A Hierarchical Adaptive Approach to Optimal Experimental Design

Woojae Kim[1], **Mark A. Pitt**[1], **Zhong-Lin Lu**[1], **Mark Steyvers**[2], **and Jay I. Myung**[1]

[1]Department of Psychology, Ohio State University, Columbus, OH 43210
[2]Department of Cognitive Sciences, University of California, Irvine, CA 92697

March 31, 2014

Abstract

Experimentation is at the core of research in the behavioral and neural sciences, yet observations can be expensive and time-consuming to acquire (e.g., MRI scans, responses from infant participants). A major interest of researchers is designing experiments that lead to maximal accumulation of information about the phenomenon under study with the fewest possible number of observations. In addressing this challenge, statisticians have developed adaptive design optimization methods. This paper introduces a hierarchical Bayes extension of adaptive design optimization that provides a judicious way to exploit two complementary schemes of inference (with past and future data) to achieve even greater accuracy and efficiency in information gain. We demonstrate the method in a simulation experiment in the field of visual perception.

Keywords: optimal experimental design, hierarchical Bayes, mutual information, visual spatial processing

1 Introduction

Accurate measurement is essential in the behavioral and neural sciences to ensure proper model inference. Efficient measurement in experimentation can also be critical when observations are costly (e.g., MRI scan fees) or time-consuming, such as requiring hundreds of observations from an individual to measure sensory (e.g., eyes, ears) abilities or weeks of training (e.g., mice). The field of design optimization (Atkinson & Donev, 1992; see Section 2 for a brief review) pursues methods of improving both, with *adaptive* optimization (e.g., DiMattina & Zhang, 2008, 2011) being one of the most promising approaches to date. These adaptive design optimization (ADO) methods capitalize on the sequential nature of experimentation by seeking to gain as much

information as possible from data across the testing session. Each new measurement is made using the information learned from previous measurements of a system so as to achieve maximal gain of information about the processes and behavior under study.

Hierarchical Bayesian modeling (HBM) is another approach to increasing the efficiency and accuracy of inference (e.g., Gelman, Carlin, Stern, & Rubin, 2004; Jordan, 1998; Koller & Friedman, 2009; Rouder & Lu, 2005). It seeks to identify structure in the data-generating population (e.g., the kind of groups to which an individual belongs) in order to infer properties of an individual given the measurements provided. It is motivated by the fact that data sets, even if not generated from an identical process, can contain information about each other. Hierarchical modeling provides a statistical framework for fully exploiting such mutual informativeness.

These two inference methods, ADO and HBM, seek to take full advantage of two different types of information, future and and past data, respectively. Because both can be formulated in a Bayesian statistical framework, it is natural to combine them to achieve even greater information gain than either alone can provide. Suppose, for instance, that one has already collected data sets from a group of participants in an experiment measuring risk tolerance, and data are about to be collected from another person. A combination of HBM and ADO allows the researcher to take into account the knowledge gained about the population in choosing optimal designs. The procedure should propose designs more efficiently for the new person than ADO alone, even when no data for that person have been observed.

Despite the intuitive appeal of this dual approach, to the best of our knowledge, a general, fully Bayesian framework integrating the two methods has not been published. In this letter, we provide one. In addition, we show how each method and their combination contribute to gaining the maximum possible information from limited data, in terms of Shannon entropy, in a simulation experiment in the field of visual psychophysics.

2 Paradigm of Adaptive Design Optimization (ADO)

The method for collecting data actively for best possible inference, rather than using a data set observed in an arbitrarily fixed design, is known as *optimal experimental design* in statistics, which goes back to the pioneering work in the 1950s and 1960s (Lindley, 1956; Chernoff, 1959; Kiefer, 1959; Box & Hill, 1967). Essentially the same technique has been studied and applied in machine learning as well, known as *query-based learning* (Seung, Opper, & Sompolinsky, 1992) and *active learning* (Cohn, Ghahramani, & Jordan, 1996). Since in most cases data collection occurs sequentially and optimal designs are best chosen upon immediate feedback from each data point, the algorithm is by nature adaptive, hence the term *adaptive design optimization* (ADO) that we use here.

The recent surge of interest in this field can be attributed largely to the advent of fast computing, which has made it possible to solve more complex and a wider range of optimization problems, and in some cases do so in real-time experiments. ADO is gaining traction in neuroscience (Paninski, 2003, 2005; Lewi, Butera, & Paninski, 2009; DiMattina & Zhang, 2008, 2011), and a growing number of labs are applying it in various areas of psychology and cognitive science, including retention memory

(Cavagnaro, Pitt, & Myung, 2009; Cavagnaro, Myung, Pitt, & Kujala, 2010), decision making (Cavagnaro, Gonzalez, Myung, & Pitt, 2013; Cavagnaro, Pitt, Gonzalez, & Myung, 2013), psychophysics (Kujala & Lukka, 2006; Lesmes, Jeon, Lu, & Dosher, 2006), and the development of numerical representation (Tang, Young, Myung, Pitt, & Opfer, 2010). In what follows, we provide a brief overview of the ADO framework.

ADO is formulated as a Bayesian sequential optimization algorithm that is executed over the course of an experiment.[1] Specifically, on each trial of the experiment, on the basis of the present state of knowledge (prior) about the phenomenon under study, which is represented by a statistical model of data, the optimal design with the highest expected value of a utility function (defined below) is identified. The experiment is then carried out with the optimal design, and measured outcomes are observed and recorded. The observations are subsequently used to update the prior to the posterior using Bayes' theorem. The posterior in turn is used to identify the optimal design for the next trial of the experiment. As depicted in the shaded region of Figure 1, these alternating steps of design optimization, measurement, and updating of the individual-level data model are repeated in the experiment until a suitable stopping criterion is met.

In formal statistical language, the first step of ADO, design optimization, entails finding the experimental design (e.g., stimulus) that maximizes a utility function of the following form (Chaloner & Verdinelli, 1995; Nelson, McKenzie, Cottrell, & Sejnowski, 2011; Myung, Cavagnaro, & Pitt, 2013):

$$U(d_t) = \iint \left[\log \frac{p(\theta|y^{(1:t)}, d_t)}{p(\theta|y^{(1:t-1)})} \right] p(y^{(t)}|\theta, d_t)\, p(\theta|y^{(1:t-1)})\, \mathrm{d}y^{(t)}\, \mathrm{d}\theta, \tag{1}$$

where θ is the parameter of a data model (or measurement model) that predicts observed data given the parameter, and $y^{(1:t)}$ is the collection of past measurements made from the first to $(t-1)$-th trials, denoted by $y^{(1:t-1)}$, *plus* an outcome, $y^{(t)}$, to be observed in the current, t-th trial conducted with a candidate design, d_t. In this equation, note that the function $p(y^{(t)}|\theta, d_t)$ specifies the model's probabilistic prediction of $y^{(t)}$ given the parameter θ and the design d_t, and $p(\theta|y^{(1:t-1)})$ is the posterior distribution of the parameter given past observations, which has become the prior for the current trial. Finally, $\log \frac{p(\theta|y^{(1:t)}, d_t)}{p(\theta|y^{(1:t-1)})}$, referred to as the *sample* utility function, measures the utility of design d_t, assuming an outcome, $y^{(t)}$, and a parameter value (often a vector), θ.

$U(d_t)$ in Eq. (1) is referred to as the *expected* utility function, and is defined as the expectation of the sample utility function with respect to the data distribution $p(y^{(t)}|\theta, d_t)$ and the parameter prior $p(\theta|y^{(1:t-1)})$. Under the above particular choice of the sample utility function, the expected utility $U(d_t)$ admits an information theoretic interpretation. Specifically, the quantity becomes the mutual information between the

[1] In the present study, we consider a particular form of ADO that assumes the use of a Bayesian model and the information-theoretic utility (discussed further in the text). While this choice has straightforward justification from the Bayesian perspective as the quality of inference is evaluated on the level of a posterior distribution, there are other forms of ADO that assumes a non-Bayesian model or achieves other types of optimality (e.g., minimum quadratic loss of a point estimate). Chaloner and Verdinelli (1995) provide a good review of various approaches to design optimization.

