*LETTER* Communicated by Naftali Tishby

# Asymptotic Theory of Information-Theoretic Experimental Design

**Liam Paninski**
*liam@gatsby.ucl.ac.uk*
*Gatsby Computational Neuroscience Unit,*
*University College London, London, WC1N 3AR, U.K.*

**We discuss an idea for collecting data in a relatively efficient manner. Our point of view is Bayesian and information-theoretic: on any given trial, we want to adaptively choose the input in such a way that the mutual information between the (unknown) state of the system and the (stochastic) output is maximal, given any prior information (including data collected on any previous trials). We prove a theorem that quantifies the effectiveness of this strategy and give a few illustrative examples comparing the performance of this adaptive technique to that of the more usual nonadaptive experimental design. In particular, we calculate the asymptotic efficiency of the information-maximization strategy and demonstrate that this method is in a well-defined sense never less efficient—and is generically more efficient—than the nonadaptive strategy. For example, we are able to explicitly calculate the asymptotic relative efficiency of the staircase method widely employed in psychophysics research and to demonstrate the dependence of this efficiency on the form of the psychometric function underlying the output responses.**

## 1 Introduction

Many experiments are undertaken with the hope of elucidating some kind of "input-output" relationship: the experimenter presents some stimulus to the system under study and records the response. More generally, the experimenter places some observational apparatus in some state—for example, by pointing a microscope to a given location or selecting some subfield in a database stream—and records the subsequent observation. If the system is simple enough and a sufficient number of observations are made, the resulting collection of data should provide an acceptably precise description of the system's overall behavior.

Given this basic paradigm, in which the experimenter has some kind of control over what stimulus is chosen or what kind of data are collected, how do we design experiments to be as efficient as possible? How can we learn the most about the system under study in the least amount of time? This question becomes especially pressing in the context of high-dimensional,

 

complex systems, where each input-output pair typically provides a small amount of information about the behavior of the system as a whole and opportunities to record responses are rare or expensive (or both). In such cases, good experimental design can play an essential role in making the benefits of the experiment worth the cost.

How can we precisely define this intuitive concept of the efficiency of an experiment? First, we have to define what exactly we mean by *experiment*. We use the following simple model of experimental design here (we have neurophysiological experiments in mind, but our results are all general with respect to the identity of the system under study). The basic idea is that we have some set of models $\Theta$, where each model $\theta$ indexes a given probabilistic input-output relationship. More precisely, a model is a set of regular conditional probability distributions $p(y|x, \theta)$ on $Y$, the set of possible output responses, given any input stimulus $x$ in some space $X$. Therefore, if we know the identity of the model $\theta$, we know the probability of observing any output $y$ given any input $x$. Of course, we do not know $\theta$ exactly (otherwise we would not need to perform any experiments); our knowledge of the system is summarized in the form of a prior probability measure, $p_0(\theta)$, on $\Theta$, and our goal is to reduce the uncertainty of this distribution as much as possible. To put everything together, the joint probability of $\theta$, $x$, and $y$ is given by the following simple equation:

$$p(x, y, \theta) = p_0(\theta)\, p(x)\, p(y|x, \theta).$$

Now we can define the "design" of our experiment in a straightforward way: on any given trial, the design is specified completely by the choice of the input probability $p(x)$, the only piece of the above equation over which we have control. One common approach is to fix some $p(x)$ at the beginning of the experiment and then sample from this distribution in an independent and identically distributed (i.i.d.) manner for all subsequent trials, independently of which input-output pairs might have been observed on any previous trial. Alternatively, we could try to design our experiment—choose $p(x)$—optimally in some sense, updating $p(x)$ online, on each trial, as more input-output data are collected and our understanding of the system increases. (The simplest special case of this would be to choose $p(x)$ to put all probability mass on a single $x$, where $x$ is optimized on each new trial.) One natural idea would be to choose $p(x)$ in such a way that we learn as much as possible about the underlying model, on average. Information theory (Cover & Thomas, 1991) thus suggests we choose $p(x)$ to optimize the following objective function,

$$I(\{x, y\}; \theta), \tag{1.1}$$

where $I(.;.)$ denotes mutual information. In other words, we want to choose $p(x)$ adaptively to maximize the information provided about $\theta$ by the pair $\{x, y\}$, given our current knowledge of the model as summarized in the posterior distribution given $N$ samples of data:

$$p_N(\theta) = p(\theta|\{x_i, y_i\}_{1 \le i \le N}).$$

We will take this information-theoretic concept of efficiency as our starting point. We note, however, that similar ideas have seen application in a wide and somewhat scattered literature: in statistics (Lindley, 1956), computer vision (Denzler & Brown, 2000; Lee & Yu, 1999), machine learning (Luttrell, 1985; Mackay, 1992; Cohn, Ghahramani, & Jordan, 1996; Sollich, 1996; Freund, Seung, Shamir, & Tishby, 1997; Axelrod, Fine, Gilad-Bachrach, Mendelson, & Tishby, 2001), conceptual psychology (Nelson & Movellan, 2000), psychophysics (Watson & Pelli, 1983; Pelli, 1987; Watson & Fitzhugh, 1990; Kontsevich & Tyler, 1999), medical applications (Parmigiani, 1998; Parmigiani & Berry, 1994), and neuroscience (Sahani, 1997). These references all discuss, to some degree, the motivation behind various different design criteria, of which the information-theoretic criterion is well motivated but certainly not unique. For more general reviews of the theory of experimental design, see, for example, Chaloner and Verdinelli (1995) and Fedorov (1972). In addition, several attempts have been made to devise algorithms to find the "optimal stimulus" of a neuron, where optimality is defined in terms of firing rate (Tzanakou, Michalak, & Harth, 1979; Nelken, Prut, Vaadia, & Abeles, 1994; Foldiak, 2001), but we should emphasize that the two concepts of optimality are not related in general and turn out to be typically at odds (maximizing the firing rate of a cell does not maximize—and in fact often minimizes—the amount we can expect to learn about the cell; see sections 3 and 4). Most recently, Machens (2002) proposed the maximization of the mutual information between the stimulus $x$ and response $y$; again, though, this procedure does not directly maximize the amount of information we gain about the underlying system $\theta$.

Somewhat surprisingly, we have not seen any applications of the information-theoretic objective function, equation 1.1, to the design of neurophysiological experiments (although see the abstract by Mascaro & Bradley, 2002, who seem to have independently implemented the same idea in a simulation study). One major reason for this might be the computational demands of this kind of design (particularly for real-time applications), although these problems certainly do not appear to be intractable given modern computing power (see, e.g., Kontsevich & Tyler, 1999, for a real-time application in which $\Theta$ is two-dimensional). We hope to address these important computational questions elsewhere.

The primary goal of this letter is to elucidate the asymptotic behavior of the a posteriori density $p_N$ when we choose $x$ according to the recipe outlined above; in particular, we want to compare the adaptive case to

the more usual (i.i.d. $x$) case. Our main result (in section 2) states that under acceptably weak conditions on the models $p(y|x, \theta)$, the information-maximization strategy leads to consistent and efficient estimates of the true underlying model, in a natural sense. In particular, the information-maximization strategy is never less efficient, in a well-defined sense—and is generically more efficient—than the simpler, nonadaptive, i.i.d. $x$ strategy. We also give a few examples to illustrate the applicability of our results (see sections 3 and 4), including a couple of surprising negative examples that demonstrate the nontriviality of our mathematical results (see section 5). We close by briefly noting the relevance of our results to noninformation-theoretic (e.g., mean-square-error based) design and describing a few open avenues for further research.

## 2 Results

First, we note that the problem as posed in section 1 turns out to be slightly simpler than one might have expected, because $I(\{x, y\}; \theta)$ is linear in $p(x)$:

$$\begin{aligned} I(\{x, y\}\,;\,\theta) &= \int_X \int_Y \int_\Theta p(x, y, \theta) \log \frac{p(x, y, \theta)}{p(x, y)\, p_N(\theta)} \\ &= \int_X \int_Y \int_\Theta p(x, y, \theta) \log \frac{p(x)\, p_N(\theta)\, p(y|x, \theta)}{p(y|x)\, p(x)\, p_N(\theta)} \\ &= \int_X \int_Y \int_\Theta p(x, y, \theta) \log \frac{p(y|x, \theta)}{p(y|x)} \\ &= \int_X p(x) \int_Y \int_\Theta p_N(\theta)\, p(y|x, \theta) \log \frac{p(y|x, \theta)}{\int_\Theta p_N(\theta)\, p(y|x, \theta)}. \end{aligned}$$

This, in turn, implies that the optimal $p(x)$ must be degenerate, concentrated on the points $x$ where $I$ is maximal. Thus, instead of finding optimal distributions $p(x)$, we need only find optimal inputs $x$, in the sense of maximizing the conditional information between $\theta$ and $y$, given a single input $x$:

$$I(y; \theta|x) \equiv \int_Y \int_\Theta p_N(\theta)\, p(y|x, \theta) \log \frac{p(y|x, \theta)}{\int_\Theta p_N(\theta)\, p(y|x, \theta)}.$$

(We will assume throughout the article that this function attains its supremum in $X$—a condition guaranteeing that this is so will be given below—and that some reasonable, though possibly nondeterministic, tie-breaking stategy exists when this maximum is not unique.)

