Parameter Estimation in Softmax Decision-Making Models With Linear Objective Functions

Paul Reverdy and Naomi Ehrich Leonard

***Abstract*—We contribute to the development of a systematic means to infer features of human decision-making from behavioral data. Motivated by the common use of softmax selection in models of human decision-making, we study the maximum-likelihood (ML) parameter estimation problem for softmax decision-making models with linear objective functions. We present conditions under which the likelihood function is convex. These allow us to provide sufficient conditions for convergence of the resulting ML estimator and to construct its asymptotic distribution. In the case of models with nonlinear objective functions, we show how the estimator can be applied by linearizing about a nominal parameter value. We apply the estimator to fit the stochastic Upper Credible Limit (UCL) model of human decision-making to human subject data. The fits show statistically significant differences in behavior across related, but distinct, tasks.**

***Note to Practitioners*—Many problems in online planning and control can be formulated as sequential decision-making tasks in which an agent seeks to maximize rewards gained (or equivalently, minimize costs incurred) from a series of choices among control actions. When the task is highly structured, methods from optimal control can provide effective automated solutions to the control problem. However, when the uncertainties associated with the task are significant, solutions to the control problem generally require some input from human supervisors because of the humans' greater flexibility, for example, to adapt to unforeseen events. For human-centered automation, one seeks to combine the computational abilities of machine automation with the flexibility of a human supervisor in an effective way. In a previous paper (Reverdy *et al.*, Proc. IEEE, vol. 102, no. 4, pp. 544–571, 2014), we studied human decision-making behavior in a reward-driven decision-making task and showed that a significant fraction of subjects exhibited very high performance, which we ascribed to their intuition about the task. We developed a model (UCL) of this behavior that represents the human subject's intuition in terms of a small number of parameters. Estimating the model parameters from observed choice behavior would allow an automated system to quantify and learn the human's intuition, which the system could use to improve its own performance. To that end, this paper addresses the parameter estimation problem for the UCL model. The softmax functional form of the UCL model is a common feature of models of human decision-making, which makes the estimator we develop relevant to a wide range of decision-making models.**

***Index Terms*—Automation, decision-making, estimation.**

I. Introduction

In a variety of decision-making scenarios an agent selects one among a discrete set of options $i \in \{1,\ldots,m\}$ and receives a reward associated with the selection. The agent's goal is to make a selection or a sequence of selections to maximize reward. For example, a human air traffic controller selects among options for allocating aircraft for takeoff, and the reward is a measure of efficiency of flight departures associated with the selected option [24]. Often the decision-making task is challenging, especially when there is uncertainty or there are complex dependencies associated with options and rewards, as in the air traffic control example.

Much research has gone into studying how humans decide among options and what conditions lead to good decision-making performance. In this research, decision-making models are used together with empirical data. One common approach is to derive a decision-making model as the solution of an optimization problem. An objective function Q_i is defined for each option i, and the model agent selects the option i^* that maximizes the objective function

$$i^* = \arg\max_i Q_i.$$

The maximum operation is deterministic and non-differentiable, so for many applications it is replaced by the so-called "softmax" operation, in which option i is chosen with probability

$$\Pr[i] = \frac{\exp(Q_i)}{\sum_{j=1}^m \exp(Q_j)}.$$

The softmax operation, which we adopt in this paper, is a stochastic, biologically plausible approximation of the maximum operation [33]. Furthermore, it is differentiable with respect to its argument Q_i, which makes it more analytically tractable. Numerous works in the psychology and neuroscience literature, e.g., [4], [5], [7], and [37], have developed models of human decision-making behavior that apply the softmax operation to various objective functions Q_i.

In contexts such as inverse reinforcement learning [28], [23] and neuroscience [20], a central goal is to understand the decision-making process by finding the objective function values

Manuscript received February 16, 2015; revised August 18, 2015; accepted November 03, 2015. Date of publication November 24, 2015; date of current version January 01, 2016. This paper was recommended for publication by Associate Editor C. Tomlin and Editor D. Tilbury upon evaluation of the reviewers' comments. This work was supported in part by Office of Naval Research (ONR) under Grant N00014-09-1-1074 and Grant N00014-14-1-0635, and was conducted with Government support under and awarded by DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a.

P. Reverdy is with the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: preverdy@seas.upenn.edu).

N. E. Leonard is with the Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: naomi@princeton.edu).

Digital Object Identifier 10.1109/TASE.2015.2499244

$\{Q_i\}$ that explain observed decisions. In this paper, we consider this problem in the case that each objective function value Q_i is linear in a set of known variables $\mathbf{x}$, i.e.,

$$Q_i = \boldsymbol{\theta}^T \mathbf{x}_i, \quad \boldsymbol{\theta}, \mathbf{x}_i \in \mathbb{R}^{n_{\text{obj}}}. \tag{1}$$

Models of this form are often used in studies of human decision-making behavior, e.g., [7], [19], [4], [10], and are therefore of interest in developing principled methods for human-centered automation. Further, by assuming the functional form (1), we reduce the problem of finding the objective function values to that of learning the vector of parameters $\boldsymbol{\theta}$, which we assume to be constant across options and decisions. We call the reduced problem the *parameter estimation problem* for softmax decision-making models with linear objective functions. The linear functional form of (1) allows us to derive conditions for convergence of the parameter estimator. In the more general case where the objective function is nonlinear, it can be locally approximated with a linear function of the form (1).

The problem of learning the objective function that can explain observed decision-making behavior is relevant for several different disciplines. In the behavioral sciences, it is often of interest to develop models that quantify the various factors that contribute to the decision-making process. Similarly, in engineering, system identification seeks to develop models of dynamic systems that can be used for engineering design. In either case the problem is generally solved in two steps. The first step is to determine which variables affect the process or system in question. In the context of the present paper, this is equivalent to determining the variables $\mathbf{x}$ in (1). The second step is to quantify the effect of each variable on the system. This is equivalent to learning the value of the parameters $\boldsymbol{\theta}$ in (1), i.e., solving the parameter estimation problem. We call the two-step process *fitting*. This paper develops an estimator with rigorous performance guarantees for the softmax decision-making model, which provides a tool for the second step in the fitting process.

For human-centered automation, one seeks to combine the computational abilities of machine automation with the flexibility of a human supervisor in an effective way. One approach is to design an automated system that can infer the intuition or the intent of a human operator and use the intuition to improve its own performance. This could be done if a decision-making model with parameters representing intuition could be fitted to observed human choice data. The estimator developed in the present paper makes this possible when applied to an appropriate decision-making model.

We demonstrate the estimator using an algorithmic model of human decision-making in a spatial search task, derived in [26]. The model, called the stochastic Upper Credible Limit (UCL) model, was derived by generalizing results in the neuroscience [37] and machine learning [13] literature concerning multi-armed bandit tasks in a Bayesian setting. In [26], the stochastic UCL model was shown to qualitatively reproduce experimentally observed human behavior. We use our estimator to infer from these experimental data the human decision-maker's intuition in terms of a set of prior beliefs about the task. The estimator is applicable to a more general class of decision-making tasks for which a softmax decision-making model can be developed.

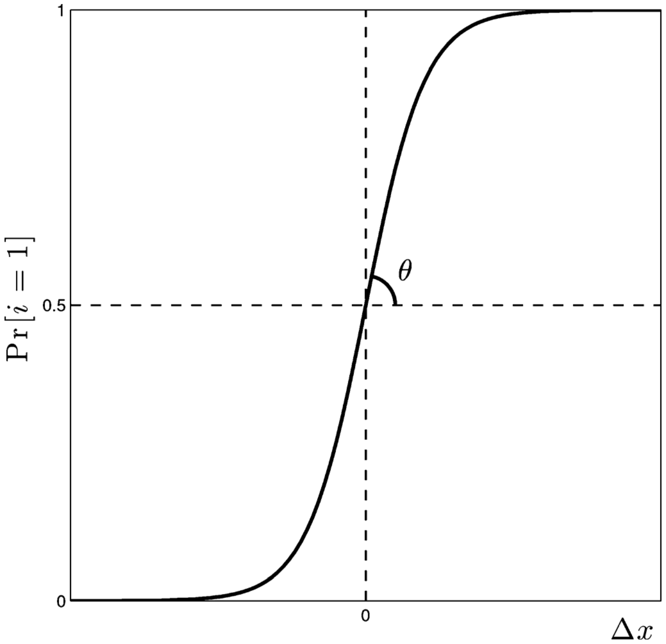


Fig. 1. The probability (2) from the model (1) with $m = 2$ options and a scalar ($n_{\text{obj}} = 1$) parameter θ. The probability of picking option 1 is a logistic function of $\Delta x = x_1 - x_2$ and the sensitivity to Δx is controlled by θ, which sets the slope at $\Delta x = 0$.

