



## Optimal experimental design for a class of bandit problems

Shunan Zhang\*, Michael D. Lee

Department of Cognitive Sciences, University of California, Irvine, CA, 92697-5100, United States

### ARTICLE INFO

#### Article history:

Received 18 March 2010

Received in revised form

27 July 2010

Available online 24 September 2010

### ABSTRACT

Bandit problems are a class of sequential decision-making problems that are useful for studying human decision-making, especially in the context of understanding how people balance exploration with exploitation. A major goal of measuring people's behavior using bandit problems is to distinguish between competing models of their decision-making. This raises a question of experimental design: How should a set of bandit problems be designed to maximize the ability to discriminate between models? We apply a previously developed design optimization framework to the problem of finding good bandit problem experiments, and develop computational sampling schemes for implementing the approach. We demonstrate the approach in a number of simple cases, varying the priors on parameters for some standard models. We also demonstrate the approach using empirical priors, inferred by hierarchical Bayesian analysis from human data, and show that optimally designed bandit problems significantly enhance the ability to discriminate between competing models.

© 2010 Elsevier Inc. All rights reserved.

### 1. Experimental design for model discrimination

A basic challenge for experimentation in psychology is to design experiments that will provide the most useful data. Often, the most useful data will be those that address underlying theoretical issues, by providing clear evidence for and against competing models implementing alternative theoretical positions. Traditionally, psychological experiments have been designed to meet these goals based on a mixture of previous results, pilot information, and the intuition of the experimenter.

Formal approaches to experimental design optimization, however, have received considerable attention in statistics and engineering (e.g., Atkinson & Donev, 1992; Box & Hill, 1967; Chaloner & Verdinelli, 1995; Kiefer, 1959). Recently, psychologists have also started to search for approaches that allow the formal optimization of the design of an experiment (e.g., Myung & Pitt, 2009). Of course, it is not possible to quantify every aspect of most experimental designs, and so there will likely always remain a role for intuition and experience. But, for those aspects of a design that are amenable to quantification, formal methods for finding the best designs can potentially make a major contribution. A good experimental design should allow competing models to be better discriminated once a fixed number of data are collected, or, alternatively, allow fewer data to be collected to achieve a required level of discriminability.

In this paper, we adapt the formal framework for experimental design optimization described by Myung and Pitt (2009) to a

research area where it has not previously been applied. Our research area involves a class of sequential decision-making problems, known as bandit problems, which have been widely studied in the cognitive science and machine-learning literatures. Because they have been widely studied, there are many different formal models of decision-making on bandit problems, many of which are sensible candidates as accounts of how people solve the problems. Thus, it is natural to ask how bandit problem experiments with people should be designed, so as to maximize the usefulness of the data in distinguishing these competing models.

The plan of the paper is as follows. In the next section, we formally define the bandit problems we study, and the models we want to be able to test experimentally. We then present the general framework for design optimization developed by Myung and Pitt (2009), and apply it to bandit problem experiments. We demonstrate and evaluate the approach on some simple cases, and then in some more realistic comparisons using empirically inferred priors from human data. We conclude with a discussion of the contribution of our work, directions for future work, and the role of design optimization in helping understand human cognition.

### 2. Two-armed finite-horizon bandit problems

In the bandit problems we consider, a decision-maker chooses between two alternatives over a fixed and known number of trials. Choosing an alternative results in either a success or a failure. This is determined probabilistically using a reward rate for each alternative, which is hidden from the decision-maker, but fixed over all the trials. The goal of the decision-maker is simply to maximize the number of successes over all the trials. Among the

\* Corresponding author.

E-mail address: [szhang@uci.edu](mailto:szhang@uci.edu) (S. Zhang).

broader class of bandit problems (e.g., Sutton & Barto, 1998), the versions we are studying are two-armed finite-horizon problems, because there are just two alternative choices, and the number of trials is fixed.

In terms of human decision-making, bandit problems provide an interesting formal setting for studying the balance between exploration and exploitation in decision-making. In early trials, it makes sense to explore different alternatives, searching for those with the highest reward rates. In later trials, it makes sense to exploit those alternatives known to be good, by choosing them repeatedly. How exactly this balance between exploration and exploitation should be managed, and should be influenced by factors such as the distribution of reward rates, the total number of trials, and so on, are challenging questions. The interplay between optimal decision-making and these factors raises basic questions about adaptation, planning, and learning in intelligent systems. It is for this reason that bandit problems have been widely studied in the machine-learning (Berry & Fristedt, 1985; Gittins, 1979; Kaebbling, Littman, & Moore, 1996; Macready & Wolpert, 1998; Sutton & Barto, 1998) and cognitive science (Cohen, McClure, & Yu, 2007; Daw, O’Doherty, Dayan, Seymour, & Dolan, 2006; Steyvers, Lee, & Wagenmakers, 2009) literatures, and many models of decision-making strategies have been proposed.

Formally, we denote the reward rates for the two alternatives as  $p_1$  and  $p_2$ , and assume they are both drawn independently from the same environmental distribution. A natural, flexible and tractable choice for environmental distributions is given by the family of Beta distributions with parameters  $\alpha$  and  $\beta$ . Thus, we assume that

$$p_i \sim \text{Beta}(\alpha, \beta), \quad i = 1, 2. \tag{1}$$

We denote the observed data of the  $i$ th alternative being chosen on the  $j$ th trial as  $y_j = i$ . The observed success or failure resulting from this choice is represented by the binary variable  $r_j$ , and is determined as

$$r_j \sim \text{Bernoulli}(p_i). \tag{2}$$

In the specific experimental setting we consider, the decision-maker completes a set of bandit problems, within the same environment. In other words, the reward rates for each problem are independent draws from the same Beta distribution. This means that there are two ways to conceive a problem of optimal design. One focuses on the individual reward rates  $p_1$  and  $p_2$  for the alternatives within problems, and the other focuses on the parameters  $\alpha$  and  $\beta$  that characterize the reward environment. In this paper, we address just the first, more basic, optimization problem regarding individual reward rates, but note that it is a building block to addressing the more general design optimization problem of specifying the best reward environment.

### 3. Models of human decision-making

To demonstrate how design optimization can be applied to bandit problems, we consider three potential models of human decision-making. All three have been widely considered as accounts of human decision-making, and correspond to different theoretical assumptions about how people balance exploration and exploitation (e.g., Cohen et al., 2007; Daw et al., 2006; Lee, Zhang, Munro, & Steyvers, in press).

The first model we consider is known as Win–Stay Lose–Shift (WSLS; Robbins, 1952), which is perhaps the simplest reasonable account. In its deterministic form, it assumes that the decision-maker continues to choose an alternative following a reward, but shifts to the other alternative following a failure to reward. In the stochastic form we use, the probability of staying after winning, and the probability of shifting after losing, are both

parameterized by the same probability  $\gamma$ . Formally, the WSLS model has likelihood function

$$p_{\text{WSLS}}(y_j = i) = \begin{cases} 1/2 & \text{if } j = 1 \\ \gamma & \text{if } j > 1, r_{j-1} = 1 \text{ and } y_{j-1} = i \\ 1 - \gamma & \text{if } j > 1, r_{j-1} = 0 \text{ and } y_{j-1} = i \\ \gamma & \text{if } j > 1, r_{j-1} = 0 \text{ and } y_{j-1} \neq i \\ 1 - \gamma & \text{if } j > 1, r_{j-1} = 1 \text{ and } y_{j-1} \neq i, \end{cases} \tag{3}$$

and is completed by placing a prior on the probability  $\gamma$ . Psychologically, the WSLS model does not require a memory, because its decisions only depend on the presence or absence of a reward on the previous trial. Nor is the heuristic sensitive to the horizon, because its decision process is the same for all trials.