Our main result is a Bernstein–von Mises type of theorem (van der Vaart, 1998). The classical form of this kind of result says, basically, that if the posterior distributions are consistent (in the sense that $p_N(U) \to 1$ for any neighborhood $U$ of the true parameter $\theta_0$) and the likelihood ratios are sufficiently smooth on average, then the posterior distributions $p_N(\theta)$ are

asymptotically normal, with easily calculable asymptotic mean and variance. In particular, it is well known that a result of this type holds in the i.i.d. $x$ case: under the smoothness conditions stated below, the posterior distribution $p_N$ is asymptotically normal, with covariance matrix $\sigma^2_{iid}/N$,

$$\sigma^2_{iid} \equiv \left( \int_X dp(x) I_{\theta_0}(x) \right)^{-1},$$

and a mean that itself is a normal random variable with mean $\theta_0$ and covariance $\sigma^2_{iid}/N$. Here we have denoted the Fisher information matrices,

$$I_\theta(x) = \int_Y \left( \frac{\dot{p}(y|x,\theta)}{p(y|x,\theta)} \right)^t \left( \frac{\dot{p}(y|x,\theta)}{p(y|x,\theta)} \right) dp(y|x,\theta),$$

where the differential $\dot{p}$ is taken with respect to $\theta$. In other words, the asymptotic variance decays as $1/N$, with the exact rate $\sigma^2_{iid}$ defined as the inverse of the average Fisher information, where the average is taken over $p(x)$.

We adapt this result to the present case, where $x$ is chosen according to the information-maximization recipe. Our main result will allow us to compute the asymptotic variance $\sigma^2_{info}/N$ and in particular will demonstrate that $|\sigma^2_{info}| \leq |\sigma^2_{iid}|$, with $|.|$ denoting the determinant of a matrix; that is, the information-maximization strategy is more efficient in general than i.i.d. sampling, at least in the sense measured by the determinant $|.|$. It turns out that the hard part is proving consistency (see section 5); we give the basic consistency lemma (interesting in its own right) first, from which the main theorem follows fairly easily. The proofs appear in appendix A.

**Lemma 1 (Consistency).** *Assume the following conditions:*

1. *The parameter space $\Theta$ is a compact metric space.*
2. *The log likelihood $\log p(y|x,\theta)$ is uniformly Lipschitz in $\theta$ with respect to some dominating measure on $Y$.*
3. *The prior measure $p_0$ assigns positive measure to any neighborhood of $\theta_0$.*
4. *The maximal Kullback-Leibler divergence,*

$$\sup_x D_{KL}(\theta_0; \theta|x) \equiv \sup_x \int_Y dp(y|x,\theta_0) \log \frac{p(y|x,\theta_0)}{p(y|x,\theta)}$$

*is positive for all $\theta \neq \theta_0$.*

*Finally, assume that the set of log likelihood functions $\log p(y|x,\theta)$, indexed by $x$, is compact in the sup-norm topology on $\theta$-continuous functions on $Y \times \Theta$. Then the posteriors are consistent: $p_N(U) \to 1$ in probability for any neighborhood $U$ of $\theta_0$.*

**Theorem 1 (Asymptotic normality).** *Assume the conditions of lemma 1, strengthened as follows:*

1. $\Theta$ *has a smooth, finite-dimensional manifold structure in a neighborhood of* $\theta_0$.
2. *The log likelihood* $\log p(y|x,\theta)$ *is uniformly* $C^2$ *in* $\theta$. *In particular, the Fisher information matrices* $I_\theta(x)$ *are well defined and continuous in* $\theta$, *uniformly in* $(x,\theta)$ *in some neighborhood of* $\theta_0$.
3. *The prior measure* $p_0$ *is absolutely continuous in some neighborhood of* $\theta_0$, *with a continuous positive density at* $\theta_0$.
4. $$max_{C\in co(I_{\theta_0}(x))}|C| > 0,$$

   *where co* $(I_{\theta_0}(x))$ *denotes the convex closure of the set of Fisher information matrices* $I_{\theta_0}(x)$.

*Then*

$$||p_N - \mathcal{N}(\mu_N, \sigma_N^2)|| \to 0$$

*in probability, where* $||.||$ *denotes variation distance,* $\mathcal{N}(\mu_N, \sigma_N^2)$ *denotes the normal density with mean* $\mu_N$ *and covariance* $\sigma_N^2$, *and* $\mu_N$ *is asymptotically normally distributed, with mean* $\theta_0$ *and variance* $\sigma_N^2$. *Here*

$$N\sigma_N^2 \to \sigma_{info}^2 \equiv \left(\text{argmax}_{C\in co(I_{\theta_0}(x))}|C|\right)^{-1}.$$

*The maximum in the above expression is well defined and unique.*

**Corollary 1.** *If, in addition, the prior* $p_0$ *is absolutely continuous, with density bounded on the parameter space* $\Theta$, *then the maximum a posteriori (MAP) estimator is consistent almost surely, with asymptotic distribution* $\mathcal{N}(\theta_0, \sigma_N^2)$.

Thus, under these conditions, the information-maximization strategy works; moreover, since the asymptotic i.i.d. variance $\sigma_{iid}^2$ is inversely related to an average over $x$ and the information-maximization variance $\sigma_{info}^2$ to a maximum over $\text{co}(I_{\theta_0}(x))$, we have by the definition of $\text{co}(I_{\theta_0}(x))$—the closure of the set of all possible averages over $I_{\theta_0}(x)$ with respect to arbitrary $p(x)$—that $|\sigma_{info}^2|$ is never larger than $|\sigma_{iid}^2|$. For one-dimensional $\theta$, $\sigma_{info}^2$ is strictly smaller than $\sigma_{iid}^2$ except in the somewhat exceptional case that $I_{\theta_0}(x)$ is constant almost surely in $p(x)$. Thus, information maximization is in a rigorous sense asymptotically more efficient than the i.i.d. sampling strategy.

A few words about the assumptions are in order. Most should be fairly self-explanatory: the conditions on the priors, as usual, are there to ensure that no matter how mistaken our original prior beliefs are, in the face of sufficient posterior evidence, we will come around to agreeing that the data

are in fact generated by the true underlying model $\theta_0$. The smoothness assumptions on the likelihood permit the local expansion that is the source of asymptotic normality, and the condition on the maximal divergence function $\sup_x D_{KL}(\theta_0; \theta|x)$ ensures that distinct models $\theta_0$ and $\theta$ are identifiable (that is, for any $\theta \neq \theta_0$, there is some input $x$ that will reliably distinguish between $\theta$ and $\theta_0$ given enough output samples $y_i$). The assumption that the set of log likelihood functions $\log p(y|x, \theta)$ is compact will guarantee that the objective function $I(y, \theta|x)$ always attains its maximum in $x$. Finally, some form of monotonicity or compactness on $\Theta$ is necessary, mostly to bound the maximal divergence function $\sup_x D_{KL}(\theta_0; \theta|x)$ and its inverse away from zero (the lower bound, again, is to uniformly ensure identifiability; the necessity of the upper bound will become clear in section 5). Also, compactness is useful (though not necessary) for adapting certain Glivenko-Cantelli bounds (van der Vaart, 1998) for the consistency proof.

It should also be clear that we have not stated the results as generally as possible. We have chosen instead to use assumptions that are simple to understand and verify and to leave the technical generalizations to the interested reader. Our assumptions should be weak enough for most neurophysiological and psychophysical situations, for example, by assuming that parameters take values in bounded (though possibly large) sets and that tuning curves are not infinitely steep.

## 3 Applications

**3.1 Psychometric Model.** As noted in section 1, psychophysicists have employed versions of the information-maximization procedure for some years (Watson & Pelli, 1983; Pelli, 1987; Watson & Fitzhugh, 1990; Kontsevich & Tyler, 1999). References in Watson and Fitzhugh (1990), for example, go back four decades, and while these earlier investigators usually couched their discussion in terms of variance instead of entropy, the basic idea is the same (note, for example, that in the one-dimensional $\theta$ case, minimizing entropy is asymptotically equivalent to minimizing variance, by our main theorem). Our results above allow us to quantify the effectiveness of this stategy precisely.

One general psychometric model is as follows. The response space $Y$ is binary, corresponding to subjective yes or no detection responses. Let $f$ be sigmoidal: a uniformly smooth, monotonically increasing function on the line, such that $f(0) = 1/2$, $\lim_{t\to-\infty} f(t) = 0$ and $\lim_{t\to\infty} f(t) = 1$ (this function represents the detection probability when the subject is presented with a stimulus of strength $t$). Let $f_{a,\theta} = f((t - \theta)/a)$; $\theta$ here serves as a location ("threshold") parameter, while $a$ sets the scale (we assume $a$ is known for now, although this can be relaxed; (Kontsevich & Tyler, 1999). Finally, let $p(x)$ and $p_0(\theta)$ be some fixed sampling and prior distributions, respectively, both equivalent to Lebesgue measure on some interval $\Theta$.