As a motivating example of the softmax model, consider the case of deciding between $m = 2$ options each with a single ($n_{\text{obj}} = 1$) known variable $\mathbf{x}_i = x_i$, $i = 1, 2$, representing the value of the option, and $\boldsymbol{\theta} = \theta$ a scalar. Then, the probability of picking option 1 is

$$\Pr[\text{pick option } 1] = \frac{1}{1 + \exp(-\theta(x_1 - x_2))}. \tag{2}$$

Fig. 1 plots the probability (2) as a function of the difference in value of the two options $\Delta x = x_1 - x_2$. When the values of the two options are identical, the probability is equal to 0.5 and it increases monotonically with increasing Δx. The rate of the increase is controlled by θ, which sets the slope of the function at $\Delta x = 0$. Large values of θ increase the slope and make the choice represented by (2) discriminate between x_1 and x_2 with more sensitivity, while small values of θ decrease the slope and make the choice less sensitive to Δx. Models of this form have been used to study a variety of decision-making tasks [16], [29], [7], [21], [32], where finding the value of θ that explains a given set of decisions is an important problem.

The parameter estimation problem for softmax decision-making models is related to other problems previously studied in the literature, in particular, the multinomial logistic regression problem [1], [15] and the conditional log-likelihood model learning problem [9]. With the linear functional form (1), the softmax decision-making model and the conditional log-likelihood model are formally equivalent, meaning that the parameter estimation problem has been studied in previous work, e.g., [9]. The novelty of the present paper comes in the application of parameter estimators to a formal model of human decision-making (the stochastic UCL model) and its use in quantifying a human subject's intuition about a decision-making task.

The stochastic UCL model for human decision-making in spatial search tasks [26] is a softmax decision model with an objective function Q_{UCL} that is a nonlinear function of several parameters. We show how Q_{UCL} can be transformed into a linear function of the form (1) by linearizing about a point in parameter space.

We adopt a maximum-likelihood (ML) approach to parameter estimation. In this framework, the convexity of a model implies asymptotic convergence of estimators and the convexity of the associated optimization problem. The convexity of the conditional log-likelihood model is an accepted fact in the natural language processing literature [9], so we do not focus on it here. We apply standard optimization algorithms to the stochastic UCL parameter estimation problem and demonstrate our results.

There are two major contributions of this paper. First, we show how to apply standard parameter estimation techniques to the stochastic UCL model, a rigorously derived model of human choice behavior. Models with a similar softmax functional form are commonly used in the neuroscience literature to model choice behavior and are likely to be widely applicable in the context of human-centered automation. Estimating the parameters of such models provides a method to quantify human intention and intuition in choice tasks, which can be leveraged in human-centered automation systems. Second, we apply the parameter estimation techniques to empirical human choice data and find statistically significant differences between groups of subjects presented with different experimental conditions.

The remainder of this paper is structured as follows. Section II defines the softmax decision model. Section III defines the parameter estimation problem for the softmax model and reviews convergence results from the literature. Section IV summarizes conditions under which the ML estimator converges. Section V provides a numerical example of the estimator. Section VI linearizes the stochastic UCL model about a nominal parameter $\bar{\boldsymbol{\theta}}$ to yield a softmax decision model with a linear objective function, and applies the estimator to simulated data. Section VII applies the linearization procedure to fit the stochastic UCL model to human subject data. Section VIII concludes.

II. The Softmax Decision Model

In this section, we define our notation and the specific softmax decision model for which we derive estimator convergence bounds. We also provide several examples of this model that appear in related literature.

A. Notation

In the spirit of [15], we set the following notation. We assume we have n observations. For each observation k, we have data consisting of $d = m \cdot n_{\mathrm{obj}}$ explanatory variables and a response, corresponding to the assignment of one of m classes. Specifically, for each observation $k \in \{1, \ldots, n\}$ we have data $(\mathbf{x}^k, \mathbf{y}^k)$. For each class $i \in \{1, \ldots, m\}$, we have n_{obj} explanatory variables $\mathbf{x}_i^k \in \mathbb{R}^{n_{\mathrm{obj}}}$. The vector of explanatory variables $\mathbf{x}^k \in \mathbb{R}^d$ is composed of the concatenation of the $\mathbf{x}_i^k$

$$\mathbf{x}^k = \left[\mathbf{x}_1^{k\,T};\ \mathbf{x}_2^{k\,T};\ \cdots \mathbf{x}_m^{k\,T}\right]^T.$$

The response variable $\mathbf{y}^k = (y_1^k, \ldots, y_m^k)^T$ represents the class assignment, where the element $y_i^k = 1$ if the observation corresponds to class i and zero otherwise.

Motivated by models of decision-making [26], we consider the following statistical model:

$$p_i^k(\boldsymbol{\theta}) = \Pr\left[y_i^k = 1 \,|\, \mathbf{x}^k, \boldsymbol{\theta}\right] = \frac{\exp\left(\boldsymbol{\theta}^T \mathbf{x}_i^k\right)}{\sum_{j=1}^m \exp\left(\boldsymbol{\theta}^T \mathbf{x}_j^k\right)} \tag{3}$$

for $i \in \{1, \ldots, m\}$, where $\boldsymbol{\theta} \in \mathbb{R}^{n_{\mathrm{obj}}}$ is a weight vector that is the same for all classes. This is the softmax decision-making model with linear objective function (1) introduced above, which has been studied in other literatures under other names. In the natural language processing literature, (3) is known as the *conditional log-likelihood* model, while in the econometrics literature, it is known as the *conditional logit* model [17].

B. Example Softmax Decision Models

In this section, we provide several concrete examples of the softmax decision model (3). The goal is to make the connection between this functional form and others that appear in the literature.

Example 1 (Softmax With Unknown Temperature): A standard decision model in reinforcement learning [33] is the so-called softmax action selection rule, which selects an option i with probability

$$\Pr[i] = \frac{\exp(V_i/\tau)}{\sum_{j=1}^n \exp(V_j/\tau)}$$

where V_i is the value associated with option i and τ is a positive parameter known as the *temperature*. This rule selects options stochastically, preferentially selecting those with higher values. The degree of stochasticity is controlled by the temperature τ. In the limit $\tau \to 0^+$, the rule reduces to the standard maximum and deterministically selects the option with the highest value of V_i. In the limit $\tau \to +\infty$, all options are equally probable and the rule selects options according to a uniform distribution.

This model is in the form of (3) with $n_{\mathrm{obj}} = 1$. Specifically, assume that the temperature τ is constant but unknown, and the values V_i are known. Then, the two models are identical if we identify

$$\theta = 1/\tau, \quad \mathbf{x}_i = V_i.$$

In the reinforcement learning literature, the quantity $1/\tau$ is sometimes known as the *inverse temperature* and referred to by the symbol β. Our methods allow one to estimate $\theta = 1/\tau = \beta$ from observed choice data.

Example 2 (Softmax With Known Cooling Schedule Form): A slightly more complicated model might let the softmax temperature τ of Example 1 follow a known functional form, called a *cooling schedule*, that depends on an unknown parameter. For example, in simulated annealing, Mitra *et al.* [18] showed that good cooling schedules follow a logarithmic functional form:

$$\tau(t) = \frac{\nu}{\log t}$$

where t is the decision time and $\nu > 0$ is a parameter.

If ν is constant but unknown, this model can be represented in the form of (3) with $n_{\mathrm{obj}} = 1$ if we identify

$$\theta = 1/\nu, \quad \mathbf{x}_i = V_i \log t.$$

Example 3 (Softmax Q-Learning With Unknown Temperature and Learning Rate): According to a simple Q-learning model [35], for each choice time t the agent assigns an expected value V_i^t to each option i. The values are initialized to 0 at $t = 1$ and then for each subsequent time, the agent picks option i_t, receives reward r_t, and updates the value of the chosen option i_t according to

$$V_{i_t}^{t+1} = V_{i_t}^t + \alpha\delta_t$$

where $\alpha \in [0, 1]$ is a free parameter called the *learning rate* and $\delta_t = r_t - V_{i_t}^t$ is the *prediction error* at time t.

A common model in reinforcement learning [6] has the agent make decisions using a softmax rule on the value function V_i^t, so the probability of selecting an option i at time t is

$$\begin{aligned}\Pr[i_t = i] &= \frac{\exp\left(V_i^t/\tau\right)}{\sum_{j=1}^n \exp\left(V_j^t/\tau\right)}\\ &= \frac{\exp\left(V_i^{t-1}/\tau + \alpha\delta_{t-1}\mathbf{1}(i = i_{t-1})/\tau\right)}{\sum_{j=1}^n \exp\left(V_j^{t-1}/\tau + \alpha\delta_{t-1}\mathbf{1}(i = i_{t-1})/\tau\right)}\end{aligned}$$

where $\mathbf{1}(\,\cdot\,)$ is the indicator function, equal to 1 if its argument is a true statement, and 0 otherwise. Similar models are used in the analysis of fMRI data, e.g., [38]. If V_i^{t-1}, V_i^{t-2}, and r_t are known while τ and α are unknown, the model is in the form of (3) with $n_{\mathrm{obj}} = 2$ if we identify

$$\boldsymbol{\theta} = [1/\tau; \quad \alpha/\tau], \quad \mathbf{x}_i = \left[V_i^{t-1}; \quad \delta_{t-1}\mathbf{1}(i = i_{t-1})\right].$$

If only the initial value $V_i^{t=1} = 0$ is known, then the value function V_i^t becomes a nonlinear function of the parameters α, τ and the model is not of the form (3), although it may be possible to find a transformation that puts it in such a form.

