

Adaptive Design Optimization: A Mutual Information-Based Approach to Model Discrimination in Cognitive Science

Daniel R. Cavagnaro

cavagnaro.2@osu.edu

Jay I. Myung

myung.1@osu.edu

Mark A. Pitt

pitt.2@osu.edu

Department of Psychology, Ohio State University, Columbus, OH 43201, U.S.A.

Janne V. Kujala

vkujala@jyu.fi

University of Jyväskylä, Agora Center, Jyväskylä FIN-40014, Finland

Discriminating among competing statistical models is a pressing issue for many experimentalists in the field of cognitive science. Resolving this issue begins with designing maximally informative experiments. To this end, the problem to be solved in adaptive design optimization is identifying experimental designs under which one can infer the underlying model in the fewest possible steps. When the models under consideration are nonlinear, as is often the case in cognitive science, this problem can be impossible to solve analytically without simplifying assumptions. However, as we show in this letter, a full solution can be found numerically with the help of a Bayesian computational trick derived from the statistics literature, which recasts the problem as a probability density simulation in which the optimal design is the mode of the density. We use a utility function based on mutual information and give three intuitive interpretations of the utility function in terms of Bayesian posterior estimates. As a proof of concept, we offer a simple example application to an experiment on memory retention.

1 Introduction ---

Experimentation is fundamental to the advancement of science, whether one is interested in studying the neuronal basis of a sensory process in cognitive science or assessing the efficacy of a new drug in clinical trials. In an adaptive experiment, the information learned from each test is used to adapt subsequent tests to be maximally informative in an appropriately defined sense. The problem to be solved in adaptive design optimization (ADO) is to identify an experimental design under which one can infer the

underlying model in the fewest possible steps. This is particularly important in cases where measurements are costly or time-consuming.

Because of its flexibility and efficiency, the use of adaptive designs has become popular in many fields of science. For example, in astrophysics, ADO has been used in the design of experiments to detect extrasolar planets (Loredo, 2004). ADO has also been used in designing phase I and phase II clinical trials to ascertain the dose-response relationship of experimental drugs (Haines, Perevozskaya, & Rosenberer, 2003; Ding, Rosner, & Müller, 2008), as well as in estimating psychometric functions (Kujala & Lukka, 2006; Lesmes, Jeon, Lu, & Doshier, 2006).

Bayesian decision theory offers a principled approach to the ADO problem. In this framework, each potential design is treated as a gamble whose payoff is determined by the outcome of an experiment carried out with that design. The idea is to estimate the utilities of hypothetical experiments carried out with each design so that an "expected utility" of each design can be computed. This is done by considering every possible observation that could be obtained from an experiment with each design and then evaluating the relative likelihoods and statistical values of these observations. The design with the highest expected utility value is then chosen as the optimal design.

Natural metrics for the utility of an experiment can be found in information theory. This was first pointed out by Lindley (1956), who suggested maximization of Shannon information as a sensible criterion for design optimization. MacKay (1992) was one of the first to apply such a criterion to ADO, using the expected change in entropy from one stage of experimentation to the next as the utility function. A few other information-based utility functions have been proposed, including cross-entropy, Kullback-Leibler divergence, and mutual information (Cover & Thomas, 1991). In particular, the desirability and usefulness of the latter was formally justified by Paninski (2005) who proved that under acceptably weak modeling conditions, the adaptive approach with a utility function based on mutual information leads to consistent and efficient parameter estimates.

Despite its theoretical appeal, the complexity of computing mutual information directly has proved to be a major implementational challenge (Bernardo, 1979; Paninski, 2003, 2005). Consequently, most design optimization research has been restricted to special cases such as linear gaussian models. For example, Lewi, Butera, and Paninski (2009) offer a fast algorithm for finding the design of a neurophysiology experiment that maximizes the mutual information between the observed data and the parameters of a generalized linear model. Using a gaussian approximation of the posterior distribution to facilitate estimation of the mutual information, the algorithm decreases the uncertainty of the parameter estimates much faster than an independent and identically distributed (i.i.d.) design and converges to the asymptotically optimal design. Other special cases

can also facilitate the implementation of the mutual information-based approach. For example, Kujala and Lukka (2006) and Kujala, Richardson, and Lyytinen (in press) successfully implemented mutual information-based utility functions for estimating psychometric functions and for the design of adaptive learning games, respectively, with direct computation made possible by the binary nature of the experimental outcomes.

The need for fast and accurate design optimization algorithms that can accommodate nonlinear models has grown with recent developments of such models in cognitive science, such as those found in memory retention (Rubin & Wenzel, 1996; Wixted & Ebbesen, 1991), category learning (Nosofsky & Zaki, 2002; Vanpaemel & Storms, 2008), and numerical estimation (Opfer & Siegler, 2007). This problem has also been approached in the astrophysics literature by Loredo (2004), who shows that so-called maximum entropy sampling can be used to find the design that maximizes the expected Shannon information of the posterior parameter estimates. This approach addresses the problem of ADO for parameter estimation but not for model discrimination. The latter problem is significantly more complex because it requires integration over the space of models in addition to the integration over each model's parameter space.

The problem of design optimization for discrimination of nonlinear models is considered in a nonadaptive setting by Heavens, Kitching, and Verde (2007) and by Myung and Pitt (2009). Heavens et al. used a Laplace approximation of the expected Bayes' factor as their utility function and compared only nested models. Myung and Pitt consider the problem much more generally. Rather than using an information-theoretic utility function, they use a utility function based on the minimum description length principle (Grünwald, 2005). They bring to bear advanced stochastic Bayesian optimization techniques that allow them to find optimal designs for discriminating among even highly complex, nonnested, nonlinear models.

