(\alpha_* \parallel \alpha(T'))) \int d\theta \exp(-TD_{KL}(\alpha_* \parallel \theta)) \tag{A.3}$$

It is natural to define the integral appearing in (A.3) as the Partition Function of an unnormalized density:

$$\mathcal{Z}(T) := \int d\theta e^{-TD_{KL}(\alpha_* \| \theta)} \tag{A.4}$$

This gives the Dynamical Bayesian Posterior the complexion of a Bolzmann weight with "energy" $D_{KL}(\alpha_* \parallel \theta)$. It also suggests that we should regard T as an inverse temperature, or imaginary time parameter as is typical in statistical field theory contexts.

Referring back to (A.3), we conclude that the role of the θ independent term in the posterior density (A.1) is explicitly to maintain the normalization of the posterior density at all T. Indeed, we can write:

$$\mathcal{Z}(T) = \exp\left(-\int_0^T dT'\, D_{KL}(\alpha_* \parallel \alpha(T'))\right) \tag{A.5}$$

Or, equivalently:

$$-\ln(\mathcal{Z}(T)) = \int_0^T dT'\, D_{KL}(\alpha_* \parallel \alpha(T')) \tag{A.6}$$

This equation relates the KL-Divergence of the T-dependent parameter estimate $\alpha(T)$ with the cumulant generating functional of the posterior distribution. Taking the first derivative of this equation with respect to T we find:

$$D_{KL}(\alpha_* \parallel \alpha(T)) = \langle D_{KL}(\alpha_* \parallel \theta)\rangle_{\pi(\theta;T)} \tag{A.7}$$

Which is precisely equation (2.23)! More generally, notice that:

$$-\left(\frac{d}{dT}\right)^n \ln(\mathcal{Z}(T)) = (-1)^{n+1}\mathcal{C}^n_{\pi(\theta;T)}\left(D_{KL}(\alpha_* \parallel \theta)\right) \tag{A.8}$$

Where here $\mathcal{C}^n_{\pi(\theta;T)}(Q(\theta))$ denotes the n^{th} cumulant of $Q(\theta)$ with respect to the time T posterior distribution, $\pi(\theta;T)$. We therefore obtain the expression:

$$\left(\frac{d}{dT}\right)^{n-1} D_{KL}(\alpha_* \parallel \alpha(T)) = (-1)^{n+1}\mathcal{C}^n_{\pi(\theta;T)}\left(D_{KL}(\alpha_* \parallel \theta)\right) \tag{A.9}$$

One may interpret this equation as saying that all of the relevant connected correlation functions associated with the statistical inference are encoded in the path $\alpha(T)$. Once $\alpha(T)$ is known these cumulants can extracted through equation (A.9).

References

[1] T. Bayes, Rev., "An essay toward solving a problem in the doctrine of chances," *Phil. Trans. Roy. Soc. Lond.* **53** (1764) 370–418.

[2] M. Gell-Mann and F. E. Low, "Quantum electrodynamics at small distances," *Phys. Rev.* **95** (1954) 1300–1312.

[3] L. P. Kadanoff, "Scaling laws for Ising models near T_c," *Physics Physique Fizika* **2** (1966) 263–272.

[4] K. G. Wilson, "Renormalization group and critical phenomena. 1. Renormalization group and the Kadanoff scaling picture," *Phys. Rev. B* **4** (1971) 3174–3183.

[5] K. G. Wilson, "Renormalization group and critical phenomena. 2. Phase space cell analysis of critical behavior," *Phys. Rev. B* **4** (1971) 3184–3205.

[6] J. Polchinski, "Renormalization and Effective Lagrangians," *Nucl. Phys. B* **231** (1984) 269–295.

[7] V. Balasubramanian, "Statistical inference, occam's razor and statistical mechanics on the space of probability distributions," arXiv:cond-mat/9601030.

[8] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics* **22** no. 1, (1951) 79 – 86.

[9] C. R. Rao, "Information and accuracy attainable in the estimation of statistical parameters"," *Bulletin of the Calcutta Mathematical Society* **37** (1945) 81–91.

[10] S. Amari, *Differential-Geometrical Methods in Statistics*. Lecture Notes in Statistics. Springer-Verlag, 1985.

[11] W. Bialek, C. G. Callan, Jr., and S. P. Strong, "Field theories for learning probability distributions," *Phys. Rev. Lett.* **77** (1996) 4693–4697, arXiv:cond-mat/9607180.

[12] M. Blau, K. S. Narain, and G. Thompson, "Instantons, the information metric, and the AdS / CFT correspondence," arXiv:hep-th/0108122.

[13] U. Miyamoto and S. Yahikozawa, "Information metric from a linear sigma model," *Phys. Rev. E* **85** (2012) 051133, arXiv:1205.3211 [math-ph].

[14] J. J. Heckman, "Statistical Inference and String Theory," *Int. J. Mod. Phys. A* **30** no. 26, (2015) 1550160, arXiv:1305.3621 [hep-th].

[15] J. J. Heckman, J. G. Bernstein, and B. Vigoda, "MCMC with Strings and Branes: The Suburban Algorithm (Extended Version)," *Int. J. Mod. Phys. A* **32** no. 22, (2017) 1750133, arXiv:1605.05334 [physics.comp-ph].