In the following section, we define the parameter estimation problem for the softmax model (3). We then analyze the problem to develop conditions under which this parameter estimation problem can be solved with provable guarantees about its convergence.

III. Parameter Estimation for Softmax Decision-Making Models

In this section, we define the parameter estimation problem for softmax decision-making models using a likelihood framework, and we review relevant results from the literature. Key to these results is the concept of concavity, which is a property of functions that can guarantee the uniqueness of a maximum. When the likelihood function is concave, the ML estimation problem can be solved by off-the-shelf optimization algorithms. Concavity is also central to several results from the econometrics literature that provide conditions under which the estimator is guaranteed to converge asymptotically.

In the optimization literature, it is traditional to consider minimization problems, for which convexity plays the same role as concavity does for maximization problems: a function f is concave if the function $-f$ is convex, and maximizing f is equivalent to minimizing $-f$. Following the literature, we refer to concavity and convexity when discussing results from econometrics and optimization, respectively. We distinguish between two notions of concavity: a function $f : \mathbb{R}^n \to \mathbb{R}$ is *weakly* concave if its Hessian is negative semidefinite, and *strongly* concave if its Hessian is strictly negative definite.

A. The Softmax Model Parameter Estimation Problem

In the parameter estimation problem for softmax decision-making models, we wish to estimate the values of $\boldsymbol{\theta}$ based on the observed data $(\mathbf{x}^k, \mathbf{y}^k)$. A standard way to perform parameter estimation is using the ML method [14]. To perform ML estimation of $\boldsymbol{\theta}$, one maximizes the log-likelihood function $\ell(\boldsymbol{\theta})$.

Problem 1: The ML *parameter estimation problem* for the softmax decision model (3) is the optimization problem

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) \tag{4}$$

where $\ell(\boldsymbol{\theta})$ is the logarithm of the likelihood function of the model (3), defined as

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_{k=1}^n \log \Pr[\mathbf{y}^k \mid \mathbf{x}^k, \boldsymbol{\theta}]\\ &= \sum_{k=1}^n \left[\sum_{i=1}^m y_i^k \boldsymbol{\theta}^T \mathbf{x}_i^k - \log \sum_{i=1}^m \exp\left(\boldsymbol{\theta}^T \mathbf{x}_i^k\right)\right].\end{aligned} \tag{5}$$

The ML estimate $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ can be interpreted as the parameter value $\boldsymbol{\theta}$ that makes the observed data most likely under the given model.

A prior on $\boldsymbol{\theta}$ can be incorporated by adopting a maximum *a posteriori* (MAP) estimate

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} [\ell(\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})] \tag{6}$$

with $p(\boldsymbol{\theta})$ being the prior on $\boldsymbol{\theta}$. The MAP estimate penalizes ML estimates that are considered unlikely under the prior.

B. Asymptotic Behavior of the ML Estimator

The ML estimator $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ solves the estimation problem in the frequentist framework, which posits that there is a true value $\boldsymbol{\theta}_0$ of the parameters that we attempt to recover from analyzing the given data. In this framework, natural questions to be asked are: 1) does $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} \to \boldsymbol{\theta}_0$ as the number of observations n grows and 2) how dispersed is the difference $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}_0$? These questions have been studied in the econometrics literature, for which [22] is a standard reference. The remainder of this section summarizes the relevant results from [22]. The answers to these two questions depend on two properties of the model, identification and concavity, defined as follows.

Definition 1 (Identification): A statistical model with likelihood function $\ell : \mathbb{R}^q \to \mathbb{R}$ and observed data $\mathbf{x}$ is said to be *identified* if, for all $\boldsymbol{\theta}, \boldsymbol{\theta}_0 \in \mathbb{R}^q$

$$\boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \Rightarrow \ell(\boldsymbol{\theta}_0; \mathbf{x}) \neq \ell(\boldsymbol{\theta}; \mathbf{x}).$$

Definition 2 (Concavity): A statistical model with likelihood function $\ell : \mathbb{R}^q \to \mathbb{R}$ is said to be *concave* if $\ell(\boldsymbol{\theta}; \mathbf{x})$ is strictly concave in $\boldsymbol{\theta}$.

If a model with likelihood function ℓ is identified and concave (see [22, Th. 2.7]) for details), the answer to question 1) is yes. These two properties imply that the true value $\boldsymbol{\theta}_0$ of the parameter is the unique maximum of the expected value of the log-likelihood $\ell(\boldsymbol{\theta})$.

Concavity and identification can depend on both the functional form of $\ell(\boldsymbol{\theta})$ and the observed data $\mathbf{x}$. As an example of how the identification property may fail due to data, consider the model (3) with $\mathbf{x}_i$ being the zero vector for each i. In this case, $\Pr[y_i = 1 \,|\, \mathbf{x}, \boldsymbol{\theta}] = 1/m$ for each i independent of $\boldsymbol{\theta}$ and the estimation procedure will be unable to distinguish among the possible parameter values. In the following section, we derive conditions on the data that ensure identification. These conditions also ensure that $\ell(\boldsymbol{\theta})$ is strictly concave and provide guidelines for the design of experiments for estimating $\boldsymbol{\theta}$.

The answer to question 2) is that, under mild regularity conditions, the distribution of $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ approaches a normal distribution as the number of samples n grows. In particular, the following limit holds:

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} \xrightarrow{d} \mathcal{N}(\boldsymbol{\theta}_0, \mathbf{J}^{-1}/n) \tag{7}$$

where $\xrightarrow{d}$ signifies a limit in distribution as $n \to \infty$ and $\mathbf{J} = -\mathbb{E}[\mathbf{H}(\boldsymbol{\theta}_0)]$ is the negative of the expected value of the Hessian of $\ell(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. See [11, Chap. 9] for more details about the concept of a limit in distribution and see [22, Th. 3.3] for full details of the conditions under which (7) holds. In practice, one uses $\hat{\mathbf{J}} = -\mathbf{H}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})/n$ as an estimate of $\mathbf{J}$. This permits construction of standard frequentist analysis tools, such as confidence intervals for the parameter estimates and hypothesis tests. The estimate $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is efficient in the sense that it obeys the Cramér-Rao lower bound [14] on the variance of estimators $\hat{\boldsymbol{\theta}}$, so no other unbiased estimator can have lower variance than $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$.

IV. Analysis of the Maximum-Likelihood Estimator for Softmax Decision Models With Linear Objective Functions

In this section, we present conditions under which the model (3) is identified and concave. These conditions imply that the ML estimator $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ converges and that the ML optimization problem (4) is convex. The concavity of the model is an accepted fact in the natural language processing literature [9]; we summarize the result in Theorem 1.

A. Asymptotic and Finite-Sample Behavior

Recall from Section III-B that two properties that guarantee asymptotic convergence of the ML estimator are identification and concavity. Whether or not the model (3) satisfies these properties can be a function of the data $\mathbf{x}^k, k \in \{1, \ldots, n\}$. Recall our example where $\mathbf{x}_i^k = \mathbf{0}$ for each i and k. In this case, the probability $\Pr[y_i^k \,|\, \mathbf{x}^k, \boldsymbol{\theta}] = 1/m$ for each i and k for all values of $\boldsymbol{\theta}$ and the likelihood function $\ell(\boldsymbol{\theta})$ is flat, so neither identification nor concavity is satisfied.

However, a sufficient condition for identification is as follows. Define the $n_{\mathrm{obj}} \times m$ matrix $\mathbf{X}^k$ by transforming the explanatory variable $\mathbf{x}^k$ of a single observation k

$$\mathbf{X}^{\mathbf{k}} = \begin{bmatrix} \mathbf{x}_1^k & \mathbf{x}_2^k & \cdots & \mathbf{x}_{m-1}^k & \mathbf{0} \end{bmatrix}. \tag{8}$$

Note that $\mathbf{X}^k \mathbf{X}^{k^T} = \sum (\mathbf{x}^{k^T} \mathbf{x}^k)$. Considering $\mathbf{X}^k$ as a random variable, the following lemma ensures identification.

Lemma 1: Let $\mathbf{x}$ be the explanatory variable for an arbitrary observation and let $\mathbf{X}$ be the transformation of $\mathbf{x}$ defined in (8). If the second-moment matrix $\mathbb{E}[\mathbf{X}\mathbf{X}^T]$ exists and is positive definite, then the model (3) is identified.