In this letter, we address the design optimization problem for discrimination of nonlinear models in an adaptive setting. Following Paninski (2003, 2005), Kujala and Lukka (2006), Lewi et al. (2009), and Kujala et al. (in press), we use a utility function based on mutual information. That is, we measure the utility of a design by the amount of information, about the relative likelihoods of the models in question that would be provided by the results of an experiment with that given design. Further, following Myung and Pitt (2009), we apply a simulation-based approach for finding the full solution to the design optimization problem without relying on linearization, normalization, or approximation, as has often been done in the past. We apply a Bayesian computational trick that was introduced in the statistics literature (Müller, Sanso, & De Iorio, 2004), which allows the optimal design to be found without evaluating the high-dimensional integration and optimization directly. Briefly, the idea is to recast the problem as a density simulation in which the optimal design corresponds to the mode

of the density. The density is simulated with an interacting particle filter, and the mode is found by gradually sharpening up the distribution with simulated annealing. We also give several intuitive interpretations of the mutual information-based utility function in terms of Bayesian posterior estimates, which both elucidates the logic of the algorithm and connects it with common statistical approaches to model selection in cognitive science. Finally, we demonstrate the approach with a simple example application to an experiment on memory retention. In simulated experiments, the optimal adaptive design outperforms all other competitors at identifying the data-generating model.

2 Bayesian ADO Framework

Adaptive design optimization within a Bayesian framework has been considered at length in the statistics community (Kiefer, 1959; Box & Hill, 1967; Chaloner & Verdinelli, 1995; Atkinson & Donev, 1992) as well as in other science and engineering disciplines (e.g., El-Gamal & Palfrey, 1996; Bardsley, Wood, & Melikhova, 1996; Allen, Yu, & Schmitz, 2003). The issue is essentially a Bayesian decision problem where, at each stage of experimentation, the most informative design (i.e., the design with the highest expected utility) is chosen based on the outcomes of the previous experiments. The criterion for the informativeness of a design often depends on the goals of the experimenter. The experiment that yields the most precise parameter estimates may not be the most effective at discriminating among competing models, for example (see Nelson, 2005, for a comparison of several utility functions that have been used in cognitive science research).

Whatever the goals of the experiment may be, solving for the optimal design is a highly nontrivial problem. The computation requires simultaneous optimization and high-dimensional integration, which can be analytically intractable for the complex, nonlinear models as often used in many real-world problems. Formally, ADO for model discrimination entails finding an optimal design that maximizes a utility function $U(d)$,

$$d^* = \operatorname{argmax}_d \{U(d)\}, \quad (2.1)$$

with the utility function defined as

$$U(d) = \sum_{m=1}^K p(m) \int \int u(d, \theta_m, y) p(y|\theta_m, d) p(\theta_m) dy d\theta_m, \quad (2.2)$$

where $m = \{1, 2, \dots, K\}$ is one of a set of K models being considered, d is a design, y is the outcome of an experiment with design d under model m , and θ_m is a parameterization of model m . We refer to the function $u(d, \theta_m, y)$

in equation 2.2 as the local utility of the design d . It measures the utility of a hypothetical experiment carried out with design d when the data-generating model is m , the parameters of the model take the value θ_m , and the outcome y is observed. Thus, $U(d)$ represents the expected value of the local utility function, where the expectation is taken over all models under consideration, the full parameter space of each model, and all possible observations given a particular model and parameter pair with respect to the model prior probability $p(m)$, the parameter prior distribution $p(\theta_m)$, and the sampling distribution $p(y|\theta_m, d)$, respectively.

The model and parameter priors are being updated on each stage $s = \{1, 2, \dots\}$ of experimentation. Specifically, on the specific outcome z_s observed at stage s of an actual experiment carried out with design d_s , the model and parameter priors to be used to find an optimal design at the next stage are updated by Bayes' rule and Bayes' factor calculation (e.g., Gelman, Carlin, Stern, & Rubin, 2004) as

$$p_{s+1}(\theta_m) = \frac{p(z_s|\theta_m, d_s) p_s(\theta_m)}{\int p(z_s|\theta_m, d_s) p_s(\theta_m) d\theta_m} \quad (2.3)$$

$$p_{s+1}(m) = \frac{p_0(m)}{\sum_{k=1}^K p_0(k) BF_{(k,m)}(z_s)_{p_s(\theta)}}, \quad (2.4)$$

where $BF_{(k,m)}(z_s)_{p_s(\theta)}$ denotes the Bayes' factor defined as the ratio of the marginal likelihood of model k to that of model m given the realized outcome z_s , where the marginal likelihoods are computed with the updated priors from the preceding stage. This updating scheme is applied successively on each stage of experimentation, after an initialization with equal model priors $p_{(s=0)}(m) = 1/K$ and a noninformative parameter prior $p_{(s=0)}(\theta_m)$.

3 Computational Methods

To find the optimal design d^* in a general setting is exceedingly difficult. Given the multiple computational challenges involved, standard optimization methods such as Newton-Raphson are out of question. However, a promising new approach to this problem has been proposed in statistics (Müller et al., 2004). It is a simulation-based approach that includes an ingenious computational trick that allows one to find the optimal design without having to evaluate the integration and optimization directly in equations 2.1 and 2.2. The basic idea is to recast the design optimization problem as a simulation from a sequence of augmented probability models.