The second model we consider is a natural extension of the WSLS model. Rather than having the same probability for winning and staying as for losing and shifting, the extended-WSLS model has a separate probability for each, given by  $\gamma_w$  for winning and  $\gamma_l$  for losing. This model has likelihood function

$$p_{\text{e-WSLS}}(y_j = i) = \begin{cases} 1/2 & \text{if } j = 1 \\ \gamma_w & \text{if } j > 1, r_{j-1} = 1 \text{ and } y_{j-1} = i \\ 1 - \gamma_l & \text{if } j > 1, r_{j-1} = 0 \text{ and } y_{j-1} = i \\ \gamma_l & \text{if } j > 1, r_{j-1} = 0 \text{ and } y_{j-1} \neq i \\ 1 - \gamma_w & \text{if } j > 1, r_{j-1} = 1 \text{ and } y_{j-1} \neq i, \end{cases} \tag{4}$$

and is completed by placing a prior on the probabilities  $\gamma_w$  and  $\gamma_l$ . Psychologically, the extension of the WSLS model corresponds to the theoretical idea that successes and failures are different, in the sense that obtaining a reward is different from not obtaining a failure. This is a basic theoretical distinction throughout psychology, especially in fields like operant conditioning, which have previously studied bandit problems (e.g., Brand, Sakoda, & Woods, 1957; Brand, Wood, & Sakoda, 1956). The WSLS and e-WSLS models formalize this theoretical distinction, and it becomes natural to ask which is able to model human behavior better. Note that the original WSLS model is a special case of the e-WSLS model, obtained by setting  $\gamma_w = \gamma_l$ , and so the evaluation of the two models constitutes a nested model comparison.

The third model we consider is the  $\epsilon$ -greedy model, which is probably the most standard account of bandit problem decision-making from the field of reinforcement learning (Sutton & Barto, 1998). It assumes that decision-making is driven by a parameter  $\epsilon$  that controls the balance between exploration and exploitation. On each trial, with probability  $1 - \epsilon$  the decision-maker chooses the alternative with the greatest estimated reward rate (i.e., the greatest proportion of rewards obtained for previous trials where the alternative was chosen). This can be conceived as an ‘exploitation’ decision. With probability  $\epsilon$ , the decision-maker chooses randomly. This can be conceived as an ‘exploration’ decision. Formally, for our two-alternative problems, denoting the number of successes and failures for the  $i$ th alternative up until the  $j$ th trial as  $s_{ij}$  and  $f_{ij}$  respectively, the likelihood function is

$$p_{\epsilon\text{-greedy}}(y_j = 1) = \begin{cases} 1/2 & \text{if } j = 1 \\ 1/2 & \text{if } j > 1, \frac{s_{1j} + 1}{s_{1j} + f_{1j} + 1} = \frac{s_{2j} + 1}{s_{2j} + f_{2j} + 1} \\ \epsilon & \text{if } j > 1, \frac{s_{1j} + 1}{s_{1j} + f_{1j} + 1} < \frac{s_{2j} + 1}{s_{2j} + f_{2j} + 1} \\ 1 - \epsilon & \text{if } j > 1, \frac{s_{1j} + 1}{s_{1j} + f_{1j} + 1} > \frac{s_{2j} + 1}{s_{2j} + f_{2j} + 1}, \end{cases} \tag{5}$$

and is completed by placing a prior on the probability  $\epsilon$ . Psychologically, the  $\epsilon$ -greedy model does require a limited form of memory, because it has to remember counts of previous successes and failures for each alternative. This corresponds to a very different theoretical position from the WSLS models, and again makes the empirical evaluation of the model important. Note that the  $\epsilon$ -greedy model is quite different from the WSLS models, and so involves a non-nested model comparison.

#### 4. Optimizing bandit problem designs

##### 4.1. General framework for design optimization

The framework for design optimization proposed by Myung and Pitt (2009) involves maximizing the expected value of a utility function over a design space. This maximization is done in terms of the models to be discriminated, and so requires likelihood functions and prior distributions over parameters for the models of interest.

Most generally, we have a space of (potentially infinite) models,  $\mathcal{M}$ , and specific designs  $\mathbf{d}$  within an experimental space,  $\mathbf{d} \in \mathcal{D}$ , that are intended to evaluate the models  $M \in \mathcal{M}$ . The design  $\mathbf{d}$  quantifies just those aspects of the experimental design that are amenable to quantification, and that we want to optimize. We also have parameters  $\theta$  for all of the possible models in  $\mathcal{M}$ . An experiment will produce a distribution over observed data  $\mathbf{y}$  that depends on both the model parameters and the design of the experiment, which we can write as  $p(\mathbf{y} | \theta, \mathbf{d})$ . It is then possible to maximize a utility function  $u(\mathbf{d}, \theta, \mathbf{y})$ , to find the optimal experimental design, as the one that maximizes the expected utility of the observed data, as averaged over all data, model parameters, and models. This optimal design is given by  $\mathbf{d}^* = \arg \max_{\mathbf{d} \in \mathcal{D}} U(\mathbf{d})$ , where

$$U(\mathbf{d}) = \iiint u(\mathbf{d}, \theta, \mathbf{y}) p(\mathbf{d}, \theta, \mathbf{y}) d\theta d\mathbf{y} d\mathcal{M} \\ = \iiint u(\mathbf{d}, \theta, \mathbf{y}) p(\mathbf{y} | \theta, \mathbf{d}) p(\theta | \mathbf{d}) d\theta d\mathbf{y} d\mathcal{M}. \quad (6)$$

Myung and Pitt (2009) consider a specific case of this very general framework, in which the experimental goal is to discriminate between just two models,  $M_a$  and  $M_b$ . In this case, there are two sets of model parameters  $\theta_a$  and  $\theta_b$ , so  $\theta = (\theta_a, \theta_b)$ . The intuitive approach they adopt is to fit each model to data generated by the other, and search for the design under which the maximal lack-of-fit is achieved. The integration over all models in Eq. (6) is replaced by a weighted sum over the two competing models, combining the lack-of-fit measure for each model with respect to data generated by the other. Under this scheme, the global utility function  $U(\mathbf{d})$  can be rewritten as

$$U(\mathbf{d}) = p(M_a) \iint u(\mathbf{d}, \theta_a, \mathbf{y}_a) p(\mathbf{y}_a | \theta_a, \mathbf{d}) p(\theta_a | \mathbf{d}) d\theta_a d\mathbf{y}_a \\ + p(M_b) \iint u(\mathbf{d}, \theta_b, \mathbf{y}_b) p(\mathbf{y}_b | \theta_b, \mathbf{d}) p(\theta_b | \mathbf{d}) d\theta_b d\mathbf{y}_b, \quad (7)$$

where  $p(M_a)$  and  $p(M_b)$  are the prior probabilities of the two models.

The optimal design thus depends on both the models and the nature of the local utility functions. The local utility should in some sense reflect the likelihood of the “correct” model based on the observed data. We use the Bayes factor, which is a standard and principled Bayesian model selection measure (Kass & Raftery, 1995).<sup>1</sup> The Bayes factor is the ratio of marginal likelihoods (or, equivalently the ratio of posterior to prior model odds), and so, for example, for data generated by  $M_a$  is given by

$$BF_{a/b} = \frac{\int p(\mathbf{y}_a | \theta_a) p(\theta_a) d\theta_a}{\int p(\mathbf{y}_a | \theta_b) p(\theta_b) d\theta_b}. \quad (8)$$

<sup>1</sup> Of course, any monotonic transformation of the Bayes factor could be used, if the goal was to find the single best design. It is possible some such transformation allows for improved computational efficiency in terms of convergence.