Proof: The probability of choosing an option i under the model (3) is a monotonic function of the objective value Q_i, so it suffices to show that the data provides a one-to-one mapping between the parameter vector $\boldsymbol{\theta}$ and the objective values $Q_1, \ldots, Q_m$.

Let $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^{n_{\mathrm{obj}}}$ and define the vectors of objective function values $\mathbf{Q} = \boldsymbol{\theta}^T \mathbf{X}$ and $\mathbf{Q}' = \boldsymbol{\theta}'^T \mathbf{X}$. Define $\Delta \mathbf{Q} = \mathbf{Q} - \mathbf{Q}' = (\boldsymbol{\theta} - \boldsymbol{\theta}')^T \mathbf{X} \in \mathbb{R}^m$. The magnitude of $\Delta \mathbf{Q}$ satisfies $\mathbb{E}[\|\Delta \mathbf{Q}\|^2] = (\boldsymbol{\theta} - \boldsymbol{\theta}')^T \mathbb{E}[\mathbf{X}\mathbf{X}^T](\boldsymbol{\theta} - \boldsymbol{\theta}')$. Then, by the assumption that $\mathbb{E}[\mathbf{X}\mathbf{X}^T]$ is positive definite, $\mathbb{E}[\|\Delta \mathbf{Q}\|^2] = 0$ implies $(\boldsymbol{\theta} - \boldsymbol{\theta}') = 0$, so $\boldsymbol{\theta} = \boldsymbol{\theta}'$ and $\mathbf{Q} = \mathbf{Q}'$. Therefore, the mapping between the parameters $\boldsymbol{\theta}$ and the objective values $Q_1, \ldots, Q_m$ is one-to-one, which implies that $\ell(\boldsymbol{\theta} \,|\, \mathbf{x}, \mathbf{y}) \neq \ell(\boldsymbol{\theta}' \,|\, \mathbf{x}, \mathbf{y})$ for $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ and the model is identified. ■

The condition of Lemma 1 is given in terms of an expectation, but in practice one has a given sample of data. In this case, the expectation can be replaced by the sample average. Specifically, define $\mathbf{X}^k$ for each observation k following (8). Then, $\mathbb{E}[\mathbf{X}\mathbf{X}^T]$ is estimated by

$$\mathbb{E}[\mathbf{X}\mathbf{X}^T] \approx \frac{1}{n} \sum_{j=1}^{n} \mathbf{X}^k \mathbf{X}^{kT}.$$

If this sample average is positive definite, then the model is identified. For the sample average to be positive definite it must be full rank $= n_{\mathrm{obj}}$, and each observation k can add at most m to the rank. Therefore, the following inequality must be satisfied for the model to be identified:

$$m \cdot n \geq n_{\mathrm{obj}}.$$

This gives a lower bound $n \geq \lceil n_{\mathrm{obj}}/m \rceil$ on the minimum number of observations required for identification. For most applications, the number of options m will be larger than the number of parameters n_{obj}, so the lower bound is trivial. However, for cases with a large number of parameters the bound can be useful for experimental design.

The following theorem summarizes the conditions under which the ML estimator (4) converges.

Theorem 1 (Convergence of the ML Estimator): Let $\mathbf{X}^k$ be defined as in (8). If the second-moment matrix

$$\frac{1}{n} \sum_{k=1}^{n} \mathbf{X}^k \mathbf{X}^{k^T}$$

exists and is positive definite, then

1) The ML optimization problem (4) is convex.
2) The ML estimator $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ for the model (3) is asymptotically approximately distributed as

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} \sim \mathcal{N}(\boldsymbol{\theta}_0, \hat{\mathbf{J}}^{-1}/n) \tag{9}$$

where $\hat{\mathbf{J}} = -\mathbf{H}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})/n$ is the empirical mean Hessian of the likelihood function evaluated at the estimated parameter value.

Proof: See [17] and [25]. ■

Theorem 1 proves convergence of the parameter estimate $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ and provides its asymptotic distribution. This distribution can be used to formulate frequentist confidence intervals for the parameter estimate $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$. Furthermore, the theorem proves that the optimization problem (4) is convex, which allows us to solve it using off-the-shelf optimization algorithms. In the following, we use the phrase *the estimator* to refer to the procedure of using an off-the-shelf convex optimization algorithm to solve the ML problem (4). We use the phrase *the estimate* to refer to the solution $\hat{\boldsymbol{\theta}}$ of (4) thus obtained. The next three sections apply the estimator to increasingly complex data sets, building towards the application to experimental human subject data.

V. Numerical Examples

In this section, we present several numerical examples to demonstrate the theory developed in the previous sections for solving the parameter estimation problem (4). In all cases, the explanatory variables $\mathbf{x}$ were drawn randomly according to Gaussian distributions and the response variables $\mathbf{y}$ were drawn according to the model (3) conditional on the explanatory variables $\mathbf{x}$. Application to data generated from simulations of the stochastic UCL model is presented in Section VI, and application to data collected from human subjects is presented in Section VII.

A. Scalar Parameter

First, we consider model (3) with $m = 10$ options and a scalar parameter $\boldsymbol{\theta} = \theta = \theta_0$ that we wish to estimate. This could correspond to a decision-maker choosing among ten options using a softmax model with unknown constant inverse temperature $\theta = \theta_0$, as in Example 1. Alternatively, it could correspond to a temperature varying with observation number $k = 1, \ldots, n$ according to a known function with a single unknown parameter $\theta = \theta_0$, e.g., $\tau_k = \theta/\log k$, as in Example 2. In this case, the $\log k$ term can be absorbed into the explanatory variables and we proceed as in the first case.

Fig. 2 shows results of applying the estimator to simulated data. For every k, when an observation was taken and a decision made, the model was simulated 100 times. For each of the 100 simulations, the estimator was applied to estimate the parameter θ based on the first k observations. Running 100 simulations made it possible to examine convergence of the estimate in distribution. Fig. 2 illustrates how the estimates converge in distribution to the normal distribution (9) as the number of observations n increases. For the simulations, the explanatory variables $\mathbf{x}^k$ were drawn from a standard Gaussian distribution $\mathcal{N}(0,1)$ (mean zero and unit variance), and the response variables $\mathbf{y}^k$ were drawn according to probability distribution (3) conditional on $\mathbf{x}^k$ and $\theta_0 = 4$. The estimates were computed by solving the optimization problem (4) using a BFGS quasi-Newton algorithm [2], [8], [12], [30] (Matlab function `fminunc` [34]). Theorem 1 guarantees that the optimization problem is convex, so the algorithm will converge.

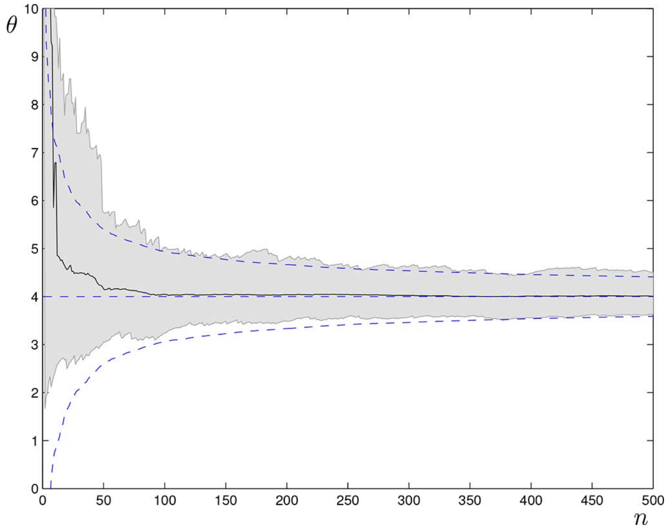


Fig. 2. Scalar parameter estimation example. The parameter estimates converge to the asymptotic normal distribution (9) as the number of observations n grows. The dashed lines show the true value of the scalar parameter $\theta_0 = 4$ and the accompanying 95% confidence intervals implied by the asymptotic normal distribution (9). For each value of n, an ensemble of 100 parameter estimates was formed by repeatedly simulating the data $\mathbf{y}$ while holding the explanatory variables $\mathbf{x}$ fixed, and using the estimator to compute the value of the parameter. The solid black line shows the mean parameter estimate and the shaded region the empirical 95% confidence interval.

The convergence behavior can be seen in Fig. 2 by observing the mean parameter estimate $\hat{\theta}_{\mathrm{ML}}$ as well as its confidence intervals. The mean parameter estimate $\hat{\theta}_{\mathrm{ML}}$, represented by the solid black line, converges to the true parameter value $\theta_0 = 4$, represented by the horizontal dashed line. However, Theorem 1 guarantees convergence in distribution, which is a stronger result. To illustrate this behavior we plot 95% confidence intervals for both the empirical distribution of estimates $\hat{\theta}_{\mathrm{ML}}$ and the asymptotic distribution (9), computed from the ensemble of 100 parameter estimates. For values of n greater than 100, the two intervals overlap closely, showing that the distribution of estimates has converged. Importantly, this shows that statistical hypothesis tests based on the asymptotic distribution (9) will be accurate.

