Fluctuation-Dissipation Theorem and Models of Learning

Ilya Nemenman
ilya.nemenman@columbia.edu
Kavli Institute for Theoretical Physics, University of California, Santa Barbara,
CA 93106, and Joint Centers for Systems Biology, Columbia University, New York,
NY 10032, U.S.A.

Advances in statistical learning theory have resulted in a multitude of different designs of learning machines. But which ones are implemented by brains and other biological information processors? We analyze how various abstract Bayesian learners perform on different data and argue that it is difficult to determine which learning–theoretic computation is performed by a particular organism using just its performance in learning a stationary target (learning curve). Based on the fluctuation–dissipation relation in statistical physics, we then discuss a different experimental setup that might be able to solve the problem.

1 Introduction

Learning based on experience (variously known as sensing, information processing, or adaptation) is ubiquitous on all scales in biology. For example, on the molecular scale, the Lac operon in *Escherichia coli* learns the lactose concentration to produce β–galactosidase (the lactose–metabolizing enzyme) in proper quantities (Cohn & Horibata, 1959). Similarly, in the sensory system, the retinal phototransduction cascade uses the information in the arrivals of photons to learn the instantaneous light intensity and thus the current visual scene (Detwiler, Ramanathan, Sengupta, & Shraiman, 2000). Additionally, it also learns the ambient light level (to adapt to it) and the temporal correlations (to estimate motion) (Reichardt, 1961; Bialek & de Ruyter van Steveninck, 2005). On the scale of cellular (neuronal) networks, learning, memory, and adaptation in the neural code are textbook knowledge see, e. g., Brenner, Bialek, & de Ruyter van Steveninck, 2000; Fairhall, Lewen, Bialek, & de Ruyter van Steveninck, 2001). At yet larger scales, experiments on rodents are revealing how they learn and respond to changes in their environments (Gallistel, Mark, King, & Latham, 2001); this is a simple, albeit quantifiable, example of the general phenomenon we call "learning" in everyday life. Finally, we may also view evolution as an example of learning, where entire species adapt to the world by means of natural selection.

The creativity of theorists matches that of nature, and the number of various learning paradigms different in their goals, assumptions, methods, and performance guarantees is astonishing—too large to enumerate here. Fortunately, it is possible to build uniform foundations for many of these learning machines (Vapnik, 1998; Nemenman, 2000; Bialek, Nemenman, & Tishby, 2001) and to find analogs among, say, structural risk minimization (Vapnik, 1998) and Bayesian (Press, 1989) models. However, while one might argue that biological systems are (efficiently) implementing one of many abstract learning–theoretic computations (Attneave, 1954; Barlow, 1959, 1961; Atick, 1992; Bialek et al., 2001), it is often unclear which exact computation is performed in a particular case. For example, what is a learning–theoretic model equivalent to a rat (Gallistel et al., 2001) or to a simple neural network that tries to maximize its reward (Seung, 2003)? Answering such questions may explain some of animal behaviors, uncover which assumptions they make about the surrounding world, and establish quantitative limits on their learning performance.

To attack the problem, one can construct a biologically plausible computing machine with a known learning–theoretic equivalent (Rao, 2004) and then search for a structural similarity with a real living organism. We do not pursue this approach, but choose a more traditional route to establish the equivalence: comparison of the performance of real creatures to that of abstract learning machines. As we will argue, analysis of paradigmatic learning curves is not always easy. Thus, one of the most important results of the article is a suggestion of a new protocol for making such comparisons. The intuition behind the suggestion comes from the famous fluctuation–dissipation theorem (Ma, 1985). Based on this analysis and on plausible assumptions about statistics of natural stimuli, we also suggest that a particular learning–theoretic model might be better suited for biological learning than the alternatives, and thus it should be realized often in reality if optimization of learning is desired.

To follow this route, we need to understand characteristics of learning within different mathematical models fairly well, and a large part of the article is devoted to this. Analysis is done in the framework of unsupervised Bayesian learning of probability distributions since (1) evidently, Bayesian paradigm is relevant in neuroscience (Kording & Wolpert, 2004), (2) different learning frameworks are often equivalent, and (3) according to Bialek et al. (2001), other learning problems usually can be reduced to unsupervised learning of distributions. Much of this first part of the article is an abridged review, which follows the spirit and the notation of Bialek et al. (2001) and often prefers clarity to mathematical rigor. We do not try to make the review self-contained, but instead want to elucidate and emphasize some important points that might have been of lesser interest in other contexts and also to present some novel results, mostly developed in the appendixes. After these developments, we return to the main question of this work: How

can one realistically determine an equivalent learning–theoretic model for a biological organism?

2 The Basics of Learning

Learning machines should be powerful enough to explain complex phenomena. However, when data are scarce, this power leads to overfitting and poor generalization. Thus, a balance must be struck between the abilities to explain and to overfit, and this balance will depend on the number of data available. In accord, much of statistical learning theory (Jeffreys, 1936; Schwartz, 1978; Janes, 1979; Rissanen, 1989; Clarke & Barron, 1990; MacKay, 1992; Balasubramanian, 1997; Vapnik, 1998; Nemenman, 2000; Bialek et al., 2001) has been devoted to putting the famous paradigm of William of Ockham, Pluralitas non est ponenda sine neccesitate, on firm mathematical footing in various theoretical frameworks. In particular, in Bayesian formulation (Press, 1989; Bernardo, 2003), we know now how proper Bayesian averaging creates Occam factors that punish for complexity and weigh posterior probabilities toward those estimates among a finite set of parametric model families that have the best overall predictive power (Bialek et al., 2001) but do not necessarily produce the best fit to the observed data. This has been called Bayesian model selection.[1]

The waters get murkier in a nonparametric or infinite parameter setting when the whole functional form of an unknown object is to be inferred. Bayesian nonparametric developments generally parallel parametric ones, and techniques of quantum field theory (QFT) help in computations (Bialek, Callan, & Strong, 1996; Holy, 1997; Bialek et al., 2001; Cucker & Smale, 2001; Nemenman & Bialek, 2002; Lemm, 2002). However, the exact relationship between the two settings is unknown, and some results suggest subtle logarithmic differences between the cases (Hall & Hannan, 1988; Rissanen, Speed, & Yu, 1992; Bialek et al., 2001).

We now review these and other Bayesian learning machines. We start with an introduction of some important and useful quantities.

Suppose we observe independently and identically distributed (i.i.d.) samples x_i, $i = 1, \ldots, N$. For simplicity, we assume that x is a scalar, but this does not affect most of the discussion. We need to estimate the probability density that generates the samples. A priori we know that this density, $Q(x|\alpha)$, can be indexed by some (possibly infinite-dimensional) vector of parameters α, and the probability of each parameter value is $\mathcal{P}(\alpha)$. Then we

---

[1] With the creationism–evolution tension mounting in teaching of biology in U.S. schools, it is amusing to see how two friars, William of Ockham and Thomas Bayes, teamed up with modern mathematicians to produce, in my view, the clearest formulation of the theory of learning from past experiences. If this approach results in a better understanding of biological designs, the situation will be even more peculiar.

define the density of models (solutions) at a given distance (dissimilarity, or divergence) $D(\bar{\boldsymbol{\alpha}}, \boldsymbol{\alpha}) = \epsilon$ away from the unknown true target $\bar{\boldsymbol{\alpha}}$, which is being learned:

$$\rho(\epsilon; \bar{\boldsymbol{\alpha}}) = \int d\boldsymbol{\alpha} \, \mathcal{P}(\boldsymbol{\alpha}) \, \delta \left[D(\bar{\boldsymbol{\alpha}}, \boldsymbol{\alpha}) - \epsilon \right]. \tag{2.1}$$

For Bayesian inference of probability densities, the correct measure of dissimilarity is the Kullback–Leibler (KL) divergence, $D_{\text{KL}}(\bar{\boldsymbol{\alpha}}||\boldsymbol{\alpha}) = \int dx \, Q(x|\bar{\boldsymbol{\alpha}}) \log \frac{Q(x|\bar{\boldsymbol{\alpha}})}{Q(x|\boldsymbol{\alpha})}$ (Bialek et al., 2001), which has an important information–theoretic interpretation (Cover & Thomas, 1991). However, in other situations, different choices of D can and should be made.

Performance of a Bayesian learner is usually measured by the speed with which the posterior probability concentrates for $N \to \infty$ (the learning curve) and by whether the point of concentration is the true unknown target (consistency). These characteristics illuminate the importance of ρ as it relates to both of them. First, it has been proven that if for $\epsilon \to +0$, the density, $\rho(\epsilon; \bar{\boldsymbol{\alpha}})$, is not zero, then the Bayesian problem is consistent (Nemenman, 2000; Bialek et al., 2001). Intuitively this is because, for large density, statistical fluctuations of the sample and of the estimated parameters result in small $D_{\text{KL}}(\bar{\boldsymbol{\alpha}}||\text{estimate})$, making convergence to the target almost certain.

Relation of ρ to the learning curve is more complicated. We can calculate the average (over samples) Occam factor for a given target (the generalization error, or the fluctuation determinant) to the leading order in $1/N$:

$$\mathcal{D}(\bar{\boldsymbol{\alpha}}; N) \approx - \log \int d\epsilon \, \rho(\epsilon; \bar{\boldsymbol{\alpha}}) \mathrm{e}^{-N\epsilon}. \tag{2.2}$$

This is the term that emerges as the penalty for complexity in Bayesian model selection (Balasubramanian, 1997; Bialek et al., 2001). If averaged over $\bar{\boldsymbol{\alpha}}$, the Occam factor becomes predictive information (Bialek et al., 2001), which is the average number of bits that N samples provide about the unknown parameters,

$$I_{\text{pred}}(N) = \int d\bar{\boldsymbol{\alpha}} \, \mathcal{P}(\bar{\boldsymbol{\alpha}}) \mathcal{D}(\bar{\boldsymbol{\alpha}}; N). \tag{2.3}$$

Finally, one can define the universal learning curve, which measures the expected D_{KL} between the target and the estimate after N observations (Bialek et al., 2001). Up to the first order in the large parameter N, this is

$$\Lambda(\bar{\boldsymbol{\alpha}}; N) \approx \frac{d\mathcal{D}(\bar{\boldsymbol{\alpha}}; N)}{dN}, \tag{2.4}$$

$$\Lambda(N) = \int d\bar{\alpha}\, \mathcal{P}(\bar{\alpha})\Lambda(\bar{\alpha}; N) \approx \frac{d\, I_{\text{pred}}}{d\, N}. \tag{2.5}$$

Many of these quantities, especially I_{pred}, are also natural objects when analyzing the complexity of a time series (Bialek et al., 2001).

