Ambiguous model learning made unambiguous with 1/f priors

G. S. Atwal
Department of Physics
Princeton University
Princeton, NJ 08544
gatwal@princeton.edu

William Bialek
Department of Physics
Princeton University
Princeton, NJ 08544
wbialek@princeton.edu

Abstract

What happens to the optimal interpretation of noisy data when there exists more than one equally plausible interpretation of the data? In a Bayesian model-learning framework the answer depends on the prior expectations of the dynamics of the model parameter that is to be inferred from the data. Local time constraints on the priors are insufficient to pick one interpretation over another. On the other hand, nonlocal time constraints, induced by a $1/f$ noise spectrum of the priors, is shown to permit learning of a specific model parameter even when there are infinitely many equally plausible interpretations of the data. This transition is inferred by a remarkable mapping of the model estimation problem to a dissipative physical system, allowing the use of powerful statistical mechanical methods to uncover the transition from indeterminate to determinate model learning.

1 Introduction

The estimation of a model underlying the production of noisy data becomes highly nontrivial when there exists more than one equally plausible model that could be responsible for the output data. The viewing of ambiguous figures, such as the Necker cube [1], is a classical problem of this type in the field of visual psychology. Pitch perception when hearing a number of different harmonics is another example of ambiguous perception [2].

Previous studies [3] have reduced the problem of optimal interpretation of an ambiguous stimulus to the problem of estimating a single variable which may vary in time $\alpha(t)$, given a time sequence of noisy data. Enforcing a prior belief that the local dynamics $\alpha(t)$ should not vary too rapidly embodies the observer's knowledge that rapid variations in $\alpha(t)$ are unlikely in the natural world or in a given experiment. Such a prior prevents overfitting the model estimate to the data as it arrives. The statistically optimal interpretation of the data was then found to consist of $\alpha(t)$ hopping randomly from one possible interpretation to another. The rate of random switching between interpretations was found to be controlled not by the noise level (e.g. in the neural hardware), as previously thought, but rather by the observer's prior hypotheses. This hopping persists indefinitely despite the fact that the probability distribution of the incoming data remains the same. In such cases it is impossible to learn a specific model parameter.

In this paper we introduce another prior over the dynamics of $\alpha(t)$. We assume that fluctuations in $\alpha(t)$ have a $1/f$ spectrum, as observed ubiquitously in nature. Such a prior is shown to induce nonlocal time constraints on the trajectories of $\alpha(t)$ and, unlike the local constraints, can result in specific model learning in the case of ambiguous models. The fact that $1/f$ priors can induce unambiguous model learning is the central result of this work.

The analyses of the long-time dynamics with nonlocal priors is permitted by a surprising and remarkable mapping to a dissipative quantum system. This mapping not only guides our intuition of the optimal trajectories of $\alpha(t)$ but also permits the usage of powerful statistical mechanical techniques. In particular, the renormalization group (RG) can be employed to uncover the conditions in which there is a transition from non-specific model learning to specific model learning.

2 Formalism

Suppose that we are given a series of N measurements $\{x_t\}$ at discrete times t. Then Bayes rule gives us the conditional probability of $\{\alpha_t\}$ giving rise to those data

$$P[\{\alpha_t\}|\{x_t\}] = \frac{P[\{x_t\}|\{\alpha_t\}]\, P[\{\alpha_t\}]}{P[\{x_t\}]}, \tag{1}$$

where the probability of making the observations $\{x_i\}$ is given by summing up all the possible models that may give rise to them,

$$P(\{x_t\}) = \int d\alpha\, P[\{x_t\}|\{\alpha_t\}]P[\{\alpha_t\}]. \tag{2}$$

We further assume conditional independence of signals,

$$P[\{x_t\}|\{\alpha_t\}] = P[x_1 x_2 ... x_N|\{\alpha_t\}] = \prod_{t=1}^{N} P[x_t|\alpha_t]. \tag{3}$$

A natural step is then to consider how close our estimate of the model $\alpha(t)$ lies to the true underlying model $\overline{\alpha}(t)$, which we take to be stationary $\overline{\alpha}(t) = \overline{\alpha}$. We can think of these probability distributions as Boltzmann distributions in which some effective potential acts to hold α close to $\bar{\alpha}$; thus we envision an energy landscape in the α space with a minimum at $\bar{\alpha}$.