Now, for any fixed scale $a$, we want to compare the performance of the information-maximization strategy to that of the i.i.d. $p(x)$ procedure. We have by theorem 1 that the most efficient estimator of $\theta$ is asymptotically unbiased with asymptotic variance $\sigma^2_{info}/N$, with

$$\sigma^2_{info} = \left(\sup_x I_{\theta_0}(x)\right)^{-1},$$

while the usual calculations show that the asymptotic variance of any efficient estimator based on i.i.d. samples from $p(x)$ is given by $\sigma^2_{iid}/N$, with

$$\sigma^2_{iid} = \left(\int_X dp(x) I_{\theta_0}(x)\right)^{-1}.$$

The Fisher information is easily calculated here to be

$$I_\theta = \frac{(\dot{f}_{a,\theta})^2}{f_{a,\theta}(1 - f_{a,\theta})}.$$

We can immediately derive two easy but important conclusions. First, there is just one function $f^*$ satisying the assumptions stated above for which the i.i.d. sampling strategy is as asymptotically efficient as the information-maximization strategy; for all other $f$, information maximization is strictly more efficient. This extremal function $f^*$ is the unique solution of the following differential equation, derived by setting $I_\theta$ to a constant (and therefore making the expected Fisher information equal to the maximal Fisher information),

$$\frac{df^*}{dt} = c\left(f^*(t)(1 - f^*(t))\right)^{1/2},$$

where the auxiliary constant $c = \sqrt{I_\theta}$ uniquely fixes the scale $a$. After some calculus, we obtain

$$f^*(t) = \frac{\sin(ct) + 1}{2}$$

on the interval $[-\pi/2c, \pi/2c]$ (and defined uniquely, by monotonicity, as 0 or 1 outside this interval). Since the support of the derivative of this function is compact, this result is not independent of the sampling density $p(x)$; if $p(x)$ places any of its mass outside the interval $[-\pi/2c, \pi/2c]$, then $\sigma^2_{iid}$ is always strictly greater than $\sigma^2_{info}$ (since $\dot{f}$, and therefore $I_{\theta_0}(x)$, is zero outside this interval). This recapitulates a basic theme from the psychophysical literature comparing adaptive and nonadaptive techniques. When the scale

of the nonlinearity $f$ is either unknown or smaller than the scale of the i.i.d. sampling density $p(x)$, adaptive techniques are generally preferable.

Second, a crude analysis shows that as the scale of the nonlinearity $a$ shrinks, the ratio $\sigma^2_{iid}/\sigma^2_{info}$ grows approximately as $1/a$. This gives quantitative support to the intuition that the sharper the nonlinearity with respect to the scale of the sampling distribution $p(x)$, the more we can expect the information-maximization strategy to help. In fact, in the limit as $a \to 0$, samples from the model become perfectly deterministic (with the response curve $f_{0,\theta}$ changing discontinuously from 0 to 1 at $\theta$), and the information-maximization strategy becomes infinitely more efficient than i.i.d. sampling. Information-maximal sampling is a version of the "twenty questions" game here, with each query $x$ decreasing the entropy of $p_N(\theta)$ by one bit, which in turn leads to exponential convergence in $N$ instead of the $N^{-1/2}$ rate guaranteed in the smoothly varying $f$ case.

**3.2 Linear-Nonlinear Cascade Model.** We now consider a model that has received growing attention from the neurophysiology community (see, e.g., Simoncelli, Paninski, Pillow, & Schwartz, 2004, for a recent review). The model is of cascade form, with a linear stage followed by a nonlinear stage: the input space $X$ is a compact subset of $d$-dimensional Euclidean space (take $X$ to be the unit sphere, for concreteness), and the firing rate of the model cell, given input $\vec{x} \in X$, has the simple form

$$E(y|\vec{x}, \theta) = f(\langle \vec{\theta}, \vec{x} \rangle).$$

Here the linear filter $\vec{\theta}$ is some unit vector in $X'$, the dual space of $X$ (thus, $\Theta$ is isomorphic to $X$, as in the previous example), while the nonlinearity $f$ is some nonconstant, nonnegative function on $[-1, 1]$. We assume that $f$ is uniformly smooth, to satisfy the conditions of theorem 1; we also assume $f$ is known, although, again, this can be relaxed. The response space $Y$—the space of possible spike counts, given the stimulus $\vec{x}$—can be taken to be some large, bounded set of the nonnegative integers. For simplicity, let the conditional probabilities $p(y|\vec{x}, \theta)$ be parameterized uniquely by the mean firing rate $f(\langle \vec{\theta}, \vec{x} \rangle)$; the most convenient model, as usual, is to assume that $p(y|\vec{x}, \theta)$ is Poisson with mean $f(\langle \vec{\theta}, \vec{x} \rangle)$. Finally, we assume that the sampling density $p(x)$ is uniform on the unit sphere (this choice is natural for several reasons, mainly involving symmetry; see, e.g., (Chichilnisky, 2001; Simoncelli et al., 2004), and that the prior $p_0(\theta)$ is positive and continuous (and is therefore bounded above and away from zero by the compactness of $\Theta$).

The Fisher information for this model is easily calculated as

$$I_\theta(x) = \frac{(\dot{f}(\langle \vec{\theta}, \vec{x} \rangle))^2}{f(\langle \vec{\theta}, \vec{x} \rangle)} P_{\vec{x},\theta},$$

where $\dot{f}$ is the usual derivative of the real function $f$ and $P_{\vec{x},\theta}$ is the projection operator corresponding to $\vec{x}$, restricted to the $(d-1)$-dimensional tangent space to the unit sphere at $\theta$. (We have assumed that the bounded set $Y$ of allowed spike counts has been taken sufficiently large to ignore the deviations from exact Poisson behavior due to the finite spike count cutoff.) Theorem 1 now implies that

$$\sigma^2_{info} = \left( \max_{t \in [-1,1]} \frac{\dot{f}(t)^2 g(t)}{f(t)} \right)^{-1},$$

while

$$\sigma^2_{iid} = \left( \int_{[-1,1]} dp(t) \frac{\dot{f}(t)^2 g(t)}{f(t)} \right)^{-1},$$

where $g(t) = 1 - t^2$, $p(t)$ denotes the one-dimensional marginal measure induced on the interval by the uniform measure $p(x)$ on the unit sphere, and $\sigma^2$ in each of these two expressions multiplies $(d-1)I_{d-1}$, with $I_{d-1}$ denoting the $(d-1)$-dimensional identity matrix.

Clearly, the arguments of section 3.1 apply here as well: the ratio $\sigma^2_{iid}/\sigma^2_{info}$ grows roughly linearly in the inverse of the scale of the nonlinearity. The more interesting asymptotics here, though, are in $d$. This is because the unit sphere has a measure concentration property (Milman & Schechtman, 1986; Talagrand, 1995): as $d \to \infty$, the measure $p(t)$ becomes exponentially concentrated around 0. In fact, it is easy to show directly that in this limit, $p(t)$ converges in distribution to the normal measure with mean zero and variance $d^{-2}$. The most surprising implication of this result is seen for nonlinearities $f$ such that $\dot{f}(0) = 0$, $f(0) > 0$; we have in mind, for example, symmetric nonlinearities like those often used to model complex cells in visual cortex. For these nonlinearities,

$$\frac{\sigma^2_{info}}{\sigma^2_{iid}} = O(d^{-2}) :$$

that is, in this case, the information-maximization strategy becomes infinitely more efficient than the usual i.i.d. approach as the dimensionality of the spaces $X$ and $\Theta$ grows.

## 4 Illustrations

Next we give some illustrations of the behavior of the information-optimization strategy, as compared to the nonadaptive i.i.d. case.

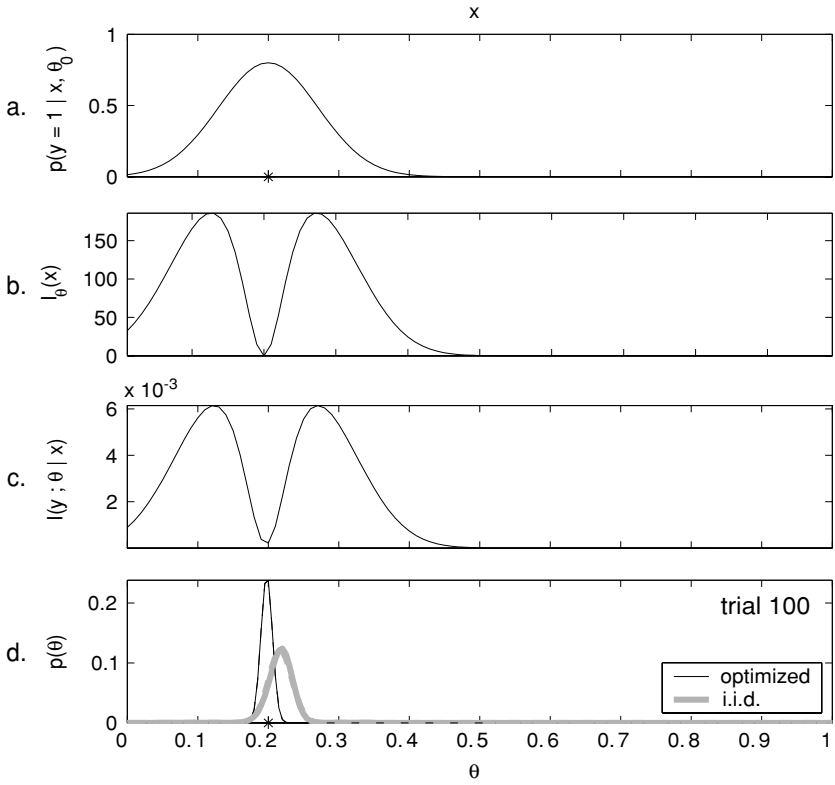


Figure 1: Snapshot of behavior of information-maximizing and i.i.d. experimental designs. (a) True underlying conditional response probabilities given input $x$. Model space $\Theta$ includes all translates of firing rate curve shown here. (b) Fisher information $I_{\theta_0}(x)$ as a function of $x$. (c) Mutual information between the response $y$ and the underlying location parameter $\theta$, as a function of $x$, after 100 samples. (d) Posterior distributions $p_N(\theta)$ after $N = 100$ samples. The asterisk indicates the location of true model parameter $\theta_0$. The dashed lines give gaussian approximations to true observed posteriors, though the dashed curves are obscured by the quality of the fit.