3 Different Models of Learning

Since one of the goals of this work is to investigate if learning machines can be discriminated by means of their learning performance, specifically $\Lambda(N)$, here we discuss how Λ depends on N for different scenarios.

3.1 Learning in a Finite Set of Parameters. Consider a setup where α can take M discrete values $a_1, a_2, \ldots, a_M$ with a priori probabilities $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_M$, and their divergences from the target a_1 are $0 = d_1 < d_2 < \cdots < d_M$. The density is $\rho(\epsilon; a_1) = \sum_{i=1}^{M} \mathcal{P}_i \delta(d_i - \epsilon)$. For $N \to \infty$, we have

$$\mathcal{D}(a_1; N) = -\log \sum_{i=1}^{M} \mathcal{P}_i \exp[-Nd_i] \approx -\log \mathcal{P}_1 - \mathcal{P}_2/\mathcal{P}_1 \exp[-Nd_2], \tag{3.1}$$

$$\Lambda(a_1; N) \approx d_2 \mathcal{P}_2/\mathcal{P}_1 \exp[-Nd_2]. \tag{3.2}$$

So exponential learning curves (and asymptotically finite $\mathcal{D}$ and I_{pred}) correspond to learning a possibility in a finite set. Similarly, we can construct models with $\Lambda(N) \propto 1/N^\nu$, $\nu > 1$, and they will also have asymptotically finite I_{pred}.

3.2 Finite Parameter Learning. Now let the target probability density $Q(x|\bar{\alpha})$, or a model, belong to a set of densities A, a model family, that can be indexed by a vector of parameters $\alpha \in A$, $\dim \alpha = K < \infty$, and $\forall \alpha \in A$, $\mathcal{P}(\alpha) > 0$. Then if A is not compact, or if the KL divergence between $\bar{\alpha}$ and the boundary of A is larger than ϵ, then the density of solutions for such K–parametric family is (Bialek et al., 2001)[2]

$$\rho(\epsilon; \bar{\alpha}) \approx \mathcal{P}(\bar{\alpha}|r) \frac{2\pi^{K/2}}{\Gamma(K/2)} \frac{\epsilon^{(K-2)/2}}{\sqrt{\det \mathcal{F}_K}}, \tag{3.3}$$

---

[2] A different scaling dimension d_K may appear in these formulas instead of K, the number of parameters. For example, for a redundant parameterization, $d_K < K$. Opposite situations, $d_K > K$ and even $d_K \to \infty$, are also possible (Bialek et al., 2001).

where

$$\mathcal{F}_K^{\mu\nu}(\bar{\boldsymbol{\alpha}}) = \frac{\partial^2 D_{\mathrm{KL}}(\bar{\boldsymbol{\alpha}}||\boldsymbol{\alpha})}{\partial \alpha_\mu \partial \alpha_\nu}\bigg|_{\boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}}}. \tag{3.4}$$

$N\mathcal{F}_K$ is the Fisher information matrix (Cover & Thomas, 1991); its eigenvectors are the principal axes of the error ellipsoid in the parameter space, and the (inverse) eigenvalues are variances of parameter estimates along each of these directions. The prefactor $2\pi^{K/2}/\Gamma(K/2)$ is the area of the K-sphere, and it has to be multiplied by the fraction of the sphere that is inside A if the latter is (semi)compact. Equation 3.3 now gives

$$I_{\mathrm{pred}}(N) \approx \mathcal{D}(\bar{\boldsymbol{\alpha}}, N) \approx K/2 \log N, \tag{3.5}$$

$$\Lambda(N) \approx \Lambda(\bar{\boldsymbol{\alpha}}, N) \approx K/(2N). \tag{3.6}$$

The situation changes slightly if $\bar{Q} \notin A$ (here $\bar{Q}$ is the target density), and the prior assumptions about the world are wrong. Then we find the best approximation to the target within A, $\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha} \in A} D_{\mathrm{KL}}(\bar{Q}||\boldsymbol{\alpha})$, and define the distance between $\bar{Q}$ and A, $D_A(\bar{Q}) \equiv D_{\mathrm{KL}}(\bar{Q}||\hat{\boldsymbol{\alpha}})$ (this is similar to the I-projection—Csiszar, 1975—but the order of arguments in D_{KL} is different). In this case, the model density is zero for $\epsilon \leq D_A(\bar{Q})$, and the estimate concentrates near $\hat{\boldsymbol{\alpha}}$ as $N \to \infty$. Thus, if the radius of curvature of A is much larger than ϵ and $D_A(\bar{Q})$ is also small, then equations 3.3 and 3.6 generalize to

$$\rho(\epsilon; \bar{\boldsymbol{\alpha}}) \approx \begin{cases} \mathcal{P}(\hat{\boldsymbol{\alpha}})\frac{2\pi^{K/2}}{\Gamma[K/2]}\frac{[\epsilon - D_A(\bar{Q})]^{(K-2)/2}}{\sqrt{\det \mathcal{F}_K}}, & \epsilon > D_A(\bar{Q}), \\ 0, & \epsilon \leq D_A, \end{cases} \tag{3.7}$$

$$\Lambda(\bar{Q}, N) \approx D_A(\bar{Q}) + K/(2N). \tag{3.8}$$

3.3 Nested Finite Parameter Models. Suppose now the target $\bar{Q}(x)$ that generates the observations belongs to one of R model families, $A_r, r = 1 \ldots R$, with $\mathrm{Prob}(\bar{Q} \in A_r) = \mathcal{P}(r)$. Models in each of the families are indexed by parameters $\boldsymbol{\alpha}^{(r)}$, $\dim \boldsymbol{\alpha}^{(r)} = K(r) < \infty$, so that the density of observing x in a given model is $Q_r(x|\boldsymbol{\alpha}^{(r)})$. Within each family, the parameters are a priori distributed according to $\mathcal{P}(\boldsymbol{\alpha}^{(r)}|r)$.

We will assume that the families are nested. That is, $Q_r(x|\boldsymbol{\alpha}^{(r)}) \equiv Q(x|\boldsymbol{\alpha})$ are independent of r, and in each family the values of α_μ, $\mu > K(r)$ are identically zero. Further, the nonzero parameters have the same a priori distributions in all families:

$$\mathcal{P}(\alpha_\mu|r) = \begin{cases} p(\alpha_\mu), & \mu \leq K(r) \\ \delta(\alpha_\mu), & \mu > K(r) \end{cases} \tag{3.9}$$

$$\mathcal{P}(\boldsymbol{\alpha}|r) = \prod_{\mu=1}^{R} \mathcal{P}(\alpha_\mu|r). \tag{3.10}$$

Thus, a parameter α_μ is switched on (or activated) when r reaches $r_\mu \equiv \min_r\{r : K(r) \geq \mu\}$. Discussion of such nested models has been current in Bayesian (Bernardo, 2003; Raftery & Zheng, 2003) and frequentist (Neter, Kutner, Nachtsheim, & Wasserman, 1996) literature for many years. However, we are unaware of any comprehensive analysis relevant to the questions analyzed in our current presentation, such as those in appendixes A to C.

If $R \to \infty$, then we require that the union of all families forms a complete set, so that every sufficiently smooth probability density can be approximated arbitrarily closely by some member of the union (if needed, this definition can be made more precise).

For simplicity, in this letter we focus on[3]

$$p(\alpha_\mu) = \mathcal{N}(0, \sigma_\mu^2), \tag{3.11}$$

$$\sigma_\mu = cr_\mu^{-\beta}, \ \beta \geq 0, \ c = \text{const}, \tag{3.12}$$

where $\mathcal{N}(a, b)$ denotes a normal distribution with the mean of a and the variance of b. In particular, $\beta = 0$ corresponds to the same in-family a priori variances for all active parameters. This is common when discussing Bayesian model selection.

While these priors describe a set of parametric models, another view is also possible. The joint distribution of r, $\boldsymbol{\alpha}$, and $\{x\}$ is $P(\{x\}, \boldsymbol{\alpha}, r) = Q(\{x\}|\boldsymbol{\alpha})\mathcal{P}(\boldsymbol{\alpha}|r)\mathcal{P}(r)$, which results in

$$P(\{x\}, \boldsymbol{\alpha}) = \sum_r Q(\{x\}|\boldsymbol{\alpha})\mathcal{P}(\boldsymbol{\alpha}|r)\mathcal{P}(r) = Q(\{x\}|\boldsymbol{\alpha}) \sum_r \mathcal{P}(\boldsymbol{\alpha}|r)\mathcal{P}(r)$$

$$\equiv Q(\{x\}|\boldsymbol{\alpha})\mathcal{P}(\boldsymbol{\alpha}), \tag{3.13}$$

where the last equation defines $\mathcal{P}(\boldsymbol{\alpha})$, the overall prior over $\boldsymbol{\alpha}$. Unlike $\mathcal{P}(\boldsymbol{\alpha}|r)$, $\mathcal{P}(\boldsymbol{\alpha})$ is not factorizable and is not differentiable at zero for any α_μ, $\mu > K(1)$. Thus, the nested setup may be viewed as inference in a combined model family with $K(R)$ parameters. In particular, for R and $K(R) \to \infty$, the learning problem has a countable infinity of parameters leading to the common assumption of equivalence with the nonparametric inference.

---

[3] Nestedness, completeness, and normality of the priors are needed only for comparison with models discussed later; they are not essential for Bayesian learning.