So, we have finally formulated the design optimization problem as one of finding the design that maximizes

$$U(\mathbf{d}) = p(M_a) \iint BF_{a/b} p(\mathbf{y}_a | \theta_a, \mathbf{d}) p(\theta_a | \mathbf{d}) d\theta_a d\mathbf{y}_a \\ + p(M_b) \iint BF_{b/a} p(\mathbf{y}_b | \theta_b, \mathbf{d}) p(\theta_b | \mathbf{d}) d\theta_b d\mathbf{y}_b. \quad (9)$$

##### 4.2. Application to bandit problems

We consider two specific applications of design optimization to two-alternative finite-horizon bandit problems. The first involves comparing the WSLs and e-WSLS models, and the second involves comparing the WSLs and  $\epsilon$ -greedy models.

###### 4.2.1. WSLs versus e-WSLS

The model parameters are  $\theta_a = \gamma$  and  $\theta_b = (\gamma_w, \gamma_l)$ . The likelihood functions are  $p_{\text{WSLS}}(\mathbf{y}_a | \gamma)$  and  $p_{\text{e-WSLS}}(\mathbf{y}_b | \gamma_w, \gamma_l)$ , given by Eqs. (3) and (4). Because we have no a priori reason to believe one model is better than the other, we set  $p(M_a) = p(M_b) = 1/2$  throughout. We also do not believe that the priors for the parameters of the model depend on the experimental design, and so assume that  $p(\gamma | \mathbf{d}) = p(\gamma)$  and  $p(\gamma_w, \gamma_l | \mathbf{d}) = p(\gamma_w, \gamma_l)$ .

What we do assume is that the parameters for the winning-and-staying and losing-and-shifting probabilities can be represented by Beta distributions, and that the two probabilities for the e-WSLS model are independent. Formally, we assume that  $\gamma \sim \text{Beta}(\alpha, \beta)$ ,  $\gamma_w \sim \text{Beta}(\alpha_w, \beta_w)$  and  $\gamma_l \sim \text{Beta}(\alpha_l, \beta_l)$ . This assumption is convenient, because it allows us to specify a wide range of interpretable priors for the probabilities, and it also means that the required Bayes factors can be determined analytically. This can be done using four counts that are sufficient summaries of the data with respect to the two WSLs models. In particular, if the data  $\mathbf{y}_a$  involve  $k_w$  counts of staying-after-winning, from a total of  $n_w$  wins, and  $k_l$  counts of shifting-after-losing, after a total of  $n_l$  losses, then, from Eq. (8), we have  $BF_{a/b}$ , given in Box I, where  $B(\cdot, \cdot)$  is the Beta function. It is straightforward to derive  $BF_{b/a}$  similarly; it uses data  $\mathbf{y}_b$  generated by the e-WSLS model to derive the counts, and is otherwise the reciprocal of Eq. (10) in Box I.

###### 4.2.2. WSLs versus $\epsilon$ -greedy

The model parameters are now  $\theta_a = \gamma$  and  $\theta_b = \epsilon$ . The likelihood functions are  $p_{\text{WSLS}}(\mathbf{y}_a | \gamma)$  and  $p_{\epsilon\text{-greedy}}$ , given by Eqs. (3) and (5). We again set  $p(M_a) = p(M_b) = 1/2$  throughout and assume that the priors have Beta distributions, so  $\gamma \sim \text{Beta}(\alpha, \beta)$ ,  $\epsilon \sim \text{Beta}(\alpha_\epsilon, \beta_\epsilon)$ .

As before, this allows the Bayes factor to be given by a closed-form expression, although the data now need to be summarized in ways suited to both of the non-nested models. For the WSLs model, we continue summarizing the data  $\mathbf{y}_a$  as involving  $k_w$  counts of staying-after-winning, from a total of  $n_w$  wins, and  $k_l$  counts of shifting-after-losing, after a total of  $n_l$  losses. For the  $\epsilon$ -greedy model, following the logic of the model presented in Eq. (5), we count the number of times  $k_g$  the data follow the “good” choice, according to the observed success ratios, the number of times  $k_b$  the data follow the “bad” choice, and the number of times  $k_e$  the two alternatives had equivalent success ratios. These three cases exhaustively partition all of the choices for all of the trials in the data.

With these counts in place, the Bayes factor for generated data  $\mathbf{y}_a$  is

$$BF_{a/b} = \frac{\int p(k_w, k_l | \gamma) p(\gamma) d\gamma}{\iint p(k_g, k_b, k_e | \epsilon) p(\epsilon) d\epsilon} \\ = \frac{B(\alpha_\epsilon, \beta_\epsilon) B(k_w + k_l + \alpha, n_w - k_w + n_l - k_l + \beta)}{(1/2)^{k_e} B(\alpha, \beta) B(k_b + \alpha_\epsilon, k_g + \beta_\epsilon)}. \quad (11)$$

$$\begin{aligned}
 \text{BF}_{a/b} &= \frac{\int p(k_w, k_l | \gamma) p(\gamma) d\gamma}{\iint p(k_w, k_l | \gamma_w, \gamma_l) p(\gamma_w, \gamma_l) d\gamma_w d\gamma_l} \\
 &= \frac{B(\alpha_w, \beta_w) B(\alpha_l, \beta_l) B(k_w + k_l + \alpha, n_w - k_w + n_l - k_l + \beta)}{B(\alpha, \beta) B(k_w + \alpha_w, n_w - k_w + \beta_w) B(k_l + \alpha_l, n_l - k_l + \beta_l)}.
 \end{aligned}
 \tag{10}$$

**Box I.**

**5. Computational method**

Finding the best design to discriminate models involves the maximization problem in Eq. (9), which involves a double integral that, in general, will not be analytically tractable. Accordingly, computational methods are required. We briefly describe the most straightforward approach, based on a grid approximation, which we found to be too computationally inefficient, before describing a better-performing computational approach based on a Markov chain Monte Carlo (MCMC) technique.

*5.1. Grid method*

The grid approach considers a large number of candidate designs  $\mathbf{d}$ , sampled systematically from a grid over the design space. For our design problem, this means choosing some resolution  $\Delta p$ , and using the grid  $(p_1, p_2) = (0, \Delta p, \dots, 1) \times (0, \Delta p, \dots, 1)$ . For each design on the grid, parameters are drawn for the two models being evaluated, sampled from the priors for those models. Based on these parameters, and the design being considered, data are generated from each model. From these data, the Bayes factors comparing the ability of each model to fit the other's data can be calculated. Finally, these Bayes factors can be combined to give a utility. The average of many of these utilities then estimates the utility for the design. Once estimates are available for all the designs on the grid, the one with maximum estimated utility is chosen as the best design. Table 1 gives pseudo-code for the grid approach, as it is specifically applied to the WSLS and  $\epsilon$ -greedy models.

We found the grid method to be useful for visualizing the utility surface across the entire design space, and for checking the accuracy of more sophisticated methods. It was, however, too computationally inefficient to be useful in general. The obvious problem is the requirement to estimate utilities across a grid. When the grid is coarse, there is no guarantee that a design near the true optimal design will be considered. When the grid is fine, the computational effort quickly grows. Ideally, we want a method that samples good designs, rather than considering the entire design space.

*5.2. MCMC algorithm*

One obvious approach for sampling from high-utility designs is to use MCMC methods.<sup>2</sup> Myung and Pitt (2009) used a sequential Monte Carlo (or particle filter) method developed in statistics by Amzal, Bois, Parent, and Robert (2006). We used another MCMC approach, also developed in statistics (Müller, 1999; Müller, Sansó, & De Iorio, 2004).