**4.1 One-Dimensional Example.** For clarity, we start with a simple example, for which the stimulus and model spaces $X$ and $\Theta$ are both one-dimensional and the outputs are again binary. We illustrate the model in Figure 1. The system responds positively with high probability when the stimulus $x$ and model preference $\theta$ agree; this probability decays smoothly and symmetrically as the difference $|x - \theta|$ increases. (We could think of this response probability curve as a sensory neuron's receptive field for some

one-dimensional stimulus, for example, or as the place field of a cell in the hippocampus of a rat constrained to run along a one-dimensional track.) The experimenter's goal is to determine the optimal $\theta_0$, given the response curves $p(y = 1|x, \theta)$. We begin with no knowledge of the true $\theta_0$ except that $0 \leq \theta_0 \leq 1$; thus, we take the prior $p_0(\theta)$ to be uniform on $[0, 1]$. In the bottom panel of Figure 1, we show the results of an experiment in which we draw 100 input samples i.i.d. from the uniform distribution $p(x) = 1$ on $[0, 1]$, and then compare to the results given 100 samples drawn adaptively, following the information-maximization strategy. After 100 samples, we find that the mutual information curve $I(y; \theta|x)$, as a function of the input $x$ (see Figure 1c), closely resembles the Fisher information curve $I_{\theta_0}(x)$ (see Figure 1b); in particular, the two curves reach their maxima for the same values of $x$, indicating that after 100 samples, the information-maximizing strategy is indeed sampling from the $x$ that maximizes the Fisher information $I_{\theta_0}(x)$, as predicted by theorem 1. Note that sampling from the $x$ that maximizes the firing rate, $x = \theta_0 = 0.2$, asymptotically minimizes the information gain $I(y; \theta|x)$, as emphasized in section 1. Also as predicted, the posterior distributions $p_N(\theta)$ are quite well approximated as gaussian, with means near $\theta_0$ and with the posterior under the information-maximization strategy more concentrated near $\theta_0$ than in the i.i.d. $x$ case.

We look more quantitatively at the evolution of the posteriors in Figure 2. The top two panels show the posteriors $p_N(\theta)$ as a function of $N$, while the bottom three panels show the posterior mean, standard deviation, and probability mass in a small neighborhood of the true parameter $\theta_0$, respectively. In each case, we again see that the posterior under the information-maximization strategy converges more rapidly than under the i.i.d. strategy, as predicted. Moreover, the predicted standard deviations of the posterior density and of the posterior mean, $\sigma_{iid} N^{-1/2}$ and $\sigma_{info} N^{-1/2}$, accurately match the true observed behavior.

**4.2 A V1 Simple-Cell Example.** Our second example is somewhat more realistic in that the neuron we are simulating responds to stimuli that have many degrees of freedom; that is, the parameter and input spaces $\Theta$ and $X$ are multidimensional. We take what is perhaps the standard model of the response properties of a simple cell in primary visual cortex (a version of the cascade model discussed in the last section; Dayan & Abbott, 2001): $p(spike|\vec{x}, \theta) = f(\langle \vec{k}_0, \vec{x} \rangle)$, where the true receptive field $\vec{k}_0$ is taken to be a Gabor function (the product of a two-dimensional sinusoid and a gaussian whose mean determines the location of the receptive field in space; see Figure 3, top left), and the monotonic nonlinear function $f$ enforces the positivity of the firing rate and can also model the cell's saturation properties (see Figure 3, bottom left). For simplicity, we assume the nonlinearity $f$ and the spatial frequency of the Gabor $\theta_0 = \vec{k}_0$ to be known; thus, the model space $\Theta$ is three-dimensional (two dimensions for the location of the receptive field

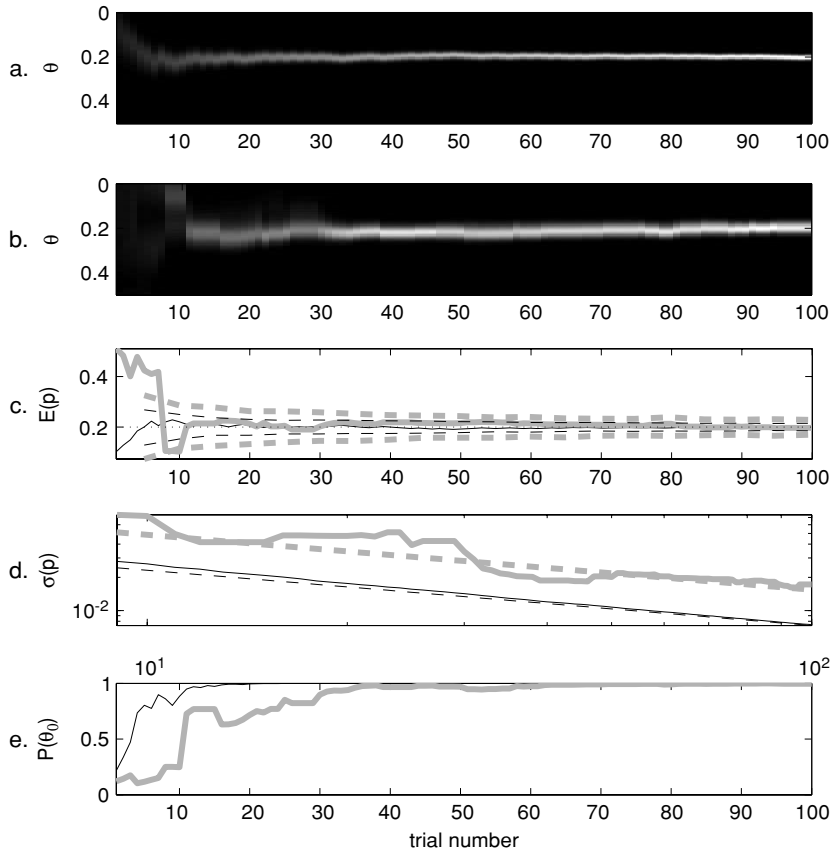


Figure 2: Evolution of posterior densities in model from Figure 1, as a function of trial number $N$. (a, b) Evolution of posteriors under information-optimizing and i.i.d. strategy, respectively. White level indicates height of probability density. Recall from Figure 1 that the true parameter is located at $\theta_0 = 0.2$. (c) Evolution of the posterior mean. The dotted line indicates true parameter location ($\theta_0 = 0.2$). The solid black and gray indicate mean given information-optimizing and i.i.d. strategies, respectively. Dashed lines show predicted 95% confidence intervals, $\theta_0 +/- 2\sigma_{info}N^{-1/2}$ and $+/- 2\sigma_{iid}N^{-1/2}$. (d) Evolution of posterior standard deviation. Solid traces are observed standard deviations; dashed traces are predicted, $\sigma_{info}N^{-1/2}$ and $\sigma_{iid}N^{-1/2}$. (e) Evolution of posterior mass contained in a small neighborhood of true parameter, $p_N([\theta_0 - 0.05, \theta_0 + 0.05])$.

and one for the orientation). As in the previous example, we start with no knowledge of the true parameter other than the fact that the center of the receptive field lies within the square shown in Figure 3, so our prior $p_0(\vec{k})$ is uniform over orientation and spatial location. The series of panels on the

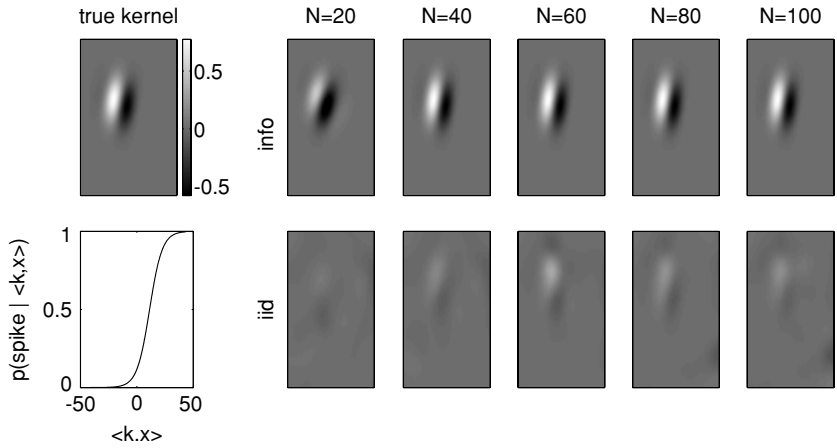


Figure 3: Evolution of posteriors in a simulated V1 simple cell experiment. (Left) True model: the simulated simple cell responds to the stimulus image $\vec{x}$ according to $p(spike|\vec{x}, \vec{k}_0) = f(\langle \vec{k}_0, \vec{x} \rangle)$, with $\langle \vec{k}_0, \vec{x} \rangle$ indicating the dot product with the Gabor kernel $\vec{k}_0$ shown in the top-left panel and $f$ the nonlinear rectification function shown in the bottom-left panel. (Right) Evolution of posteriors as a function of trial number $N$. Each panel shows the posterior mean $\int \vec{k} dp_N(\vec{k})$.

right displays the evolution of the posteriors $p_N(\vec{k})$ via the posterior mean $\int \vec{k} dp_N(\vec{k})$; as $p_N(\vec{k})$ becomes concentrated around the true receptive field $\vec{k}_0$, the image of the posterior mean resembles $\vec{k}_0$ more and more closely. We took the stimuli $\vec{x}$ here to be Gabors of varying orientations and locations, but similar results are seen if white- or colored-noise stimuli are used instead (data not shown). As before, the information-optimizing strategy leads to more rapid convergence than i.i.d. sampling from $x$.

