Inference and Learning for Active Sensing, Experimental Design and Control

Hendrik Kueck, Matt Hoffman, Arnaud Doucet, and Nando de Freitas

Department of Computer Science, UBC, Canada
{kueck,hoffmanm,arnaud,nando}@cs.ubc.ca

Abstract. In this paper we argue that maximum expected utility is a suitable framework for modeling a broad range of decision problems arising in pattern recognition and related fields. Examples include, among others, gaze planning and other active vision problems, active learning, sensor and actuator placement and coordination, intelligent human-computer interfaces, and optimal control. Following this remark, we present a common inference and learning framework for attacking these problems. We demonstrate this approach on three examples: (i) active sensing with nonlinear, non-Gaussian, continuous models, (ii) optimal experimental design to discriminate among competing scientific models, and (iii) nonlinear optimal control.

1 The Principle of Maximum Expected Utility

Broadly speaking, *utility* reflects the preferences of an agent. That is, if outcome o_1 is preferred to o_2 (*i.e.* $o_1 \succ o_2$), we say that o_1 has higher utility than o_2. More formally, let $o_1 \succeq o_2$ denote weak preference, $o_1 \succ o_2$ denote strong preference and $o_1 \sim o_2$ denote indifference. Define a *lottery* to be a random set of outcomes with corresponding probabilities: $l = [(o_1, p_1), (o_2, p_2), \ldots, (o_k, p_k)]$, where the probabilities satisfy $p_i \geq 0$ and $\sum_i^k p_i = 1$ as usual. Now consider the following axioms:

1. **Completeness:** $\forall o_1, o_2$, we have $o_1 \succ o_2$, $o_2 \succ o_1$ or $o_1 \sim o_2$.
2. **Transitivity:** If $o_1 \succeq o_2$ and $o_2 \succeq o_3$, then $o_1 \succeq o_3$.
3. **Substitutability:** If $o_1 \sim o_2$, then for all sequences of outcomes $o_3, \ldots, o_k$ and sets of probabilities $p, p_3, \ldots, p_k$ for which $p + \sum_{i=3}^k p_i = 1$, we have $[(o_1, p), (o_3, p_3), \ldots, (o_k, p_k)] \sim [(o_2, p), (o_3, p), \ldots, (o_k, p_k)]$.
4. **Decomposability:** Let $P_l(o_i)$ be the probability that outcome o_i is selected by lottery l. If for all o_i: $P_{l_1}(o_i) = P_{l_2}(o_i)$, then $l_1 \sim l_2$.
5. **Monotonicity:** If $o_1 \succ o_2$ and $p > q$, then $[(o_1, p), (o_2, 1-p)] \succ [(o_1, q), (o_2, 1-q)]$.
6. **Continuity:** If $o_1 \succ o_2$ and $o_2 \succ o_3$ then $\exists p \in [0, 1]$ such that $o_2 \sim [(o_1, p), (o_3, 1-p)]$.

Using these axioms, von Neumann and Morgenstern [16] proved the following fundamental result showing the existence of utility:

Theorem 1. *If a preference relation $\succeq$ satisfies axioms 1 to 6 above, then there exists a function u mapping outcomes to the real line with the properties that:*

1. $u(o_1) \geq u(o_2)$ *iff* $o_1 \succeq o_2$
2. $u([(o_1, p_1), (o_2, p_2), \ldots, (o_k, p_k)]) = \sum_{i=1}^{k} u(o_i)p_i$.

Expected utility does therefore arise as a rational consequence of fairly unassailable axioms. An agent expecting to behave optimally must maximize its expected utility; see [14] for a more comprehensive treatment.