To illustrate how it works, let us consider the design optimization problem to be solved at any given stage s of experimentation, and, for simplicity, we will suppress the subscript s in the remainder of this section. According to the computational trick of Müller et al. (2004), we treat the design d as a random variable and define an auxiliary distribution $h(d, \cdot)$ that admits

$U(d)$ as its marginal density. Specifically, we define

$$h(d, y_1, \theta_1, \dots, y_K, \theta_K) = \alpha \left[\sum_{m=1}^K p(m) u(d, \theta_m, y_m) \right] p(y_1, \theta_1, \dots, y_K, \theta_K | d), \quad (3.1)$$

where $\alpha(>0)$ is the normalizing constant of the auxiliary distribution and

$$p(y_1, \theta_1, \dots, y_K, \theta_K | d) = \prod_{m=1}^K p(y_m | \theta_m, d) p(\theta_m). \quad (3.2)$$

Note that the subscript m in equations 3.1 and 3.2 refers to model m , not the stage of experimentation. For instance, y_m denotes an experimental outcome generated from model m with design d and parameter θ_m .

Marginalizing $h(d, \cdot)$ over $(y_1, \theta_1, \dots, y_K, \theta_K)$ yields

$$h(d) = \int \dots \int h(d, y_1, \theta_1, \dots, y_K, \theta_K) dy_1 d\theta_1 \dots dy_K d\theta_K \quad (3.3)$$

$$= \alpha \sum_{m=1}^K p(m) \int \int u(d, \theta_m, y_m) p(y_m | \theta_m, d) p(\theta_m) dy_m d\theta_m \quad (3.4)$$

$$= \alpha U(d). \quad (3.5)$$

Consequently, the design with the highest utility can be found by taking the mode of a sufficiently large sample from the marginal distribution $h(d)$. However, finding the global optimum could potentially require a very large number of samples from $h(d)$, especially if there are many local optima or if the design space is very irregular or high-dimensional. To overcome this problem, assuming $h(d, \cdot)$ is nonnegative and bounded,¹ we augment the auxiliary distribution with independent samples of y 's and θ 's given design d as follows,

$$h_J(d, \cdot) = \alpha_J \prod_{j=1}^J h(d, y_{1,j}, \theta_{1,j}, \dots, y_{K,j}, \theta_{K,j}), \quad (3.6)$$

¹Negative values of $h(d, \cdot)$ can be handled in the implementation by adding a small constant to the distribution and truncating it at zero. This transformation does not change the location of the global maximum, provided that the truncated values are not too extremely negative. However, adding a constant does decrease the relative concentration of the distribution around the global maximum, making it more difficult to find.

for a positive integer J and $\alpha_J (>0)$. The marginal distribution of $h_J(d)$ obtained after integrating out model parameters and outcome variables will then be equal to $\alpha_J U(d)^J$. Hence, as J increases, the distribution $h_J(d)$ will become more highly peaked around its (global) mode corresponding to the optimal design d^* , thereby making it easier to identify the mode.

Following Amzal, Bois, Parent, and Robert (2006), we implemented a sequential Monte Carlo (particle filter) algorithm that begins by simulating $h_J(d, \cdot)$ in equation 3.6 for $J = 1$ and then increases J incrementally on subsequent iterations on an appropriate simulated annealing schedule (Kirkpatrick, Gelatt, & Vecchi, 1983; Doucet, de Freitas, & Gordon, 2001).

4 Mutual Information Utility

Selection of a utility function that adequately captures the goals of the experiment is an integral, often crucial, part of design optimization. A design that is optimal for parameter estimation is not necessarily optimal for model selection. Perhaps the most studied optimization criterion in the design optimization literature is minimization of the variance of parameter estimates. In the case of linear models, this is achieved by maximizing the determinant of the variance-covariance matrix, which is called the D-optimality criterion (Atkinson & Donev, 1992). For nonlinear models, a sensible choice of utility function is the negative entropy of the posterior parameter estimates after observing experimental outcomes (Loredo, 2004; Kück, de Freitas, & Doucet, 2006). It has been shown that such an entropy-based utility function also leads to D-optimality in the linear gaussian case (Bernardo, 1979).

Implicit in the preceding optimality criteria is the assumption that the underlying model is correct. Quite often, however, the researcher entertains multiple models and wishes to design an experiment that can effectively distinguish them. One way to achieve this goal is to minimize model mimicry (i.e., the ability of a model to account for data generated by a competing model). To this end, the T-optimality criterion maximizes the sum-of-squares error between data generated from a model and the best-fitting prediction of another competing model (Atkinson & Federov, 1975a, 1975b). In practice, however, sum-of-squares error is a poor choice for model discrimination because it is biased toward more complex models (e.g., Myung, 2000). As an alternative, one can use a statistical model selection criterion such as the Akaike information criterion (Akaike, 1973), the Bayes factor (Kass & Raftery, 1995), or the minimum description length principle (Grünwald, 2005; Myung & Pitt, 2009; Balasubramanian, Larjo, & Seth, 2008).

One can also construct a utility function motivated from information theory (Cover & Thomas, 1991). In particular, mutual information seems to provide an ideal measure for quantifying the value of an experiment design. Specifically, mutual information measures the reduction in uncertainty about one variable that is provided by knowledge of the value of

the other random variable. Formally, the mutual information of a pair of random variables P and Q , taking values in \mathcal{X} , is given by

$$I(P; Q) = H(P) - H(P|Q), \quad (4.1)$$

where $H(P) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$ is the entropy of P , and $H(P|Q) = \sum_{x \in \mathcal{X}} p(x) H(P|Q = x)$ is the conditional entropy of P given Q . A high mutual information indicates a large reduction in uncertainty about P due to knowledge of Q . For example, if the two distributions were perfectly correlated, meaning that knowledge of Q allowed perfect prediction of P , then the conditional distribution would be degenerate, having entropy zero. Thus, the mutual information of P and Q would be $H(P)$, meaning that all of the entropy of P was eliminated through knowledge of Q . Mutual information is symmetric in the sense that $I(P; Q) = I(Q; P)$.