## 5 Negative Examples

Our next two examples are more negative and perhaps more surprising: they show how the information-maximation strategy can fail, in a certain sense, if the conditions of the consistency lemma are not met. (Note that as emphasized above, these consistency conditions are fairly weak; therefore, the fact that they fail in the following examples implies that these examples might be interesting from a mathematical point of view but might have less practical negative relevance for psychophysical or neurophysiological situations.) In each case, the method can be fixed using ad hoc methods; it is unclear at present whether a generally applicable modification of the basic information-maximization strategy exists.

**5.1 Two-Threshold Model.** Let $\Theta$ be multidimensional, with coordinates that are "independent" in the sense that the responses of the model given one coordinate do not depend on the value of the other coordinate, and assume the expected information obtained from one coordinate remains bounded strictly away from the expected information obtained from one of the other coordinates. For instance, consider the following binary model:

$$p(1|x,\theta) = \begin{cases} .5 & -1 < x \le \theta_{-1}, \\ f_{-1} & \theta_{-1} < x \le 0, \\ .5 & 0 < x \le \theta_1, \\ f_1 & \theta_1 < x \le 1, \end{cases}$$

where $0 \le f_{-1}, f_1 \le 1$,

$$|f_{-1} - .5| > |f_1 - .5|,$$

are known and $-1 < \theta_{-1} < 0$ and $0 < \theta_1 < 1$ are the parameters we want to learn.

Let the initial prior $p_0(\theta)$ be absolutely continuous with respect to the Lebesgue measure; this implies that all posteriors $p_N$ will have the same property. Then, using the inverse cumulative probability transform and the fact that mutual information is invariant with respect to invertible mappings, it is easy to show that the maximal information we can obtain by sampling from the left is strictly greater than the maximal information obtainable from the right, uniformly in $N$. Thus, the information-maximization strategy will sample from the left side forever, leading to a linear information growth rate (and easily proven consistency) for the left parameter and nonconvergence on the right. Compare the performance of the usual i.i.d. approach for choosing $x$ (using any Lebesgue-dominating measure on the parameter space), which leads to the standard algebraic convergence rate for both parameters (i.e., is strongly consistent in posterior probability).

Note that this kind of inconsistency problem does not occur in the case of sufficiently smooth $p(y|x,\theta)$, by our main theorem. Thus, one way of avoiding this problem would be to fix a finite sampling scale for each coordinate (i.e., discretizing). Below this scale, no information can be extracted; therefore, when the algorithm hits this "floor" for one coordinate, it will switch to the other. However, the next example shows that the lack of consistency is not necessarily tied to the discontinuous nature of the conditional densities.

**5.2 White Noise Models.** We present two models of a slightly different flavor: the basic mechanism of inconsistency is the same in each case. The samples $x$ take values on the positive integers. The models live on the positive integers as well: $\theta$ is given by a standard discrete (1) normal and (2) binary white noise process (that is, $p(\theta)$ is generated by an infinite

sequence of standard normals and independent fair coins, respectively). The conditionals are defined as follows. For the first model, the observations $y$ are gaussian-contaminated versions of $\theta(x)$, that is, $y \sim \mathcal{N}(\theta(x), 1)$. For the second model, let $y$ be drawn randomly from $q_{\theta(x)}$, where $q_0$ and $q_1$ are nonidentical measures on some arbitrary space.

Then it is not hard to show, for either model, that an experimenter using the information-maximization strategy will never sample from any $x$ infinitely often. As soon as we learn something about $\theta_i$ (by sampling from $x_i$), $\theta_{i+1}$ will become more interesting, and we will begin to sample from $x_{i+1}$ instead. (For the second model, in fact, if the densities of $q_0$ and $q_1$ with respect to some dominating measure are unequal almost surely, and $I(y, \theta(1))$ reaches its unique maximum in $p(\theta(1))$ at the midpoint $p(\theta(1) = 0) = p(\theta(1) = 1)$, then we will sample from each $x$ just once, almost surely.) This again implies a lack of consistency of the posterior (although, as above, we have a linear growth of information). The basic idea is that there will always be a more informative part of the sample space $X$ to measure from, and the experimenter will never spend enough time in one place $x$ to sufficiently characterize $\theta(x)$. This emphasizes the necessity of something like the compactness condition we imposed in the statement of lemma 1.

As in the last section, the standard i.i.d. approach (using any measure $p(x)$ that does not assign zero mass to any of the integers) is consistent here. Note that in contrast to the last example, the smoothness of the conditionals $p(y|x, \theta)$ (in the gaussian model) does not rescue consistency. Nor is the inconsistency due to some pathology of differential entropy (the measures $q_i$ can be discrete, even binary). The "floor" trick suggested for the last example can be modified here by sequentially restricting our search for optimal $x$ over compacta that are allowed to grow slowly toward infinity. More generally, we can probably salvage consistency in general by not sampling exclusively from information-maximizing points (perhaps by sampling "passively" with a frequency that decreases as $N \to \infty$; this would restore consistency in many cases without a sacrifice in the asymptotic information growth rate). We leave the general formulation of such a result to the reader.

## 6 Directions

We have presented a rigorous theoretical framework for adaptive design of experiments using the information-theoretic objective function (see equation 1.1). Most important, we have offered some asymptotic results that clarify the effectiveness of this information-maximizing strategy; in addition, we expect that our results should find applications in approximative computational schemes for optimizing stimulus choice during this type of online experiment. For example, our theorem 1 might suggest the use of a

mixture-of-gaussians representation as an efficient approximation for the posteriors $p_N(\theta)$ (Deignan, Meckl, Franchek, Abraham, & Jaliwala, 2000).

We briefly describe a few more open research directions.

**6.1 Nuisance Parameters and Hypothesis Testing.** Perhaps the most obvious such open question concerns the use of noninformation-theoretic objective functions. Concerning the question of which objective function is "best" in general, our results should be useful in clarifying the exact form of the asymptotic covariance matrix $\sigma^2_{info}$: for the information-maximization case, this matrix asymptotically minimizes the log-determinant function on the class of feasible asymptotic covariance matrices $(\mathrm{co}(I_{\theta_0}(x)))^{-1}$. However, we are typically interested in some parameters more than others; as has been noted elsewhere (Mackay, 1992), mutual information as an objective function (and, by extension, its log-determinant asymptotic form) leaves little flexiblity for focusing our resources on these more interesting parameters. Alternative objective functions include weighted sums of entropies or Bayes mean-square errors. It turns out that many of our results apply with only modest changes if the experiment is instead designed to optimize these alternative objective functions: in this case, the results in sections 3 and 4.1 remain completely unchanged, while the statement of our main theorem requires only slight changes in the asymptotic covariance formula (see appendix A). The task of choosing a good objective function on input distributions $p(x)$ in the multidimensional $\theta$ case is thus reduced asymptotically to the simpler problem of choosing a suitable objective function on covariance matrices; in the one-dimensional $\theta$ case, the asymptotic variance does not depend on whether we choose to minimize entropy or variance.

An alternative approach that has received less attention involves mapping irrelevant "nuisance" parameters out of $\Theta$. In a sense, this is a special case of the weighted sum of entropies idea, for which some of the weights are set to zero, but can be defined in a slightly more general setting. We define our new objective function in a simple way as

$$I(\{x, y\}; T(\theta)),$$

where $T$ is a surjective map from $\Theta$ to some new, reduced parameter space (obviously this new definition corresponds to the original equation 1.1 if $T$ is bijective). This approach thus integrates over nuisance parameters in a completely direct way. Clearly, much of our asymptotic theory will go through under some continuity on $T$, as long as the assumptions on the maximal divergence $\sup_x D_{KL}(\theta_0; \theta|x)$ and on the positivity of the Fisher information matrices are unharmed.

It is worth addressing an extreme specialization of the above idea: the case for which $T$ maps $\Theta$ to the two points $\{0, 1\}$ corresponds to compound hypothesis testing. Again, as long as the gap on $\sup_x D_{KL}(\theta_0; \theta|x)$ is respected

by $T$, consistency will go through, but asymptotic normality stops making sense: the more relevant concept now becomes the large deviations behavior of $p_N(0)$ and $p_N(1)$, as described by the Chernoff information (Cover and Thomas, 1991; Dembo & Zeitouni, 1993). We have not addressed the optimal rates of convergence in this case, but note only that even simple hypothesis testing (i.e., the case that $\Theta = \{0, 1\}$) is aided by adaptive stimulus design, in the sense that a given $x$ that optimizes $I(y, \theta|x)$ for one value of $p_N(0)$ is not necessarily optimal for all other $p_N(0)$; thus, as before, the optimal sampling strategy varies in general with $N$.