It is of interest to calculate the combined a priori mean and variance of α. Integrating over all α_ν, $\nu \neq \mu$, we get the combined prior for α_μ:

$$\mathcal{P}(\alpha_\mu) = \delta(\alpha_\mu) \sum_{r < r_\mu} \mathcal{P}(r) + p(\alpha_\mu) \sum_{r \geq r_\mu} \mathcal{P}(r). \tag{3.14}$$

By equation 3.11, the a priori means of all parameters are zero, and the variances are

$$\langle \delta\alpha_\mu^2 \rangle = \sigma_\mu^2 \sum_{r \geq r_\mu} \mathcal{P}(r). \tag{3.15}$$

Thus, the bare variance σ_μ^2 is "renormalized" by the probability of being in a family in which the parameter is nonzero. An interesting special case is

$$\mathcal{P}(r) \propto r^{-\gamma}, \quad \gamma > 1, \quad R \to \infty, \tag{3.16}$$

$$r_\mu = \mu. \tag{3.17}$$

Then the a priori variance gets a simple form:

$$\langle \delta\alpha_\mu^2 \rangle \propto \mu^{-\beta} \sum_{r=\mu}^{\infty} r^{-\gamma} \sim \mu^{-\beta-\gamma+1}. \tag{3.18}$$

Thus, $\langle \delta\alpha_\mu^2 \rangle$ depends as much on the bare variance as on the speed of decay of $\mathcal{P}(r_\mu)$. This suggests that the learning properties of the nested setup will depend equivalently on β and γ. In fact, as shown in appendix A, this is not true: while behavior of $p(\alpha_\mu)$ is important, any reasonable choice of $\mathcal{P}(r)$ does not affect success of the learning.

From equation 3.7, we can now evaluate the model density and the learning curve for the nested setup. For each value of $\bar{\alpha}$ and r, we can find the model $\hat{\alpha}_r = \arg\min_{\alpha \in A_r} D_{KL}(\bar{\alpha}||\alpha)$ that best approximates $\bar{\alpha}$ in A_r, and define $D_r(\bar{\alpha}) \equiv D_{KL}(\bar{\alpha}||\hat{\alpha}_r)$, the distance between $\bar{\alpha}$ and A_r. If $\bar{\alpha} \in A_r$, then $\bar{\alpha} = \hat{\alpha}_r$, and $D_r(\bar{\alpha}) = 0$. However, if $\bar{\alpha} \notin A_r$, then $D_r(\bar{\alpha}) > 0$. We then have

$$\rho(\epsilon; \bar{\alpha}) = \sum_{r: D_r(\bar{\alpha}) \leq \epsilon} \mathcal{P}(r)\mathcal{P}(\hat{\alpha}_r|r) \frac{2\pi^{K(r)/2}}{\Gamma[K(r)/2]} \frac{[\epsilon - D_r(\bar{\alpha})]^{[K(r)-2]/2}}{\sqrt{\det \mathcal{F}_{K(r)}}}. \tag{3.19}$$

The learning curve in this scenario strongly depends on the target. Let $\bar{\alpha}$ have $\bar{r}$ active modes (with $\bar{r}$ determined according to $\mathcal{P}(\bar{r})$), and let each

of these modes have an amplitude $\sim\sigma$ (that is, $\beta = 0$). Then $D_r(\bar{\alpha})$ is either exactly zero (for $r \geq \bar{r}$) or large (for $r < \bar{r}$). So equation 3.19 becomes

$$\rho(\epsilon \to 0; \bar{\alpha}_{\text{typ};\bar{r}}) \sim \sum_{r \geq \bar{r}} \mathcal{P}(r)\mathcal{P}(\hat{\alpha}_r|r) \frac{2\pi^{K(r)/2}}{\Gamma[K(r)/2]} \frac{\epsilon^{[K(r)-2]/2}}{\sqrt{\det \mathcal{F}_{K(r)}}}, \tag{3.20}$$

where $\bar{\alpha}_{\text{typ};\bar{r}}$ is a distribution typical in $A_{\bar{r}}$. This is dominated by $r = \bar{r}$, and for $N \gg K(\bar{r})$, the learning curve is

$$\Lambda(N) \approx \frac{K(\bar{r})}{2N}. \tag{3.21}$$

It is now clear that averaging over $\mathcal{P}(\bar{r})$ is not informative. Note also that for $N \lesssim \bar{r}$, the learning curve goes through a cascade of $K(r)/N$ behaviors, $1 \leq r \leq \bar{r}$, and changes of the prefactor of the N^{-1} scaling correspond to activations of new parameters, which happen rather abruptly (see appendix A).

3.4 Nonparametric Learning. Nonparametric learning usually refers to inferring a functional form of a probability density $Q(x)$, or rather of $\phi(x) \equiv -\log Q(x)$, with some smoothness constraints on it. The constraints may be in the form of bounding some derivatives of Q or ϕ, which was the choice of Hall & Hannan (1988) and Rissanen et al. (1992).[4] Alternatively, in the Bayesian framework followed here, the constraints may be incorporated into a functional prior that makes sense as a continuous theory, independent of discretization of x on small scales. For x in one dimension, the minimal and the most common choice is (Bialek et al., 1996; Aida, 1999; Nemenman & Bialek, 2002; Lemm, 2002)

$$\mathcal{P}[\phi(x)] = \frac{1}{\mathcal{Z}} \exp\left[-\frac{\ell^{2\eta-1}}{2} \int dx \left(\frac{\partial^\eta \phi}{\partial x^\eta}\right)^2\right]$$
$$\times \delta\left[\frac{1}{l_0} \int dx\, e^{-\phi(x)} - 1\right], \tag{3.22}$$

where $\eta > 1/2$, $\mathcal{Z}$ is the normalization constant, and the δ-function enforces normalization of Q. The hyperparameters ℓ and η are called the smoothness scale and the smoothness exponent, respectively. Fractional-order

---

[4] These authors used histogramming density estimators, which have no hierarchy of model families; this is especially true for Rissanen et al. (1992), who allowed locally varying bin widths. Therefore, these techniques cannot be referred to as nested parametric methods. On the other hand, they allow an arbitrarily precise fit to any probability density and may require an arbitrarily large number of break points and density values for complete specification. This is the reason for treating them as nonparametric.

derivatives are defined by multiplying by the wave number to the appropriate power in the Fourier representation of ϕ (we assume periodicity on $[0, 1)$).

This prior is equivalent to specifying a one-dimensional quantum field theory (Bialek et al., 1996; Holy, 1997; Nemenman & Bialek, 2002; Lemm, 2002), and QFT methods have been successful in the analysis. In particular, the maximum likelihood estimate of the distribution $Q^*(x) \equiv \exp[-\phi^*(x)]$ is given by the following differential equation,

$$\ell^{2\eta-1}\mathbb{R}_{-2\eta}\frac{\partial^{2\eta}\phi^*(x)}{\partial x^{2\eta}} - NQ^*(x) + \sum_i \delta(x - x_i) = 0, \tag{3.23}$$

where the operator $\mathbb{R}_\theta$ shifts the phase of each Fourier component of its argument by $\pi\theta/2$.[5] The equation shows that derivatives of ϕ^* and Q^* of order $2\eta - 1$ have step discontinuities. Thus, for $2\eta = 1$, the maximum likelihood solution itself, $\phi^*(x)$, is discontinuous, and for $2\eta < 1$, the singularities are even more severe. We may characterize sample-dependent fluctuations in Q^* by $D_{KL} = \int dx\, Q_1^*(x) \log Q_1^*(x)/Q_2^*(x)$, where Q_1^* and Q_2^* are maximum likelihood solutions for different sample realizations. If Q^* has at least step discontinuities at the sample points and these points are random, then D_{KL} does not fall to zero as N grows. Therefore, the QFT setup becomes inconsistent at $\eta = 1/2$, even though Bayesian formulation is proper, and the prior can still be normalized by, for example, going to the Fourier representation. This is in contrast to the nested setup, where normalizable priors guarantee consistency.

Bialek et al. (1996, 2001) have calculated the $\epsilon \to 0$ model density and the fluctuation determinant for different η's. By noticing from equation 3.23 that N and ℓ can enter the solutions only in a combination $N/\ell^{2\eta-1}$, we extend their results and recover correct dependence not only on η (for $\eta > 1/2$) but also on ℓ:

$$\rho(\epsilon; \bar{\phi}) \approx A[\bar{\phi}]\, \epsilon^\xi \exp\left[-\frac{B[\bar{\phi}]}{\ell\epsilon^{1/(2\eta-1)}}\right], \tag{3.24}$$

$$\mathcal{D}(\bar{\phi}; N) \approx C[\bar{\phi}]\left(\frac{N}{\ell^{2\eta-1}}\right)^{1/2\eta}, \tag{3.25}$$

$$\Lambda(\bar{\phi}, N) \approx \frac{C[\bar{\phi}]}{2\eta\, \ell^{2\eta-1}}\left(\frac{N}{\ell^{2\eta-1}}\right)^{1/2\eta-1}. \tag{3.26}$$

Here, ξ depends only on η, and A, B, and C are some known related functionals that do not depend on ℓ. These asymptotics kick in when $N \gg 1/\ell$.

---

[5] For a comprehensive treatment of fractional differentiation, see Samko, Kilbas, and Marichev (1987).

In particular, for smaller N, $\Lambda \sim 1$ and is barely decreasing. The dependence of $C[\bar{\phi}]$ on $\bar{\phi}$ may be significant (and, possibly, diverging for ill-behaved targets). However, from equation 3.24, the dependence on η near $2\eta - 1 \to +0$ is easier to analyze,

$$C[\bar{\phi}] \sim (2\eta - 1)^{1/2\eta - 1}, \tag{3.27}$$

with an undetermined value at $2\eta = 1$. So for $\eta \to 1/2$, $\mathcal{D}$ approaches extensivity in N and then becomes ill defined, again signaling inconsistency. As discussed by Bialek et al. (2001), problems with $\mathcal{D}(N)/N \to$ const are the most complicated correctly posed learning problems that exist, and they can be studied in the Bayesian QFT setting.