The approach we used recasts the optimization problem as a problem of augmented probability simulation, allowing one to sample designs with respect to a density where the optimal design is the mode. This requires a probability model  $h(\cdot)$  to be defined in such a way that the marginal distribution of the design  $\mathbf{d}$  is

proportional to the expected utility  $U(\mathbf{d})$ . This can be done by the definition

$$h(\mathbf{d}, \boldsymbol{\theta}, \mathbf{y}) \propto [p(M_a)\text{BF}_{a/b} + p(M_b)\text{BF}_{b/a}] p(\mathbf{y}_a, \mathbf{y}_b, \boldsymbol{\theta}_a, \boldsymbol{\theta}_b).
 \tag{12}$$

Thus, the task becomes to estimate this density using an MCMC sampling scheme.

Second, the Müller (1999) method uses an annealing procedure to improve the efficiency of sampling. Annealing is based on the product of the utility of  $J$  samples from  $h(\cdot)$ , given by

$$h_J(\cdot) \propto \prod_j [p(M_a)\text{BF}_{a/b} + p(M_b)\text{BF}_{b/a}] p(\mathbf{y}_{aj}, \mathbf{y}_{bj}, \boldsymbol{\theta}_{aj}, \boldsymbol{\theta}_{bj} | \mathbf{d}).
 \tag{13}$$

For a positive integer  $J$ , the marginal distribution of  $h_J(\cdot)$ , obtained after integrating out the outcome variable and model parameters, yields  $U(d)^J$ . Note that the higher the value of  $J$ , the more highly peaked the distribution  $U(d)^J$  will be around its global mode, and therefore the more easily the mode can be found. In our study we fixed  $J$  at 50, which we found produced good results. In theory, one could increase the value of  $J$  gradually in the annealing process, but this requires significant additional computational time, and focuses the optimization process on finding the mode. In our application, where people typically do a sequence of bandit problems in an experiment, we want to generate a distribution of good single-problem designs that can be sampled to create an experiment. Table 2 gives pseudo-code for this MCMC simulation approach, as it is specifically applied to the WSLS and  $\epsilon$ -greedy models.

**6. Demonstrations of design optimization**

In this section, we present four demonstrations of the design optimization method for bandit problems. We consider just the comparison of WSLS and e-WSLS models, because this comparison is the easiest for which to have clear intuitions about good designs, and so it makes the demonstrations more helpful for understanding the method.

The four different demonstrations involve specifying different priors on the model parameters, which are the rates  $\gamma$  for the WSLS method, and  $\gamma_w$  and  $\gamma_l$  for the e-WSLS method. The first demonstration uses a uniform distribution for all of the model parameters, so  $\gamma, \gamma_w, \gamma_l \sim \text{Beta}(1, 1)$ . The results of our design optimization analysis are shown in the top-left panel of Fig. 1. The points correspond to the sampled  $\mathbf{d} = (p_1, p_2)$  designs returned by the MCMC approach. The histograms show the corresponding marginal distributions for both  $p_1$  and  $p_2$ .<sup>3</sup> Note that there is only one optimal design, corresponding to a single point in the main panel, but that the distribution of points shown correspond to “good” designs.

Thus, the top-left panel of Fig. 1 shows that the best designs have both reward rates  $p_1$  and  $p_2$  distributed around 0.5, but negatively correlated, so that if one reward rate is relatively high, the other should be relatively low. The modal good design is at (0.5, 0.5). These results are consistent with intuition, given

<sup>2</sup> Of course, other improvements on the grid method, such as recursive extensions that move from coarse to fine grids, might also work well.

<sup>3</sup> In general, because  $p_1$  and  $p_2$  are exchangeable, their marginal distributions are the same, and the joint distribution is symmetric about the line  $p_1 = p_2$ .

**Table 1**Pseudo-code for the grid approach to design optimization for bandit problems, for the problem of comparing the WSLS model with the  $\epsilon$ -greedy model.

---

```

Choose a grid resolution  $\Delta p$ 
for all designs  $\mathbf{d} = (p_1, p_2)$  over the grid defined by  $\Delta p$  do
  for  $i = 1$  to  $J$  samples do
    sample parameter for WSLS model,  $\gamma \sim \text{Beta}(\alpha, \beta)$ 
    sample data for current experimental design using WSLS  $\mathbf{y}_a \sim p_{\text{WSLS}}(\gamma)$ 
    calculate  $\text{BF}_{a/b}$  for data  $\mathbf{y}_a$ 
    sample parameter for  $\epsilon$ -greedy model,  $\epsilon \sim \text{Beta}(\alpha_\epsilon, \beta_\epsilon)$ 
    sample data for current experimental design using  $\epsilon$ -greedy  $\mathbf{y}_b \sim p_{\epsilon\text{-greedy}}(\epsilon)$ 
    calculate  $\text{BF}_{b/a}$  for data  $\mathbf{y}_b$ 
    calculate  $U_i(\mathbf{d}) = \text{BF}_{a/b} + \text{BF}_{b/a}$ 
  end for
  estimate utility of design as  $\hat{U}(\mathbf{d}) = 1/J \sum_{i=1}^J U_i(\mathbf{d})$ 
end for
Choose the design  $\mathbf{d}^*$  with the maximum utility  $\hat{U}(\mathbf{d})$ 

```

---

**Table 2**Pseudo-code for the MCMC approach to design optimization for bandit problems, for the problem of comparing the WSLS model with the  $\epsilon$ -greedy model.

---

```

choose  $\sigma$  for proposal distribution
set  $t \leftarrow 0$ 
start with a design  $\mathbf{d}^0$ 
repeat
  increment  $t \leftarrow t + 1$ 
  for  $i = 1$  to  $J$  samples do
    sample parameter for WSLS model,  $\gamma \sim \text{Beta}(\alpha, \beta)$ 
    sample data for experimental design  $\mathbf{d}^t$  using WSLS  $\mathbf{y}_a \sim p_{\text{WSLS}}(\gamma)$ 
    calculate  $\text{BF}_{a/b}$  for data  $\mathbf{y}_a$ 
    sample parameter for  $\epsilon$ -greedy model,  $\epsilon \sim \text{Beta}(\alpha_\epsilon, \beta_\epsilon)$ 
    sample data for current experimental design using  $\epsilon$ -greedy  $\mathbf{y}_b \sim p_{\epsilon\text{-greedy}}(\epsilon)$ 
    calculate  $\text{BF}_{b/a}$  for data  $\mathbf{y}_b$ 
    calculate  $w_i(\mathbf{d}^t) = \log(\text{BF}_{a/b} + \text{BF}_{b/a})$ 
  end for
  calculate  $w(\mathbf{d}^t) = \sum_{i=1}^J w_i(\mathbf{d}^t)$ 
  generate different design from proposal distribution,  $\mathbf{d}^{t'} \sim \text{Gaussian}(\mathbf{d}^t, \sigma \mathbf{I})$ 
  for  $i = 1$  to  $J$  samples do
    sample parameter for WSLS model,  $\gamma \sim \text{Beta}(\alpha, \beta)$ 
    sample data for experimental design  $\mathbf{d}^{t'}$  using WSLS  $\mathbf{y}_a \sim p_{\text{WSLS}}(\gamma)$ 
    calculate  $\text{BF}_{a/b}$  for data  $\mathbf{y}_a$ 
    sample parameter for  $\epsilon$ -greedy model,  $\epsilon \sim \text{Beta}(\alpha_\epsilon, \beta_\epsilon)$ 
    sample data for current experimental design using  $\epsilon$ -greedy  $\mathbf{y}_b \sim p_{\epsilon\text{-greedy}}(\epsilon)$ 
    calculate  $\text{BF}_{b/a}$  for data  $\mathbf{y}_b$ 
    calculate  $w_i(\mathbf{d}^{t'}) = \log(\text{BF}_{a/b} + \text{BF}_{b/a})$ 
  end for
  calculate  $w(\mathbf{d}^{t'}) = \sum_{i=1}^J w_i(\mathbf{d}^{t'})$ 
  evaluate acceptance probability  $\pi = \min(1, \exp\{w(\mathbf{d}^t) - w(\mathbf{d}^{t'})\})$ 
  accept  $\mathbf{d}^t \leftarrow \mathbf{d}^{t'}$  with probability  $\pi$ 
until  $T$  samples have been generated
discard the first  $t_{\text{burnin}}$  design samples, and treat the remaining  $\mathbf{d}^t$  samples as draws from the optimal design distribution