Following this result, it is reasonable that our goal in decision making under uncertainty be one of finding an optimal strategy π^* that maximizes the *expected utility* $U(\pi)$ of the agent:

$$\pi^* = \arg\max_{\pi} U(\pi), \text{ with } U(\pi) = \int u(x, \pi)p(x|\pi)\, dx \tag{1}$$

Note, we are integrating over all possible unknown states x (that is we are considering all *possible worlds* and weighting them according to how likely we deem them). $U(\pi)$ then describes how useful we expect the outcomes of adopting a policy to be based on the current beliefs encoded in $p(x|\pi)$. In the MEU view, we are assuming that we can solve the joint maximization and integration problem. In general we can't do this and are forced to make approximations. Theories that take into account these approximations have appeared under the umbrella of *bounded rationality*. The application domain of the MEU principle is very broad. The principle can be used to guide the placement and control of a network of sensors in a changing environment, to decide what data must be gathered to improve a classifier, to plan a sequence of gazes to dynamically understand a visual scene, to plan the trajectory of a robot so as to minimize resources, and so on. In this paper, we will present an inference and learning approach for solving MEU problems. This single approach will suffice to solve difficult nonlinear, non-Gaussian problems arising in myopic and sequential (multi-stage) decision making. We discuss these two problems in the following subsections.

1.1 Myopic Decision Making

We will illustrate myopic decision making in the context of Bayesian experimental design. We don't lose much generality doing this because Bayesian experimental design is a broad field of study that is applicable to many problems, including active vision, sensor network management and active learning. We assume that we have a measurement model $p(y|\theta, \pi)$ of experimental outcomes $y \in \mathcal{Y}$ given a design π as well as a prior $p(\theta)$ on the model parameters $\theta \in \Theta$. The prior could be based on expert knowledge or previous experiments. We recover the general model presented in the previous section by noting that $x = \{y, \theta\}$.

The model parameters as well as future observations are unknown so we have to integrate out over their probability distributions. (In the simple case of learning a regression function or a classifier, which is widely studied in active

learning, π would correspond to the predictors (inputs) and y to the corresponding covariates (outputs).) The general goal is then to choose the optimal design $\pi^* \in \mathbb{R}^p$, which maximizes the expected utility

$$U(\pi) = \iint u(y, \pi, \theta)\, p(\theta) p(y|\theta, \pi)\, dy\, d\theta \tag{2}$$

with respect to some measure of utility $u(y, \pi, \theta)$. When the model parameters are the objects of interest, which will be the case in Section 3.1, the negative posterior entropy is commonly chosen as the utility function. That is, one aims to maximize

$$u(y, \pi, \theta) = \int p(\theta'|y, \pi) \log p(\theta'|y, \pi) d\theta'.$$

This measures how concentrated the belief distribution over the parameters is after conducting an experiment with design π and observing outcome y. Note the difference between θ and θ' here. θ represents the true model parameters of the possible world under consideration in which the hypothetical experiment is conducted. The outcome y is generated according to $p(y|\theta, \pi)$. $p(\theta'|y, \pi)$ then is the belief distribution over the model parameters that we would have after observing y. Note that this particular utility function does not actually depend on θ. It merely measures how peaked the posterior belief $p(\theta'|y, \pi)$ is, not how close it is to the 'real' θ. In the case of a linear-Gaussian model, this entropy based utility function is referred to as Bayesian D-optimality [1].

In general, the choice of utility function should reflect the objective of the experiment as well as costs and risks related to the experiment as well as possible. For example in a medical trial the goal might be to gain the maximum amount of information about the effects of a new drug while at the same time keeping the risk of people dying (or suffering severe side effects) to a minimum and also minimizing the monetary cost of the trial. The utility function would then consist of several terms representing these (possibly conflicting) objectives. Another interesting choice of utility is given in Section 3.2 of this paper.

1.2 Sequential Decision Making

In the previous section, we only considered integrating over the outcome at the next step of decision making. In general, we would like to plan several steps ahead. This type of planning can be modeled with a Markov decision process (MDP); illustrated in Figure 1. Here we are integrating over an infinite sequence $z_{0:\infty} = \{z_0, z_1, \dots\}$, where each $z_n = (s_n, a_n, r_n)$ represents a tuple of state, action, and reward at time n.