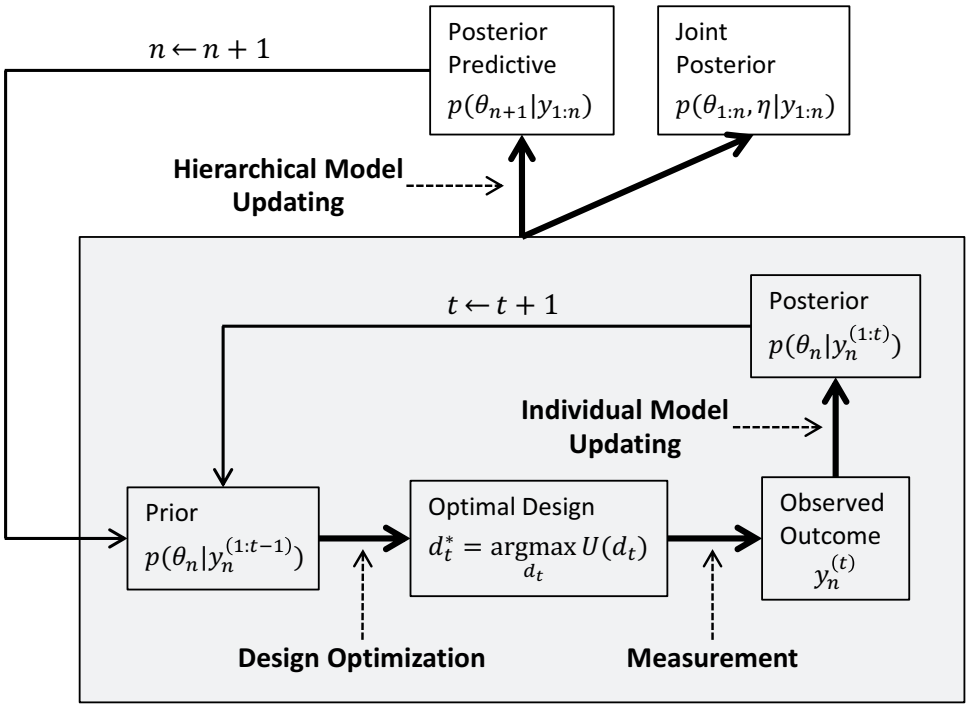

Figure 1: Schematic illustration of the steps involved in adaptive design optimization (ADO; shaded region only) and hierarchical ADO (HADO; whole diagram). See text for further details.

parameter variable Θ and the outcome variable $Y^{(t)}$ conditional upon design d_t, i.e., $U(d_t) = I(\Theta; Y^{(t)}|d_t)$ (Cover & Thomas, 1991), which also represents the so-called Bayesian *D*-optimality (Chaloner & Verdinelli, 1995). Accordingly, the optimal design that maximizes $U(d_t)$, or $d_t^* = \arg\max U(d_t)$, is the one that yields the largest information gain about the model parameter(s) upon the observation of a measurement outcome.[2]

The second, measurement step of ADO involves administering the optimal design d_t^* and observing the measurement outcome $y^{(t)}$, as illustrated in Figure 1. The final, third step of the ADO application is updating the prior $p(\theta|y^{(1:t-1)})$ to the posterior $p(\theta|y^{(1:t)})$ by Bayes' theorem on the basis of the newly observed outcome $y^{(t)}$.

[2]In defining the mutual information here, we assume that the goal of ADO is to maximize the information about all parameter elements of a model jointly, rather than some of them. In another situation, for example, the model may be a mixture model whose parameter θ contains an indicator to a sub-model, and the goal of ADO may be to maximize the information about the indicator variable (i.e., the problem of model discrimination; e.g., Cavagnaro et al., 2010). In this case, the required change is to redefine the sample utility function in Eq. (1) by integrating out the parameters of no interest (e.g., sub-model parameters) from each of the distributions inside the logarithm.

In implementing ADO, a major computational burden is finding the optimal design d^*, which involves evaluating the multiple integrals in both the sample and the expected utility functions in Eq. (1) (integral is implicit in the sample utility). The integrals generally have no closed-form solutions and need to be calculated many times with candidate designs substituted during optimization. Further, online data collection requires that the integration and optimization be solved numerically on computer in real time. Advances in parallel computing (e.g., general purpose GPU computing) have made it possible to solve some problems using grid-based algorithms. In situations in which grid-based methods are not suitable, several promising Markov chain Monte Carlo (MCMC) methods have been developed to perform the required computation (Müller, Sanso, & De Iorio, 2004; Amzal, Bois, Parent, & Robert, 2006; Cavagnaro et al., 2010; Myung et al., 2013).

3 Hierarchical Adaptive Design Optimization (HADO)

As currently used, ADO is tuned to optimizing a measurement process at the individual participant level, without taking advantage of information available from data collected from previous testing sessions. Hierarchical Bayesian modeling (HBM; for theory, Good, 1965; de Finetti, 1974; Bernado & Smith, 1994; for application examples, Jordan, 1998; Rouder, Speckman, Sun, & Jiang, 2005; Rouder & Lu, 2005; Lee, 2006) not only provides a flexible framework for incorporating this kind of prior information but is also well suited for being integrated within the existing Bayesian ADO paradigm to achieve even greater efficiency of measurement.

The basic idea behind HBM is to improve the precision of inference (e.g., power of a test) by taking advantage of statistical dependencies present in data. For example, suppose that there are previous measurements taken from different individuals who are considered a random sample from a certain population. It is highly likely that measurements taken from a new individual drawn from the same population will share similarities with others. In this situation, adaptive inference will enjoy greater benefit when taking the specific data structure into account rather than starting with no such information. That is, data sets, as a collection, contain information about one another, lending themselves to more precise inference. Since individual data sets require themselves to be modeled (i.e., a measurement model), the statistical relationship among them needs to be modeled on a separate level, hence the model being hierarchical (for more examples of the upper-level structure in a hierarchical model, see Koller & Friedman, 2009; Gelman et al., 2004).

From the perspective of Bayesian inference, HBM is a way, given a certain data model, to form an *informative prior* for model parameters by learning from data. An informative prior, however, may be obtained not only by learning newly from empirical observations but also by incorporating established knowledge about the data-generating structure. Since the use of prior information is one of the major benefits of Bayesian optimal experimental design (Chaloner & Verdinelli, 1995), it is no surprise to find examples of using informative priors in the literature of design optimization. These applications focus on imposing theoretically sensible constraints on the prior in a conservative manner, in which the constraints are represented by a restricted support of

the prior (Tulsyan, Forbes, & Huang, 2012), regularization (Woolley, Gill, & Theunissen, 2006), structured sparsity (Park & Pillow, 2012), and modeled covariance structure (Ramirez et al., 2011). Some of these studies employ hierarchical models because modeling a prior distribution with hyper-parameters naturally entails hierarchical structure. The present study, by contrast, focuses on *learning* prior knowledge from data, which is useful when the phenomenon being modeled has yet to permit effective, theoretical (or algorithmic) constraints to be used as a prior or when, if certain constraints have already been incorporated, inference can further benefit from information elicited from a specific empirical condition.

3.1 Formulation

To integrate HBM into ADO, let us first specify a common form of a hierarchical Bayes model. Suppose that an individual-level measurement model has been given as a probability density or mass function, $p(y_i|\theta_i)$, given the parameter (vector), θ_i, for individual i, and the relationship among individuals is described by an upper-level model, $p(\theta_{1:n}|\eta)$ (e.g., a regression model with η as coefficients), where $\theta_{1:n} = (\theta_1, \cdots, \theta_n)$ is the collection of model parameters for all n individuals. Also commonly assumed is conditional independence between individuals such that $p(y_i|\theta_{1:n}, y_{-.i}) = p(y_i|\theta_i)$ where $y_{-.i}$ denotes the collection of data from all individuals except individual i (i.e., $y_{1:n} = (y_1, \cdots, y_n)$ minus y_i). Then, the joint posterior distribution of the hierarchical model given all observed data is expressed as

$$\begin{aligned} p(\theta_{1:n}, \eta|y_{1:n}) &= \frac{1}{p(y_{1:n})} p(y_{1:n}|\theta_{1:n}) p(\theta_{1:n}|\eta) p(\eta) \\ &= \frac{1}{p(y_{1:n})} \left[\prod_{i=1}^{n} p(y_i|\theta_i)\right] p(\theta_{1:n}|\eta) p(\eta), \end{aligned} \tag{2}$$

where $p(\eta)$ is the prior distribution for the upper-level model's parameter, η, and the marginal distribution $p(y_{1:n})$ is obtained by integrating the subsequent expression over $\theta_{1:n}$ and η.