```

---

the priors on model parameters correspond to knowing nothing about winning-and-staying or losing-and-shifting behavior. In these circumstances, to compare the WSLS and e-WSLS models – which could differ in one or both of these behaviors – it is best to see an equal number of success and failure trials. This is achieved, as the joint distribution shows, by setting  $p_1 = p_2 = 0.5$ , or by decreasing  $p_2$  for  $p_1 > 0.5$ , or by increasing  $p_2$  for  $p_1 < 0.5$ .

The second demonstration makes some more realistic prior assumptions about model parameters. For the WSLS method, it is assumed that there is likely a high rate of winning-and-staying and losing-and-shifting. For the e-WSLS method, it is assumed that there is likely a quite high rate of winning-and-staying, and a high rate of losing-and-shifting. Quantitatively, the priors used are  $\gamma \sim \text{Beta}(5, 1)$ ,  $\gamma_w \sim \text{Beta}(4, 2)$ , and  $\gamma_l \sim \text{Beta}(5, 1)$ . The design optimization results for this case are shown in the top-right panel of Fig. 1. It is clear that designs with high reward rates for the alternatives are favored. This makes intuitive sense, since the key difference between the models is in how they respond to success (i.e., the WSLS and e-WSLS models have the same prior on losing-and-shifting, but different priors on winning-and-staying). Thus,

distinguishing the models requires observing successes, and this is achieved by an experimental design with high reward rates.

The third demonstration uses parameter priors  $\gamma \sim \text{Beta}(5, 1)$ ,  $\gamma_w \sim \text{Beta}(5, 1)$ , and  $\gamma_l \sim \text{Beta}(4, 2)$ . These assumptions are a slight variant on the preceding demonstration, now making it most likely that losing-and-shifting is the diagnostic behavior. Accordingly, the designs sampled in the bottom-left panel of Fig. 1 correspond to having low reward rates for both alternatives.

The fourth and final demonstration uses  $\gamma \sim \text{Beta}(5, 1)$ ,  $\gamma_w \sim \text{Beta}(1, 1)$ , and  $\gamma_l \sim \text{Beta}(1, 1)$ . This corresponds to having clear prior assumptions about the WSLS model, but being much less certain about the expected behavior of the alternative e-WSLS model. The results are shown in the bottom-right panel of Fig. 1, and they show an interesting nonlinear pattern. Good designs are those that have one moderately high reward rate and one low reward rate, or both reward rates moderately high. It makes intuitive sense that having a high and low reward rate should help discriminate the models, since it is equally possible that winning-and-staying or losing-and-shifting might be diagnostic.

All of these examples are based on three independent sampling chains, within different initial values. Each had a burn-in of 500

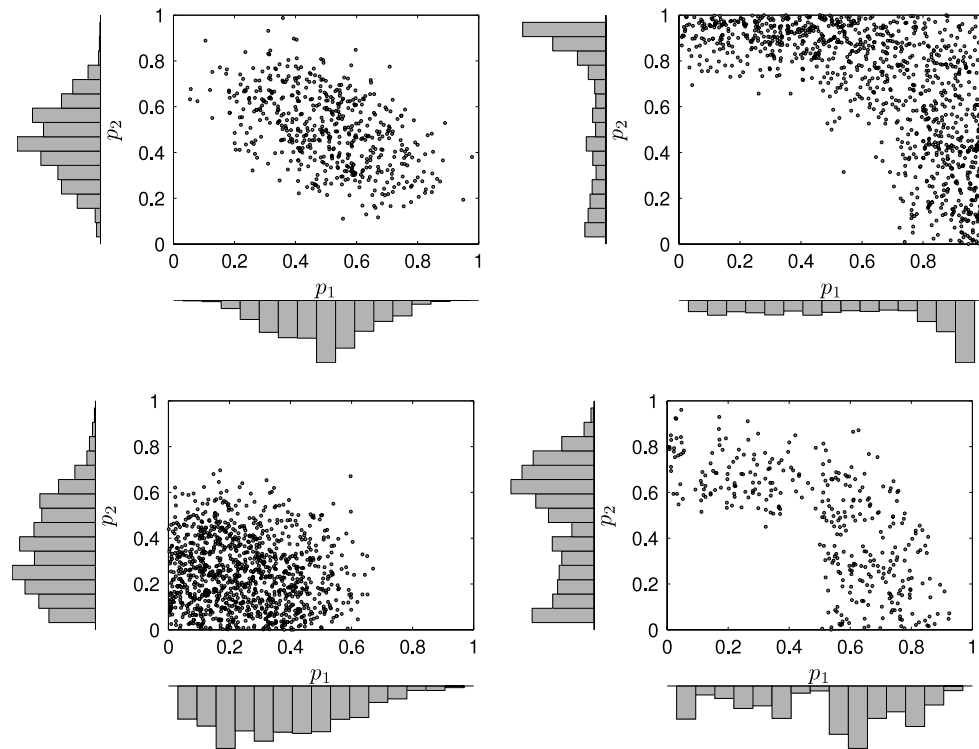


Fig. 1. Samples of “good” designs returned by the MCMC sampler under four different prior distributions of the model parameters.

samples, and we used the remaining 5000 samples. The standard  $\hat{R}$  measure of convergence (Brooks & Gelman, 1997) gave values of 1.0034, 1.0009, 1.0025 and 1.0234 for the four examples, respectively, indicating good convergence.

## 7. Design optimization using empirical priors

In most practical situations involving questions of experimental design optimization, it is possible to do better than simply making assumptions about priors on model parameters. In particular, for most tasks, previous experimental data are available that can help inform the specification of reasonable priors. In this section, we analyze an existing bandit problem data set, using the WSLs, e-WSLS, and  $\epsilon$ -greedy models. This analysis gives us empirically justified priors, which we then use in the design optimization analysis.

### 7.1. Previous data

We used the data presented by Steyvers et al. (2009). A total of 451 participants who completed a series of 20 bandit problems, each having 4 alternatives and 15 trials. The reward rates were drawn independently from a Beta(2, 2) distribution. Steyvers et al. (2009) presented a variety of model-based analyses for these data. Most relevantly, they found clear evidence for individual differences, with the largest proportion of participants using a simple decision-making process compatible with the WSLs method. They did not, however, consider the e-WSLS variant, nor the simple  $\epsilon$ -greedy model, and so it is interesting to ask what the best experimental designs would be to discriminate those models.