The design parameter π in this setting determines a policy for choosing an action a_n based on the current state s_n according to $p(a_n|s_n, \pi)$. Given a policy, the states and rewards of the Markov process evolve according to an initial-state model: $s_0 \sim \mu(s_0)$, a transition model: $s_{n+1} \sim p(s_{n+1}|s_n, a_n)$, and rewards: $r_n \sim p(r_n|s_n, a_n)$. We can then define the utility function for a single trajectory as its discounted reward, $u(z_{0:\infty}) = \sum_{n=0}^{\infty} \gamma^n r_n$. Intuitively the discount factor

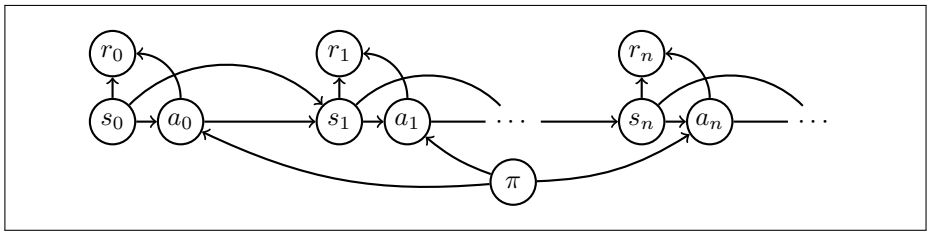


Fig. 1. A graphical model depicting the random variables for an MDP.

$\gamma \in [0, 1]$ emphasizes immediate rewards over more distant future rewards. Our goal is then to find the policy π^* which maximizes this expected utility:

$$U(\pi) = \int \left[\sum_{n=0}^{\infty} \gamma^n \, r_n \right] p(z_{0:\infty}|\pi) \, dz_{0:\infty}.$$

This integration problem is made difficult, however, by the fact that $z_{0:\infty}$ is an infinite dimensional object.

Following [15], it is possible to move the expectation *inside the summation* and rewrite the expected utility as

$$U(\pi) = (1-\gamma)^{-1} \sum_{n=0}^{\infty} \int r_n \, p(n, z_{0:n}|\pi) \, dz_{0:n}$$

where this expectation is taken with respect to the trans-dimensional distribution

$$p(n, z_{0:n}|\pi) = (1-\gamma) \, \gamma^n \, p(z_0|\pi) \prod_{j=1}^{n} p(z_j|z_{j-1}, \pi). \tag{3}$$

With this formulation we are still integrating[1] over an infinite-dimensional quantity, but now we have broken this into a joint integral where one of our random variables is the dimensionality n. Note, however, that the general form of Equation (1) can be recovered by letting $x = (n, z_{0:n})$ and defining our utility as the final reward in each finite-length trajectory $u(n, z_{0:n}) = r_n$.

It is also possible to further generalize this model to take into account situations where the states s_n are not visible, and instead some observations $y_n \sim p(y_n|s_n)$ are given. This model, known as a partially observable MDP (POMDP), is a much more difficult problem and is not one currently tackled by this framework.

2 A Common Solution Framework

The joint optimization and nested integration problem in equation (1) is computationally challenging. For this reason, much of the research in experimental

[1] We note that the sum over n is really an integral over the natural numbers.

design and control has focused on the simple linear-Gaussian models, for which closed form solutions exist [2, 3]. Recently, however, there has been a flurry of work applying inference and learning techniques to more difficult nonlinear problems. In the context of control this seems to have originated with [4], although only immediate rewards are considered, thus making it perhaps more applicable to the setting of myopic experimental design. In this section we will focus on a promising sample-based technique originating in the experimental design literature [10].

One possible strategy for solving these problems involves sampling policies π and hidden states x from the artificial target distribution

$$h(\pi, x) \propto u(x, \pi)\, p(x|\pi). \tag{4}$$

We can see by marginalizing over x that this produces a distribution

$$h(\pi) \propto \int u(x, \pi)\, p(x|\pi)\, dx = U(\pi)$$

proportional to the expected utility. In order to sample from this distribution we must assume that $u(x, \pi)$ is positive and finite but this is easy to ensure so long as the utility is bounded. We may also be required to introduce a prior $p(\pi)$ to ensure that this distribution is well defined, but typically a uniform distribution over some bounded region is sufficient.