For comparison with the nested case (see appendix B), we may replace ϕ by its Fourier series,

$$\phi(x|\boldsymbol{\alpha}) \equiv -\log Q(x|\boldsymbol{\alpha}) = \alpha_0 + \sum_{\mu=1}^{r} \left(\alpha_\mu^+ \cos 2\pi\mu x + \alpha_\mu^- \sin 2\pi\mu x \right), \tag{3.28}$$

$$\alpha_0 = \log \int dx \, \exp\left[-\sum_{\mu=1}^{r} \left(\alpha_\mu^+ \cos 2\pi\mu x + \alpha_\mu^- \sin 2\pi\mu x \right) \right], \tag{3.29}$$

with $r \to \infty$ (finite r results in a finite parameter model). The last equation enforces normalization, $\int dx \, Q(x|\boldsymbol{\alpha}) = 1$, and it is equivalent to the constraint $\delta\left(\int \exp[-\phi(x|\boldsymbol{\alpha})]dx - 1 \right)$ in the prior $\mathcal{P}(\boldsymbol{\alpha}|r)$ or $\mathcal{P}[\phi(x)]$. Since the Jacobian of the transformation $\phi(x) \to \{\alpha_\mu^\pm\}$ is a constant, equation 3.22 amounts to zero-mean gaussian priors over $\alpha_\mu^\pm$ with the variance (Nemenman & Bialek, 2002),

$$\left\langle \left(\delta\alpha_\mu^\pm\right)^2 \right\rangle = \frac{2}{\ell^{2\eta-1}} \frac{1}{(2\pi\mu)^{2\eta}}, \qquad \mu > 0. \tag{3.30}$$

Equations 3.18 and 3.30 suggest that the nested and nonparameteric case are similar: the a priori means of the amplitudes are zero, and the variances fall off as power laws in μ. However, the field theory model requires the variance to decrease at least as fast as $1/\mu$ (recall that $\eta > 1/2$), while the finite parameter case does not impose such constraints. This is an indication of an essential difference between the models: in the nested case, the priors, specifically the a priori variances of parameters, have less of an influence on learning. This can be easily explained. QFT nonparametric models do not have a sharp separation between active and passive modes. The modes with low μ are determined by the data, but fluctuations for larger μ are inhibited only due to the small a priori variances, equation 3.30. The exact attenuation of the fluctuation depends on the values of η and ℓ, and the cumulative contribution to posterior variance of the estimator may be substantial. In

contrast, for the finite parameter nested case, once the most probable model family is determined, fluctuations of the higher-order parameters are inhibited exponentially (see appendix A). The cumulative fluctuations are then small and almost independent of the a priori parameter variances, and the learning may succeed even for $\mathcal{P}(r)$ with a long tail.

The dependence of the QFT model on the prior can be weakened by treating ℓ as an unknown random variable and averaging over it (Bialek et al., 1996; Nemenman & Bialek, 2002) (similar averaging over η has not yet been performed). This is akin to nesting of finite-dimensional models and improves learning curves for a wide range of targets. On the other hand, integration over ℓ produces the theory that is not necessarily local in $\phi(x)$, couples all of the Fourier amplitudes, and is difficult to compare to the nested finite parameter setup directly. Therefore, we do not discuss the averaging in what follows, but assume that the values of η and ℓ used for learning are the best for a particular target being learned.

3.5 Comparing the Performance. One of the goals of the letter is to decide if learning curves can be used to distinguish which learning machine is a good description of a particular biological system. To this extent, we need to analyze responses of various learners to data that they are not expecting. Thus, in this section, we derive learning curves for a finite parameter, a nested ($\beta = 0$), and a QFT machine on data that is typical in the prior of one of the two others. With mismatched data and expectations, the learning curve cannot be optimal but may come quite close.

First, consider $\bar{Q}$ taken from the QFT prior. The learning curve for the finite parameter model is given by equation 3.8—a N^{-1} decay toward some approximate target. Further, as shown in appendix C, the learning curves for complete nested models, equation C.4, and for the QFT machine, equation 3.26, which is the best possible machine for such data, differ only logarithmically.

If instead we study a distribution that is typical in the nested case for some $\bar{r}$ (equivalently, a finite parameter distribution with $K(\bar{r})$ parameters), then a finite parameter model again gives equation 3.8. On the other hand, for $N \lesssim K(\bar{r})$, no complete learning machine can estimate all required unknown parameters, and Λ does not have a well-defined scaling (Nemenman & Bialek, 2002). The differences between the machines emerge for $N \gg \bar{r}$. The nested machine eventually asymptotes to equation 3.21, and starts learning at the rate of $1/N$. However, the QFT setup performs differently: when $\Lambda \to 0$ and all $\bar{r}$ modes are well approximated, the machine continues trying to fit higher-order modes, which it expects to be present even though they are not. This will result in the same fluctuation determinant as in equation 3.25, switching to the usual asymptotic $\Lambda \propto (N\ell)^{1/2\eta-1}$ instead of the optimal N^{-1}.

So, surprisingly, when the target has a finite number of degrees of freedom, the nested setup is qualitatively faster than the QFT learning machine.

4 Learning a Changing Target

One never needs to know the distribution that generated the data to an infinite precision, and some $\epsilon > 0$ approximation is usually enough. Further, if learning in biological systems is stochastic, as argued, for example, by Seung (2003), then ϵ is bounded from below by the noise variance. As shown by Fairhall et al. (2001) and especially by Gallistel et al. (2001), convergence to the "good enough" estimate happens so quickly that the transient learning curves are difficult to resolve. Is the performance difference between the nested and the QFT scenarios seen in the previous section important? And can it be used to discriminate between the models?

4.1 Model Density and Variable Stimuli. Notice that often the target itself changes while being learned. The ambient light intensity may be fluctuating while our eye estimates it, or the variance of angular velocities measured by a fly motion-sensitive neuron can be varied by an experimenter while the fly tries to adapt to it (Fairhall et al., 2001). In these cases, one has to learn constantly to stay at the allowed ϵ-error, and then a faster learning machine may be truly advantageous. However, even for a variable target, the nested learner will not be helpful if a small change of the target parameters throws it back to a very large Λ, or the changing target may drift to a region where $\bar{r}$ is so large that the nested setup is not better than the nonparametric one anymore.

To answer these concerns, instead of focusing on the density of solutions as a function of the allowed error ϵ, we will keep ϵ fixed and vary $\bar{\alpha}$. For some small ϵ, a schematic drawing of dependence of ρ on $\bar{\alpha}_1^+$ and $\bar{\alpha}_2^+$ with the other parameters fixed at 0 is shown in Figure 1. In the nested case, there is a ridge along $\alpha_2^+ \approx 0$, where the density is, at least, $\sim 1/\sqrt{\epsilon}$ larger than anywhere else (cf. equation 3.19). The ridge comes from the prior, equation 3.13, for $\bar{\alpha}_2^+ = 0$ being singularly larger than for $\bar{\alpha}_2^+ \neq 0$, and the singularity is then smoothed out by ϵ-approximation. In comparison, the nonparametric prior has a bivariate normal shape, which, after ϵ-smearing, results in a weak target dependency of ϵ-independent prefactors in equation 3.24; thus, $\rho(\bar{\alpha})$ varies slowly.[6]

Figure 1 answers both of the concerns mentioned above. For a QFT machine, densities everywhere are comparatively small. So a small change of the target means vast and slow relearning. In contrast, if, for a nested case,

---

[6] The plots of $\mathcal{P}(\bar{\alpha})$ and $\rho(\epsilon; \bar{\alpha})$ have very different meanings. The volume under the $\mathcal{P}(\bar{\alpha})$ surface is fixed by normalization, $\int d\bar{\alpha} \, \mathcal{P}(\bar{\alpha}) = 1$. Thus, high a priori probability on any singular line, for example, $\alpha_2^+ = 0$, necessarily means a lower prior elsewhere. Such considerations are the reason for no-free-lunch theorems (Wolpert, 1995). In the language of the model density, the normalization condition is $\int d\epsilon \, \rho(\bar{\alpha}; \epsilon) = 1$. However, there are no constraints on the density integrated over $\bar{\alpha}$, and a large density for some target does not necessarily result in a lower density elsewhere.
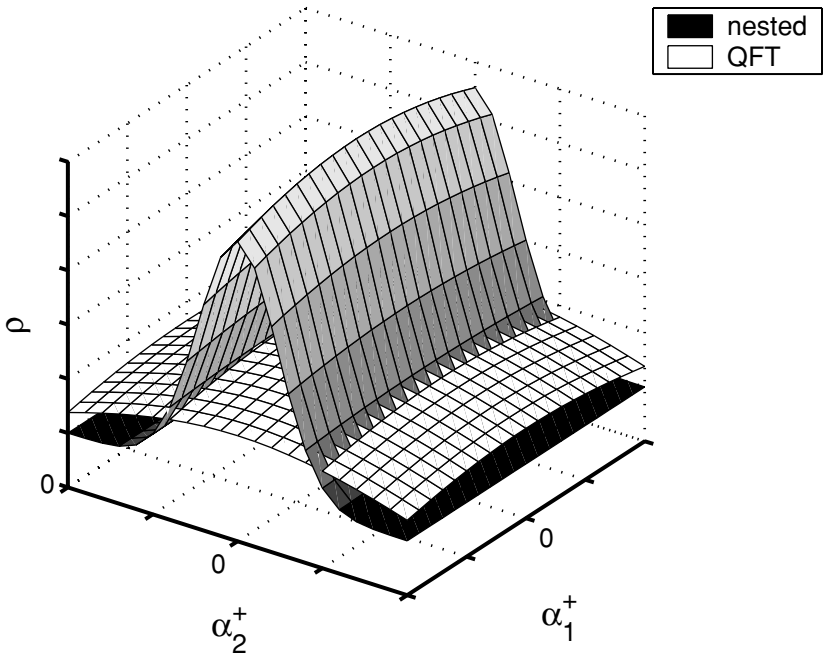
Figure 1: Schematic density of models as a function of the target location.

$\bar{\alpha}$ is in the large-density region, then there are many other models in the vicinity. Small parameter changes likely leave the target close, and not much needs to be relearned. Further, since the ridge drops off smoothly, models in the vicinity of a large-density target also have large densities, and thus are learned fast as well. Of course, this holds only when the target varies mostly along a small set of directions, and density ridges are aligned with those. Importantly, since at a finite ϵ, the ridge has a finite width, a perfect alignment is not necessary.