### 7.2. Inferring empirical priors

We infer the empirical prior distribution for all of the parameters in all three models using hierarchical Bayesian methods. Fig. 2 shows our three analyses, using the formalism

provided by graphical models, as widely used in statistics and computer science (e.g., Koller, Friedman, Getoor, & Taskar, 2007). A graphical model is a graph with nodes that represents the probabilistic process by which unobserved parameters generate observed data. Details and tutorials that are aimed at cognitive scientists are provided by Lee (2008) and Shiffrin, Lee, Kim, and Wagenmakers (2008). The practical advantage of graphical models is that sophisticated and relatively general-purpose MCMC algorithms exist that can sample from the full joint posterior distribution of the parameters conditional on the observed data.

The left panel in Fig. 2 corresponds to the WSLs model. It shows the  $\gamma$  parameter for each individual being drawn from the group distribution parameterized by  $\alpha$  and  $\beta$ . For each trial on each game,  $\gamma$  combines with the existing history of reward  $\mathbf{r}$  to determine  $\xi = p_{\text{WSLS}}(y_i)$ , which then makes a probabilistic prediction about the observed data. Using the observed data from Steyvers et al. (2009), and standard priors  $\alpha, \beta \sim \text{Poisson}(10)$  then allows the joint posterior distribution  $\alpha, \beta \mid \mathbf{y}$  to be sampled. The expected value of this posterior distribution was at  $\hat{\alpha} = 18.9$  and  $\hat{\beta} = 7.8$ . We tried several other hyper-parameters for the Poisson prior – including Poisson(1), Poisson(5), Poisson(50), and Poisson(100) – and observed that these changes had very little effect on the posterior inference.

Similarly, the middle panel in Fig. 2 corresponds to the e-WSLS model, and it allows inferences for  $\alpha_w, \beta_w, \alpha_l$  and  $\beta_l$ . Their posterior expectations for the same data are  $\hat{\alpha}_w = 2.9, \hat{\beta}_w = 0.7, \hat{\alpha}_l = 2.3$  and  $\hat{\beta}_l = 1.8$ . The right panel in Fig. 2 corresponds to the  $\epsilon$ -greedy model, and it finds  $\hat{\alpha}_\epsilon = 5.8$  and  $\hat{\beta}_\epsilon = 11.3$ .

The empirical priors corresponding to these posterior expectations are shown graphically in Fig. 3. These distributions are used, in the process of optimization the experimental design, to sample parameter values for the models. Thus, for example, the distribution for  $\gamma$  shown in Fig. 3 is Beta(18.9, 7.8). The distributions show that, under the WSLs model, the probability of winning-and-staying or losing-and-shifting lies between about 0.5 and 0.9, with the most likely probabilities just over 0.7. According to the analysis for the e-WSLS model, which separates these probabilities,

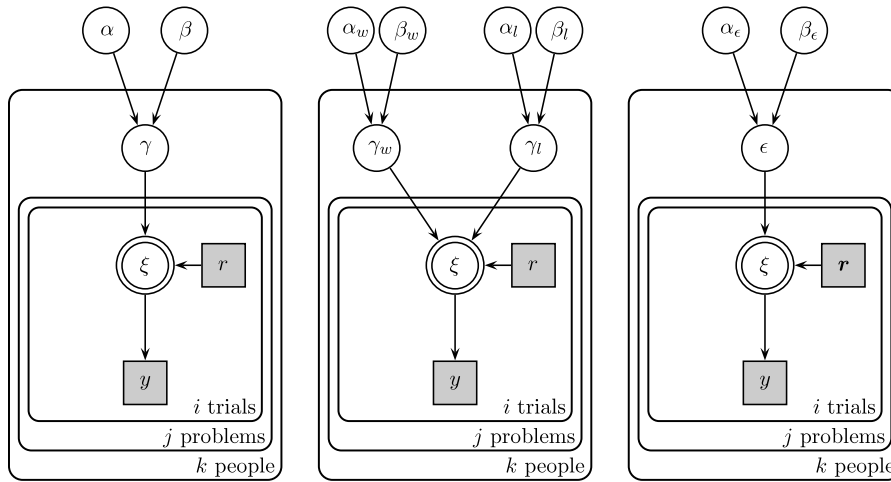


Fig. 2. Hierarchical Bayesian graphical model for inferring empirical priors, for the (left) WSL, (middle) e-WSL and (right)  $\epsilon$ -greedy models.

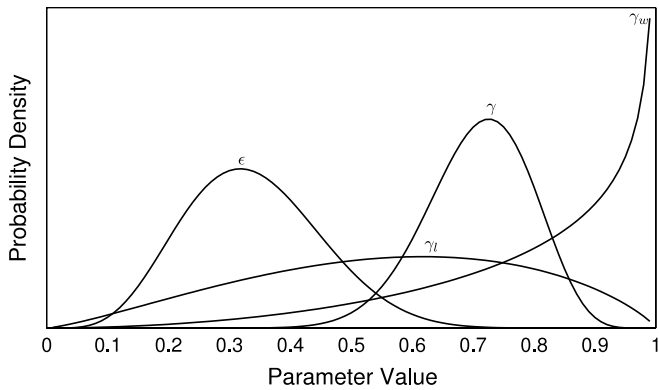


Fig. 3. The empirical priors for model parameters, based on a hierarchical Bayesian analysis of the Steyvers et al. (2009) data.

winning-and-staying, as per  $\gamma_w$ , is more likely than losing-and-shifting, as per  $\gamma_l$ . There is relatively less certainty about both of these probabilities, however, as shown by their broader distributions. The distribution for  $\epsilon$  shows the empirically based probability of exploration choices under the  $\epsilon$ -greedy model, which is anywhere between 0.1 and 0.6, with most likely values around 0.3.

### 7.3. Optimal designs

#### 7.3.1. WSL versus e-WSL

The left panel of Fig. 4 shows the results of our design optimization approach, comparing the WSL and e-WSL models with the empirical priors for model parameters. We used just a single chain of 20,000 samples after a burn-in of 1000 samples, and checked the convergence using the method proposed by Geweke (1992). The clear result is that the best designs have one alternative with a high reward rate, and the other alternative with a low reward rate. This makes intuitive sense, looking at the priors in Fig. 3, which show that either or both  $\gamma_w$  and  $\gamma_l$  might be different from  $\gamma$ . Thus, it is possible that either different rates of winning-and-staying or different rates of losing-and-shifting might be the key to distinguishing the models. Good designs need to accommodate both possibilities, and so generate both successes and failures by using both high and low rewarding alternatives. The designs sampled in the left panel of Fig. 4 achieve this.

A more subtle observation is that, while designs are necessarily symmetric about the identity line, they are asymmetric about the other diagonal. Designs from the bottom-left corner are preferred

over those in the top-right corner, indicating that large values for both reward rates is the worst type of design for telling apart the two models. This is a reflection of the fact that the prior for  $\gamma_w$  is slightly left skewed compared to the prior for  $\gamma_l$ . Of course, this subtlety is relatively unimportant compared to the main result of choosing one high and one low reward rate. But it does highlight the way in which the formal quantitative approach naturally incorporates all of the information about the models in searching for good experimental designs.

#### 7.3.2. WSL versus $\epsilon$ -greedy

The right panel of Fig. 4 shows the results of design optimization for comparing the WSL and  $\epsilon$ -greedy models. Again, we used a single chain of 20,000 samples after a burn-in of 1000 samples, and checked convergence using the method proposed by Geweke (1992). The best designs are found to be those with comparatively small values for both reward rates, centered around about 0.3. This means that a good bandit experiment for distinguishing these models is one that produces a relatively large number of failures.