If $U(\pi)$ happens to have a strongly dominant and highly peaked mode around the global maximum π^*, we can justify sampling from (4) and deriving a point estimate by averaging these samples. In the context of sequential decision making this is the approach taken in [6]. However, in general the assumption of such a favorable $U(\pi)$ is unrealistic, and applying this strategy to a multimodal or fairly flat utility landscape will yield poor estimates.

Other strategies involve discretizing the policy space $\mathbb{R}^p$ and approximating the integrals with direct Monte Carlo methods. However, these approaches are expensive and inadequate for high dimensional spaces. To eliminate the need for discretization, Müller et al. [10] proposed a Markov chain Monte Carlo annealing technique for simultaneous maximization and integration. They define the following artificial target distribution

$$h_J(\pi, x_{1:J}) \propto \prod_{j=1}^{J} u(\pi, x_j) p(x_j|\pi)$$

Marginalizing this distribution over the unknown outcomes and parameters gives

$$\begin{aligned} h_J(\pi) &= \iint h_J(\pi, x_{1:J})\, dx_{1:J} = \iint \prod_{j=1}^{J} u(\pi, x_j) p(x_j|\pi)\, dx_{1:J} \\ &= U(\pi) \iint \prod_{j=2}^{J} u(\pi, x_j) p(x_j|\pi)\, dx_{2:J} = \prod_{j=1}^{J} U(\pi) = U^J(\pi). \end{aligned}$$

For large exponents J, the probability mass of this distribution will concentrate on the global maximum for π^*. However, because the modes of this distribution will typically be narrow and widely separated for large J, sampling from this distribution using Markov chain Monte Carlo techniques directly is difficult. We must therefore take a simulated annealing approach in which we start sampling from $U^J(\pi)$ for $J = 1$ and slowly increase this exponent over time according to some annealing schedule. Increasing J slowly enough allows the chain to efficiently explore the whole parameter space before becoming more constrained to the major modes.

In order to apply this technique to problems of sequential decision making we must use reversible jump MCMC [5] to sample from trans-dimensional distributions such as (3); see [6, 7] for more details on applying these ideas to control problems. In the myopic experimental design setting we advise the use of sequential Monte Carlo samplers [8] (see also Figure 2).

3 Demonstrations

3.1 Active Sensing Example

As a first example we study a synthetic problem that, despite its apparent simplicity, exhibits complex multi-modality. In particular, we address the problem of inferring the parameters of a sine wave. This non-linear experimental design example is motivated by the problem of scheduling expensive astronomical observations [9]. The sine wave is parameterized by $\theta = \{A, \omega, \rho\}$ where A is the amplitude, ω the frequency and ρ the phase

$$y = f(\pi; A, \omega, \rho) + \epsilon = A \sin(\pi\,\omega + \rho) + \epsilon.$$

Here ϵ denotes the normally distributed measurement noise. The objective is to find the optimal location π^* along the x-axis at which to make the next noisy y measurement in order to maximally reduce uncertainty about the parameters θ. In the example shown in Figure 2, two prior observations have already been made and the design problem consists of choosing the optimal third measurement. That is, the prior belief $p(\theta)$ here is actually the posterior parameter distribution after these first two measurements. Figure 2(a) shows some sine waves corresponding to samples from this $p(\theta)$, visualizing the belief about possible sine waves.

The corresponding expected utility function $U(\pi)$ shown in Figure 2(b) is proportional to the uncertainty about y at a given point π along the x-axis. This function is highly multi-modal, with most of the modes having similar magnitudes. However exponentiating this function to a power of 50 concentrates most of the probability mass on the major mode, so that samples distributed proportional to this function provide a good basis for estimating π^*. When using a single MCMC chain using the approach of [10], the chain often gets trapped in the minor modes, as can be seen in Figure 2(c). The interaction between multiple particles in the SMC samplers algorithm we proposed in [8] helps avoid this and yields a much better estimate as shown in Figure 2(d).