**6.2 "Batch Mode" and Stimulus Dependencies.** Perhaps our strongest assumption here is that the experimenter will be able to freely choose the stimuli on each trial. This might be inaccurate for a number of reasons: for example, computational demands might require that experiments be run in batch mode, with stimulus optimization taking place not after every trial, but perhaps only after each batch of $k$ stimuli, all chosen according to some fixed distribution $p(x)$. Another common situation involves stimuli that vary temporally, for which the system is commonly modeled as responding not just to a given stimulus $x(t)$, but also to some or all of its time-translates $x(t - \tau)$. Finally, if there is some cost $C(x_0, x_1)$ associated with changing the state of the observational apparatus from the current state $x_0$ to $x_1$, the experimenter may wish to optimize an objective function that incorporates this cost: $I(y; \theta|x_1) - C(x_0, x_1)$, for example.

Each of these situations is clearly ripe for further study. Here we restrict ourselves to the first setting and give a simple conjecture, based on the asymptotic results presented above and inspired by results like those of Berger, Bernardo, and Mendoza (1989), Clarke and Barron (1994), and Scholl (1998). First, we state more precisely the optimization problem inherent in designing a batch experiment: we wish to choose some sequence, $\{x_i\}_{1 \le i \le k}$, to maximize

$$I(\{x_i, y_i\}_{1 \le i \le k}; \theta).$$

The main difference here is that $\{x_i\}_{1 \le i \le k}$ must be chosen nonadaptively, that is, without sequential knowledge of the responses $\{y_j\}_{j<k}$. Clearly, the order of any sequence of optimal $\{x_i\}_{1 \le i \le k}$ is irrelevant to the above objective function; in addition, it should be apparent that if no given datum $(x, y)$ is too strong (for example, under Lipschitz conditions like those in lemma 1), any given elements of such an optimal sequence $\{x_i\}_{1 \le i \le k}$ should be asymptotically independent in some sense. (Without such a smoothness condition—for example, if some input $x$ could definitively decide between some given $\theta_0$ and $\theta_1$—then no such asymptotic independence statement can hold, since no more than one sample from such an $x$ would be necessary.) Thus, we can hope that we should be able to asymptotically approximate

this optimal experiment by sampling in an i.i.d. manner from some well-chosen $p(x)$. Moreover, we can make a guess as to the identity of this putative $p(x)$:

**Conjecture (Batch mode).** *Under suitable conditions on the topology of $X$, the empirical distribution corresponding to any optimal sequence $\{x_i\}_{1 \le i \le k}$,*

$$\hat{p}(x) \equiv \frac{1}{k} \sum_{i=1}^{k} \delta(x_i),$$

*converges weakly as $k \to \infty$ to $S$, the convex set of maximizers in $p(x)$ of*

$$E_\theta \log \left( \left| \int dp(x) I_\theta(x) \right| \right). \tag{6.1}$$

Thus, instead of a very difficult (in particular, nonconvex in general) optimization over the sequence space $X^k$, we can optimize over distributions $p(x)$ to find good experiments (assuming $k$ is large enough). In particular, this latter optimization is tractable by the concavity of equation 6.1 in $p(x)$ (this follows from the concavity of the function $\log |C|$ as a function of the matrix $C$; Cover & Thomas, 1991; Lewis, 1996): simple ascent methods will find the global maximum without fear of becoming trapped in local optima. Expression 6.1 is an average over $p(\theta)$ of terms proportional to the negative entropy of the asymptotic gaussian posterior distribution corresponding to each $\theta$, and thus should be maximized by any optimal approximant distribution $p(x)$. In fact, it is not difficult, using the results of Clarke and Barron (1990), to prove the above conjecture under conditions like those of theorem 1, assuming that $X$ is finite (in which case, weak convergence is equivalent to pointwise convergence). We leave generalizations (in particular, the formulation of suitable conditions on the topology of more general $X$) for future work.

We should note that maximization of terms like expression 6.1 has been previously studied not only in the context of experimental design (where designs that maximize that equation are commonly called "D-optimal"; (Fedorov, 1972; Clyde & Chaloner, 1996), but also elsewhere. For example, when $\theta$ is one-dimensional (and thus the information matrices are simply scalar weights), equation 6.1 is mathematically equivalent to the criterion for weighted log optimality in the theory of optimal financial portfolio selection (Cover & Thomas, 1991). In this case, the Kuhn-Tucker conditions for optimality of $p(x)$ are well known and can be generalized easily to the multidimensional case once the directional derivatives of equation 6.1 with respect to $p(x)$ have been identified. We leave the details for appendix B.

## Appendix A: Proofs

We sketch proofs for the main results here.

**A. 1 Posterior Consistency.** We follow the basic technique of Wald (van der Vaart, 1998). The main idea is that $D_{KL}(\theta_0; \theta|\{x_i\}_{i<N})$ increases in $N$ for all $\theta \neq \theta_0$, and this expected log likelihood provides a suitable approximation to the observed posterior log likelihood $\log p_N(\theta_0/\theta)$. In the i.i.d. $p(x)$ case, this $D_{KL}$ term increases linearly with $N$, and this is by itself enough to prove consistency under weaker conditions than those stated here (Schwartz, 1967; van der Vaart, 1998; Barron, Schervish, & Wasserman, 1999). In the non-i.i.d. case, this linear growth does not necessarily hold, and we have to make sure that this function actually increases quickly enough off any given neighborhood $U$ of $\theta_0$ (i.e., that the sampler's attention does not get absorbed by some proper subset of $\Theta$).

We need to prove that

$$\frac{\int_{\Theta \cap U^c} dp_0(\theta) f_N(\theta)}{\int_U dp_0(\theta) f_N(\theta)} \to 0$$

for any neighborhood $U$, where $f_N(\theta)$ denotes the (random) likelihood ratio,

$$f_N(\theta) = \prod_{i=1}^{N} \frac{p(y_i|x_i, \theta)}{p(y_i|x_i, \theta_0)}.$$

The first step is to demonstrate that $\int_U dp_0(\theta) f_N(\theta)$ decreases at a slower-than-exponential rate, that is,

$$\liminf_N \frac{1}{N} \log \int_U dp_0(\theta) f_N(\theta) > -\epsilon$$

almost surely for any $\epsilon > 0$. This follows by exactly the usual proof (Schwartz, 1967; van der Vaart, 1998; Barron et al., 1999). The key step is to approximate

$$\log f_N(\theta) \approx \sum_i D_{KL}(\theta_0; \theta|x_i)$$

by a uniform law of the large numbers argument (van der Vaart, 1998) (the term on the right is the expectation of that on the left, where the expectation is taken under the true parameter $\theta_0$). Once this is done, the statement is proven by using the fact that

$$D_{KL}(\theta_0; \theta|x) \to 0$$

as $\theta \to \theta_0$, uniformly in $x$ (by assumption 2 of lemma 1), and that $p_0(U) > 0$ for any neighborhood $U$ (assumption 3).

The next step in the usual proof is to demonstrate that $\int_{\Theta \cap U^c} dp_0(\theta) f_N(\theta)$ decreases at an exponential rate, that is,

$$\limsup_N \frac{1}{N} \log \int_{\Theta \cap U^c} dp_0(\theta) f_N(\theta) < -\epsilon(U) < 0$$

almost surely for some positive $\epsilon$ that depends on $U$. Unfortunately, this exponential decay of $p_N(\Theta \cap U^c)$ does not necessarily hold in the information-maximization case. An example of this nonexponential decay is given after the proof.

To deal with this, first note that the set of functions $D_{KL}(\theta_0; \theta|x)$ is compact in the sup-norm topology on functions on $\Theta$. This follows from the Arzela-Ascoli theorem (Rudin, 1973), given the equicontinuity of the log likelihoods $\log p(y|x, \theta)$ in $\theta$ and the assumption that the set of these likelihoods is closed in the sup-norm topology. (A similar compactness argument guarantees that the maximum of $I(y, \theta|x)$ in $x$ is always attained in $X$.) Thus, the full set of stimuli $X$ may be replaced by a finite subset, $X' = \{x_j\}_{0<j<k<\infty} \subset X$, which satisfies the assumptions of lemma 1 and approximates $X$ arbitrarily well. For any $\epsilon > 0$, we may choose $X'$ such that

$$\sup_{x \in X} \min_{x' \in X'} \sup_{\theta \in \Theta} \left| D_{KL}(\theta_0; \theta|x) - D_{KL}(\theta_0; \theta|x') \right| < \epsilon.$$

The lemma will follow if we can prove the result for any such finite approximating set $X'$.

Thus, we may restrict our attention below to finite $X'$. We can immediately dispose of any set

$$Z_{>\delta} = \{\theta : D_{KL}(\theta_0; \theta|x) > \delta \; \forall \; x \in X'\}$$

for any $\delta > 0$, since the posterior mass of such a set decays exponentially, by the standard proof. This leaves us with the compact subset

$$Z_0 = \{\theta : \min_{x \in X'} D_{KL}(\theta_0; \theta|x) = 0\},$$

the set of $\theta$ where $D_{KL}(\theta_0; \theta|x) = 0$ for at least one $x \in X'$. Clearly, $\theta_0 \in Z_0$. If $Z_0 = \theta_0$, the proof is complete; thus, assume otherwise.