A more interesting, and generalized, question arises when we consider the global properties of the extended energy landscape. In particular there may be $M > 1$ equally plausible interpretations consistent with the input data[1] in which case there exist degenerate minima at $\overline{\alpha}_m$ ($m = 1, 2....M$),

$$P[x_t|\overline{\alpha}_1] = P[x_t|\overline{\alpha}_2] = ... = P[x_t|\overline{\alpha}_M]. \tag{4}$$

Therefore we may write Eq. (3) as

$$P[\{x_t\}|\{\alpha_t\}] = \prod_{t=1}^{N} \left(\prod_{m=1}^{M} P[x_t|\overline{\alpha}_m]^{1/M} \right) \exp\left[\frac{1}{M} \sum_{m=1}^{M} \sum_{t=1}^{N} \ln \frac{P[x_t|\alpha_t]}{P[x_t|\overline{\alpha}_m]} \right]. \tag{5}$$

On average, the term in square brackets is related to the Kullback-Leibler divergences between distributions conditional on $\alpha(t)$ and distributions conditional on the true $\bar{\alpha}$. If the

[1] Of course it may be the case that some interpretations may be more plausible than others, resulting in a non uniform probability distribution over possible models. In this paper we illustrate the case where all interpretations are equally likely, $P[\overline{\alpha}_m] = 1/M$.

time variation of α is slow, we effectively collect many samples of x before α changes, and it makes sense to replace the sum over samples by its average:

$$\begin{aligned}\lim_{N\to\infty}\sum_{m=1}^{M}\sum_{t=1}^{N}\ln\frac{P[x_t|\alpha_t]}{P[x_t|\overline{\alpha}_m]} &\approx \frac{1}{\tau_0}\sum_{m=1}^{M}\int dt\int dx P[x(t)|\overline{\alpha}_m]\ln\frac{P[x(t)|\alpha(t)]}{P[x|\overline{\alpha}_m]}, \\ &\equiv -\frac{1}{\tau_0}\sum_{m=1}^{M}\int dt D_{KL}[\overline{\alpha}_m||\alpha(t)]. \end{aligned} \qquad (6)$$

where τ_0 is the average time between observations, and we take the continuum limit.

2.1 Priors

We need to have some prior hypotheses about how $\alpha(t)$ can vary in time, serving as our prior probability distribution $P[\alpha(t)]$. We introduce two different types of priors characterized by whether they constrain the local or nonlocal time dynamics,

$$P[\alpha(t)] = P_{\text{local}}[\alpha(t)]P_{\text{nonlocal}}[\alpha(t)]. \qquad (7)$$

To summarize our prior expectation that the local dynamics of $\alpha(t)$ vary slowly, we assume that the time derivative of $\alpha(t)$ is chosen independently at each instant of time from a Gaussian distribution,

$$P_{\text{local}}[\alpha(t)] \propto \exp\left[-\frac{1}{4D}\int dt\left(\frac{\partial\alpha}{\partial t}\right)^2\right]. \qquad (8)$$

Note that this distribution corresponds to random walk with effective diffusion constant D.

Motivated by the ubiquitous occurrence of $1/f$ fluctuations in nature we chose to encapsulate the nonlocal dynamics by a Gaussian distribution with a $1/f$ power spectrum of noise, conveniently expressed in Fourier coordinates ω as

$$P_{\text{nonlocal}}[\alpha(t)] \propto \exp\left[-\frac{1}{2}\int\frac{d\omega}{2\pi}\frac{|\alpha(\omega)|^2}{S(\omega)}\right], \qquad (9)$$

where the spectral noise function takes the form

$$S(\omega) = \frac{1}{\eta|\omega|}. \qquad (10)$$

Note that the spectrum must be even in ω since for any stationary process $S(\omega) = S(-\omega)$. The parameter η determines the strength of *a priori* belief in nonlocal dynamics, or as we will see later, it can be equivalently viewed as a frictional constant determining the dissipation of the time trajectories of $\alpha(t)$. In the time-domain Eq. (9) becomes

$$P_{\text{nonlocal}}[\alpha(t)] \propto \exp\left[-\frac{\eta}{4\pi}\int dtdt'\left(\frac{\alpha(t)-\alpha(t')}{t-t'}\right)^2\right]. \qquad (11)$$