We believe that many natural signals have such structure. For example, in phototransduction, instantaneous intensity is determined by the statistics of reflectivities of objects that come in the view and by the mean ambient light intensity. The statistics barely change over long timescales, while the mean intensity depends on, for example, clouds shading the sun and varies a lot and rapidly. The photoreceptor may want to adapt to intricate details of the distribution of reflectivities, but only after it accurately learns the mean light level. A similar separation of timescales is observed in transcriptional regulation, where, for example, changes in the lactose concentration happen on the scale of minutes, while statistics of lactose bursts depend on the environment and is constant for generations. In neuroscience, when

estimating an angular velocity, a fly takes into the account the preceding velocity variance (Fairhall et al., 2001), but it may not have time for reaction to higher-order moments. Thus, we believe that many natural learners that have a need to learn fast, but also to be able to learn a very wide class of models accurately on longer timescales, will be organized as nested learning machines with the density ridges approximately adjusted to fast variable directions.

4.2 Fluctuation-Dissipation and Determining the Model. The prediction in the last section brings us back to the main question of this work: How can an underlying learning–theoretic computation be inferred? For many reasons, analysis of learning curves is not always a good idea. First, learning may happen so fast that resolving it might present a problem (Gallistel et al., 2001). Second, to estimate $\Lambda(N)$ reliably, we need to average, and a complete instance of the learning curve is just one sample. Such averaging may require prohibitively long experiments. Third, it is well known that animals adapt. Thus, eliciting the same response to the same target requires large intertrial time delays, further increasing the experimental duration. These problems can be traced to learning being an inherently transient behavior, and they might become less severe if we can characterize learning machines by some stationary response properties. A hint comes from the fluctuation–dissipation theorem in statistical physics (Ma, 1985), which states that if a system fluctuates in the presence of a linear dissipative restoring force, then the variance of fluctuations (a stationary property) is linearly related to the dissipation coefficient (a feature of the transient response). In our case, we may hope that response to a variable target (fluctuations) reveals information about the learning curve (dissipation).

In view of this suggestion, let us now analyze a few examples of a variable target learning.[7] We now denote by α an estimate of $\bar\alpha$ averaged over many presentations of the same data. We keep almost all parameters fixed (or changing very slowly), while $\bar\alpha_1(t)$, which is approximately the direction of the ridge in the density of solutions, is allowed to vary. If data are observed for a long time, then $\bar\alpha_\mu \approx \alpha_\mu$ for $\mu \neq 1$ (provided $\bar\alpha = \hat\alpha$). Now remember that Λ is the expected KL divergence between $\bar\alpha$ and α, which converges to the χ^2 distance when it is small. Thus, if $\bar\alpha_1 - \alpha_1$ is not large,

$$\Lambda \propto (\alpha_1 - \bar\alpha_1)^2. \tag{4.1}$$

---

[7] It is clear that stretching the theory of learning a fixed target to the fluctuating case may hide many potential pitfalls. We do this because we are unaware of any comprehensive treatments of the latter problem, though some progress is being made (cf. DeWeese & Zador (1998); Atwal & Bialek (2004).

For a fixed target and $\Lambda \to 0$ (that is, for $N \to \infty$, $\hat{\alpha} = \bar{\alpha}$), all learning curves we studied can be summarized as

$$\frac{\partial \Lambda}{\partial N} = -\zeta_N \Lambda^\nu. \tag{4.2}$$

Here, in particular, $\nu = 1$ corresponds to a finite set of solutions along the direction of α_1, $\nu = 2$ is the finite–parameter or nested case, and $\nu = 3$ is the $\eta = 1$ QFT model. In principle, other values of $\nu \in (0; \infty)$ are possible. The constant $\zeta_N \sim 1$ depends on the details of the learning setup. For example, for parametric cases, $\zeta_N = 2/K(\bar{r})$.

For equation 4.2, which is manifestly true for a fixed $\bar{\alpha}$, to also hold in the fluctuating target case, the learning machine must quickly notice the target's variation and disregard old samples as soon as they become outdated. Gallistel et al. (2001) show that a rat reacts to changes in the reward rates as fast an ideal detector would. Therefore, this assumption is reasonable for biological systems.[8]

If measurements are taken at a fixed rate, so that $dN/dt = $ const, we can combine equations 4.1 and 4.2 to get

$$\frac{d\Delta}{dt} = -\zeta \, \mathrm{sign}(\Delta) \, |\Delta|^{2\nu-1} - v_{\bar{\alpha}} \,, \tag{4.3}$$

where $\Delta = \alpha_1 - \bar{\alpha}_1$ is the average error of the estimation, ζ is some unknown constant with the dimensionality of $1/t$ and is basically the scaled sampling rate, and $v_{\bar{\alpha}}$ is the drift velocity of the target. Equation 4.3 is a clear example of a dissipative system, and it has many analogs in the theories of classical and quantum dissipation (Weiss, 1995). Note also that unlike in the fluctuation-dissipation analysis in statistical physics, the spectrum of fluctuations, $v_{\bar{\alpha}}$, is not necessarily white and can be controlled by an experimentalist, potentially providing more ways to probe the underlying dissipative dynamics.

If the target's variation cannot be learned (incomplete or mismatched machine), then equation 4.3 still holds. However, because of equation 3.8, we now have $\Delta = \alpha_1 - \hat{\alpha}_1$ (recall that $\hat{\alpha}$ is the best approximation to the target by a particular learning machine). Thus, to trace the evolution of α_1 using equation 4.3, one would need to evaluate $\hat{\alpha}(\bar{\alpha})$, which can be done from the stationary target analysis. Further, if the target varies along many learnable directions, then for each such direction, we have an analog of equation 4.3, possibly with different ξ. So the dynamics of Λ is still given

---

[8] We leave aside important comments by DeWeese and Zador (1998), who argued that time needed to notice a change may be not invariant with respect to the direction of the change.

by equation 4.2 with forcing, but the dissipation constant depends on the number of varying parameters.

Let us now consider a few different examples of $v_{\bar{\alpha}}$. If $v_{\bar{\alpha}} = \mathcal{A}$ is a constant, then asymptotically for $t \to \infty$, setting $d\Delta/dt = 0$, we find

$$\Delta \to \Delta_\infty = -\left(\frac{\mathcal{A}}{\zeta}\right)^{1/(2\nu-1)}. \tag{4.4}$$

The ratio $v_{\bar{\alpha}}/\zeta$ must be $\ll 1$; otherwise, Δ is outside the $\Lambda \to 0$ asymptotic, for which equation 4.2 is valid. Thus, for small drifts, setups with smaller ν win qualitatively.

It is also of interest to consider the situation when $\bar{\alpha}_1$ undergoes a Brownian motion, $\langle v_{\bar{\alpha}}(t)v_{\bar{\alpha}}(t')\rangle = \Omega\,\delta(t-t')$. Writing the Fokker-Planck equation for this Langevin dynamics, we easily find the stationary distribution of Δ,

$$P(\Delta) = \frac{\nu}{\Gamma\left(\frac{1}{2\nu}\right)}\left(\frac{\zeta}{\nu\Omega}\right)^{1/(2\nu)}\exp\left\{-\frac{\zeta|\Delta|^{2\nu}}{\nu\Omega}\right\}, \tag{4.5}$$

which results in the root mean square (rms) fluctuations

$$\Delta_{\text{rms}} = \left\{\nu^{1/\nu}\frac{\Gamma\left(\frac{3}{2\nu}\right)}{\Gamma\left(\frac{1}{2\nu}\right)}\right\}^{1/2}\left(\frac{\Omega}{\zeta}\right)^{1/(2\nu)}. \tag{4.6}$$

Again, these results are true only if $\Delta_{\text{rms}} \ll 1$, and again smaller ν provides for better trailing of the target.

Finally, inspired by Fairhall et al. (2001), let us examine the case of a periodic motion of $\bar{\alpha}_1$ and take, for simplicity, $\bar{\alpha}_1 = \mathcal{A}\sin\omega t$ and $v_{\bar{\alpha}} = \mathcal{A}\omega\cos\omega t$. Now equation 4.3 does not have a simple solution. However, we search for an asymptotically periodic $\Delta(t)$ with the same angular frequency of ω. Therefore, if we multiply equation 4.3 by $\cos\omega t$, integrate over a full period, and exchange the order of the differentiation and the integration, we get

$$\frac{d\langle\Delta\,\cos\omega t\rangle}{dt} = -\zeta\langle\text{sign}(\Delta)\,|\Delta|^{2\nu-1}\cos\omega t\rangle - \mathcal{A}\omega\langle\cos^2\omega t\rangle, \tag{4.7}$$

where $\langle\ldots\rangle$ denotes averaging over the period. Since we are looking for stationary oscillations, time derivative applied to any average is zero. This gives

$$\langle\text{sign}(\Delta)\,|\Delta|^{2\nu-1}\cos\omega t\rangle = -\frac{\mathcal{A}\omega}{2\zeta}. \tag{4.8}$$

Now multiplying equation 4.3 by $\mathrm{sign}(\Delta)\,\Delta^{2\nu-1}$ and averaging again results in

$$\langle |\Delta|^{4\nu-2} \rangle = \frac{(\mathcal{A}\omega)^2}{2\zeta^2}, \tag{4.9}$$

which is the same scaling as in equation 4.4. However, now we also have a dependence on ω.

There are other cases that can be analyzed, such as a step jump in the target, $\bar{\alpha}_1$, its square wave modulation, or its diffusion in a potential (Ornstein-Uhlenbeck process). Interestingly, the last two of these cases were used experimentally by Fairhall et al. (2001). However, we leave the analysis for the future, when it will be answering some specific question and will not be just a mathematical exercise. Even with the three examples already discussed, it is clear that letting the target move maps the scaling of the learning curve into a stationary property (e.g., variance of the estimation error), which might be easier to analyze experimentally.