The model also needs to be expressed in terms of an entity about which the measurement seeks to gain maximal information. In most measurement situations, it is sensible to assume that the goal is to estimate the traits of a newly measured individual most accurately. Suppose that a measurement session is currently underway on the n-th individual, and data from previous measurement sessions, $y_{1:n-1}$, are available. Then, the posterior distribution of θ_n for this particular individual given *all* available data is derived from (2) as

$$p(\theta_n|y_{1:n}) = \frac{1}{p(y_{1:n})} \iint \left[\prod_{i=1}^{n} p(y_i|\theta_i)\right] p(\theta_{1:n}|\eta) p(\eta) \, \mathrm{d}\eta \, \mathrm{d}\theta_{1:n-1} \tag{3}$$

where the marginal distribution $p(y_{1:n})$ is obtained by integrating the integrand further over θ_n. From a computational standpoint, it is advantageous to turn the above posterior distribution into a sequentially predictive form. Under the assumption of conditional

independence, Eq. (3) can be rewritten as

$$p(\theta_n|y_{1:n}) = \frac{p(y_n|\theta_n)p(\theta_n|y_{1:n-1})}{\int p(y_n|\theta_n)p(\theta_n|y_{1:n-1})\,\mathrm{d}\theta_n}, \tag{4}$$

where

$$p(\theta_n|y_{1:n-1}) = \frac{1}{p(y_{1:n-1})} \iint \left[\prod_{i=1}^{n-1} p(y_i|\theta_i)\right] p(\theta_{1:n}|\eta)p(\eta)\,\mathrm{d}\eta\,\mathrm{d}\theta_{1:n-1} \tag{5}$$

is the posterior predictive distribution of θ_n given the data from previous measurement sessions, $y_{1:n-1}$ (assuming that y_n is yet to be observed).[3] An interpretation of this form is that, as far as θ_n is concerned, the predictive distribution in Eq. (5) fully preserves information in the previous data $y_{1:n-1}$ and, in turn, serves as an informative prior for the current, n-th individual, which is updated upon actually observing y_n.

Having established the basic building blocks of hierarchical adaptive design optimization (HADO), we now describe how measurement within the HADO framework is carried out. Suppose that a measurement has been taken in trial $t-1$ for the n-th individual, and the session is in need of an optimal design to make the next observation, $y_n^{(t)}$, in trial t. Then, the optimal design, d_t^*, is the one that maximizes the following mutual-information utility:

$$U(d_t) = \iint \left[\log \frac{p(\theta_n|y_{1:n-1}, y_n^{(1:t)}, d_t)}{p(\theta_n|y_{1:n-1}, y_n^{(1:t-1)})}\right] p(y_n^{(t)}|\theta_n, d_t)p(\theta_n|y_{1:n-1}, y_n^{(1:t-1)})\,\mathrm{d}y_n^{(t)}\,\mathrm{d}\theta_n, \tag{6}$$

where $y_{1:n-1}$ denotes the data from previous $n-1$ measurement sessions, and $y_n^{(1:t)}$ contains the n-th individual's measurements from past $t-1$ trials (i.e., $y_n^{(1:t-1)}$) plus an observation, $y_n^{(t)}$, that is to be made in trial t using a candidate design, d_t. Note that this utility function of HADO, similar as it may seem in its form to that of ADO in Eq. (1), takes all previously observed data into account through the hierarchical model, not just that from the current measurement session.

For HADO to be adaptive, Bayesian updating for posterior distributions inside the above utility function is performed recursively on two different levels (Figure 1). First, on the individual level (shaded region), updating is repeated over each measurement trial (i.e., to find the optimal design d_{t+1}^* after observing $y_n^{(t)}$) using Eq. (4) (i.e., Bayes' theorem). Note that what is modified in Eq. (4) is only the individual data model (i.e., $p(y_n|\theta_n)$) with $y_n = y_n^{(1:t-1)}$ augmented with a new measurement, $y_n^{(t)}$. Next, when the session ends and a new one is to begin for the next participant (outside the shaded region), the hierarchical model is updated, again using Bayes' theorem, on the basis of all n sessions' data, $y_{1:n}$, and expressed in a posterior predictive form for θ_{n+1} (Eq. (5) with $n+1$). The session counter n now shifts to $n+1$, the trial counter t is reset

[3] Although the term *predictive distribution* is usually associated with a Bayesian model's prediction of a future observation, it may also be used to mean the prediction of a future, latent variable in a hierarchical model, such as θ_n in the present context.

to 1, and the posterior predictive distribution becomes the prior for the new session to start with (i.e., $p(\theta_{n+1}|y_{n+1}^{(1:0)}) = p(\theta_{n+1}|y_{1:n})$). This two-stage adaptation is a defining characteristic of HADO, hence the term "hierarchical adaptive."

Although not implemented as an application example in the present study, there are additional forms of HADO that are worth mentioning. The idea of *combining* the techniques of hierarchical Bayes and optimal experimental design is more general than described above. For example, suppose that one wants to understand the population-level parameters but it is difficult to collect a sufficient amount of data from each individual (e.g., in developing a human-computer interaction model that functions robustly in a general setting). This problem is best addressed by hierarchical modeling but the application of hierarchical modeling alone is merely ad hoc in the sense that the acquisition of data is not planned optimally. In this case, introduction of ADO will make it possible to choose optimal designs adaptively, not only within but also across individual measurement sessions, so that the maximum possible information is gained about the population-level parameters. That is, it is possible for the algorithm to probe different aspects of individuals across sessions that best contribute to the goal of learning the common functioning, not necessarily learning that particular individual. In achieving this, the optimal design maximizes the following information-theoretic utility:

$$U'(d_t) = \iint \left[\log \frac{p(\eta|y_{1:n-1}, y_n^{(1:t)}, d_t)}{p(\eta|y_{1:n-1}, y_n^{(1:t-1)})} \right] p(y_n^{(t)}|\theta_n, d_t) p(\theta_n, \eta|y_{1:n-1}, y_n^{(1:t-1)}) \, \mathrm{d}y_n^{(t)} \, \mathrm{d}\theta_n \, \mathrm{d}\eta, \tag{7}$$

which measures the expected information gain from a deign d_t of the next trial about the population-level parameter(s) η. As with the preceding formulation, Bayesian updating needs to be performed on both individual and higher levels, but in this case, updating $p(\theta_n, \eta|\cdot)$.

One may also want to optimize an experiment to infer both the higher-level structure and the individual-level attributes. The formal framework employed in the present study is general enough to address this problem (i.e., meeting seemingly multiple goals of inference). The utility function to maximize for an optimal design in the next trial is a slight modification of Eq. (7):

$$U''(d_t) = \iint \left[\log \frac{p(\theta_n, \eta|y_{1:n-1}, y_n^{(1:t)}, d_t)}{p(\theta_n, \eta|y_{1:n-1}, y_n^{(1:t-1)})} \right] p(y_n^{(t)}|\theta_n, d_t) p(\theta_n, \eta|y_{1:n-1}, y_n^{(1:t-1)}) \, \mathrm{d}y_n^{(t)} \, \mathrm{d}\theta_n \, \mathrm{d}\eta, \tag{8}$$

which equals $I(\Theta_n, H; Y_n^{(t)}|d_t)$ by the notation of mutual information (H denotes the random variable corresponding to η). A simple yet notable application example of this formulation is a situation in which the goal of an experiment is to select among multiple, alternative models, assuming that one of them is the underlying data-generating process for all individuals, and at the same time to estimate distinct parameter values for each individual. The utility that captures this goal is a special case of Eq. (8) in which the

higher-level parameter η turns into a model index m and the corresponding integration is replaced by summation over the indexes. In fact, a similar approach to choosing optimal designs for model selection and parameter learning has been proposed previously (Sugiyama & Rubens, 2008) but the current framework is more general in that any type of hierarchical structure can be inferred and the optimality of a design with respect to the goal is understood from a unified perspective.

3.2 Implementation Considerations

In typical applications of hierarchical Bayes, posterior inference is conducted mainly to incorporate the data that have already been collected, and all the parameters of interest are updated jointly in a simulation-based method (e.g., via MCMC). This approach, however, is not well suited to HADO. Many applications of adaptive measurement require the search for an optimal design between trials to terminate in less than a second. To circumvent this computational burden, we formulated HADO, as described in the preceding section, in a natural way that suits its domain of application (experimentation), allowing the required hierarchical Bayes inference to be performed in two stages. Below we describe specific considerations for implementing these steps.

Once a numerical form of the predictive distribution (Eq. (5)) is available, updating the posterior distribution (Eq. (4)) within each HADO measurement session concerns only the current individual's parameter and data just as in the conventional ADO. Accordingly, the recursive updating on the individual level will be no more demanding than the corresponding operation in conventional ADO since they involve essentially the same computation. Beyond the individual level, an additional step is required to revise the posterior predictive distribution of θ_n given all previous data upon the termination of each measurement session, which is shown outside the shaded area in Figure 1. The result becomes a prior for the next session, serving as an informative prior for the individual to be newly measured.[4]

Critical, then, to the implementation of HADO is a method for obtaining a numerical form of the predictive distribution of θ_n before a new measurement session begins for individual n. Fortunately, in most cases this distribution conforms to smoothness and structured sparsity (a prior distribution with a highly irregular shape is not sensible), being amenable to approximation. Furthermore, in modeling areas dealing with high-dimensional feature space, certain theoretical constraints that take advantage of such regularity are often already studied and modeled into a prior (e.g., Park & Pillow, 2012; Ramirez et al., 2011), which can also be utilized to represent the predictive distribution. Otherwise, various density estimation techniques with built-in regularization mechanisms (e.g., kernel density estimator) may be used to approximate the distribution (Scott, 1992). For a lower-dimensional case, a grid representation may be useful. In fact, grid-based methods can handle multidimensional problems with high efficiency

[4]The same, two-level updating can also apply to the case where the inference involves the population-level parameters with optimal designs satisfying the utility function shown in Eq. (7) or Eq. (8), as long as the predictive distribution $p(\theta_n, \eta|y_{1:n-1})$ is computable.

when combined with a smart gridding scheme that exploits regularity (e.g., sparse grids; Pflüger, Peherstorfer, & Bungartz, 2010).