Combining Eq. (8) and Eq. (11) we then obtain the total prior expectation of the probability distribution over the time-dependence of the model parameter $\alpha(t)$

$$P[\alpha(t)] \propto \exp\left[-\frac{1}{4D}\int dt\left(\frac{\partial\alpha}{\partial t}\right)^2 - \frac{\eta}{4\pi}\int dtdt'\left(\frac{\alpha(t)-\alpha(t')}{t-t'}\right)^2\right]. \qquad (12)$$

Taken together, the local and non-local terms describe fluctuations in α which are $1/f$ up to a cutoff frequency, $\omega_c \sim D\eta$. Returning to the Bayesian conditional probability Eq. (1) we then obtain a path-integral expression

$$P[\alpha(t)|\{x_i\}] \propto \exp(-S[\alpha(t)]), \qquad (13)$$

where the action $S[\alpha(t)]$ is given by

$$S[\alpha(t)] = \int dt \left[\frac{1}{4D} \left(\frac{\partial \alpha}{\partial t} \right)^2 + \eta \int \frac{dt'}{4\pi} \left(\frac{\alpha(t) - \alpha(t')}{t - t'} \right)^2 + V_{\text{eff}}[\alpha(t)] \right], \quad (14)$$

$$V_{\text{eff}}[\alpha(t)] = \frac{1}{\tau_0 M} \sum_{m=1}^{M} D_{KL}[\overline{\alpha}_m || \alpha(t)]. \quad (15)$$

This is equivalent to the imaginary time path-integral for a quantum mechanical particle [4] of mass $1/2D$, with coordinates given by $\alpha(t)$, moving in an effective potential $V_{\text{eff}}[\alpha(t)]$ and subject to (linear) frictional forces with a damping constant η. This mapping provides an extremely useful guide to our intuition for the probable trajectories of $\alpha(t)$. Just as in the analyses of particle dynamics in dissipative quantum mechanics [4] we anticipate that the time-course of $\alpha(t)$ may exhibit qualitatively different types of behavior depending on the strength of the non-local terms. In addition, the equivalence to a physical system permits exploitation of powerful techniques developed in the study of quantum mechanical systems with infinite degrees of freedom.

In the following we consider the cases of $m = 1$ and $m = 2$ and use the RG transformations to consider localization-delocalization transitions.

2.2 M=1 : One true interpretation of data

Now if $\alpha(t)$ differs from $\overline{\alpha}$ by a small $\Delta\alpha(t)$ we can Taylor expand the Kullback-Leibler divergence to give a quadratic distance measure

$$D_{KL}(\overline{\alpha}||\alpha) = \frac{1}{2} F[\overline{\alpha}(t)] \Delta\alpha(t)^2 + O(\Delta\alpha^3), \quad (16)$$

where the metric is the Fisher information

$$F[\alpha(t)] = \int dx \frac{1}{P[x|\alpha(t)]} \left(\frac{\partial P[x|\alpha(t)]}{\partial \alpha(t)} \right)^2 . \quad (17)$$

Thus, close to the true parameter $\overline{\alpha}$ the potential energy term in Eq. (14) is simply a harmonic oscillator with stiffness given by the Fisher information. Guided by the mapping to a dissipative quantum mechanical system we expect that if the initial distribution of α already happens to be closely centered around the correct value then the most likely trajectory will be simply to move closer to the minima of the potential energy at $\overline{\alpha}_1$.