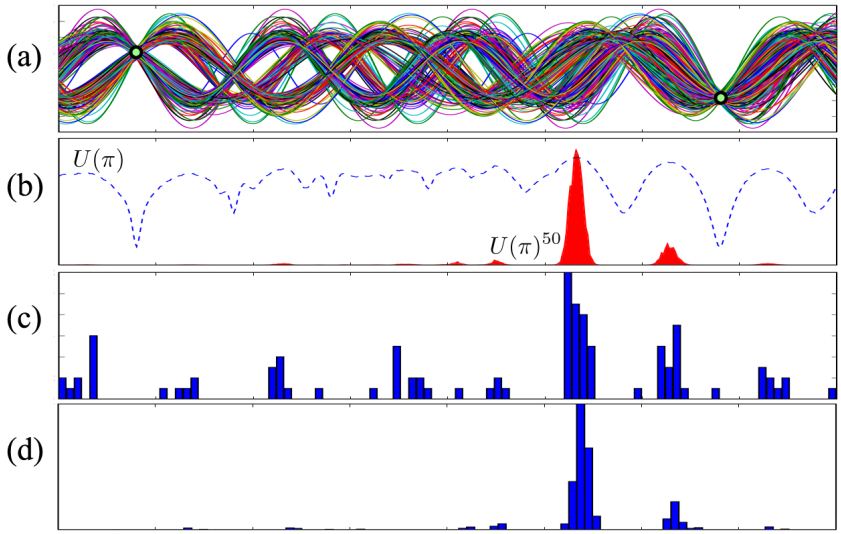


Fig. 2. Plot (a) shows sine waves visualizing the belief after 2 initial observations. The corresponding expected utility $U(\pi)$ for the maximum entropy criterion from Equation 1.1 is shown as a dashed blue line in (b) while $U(\pi)^{50}$ is displayed in solid red. Plot (c) presents a histogram of the final samples of 100 independent MCMC chains using the approach of [10] when annealing to $U(\pi)^{50}$, while (d) shows the result achieved using 100 interacting particles using our SMC samplers algorithm proposed in [8].

3.2 Experimental Design to Choose Among Scientific Models

Often in science and economics, several mathematical models are proposed for describing a common phenomenon. It is therefore important to have a sound mechanism for gathering evidence so as to validate the various model options and assess their merits. For example, in mathematical psychology — a branch of psychology concerned with the mathematical modeling of various aspects of human cognitive performance — researchers have proposed the use of automatic experimental design techniques to find the most plausible model from a set of model alternatives [11]. In this domain, the goal is to choose an optimal experiment for maximally discriminating among several given models. Since such experiments tend to be very costly and work intensive, it is crucial to carefully design them to gain the most information from them and make the most efficient use of the resources involved.

There exists a large body of research in psychology on how we remember and/or forget things. Typically this research involves experiments in which subjects initially memorize some material (such as word lists) and are subsequently tested for recall after several different time intervals. A survey of many such studies is presented in [13]. The percentage of recalled items monotonically decreases over time according to a function of roughly logarithmic shape. The question

that researchers in mathematical psychology are interested in is exactly which mathematical function best describes human retention performance.

Following [11], we are concerned in this example with differentiating among two previously proposed models. More specifically, we assume that a trial in which a single item has to be recalled after time t is repeated n times. The probability that a subject will remember k out of the possible n objects is given by the Binomial distribution: $p(k|n,\rho) = \binom{n}{k}\rho^k(1-\rho)^{n-k}$.

The two models considered for predicting the probability of retention ρ after elapsed time t have 2 parameters $\phi = \{a, b\}$. The first is an exponential model, $\rho_{M_e}(t,\phi) = a\,e^{-bt}$, while the second one is a power model $\rho_{M_p}(t,\phi) = a\,(t+1)^{-b}$.