5 Discussion

We have shown that with a moving target, transient learning curves are replaced by different scaling dependences of the estimation errors on the amplitude of the target's motion. This effect is stationary and may be easier to observe experimentally. However, since we do not have a comprehensive theory of variable target learning yet, a few precautions are in order when designing and analyzing experiments along these lines:

1. Target velocities must be kept small so that the asymptotic analysis presented in this work holds.

2. The analysis is valid only when the learner forgets past observations as soon as the target changes appreciably. Learning will be much slower if such outdated samples are kept.

3. When varying the stimulus, we have to be reasonably sure that the animal only tracks it, and does not predict it. White noise $v_{\bar{\alpha}}$ or a multiparameter representation of the target in terms of the position, velocity, or acceleration, for example, might be a solution.

4. Finally, we have to keep in mind that in a behaving animal, learning a change in a signal and reacting to it may be separated by a long delay, and special care is needed to observe the former, but not the latter.

This being said, it nevertheless is possible that all these and other disadvantages will be outweighed by the ability to determine the correct learning–theoretic model of the organism by varying the amplitude and the nature

(say, stochastic or periodic) of the target's motion and studying typical responses as functions of these parameters.

Consider, for example, the experiment described in Figure 4 of Fairhall et al. (2001). There, the input signal (the standard deviation of the angular velocity, $\sigma(t)$) undergoes a finite variance Σ^2 and a finite correlation time τ random motion. The instantaneous neuron firing rate $r(t)$ is the estimate of $\sigma(t)$. Repeating exactly the same randomly generated stimulus many times and averaging over spike trains, one may estimate $r(t)$ and, consequently, the rms estimation error $\Delta_{rms} = \langle (\sigma(t) - r(t)^2)_t \rangle^{1/2}$. Studying dependence of Δ_{rms} on Σ and τ along the lines of equation 4.6, one can estimate ν. Any $\nu \neq 2$ uniquely determines the underlying computational model. For $\nu = 2$, to distinguish a usual finite parameter model from the one that is nested, one makes the signal multidimensional (other parameters of the angular velocity, such as the mean and the skewness, vary together with σ). For at least some signal extensions, the nested model will change the magnitude (but not the scaling) of Δ_{rms} since $\zeta \propto 1/K(\bar{r})$. In contrast, the simpler model will keep the same prefactor but will be converging only to an approximation of the target.

In the cognitive experiments of Gallistel et al. (2001), a rat was trying to learn reward rates on different terminals and match its foraging habits correspondingly. It was determined to be an ideal change detector. To build a more detailed model of the animal, one can vary the reward rates continuously, repeat experiments many times, and then look at the average mismatch between the stimulus and the response. Then dependence of the mismatch on the parameters of the rate changes will point to a proper class of learning–theoretic models to compare the rat to. Similarly, one can do this type of analysis on artificial neural networks designed explicitly to model particular animal behavior (Seung, 2003); this will build connections between network architectures and types of inference tasks performed by them.

Another conclusion of our work is that the nested setup may learn faster than the QFT one under some conditions. Thus, if one desires a complete learning machine, a nested machine should be built unless there is some specific reason to do the opposite (such as knowing that the world is unlikely to have sharp cutoffs). With experiments along the lines suggested above, this prediction should be testable. We should be able to see if our intuitive beliefs about appropriate complexities of learners for particular tasks match nature's choices. It would also be interesting to study if structural characteristics of a learner are correlated with its learning-theoretic description. That is, could it be that modular, irregular networks, like those seen in biochemistry, often compute like parametric or nested machines? And could layered, regular networks in our brains, which are believed to be able to solve the most complicated learning problems, be realizing QFT machines instead?

Appendix A: Model Family Selection in the Nested Setup ⎯⎯⎯⎯

Inference in Bayesian i.i.d. setup is quite standard (Press, 1989; Bialek et al., 1996; Balasubramanian, 1997; Bernardo, 2003; Raftery & Zheng, 2003), and the nested case is not very different. For example, a posteriori expectations of parameter values are given by a derivative of the posterior moment generating function (or the partition function), $Z(\mathbf{J})$:

$$\langle \alpha_\mu \rangle = \left. \frac{\partial}{\partial J_\mu} \right|_{\mathbf{J}=0} \log Z(\mathbf{J}), \tag{A.1}$$

$$Z(\mathbf{J}) \equiv \int d\boldsymbol{\alpha} \, \mathcal{P}(\boldsymbol{\alpha}) \, e^{-\mathcal{L}(\boldsymbol{\alpha})+\mathbf{J}\cdot\boldsymbol{\alpha}} \tag{A.2}$$

$$= \sum_r \mathcal{P}(r) Z_r(\mathbf{J}) r, \tag{A.3}$$

$$Z_r(\mathbf{J})r \equiv \int d^{K(r)}\alpha \, e^{-\mathcal{L}_r(\boldsymbol{\alpha})+\sum_{\mu=1}^{K(r)} J_\mu \alpha_\mu}, \tag{A.4}$$

$$\mathcal{L}(\boldsymbol{\alpha}) \equiv \sum_{i=1}^{N} \phi(x_i|\boldsymbol{\alpha}), \tag{A.5}$$

$$\mathcal{L}_r(\boldsymbol{\alpha}) \equiv -\sum_{\mu=1}^{K(r)} \log p(\alpha_\mu) + \sum_{i=1}^{N} \phi(x_i|\boldsymbol{\alpha}), \tag{A.6}$$

$$\phi(x|\boldsymbol{\alpha}) \equiv -\log Q(x|\boldsymbol{\alpha}). \tag{A.7}$$

The posterior expectations are thus determined by the properties of the $Z(\mathbf{J})$, which can be calculated using the saddle point analysis for $N \gg 1$. This is difficult for the first form of $Z(\mathbf{J})$, equation A.2, due to the singularity at $\alpha_\mu = 0$ (the singularity was also the reason that we left $\mathcal{P}(\boldsymbol{\alpha})$ out of the combined Lagrangian, equation A.5). Hence, we return to the nested form, equations A.3 and A.4, but the equivalence between the representation should be kept in mind. Exchanging the order of integration and summation in equations A.2 and A.3 and similar is possible if the priors decay sufficiently fast at $r \to \infty$, or are regularized with regularization lifted after averages are calculated. Unless mentioned otherwise, this is always assumed.

The expectation of α_μ in the model families with $K(r) < \mu$ is necessarily zero, and a similar bias toward smaller magnitudes of parameters will be present when we average over families. Therefore, the a priori decrease of the variances with μ, equations 3.15 and 3.18, will persist a posteriori for finite N. This is the famous James and Stein (1961) shrinkage.

The saddle point, also called classical or maximum likelihood, values of parameters in each family, $\boldsymbol{\alpha}_r^* \equiv \{\alpha_{\mu;r}^*\}$, and the second derivatives matrix

at the saddle, $\mathbb{F}_r$, are determined by (remember that $\alpha^*_{\mu;r} \equiv 0$ for $\mu > K(r)$)[9]

$$\left.\frac{\partial \mathcal{L}_r(\boldsymbol{\alpha})}{\partial \alpha_\mu}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*_r} = 0\,,\ \mu \leq K(r)\,, \tag{A.8}$$

$$\left.\frac{\partial^2 \mathcal{L}_r(\boldsymbol{\alpha})}{\partial \alpha_\mu \partial \alpha_\nu}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*_r} = \mathbb{F}^{\mu\nu}_r\,,\ \mu, \nu \leq K(r). \tag{A.9}$$

To the first order in $1/N$, this gives

$$Z(\mathbf{J}) = \sum_r \mathcal{P}(r) \frac{\mathcal{P}(\boldsymbol{\alpha}^*_r|r)(2\pi)^{K(r)/2}}{N^{K(r)/2} \det^{1/2} \frac{\mathbb{F}_r}{N}} Q(\{x\}|\boldsymbol{\alpha}^*_r) \mathrm{e}^{\frac{1}{2}\mathbf{J}_r \mathbb{F}_r^{-1}\mathbf{J}_r + \mathbf{J}_r \cdot \boldsymbol{\alpha}^*_r}\,, \tag{A.10}$$

where $\mu \leq K(r)$ components of $\mathbf{J}_r$ are the same as those of $\mathbf{J}$, and all higher-order components are zero. Differentiating, we get:

$$\langle \alpha_\mu \rangle = \frac{\sum_{r=1}^R \alpha^*_{\mu;r} \mathrm{e}^{-\mathcal{L}(r)}}{\sum_{r=1}^R \mathrm{e}^{-\mathcal{L}(r)}}, \tag{A.11}$$

$$\mathcal{L}(r) \equiv -\log \mathcal{P}(r) - \sum_{\mu=1}^{K(r)} \log p(\alpha^*_{\mu;r})$$

$$+ \sum_{i=1}^N \phi(x_i|\boldsymbol{\alpha}^*_r) + \frac{K(r)}{2} \log \frac{N}{2\pi} + \mathrm{Tr} \log \frac{\mathbb{F}_r}{N}. \tag{A.12}$$

For finite R and $\beta = 0$, this is the usual Bayesian model family selection: a posteriori expectations are weighted sums over posterior probabilities of families defined by $\mathrm{e}^{-\mathcal{L}(r)}$. This posterior includes the negative maximum likelihood term, $\sum_{i=1}^N \phi(x_i|\boldsymbol{\alpha}^*_r)$, which grows in magnitude linearly with N but decreases as r grows due to nestedness. It also incorporates the fluctuation determinant $\frac{K(r)}{2} \log \frac{N}{2\pi} + \mathrm{Tr} \log \frac{\mathbb{F}_r}{N}$, which grows logarithmically in N but increases with r. Depending on the value of N, there will be some r^*, for which $\mathcal{L}(r)$ is minimal. For large N, as a discrete analog of the saddle point argument, this value will dominate the sums in equation A.11, and hence some model family will be "selected."

However, equations A.11 and A.12 become more interesting if one lets $R \to \infty$. The completeness condition ensures that for large enough r, one will be overfitting the data, and $Q(x|\boldsymbol{\alpha}^*_r) \to 1/N \sum \delta(x - x_i)$. Therefore, if

---

[9] There are possibilities of more than one saddle point and of other anomalies. This was analyzed by Bialek et al. (2001). The conditions to prevent such problems are mild, and we assume them to hold in what follows.

the sums are dominated by $r \to \infty$, then consistency breaks and the learning fails. One would thus expect two features to influence the success of the learning. First, it is the prior $\mathcal{P}(r)$, which switches on extra degrees of freedom: for slowly decaying priors, one would expect $r \to \infty$ terms to win. Second, it is the dependence of the likelihood term on r, which measures how capable are the newly activated degrees of freedom of overfitting, or, equivalently, how fast $\max_i Q(x_i | \boldsymbol{\alpha}_r^*)$ grows.

