Design of experiments via information theory

Liam Paninski
Center for Neural Science
New York University
New York, NY 10003
liam@cns.nyu.edu

February 19, 2003

Abstract

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 the more usual nonadaptive experimental design.

1 Introduction

One simple model of experimental design (we have neurophysiological experiments in mind, but our results are all completely general with respect to the identity of the system under study) is as follows. We have some set of input stimuli, X, and some knowledge of how the system should respond to every stimulus, x, in X. Our knowledge of how the system responds is summarized in the form of a prior distribution, $p_0(\theta)$, on the space Θ of all models θ. A model is a set of probabilistic input-output relationships: regular conditional distributions $p(y|x, \theta)$ on Y, the set of possible output responses, given each x in X. Thus the joint probability of θ, x, and y is specified by the following:

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

The "design" of an experiment is given by the choice of input probability $p(x)$. We want to design our experiment — choose $p(x)$ — optimally in some sense. 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 thus suggests we choose $p(x)$ to optimize the following objective function:

$$I(\{x, y\}; \theta) \tag{1}$$

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

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

We should note that similar ideas have seen application in a wide and somewhat scattered literature; for a partial bibliography, see the longer draft of this paper at http://www.cns.nyu.edu/∼liam. Somewhat surprisingly, we have not seen any applications of the information-theoretic objective function (1) to the design of neurophysiological experiments (although see the abstract by [2], who seem to have independently implemented the same idea in a simulation study).

The primary goal of this paper 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 (section 2) states that, under acceptably weak conditions on the models $p(y|\theta, x)$, the information-maximization strategy leads to consistent and efficient estimates of the true underlying model, in a natural sense. We also give a few simple examples to illustrate the applicability of our results (section 3).

2 Results

First, we note that the problem as posed in the introduction turns out to be slightly easier than one might have expected, because $I(\{x, y\}; \theta)$ is linear in $p(x)$. This, in turn, implies that $p(x)$ 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 θ and y, given a single input x:

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

Our main result is a "Bernstein-von Mises" - type theorem [4]. 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 θ_0) and the likelihood ratios are sufficiently smooth on average, then the posterior distributions $p_N(\theta)$ are asymptotic normal, with easily calculable asymptotic mean and variance. We adapt this result to the present case, where x is chosen according to the information-maximization recipe. It turns out that the hard part is proving consistency (c.f. section 3); we give the basic consistency lemma (interesting in its own right) first, from which the main theorem follows fairly easily.

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

1. *The parameter space Θ is compact.*
2. *The loglikelihood $\log p(y|\theta, x)$ is Lipschitz in θ, uniformly in θ, x, and y, with respect to some dominating measure.*
3. *The prior measure p_0 assigns positive measure to any neighborhood of θ_0.*
4. *The maximal divergence $\sup_x D_{KL}(\theta_0; \theta|x)$ is positive for all $\theta \neq \theta_0$.*
5. *The set $Z_x = \{\theta : D_{KL}(\theta_0; \theta|x) = 0\}$ has zero prior measure for all $x \in X$.*

Then the posteriors are consistent: $p_N(U) \to 1$ in probability for any neighborhood U of θ_0.

Theorem 2 (Asymptotic normality). *Assume the conditions of Lemma 1, stengthened as follows:*

1. *Θ has a smooth, finite-dimensional manifold structure in a neighborhood of θ_0.*
2. *The loglikelihood $\log p(y|\theta, x)$ is continuously differentiable in θ, uniformly in θ, x, and y, with respect to some dominating measure. Moreover, the Fisher information matrices*

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

   where the differential $\dot{p}$ is taken with respect to θ, are well-defined and continuous in θ, uniformly in x, θ in some neighborhood of θ_0.

3. *The prior measure p_0 is absolutely continuous in some neighborhood of θ_0, with a continuous positive density at θ_0.*

Then

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

in probability, where $||.||$ denotes variation distance and $\mathcal{N}(\mu_N, \sigma_N^2)$ denotes the normal density with mean μ_N and variance σ_N^2. Here

$$\sigma_N^2 = \left(\sum_{i=1}^{N} I_{\theta_0}(x) \right)^{-1},$$

and μ_N is asymptotically normal with mean θ_0 and variance σ_N^2.