[16] J. J. Heckman, J. G. Bernstein, and B. Vigoda, "MCMC with Strings and Branes: The Suburban Algorithm," arXiv:1605.06122 [stat.CO].

[17] T. Clingman, J. Murugan, and J. P. Shock, "Probability Density Functions from the Fisher Information Metric," arXiv:1504.03184 [cs.IT].

[18] E. Malek, J. Murugan, and J. P. Shock, "The Information Metric on the moduli space of instantons with global symmetries," *Phys. Lett. B* **753** (2016) 660–663, arXiv:1507.08894 [hep-th].

[19] H. Dimov, I. N. Iliev, M. Radomirov, R. C. Rashkov, and T. Vetsov, "Holographic Fisher information metric in Schrödinger spacetime," *Eur. Phys. J. Plus* **136** no. 11, (2021) 1128, arXiv:2009.01123 [hep-th].

[20] J. Erdmenger, K. T. Grosvenor, and R. Jefferson, "Information geometry in quantum field theory: lessons from simple examples," *SciPost Phys.* **8** no. 5, (2020) 073, arXiv:2001.02683 [hep-th].

[21] A. Tsuchiya and K. Yamashiro, "A geometrical representation of the quantum information metric in the gauge/gravity correspondence," *Phys. Lett. B* **824** (2022) 136830, arXiv:2110.13429 [hep-th].

[22] R. Fowler and J. J. Heckman, "Misanthropic Entropy and Renormalization as a Communication Channel," arXiv:2108.02772 [hep-th].

[23] R. Neal, *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics. Springer-Verlag, 1996.

[24] P. Mehta, M. Bukov, C.-H. Wang, A. G. R. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, "A high-bias, low-variance introduction to Machine Learning for physicists," *Phys. Rept.* **810** (2019) 1–124, arXiv:1803.08823 [physics.comp-ph].

[25] G. Mack, "All unitary ray representations of the conformal group SU(2,2) with positive energy," *Commun. Math. Phys.* **55** (1977) 1.

[26] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[27] V. Balasubramanian, J. J. Heckman, and A. Maloney, "Relative Entropy and Proximity of Quantum Field Theories," *JHEP* **05** (2015) 104, arXiv:1410.6809 [hep-th].

[28] V. Balasubramanian, J. J. Heckman, E. Lipeles, and A. P. Turner, "Statistical Coupling Constants from Hidden Sector Entanglement," *Phys. Rev. D* **103** no. 6, (2021) 066024, arXiv:2012.09182 [hep-th].

[29] R. K. Pathria, *Statistical Mechanics*. Butterworth-Heinemann, 1996.

[30] D. A. Roberts, S. Yaida, and B. Hanin, "The Principles of Deep Learning Theory," arXiv:2106.10165 [cs.LG].

[31] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv e-prints* (Dec., 2014) arXiv:1412.6980, arXiv:1412.6980 [cs.LG].

[32] J. Halverson, A. Maiti, and K. Stoner, "Neural Networks and Quantum Field Theory," *Mach. Learn. Sci. Tech.* **2** no. 3, (2021) 035002, arXiv:2008.08601 [cs.LG].

[33] J. Halverson, "Building Quantum Field Theories Out of Neurons," arXiv:2112.04527 [hep-th].

[34] D. Bachtis, G. Aarts, and B. Lucini, "Quantum field-theoretic machine learning," *Phys. Rev. D* **103** no. 7, (2021) 074510, arXiv:2102.09449 [hep-lat].

[35] D. Bachtis, G. Aarts, and B. Lucini, "Quantum field theories, Markov random fields and machine learning," in *32nd IUPAP Conference on Computational Physics*. 10, 2021. arXiv:2110.10928 [cs.LG].

[36] G. Aarts, D. Bachtis, and B. Lucini, "Interpreting machine learning functions as physical observables," in *38th International Symposium on Lattice Field Theory*. 9, 2021. arXiv:2109.08497 [hep-lat].

[37] D. Bachtis, G. Aarts, and B. Lucini, "Machine learning with quantum field theories," in *38th International Symposium on Lattice Field Theory*. 9, 2021. arXiv:2109.07730 [cs.LG].

[38] D. Bachtis, G. Aarts, F. Di Renzo, and B. Lucini, "Inverse Renormalization Group in Quantum Field Theory," *Phys. Rev. Lett.* **128** no. 8, (2022) 081603, arXiv:2107.00466 [hep-lat].

[39] K. Hashimoto, S. Sugishita, A. Tanaka, and A. Tomiya, "Deep Learning and Holographic QCD," *Phys. Rev. D* **98** no. 10, (2018) 106014, arXiv:1809.10536 [hep-th].

[40] K. Hashimoto, "AdS/CFT correspondence as a deep Boltzmann machine," *Phys. Rev. D* **99** no. 10, (2019) 106017, arXiv:1903.04951 [hep-th].

[41] Y. Gal, V. Jejjala, D. K. Mayorga Pena, and C. Mishra, "Baryons from Mesons: A Machine Learning Perspective," arXiv:2003.10445 [hep-ph].

[42] J. M. Maldacena, "The Large N limit of superconformal field theories and supergravity," *Adv. Theor. Math. Phys.* **2** (1998) 231–252, arXiv:hep-th/9711200.