Another consideration is that the predictive distribution of θ_n must be obtained by integrating out all other parameters numerically, particularly other individuals' parameters θ_i's. If θ_i's (or groups of θ_i's) are by design conditionally independent in the upper-level model ($p(\theta_{1:n}|\eta)p(\eta)$ in Eq. (5)), it is possible to phrase the integral as repeated integrals that are easier to compute. Also, note that the shape of the integrands is highly concentrated with a large number of observations per individual (i.e., large t) and the posterior predictive of θ_n tends to be localized as well with accumulation of data over many sessions (i.e., large n). Various techniques for multidimensional numerical integration are available that can capitalize on these properties. *Monte Carlo integration* based on a general sampling algorithm such as MCMC is a popular choice for high-dimensional integration problems (Robert & Casella, 2004). However, unless the integrand is highly irregular, *multivariate quadrature* is a viable option because, if applicable, it generally outperforms Monte Carlo integration in regard to efficiency and accuracy and, with recent advances, scales well to high-dimensional integration depending on the regularity (Griebel & Holtz, 2010; Holtz, 2011; Heiss & Winschel, 2008).

Note that, although an estimate of θ_n (e.g., posterior mean) is obtained at the end of the measurement session, the main purpose of posterior updating for θ_n within the session is to generate optimal designs. Thus, the resulting estimate of θ_n may not necessarily be taken as a final estimate, especially when the employed posterior predictive approximation is not highly precise. If needed, additional Bayesian inference based on the joint posterior distribution in Eq. (2) may be conducted after each test session with added data (top right box in Figure 1). This step will be particularly suitable when the upper-level structure (i.e., η) needs to be analyzed, or precise estimates of all previously tested individuals' parameters are required for a certain type of analysis (e.g., to build a classifier that categorizes individuals based on modeled traits in the parameters).

It is also notable, from the computational perspective, that the procedure inside the shaded area in Figure 1 requires online computation during the measurement session, whereas the posterior predictive calculation outside the area (i.e., computing its grid representation) is performed offline between sessions. In case multiple sessions need to be conducted continually without an interval sufficient for offline computation, the same predictive distribution may be used as a prior for these sessions; for example, offline computation is performed overnight to prepare for the next day's measurement sessions. This approach, though not ideal, will provide the same benefit of hierarchical modeling as data accumulate.

Lastly, in applying HADO we may want to consider two potential use cases. One is a situation in which there is no background database available a priori and therefore the hierarchical model in HADO might learn some idiosyncrasies from the first few data sets (i.e., small n). The other, more likely use case is where there is a fairly large number of pretested individuals that can be used to build and estimate the hierarchical model. While HADO can be applied to both cases, it would be no surprise that its benefit should be greater in the latter situation. Even so, the behavior of HADO with small n is worth noting here. First, if there exists a prior that has been used conventionally for the modeling problem, the prior of the upper-level structure in HADO should be

set in such a way that when $n = 0$ or 1 it becomes comparable to that conventional prior, if the hyper-parameters are marginalized out. Second, unless the model is overly complex (e.g., in this context, the higher-level structure is highly flexible with too many parameters), Bayesian inference is generally robust against overfittting to idiosyncrasies in a small data sample because the posterior of model parameters given the data would not deviate much from the prior. Otherwise, if overfitting is suspected, HADO inference should start being applied and interpreted once an adequate sample is accumulated.

In sum, ADO for gaining maximal information from sequential measurements has been extended to incorporate the hierarchical Bayes model to improve information gain further. Conceptually, HADO improves the estimation of an individual data model by taking advantage of the mutual informativeness among individuals tested in the past. While there may be alternative approaches to forming an informative prior from past data for a Bayesian analysis, hierarchical Bayes is the method that enables both the generations of individual-level data and the relationship among them to be modeled and inferred jointly in a theoretically justified manner. The formulation and implementation of HADO provided above exploit the benefits of both hierarchical Bayes and ADO by integrating them within a fully Bayesian framework.

4 Application Example

The benefits of HADO were demonstrated in a simulated experiment in the domain of visual perception. Visual spatial processing is most accurately measured using a contrast sensitivity test, in which sinewave gratings are presented to participants at a range of spatial frequencies (i.e., widths) and luminance contrasts (i.e., relative intensities). The objective of the test is to measure participants' contrast threshold (detectability) across a wide range of frequencies, which together create a participant's contrast sensitivity function (CSF). The comprehensiveness of the test makes it useful for detecting visual pathologies. However, because the standard methodology can require many hundreds of stimulus presentations for accurate threshold measurements, it is a prime candidate for the application of ADO and HADO.

Using the Bayesian framework described in Section 2, Lesmes, Lu, Baek, and Albright (2010) introduced an adaptive version of the contrast sensitivity test called qCSF. Contrast sensitivity, $S(f)$, against grating frequency, f, was described using the truncated log-parabola with four parameters (Watson & Ahumada, 2005):

$$S(f) = \begin{cases} \gamma^{\max} - \delta & \text{if } f < f^{\max} - \dfrac{\beta}{2}\sqrt{\dfrac{\delta}{\log_{10} 2}}; \\ \gamma^{\max} - (\log_{10} 2)\left(\dfrac{f - f^{\max}}{\beta/2}\right)^2 & \text{otherwise,} \end{cases} \tag{9}$$

where $\gamma^{\max}$ is the peak sensitivity at the frequency $f^{\max}$, β denotes the bandwidth of the function (full width at half the peak sensitivity), δ is the low-frequency truncation level, and all variables and parameters are on base-10 log scales. The optimal stimulus selection through ADO, along with the parametric modeling, was shown to reduce the number of trials ($<$100) required to obtain a reasonably accurate estimate of CSF at only a minimal cost in parameter estimation compared to non-adaptive methods.

To demonstrate the benefits of HADO, the current simulation study considered four conditions in which simulated subjects were tested for their CSFs by means of four different measurement methods. We begin by describing how these conditions were designed and implemented.

4.1 Simulation Design

The two most interesting conditions were the ones in which ADO and HADO were used for stimulus selection. In the first, *ADO* condition, the qCSF method of Lesmes et al. (2010) was applied and served as the existing, state-of-the-art technique against which, in the second, *HADO* condition, its hierarchical counterpart developed in the present study was compared. If the prior information captured in the upper-level structure of the hierarchical model can improve the accuracy and efficiency of model estimation, then performance in the HADO condition should be better than that in the ADO (qCSF) condition. Also included for completeness were two other conditions to better understand information gain achieved by each of the two components of HADO: hierarchical Bayes modeling (HBM) and ADO. To demonstrate the contribution of HBM alone to information gain, in the third, *HBM* condition, prior information was conveyed through HBM but no optimal stimulus selection was performed during measurement (i.e., stimuli were not selected by ADO but sampled randomly). In the fourth, *non-adaptive* condition, neither prior data nor stimulus selection was utilized, so as to provide a baseline performance level against which improvements of the other methods could be assessed.

The hierarchical model in the HADO condition comprised two layers. On the individual level, each subject's CSF was modeled by the four-parameter, truncated log-parabola specified in Eq. (9). The model provided a probabilistic prediction through a psychometric function so that the subject's binary response to a presented stimulus (i.e., detection of a sinusoidal grating with chosen contrast and frequency) could be predicted as a Bernoulli outcome. The log-Weibull psychometric function in the model has the form:

$$\Psi(c, f) = .5 + (.5 - \lambda/2)\left[1 - \exp\left(-10^{\kappa(\log_{10} c + \log_{10} S(f))}\right)\right], \tag{10}$$

where c and f denote the contrast and the spatial frequency, respectively, of a stimulus being presented (i.e., design variables), and $S(f)$ is the contrast sensitivity (or the reciprocal of the threshold) at the frequency f (i.e., CSF) modeled by the truncated log-parabola in Eq (9). The two parameters of the psychometric function, λ (lapse rate; set to .04) and κ (psychometric slope; set to 3.5), were given particular values following the convention in previous studies (Lesmes et al., 2010; Hou et al., 2010).

On the upper level, the generation of a subject's CSF parameters was described by a two-component, four-variate Gaussian mixture distribution, along with the usual, normal-inverse-Wishart prior on each component and the beta prior on mixture weights.