Mutual information can also be defined by a Kullback-Leibler (KL) divergence between a joint distribution and the product of marginal distributions such as $I(P; Q) = D_{KL}((P, Q), PQ)$, where $D_{KL}(P, Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$ is the KL divergence between the two distributions P and Q . The product PQ represents the hypothetical joint distribution of P and Q in the case that they were independent. Thus, the mutual information of P and Q measures how "far" (in terms of KL divergence) the actual joint distribution is from what it would be if the distributions were independent. For example, if the distributions actually were independent, then the actual and hypothetical joint distributions would be identical and, hence, the KL divergence would be zero, meaning that Q provides no information about P .

Mutual information can be implemented as an optimality criterion in ADO for model discrimination on each stage s ($= 1, 2, \dots$) of experimentation in the following way. (For simplicity, we omit the subscript s in the equations below.) Let M be a random variable defined over a model set $\{1, 2, \dots, K\}$, representing uncertainty about the true model, and let Y be a random variable denoting an experimental outcome. Hence, $\text{Prob.}(M = m) = p(m)$ is the prior probability of model m , and $\text{Prob.}(Y = y|d) = \sum_{m=1}^K p(y|m, d) p(m)$, where $p(y|m, d) = \int p(y|\theta_m, d) p(\theta_m) d\theta_m$, is the associated prior over experimental outcomes given design d . Then $I(M; Y|d) = H(M) - H(M|Y, d)$ measures the decrease in uncertainty about which model drives the process under investigation given the outcome of an experiment with design d . Since $H(M)$ is independent of the design d , maximizing $I(M; Y|d)$ on each stage of ADO is equivalent to minimizing $H(M|Y, d)$, which is the expected posterior entropy of M given d .

Implementing this ADO criterion requires identification of an appropriate local utility function $u(d, \theta_m, y)$ in equation 2.2; specifically, a function

whose expectation over models, parameters, and observations is $I(M; Y|d)$. Such a function can be found by writing

$$I(M; Y|d) = \sum_{m=1}^K p(m) \int \int p(y|\theta_m, d) p(\theta_m) \log \frac{p(m|y, d)}{p(m)} dy d\theta_m, \quad (4.2)$$

from whence it follows that setting $u(d, \theta_m, y) = \log \frac{p(m|y, d)}{p(m)}$ yields $U(d) = I(M; Y|d)$. Thus, the local utility of a design for a given model and experiment outcome is the log ratio of the posterior probability to the prior probability of that model. Put another way, the above utility function prescribes that a design that increases our certainty about the model on the observation of an outcome is more valued than a design that does not.

Another interpretation of this local utility function can be obtained by rewriting it, applying Bayes' rule, as $u(d, \theta_m, y) = \log \frac{p(y|d, m)}{p(y|d)}$. In this form, the local utility can be interpreted as the net informational loss (in terms of KL divergence) incurred from estimating the true distribution P^* over $Y|d$ with the distribution $p(y|d)$ (Haussler & Oppen, 1997). This net loss, or "regret," is the additional loss over that which would have been incurred from estimating P^* if the true model were known (i.e., with $p(y|d, m)$). What this means for ADO is that the observation that is to be made at each stage is the one whose result is the least expected or, equivalently, the most surprising. In a manner of speaking, to learn the most, we should test where we know the least.

This local utility function can be interpreted in yet another way, in terms of Bayes' factors, by rewriting it as

$$u(d, \theta_m, y) = -\log \sum_{k=1}^K p(k) BF_{(k,m)}(y), \quad (4.3)$$

where $BF_{(k,m)}(y) = \frac{p(y|k)}{p(y|m)}$ is the marginal likelihood of model k divided by the marginal likelihood of model m , that is, the Bayes' factor for model k over model m for y .² Examining equation 4.3 more closely, the weighted sum of Bayes' factors quantifies the evidence against m , provided by an observation y , aggregated across head-to-head comparisons of m with each of the models under consideration. Further, the negative sign means that to maximize the local utility is to minimize the aggregate evidence against m . Accordingly, the designs that are favored by the utility function in equation 4.2 are those

²Bayes' factor evaluations within each utility estimate can be done by grid discretization if each model has only a few parameters. More generally, Monte Carlo estimates can be used, but care must be taken to limit sampling error (see Han & Carlin, 2001, for example).

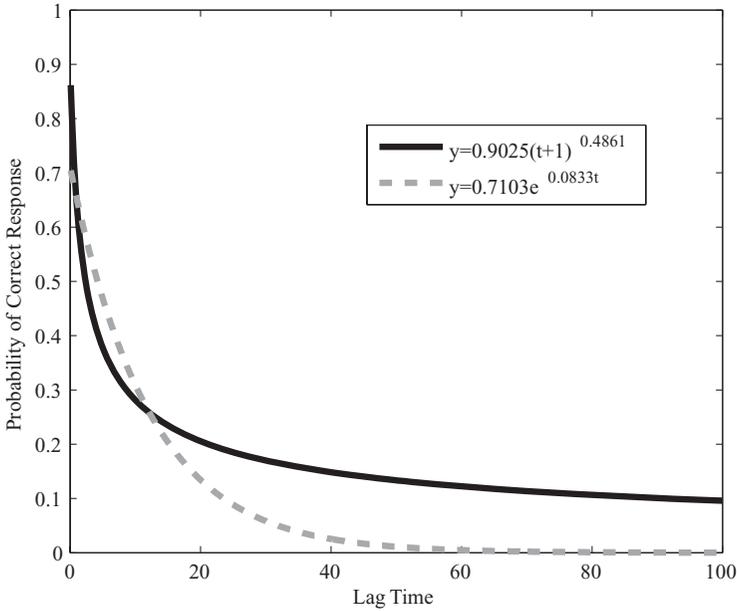


Figure 1: Maximum likelihood estimates for POW (solid lines) and EXP (dashed lines) obtained by Rubin et al. (1999).

that, on average, are expected to produce the least amount of evidence against the true model or, equivalently, the largest amount of evidence for the true model relative to the other models under consideration.