The important point to note is that had we chosen just the local constraints on our priors Eq. (8) then the trajectory of $\alpha(t)$ would persistently fluctuate around $\overline{\alpha}_1$, representing a trade-off between avoiding overfitting the data and inertia of our estimate. In the quantum mechanical picture this corresponds to the zero point fluctuations around the minima. Adding the dissipative term reduces the fluctuations around $\overline{\alpha}_1$ by an amount monotonically dependent on η, thus improving on the optimal estimate.

A RG treatment of the single-well problem, within the harmonic approximation, renormalizes the Fisher information such that the curvature of the potential well increases for all values of the η, and thus the fixed point of the dynamics is simply the convergence of $\alpha(t)$ to reduced fluctuations around the true parameter $\overline{\alpha}_1$. We explicitly carry out the RG calculation in the more interesting case where we have two global minima in the next section.

2.3 M=2 : Two equally possible interpretations of the data

In the case of two equally viable interpretations of the data, the potential energy term becomes that of a double-well potential with degenerate minima at $\overline{\alpha}_1$ and $\overline{\alpha}_2$ and energy barrier h

$$h = \frac{1}{2\tau_0} \left(D_{KL}(\overline{\alpha}_1 || (\overline{\alpha}_1 + \overline{\alpha}_2)/2) + D_{KL}(\overline{\alpha}_2 || \overline{\alpha}_1 + \overline{\alpha}_2)/2) \right) \tag{18}$$

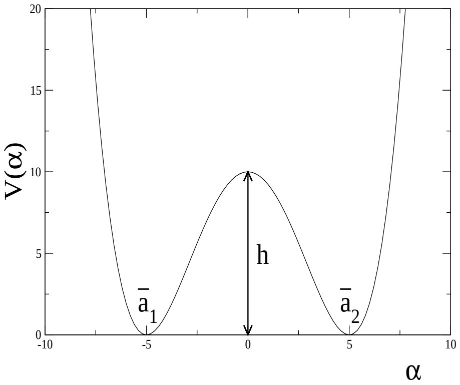

Figure 1: Potential energy landscape for α where there exist two equally valid interpretations. Eq. (19)

Without any dissipative dynamics, the optimal estimate of $\alpha(t)$ will switch between the two minima, representing instanton trajectories of a quantum particle tunnelling through the energy barrier backwards and forwards [3]. In contrast, it is well known that, at least in some regimes, the problem with dissipation has a phase transition to a truly localized state. Previous work has demonstrated such a dynamical phase transition in the strong-coupling limit (i.e. large barrier height limit) using semi-classical approximations for the dynamics [4,5,6], and in this section we will show that a perturbative RG treatment yields similar results in the opposite weak-coupling limit.

For the sake of simplicity we employ the following simple quartic potential (see Fig.1), although the results will be independent of its exact form,

$$V(\alpha) = \frac{h}{\overline{\alpha}_1^4} \left(\alpha^2 - \overline{\alpha}_1^2 \right)^2. \tag{19}$$

The α coordinates have been shifted such that $\overline{\alpha}_1 = -\overline{\alpha}_2$, and the height h of the energy barrier located at $\alpha = 0$ sets the overall energy scale. It is useful to write the effective action of Eq. (14) in dimensionless parameters

$$a = \frac{\alpha}{\overline{\alpha}_1}, \qquad b = \eta \overline{\alpha}_1^2, \qquad c = \frac{h}{\Lambda}, \tag{20}$$

where $\Lambda = D/\overline{\alpha}_1^2$ is the energy/frequency scale [2]

$$\begin{aligned} S &= \frac{1}{2} \int \frac{d\omega}{2\pi} \left(\frac{1}{2\Lambda} \omega^2 + b|\omega| \right) |a(\omega)|^2 + c\Lambda \int dt\, V'(a), &(21) \\ V'(a) &= (a^2 - 1)^2. &(22) \end{aligned}$$

[2] The constant of proportionality between energy and frequency is set to 1, akin to the common physics computation setting of $\hbar = 1$.