Intuitively, it may not be obvious why this is the case, but some analysis gives an insight that makes the basic result interpretable. The key observation involves a simple inequality, comparing the observed success rate before and after a winning-and-staying decision. This is given by

$$\frac{(s_{ij} + 1) + 1}{(s_{ij} + f_{ij} + 1) + 1} > \frac{s_{ij} + 1}{s_{ij} + f_{ij} + 1}, \tag{14}$$

and shows that continuing to choose a rewarding alternative will increase the observed success rate, and therefore continue to encourage the same decision from the  $\epsilon$ -greedy model. The same is obviously true of the WSL model, which has a probability of winning-and-staying given by  $\gamma$ . Accordingly, to distinguish the models, it is more informative to focus on decisions following failures. This is what the designs found to be the best ones in the right panel of Fig. 4 achieve.

### 7.4. Performance of optimal designs

It is natural to ask whether, and to what extent, the optimal designs just derived improve upon the original design used by Steyvers et al. (2009). Fig. 5 shows the results of a series of analyses that address this question. In these analyses, one model is used to generate decision-making data for a sequence of bandit problems following either the optimal design, or the original design. The data are generated for 10 simulated participants, using model

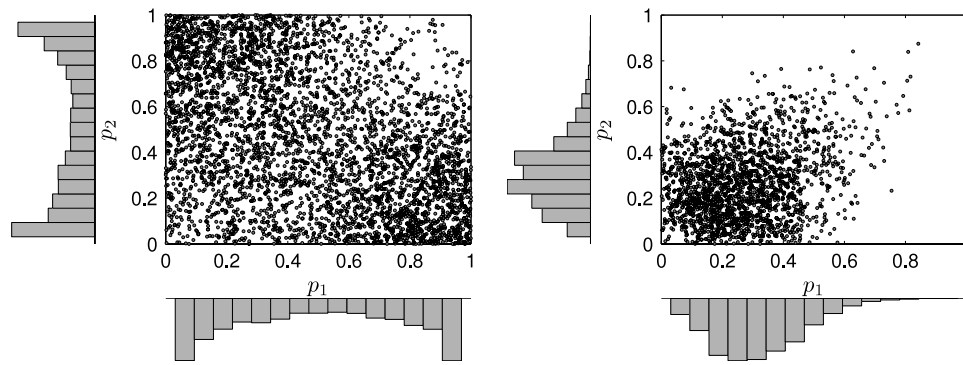


Fig. 4. Samples of “good” designs for discriminating between the (left panel) WSLs and e-WSLs models, and (right panel) the WSLs and  $\epsilon$ -greedy models.

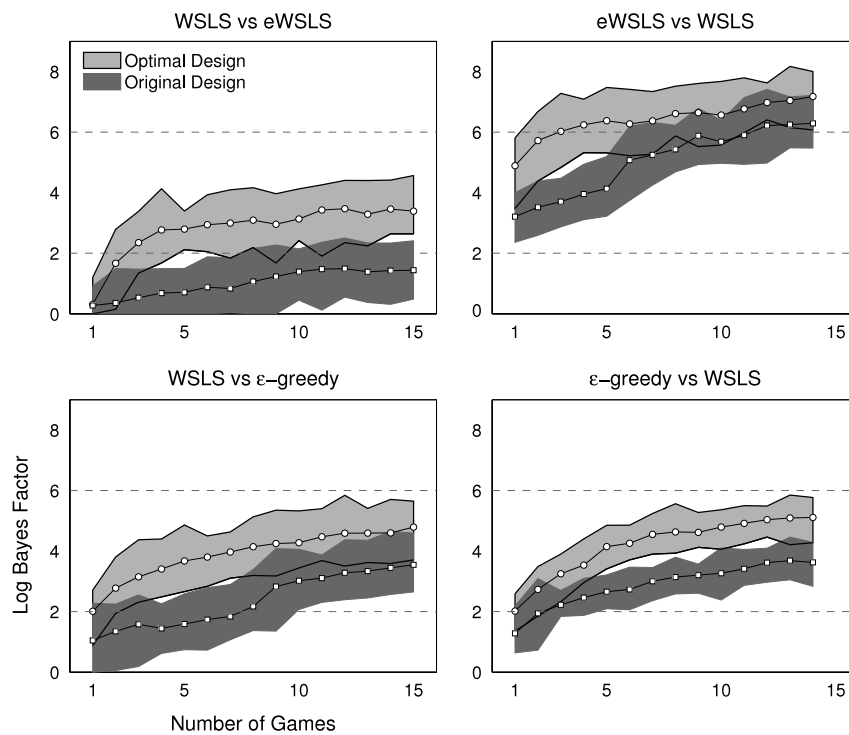


Fig. 5. Performance of optimal experiments relative to the original Beta(2, 2) design used by Steyvers et al. (2009).

parameters drawn from the empirical priors, and experiments with different numbers of problems are considered. Each experimental design is either sampled from the posterior distributions shown in Fig. 4, or from the Beta(2, 2) originally used.

Under both designs, the log Bayes factor in favor of the correct generating model is used to measure the effectiveness of the experimental design. Fig. 5 shows the mean (by lines and markers) and the range (by bounded shaded regions) for the log Bayes factors, in four different analyses. These consider both the WSLs versus e-WSLs and WSLs versus  $\epsilon$ -greedy model comparisons, and consider both assumptions about which model generated the data. The means, minima and maxima shown are based on 100 independent runs of each simulated experiment. It is clear from Fig. 5 that the optimal design almost always outperforms the original design on average. Even more compellingly, the worst observed optimal design is always better than the mean original design, and is often better than the best-performed original design.

Some additional conclusions can be drawn from Fig. 5 if the magnitude of evidence is interpreted in terms of standard calibrations of the log Bayes factor scale (e.g., Kass & Raftery, 1995). Fig. 5 encourages this interpretation, showing broken lines for log Bayes factors of 2 and 6, corresponding to the thresholds

for “positive” and “strong” evidence, respectively. This absolute, rather than relative, assessment of evidence suggests that optimal designs are especially beneficial when a simple model is generating data, and is being compared to a more complicated model within which it is nested. The WSLs and e-WSLs models provide such an example, and the top two panels of Fig. 5 show that both standard and optimal designs can find evidence for the e-WSLs model for a small number of problems, but only the optimal design even achieves “positive” evidence when the WSLs model generates the data. It is generally true that, from sparse and noisy data, it is hard to find conclusive evidence for models that are special cases of more general accounts, and our evaluations show how design optimization can potentially help tackle that challenge in finding simple models of cognition.

Of course, this evaluation of the optimal design depends on the reasonableness of the models used to generate data as accounts of human decision-making. While the results in Fig. 5 are comforting, and show that the statistical methods we have employed to find optimal designs are working well, they do not guarantee empirical success. The ultimate test of our methods is to run experiments on human subjects using the optimal designs. This is an important final verification step we leave for future research.

## 8. Discussion

### 8.1. Contributions of current work

Our application of the Myung and Pitt (2009) framework for design optimization makes a number of new contributions. Most obviously, this is the first time that the framework has been applied to bandit problems. Bandit problems are among the most widely used sequential decision-making problems in cognitive science. They are studied in fields as diverse as neuroscience (e.g., Cohen et al., 2007; Daw et al., 2006), reinforcement and machine learning (e.g., Kaelbling et al., 1996; Sutton & Barto, 1998), and statistics, game theory and computer science (e.g., Berry & Fristedt, 1985; Gittins, 1979). Given their widespread importance as an empirical task, the development of a method that is able to optimize important aspects of their design is a worthwhile contribution.