In what follows, we demonstrate the application of the ADO framework for discriminating retention models in cognitive science.

5 Application

A central issue in memory research is the rate of forgetting over time. Of the dozens of retention functions (so-called because the amount of information retained after study is measured) that have been evaluated by researchers, two models, power (POW) and exponential (EXP), have received considerable attention. Both are Bernoulli models, defined by $p = a(t+1)^{-b}$ and $p = ae^{-bt}$, respectively, where p is the probability of correct recall of a stimulus item (e.g., word) at time t , and a , b are model parameters. The maximum likelihood estimates for a data set collected by Rubin, Hinton, and Wenzel (1999) are depicted in Figure 1.

Many experiments have been performed to precisely identify the functional form of retention (see Rubin & Wenzel, 1996, for a thorough review). In a typical retention experiment, data are collected through a sequence of

trials, each of which assesses retention at a single time point, and the data are then aggregated across trials so that a retention curve can be estimated. Each trial consists of study phase, in which a participant is given a list of words to memorize, followed by a test phase, in which retention is assessed by testing how many words the participant can correctly recall from the study list. The length of time between the the study phase and the test phase is called the lag time. The lag times are design variables that can be controlled by the experimenter. Thus, the goal of design optimization is to find the most informative set of lag times for the purpose of discriminating between the power and exponential models.

We conducted computer simulations to illustrate the ADO procedure for discriminating between the power and exponential models of retention, in which optimal designs were sought over a series of stages of experimentation. For simplicity, we considered only designs in which one lag time was tested in each stage of experimentation. This luxury was afforded by two considerations. First, unlike the nonadaptive setting in which all of the lag times must be chosen before experimentation begins, in the adaptive setting we can choose a new lag time after each set of observations. Second, unlike utility functions based on statistical model selection criteria such as minimum description length (MDL), the mutual-information-based utility function does not require computation of the maximum likelihood estimate (MLE) for each model. For these two-parameter models, observations at no fewer than three distinct time points would be required to compute the MLE; hence, an MDL-based utility function would be undefined for a design with fewer than three test phases.³

We used parameter priors $a \sim \text{Beta}(2, 1)$ and $b \sim \text{Beta}(1, 4)$ for POW, and $a \sim \text{Beta}(2, 1)$ and $b \sim \text{Beta}(1, 80)$ for EXP.⁴ Figure 2 depicts a random sample of curves generated by each model with parameters drawn from these priors. At each stage of the simulated experiment, the most informative lag time for discriminating the models was computed, data were generated from POW with $a = 0.9025$ and $b = 0.4861$ (i.e., the MLE for EXP from Rubin et al.) and 10 Bernoulli trials at that time point, and the predictive distributions were updated accordingly. We continued the process for 10 stages of the experiment. A typical profile of the posterior model probability $p_s(\text{POW})$ as a function of stage s is shown by the solid black line in Figure 3.

³The MDL-based utility function is implemented with the Fisher information approximation (FIA) defined as $FIA = -\ln f(y|\hat{\theta}) + \frac{k}{2} \ln \frac{n}{2} + \ln \int \sqrt{|I(\theta)|} d\theta$, where $f(y|\hat{\theta})$ is the maximum likelihood, k is the number of parameters n is the sample size, and $I(\theta)$ is the Fisher information matrix of sample size 1 (Myung & Pitt, 2009).

⁴The priors reflect conventional wisdom about these retention models based on many years of investigation. The choice of priors does indeed change the optimal solution, but the importance of this example is the process of finding a solution, not the actual solution itself.

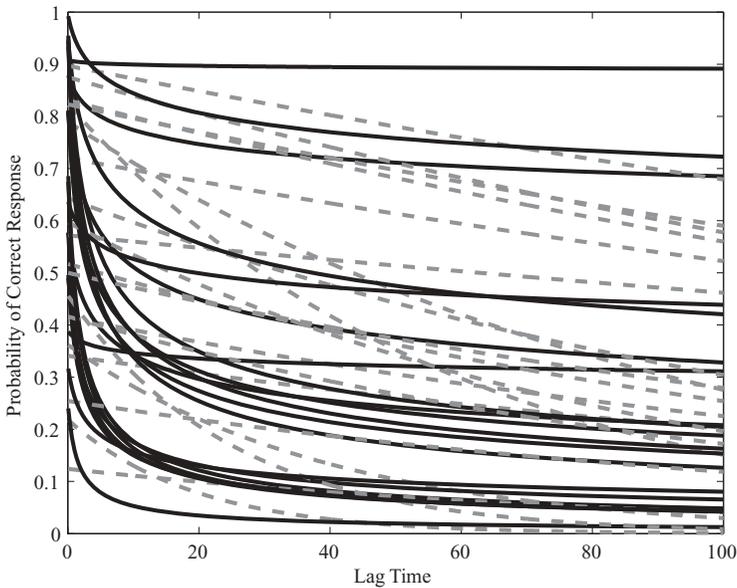


Figure 2: Random sample of curves generated from POW (solid lines) and EXP (dashed lines) illustrating the ability of these models to mimic one another. The models also include binomial error (not shown), which complicates the task of discriminating them.