For small amounts of data, i.e., $n < 50$, the mean parameter estimate is biased above the true value. The bias is due to an insufficient amount of data being used in the estimation procedure, and the direction of the bias can be explained as follows. Larger values of the parameter θ correspond to more deterministic choice behavior. When $\theta_0 > 0$, for any given choice, the model is more likely to pick the option with a larger objective value, resulting in a parameter estimate that is biased upwards.

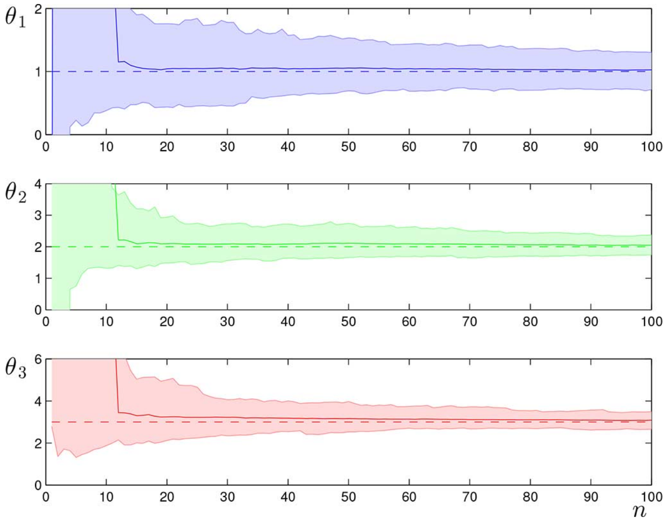


Fig. 3. Vector parameter estimation example. The parameter estimates converge to the asymptotic normal distribution (9) as the number of observations n grows. The dashed lines show the true value of each element of the vector parameter $\boldsymbol{\theta}_0 = [1,2,3]^T$. For each value of n, an ensemble of 100 parameter estimates was formed by repeatedly simulating the data $\mathbf{y}$ while holding the explanatory variables $\mathbf{x}$ fixed, and using the estimator to compute the value of the parameter. The solid lines show the mean parameter estimate and the shaded regions the empirical 95% confidence interval.

This bias can be seen in Fig. 3 as well, which treats a case with a vector parameter.

B. *Vector Parameter*

Next, we consider the model (3) with $m = 100$ options and a vector parameter $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ with $n_{\text{obj}} = 3$ elements that we wish to estimate. Fig. 3 shows results of applying the estimator to simulated data in this vector parameter estimation example. As in the scalar parameter estimation case above, the model was simulated 100 times for every $k = 1, \ldots, n$. Fig. 3 shows how the estimate converges to the true value $\boldsymbol{\theta}_0$ as the total number of observations n increases. The explanatory variables $\mathbf{x}^k$ were drawn according to independent standard Gaussian distributions, and the response variables $\mathbf{y}^k$ drawn according to the model (3) conditional on $\mathbf{x}^k$ and true vector parameter value $\boldsymbol{\theta}_0 = [1,2,3]^T$. The estimates were computed as in the scalar case.

The convergence behavior can be seen in Fig. 3 by observing the mean parameter estimate $\hat{\boldsymbol{\theta}}_{\text{ML}}$ as well as its confidence intervals. For each of the three parameters θ_i, $i = 1,2,3$, the corresponding mean parameter estimate $(\hat{\theta}_{\text{ML}})_i$, represented as a solid line, converges to the true parameter value $(\boldsymbol{\theta}_0)_i$, represented by a horizontal dashed line. The shaded regions represent the empirical 95% confidence interval around the corresponding mean value, computed from the ensemble of 100 parameter estimates. For clarity, we omit the confidence intervals implied by the asymptotic normal distribution (9) from the figure, but the behavior is similar to that shown in Fig. 2.

There is an upwards bias in the parameter estimates for small numbers of observations n, as in Fig. 2. The width of the confidence intervals for the three parameters scales roughly with their true value $(\boldsymbol{\theta}_0)_i$. This behavior can be seen in the figures in the next section as well.

VI. Application to Nonlinear Objective Functions Using Linearization

The development up to this point for addressing the parameter estimation problem (4) has assumed that the objective function takes the linear form (1). However, many relevant objective functions are nonlinear functions of the unknown parameter $\boldsymbol{\theta}$. One approach is to linearize the nonlinear objective function about a nominal parameter value, and then apply the estimator to the linearized objective function. We apply this approach to the nonlinear objective function from the stochastic UCL algorithm [26], an algorithm that models human decision-making in multi-armed bandit tasks in a Bayesian setting, and show how its parameters can be estimated.

A. *The Multi-Armed Bandit Problem*

The multi-armed bandit problem, introduced by Robbins [27] is a sequential decision-making problem which consists of a set of N options (each option is also called an *arm* in analogy with the lever of a slot machine). Each option $i \in \{1, \ldots, N\}$, has an associated probability distribution p_i with mean m_i, unknown to the agent solving the problem. At each sequential decision time $t \in \{1, \ldots, T\}$, the agent picks an arm i_t and receives a stochastic reward $r_t \sim p_{i_t}(r)$ drawn from the probability distribution associated with that arm. This is a special case of the notation introduced in Section II-A, with $m = N$ options indexed by i and $n = T$ decisions indexed by t. The agent's objective is to maximize the expected value of the cumulative rewards received from the T decisions

$$\max_{\{i_t\}} J, \quad J = \mathbb{E}\left[\sum_{t=1}^{T} r_t\right] = \sum_{t=1}^{T} m_{i_t}.$$

Each choice of i_t is made conditional on the information available to the agent at time t. If the mean rewards m_i were known to the agent, the optimal policy would be trivial: pick arm $i_t \in \arg\max_i m_i$ for each t. However, since the mean rewards are unknown, the agent must simultaneously select arms where the reward value is uncertain to gain information about the rewards and preferentially select arms with high rewards to accumulate reward. The tension between selecting arms with uncertain (but possibly high) rewards and selecting arms that appear to have high rewards based on current information is known as the *explore-exploit* tradeoff. This tradeoff is common to a variety of problems in machine learning and adaptive control.

The multi-armed bandit problem is the subject of active research in machine learning as well as in neuroscience. In [26], we showed that a significant fraction of human subjects exhibited excellent performance in solving a multi-armed bandit problem, even outperforming algorithms known to have optimal performance in some cases. We attributed this good performance to the human subjects' having good priors on the structure of the rewards m_i, and we designed the stochastic UCL algorithm as a model of human behavior to capture this dependence on priors. Estimating the parameters of this model from observations of a human solving the multi-armed bandit task would allow a machine to learn the human's belief priors. This could in turn facilitate the design of a human-machine system that could achieve better performance than either the human or the machine could on its own.

B. The Stochastic UCL Algorithm

The stochastic UCL algorithm [26] is designed to solve multi-armed bandit problems with Gaussian rewards, i.e., where the reward distribution $p_i(r) = \mathcal{N}(m_i, \sigma_s^2)$ is Gaussian with unknown mean m_i and known variance σ_s^2. The algorithm consists of two parts: Bayesian inference that maintains the agent's belief state and a softmax decision model that uses an objective function Q that depends on the belief state. Both the inference and the decision parts introduce nonlinear dependencies on the parameters of the algorithm.

As a model of human behavior, the stochastic UCL algorithm assumes that the agent's prior distribution of $\mathbf{m}$ (i.e., the agent's initial beliefs about the mean reward values $\mathbf{m}$ and their covariance) is multivariate Gaussian with mean $\boldsymbol{\mu}_\mathbf{0}$ and covariance Σ_0

$$\mathbf{m} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{0}, \Sigma_0)$$

where $\boldsymbol{\mu}_\mathbf{0} \in \mathbb{R}^N$ and $\Sigma_0 \in \mathbb{R}^{N\times N}$ is a positive-definite matrix.

In [26], we use a minimal set of three parameters to specify $(\boldsymbol{\mu}_0, \Sigma_0)$. For the mean we use a uniform prior $\boldsymbol{\mu}_0 = \mu_0 \mathbf{1}_N$,, where $\mu_0 \in \mathbb{R}$ is a single parameter that encodes the agent's belief about the mean value of the rewards and $\mathbf{1}_N$ is the vector with each element equal to 1. For the problems considered in [26], the arms are spatially embedded with each arm at a different location in space (see Fig. 8 in the next section). It is reasonable to assume that arms that are spatially close will have similar mean rewards. Therefore, for the covariance Σ_0 we set $\Sigma_0 = \sigma_0^2 \Sigma$, where Σ represents a prior that is exponential in distance, i.e., each element has the form

$$\Sigma_{ij} = \exp(-\|z_i - z_j\|/\lambda) \tag{10}$$

where z_i is the location of arm i and $\lambda \geq 0$ is the correlation length scale. The parameter $\sigma_0 \geq 0$ can be interpreted as a confidence parameter, with $\sigma_0 = 0$ representing absolute confidence in the beliefs about the mean $\boldsymbol{\mu}_0$, and $\sigma_0 = +\infty$ representing complete lack of confidence.