In our example, the goal will be to compute the optimal 2 point design $\pi \equiv t_{1:2} = \{t_1, t_2\}$. That is, we need to choose the two time lags after which the subject's retention will be tested. We are seeking the design that will allow us to best distinguish between the two models. We adopt a Bayesian model comparison criterion for our utility function. The utility of an experiment with design $t_{1:2}$ and experiment outcomes $k_{1:2}$ is given by the posterior marginal probability of the true model M which generated the data

$$u(t_{1:2}, k_{1:2}, M) = p(M|t_{1:2}, k_{1:2}) \propto \int \prod_{j=1}^{2} p(k_j|n, \rho_M(t_j,\phi))\, p(\phi|M)\, d\phi$$

where $p(\phi|M)$ is the prior on the parameters for the model under consideration; In our experiments we used an empirical prior based on data from a previously conducted study [12]. Intuitively, an experiment is of high utility if we strongly favor the true model after observing the experiment outcome.

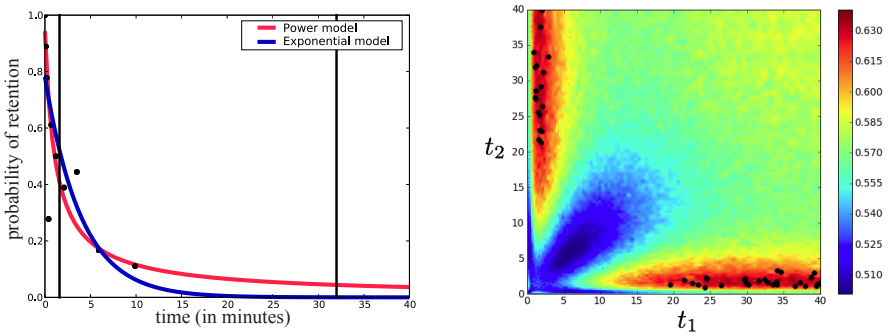
To compute the expected utility according to Equation (2) we need to integrate over the unknown model parameters θ, wich in this case consist of $\{M, \phi\}$ as well as over the possible experiment outcomes $y \equiv k_{1:2}$. The expected utility of an experiment design $t_{1:2}$ is then

$$U(t_{1:2}) = \frac{1}{2} \sum_{M \in \{M_e, M_p\}} \int u\left(t_{1:2}, k_{1:2}, M\right) p\left(k_{1:2}|n, \rho_M(t_{1:2},\phi)\right) p\left(\phi|M\right) dk_{1:2}\, d\phi$$

As in Section 3.1 we are using an SMC samplers approach, employing a system of interacting particles to efficiently sample from $U(t_{1:2}|n)$[100]. The resulting samples as well as the derived optimal design are shown in Figure 3.

3.3 Control Example: Particles with Force-Fields

Consider a physical system consisting of particles moving in a 2-dimensional space. The particles are released from some stochastic start region, fall downwards under the force of gravity, and are slowed by a frictional force resisting movement. At each discrete time step the particles receive some reward based on their current position and the position and velocity are then updated using a simple forward simulation. The goal is to direct the particles, using additional



(a) example model fits and optimal design (b) $U(t_{1:2})$ and final particle locations

Fig. 3. Figure (a) shows the computed optimal design for discriminating between the power and exponential model for explaining human memory performance; the two black bars are the optimal time points for testing a subject's retention. For illustration we show fits of the two models (red and blue curve) to an example subject's data from a previous study [12] (black dots). Figure (b) visualizes both $U(t_{1:2}|n)$ (background colors) as well as the final samples from $U(t_{1:2}|n)^{100}$ (black points) that the optimal design in (a) was derived from.