For comparison, we also conducted several simulated experiments with randomly generated designs. These experiments with random designs proceeded in the manner described above, except that the lag time at each stage was chosen randomly (i.e., from a continuous, uniform distribution) between 0 and 100 seconds. The solid gray line in Figure 3 shows a typical posterior model probability curve obtained in these random experiments.

The results of the experiments with random designs show the advantage of ADO over a less principled approach to designing a sequential experiment, but they do not show how ADO compares with the current standard in retention research. To do that, we conducted additional simulations using a typical design from the retention literature. While there is no established standard for the set of lag times to test in retention experiments, a few conventions have emerged. For one, previous experiments utilize what we call fixed designs, in which the set of lag times at which to assess memory is specified before experimentation begins and held fixed for the duration of the experiment. Thus, there is no Bayesian updating between stages as there would be in a sequential design, such as what would be prescribed by ADO. The lag times are typically concentrated near zero and spaced roughly geometrically. For example, the data set collected by Rubin et al.

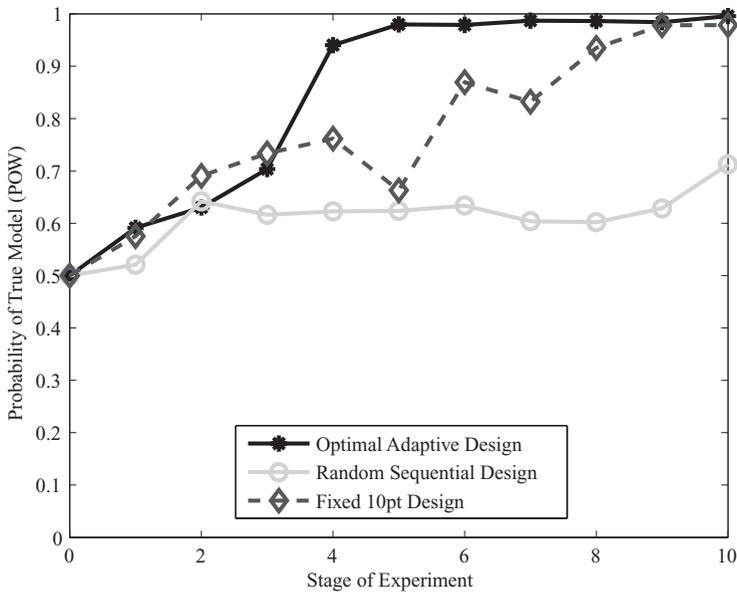


Figure 3: Posterior model probability curves from simulated experiments with each of the three designs, in which data were generated from POW with $a = 0.9025$, $b = 0.4861$, and 10 Bernoulli trials per stage. As suggested by the theory, the optimal adaptive design accumulates evidence for POW faster than either of the competing designs.

(1999) used a design consisting of 10 lag times: (0 s, 1 s, 2 s, 4 s, 7 s, 12 s, 21 s, 35 s, 59 s, 99 s). To get a meaningful comparison between this fixed design and the sequential designs, we generated data at each stage from the same model as in the previous simulations, but with just one Bernoulli trial at each of the 10 lag times in the Rubin et al. design. That way, the “cost” of each stage, in terms of the number of trials, is the same as in the adaptive design. The posterior model probabilities and parameter estimates were computed after each stage from all data up to that point. The obtained posterior model probabilities from a typical simulation are shown by the dashed line in Figure 3.

The results of these simulations clearly demonstrate the efficiency of the optimal adaptive design. The optimal-adaptive-design simulation identifies the correct model with over 0.95 probability after just 4 stages or 40 Bernoulli trials. In contrast, the fixed-design simulation requires twice as many observations (8 stages, or 80 Bernoulli trials) to produce a similar level of evidence in favor of the true model. The random-design simulation does not conclusively discriminate the models even after all 10 stages were complete.

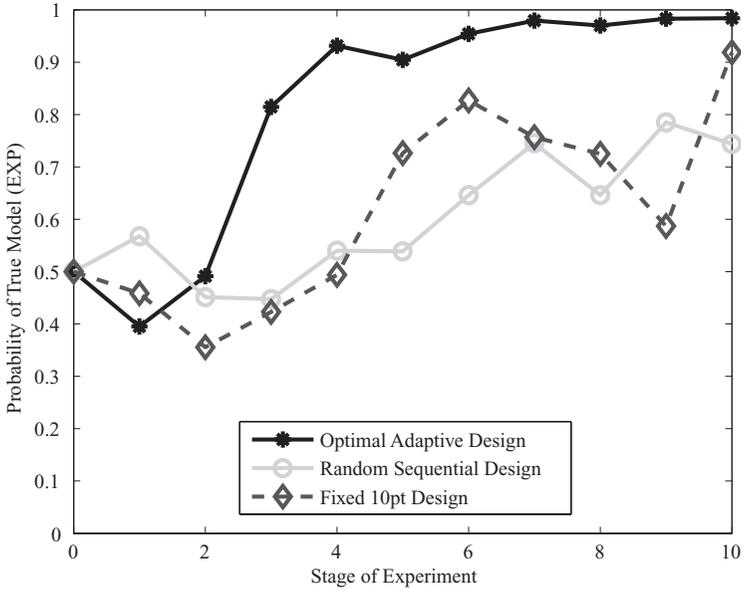


Figure 4: Posterior model probability curves from simulated experiments with each of the three designs, in which data were generated from EXP with $a = 0.7103$, $b = 0.0833$, and 10 Bernoulli trials per stage. Again, the optimal adaptive design accumulates evidence for the true model much faster than either of the competing designs. The nonmonotonic behavior results from the data observed at a given stage being more likely, according to the priors at that stage, under POW than under EXP. Although the data are always generated by EXP in these simulations, such behavior is not surprising given how closely POW can mimic EXP, as shown in Figure 2.