With this prior, the posterior distribution is also Gaussian, so the Bayesian optimal inference algorithm is linear and can be written down as follows. At each time t, the agent selects option i_t and receives a reward r_t. Let $\mathbf{r}^t$ be the $t \times 1$ vector composed of the r_t. Let n_i^t be the number of times the agent has selected option i up to time t, let $\bar{m}_i^t$ be the empirical mean reward observed for option i, and let $\mathbf{n}^t$ and $\bar{\mathbf{m}}^t$ be the vectors composed of the n_i^t and $\bar{m}_i^t$, respectively. For each time t, define the precision matrix $\Lambda_t = \Sigma_t^{-1}$. Then, the belief state at time t is ([14, Th. 10.3])

$$\Lambda_t = \frac{\operatorname{diag}(\mathbf{n}^t)}{\sigma_s^2} + \Lambda_0, \quad \Sigma_t = \Lambda_t^{-1} \tag{11}$$

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_0 + \Sigma_0 H_t^T \left(H_t \Sigma_0 H_t^T + \sigma_s^2 I_t\right)^{-1} (\mathbf{r}^t - H_t \boldsymbol{\mu}_0) \tag{12}$$

where diag maps a vector to a diagonal matrix, H_t is the $t \times N$ observation matrix with $H_t(t,j) = 1$ if $i_t = j$ and zero otherwise, and I_t is the t-dimensional identity matrix.

Based on the belief state $(\boldsymbol{\mu}_t, \Sigma_t)$, the stochastic UCL algorithm chooses arm i_t with probability

$$\Pr[i_t = i \,|\, \tilde{Q}^t, \upsilon_t] = \frac{\exp\left(\tilde{Q}_i^t/\upsilon_t\right)}{\sum_{j=1}^N \exp\left(\tilde{Q}_i^t/\upsilon_t\right)}, \tag{13}$$

where $\tilde{Q}_i^t$ is the heuristic function value for arm i at time t and υ_t is the temperature corresponding to the cooling schedule at time t. The cooling schedule is assumed to take the form $\upsilon_t = \nu/\log t$, ν a constant, so the probabilities (13) become

$$\Pr[i_t = i \,|\, \tilde{Q}^t, \nu] = \frac{\exp\left(\left(\tilde{Q}_i^t \log t\right)/\nu\right)}{\sum_{j=1}^N \exp\left(\left(\tilde{Q}_i^t \log t\right)/\nu\right)}. \tag{14}$$

The heuristic function is

$$\tilde{Q}_i^t = \mu_i^t + \sigma_i^t \Phi^{-1}(1 - \alpha_t) \tag{15}$$

where $\mu_i^t = (\boldsymbol{\mu}_t)_i$ is the posterior mean reward of arm i at time t and $\sigma_i^t = \sqrt{(\Sigma_t)_{ii}}$ its associated standard deviation. The quantity $\Phi^{-1}(\,\cdot\,)$ is the inverse cumulative distribution function of the standard normal distribution and $\alpha_t = 1/\sqrt{2\pi e t}$ is a decreasing function of time.

This is a softmax decision model with unknown parameters $(\mu_0, \sigma_0, \lambda, \nu)$, but it is not yet in the form (3) since the quantity $(\tilde{Q}_i^t \log t)/\nu$ is a nonlinear function of the parameters. However, we can locally approximate $(\tilde{Q}_i^t \log t)/\nu$ with a linear function by linearizing about a nominal value of the prior. By estimating the parameter values of the linearized model, we can recover the parameters of the original nonlinear model (14) near the nominal prior.

C. Linearization

Let $\delta_0^2 = \sigma_s^2/\sigma_0^2$ be the relative precision of a reward measurement compared to the certainty of the prior. Fix a nominal prior with parameter values $(\bar{\mu}_0, \bar{\delta}_0^2, \lambda)$ and consider small deviations Δ_μ and Δ_δ about $\bar{\mu}_0$ and $\bar{\delta}_0^2$, respectively

$$\mu_0 = \bar{\mu}_0 + \Delta_\mu, \quad \delta_0^2 = \bar{\delta}_0^2 + \Delta_\delta.$$

In the case that the true value of λ is unknown, this method is easily generalized to include deviations in λ, but for simplicity of exposition we consider it fixed. Recall that the covariance prior is $\Sigma_0 = \sigma_0^2 \Sigma$, where Σ is defined by (10), and its inverse is denoted by $\Lambda = \Sigma^{-1}$.

In terms of the nominal value $\bar{\delta}_0^2$, (11) becomes

$$\Lambda_t = \frac{1}{\sigma_s^2}\left(\operatorname{diag}(\mathbf{n}^t) + \bar{\delta}_0^2 \Lambda + \Delta_\delta \Lambda\right).$$

Therefore, to first order in Δ_δ, Σ_t is given by

$$\Sigma_t = \sigma_s^2 A_t^{-1} - \sigma_s^2 A_t^{-1} B A_t^{-1} \Delta_\delta + \mathcal{O}\left(\Delta_\delta^2\right) \tag{16}$$

where $A_t = \bar{\delta}_0^2 \Lambda + \operatorname{diag}(\mathbf{n}^t)$ and $B = \Lambda = \Sigma^{-1}$. Expanding the square root in the following, we get:

$$\sigma_i^t = \sqrt{(\Sigma_t)_{ii}} = \sqrt{c_i^t} - \frac{d_i^t}{2\sqrt{c_i^t}} \Delta_\delta + \mathcal{O}\left(\Delta_\delta^2\right) \tag{17}$$

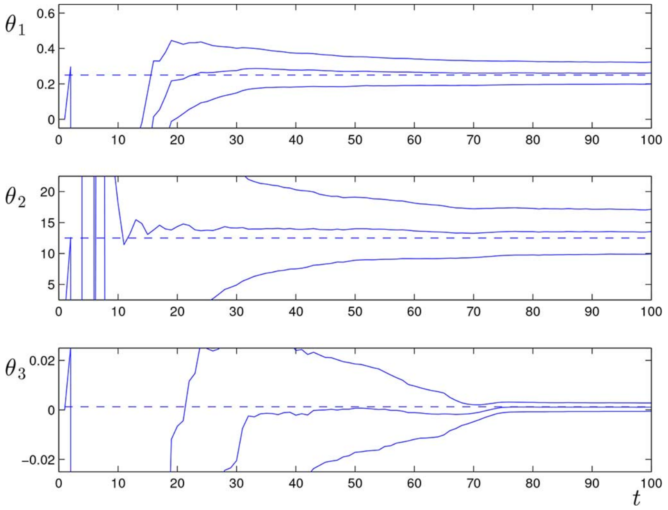


Fig. 4. Estimate of the vector of parameters $\boldsymbol{\theta}$ based on simulated data from the stochastic UCL algorithm. The linearization point was taken to be $\bar{\mu}_0 = 150, \bar{\sigma}_0^2 = 2$. The true algorithm parameters were $\mu_0 = 200, \sigma_0^2 = 1, \lambda = 1$, and $\nu = 4$. The estimate converges as the number of observations t grows. The dashed lines show the true value of each parameter θ_i. For each value of t, an ensemble of 100 parameter estimates was formed by repeatedly simulating the data $\{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^T$, while holding the parameters $\boldsymbol{\theta}$ fixed, and using the estimator to compute the value of the parameters. The solid lines show the mean parameter estimate and the 95% confidence interval implied by the asymptotic normal distribution (9).

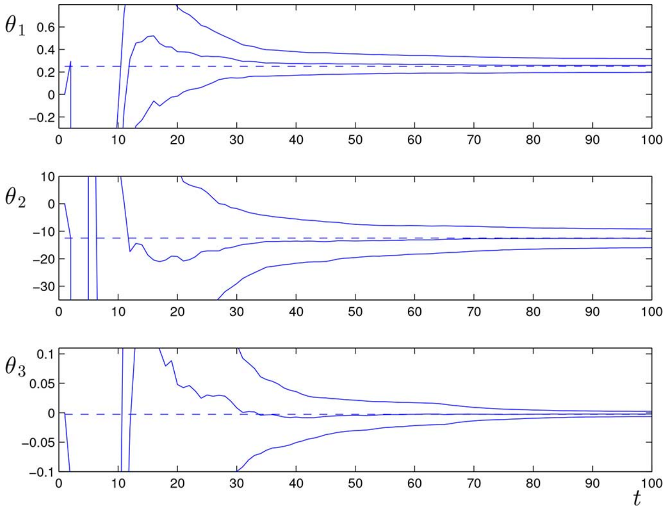


Fig. 5. Estimate of the vector of parameters $\boldsymbol{\theta}$ based on simulated data from the stochastic UCL algorithm. Everything is the same as in Fig. 4 except that the linearization point was taken to be $\bar{\mu}_0 = 250, \bar{\sigma}_0^2 = 0.5$.

where c_i^t is the ith element on the diagonal of $C_t = \sigma_s^2 A_t^{-1}$ and d_i^t is the ith element on the diagonal of $D_t = \sigma_s^2 A_t^{-1} B A_t^{-1}$. The standard deviation σ_i^t must be nonnegative, which implies an upper bound on Δ_δ. Similarly, δ_0^2 must be nonnegative, which implies a lower bound on Δ_δ, which is already assumed to be small. The implied bounds on Δ_δ are

$$-\bar{\delta}_0^2 = -\frac{\sigma_s^2}{\bar{\sigma}_0^2} \leq \Delta_\delta \leq \frac{2c_i^t}{d_i^t}$$

which, together with the requirement that Δ_δ be small with respect to $\bar{\delta}_0^2$, gives a bound on the values of Δ_δ for which the linearization is valid.