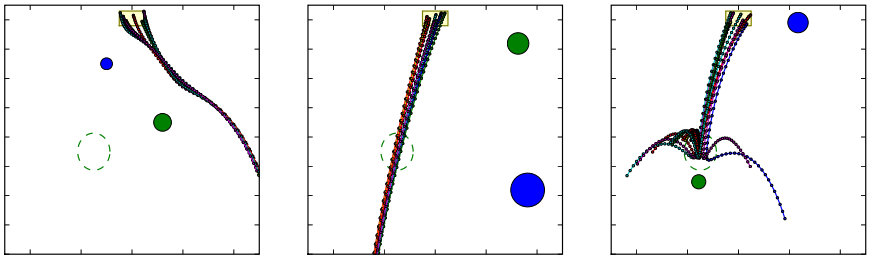
forces, through high reward regions of the state space in order to maximize their expected utility.

The four-dimensional state-space in this problem consists of a particle's position and velocity $s_n = (p, \dot{p})$ for $p \in \mathbb{R}^2$, and actions a_n consist of external forces acting on the particles. In particular, we will use a policy defined by a set of "repellers" which push each particle directly away from themselves with a force inversely proportional to their distance from the particle. More precisely, the force acting on a particle at position p is given by $a_n = f_\pi(p) = \sum_i w_i \frac{p-c_i}{\|p-c_i\|^3}$, where this policy is parameterized by $\pi = \{(c_1, w_1), \dots\}$ for repeller centers c_i and strengths w_i. Example trajectories for different policies are shown in Figure 4. In these examples rewards r_n are defined using a simple Gaussian defined over the particles' position p.

This model is particularly interesting because it is highly multimodal with large flat regions in the expected utility surface. In Figure 4(b) we can see one policy which very quickly moves particles into the reward region. A better policy can be seen in 4(c), in which one repeller is used to direct particles towards the reward region while another repeller slows particles so that they stay in this region as long as possible. This better policy is the one ultimately found by the procedure described in this paper.

References

1. J. Bernardo. Expected information as expected utility. *The Annals of Statistics*, 7(3):686–690, 1979.



(a) a bad policy (b) a local maximum (c) a better maximum

Fig. 4. Example trajectories from the repellers model. The two colored circles denote the repellers' location and strength, the dotted circle denotes the reward region, and the yellow rectangle is the initial state region. Shown are (a) a low utility policy, (b) a low utility local maximum, and (c) a higher utility maximum.

2. D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 1995.
3. K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.
4. P. Dayan and G. E. Hinton. Using EM for reinforcement learning. *Neural Computation*, 9:271–278, 1997.
5. P. Green. Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.
6. M. Hoffman, A. Doucet, N. de Freitas, and A. Jasra. Bayesian policy learning with trans-dimensional MCMC. In *NIPS*, 2007.
7. M. Hoffman, A. Doucet, N. de Freitas, and A. Jasra. On solving general state-space sequential decision problems using inference algorithms. Technical Report TR-2007-04, University of British Columbia, Computer Science, 2007.
8. H. Kueck, N. de Freitas, and A. Doucet. SMC samplers for Bayesian optimal nonlinear design. In *Nonlinear Statistical Signal Processing*, 2006.
9. T. J. Loredo. Bayesian adaptive exploration. In *Bayesian Inference And Maximum Entropy Methods In Science And Engineering*, pages 330–346, 2003.
10. P. Müller, B. Sansó, and M. de Iorio. Optimal Bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association*, 99:788–798, 2004.
11. J. I. Myung and M. A. Pitt. Optimal experimental design for model discrimination. *under review*.
12. D. Rubin, S. Hinton, and A. Wenzel. The precise time course of retention. *Journal of experimental psychology. Learning, memory, and cognition*, 25(5):1161–1176, 1999.
13. D. C. Rubin and A. E. Wenzel. One hundred years of forgetting: A quantitative description of retention. *Psychological review*, 103:734–760, 1996.
14. Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2009.
15. M. Toussaint and A. Storkey. Probabilistic inference for solving discrete and continuous state Markov Decision Processes. In *ICML*, 2006.
16. J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behaviour*. Princeton University Press, 1947.