From equation A.12 it is easy to see that large r will have an exponentially small weight in the posterior probability if

$$\lim_{r \to \infty} \frac{N \max_i \log Q(x_i | \boldsymbol{\alpha}_r^*) + \log \mathcal{P}(r)}{K(r) \log N} = 0. \tag{A.13}$$

Under this condition, $Q(x | \boldsymbol{\alpha}^*)$ will eventually approach the correct distribution but not the sum of δ–functions. Colloquially, equation A.13 requires the explanatory capacity of the new, high-order degrees of freedom to be small enough so that keeping them always "on" does not make sense. This criterion, which we have not seen explicitly presented anywhere before, is similar to the consistency condition of the structural risk minimization (SRM) theory, which requires that the Vapnik–Chervonenkis dimension, the SRM capacity measure of the selected model, grows more slowly than the number of samples to be explained (Vapnik, 1998; Nemenman, 2000).

As an example, let us analyze how the condition in equation A.13 may be violated for $K(r) \sim r$. In this case, $\mathcal{P}(r)$ must be superexponential to be relevant for finding r^*. Thus, it is not required to decay at some minimal speed as might have been expected, though a need to exchange the order of integrations and summations in arriving at equation A.12 may still force that. Due to light tails and small effective support, exponentially decaying priors are not very interesting, so we disregard the prior term in equation A.13. Then for a fixed large N, a finite r^* will be dominant if $\log Q(x_i | \boldsymbol{\alpha}_r^*)/r \to 0$. That is, the growth of the δ function-like peaks of the maximum likelihood distribution should be superlinear in K, the number of parameters in the model family, in order for $r^* \to \infty$ and Bayesian setup to be inconsistent.

Appendix B: Fourier Polynomials Nested Model

To compare nonparametric and finite parameter nested scenarios directly, we analyze the following example. Consider families of probability distributions periodic on $[0, 1]$ and with the logarithms of the distributions given by Fourier polynomials of degree $r < \infty$, as in equation 3.28. Due to the normalization condition, equation 3.29, the number of parameters in the rth model family is $K(r) = 2r$. With an appropriate choice of priors, equations 3.9 and 3.10, these families form a nested set, and the completeness for $R \to \infty$ follows from the Fourier theorem.

The classical solution for this parameterization is ($1 < \mu \le r$)

$$\frac{\alpha_\mu^{*\pm}}{\sigma_\mu^2} + \sum_i \binom{\cos}{\sin} 2\pi\mu x_i - N \int dx \, Q(x|\boldsymbol{\alpha}^*) \binom{\cos}{\sin} 2\pi\mu x \equiv$$

$$\frac{\alpha_\mu^{*\pm}}{\sigma_\mu^2} + \frac{N}{2}\Delta_\mu^\pm - \frac{N}{2}Q_\mu^{*\pm} = 0 \,. \tag{B.1}$$

Here, $Q_\mu^\pm$ are the cosine (sine) amplitudes of the μth mode in the Fourier expansion of $Q(x|\boldsymbol{\alpha})$, and $\Delta_\mu^\pm$ are the same for the empirical probability density, $1/N \sum \delta(x - x_i)$. $\Delta_\mu^\pm$ are also the stochastic Fourier transform of $Q(x)$. The cosine–cosine components of the second derivative matrix at the saddle point are

$$\left.\frac{\partial^2 \mathcal{L}}{\partial \alpha_\mu^+ \partial \alpha_\nu^+}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*} = \frac{\delta_{\mu\nu}}{\sigma_\mu^2} + N \int dx \, Q(x|\boldsymbol{\alpha}^*) \cos 2\pi\mu x \, \cos 2\pi\nu x$$

$$- N \int dx \, Q(x|\boldsymbol{\alpha}^*) \cos 2\pi\mu x \int dy \, Q(y|\boldsymbol{\alpha}^*) \cos 2\pi\nu y \quad \text{(B.2)}$$

$$= \frac{\delta_{\mu\nu}}{\sigma_\mu^2} + \frac{N}{4}\left(Q_{\mu+\nu}^{*+} + Q_{\mu-\nu}^{*+}\right) + \frac{N}{2}Q_\mu^{*+}Q_\nu^{*+} \,, \tag{B.3}$$

and the sine-sine and the sine-cosine components are written similarly. This matrix is provably positive definite. Thus, for $N \to \infty$, we can perform the saddle point analysis.

For $\beta = 0$, the variance σ_μ^2 is constant, and we can neglect the first term in equation B.1 in the limit of large N. This leads to the following solution of the saddle point equations:

$$Q_\mu^{*\pm} \approx \Delta_\mu^\pm, \qquad \mu = 1, \ldots, r. \tag{B.4}$$

For $\beta > 0$, $Q_\mu^{*\pm}$ will be corrected by a systematic β-dependent bias, which will tend to 0 for fixed μ as N grows. This will decrease the posterior variance of the estimator Q^*.

Equation B.4 says that the first r pairs of coefficients $\alpha_\mu^{*\pm}$ are such that the corresponding Fourier amplitudes of the classical solution Q^* match those of the empirical one. By Nyquist theorem and the law of large numbers, for $r < N/2$, $\Delta_\mu^\pm$ approach the Fourier amplitudes of the unknown target probability density $\bar{Q}$. Thus, the low-frequency modes will be learned well. However, if $r > N/2$, the saddle point solution will start to overfit and develop δ-like spikes at each observed data point. This is in accord with the observation we have already mentioned: to guarantee consistency, the capacity of models, as measured by either the VC dimension or the scaling dimension of Bialek

et al. (2001), which in this case is equal to the number of free parameters, must grow slower than N.

To avoid overfitting when averaging over r, we must make sure that the contribution of $r \to \infty$ to the posterior log probability, equation A.12, is negligible. In this regime, according to equation B.4, the r available modes will create peaks of height $\sim r$ (recall the Fourier expansion of the δ–function) at the observed sample points. With $K(r) = 2r$, this ensures consistency by satisfying equation A.13.

Further, we can prove that r^* is not only finite, but actually grows sublinearly in N, again paralleling results for SRM (Vapnik, 1998) and their Bayesian equivalent (Nemenman, 2000). Suppose $r \gg N$ dominates the posterior. Then, for a slowly decaying $\mathcal{P}(r)$, equation A.12 can be rewritten as

$$\mathcal{L}(r) \sim -N \log r + r \log N. \tag{B.5}$$

This is minimized (and the posterior probability is maximized) for $r^* \sim \frac{N}{\log N}$, and higher values of r are exponentially inhibited. Thus, the assumption of $r \gg N$ being dominant is incorrect, and the posterior probability is dominated by $r^* \sim N$ for all reasonable priors. This is, of course, the worst-case estimation, and in many typical applications the value of r^* is even lower.

Appendix C: Fourier Nested Model and QFT Targets

As shown above, r^* that minimizes equation A.12 for a Fourier nested setup is much smaller than N. This is true for any target, including QFT–typical targets. For r of such magnitude, the first r modes of the target are well approximated by the estimate, and they contribute $O(r/N)$ to the leading data-dependent term in equation A.12. The modes of the target above the rth are not fitted by the estimate, and each of them contributes its variance of about $\ell^{-2\eta+1}\mu^{-2\eta}$ to the data-dependent term, adding up to $\sum_{\mu=r+1}^{\infty} \ell^{-2\eta+1}\mu^{-2\eta} \propto (r\ell)^{-2\eta+1}$. Combined with the fluctuation determinant, this gives

$$\mathcal{L}(r) \sim -N(r\ell)^{-2\eta+1} + r \log N \tag{C.1}$$

for determining the most probable r. Thus, for $N \gg 1$ (or $\Lambda \ll 1$),

$$r^* \propto \left(\frac{N}{\log N} \right)^{1/2\eta} \ell^{1/2\eta - 1}, \text{ and} \tag{C.2}$$

$$\mathcal{D} \propto N^{1/2\eta} \left(\frac{\log N}{\ell} \right)^{1 - 1/2\eta}. \tag{C.3}$$

$$\Lambda \sim \left(\frac{\log N}{N\ell} \right)^{1-1/2\eta}. \tag{C.4}$$

Due to the many simplifications made here, the exact form of the logarithmic terms in these expression is questionable,[10] and, in practice, they are impossible to observe for realistic N due to the target-dependent prefactors in front of the universal scaling term and various statistical fluctuations. However, the power law in equation C.4, which is definitely correct, suggests that the performance of the nested model is comparable to that of the true QFT one. In particular, the nested learning machine also can solve arbitrarily complex inference problems.