Similarly, the expression (12) for $\boldsymbol{\mu}_t$ becomes

$$\boldsymbol{\mu}_t = E_t + F_t \Delta_\mu + G_t \Delta_\delta + \mathcal{O}(\Delta^2) \tag{18}$$

where Δ^2 denotes second-order terms in the deviation variables Δ_δ and Δ_μ, and E_t, F_t, and G_t are the $N \times 1$ vectors

$$E_t = \bar{\mu}_0 \mathbf{1}_N + \frac{\Sigma H_t^T}{\bar{\delta}_0^2} \left(I_t - H_t A_t^{-1} H_t^T \right) \left(\bar{\mathbf{m}}^t - H_t \bar{\mu}_0 \mathbf{1}_N \right) \tag{19}$$

$$F_t = \mathbf{1}_N - \frac{\Sigma H_t^T}{\bar{\delta}_0^2} \left(I_t - H_t A_t^{-1} H_t^T \right) H_t \mathbf{1}_N \tag{20}$$

$$G_t = -A_t^{-1} B A_t^{-1} \left(H_t^T \mathbf{m}^t - \mathbf{n}^t \bar{\mu}_0 \right). \tag{21}$$

Define e_i^t, f_i^t, and g_i^t as the ith components of E_t, F_t, and G_t, respectively. Then, the linearized heuristic is

$$\frac{\tilde{Q}_i^t \log t}{\nu} \approx Q_i^t = \boldsymbol{\theta}^T \mathbf{x}_i^t = \theta_1 x_{i,1}^t + \theta_2 x_{i,2}^t + \theta_3 x_{i,3}^t \tag{22}$$

where the parameters $\boldsymbol{\theta}$ are defined by

$$\theta_1 = \frac{1}{\nu}, \quad \theta_2 = \frac{\Delta_\mu}{\nu}, \quad \theta_3 = \frac{\Delta_\delta}{\nu} \tag{23}$$

and the explanatory variables $\mathbf{x}_i^t$ are defined as

$$x_{i,1}^t = \left(e_i^t + \sqrt{c_i^t} \Phi^{-1}(1 - \alpha_t) \right) \log t \tag{24}$$

$$x_{i,2}^t = f_i^t \log t \tag{25}$$

$$x_{i,3}^t = \left(g_i^t - \frac{d_i^t}{2\sqrt{c_i^t}} \Phi^{-1}(1 - \alpha_t) \right) \log t. \tag{26}$$

The linearized heuristic (22) defines a softmax decision-making model with a linear objective function of the form (3). Thus, we can apply our estimation algorithm to estimate the parameters $\boldsymbol{\theta}$. Using (23), we can then use the estimate of $\boldsymbol{\theta}$ to provide an estimate of the parameters (μ_0, σ_0^2, ν).

D. *Example Estimates*

We tested the estimation procedure described above by simulating runs of the stochastic UCL algorithm for various parameter values. Figs. 4 and 5 show two examples of estimates computed using simulated data from the stochastic UCL algorithm with the nonlinear objective function $(\tilde{Q}_i^t \log t)/\nu$ and true parameters $(\mu_0, \sigma_0^2, \lambda, \nu) = (200, 1, 1, 4)$. These parameters result in the algorithm achieving high performance (specifically, logarithmic regret, see [26] for details). Fig. 4 shows estimates based on linearization about the point $(\bar{\mu}_0, \bar{\sigma}_0^2) = (150, 2)$. Following (23), the linearized objective function corresponds to parameters θ_1, θ_2, and θ_3 having true values $\theta_1 = 1/\nu = 0.25, \theta_2 = (\mu_0 - \bar{\mu}_0)/\nu = 12.5$, and $\theta_3 = 1.25 \times 10^{-3}$. These are the values to which the estimates should converge. Fig. 5 shows estimates based on linearization about the point $(\bar{\mu}_0, \bar{\sigma}_0^2) = (250, 0.5)$. The linearized objective function in this case corresponds to the three parameters taking true values $\theta_1 = 0.25, \theta_2 = -12.5$, and $\theta_3 = -2.5 \times 10^{-3}$.

In both cases the estimator converges to the true value of $\boldsymbol{\theta}$ within the horizon $T = 100$ of the decision task. Further, the

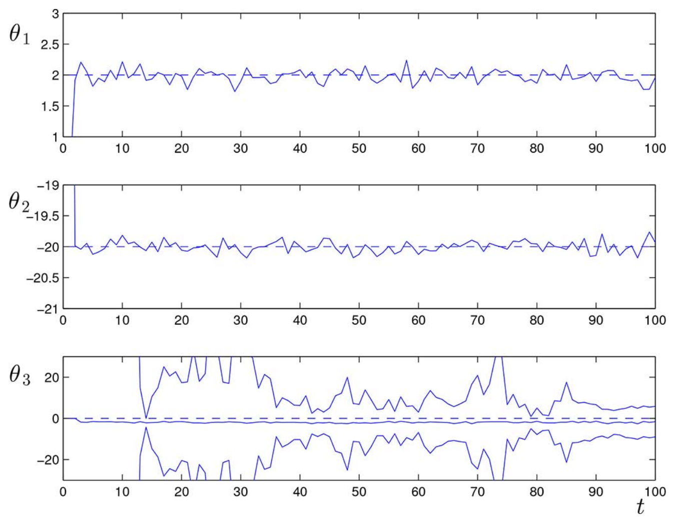


Fig. 6. Estimate of the vector of parameters $\boldsymbol{\theta}$ based on simulated data from the UCL algorithm with a weakly-informative prior. This prior makes the algorithm's choice behavior more random, which makes the estimation problem more difficult. Everything is the same as in Fig. 4 except that the linearization point was taken to be $\bar{\mu}_0 = 40, \bar{\sigma}_0^2 = 950$ and the true algorithm parameters were $\mu_0 = 30, \sigma_0^2 = 10^3, \lambda = 0$, and $\nu = 0.5$. The 95% confidence interval implied by the asymptotic normal distribution (9) is shown only in the plot of θ_3. For parameters θ_1 and θ_2, the width of the confidence intervals are much greater than the magnitudes of the parameter estimates and are omitted for legibility.

true value of the parameter is within the 95% confidence interval after 30 observed choices. There are two implications from this result. First, the estimation procedure is at least somewhat robust to the choice of linearization point for this set of algorithm parameters. Second, the estimator is useful for realistic empirical data sets, such as those reported in [26] and studied in the following section. For these data sets, the horizon is $T = 90$ choices. For this amount of data, the simulations show that the estimation procedure can identify the true value of the parameter in a statistically significant way. This result is valuable because the rigorous convergence result from Theorem 1 does not directly guarantee convergence in the more general case of nonlinear objective functions.

The amount of data required to get a reliable estimate can depend on the true value of the algorithm parameters, as shown in Fig. 6. In this case, the true value of the algorithm parameters are $(\mu_0, \sigma_0^2, \lambda, \nu) = (30, 10^3, 0, 0.5)$ and the linearization is made about the point $(\bar{\mu}_0, \bar{\sigma}_0^2) = (40, 950)$. The linearized objective function corresponds to the three parameters taking true values $\theta_1 = 2, \theta_2 = -20$, and $\theta_3 = -1.05 \times 10^{-6}$. With the true values of the prior in the algorithm, the agent is sufficiently uncertain about the rewards and makes most of its initial 100 choices at random in order to gain information about the rewards. This choice behavior results in low performance (specifically, linear regret, see [26] for details). Since the initial choices are effectively made at random, they do not provide useful information about the parameter values (except that they represent some combination of an uncertain prior and high decision noise). The uncertainty in the parameter values can be seen from the width of the confidence interval around the mean parameter estimates shown in Fig. 6. For θ_1 and θ_2 their width is many orders of magnitude larger than the magnitude of the parameter and they are not displayed. For θ_3, the estimate exhibits persistent bias away from the true value, but the width of the associated confidence interval is significantly larger than the bias. Therefore, for such parameter values, one must observe more data to be able to shrink the confidence intervals and provide precise estimates of the parameter values.