Symbolically,

$$
\begin{aligned}
(\gamma_i^{\max}, f_i^{\max}, \beta_i, \delta_i) &\sim \sum_{j=1}^{2} \phi_j \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \quad i = 1, \cdots, n \\
(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) &\sim NIW(\boldsymbol{\mu}_0, \kappa_0, \boldsymbol{\Lambda}_0, \nu_0), \quad j = 1, 2 \\
\phi_1 &\sim Beta(\alpha_0, \beta_0), \;\; \phi_2 = 1 - \phi_1,
\end{aligned} \tag{11}
$$

where the parameter values of the normal-inverse-Wishart prior ($\boldsymbol{\mu}_0$ = (2, 0.40, 0.78, 0.5), $\kappa_0 = 2$, $\boldsymbol{\Lambda}_0 = \frac{1}{3}\pi^2 \mathbf{I}$, $\nu_0 = 5$) were chosen on the following grounds: When there is little accumulation of data, the predictive distribution of CSF parameters should be comparable to the prior distribution used in the previous research (i.e., the prior of the non-hierarchical CSF model in Lesmes et al., 2010). The beta prior was set to $\alpha_0 = \beta_0 = 0.5$. The choice of a two-component mixture was motivated by the nature of the data which are assumed to be collected from two groups under different ophthalmic conditions. In practice, when this type of information (i.e., membership to distinct groups) is available, the use of a mixture distribution will be a sensible approach to lowering the entropy of the entity under estimation. While a more refined structure might be plausible (e.g., CSFs covary with other observed variables), we did not further investigate the validity of alternative models since the current hypothesis (i.e., individuals are similar to each other in the sense that their CSFs are governed by a Gaussian component in a mixture model) was simple and sufficient to show the benefits of HADO.

The procedure for individual-level measurement with optimal stimuli (i.e., shaded area in Figure 1) followed the implementation of qCSF (Lesmes et al., 2010) in which all required computations for design optimization and Bayesian updating were performed on a grid in a fully deterministic fashion (i.e., no Monte Carlo integration; *see* Lesmes et al., 2010 for detail). The posterior inference of the upper-level model, or the formation of a predictive distribution given the prior data (i.e., outside the shaded region in Figure 1), also involved no sampling-based computation. This was possible because the upper-level model (i.e., Gaussian mixture) allowed for conditional independence between individuals so that the posterior predictive density (Eq. (5)) of a particular θ_n value could be evaluated as repeated integrals over individual θ_i's. To increase the precision of grid representations of prior and posterior distributions, which are constantly changing with data accumulation, the grid was defined dynamically on a four-dimensional ellipsoid in such a way that the support of each updated distribution with at least 99.9% probability is contained in it. The grid on the ellipsoid was obtained by linearly transforming a grid on a unit 4-ball that had 20,000 uniformly spaced points.

The ADO (qCSF) condition shared the same individual data model as specified in the HADO condition, but the variability among individuals was not accounted for by an upper-level model. Instead, each individual's parameters were given a diffuse, Gaussian prior comparable to the non-informative prior used previously in the field. The HBM condition took the whole hierarchical model from HADO, but the measurement for each individual was made with stimuli randomly drawn from a prespecified set. Finally, the non-adaptive method was based on the non-hierarchical model in ADO (qCSF) and used random stimuli for measurement.

To increase the realism of the simulation, we used real data collected from adults who underwent CSF measurement. There were 147 data sets, 67 of which were from individuals whose tested eye was diagnosed as amblyopic (poor spatial acuity). The remaining 80 data sets were from tests on non-diseased eyes. Thirty-six of these individuals took the qCSF test (300 trials with optimal stimuli) and 111 were administered the non-adaptive test (700 to 900 trials with random stimuli). The number of measurements obtained from each subject was more than adequate to provide highly accurate estimates of their CSFs.

To compare the four methods, we first used a leave-one-out paradigm, treating 146 subjects as being previously tested and the remaining subject as a new individual to be measured subsequently. We further assumed that, in each simulated measurement session, artificial data are generated from an underlying CSF (taken from the left-out subject's estimated CSF) with one of the four methods providing stimuli. If HADO is applied, this situation represents a particular state in the recursion of measurement sessions shown in Figure 1; that is, the session counter is changing from $n = 146$ to $n = 147$ to test a new, 147th subject. It does not matter whether the previously collected data were obtained by using HADO or not, since their estimation precision was already very high as a result of using the brute-force, large number of trials.

One may wonder how HADO would perform if it were applied when there is a small accumulation of data (i.e., when n is small). As mentioned earlier, Bayesian inference is robust against overfitting to idiosyncrasies in a small sample, especially when the model is not very complex (here, the higher-level structure is relatively simple). To demonstrate this, an additional simulation in the HADO condition was performed with small n's being assumed.

Finally, since the observations from each simulated measurement session were random variates generated from a probabilistic model, to prevent the comparison of performance measures from being masked by idiosyncrasies, ten replications of the 147 leave-one-out sessions were run independently and the results were averaged over all individual sessions (i.e., 10×147 = 1,470 measurement sessions were conducted in total).

4.2 Results

The whole simulation procedure was implemented on a machine with two, quad-core Intel 2.13GHz XEON processors and one Nvidia Tesla C2050 GPU computing processor running Matlab. Grid-based computing for utility function evaluations and Bayesian updating was parallelized through large GPUArray variables in Matlab. As a result, each inter-trial computing process, including stimulus selection, Bayesian updating and grid adaptation, took 90 milliseconds on average, and hierarchical model updating with 146 previous data sets took about 11 seconds, which was six to eight times faster than the same tasks processed by fully vectorized Matlab codes running on CPUs.

Performance of the four methods of measurement and inference described in the preceding section was assessed in three ways: information gain, accuracy of parameter estimation, and accuracy of amblyopia classification. These evaluation measures were calculated across all trials in each simulated measurement session. For information gain, the degree of uncertainty about the current, n-th subject's parameter(s) upon

observing trial t's outcome was measured by the *differential entropy* (extension of the Shannon entropy to the continuous case):

$$H_t(\Theta_n) = -\int p(\theta_n|y_{1:n-1}, y_n^{(1:t)}) \log p(\theta_n|y_{1:n-1}, y_n^{(1:t)})\, \mathrm{d}\theta_n. \tag{12}$$

Use of the differential entropy, which is not bounded in either direction on the real line, is often justified by choosing a baseline state and defining the observed information gain as the difference between two states' entropies. In the present context, it is

$$IG_t(\Theta_0, \Theta_n) = H_0(\Theta_0) - H_t(\Theta_n), \tag{13}$$

where $H_0(\Theta_0)$ denotes the entropy of a baseline belief about θ in a prior distribution so that $IG_t(\Theta_0, \Theta_n)$ may be interpreted as the information gain achieved upon trial t during the test of subject n relative to the baseline state of knowledge. In the current simulation, we took the entropy of the non-informative prior used in the conditions with no hierarchical modeling (i.e., ADO and non-adaptive) as $H_0(\Theta_0)$. Note that the information gain defined here is a cumulative measure over the trials in a session in the sense that $IG_t(\Theta_0, \Theta_n) = H_0(\Theta_0) - H_1(\Theta_n) + \sum_{s=2}^{t} [H_{s-1}(\Theta_n) - H_s(\Theta_n)]$ where the quantity being summed is information gain upon trial s relative to the state before that trial.

Shown in Figure 2 is the cumulative information gain observed in each simulation condition designed to evaluate the performance of the four different methods. Each of the four curves corresponds to information gain (y-axis) in each condition over 200 trials (x-axis) relative to the non-informative, baseline state (0 on the y-axis). The information gain measures were averaged over all 1,470 individual measurement sessions in each condition. Then, we further normalized the measures by dividing them by the average information gain at the 200th trial achieved by the crude, non-adaptive method in order to take the value of 1 as a baseline level of performance against which to compare performance of the other methods.

First, the results demonstrate that the hierarchical adaptive methodology (HADO) achieves higher information gain than the conventional adaptive method (ADO). The contribution of hierarchical modeling is manifested at the start of each session as a considerable amount of information (0.4) in the HADO condition (solid curve) than no information (zero) in the ADO condition (dashed curve). As expected, this is because HADO benefits from the mutual informativeness between individual subjects, which is captured by the upper-level structure of the hierarchical model and makes it possible for the session to begin with significantly greater information. As the session continues, HADO needs 43 trials on average to reach the baseline gain level (dotted, horizontal line) whereas ADO (qCSF) requires 62 trials. The clear advantage diminishes as information accumulates further over the trials since the measure would eventually converge to a maximum as data accumulate.