One aspect of the design optimization framework that we have emphasized is the ability to generate a number of different good designs, rather than focus too narrowly on a single best design. All of our results take the form of many sampled designs with high utility for discriminating the relevant models. Given that we are only optimizing one aspect of bandit problems – the underlying reward rates for the alternatives – it seems unnecessary, and perhaps even inappropriate, to target a single optimal design. Instead, we think that it makes sense to consider a set of candidate good designs, giving some variability and robustness to the experiments done, and acknowledging that other relevant aspects of the design are not being taken into account.

The use of hierarchical Bayesian methods to infer empirical priors is a third useful contribution. In practice, an approach like this seems needed to use what an experimenter knows about models in developing optimal designs. Rarely will model parameters be completely unknown, and often they will not be able to be specified by theoretical argument. Most intuitions and expectations about model parameters in the cognitive sciences come from previous fitting of the models to relevant data. Our empirical prior approach formalizes this information in a principled way, and allows it to be inserted directly into the design optimization process.

### 8.2. Future work

There are four, successively more ambitious, lines of future work we would like to pursue. The first, and most obvious, extension is to consider additional elements of bandit problems as part of the formal design optimization process. In this paper, we concentrated on the reward probabilities for the alternatives. It would also be possible, however, to consider the number of alternatives, and the number of trials in a problem, as variables to be optimized jointly with the reward probabilities. One particularly interesting problem would be to assume that participants are available for a fixed number of trials, and optimize how those trials are divided into problems. There is a natural trade-off between the number of problems and the number of trials per problem. It is not immediately clear whether choosing to administer a few problems, each with many trials, or many problems, each with a few trials, or something in between, is optimal. An even more general formulation would allow for each problem to have a potentially different number of trials, as long as the total number was constant. All of these questions, in principle, could be addressed within the statistical framework for design optimization proposed by Myung and Pitt (2009), and are worth pursuing.

The second extension of the current work, which we mentioned earlier, involves solving a more general design optimization for bandit problems. Our current approach focuses on choosing

good reward rates for the two alternatives within each bandit problem. It would also be useful, however, to solve the problem of optimizing the environment from which these reward rates are drawn. The goal would be to find a parameterized distribution for representing environments – the obvious one is given by the Beta family – and then solving the quantitative problem of finding the best environment for generating a sequence of individual bandit problems that best discriminate two models. Conceptually, this is a straightforward hierarchical extension of our current approach. Computationally, it involves an extra level of integration, and so may prove challenging.

A third extension involves considering alternative forms of the local utility function  $u(\mathbf{d}, \theta, \mathbf{y})$ . In this study, we used the Bayes factor, as a theoretically justifiable and sensible model selection criterion. But Myung and Pitt (2009) have also considered some alternative forms, such as the simple sum-squared error, and metrics from information theory (Cavagnaro, Myung, Pitt, & Kujala, 2010). Pursuing ideas from information theory – including Shannon information, Kullback–Leibler divergence, Hellinger distance, and mutual information – to improve the design of human experimentation seems especially worthwhile, given their rich history of application to statistical design optimization problems (e.g., Bingham & Chipman, 2002; Box & Hill, 1967; Cover & Thomas, 1991; Lindley, 1956; Mackay, 1992; Paninski, 2005).

A final direction for future research involves considering adaptive sequential design optimization. Bandit problem experiments usually consist of a long sequence of individual problems. An improved ability to discriminate models should result from a capability to adapt or adjust the design of the individual problems not yet completed. Myung and Pitt (2009) present a natural and powerful extension of the current design optimization approach to allow adaptive design optimization problems to be considered. Once again, there are significant additional demands, in the form of fast computation requirements, needed to apply these ideas to bandit problems.

### 8.3. Conclusion

A common claim is that scientists seeking quantitative explanations are now drowning in data (e.g., Han & Kamber, 2006). For many endeavors, this may be true. But it remains the case that it is very expensive and time-consuming to collect experimental data, in controlled laboratory settings, that measure human cognition. For this reason, design optimization is an important capability for experimentation in the cognitive sciences. Being able to design experiments that have the greatest ability to distinguish competing models is an important way to further theoretical progress in understanding human cognition.

In this paper, we have taken a recently developed framework for design optimization, and begun to apply it to a class of well-studied sequential decision-making tasks known as bandit problems. Our results show that the method is able to generate a set of good designs for both nested and non-nested model comparisons, under a range of different assumptions about the parameters of the models. We have also shown that the approach is naturally applied to empirical priors for these parameters, as inferred from data. While there remain a series of extensions and challenges to realize the full potential of the approach, we have already reached a stage where quantitative analysis is producing sound guidance for experimental design, and helping to measure human cognition effectively and efficiently.

### Acknowledgments

We thank Jay Myung, and members of the Memory and Decision-Making Laboratory at UC Irvine for very helpful discussions. We also thank two anonymous reviews for useful comments. This work is supported by an award from the Air Force Office of Scientific Research (FA9550-07-1-0082).

## 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. C., & Donev, A. N. (1992). *Optimum experimental designs*. Oxford University Press.
- Berry, D. A., & Fristedt, B. (1985). *Bandit problems: sequential allocation of experiments*. London: Chapman & Hall.
- Bingham, D., & Chipman, H. (2002). Optimal design for model selection. *Technical report 388*. University of Michigan.
- Box, G. E. B., & Hill, W. J. (1967). Discrimination among mechanistic models. *Technometrics*, 9, 5–21.
- Brand, H., Sakoda, J. M., & Woods, P. J. (1957). Effects of a random versus pattern reinforcement instructional set in a contingent partial reinforcement situation. *Psychological Reports*, 3, 473–479.
- Brand, H., Wood, P. J., & Sakoda, J. M. (1956). Anticipation of reward as a function of partial reinforcement. *Journal of Experimental Psychology*, 52(1), 18–22.
- Brooks, S. P., & Gelman, A. (1997). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434–455.
- 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, 831–886.
- Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: a review. *Statistical Science*, 10, 274–304.
- Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? Exploration versus exploitation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362, 933–942.
- Cover, T., & Thomas, J. (1991). *Elements of information theory*. Wiley.
- Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, 441, 876–879.
- Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics*, 4, 169–193.
- Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, 41, 148–177.
- Han, J., & Kamber, M. (2006). *Data mining* (second ed.). New York: Morgan Kaufman.
- Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, 4, 237–285.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 377–395.
- Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society: Series B*, 21, 272–319.
- Koller, D., Friedman, N., Getoor, L., & Taskar, B. (2007). Graphical models in a nutshell. In L. Getoor, & B. Taskar (Eds.), *Introduction to statistical relational learning*. Cambridge, MA: MIT Press.
- Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, 15(1), 1–15.
- Lee, M. D., Zhang, S., Munro, M. N., & Steyvers, M. (in press). Psychological models of human and optimal performance on bandit problems. *Cognitive Systems Research*.
- Lindley, D. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27, 986–1005.
- Mackay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Computation*, 4, 590–604.
- Macready, W. G., & Wolpert, D. H. (1998). Bandit problems and the exploration/exploitation tradeoff. *IEEE Transactions on Evolutionary Computation*, 2(1), 2–22.
- Müller, P. (1999). Simulation-based optimal design. *Bayesian Statistics*, 6, 459–474.
- Müller, P., Sansó, 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., & Pitt, M. (2009). Optimal experimental design for model discrimination. *Psychological Review*, 116, 499–518.
- Paninski, L. (2005). Asymptotic theory of information-theoretic experimental design. *Neural Computation*, 17, 1480–1507.
- Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58, 527–535.
- Shiffrin, R. M., Lee, M. D., Kim, W.-J., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, 32, 1248–1284.
- Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, 53, 168–179.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. Cambridge, MA: The MIT Press.