To ensure that the advantage of the optimal adaptive design was not due to the choice of POW as the true model, we repeated each of the simulated experiments with data generated from EXP, with $a = 0.7103$ and $b = 0.0833$ (i.e., the MLE for EXP from Rubin et al. (1999)). The results of these simulations are given in Figure 4, and the advantage of the optimal adaptive design is apparent once again. The optimal-adaptive-design simulation identifies the true model with over 0.93 probability after just 4 stages or 40 Bernoulli trials. This is much quicker accumulation of evidence than in the fixed-design simulation, which requires all 10 stages, or 100 Bernoulli trials, to identify the true model with 0.92 probability. Again, the random design simulation does not conclusively discriminate the models even after all 10 stages were complete.

This example is intended as a proof of concept. In this simple case, an optimal design could have been found by comprehensive grid searches. However, the approach that we have demonstrated here generalizes easily

to much more complex problems in which a brute force approach would be impractical or impossible. Moreover, this example shows that the methodology does not necessarily require state-of-the-art computing hardware, as all of the computations were performed in one night on a personal computer.

6 Conclusions

ADO is an example of a large class of problems that can be framed as Bayesian decision problems with expected information as expected utility. For example, current work in neurophysics aims to continuously optimize a stimulus ensemble in order to maximize mutual information between inputs and outputs (Toyoizumi, Pfister, Aihara, & Gerstner, 2005; Brunel & Nadal, 1998; Machens, 2002; Machens, Gollisch, Kolesnikova, & Herz, 2005). It is also related to the optimization of dynamic sensor networks (Hoffman, Waslander, & Tomlin, 2006) and online learning in neural networks (Oppen, 1999). In machine learning and reinforcement learning literatures, DO is known as active learning or policy decision. Essentially the same math problem is to be solved. In constructing phase portraits of dynamic systems, designs are sought to minimize the mutual information between observations (Fraser & Swinney, 1986).

The Bayesian ADO framework developed here is myopic in the sense that the optimization at each stage is done as though the current stage will be the final stage. That is, it does not take into account the potential for future stages at which a new optimal design will be sought based on the outcome at the current stage. In reality, later designs will depend on previous outcomes. Finding the globally optimal sequence of designs requires backward induction involving an exponentially increasing number of scenarios. This challenging problem is considered by Müller, Berry, Grieve, Smith, and Krams (2007), who also offer an algorithm for approximating a solution using constrained backward induction. We believe that future work should approach the ADO problem from this framework.

In the special case where the goal of experimentation is to discriminate between just two models, a natural choice for the utility of a design is the expected Bayes' factor between the two models. This was the approach employed by Heavens et al. (2007), for example. The expected Bayes' factor works well as a utility function because, as with mutual information, it is nonnegative and parameterization invariant, and it does not require computation of the MLE. However, the Bayes' factor is not appropriate for comparing more than two models. Mutual information provides a natural generalization of the expected Bayes' factor for comparing more than two models.

In sum, the growing importance of computational modeling in many disciplines has led to a need for sophisticated methods to discriminate these models. Adaptive design optimization is a principled and maximally efficient means of doing so—one that achieves this goal by increasing the

informativeness of an experiment. When combined with a utility function that is based on mutual information, the methodology increases in flexibility, being applicable to more than two models simultaneously, and provides useful insight into the model discrimination process.

Acknowledgments

This research is supported by National Institute of Health Grant R01-MH57472 to J.I.M. and M.A.P., as well as the Academy of Finland grant 121855 to J.V.K. Parts of this work were presented at the 31st annual conference of the Cognitive Science Society (2009) in Amsterdam, Netherlands, and are published in the *Proceedings*. We thank Hendrik Kück and Nando de Freitas for valuable feedback and technical help provided for project, and Michael Rosner for implementation of the design optimization algorithm in C++. Correspondence concerning this article should be addressed to Daniel Cavagnaro.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Proceedings of the Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.
- Allen, T., Yu, L., & Schmitz, J. (2003). An experimental design criterion for minimizing meta-model prediction errors applied to die casting process design. *Applied Statistics*, *52*, 103–117.
- Amzal, B., Bois, F. Y., Parent, E., & Robert, C. P. (2006). Bayesian-optimal design via interacting particle systems. *Journal of the American Statistical Association*, *101*(474), 773–785.
- Atkinson, A., & Donev, A. (1992). *Optimum experimental designs*. New York: Oxford University Press.
- Atkinson, A., & Federov, V. (1975a). The design of experiments for discriminating between two rival models. *Biometrika*, *62*(1), 57–70.
- Atkinson, A., & Federov, V. (1975b). Optimal design: Experiments for discriminating between several models. *Biometrika*, *62*(2), 289–303.
- Balasubramanian, V., Larjo, K., & Seth, R. (2008). Experimental design and model selection: The example of exoplanet detection. In P. Grünwald, P. Myllymäki, I. Tabus, M. Weinberger, & B. Yu (Eds.), *Festschrift in Honor of Jorma Rissanen*. Tampere: Tampere University Press.
- Bardsley, W., Wood, R., & Melikhova, E. (1996). Optimal design: A computer program to study the best possible spacing of design points for model discrimination. *Computers and Chemistry*, *20*, 145–157.
- Bernardo, J. M. (1979). Expected information as expected utility. *Annals of Statistics*, *7*(3), 686–690.
- Box, G., & Hill, W. (1967). Discrimination among mechanistic models. *Technometrics*, *9*, 57–71.