The HBM condition (dash-dot curve), which employs the hierarchical modeling alone and no stimulus selection technique, enjoys the prior information provided by the hierarchical structure at the start of a session and exhibits greater information gain than the ADO method until it reaches trial 34. However, due to the lack of stimulus optimization, the speed of information gain is considerably slower, taking 152 trials to

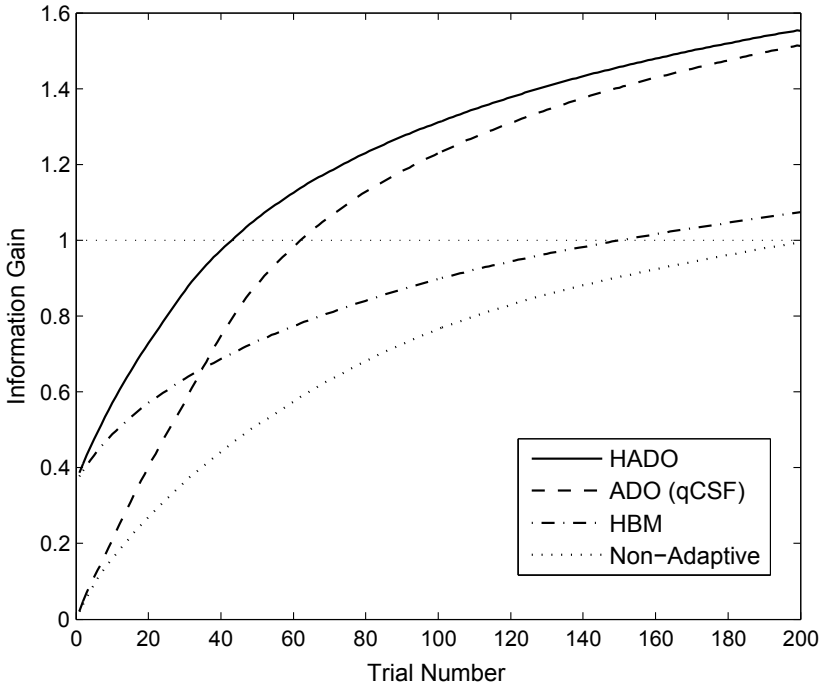

Figure 2: Information gain over measurement trials achieved by each of the four measurement methods.

attain baseline performance. The non-adaptive approach (dotted curve), with neither prior information nor design optimization, shows the lowest level of performance.

Information gain analyzed above may be viewed as a summary statistic, useful for evaluating the measurement methods under comparison. Not surprisingly, we were able to observe the same profile of performance differences in estimating the CSF parameters. The accuracy of a parameter estimate was assessed by the root mean squared error (RMSE) defined by

$$\mathrm{RMSE}\left(\hat{\psi}^{(t)}\right) = 20 \cdot \sqrt{\mathbb{E}\left[\left(\hat{\psi}^{(t)} - \psi^{\mathrm{true}}\right)^2\right]}, \tag{14}$$

where $\hat{\psi}^{(t)}$ is the estimate of one of the four CSF parameters (e.g., $\gamma^{\max}$) for a simulated subject, which was obtained as the posterior mean after observing trial t's outcome, ψ^{true} is the true, data-generating parameter value for that subject, and the factor of 20 is multiplied to read the measure on the decibel (dB) scale as the parameter values are base-10 logarithms. The expectation is assumed to be over all subjects and replications, and hence was replaced by the sample mean over 1,470 simulated sessions.

Results from the second analysis, comparing parameter estimation error for each of the four models, are shown in Figure 3. Error was quantified in terms of RMSE (y-axis; described above) over 200 trials (x-axis) for each of the four parameters. As with the case of information gain, HADO benefits from the informative prior through the hierar-

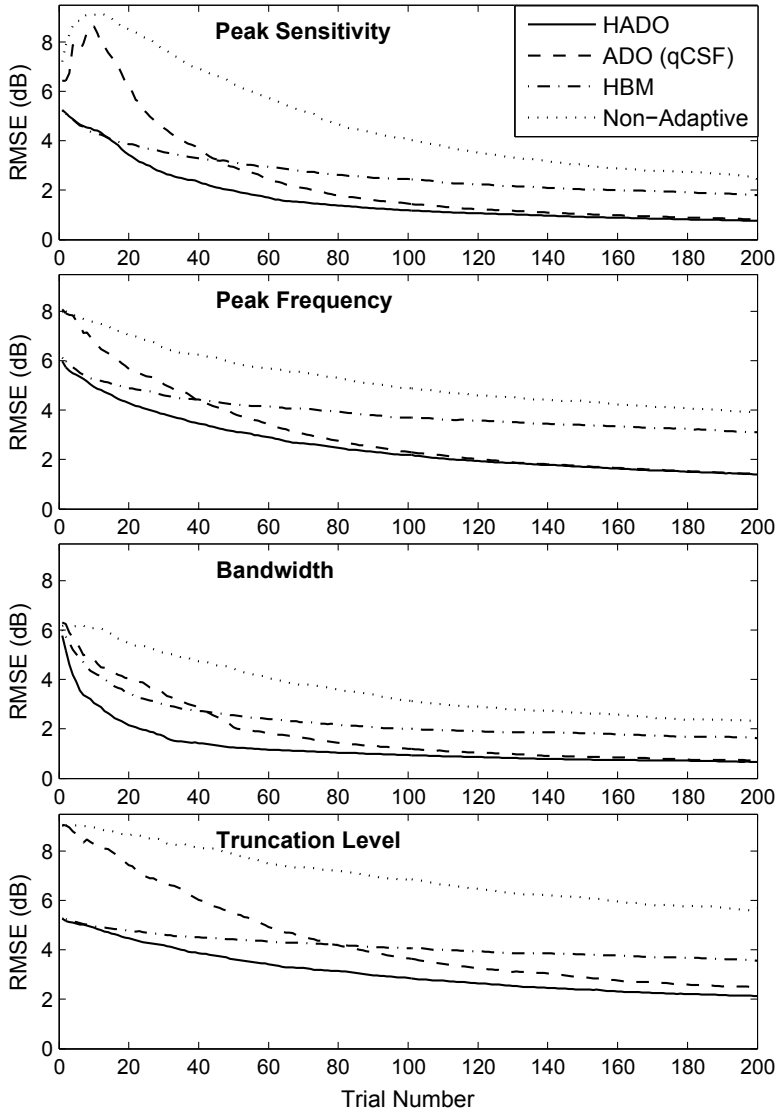

Figure 3: Accuracy of parameter estimation over measurement trials achieved by each of the four measurement methods.

chical model as well as the optimal stimuli through design optimization, exhibiting the lowest RMSE of all methods' from the start to the end of a session, and this holds for all four parameters. The benefit of the prior information is also apparent in the HBM condition, making the estimates more accurate than with the uninformed, ADO method for the initial 40 to 80 trials, but the advantage is eclipsed in further trials by the effect of design optimization in ADO.

Since accurate CSF measurements are often useful for screening eyes for disease, we performed yet another test of each method's performance, in which the estimated CSFs were put into a classifier for amblyopia. Despite various choices of a possible classifier (e.g., support vector machine, nearest neighbor, etc.), the logistic regression

model built on selected CSF traits (Hou et al., 2010), which had been shown to be effective in screening amblyopia, sufficed for our demonstration. Performance of each measurement method in classifying amblyopia was assessed in the leave-one-out fashion as well, by first fitting the logistic regression model using the remaining 146 subjects' CSF estimates (assumed to be the same regardless of the method being tested) and then entering the left-out, simulated subject's CSF estimate (obtained with the method evaluated in the simulation) into the classifier to generate a prediction. The given, actual label (i.e., amblyopic or normal eye) of the left-out subject, which had been provided by an actual clinical diagnosis, was taken as the true value against which the classification result in each simulation condition was scored.

Not surprisingly, classification accuracy increases with accumulation of measurement data in all methods. This is seen in Figure 4, which shows the percentage of correct amblyopia classifications out of all cases of amblyopic eyes over the first 100 measurement trials (i.e., hit rates).[5] As was found with the preceding tests, HADO demonstrates superior performance, requiring only a small number of trials to produce highly accurate classification results. Most notably, it takes on average 30 trials for HADO to correctly classifies an amplyopic eye 90% of the time, whereas the non-hierarchical adaptive method (ADO) requires 53 trials to achieve the same level of accuracy, otherwise reaching 82% accuracy with the same 30 trials.

Note that in the early trials of ADO and HADO, there can be considerable fluctuation in classification accuracy. This is not due to a small sample size (proportions out of 670 amblyopic eyes have sufficiently small standard errors), but rather to the adaptive method itself. Seeking the largest possible information gain, the algorithm is highly exploratory in choosing a stimulus that would yield a large change in the predicted state of the tested individual. This characteristic especially stands out in early trials of the classification task by causing some of the amblyopic eyes near the classifier's decision bound to alternate between the two sides of the bound across one trial to another. This effect remains even after taking proportions out of the large sample (670) because, with little accumulation of observations, selecting optimal stimuli in early trials is systematic without many possible paths of the selection. Although this can lead to short-term drops in accuracy, the benefits of early exploration pay dividends immediately and over the long term.

Finally, to see how this application of HADO performs when there is a small accumulation of data, an additional simulation was conducted with small n's ($n = 4, 10, 40$) assumed in the HADO condition. For each of the same, 147 simulated subjects (times 10 independent replications) as used before, HADO was used to estimate its CSF by assuming that only n, rather than all 146, subjects had been previously tested to be included in the hierarchical model estimation. Among the n (4, 10 or 40) data sets, half were randomly drawn from the normal-eye group and the other half from the amblyopic group.