Corollary 3. *If the determinant of the Fisher information matrices $I_{\theta_0}(x)$ has a unique maximum for some $I_{\theta_0}(x)'$, then*

$$N\sigma_N^2 \to (I_{\theta_0}(x)')^{-1}.$$

Thus, under these conditions, the information maximization strategy works, and works better than the i.i.d. x strategy (where the asymptotic variance σ^2 is inversely related to an average, not a maximum, over x, and is therefore generically larger).

It should also be clear that we have not stated the results as generally as possible; we have chosen instead to use assumptions which 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 examples, for example, by assuming that parameters take values in bounded (though possibly large) sets and that tuning curves are not infinitely steep and do not have too many uninformative "flat" regions.

3 Examples

We have space here for just one application of the above results; again, see http://www.cns.nyu.edu/∼liam for more examples. It is worth noting that a couple of the most surprising examples there are negative, in a sense, serving to illustrate the nontriviality of our mathematical results.

3.1 Psychometric model

As noted in the introduction, psychophysicists have employed versions of the information-maximization procedure for some years [6, 3, 5, 1]; references in [5], for example, go back four decades. Our results above allow us to precisely quantify the effectiveness of this stategy.

One general psychometric model is as follows. The response space Y is binary. 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 $f_{a,\theta} = f((t - \theta)/a)$; θ here serves as a location ("threshold") parameter, while a sets the scale (we assume a is known, for now, although of course this can be relaxed [1]). Finally, let $p(x)$ and $p_0(\theta)$ be some fixed sampling and prior distributions, respectively, both equivalent to Lebesgue measure on some interval Θ.

To apply our main result to this model, we have to relax one of the conditions in the consistency lemma: if f is constant on any interval, the condition on the uninformative sets Z_x will typically not hold. Thus, we state a modified version of the lemma:

Lemma 4 (Consistency under monotonicity). *Posterior consistency holds under the monotonicity condition on the psychometric function f; thus, the condition on the sets Z_x is unnecessary in this context.*

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 the corollary to theorem 2 that the most efficient estimator of θ is asymptotically unbiased with asymptotic variance

$$\sigma^2_{info} \approx (N \sup_x I_{\theta_0}(x))^{-1},$$

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

$$\sigma^2_{iid} \approx (N \int_X dp(x) I_{\theta_0}(x))^{-1}.$$

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 information-maximization strategy; for all other f, information maximization is strictly more efficient. The extremal function f^* is the unique solution of the following differential equation:

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

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 of the interval $[-\pi/2c, \pi/2c]$, then σ^2_{iid} is always strictly greater than σ^2_{info}. 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 greatly preferable.

Second, a crude analysis shows that, as the scale of the nonlinearity $1/a$ shrinks, the ratio $\sigma^2_{iid}/\sigma^2_{info}$ grows approximately as 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.

Acknowledgements

We thank R. Sussman, E. Simoncelli, and D. Pelli for helpful conversations. LP is supported by a predoctoral fellowship from HHMI.

References

[1] L. Kontsevich and C. Tyler. Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, 1999.

[2] M. Mascaro and D. Bradley. Optimized neuronal tuning algorithm for multichannel recording. Unpublished abstract at http://www.compscipreprints.com/comp/Preprint/bradleydavid1961/20020210.4/2/, 2002.

[3] D. Pelli. The ideal psychometric procedure. *Investigative Ophthalmology and Visual Science (Supplement)*, 28:366, 1987.

[4] A. van der Vaart. *Asymptotic statistics*. Cambridge University Press, Cambridge, 1998.

[5] A. Watson and A. Fitzhugh. The method of constant stimuli is inefficient. *Perception and Psychophysics*, 47:87–91, 1990.

[6] A. Watson and D. Pelli. Quest: a bayesian adaptive psychophysical method. *Perception and Psychophysics*, 33:113–120, 1983.