To complete the proof, we just need to show that the information-maximizing sampler does not asymptotically "ignore" any set $Z$ within $Z_0 \cap \Theta \cap U^c$ such that $p_0(Z^\epsilon) > 0$ (with $Z^\epsilon$ an arbitrary $\epsilon$-neighborhood of $Z$); that is, it does not sample so frequently from $x$ with $D_{KL}(\theta_0; \theta|x) = 0$ for $\theta \in Z$ that $q_N(Z) \equiv \int_Z e^{-\sum_i D_{KL}(\theta_0; \theta|x_i)} dp_0(\theta)$ does not decrease more quickly

than $q_N(U) = \int_U e^{-\sum_i D_{KL}(\theta_0;\theta|x_i)} d\,p_0(\theta)$. The main idea preventing this is the smoothness condition on the log likelihood functions $\log p(y|x,\theta)$; this condition guarantees that the information gain associated with increasing the concentration of $p_N$ at $\theta_0$ (the only point at which $D_{KL}(\theta_0;\theta|x) = 0$ for all $x$, by assumption) will decrease to zero, roughly as

$$\log \frac{q_N(U)}{q_{N+1}(U)} \approx \frac{q_N(U) - q_{N+1}(U)}{q_N(U)}.$$

Meanwhile, the gain associated with testing between $\theta_0 \in U$ and the alternative hypothesis will remain large whenever the posterior mass on $Z$ remains comparable to that on $U$, falling as $q_N(Z)/q_N(U)$. Thus, the information-maximizing sampler will prefer to increase the concentration at $\theta_0$ over attempting to distinguish between the hypotheses $\theta_0 \in U$ and $\theta_0 \in Z$ accordingly as $q_N(Z) < q_N(U) - q_{N+1}(U)$ or otherwise, respectively. The definition of the information-maximization strategy now implies that either $q_N(Z)$ decays exponentially (which would again complete the proof) or, alternatively,

$$\frac{q_N(U) - q_{N+1}(U)}{q_N(U)} \sim \frac{q_N(Z)}{q_N(U)},$$

that is, $q_N(U) - q_{N+1}(U) \sim q_N(Z)$. Since the posterior mass on $Z$ falls exponentially with the number of samples devoted to this hypothesis test and the posterior mass on $U$ cannot fall exponentially (as discussed above), the posterior mass on $Z$ must therefore decrease to zero relative to $q_N(U)$, because $q_N(U) - q_{N+1}(U) = o(q_N(U))$ for any sequence $q_N(U)$ that decays at a subexponential rate.

The above logic is further illustrated in the example below. We should also note that a similar result can be stated in the case that $\theta_0 \notin \Theta$, that is, when the data are not generated by a member of the hypothesized parameter space. Here, as usual, the posterior may be approximated as

$$p_N(\theta) \sim e^{-\sum_i D_{KL}(\theta_0;\theta|x_i)} p_0(\theta).$$

In the i.i.d. setting, this posterior will asymptotically concentrate around $\theta$, which are closest to the true $\theta_0$ in the sense of average $D_{KL}$ distance; however, in the information-maximization case, this notion of closeness to the true $\theta_0$ depends strongly on the stimuli $x$, and it is not clear that $e^{-\sum_i D_{KL}(\theta_0;\theta|x_i)}$ will even have a well-defined limit in general. Thus, to generalize lemma 1 to this out-of-$\Theta$ case, we would have to impose further conditions on the functions $D_{KL}(\theta_0;\theta|x)$. For example, the above proof holds if we stipulate some fixed "closest" element $\theta^*$ such that $\theta^*$ is in the set of minimizers of $D_{KL}(\theta_0;\theta|x)$ for all $x$ and is the only such member of $\Theta$ (just as in the setting

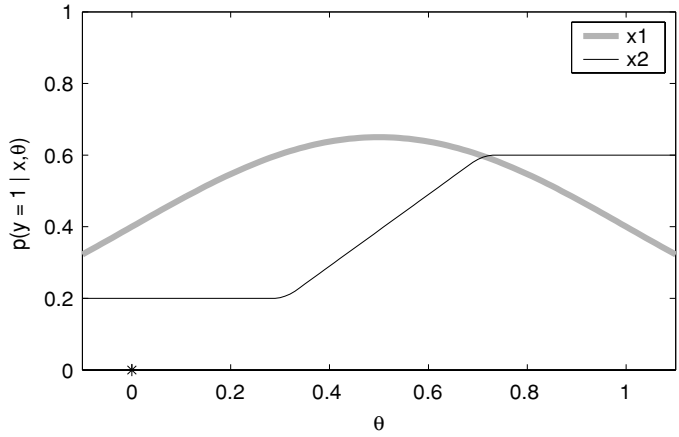


Figure 4: An example of a model in which the posterior mass off neighborhoods of the true parameter $\theta_0$ does not decay exponentially. $\theta_0 = 0$ here, as marked by an asterisk.

of lemma 2, $\theta_0$ is the unique member of $\Theta$ that minimizes $D_{KL}(\theta_0; \theta|x)$ for all $x$). In this case, $p_N$ will asymptotically concentrate on $\theta^*$.

**A.2 An Example of Subexponential Decay.** As discussed above, one major difference between the asymptotic behavior of the posteriors in the information-maximization and i.i.d. sampling cases is that the posterior mass $p_N(\Theta \cap U^c)$ generically decays exponentially in the i.i.d. setting (where $U$, again, is any neighborhood of the true parameter $\theta_0$); however, in the information-maximization case, this exponential decay does not necessarily hold. We give a simple example of this phenomenon here.

We choose both $X$ and $Y$ to be binary; the conditional distributions $p(y = 1|x_1, \theta)$ and $p(y = 1|x_2, \theta)$ are shown in Figure 4. The main important features to note are that $p(y|x_1, \theta_0) = p(y|x_1, \theta = 1)$, and $p(y|x_2, \theta_0) = p(y|x_2, \theta \in U)$, with $U$ a sufficiently small neighborhood of $\theta_0 = 0$. Thus, $x_1$ cannot distinguish between $\theta = 0$ and 1, and $x_2$ gives no information when $p(\theta)$ is sufficiently concentrated about $\theta_0$. This in turn implies that asymptotically, the information-maximization strategy will be to sample preferentially from $x_1$, as indicated by theorem 1. However, if all our samples are drawn from $x_1$, then significant posterior mass will remain on $\theta = 1$.

More precisely, the posterior $p_N(\theta)$ may be asymptotically approximated by a mixture of gaussians, one with mean at $\theta = 0$ and the other with mean at $\theta = 1$. (This approximation will hold asymptotically, by the usual argument, whenever the prior $p_0(\theta)$ has a positive, continuous density with respect to the Lebesgue measure on the $\Theta$ shown.) Both gaussians have variance of order $1/n$, where $n = n(N)$ is the number of samples from $x_1$ in the first

$N$ samples; the asymptotic ratio of masses between the two gaussians, on the other hand, behaves as $e^{-c(N-n)}$, with $c = D_{KL}(\theta_0; \theta|x_2)$. This implies that $I(y; \theta|x_1)$ scales as $1/n$, while $I(y; \theta|x_2)$ scales as $e^{-c(N-n)}$. This, in turn, means that $n(N)$ satisfies the scaling

$$n(N)^{-1} \sim e^{-c(N-n(N))},$$

leading to the conclusion that $N - n(N)$, the number of samples from $x_2$, grows sublinearly in $N$, and therefore that the posterior mass off neighborhoods of the true parameter $\theta_0$ does not decay exponentially in $N$. Conversely, it is easy to demonstrate exponential decay for any i.i.d. sampling strategy that places positive mass $p(x)$ on both $x_1$ and $x_2$.

**A.3 Asymptotic Normality.** The proof of asymptotic normality is fairly standard, and is therefore omitted (see, e.g., Schervish, 1995; van der Vaart, 1998). The only new part is the computation of the asymptotic variance $\sigma_N^2$. The classical result tells us that

$$\left(N\sigma_N^2\right)^{-1} \to \frac{1}{N}\sum_{i=1}^{N} I_{\theta_0}(x_i).$$

To obtain our result, we need to understand the tail behavior of this sum. We proceed by analyzing the dynamical system

$$A_{N-1} \to A_N = \frac{1}{N}((N-1)A_{N-1} + B_N),$$

with $A_N$ denoting the negative Hessian of a suitable gaussian approximation to the posterior likelihood after $N$ trials at $\theta_N$, its maximizer; $B_N$ denotes the negative Hessian of $\log p(y_N|x_N, \theta)$ at $\theta_N$. This map gives a rough (but asymptotically accurate) approximation of the effect of the $N$th sample on the (near-gaussian) posterior, after a suitable variance stabilizing. It is clear that the range of this map is asymptotically contained in $\mathrm{co}(I_{\theta_0}(x))$, which we have defined as the closure of all convex combinations of the available Fisher information matrices, $I_{\theta_0}(x)$, with $x$ ranging over the full sample space $X$. It is equally clear, when we examine the above dynamical system over multiple trials (thus, roughly, averaging over multiple $B_N$), that the information-maximization strategy is asymptotically doing something like gradient ascent on $\log|A_N|$ (the asymptotic negative entropy of the gaussian posterior, up to an irrelevant scale factor), with the allowed ascent directions taking values within $\mathrm{co}(I_{\theta_N}(x))$, which in turn, by the consistency lemma and the continuity of $I_\theta(x)$, converges to $\mathrm{co}(I_{\theta_0}(x))$. Our asymptotic variance formula now follows from the strict concavity of the function $\log|C|$ in $C$, where $C$ ranges over the symmetric, positive semidefinite (covariance)

matrices (Cover & Thomas, 1991; Lewis, 1996), the fact that $|C|$ and $\log|C|$ have identical maximizers, and the compactness and convexity of $\mathrm{co}(I_{\theta_0}(x))$.