[5] Classification results for normal eyes are not shown since the prior of CSF parameters was specified in a way that the classifier with any of the methods would categorize a subject as being normal when there is little or no accumulation of data (i.e., a bias was built in to avoid false alarms). In addition, the results are shown only up to 100 trials to provide a better view of performance differences across methods.

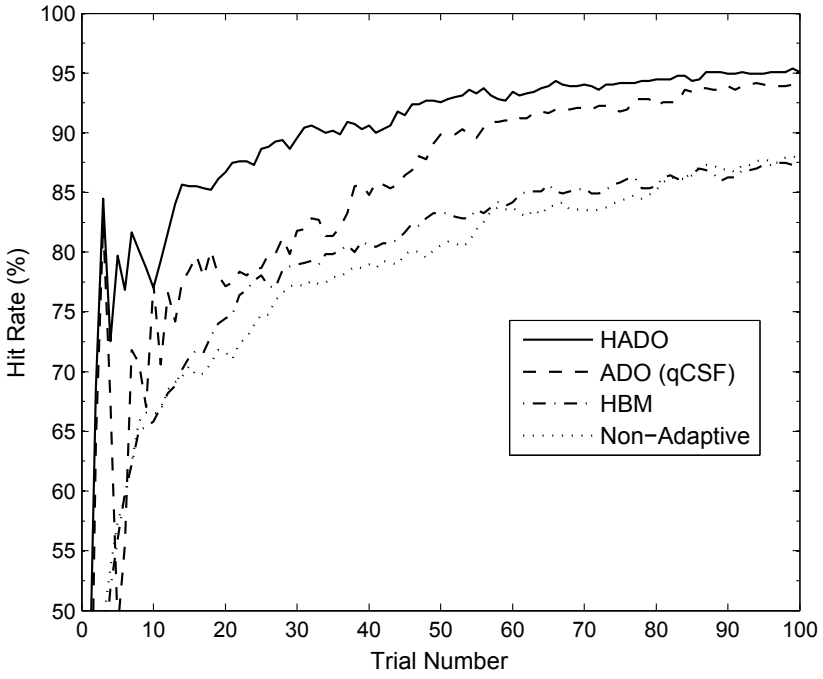

Figure 4: Accuracy of amblyopia classification over measurement trials achieved by each of the four measurement methods.

The results are in Figure 5, which displays the RMSE measures for estimating the peak sensitivity parameter (other evaluation measures exhibit a similar pattern, leading to the same interpretation). For comparison, the data from the ADO and full HADO conditions are also plotted. CSF estimation by HADO with n as small as 4 is no worse, and in fact slightly more efficient, than that of ADO with a diffuse prior, as shown by the RMSE's when $n = 4$ (dash-dot curve) being consistently lower than those of ADO (dotted curve) over trials. Though not shown here, visual inspection of the distribution of individual estimates over all subjects and replications showed no larger dispersion than the case of estimates by ADO at all trials. As n increases or more data from additional subjects are available, the efficiency of HADO estimation becomes higher (dashed and thin solid curves for $n = 10$ and $n = 40$), approaching the performance level of HADO with full data sets (thick solid curve). These results indicate that the Bayesian estimation of this hierarchical model is robust enough to take advantage of even a small sample of previously collected data. However, as noted in Implementation Considerations, the effect of small n may depend on the model employed, suggesting that the above observation would not generalize to all potential HADO applications.

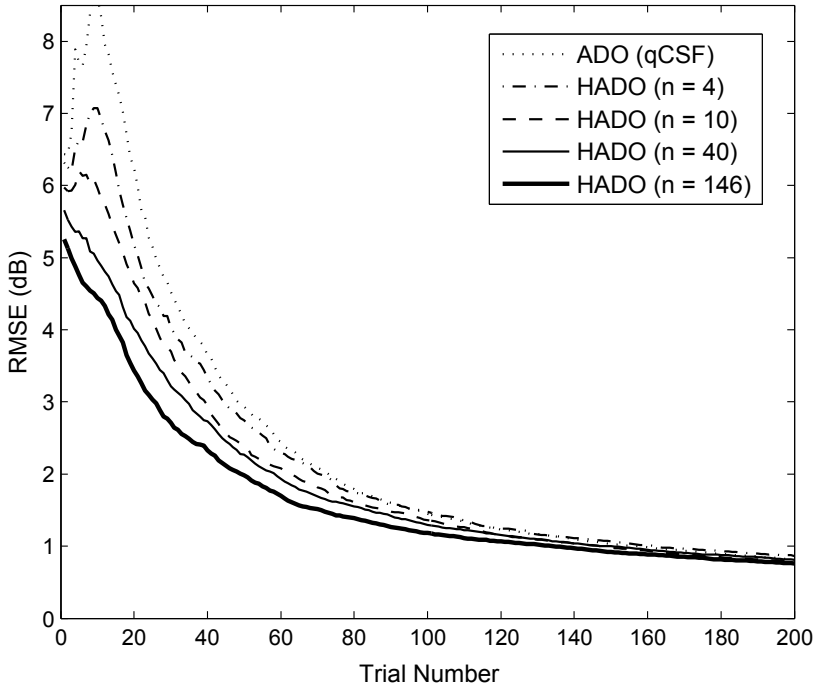

Figure 5: Effect of the size of previously collected data sets on HADO estimation accuracy of the peak sensitivity parameter.

5 Discussion

The present study demonstrates how hierarchical Bayes modeling can be integrated into adaptive design optimization to improve the efficiency and accuracy of measurement. When applied to the problem of estimating a contrast sensitivity function (CSF) in visual psychophysics, HADO achieved an average decrease of 38% (from 4.9 dB to 3.1 dB) in error of CSF parameter estimation and an increase of 10% (from 82% to 90%) in accuracy of eye disease screening over conventional ADO, under the scenario that a new session could afford to make only 30 measurement trials. In addition, efficiency of testing improved by an average of 43% in the sense that the required number of trials to reach a criterion of 90% screening accuracy decreased from 53 to 30 trials.

Although the simulation study served the purpose of demonstrating the benefit of the hierarchical adaptive methodology, the full potential of HADO should be greater than that shown in our particular example. The level of improvement possible with HADO depends on the sophistication of the hierarchical model itself. In our case, the model was based on a simple hypothesis that a newly tested individual belongs to the population from which all other individuals have been drawn. Although the model has flexibility in defining the population as a mixture distribution, it conveys no further specific information about the likely state of a new individual (e.g., his or her membership to a mixture component is unknown).

There are various situations in which hierarchical modeling can take better advantage of the data-generating structure. For example, although modeled behavioral traits vary across individuals, they may covary with other variables that can be easily observed, such as demographic information (e.g., age, gender, occupation, etc.) or other measurement data (e.g., contrast sensitivity correlates with measures of visual acuity - eye chart test). In this case, a general, multivariate regression or ANOVA model may be employed as the upper-level structure to utilize such auxiliary information to define a more detailed relationship between individuals. This greater detail in the hierarchical model should promote efficient measurement by providing more precise information about the state of future individuals.

In many areas of behavioral science, there is more than one test that measures the same condition or phenomenon (e.g., memory, depression, attitudes). Often times, these tests are related to each other and modeled within a similar theoretical framework. In such situations, a hierarchical model provides a well-justified way to integrate those models in such a way that behavioral traits inferred under one model are informative about those estimated by another. Yet another situation in which hierarchical modeling would be beneficial is when a measurement is made after some treatment and it is sensible or even well known that the follow-up test has a particular direction of change in its outcome (i.e., increase or decrease). Taking this scenario one step further, a battery of tests may be assumed to exhibit profiles that are characteristic of certain groups of individuals. The upper-level structure can also be modeled (e.g., by an autoregressive model) to account for such transitional variability in terms of the parameters of the measurement model. With these kinds of structure built in the hierarchical model, HADO can be used to infer quickly the state of new individuals.

An assumption of the approaches to higher-level modeling discussed so far is that the most suitable data-generating structure is already known. In fact, a sufficient amount of data is needed to determine which structure is best suited. To be more precise, the optimally complex structure for the best possible inference depends on the amount of information available; an arbitrarily complex model that is not validated by data will lead to sub-optimal inference. For this reason, HADO will perform best when the hierarchical model *evolves* along with the accumulation of data. Larger data sets make it possible to evaluate better alternative modeling hypotheses, and analysis methods such as Bayesian model choice (Kass & Raftery, 1995) or cross validation can be performed to guide model revision. In effect, the upper-level model will evolve by incorporating increasingly richer structure (e.g., finer subgroup distinctions or better selected predictor variables in a regression model).