By power counting in the first integral the dissipative term, at low frequencies, dominates over the kinetic energy term. In the language of RG, the kinetic energy term is an irrelevant operator and can thus be ignored if we now focus our attention to frequencies below some cut-off λ. To determine the RG flow of the dimensionless coupling parameters the high-frequency components are integrated out from $\omega = \lambda - d\lambda$ to $\omega = \lambda$ to give a new effective action $\tilde{S}$ over the low frequency modes $\omega < \lambda$. To accomplish this the function $\alpha(\omega)$ is split

$$a(\omega) = a_<(\omega)\theta(|\omega| < \lambda - d\lambda) + a_>(\omega)\theta(\lambda - d\lambda < |\omega| < \lambda), \tag{23}$$

and the new action is obtained by integrating over $a_>(\omega)$,

$$\begin{aligned} Z &= \int \mathcal{D}a \exp[-S(a)], \\ &= \int\int \mathcal{D}a_< \mathcal{D}a_> \exp[-S(a_< + a_>)], \\ &= \int \mathcal{D}a_< \exp[-\tilde{S}(a_<)]. \end{aligned} \tag{24}$$

Therefore,

$$\tilde{S}(a_<) = \frac{b}{2}\int_0^{\lambda - d\lambda} \frac{d\omega}{2\pi}|\omega||a_<(\omega)|^2 + \ln\left\langle \exp\left[c\Lambda \int dt V'(a_< + \alpha_>)\right]\right\rangle_{a_>}, \tag{25}$$

where the averaging is defined by

$$\langle A \rangle_{a_>} \propto \int \mathcal{D}a_> \exp\left\{-\frac{b}{2}\int_{\Lambda - d\Lambda}^{\Lambda} \frac{d\omega}{2\pi}|\omega||a_>(\omega)|^2\right\} A. \tag{26}$$

In the weak-coupling limit, we may expand the exponential term in Eq. (25) before performing the averaging,

$$\left\langle \exp[c\Lambda \int dt V'(a_< + a_>)]\right\rangle_{a_>} = \left\langle 1 + c\Lambda \int dt V'(a_< + a_>) + ...\right\rangle_{a_>}. \tag{27}$$

Terminating the expansion to first order in the potential represents a one-loop calculation in field theories.

Making use of

$$\left\langle a_>^2(t)\right\rangle_{a_>} = \int_{\lambda - d\lambda}^{\lambda} \frac{d\omega}{\pi}\frac{1}{b|\omega|} \approx \frac{1}{\pi b}\frac{d\lambda}{\lambda}, \tag{28}$$

we find that the potential term renormalizes as

$$(c\Lambda(a^2-1)^2)_\lambda \Rightarrow (c\Lambda(a^2-1)^2)_{\lambda - d\lambda} \approx (c\Lambda)_\lambda \left[(a_<^2 - 1)^2 + (3a_<^2 - 1)\frac{2}{\pi b}\frac{d\lambda}{\lambda}\right], \tag{29}$$

where we have ignored terms including higher powers of $d\lambda/\lambda$. To recast the new lower-frequency action into the same form as the original action the dimensionless coupling parameters must be renormalised. In particular, we observe that the dimensionless barrier height c can either grow or shrink depending on the value of the dimensionless dissipation b. Note that the coordinates must also be rescaled (also known as wavefunction renormalization) for the potential in Eq. (29) to maintain the same quartic form as in Eq. (22), thereby inducing a rescaling of b. We concentrate here on the renormalized potential coupling term and find that, up to a constant,

$$c_{\lambda - d\lambda} = c_\lambda \left[1 + \frac{d\lambda}{\lambda}\left(1 - \frac{6}{\pi b}\right)\right], \tag{30}$$

giving then the following differential RG flow equation

$$\frac{dc}{d\ln\lambda} = \left(\frac{b^*}{b} - 1\right) c. \tag{31}$$

As the (dimensionless) barrier height c renormalizes towards lower frequencies we observe two types of behavior depending on whether the parameter b is greater or smaller than the critical value $b^* = 6/\pi$ (the actual numerical value may well be slightly altered by going to higher orders in the perturbative expansion, but the important point to note that it is nonzero and thus gives rise to distinct dynamical phases). For $b > b^*$ the barrier height grows without bounds and thus effectively traps $\alpha(t)$ in one of the two minima, representing a localized phase. This localization can be brought about by increasing the magnitude of η, the numerical prefactor of our dissipative nonlocal priors, and/or increasing $\overline{\alpha}_1$ the distance between the two possible interpretations of the data. On the other hand, for $b < b^*$ the potential becomes ineffective in localizing α, and thus α freely tunnels between the two wells, representing indeterminancy of the correct true model parameter.