It is worth noting that this proof goes through essentially unchanged if we sample to optimize something like a weighted mean-square error instead of mutual information. In this case, the sampler will asymptotically attempt to minimize a matrix function of the form

$$\mathrm{tr}\left(V^t \sigma_N^2 V\right) \approx \mathrm{tr}\left(V^t A_N^{-1} V\right),$$

where $V$ is some weight matrix. Since the above function is convex in $A_N$ (Lewis, 1996), the only change we need to make is in the final form of the asymptotic variance formula: for this problem,

$$N\sigma_N^2 \to \left(\mathrm{argmin}_{C \in \mathrm{co}(I_{\theta_0}(x))} \mathrm{tr}(V^t C^{-1} V)\right)^{-1},$$

where once again the optimum is well defined (and unique when $V$ is of full rank). When $\Theta$ is one-dimensional, these two approaches clearly lead to the same asymptotic result.

## Appendix B: Kuhn-Tucker Optimality for Bayesian D-Optimal Design

We briefly describe the necessary and sufficient conditions for optimality in the batch experiment setting discussed in section 6.2. We follow Cover and Thomas (1991). We are trying to maximize the function 6.1, which is concave in $p(x)$ (implying that the maximizers we seek form a nonempty convex set). We need to compute the derivative of this function along convex lines through $p(x)$, as follows:

$$\begin{aligned}
V(p,q) &\equiv \frac{\partial}{\partial t}\left( E_\theta \log\left| \int I_\theta(x) d(tq(x) + (1-t)p(x)) \right| \right)\Bigg|_{t=0} \\
&= E_\theta \frac{\partial}{\partial t}\left( \log\left| \int I_\theta(x) d(tq(x) + (1-t)p(x)) \right| \right)\Bigg|_{t=0} \\
&= E_\theta \frac{\partial}{\partial t}\left( \log\left| (1-t)I + t\left( \int I_\theta(x) dp(x) \right)^{-1} \int I_\theta(x) dq(x) \right| \right)\Bigg|_{t=0} \\
&= E_\theta \left( \mathrm{tr}\left( \left( \int I_\theta(x) dp(x) \right)^{-1} \int I_\theta(x) dq(x) \right) \right) - \dim \Theta.
\end{aligned}$$

The interchange of derivative and expectation can be justified by dominated convergence under the conditions of theorem 1.

In the discrete $X$ case, Kuhn-Tucker now implies that for optimal $p(x)$ (and only optimal $p(x)$),

$$\frac{1}{\dim \Theta} E_\theta \left( \mathrm{tr} \left( \left( \int I_\theta(x) dp(x) \right)^{-1} I_\theta(x) \right) \right) \begin{cases} = 1 & \text{if} \quad p(x) > 0 \\ \leq 1 & \text{if} \quad p(x) = 0. \end{cases}$$

Similar results can be derived for more general $X$ by the usual approximation techniques. (For further discussion, see, e.g., Bell & Cover, 1980; Cover & Thomas, 1991; Clyde & Chaloner, 1996.)

## Acknowledgments

We thank E. Simoncelli, C. Machens, and D. Pelli for helpful conversations. This work was partially supported by a predoctoral fellowship from HHMI and by funding from the Gatsby Charitable Trust. A brief account of this work appeared in the conference proceedings of the 16th Annual NIPS meeting, Vancouver, B.C., 2003.

## References

Axelrod, S., Fine, S., Gilad-Bachrach, R., Mendelson, S., & Tishby, N. (2001). *The information of observations and application for active learning with uncertainty* (Tech. Rep.). Jerusalem: Leibniz Center, Hebrew University. Available online: citeseer.nj.nec.com/axelrod01information.html.

Barron, A., Schervish, M., & Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Annals of Statistics*, *27*, 536–561.

Bell, R., & Cover, T. (1980). Competitive optimality of logarithmic investment. *Mathematics of Operations Research*, *5*, 161–166.

Berger, J., Bernardo, J., & Mendoza, M. (1989). On priors that maximize expected information. In J. Klein & H. J. Lee (Eds.), *Recent developments of statistics and its applications* (pp. 1–20). Seoul: Freedom Academy.

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, *10*, 273–304.

Chichilnisky, E. (2001). A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, *12*, 199–213.

Clarke, B., & Barron, A. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, *36*, 453–471.

Clarke, B., & Barron, A. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning Inference*, *41*, 37–60.

Clyde, M., & Chaloner, K. (1996). The equivalence of constrained and weighted designs in multiple objective design problems. *Journal of the American Statistical Association*, *91*, 1236–1244.

Cohn, D., Ghahramani, Z., & Jordan, M. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, *4*, 129–145.

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

Dayan, P., & Abbott, L. (2001). *Theoretical neuroscience*. Cambridge, MA: MIT Press.

Deignan, P., Meckl, P., Franchek, M., Abraham, J., & Jaliwala, S. (2000). *Using mutual information to pre-process input data for a virtual sensor.* Paper presented at the American Control Conference 2000, Chicago.

Dembo, A., & Zeitouni, O. (1993). *Large deviations techniques and applications*. New York: Springer.

Denzler, J., & Brown, C. (2000). Optimal selection of camera parameters for state estimation of static systems: An information theoretic approach. (University of Rochester Tech. Rep. 732). Rochester, NY: University of Rochester.

Fedorov, V. (1972). *Theory of optimal experiments*. New York: Academic Press.

Foldiak, P. (2001). Stimulus optimisation in primary visual cortex. *Neurocomputing*, *38–40,* 1217–1222.

Freund, Y., Seung, H. S., Shamir, E., & Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Machine Learning*, *28*(2–3), 133–168.

Kontsevich, L., & Tyler, C. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, *39*, 2729–2737.

Lee, T., & Yu, S. (1998). An information-theoretic framework for understanding saccadic behaviors. In S. A. Solla, T. K. Leen, & K.-R. Muller (Eds.), *Advances in neural information processing, 12*. Cambridge, MA: MIT Press.

Lewis, A. (1996). Convex analysis on the Hermitian matrices. *SIAM Journal on Optimization*, *6*, 164–177.

Lindley, D. (1956). On a measure of information provided by an experiment. *Annals of Mathematical Statistics*, *29*, 986–1005.

Luttrell, S. (1985). The use of transinformation in the design of data sampling schemes for inverse problems. *Inverse Problems*, *1*, 199–218.

Machens, C. (2002). Adaptive sampling by information maximization. *Physical Review Letters*, *88*, 228104–228107.

Mackay, D. (1992). Information-based objective functions for active data selection. *Neural Computation*, *4*, 589–603.

Mascaro, M., & Bradley, D. (2002). *Optimized neuronal tuning algorithm for multichannel recording*. Unpublished abstract. Available online: http://www.compscipreprints.com/.

Milman, V., & Schechtman, G. (1986). *Asymptotic theory of finite dimensional normed spaces*. Berlin: Springer-Verlag.

Nelken, I., Prut, Y., Vaadia, E., & Abeles, M. (1994). In search of the best stimulus: An optimization procedure for finding efficient stimuli in the cat auditory cortex. *Hearing Research*, *72*, 237–253.

Nelson, J., & Movellan, J. (2000). Active inference in concept learning. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing, 13*. Cambridge, MA: MIT Press.

Parmigiani, G. (1998). Designing observation times for interval censored data. *Sankhya A*, *60*, 446–458.

Parmigiani, G., & Berry, D. (1994). Applications of Lindley information measure to the design of clinical experiments. In A. F. M. Smith & P. Freeman (Eds.), *Aspects of uncertainty: A tribute to D. V. Lindley* (pp. 333–352). New York: Wiley.

Pelli, D. (1987). The ideal psychometric procedure. *Investigative Ophthalmology and Visual Science (Suppl.)*, *28*, 366.

Rudin, W. (1973). *Functional analysis*. New York: McGraw-Hill.

Sahani, M. (1997). *Interactively exploring a neural code by active learning*. Poster session presented at NIC97 meeting, Snowbird, Utah. Available online: http://www.gatsby.ucl.ac.uk/~maneesh/conferences/nic97/poster/home.html.

Schervish, M. (1995). *Theory of statistics*. New York: Springer-Verlag.

Scholl, H. R. (1998, June). Shannon optimal priors on i.i.d. statistical experiments converge weakly to Jeffreys' prior. *Test*, 7(no. 1). Available online: citeseer.nj.nec.com/104699.html.

Schwartz, L. (1967). On Bayes procedures. *Z. Wahrsch. Verw. Gabiete*, *4*, 10–26.

Simoncelli, E., Paninski, L., Pillow, J., & Schwartz, O. (2004). Characterization of neural responses with stochastic stimuli. In M. Gazzaniga (Ed.), *The cognitive neurosciences*. (3rd ed.). Cambridge, MA: MIT Press.

Sollich, P. (1996). Learning from minimum entropy queries in a large committee machine. *Physical Review E*, *53*, R2060–R2063.

Talagrand, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publ. Math. IHES*, *81*, 73–205.

Tzanakou, E., Michalak, R., & Harth, E. (1979). The alopex process: Visual receptive fields by response feedback. *Biological Cybernetics*, *35*, 161–174.

van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge: Cambridge University Press.

Watson, A., & Fitzhugh, A. (1990). The method of constant stimuli is inefficient. *Perception and Psychophysics*, *47*, 87–91.

Watson, A., & Pelli, D. (1983). QUEST: A Bayesian adaptive psychophysical method. *Perception and Psychophysics*, *33*, 113–120.

---

Received June 2, 2003; accepted January 5, 2005.