- Brunel, N., & Nadal, J.-P. (1998). Mutual information, Fisher information, and population coding. *Neural Computation*, 10, 1731–1757.
- Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, 10(3), 273–304.
- Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.
- Ding, M., Rosner, G., & Müller, P. (2008). Bayesian optimal design for phase II screening trials. *Biometrics*, 64, 886–894.
- Doucet, A., de Freitas, N., & Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. Berlin: Springer.
- El-Gamal, M., & Palfrey, T. (1996). Economical experiments: Bayesian efficient experimental design. *International Journal of Game Theory*, 25, 495–517.
- Fraser, M. A., & Swinney, H. L. (1986). Independent coordinates for strange attractors from mutual information. *Physical Review A*, 33(2), 1134–1140.
- Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis*. London: Chapman & Hall.
- Grünwald, P. (2005). A tutorial introduction to the minimum description length principle. In P. Grünwald, I. J. Myung, & M. A. Pitt (Eds.), *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT Press.
- Haines, L., Perevozskaya, I., & Rosenberer, W. (2003). Bayesian optimal designs for phase I clinical trials. *Biometrics*, 59, 591–600.
- Han, C., & Carlin, B. P. (2001). MCMC methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association*, 96, 1122–1132.
- Hausler, D., & Opper, M. (1997). Mutual information, metric entropy and cumulative relative entropy risk. *Annals of Statistics*, 25, 2451–2492.
- Heavens, A., Kitching, T., & Verde, L. (2007). On model selection forecasting, dark energy and modified gravity. *Monthly Notices of the Royal Astronomical Society*, 380(3), 1029–1035.
- Hoffman, G., Waslander, S., & Tomlin, C. (2006). Mutual information methods with particle filters for mobile sensor network control. *IEEE Conference on Decision and Control*. Piscataway, NJ: IEEE Press.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
- Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society. Series B (Methodological)*, 21(2), 272–319.
- Kirkpatrick, S., Gelatt, C., & Vecchi, M. (1983). Optimization by simulated annealing. *Science*, 220, 671–680.
- Kück, H., de Freitas, N., & Doucet, A. (2006). SMC samplers for Bayesian optimal nonlinear design. *Nonlinear Statistical Signal Processing Workshop (NSSPW)*. Available online at <http://people.cs.ubc.ca/~nando/papers/smdesign.pdf>.
- Kujala, J. V., & Lukka, T. J. (2006). Bayesian adaptive estimation: The next dimension. *Journal of Mathematical Psychology*, 50(4), 369–389.
- Kujala, J. V., Richardson, U., & Lyytinen, H. (in press). A Bayesian-optimal principle for a child-friendly adaptation in learning games. *Journal of Mathematical Psychology*.
- Lesmes, L., Jeon, S.-T., Lu, Z.-L., & Doshier, B. (2006). Bayesian adaptive estimation of threshold versus contrast external noise functions: The quick *tvc* method. *Vision Research*, 46, 3160–3176.

- Lewi, J., Butera, R., & Paninski, L. (2009). Sequential optimal design of neurophysiology experiments. *Neural Computation*, *21*, 619–687.
- Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, *27*(4), 986–1005.
- Loredo, T. J. (2004). Bayesian adaptive exploration. In G. J. Erickson & Y. Zhai (Eds.), *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 23rd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering* (vol. 707, pp. 330–346). Available online at arXiv:astro-ph/409386v1.
- Machens, C. (2002). Adaptive sampling by information maximization. *Physical Review Letters*, *88*(22), 2281040.
- Machens, C., Gollisch, T., Kolesnikova, O., & Herz, A. (2005). Testing the efficiency of sensory coding with optimal stimulus ensembles. *Neuron*, *47*, 447–456.
- MacKay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Computation*, *4*(4), 590–604.
- Müller, P., Berry, D., Grieve, A., Smith, M., & Krams, M. (2007). Simulation-based sequential Bayesian design. *Journal of Statistical Planning and Inference*, *137*, 3140–3150.
- Müller, P., Sanso, B., & De Iorio, M. (2004). Optimal Bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association*, *99*(467), 788–798.
- Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*(4), 190–204.
- Myung, J. I., & Pitt, M. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*, 499–518.
- Nelson, J. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4), 979–999.
- Nosofsky, R., & Zaki, S. (2002). Exemplar and prototype models revisited: Response strategies, selective attention and stimulus generalization. *Journal of Experimental Psychology*, *28*, 924–940.
- Opfer, J., & Siegler, R. (2007). Representational change and children's numerical estimation. *Cognitive Psychology*, *55*, 165–195.
- Opper, M. (1999). A Bayesian approach to online learning. In D. Saad (Ed.), *Online learning in neural networks* (pp. 363–377). Cambridge: Cambridge University Press.
- Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, *15*, 1191–1253.
- Paninski, L. (2005). Asymptotic theory of information-theoretic experimental design. *Neural Computation*, *17*, 1480–1507.
- Rubin, D., Hinton, S., & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology*, *25*(5), 1161–1176.
- Rubin, D., & Wenzel, A. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*(4), 734–760.
- Toyoizumi, T., Pfister, J.-P., Aihara, K., & Gerstner, W. (2005). Generalized Bienenstock-Cooper-Munro rule for spiking neurons that maximizes information transmission. *Proceedings of the National Academy of Sciences*, *102*(14), 5239–5244.

- Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin and Review*, *15*, 732–749.
- Wixted, J., & Ebbesen, E. (1991). On the form of forgetting. *Psychological Science*, *2*(6), 409–415.

Received February 9, 2009; accepted July 15, 2009.