It is interesting to note that a flow equation, similar to Eq. (31), has been reported for the opposite limit (strong-coupling) using the instanton method[5,6]. Arguably what we have really shown is that even if one starts with weak coupling, so that it should be ”easy” to jump from one interpretation to another, for $b > b^*$ we will flow to strong-coupling, at which point known results about localization take over.

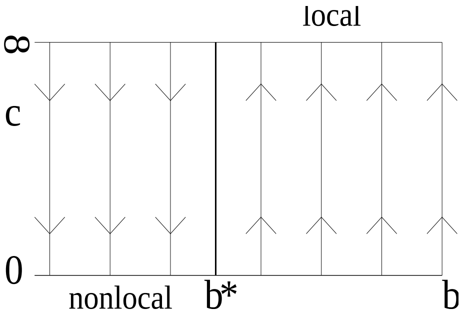

Figure 2: Schematic RG flow of the potential energy coupling parameter for $M \geq 2$. Note that the flow-lines are not expected to be strictly vertical due to wavefunction renormalization.

The qualitative picture does not change when there are more than two possible model interpretations, $M > 2$. In fact, the case of $M = \infty$ has been studied [7] where the potential energy landscape is taken to be sinusoidal, and it has been demonstrated that there again exists a critical value b^* which separates a localized phase from a nonlocalized phase. The flow of the potential energy coupling constant c is shown in Fig.2 which is expected to be qualitatively correct across the whole range $2 \leq M \leq \infty$.

3 Discussion

In summary, the optimal model estimate in the response of ambiguous signals always results in random perceptual switching when the priors only constrain the local dynamics. We have shown that when we allow the possibility of $1/f$ noise in our priors then a specific model is learnt amongst the many possible models.

The connection between estimation theory and statistical mechanics is well known. One of the key results in statistical mechanics is that local interactions in one dimension can never lead to a phase transition. Thus if we are interested in, for example, learning a single parameter by making repeated observations, then there can be no phase transition to certainty about the value of this parameter as long as our prior hypotheses about its dynamics are equivalent to local models in statistical mechanics. Markov models, Gaussian processes with rational spectra, and other common priors all fall in this local class.

The common occurrence of $1/f$ fluctuations in nature motivates the analyses of estimation theory with such priors. Crucially, $1/f$ spectra do not correspond to local models. In fact they correspond exactly to the addition of friction to the path integral describing a quantum mechanical particle, a problem of general interest in condensed matter physics and more recently in quantum computing. Here we note one important consequence of these priors, namely that we can process data in a model which admits the possibility of time variation for the underlying parameter, but nonetheless find that our best estimate of this parameter is localized for all time to one of many equally plausible alternatives. It seems that $1/f$ priors may provide a way to understand the emergence of certainty more generally as a phase transition.

References

[1] G. H. Fisher, Perception & Psychophysics **4**, 189 (1968)

[2] E. de Boer, Handbk. Sens. Physiol. **3**, 479 (1976)

[3] W. Bialek and M. DeWeese, M. Phys. Rev. Lett. **74**, 3079 (1995)

[4] A. O. Caldeira and A. J. Leggett, Phys. Rev. Lett. **46**, 211 (1981)

[5] A. J. Bray and M. A. Moore, Phys. Rev. Lett **49**, 1545 (1982)

[6] A. J. Leggett, S. Chakravarty, A. T. Dorsey, M. P. A. Fisher, A. Garg and W. Zwerger, Rev. Mod. Phys. **59**, 1 (1987)

[7] M. P. A. Fisher and W. Zwerger, Phys. Rev. Lett **32**, 6190 (1985)