The notion of model evolution fits with recent advances in nonparametric Bayes methods that essentially seek to enable a statistical model to adapt itself to the amount of information in the data by adding more and more components with no preset limit (MacEachern, 2000; Rasmussen & Williams, 2006; Teh & Jordan, 2010). This methodology can further stretch the extent of model evolution and will be especially suited to HADO because most modern measurement processes are computer-based, so data collection and organization are effortless, allowing the method to quickly exploit a massive amount of data.

The technique of optimal experimental design or active learning has been applied to a number of modeling problems in neuroscience and machine learning (Wu, David, &

Gallant, 2006; Lewi et al., 2009; DiMattina & Zhang, 2011; Cohn et al., 1996; Tong & Koller, 2002; Settles, 2010). These models usually deal with a large number of features in order to predict or describe response variables, resulting in a large number of parameters to infer (e.g., neural receptive field modeling; Wu et al., 2006). A consequence of this is the use of various methods for improving generalizability by imposing certain constraints (e.g., Ramirez et al., 2011; Park & Pillow, 2012), which may be directly or indirectly interpreted as a prior from the Bayesian perspective. In other words, a prior is used to reduce the variance of a model. However, as this type of a prior is theoretically derived, it is by nature conservative in order not to introduce bias. In this case, HADO may be employed to enhance inference by learning further prior knowledge from specific empirical conditions. This information may be encapsulated into the existing, constrained structure of a model. To this end, different forms of HADO described in the formulation section will be useful. Computational complexity, particularly numerical integration over many parameters, will be challenging. Nonetheless, this should not be considered a hindrance—as discussed in Implementation Considerations, recent technical advances in both algorithms and hardware as well as inherent regularity in each problem can be taken advantage of to achieve adequate approximations with practical running time.

To conclude, science and society benefit when data collection is efficient with no loss of accuracy. The proposed HADO framework, which judiciously integrates the best features of design optimization and hierarchical modeling, is an exciting new tool that can significantly improve upon the current state of the art in experimental design, enhancing both measurement and inference. This theoretically well-justified and widely applicable experimental tool should help accelerate the pace of scientific advancement in behavioral and neural sciences.

Acknowledgments

This research is supported by National Institute of Health Grant R01-MH093838 to J.I.M and M.A.P., as well as National Eye Institute Grant R01-EY021553-01 to Z.-L.L. We thank Fang Hou and Chang-Bing Huang for organizing the CSF measurement data.
Correspondence concerning this article should be addressed to Woojae Kim, Department of Psychology, Ohio State University, 1835 Neil Avenue, Columbus, OH 43210. Email: kim.1124@osu.edu.

References

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*. Oxford University Press.

Bernado, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York: Wiley.

Box, G. F. B., & Hill, W. J. (1967). Discrimination among mechanistic models. *Technometrics*, *9*, 57-71.

Cavagnaro, D. R., Gonzalez, R., Myung, J. I., & Pitt, M. A. (2013). Optimal decision stimuli for risky choice experiments: An adaptive approach. *Management Science*, *59*(2), 358–375.

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. V. (2010). Adaptive design optimization: A mutual information based approach to model discrimination in cognitive science. *Neural Computation*, *22*(4), 887–905.

Cavagnaro, D. R., Pitt, M. A., Gonzalez, R., & Myung, J. I. (2013). Discriminating among probability weighting functions using adaptive design optimization. *Journal of Risk and Uncertainty*, *47*(3), 255-289.

Cavagnaro, D. R., Pitt, M. A., & Myung, J. I. (2009). Adaptive design optimization in experiments with people. *Advances in Neural Information Processing Systems*, *22*, 234-242.

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, *10*(3), 273-304.

Chernoff, H. (1959). Sequential design of experiments. *Annals of Mathematical Statistics*, 755–770.

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

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. Hoboken, New Jersey: John Wiley & Sons, Inc.

de Finetti, B. (1974). *Theory of probability*. New York: Wiley.

DiMattina, C., & Zhang, K. (2008). How optimal stimuli for sensory neurons are constrained by network architecture. *Neural Computation*, *20*, 668-708.

DiMattina, C., & Zhang, K. (2011). Active data collection for efficient estimation and comparison of nonlinear neural models. *Neural Computation*, *23*, 2242–2288.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis (2nd edition)*. Boca Raton, Florida: Chapman & Hall/CRC.

Good, I. J. (1965). *The estimation of probabilities: An essay on modern Bayesian methods*. Cambridge, MA: MIT Press.

Griebel, M., & Holtz, M. (2010). Dimension-wise integration of high-dimensional functions with applications to finance. *Journal of Complexity*, *26*(5), 455–489.

Heiss, F., & Winschel, V. (2008). Likelihood approximation by numerical integration on sparse grids. *Journal of Econometrics*, *144*(1), 62–80.

Holtz, M. (2011). *Sparse grid quadrature in high dimensions with applications in finance and insurance*. New York: Springer.

Hou, F., Huang, C. B., Lesmes, L., Feng, L. X., Tao, L., Zhou, Y. F., & Lu, Z.-L. (2010). qCSF in clinical application: efficient characterization and classification of contrast sensitivity functions in amblyopia. *Investigative Ophthalmology & Visual Science*, *51*(10), 5365–5377.

Jordan, M. I. (Ed.). (1998). *Learning in graphical models*. Cambridge, MA: MIT 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.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. Cambridge, MA: MIT Press.

Kujala, J. V., & Lukka, T. J. (2006). Bayesian adaptive estimation: The next dimension. *Journal of Mathematical Psychology*, *50*(4), 369–389.

Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, *30*, 55–580.

Lesmes, L. A., Jeon, S.-T., Lu, Z.-L., & Dosher, B. A. (2006). Bayesian adaptive estimation of threshold versus contrast external noise functions: The quick TvC method. *Vision Research*, *46*, 3160–3176.

Lesmes, L. A., Lu, Z.-L., Baek, J., & Albright, T. D. (2010). Bayesian adaptive estimation of the contrast sensitivity function: The quick CSF method. *Journal of Vision*, *10*, 1–21.

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.

MacEachern, S. N. (2000). *Dependent Dirichlet processes* (Tech. Rep.). Columbus, OH: Ohio State University.

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, J. I., Cavagnaro, D. R., & Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology*, *57*, 53-67.

Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., & Sejnowski, T. J. (2011). Experience matters: Information acquisition optimizes probability gain. *Psychological Science*, *21*(7), 960–969.

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.

Park, M., & Pillow, J. W. (2012). Bayesian active learning with localized priors for fast receptive field characterization. *Advances in Neural Information Processing Systems*, *25*, 2357–2365.

Pflüger, D., Peherstorfer, B., & Bungartz, H.-J. (2010). Spatially adaptive sparse grids for high-dimensional data-driven problems. *Journal of Complexity*, *26*(5), 508–522.

Ramirez, A. D., Ahmadian, Y., Schumacher, J., Schneider, D., Woolley, S. M., & Paninski, L. (2011). Incorporating naturalistic correlation structure improves spectrogram reconstruction from neuronal activity in the songbird auditory midbrain. *Journal of Neuroscience*, *31*(10), 3828–3842.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods (2nd edition)*. New York, NY: Springer.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*,

12, 573–604.

Rouder, J. N., Speckman, P. L., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*, 195–223.

Scott, D. W. (1992). *Multivariate density estimation: theory, practice, and visualization*. New York, NY: John Wiley & Sons.

Settles, B. (2010). *Active learning literature survey* (Tech. Rep.). Madison, WI: University of Wisconsin–Madison.

Seung, H. S., Opper, M., & Sompolinsky, H. (1992). Query by committee. In *Proceedings of the fifth workshop on computational learning theory* (pp. 287–294). San Mateo, CA: Morgan Kaufman.

Sugiyama, M., & Rubens, N. (2008). A batch ensemble approach to active learning with model selection. *Neural Networks*, *21*(9), 1278–1286.

Tang, Y., Young, C., Myung, J. I., Pitt, M. A., & Opfer, J. (2010). Optimal inference and feedback for representational change. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual meeting of the cognitive science society* (p. 2572-2577). Austin, TX: Cognitive Science Society.

Teh, Y. W., & Jordan, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In N. Hjort, C. Holmes, P. Müller, & S. Walker (Eds.), *Bayesian nonparametrics.* London, U.K.: Cambridge University Press.

Tong, S., & Koller, D. (2002). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, *2*, 45–66.

Tulsyan, A., Forbes, J. F., & Huang, B. (2012). Designing priors for robust bayesian optimal experimental design. *Journal of Process Control*, *22*(2), 450–462.

Watson, A. B., & Ahumada, A. J. (2005). A standard model for foveal detection of spatial contrast. *Journal of Vision*, *5*(9), 717–740.

Woolley, S. M., Gill, P. R., & Theunissen, F. E. (2006). Stimulus-dependent auditory tuning results in synchronous population coding of vocalizations in the songbird midbrain. *Journal of Neuroscience*, *26*(9), 2499–2512.

Wu, M. C.-K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, *29*, 477–505.