A rigorous way to estimate performance of the nested learning on a nonparametric target is to calculate $\langle \rho(\epsilon) \rangle = \int d\bar{\alpha}\, \mathcal{P}(\bar{\alpha})\rho(\epsilon; \bar{\alpha})$, where ρ is of the form equation 3.19, and the averaging is done over the QFT prior, and then calculate $\mathcal{D}$ and Λ from this averaged ρ. This is difficult, and instead we may choose to replace $\langle \rho(\epsilon) \rangle$ by $\rho(\epsilon; \bar{\alpha}_{\text{typ}})$, where $\bar{\alpha}_{\text{typ}}$ is a typical target in the nonparametric prior, equation 3.22.[11] For such $\bar{\alpha}_{\text{typ}}$, $D_r(\bar{\alpha}) \sim \sum_{\mu > r} \ell^{-2\eta+1} \mu^{-2\eta} \propto (\ell r)^{-2\eta+1}$. Further, $\mathcal{P}(\hat{\alpha}^{\pm}_{\mu;\text{typ}} | r) \sim \exp[-0.5\, \mu^{-2\eta} \ell^{-2\eta+1} / \sigma_\mu^2]$. In our case, $\sigma_\mu^2 \sim \mu^{-2\beta}$. Therefore, for $2(\eta - \beta) > 1$, which is satisfied for $\eta > 1/2$ and $\beta = 0$, this gives $\mathcal{P}(\hat{\alpha}_r | r) \sim \exp\left[-\sum_{\mu=1}^{r} \mu^{-2\eta} \ell^{-2\eta+1} / \sigma_\mu^2 \right] \sim \exp\left[C_1 - C_2 r^{-2(\eta-\beta)+1} \right]$, where C_1 and C_2 are constants. For large enough r, this whole expression tends to a constant. Thus, combining with equation 3.19, we get

$$\rho(\epsilon; \bar{\alpha}_{\text{typ}}) \sim \sum_{r:\, r^{-2\eta+1} \le \epsilon} \mathcal{P}(r) \frac{2\pi^r}{\Gamma(r)} \frac{[\epsilon - (\ell r)^{-2\eta+1}]^{(r-1)}}{\sqrt{\det \mathcal{F}_{K(r)}}}. \tag{C.5}$$

---

[10] If a certain derivative of the target distribution satisfies some Lipschitz conditions, then the Occam factor and the learning curve for histogramming density estimators provably have logarithmic contributions (Hall & Hannan, 1988; Rissanen et al., 1992). In contrast, logarithmic corrections for QFT models and for parametric learning of QFT–typical targets have not yet been analyzed. However, the logarithmic differences between the cases have been expected: in discrete case, $\beta = 0$, once we know $K^* \sim N^\omega$, each of K^* parameters is free to vary with the same variance, giving familiar $N^\omega \log N$ fluctuations. For the nonparametric case, $\sigma_\mu < \sigma_\nu$ for $\mu > \nu$. Thus, each next parameter varies less, somewhat decreasing the total fluctuations (Bialek et al., 2001).

These logarithmic terms have the same roots as the difference between cross-validation, bootstrap, and Akaike's model selection criterion, on one hand, and Dawid's prequential statistics and Bayesian model selection, on the other (Stone, 1977; Dawid, 1984). There, the difference in the magnitude of the prediction error is also due to most of the parameters that are active at a given N being latent for smaller sample sizes.

[11] The benefit $\langle \rho(\epsilon) \rangle$ provides over $\rho(\epsilon; \bar{\alpha}_{\text{typ}})$ is knowing the prefactors in $\mathcal{D}$ and Λ. We do not believe that any of the priors studied in this work will be exactly realized in nature. Therefore, calculation of $\langle \rho(\epsilon) \rangle$ is not a priority.

If $\mathcal{P}(r)$ is subexponential as before, we get ρ to the leading order in small ϵ by calculating the sum in equation C.5 using the saddle point analysis and taking just the zeroth-order term. The saddle value for r is $r^* \sim \epsilon^{-1/(2\eta-1)}\ell^{-1}$, which gives

$$\rho(\epsilon; \bar{\boldsymbol{\alpha}}_{\text{typical}}) \sim \epsilon^{\epsilon^{-1/(2\eta-1)}\ell^{-1}}, \tag{C.6}$$

with the first subleading term of $O\left(\exp\left[-\epsilon^{-1/(2\eta-1)}\ell^{-1}\right]\right)$. Doing the leading-order evaluation of the integral in equation 2.2, we now get $\epsilon^* \sim \left(N\ell/\log N\right)^{1/2\eta-1}$, which again results in equation C.3.

In summary, learning a distribution typical in the nonparametric model by means of the nested setup results in, at most, a logarithmic performance loss.

Acknowledgments

References

Aida, T. (1999). Field theoretical analysis of on-line learning of probability distributions. *Phys. Rev. Lett.*, *83*, 3554–3557.

Atick, J. (1992). Could information theory provide an ecological theory of sensory processing? In W. Bialek (Ed.), *Princeton Lectures on Biophysics* (pp. 223–289). Singapore: World Scientific.

Attneave, F. (1954). Some informational aspects of visual perception. *Psych. Rev.*, *61*, 183–193.

Atwal, G., & Bialek, W. (2004). Ambiguous model learning made unambiguous with 1/f priors. In S. Thrun, L. Saul, & B. Scholkopf (Eds.), *Advances in neural information processing systems*, *16*. Cambridge, MA: MIT Press.

Balasubramanian, V. (1997). Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neur. Comp.*, *9*, 349–368.

Barlow, H. (1959). Sensory mechanisms, the reduction of redundancy and intelligence. In D. Blake & A. Uttley (Eds.), *Proc. Symp. Mechanization of Thought Processes* (Vol. 2, pp. 537–574). London: H. M. Stationery Office.

Barlow, H. (1961). Possible principles underlying the transformation of sensory messages. In W. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Bernardo, J. (2003). Bayesian statistics. In *UNESCO Encyclopedia of Life Support Systems (EOLSS)*. Oxford: EOLSS Publishers.

Bialek, W., Callan, C., & Strong, S. (1996). Field theories for learning probability distributions. *Phys. Rev. Lett.*, *77*, 4693–4697.

Bialek, W., Nemenman, I., & Tishby, N. (2001). Predictability, complexity, and learning. *Neur. Comp.*, *13*, 2409–2463.

Bialek, W., & de Ruyter van Steveninck, R. R. (2005). *Features and dimensions: Motion estimation in fly vision*. Manuscript submitted for publication.

Brenner, N., Bialek, W., & de Ruyter van Steveninck, R. (2000). Adaptive rescaling maximizes information transmission. *Neuron*, *26*, 695–702.

Clarke, B., & Barron, A. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inf. Thy.*, *36*, 453–471.

Cohn, M., & Horibata, K. (1959). Inhibition by glucose of the induced synthesis of the β-galactoside-enzyme system of *Escherichia coli*: Analysis of maintenance. *J. Bacteriol.*, *78*, 601–612.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York: Wiley.

Csiszar, I. (1975). I–divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, *3*(1), 146–158.

Cucker, F., & Smale, S. (2001). On the mathematical foundations of learning. *Bull. Amer. Math. Soc. n.s.*, *39*(1), 1–49.

Dawid, A. (1984). Present position and potential developments: Some personal views. Statistical theory: The prequential approach. *J. Roy. Stat. Soc. A*, *147* (p. 2), 278–292.

Detwiler, P., Ramanathan, S., Sengupta, A., & Shraiman, B. (2000). Engineering aspects of enzymatic signal transduction: Photoreceptors in the retina. *Biophys. J.*, *79*, 2801–2817.

DeWeese, M., & Zador, A. (1998). Asymmetric dynamics in optimal variance adaptation. *Neur. Comp.*, *10*, 1179–1202.

Fairhall, A., Lewen, G., Bialek, W., & de Ruyter van Steveninck, R. (2001). Efficiency and ambiguity in an adaptive neural code. *Nature*, *412*, 787–792.

Gallistel, C., Mark, T., King, A., & Latham, P. (2001). The rat approximates an ideal detector of changes in rates of reward: Implications for the law of effect. *J. Exper. Psych.: Animal Behav. Proc.*, *27*, 354–372.

Hall, P., & Hannan, E. (1988). On stochastic complexity and nonparametric density estimation. *Biometrika*, *75*(4), 705–714.

Holy, T. (1997). Analysis of data from continuous probability distributions. *Phys. Rev. Lett.*, *79*, 3545–3548.

James, W., & Stein, C. (1961). Estimation with quadratic loss. In J. Neyman (Ed.), editor, *Proc. Fourth Berkeley Symposium Mathematical Statistics and Probability* (Vol. 1, pp. 361–379). Berkeley: University of California Press.

Janes, E. (1979). Inference, method, and decision: Towards a Bayesian philosophy of science. *J. Amer. Stat. Assoc.*, *74*(367), 740–741.

Jeffreys, H. (1936). Further significance tests. *Proc. Camb. Phil. Soc*, *32*, 416–445.

Kording, K., & Wolpert, D. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244–247.

Lemm, J. (2002). *Bayesian field theory*. Baltimore, MD: Johns Hopkins University Press.

Ma, S. (1985). *Statistical mechanics*. Singapore: World Scientific.

MacKay, D. (1992). Bayesian interpolation. *Neur. Comp.*, *4*, 415–448.

Nemenman, I. (2000). *Information theory and learning: A physical approach*. Unpublished doctoral dissertation, Princeton University.

Nemenman, I., & Bialek, W. (2002). Occam factors and model-independent Bayesian learning of continuous distributions. *Phys. Rev. E*, *65*, 026137.

Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W. (1996). *Applied linear regression models* (3rd ed.). Chicago: Irwin.

Press, S. (1989). *Bayesian statistics: Principles, models, and applications*. New York: Wiley.

Raftery, A., & Zheng, Y. (2003). Discussion: Performance of Bayesian model averaging. *J. Amer. Stat. Assoc.*, *98*(464), 931–938.

Rao, R. (2004). Bayesian computation in recurrent neural circuits. *Neur. Comp.*, *16*(1), 1–38.

Reichardt, W. (1961). Autocorrelation, a principle for the evaluation of sensory information by the central nervous system. In W. Rosenblith (Ed.), *Principles of sensory communication* (pp. 303–317). New York: Wiley.

Rissanen, J. (1989). *Stochastic complexity and statistical inquiry*. Singapore: World Scientific.

Rissanen, J., Speed, T., & Yu, B. (1992). Density estimation by stochastic complexity. *IEEE Trans. Inf. Thy.*, *38*(2), 315–323.

Samko, S., Kilbas, A., & Marichev, O. (1987). *Integraly i proizvodnye drobnogo poriadka i nekotorye ikh prilozheniia*. Minsk, Belarus. Nauka i tekhnika. (In Russian)

Schwartz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, *6*, 461–464.

Seung, H. (2003). Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron*, *40*, 1063–1073.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Stat. Soc. B*, *39*(1), 44–47.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Weiss, U. (1995). *Quantum dissipative systems* (2nd ed.). Singapore: World Scientific.

Wolpert, D. (1995). On the Bayesian "Occam factors" argument for Occam's razor. In T. Petsche T (Ed.), *Computational learning and natural learning systems* (Vol. 3). Cambridge, MA: MIT Press.