E. *Discussion*

The linearization procedure described above yields a local linear approximation to the likelihood maximization problem (4), and Theorem 1 provides conditions under which the local approximation results in an identified model with a convex optimization problem. However, the effectiveness of the procedure is sensitive to the choice of nominal prior $(\bar{\mu}_0, \bar{\delta}_0^2)$ about which to linearize. The linearization point should be chosen such that the linear approximation is valid at the (unknown) true value of the parameters. In the worst case, there might not be any intuition for choosing the linearization point, making the above procedure no better than any other local optimization technique for which a starting point must be chosen.

Fortunately, there are several advantageous aspects of the problem. The first is generic to any heuristic function, which is the fact that the likelihood function forms a unique objective for judging the “goodness” of the estimated parameter. Without knowing in advance a good choice of linearization point, one approach is to perform the estimation assuming two different choices of linearization points and to compare the resulting estimates $\hat{\boldsymbol{\theta}}$. If the two linearization points result in identical estimates there is no conflict, while if the estimates differ, the one with the higher likelihood value is better.

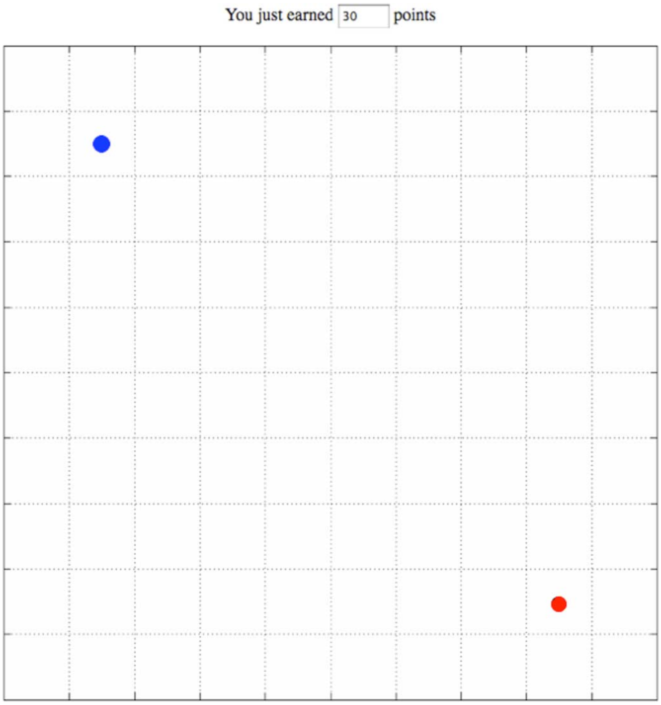
Second, there may be intuition about an appropriate choice of linearization point due to the structure of the model. In [26], we showed that behavior of the stochastic UCL model falls broadly into three classes as a function of the parameters $(\mu_0, \sigma_0^2, \lambda, \nu)$. Thus, by categorizing a given data set into one of the three classes, we narrow the search for a linearization point to the associated regions of parameter space. And, as we saw in Figs. 4 and 5, the stochastic UCL model appears to be relatively insensitive to the choice of linearization point within the region of parameter space associated with a given behavioral class. In the following section, we exploit this intuition to estimate the parameters of the stochastic UCL algorithm based on data from a human subject experiment.

VII. Application to Experimental Data

In this section, we apply the estimator to fit the stochastic UCL model (14) to experimental data studied in [26]. By *fit*, we refer to the process of selecting a nominal parameter for linearization and applying the estimator to the linearized model. The parameter estimates produced by the fitting procedure show that individuals with high performance match their behavior to the task in a statistically significant way.

A. *Experimental Setup*

This section reviews the experimental setup as presented in Reverdy *et al.* [26]. As described in [26], we collected data from a human subject experiment where we ran multi-armed bandit



You just earned 30 points

Fig. 7. The experimental interface used in the human subject experiment. Upon clicking on one of the 100 squares arranged in a 10 × 10 grid, the red dot would move to the center of the square. The subject was free to select a new square without penalty until the time allotted (1.5 or 6 s per choice) had elapsed, at which time the blue dot would move to the center of the selected square and the subject would receive a reward reported in the text box at the top of the screen. Originally appeared as Fig. 5 in [26]; reproduced with permission.

tasks through web servers at Princeton University (Princeton, NJ, USA) following protocols approved by the Princeton University Institutional Review Board. Human participants were recruited using Amazon's Mechanical Turk (AMT) web-based task platform [3]. Participants were shown instructions that told them they would be playing a simple game during which they could obtain points, and that their goal was to obtain the maximum number of total points in each part of the game.

Each participant was presented with a set of $N = 100$ options, presented as squares arranged in a 10×10 grid. See Fig. 7 for a visualization of the experimental interface. At each decision time $t \in \{1, \ldots, T\}$, the participant made a choice by moving the cursor to one square in the grid and clicking. After each choice was made, a numerical reward associated with that choice was reported on the screen. A variety of aspects of the game, including timing, game dynamics, and reward structures, were manipulated as part of the experimental design. As a result of these manipulations, only 326 of the 417 participants were assigned to a standard multi-armed bandit task for which the stochastic UCL model is appropriate. In the remainder of the section, we focus exclusively on data from these 326 participants.

The mean value of the reward associated with choosing a particular option i was m_i. Since the options were arranged in a 10×10 grid, the set of mean values can be thought of as a real-valued function on the discrete two-dimensional grid. We refer to this function as the reward landscape, and prior knowledge about the rewards in a given task corresponds to prior knowledge about the landscape. Mean rewards in each task corresponded to one of two landscapes: Landscapes A and B, shown in Fig. 8. Each landscape was flat along one dimension and followed a profile along the other dimension. The profile of Landscape A was such that a simple gradient-climbing strategy was likely to prove effective, while Landscape B was constructed to require a more sophisticated strategy. Each participant played the game with each landscape once, presented in random order. Due to the structure of the experimental design, only one of the two landscapes was associated with a standard multi-armed bandit task.

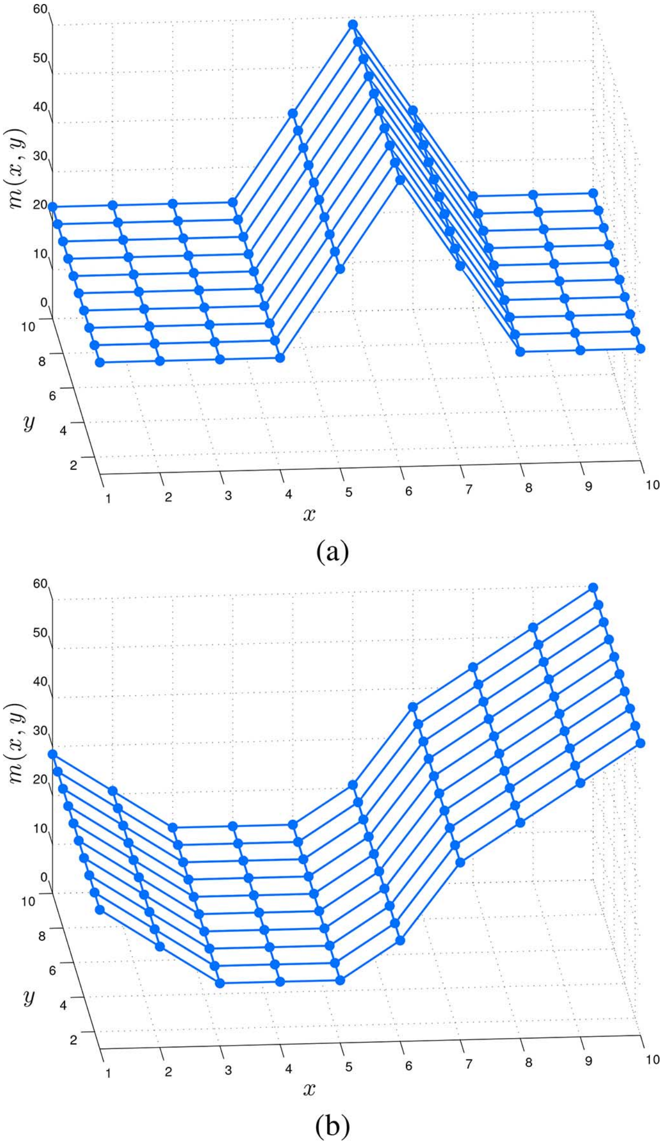


Fig. 8. The two task reward landscapes: (a) Landscape A and (b) Landscape B. The two-dimensional reward surfaces for the 10×10 set of options followed the profile along one dimension (here the x direction) and were flat along the other (here the y direction). The Landscape A profile is designed to be simple in the sense that the surface is concave and there is only one global maximum ($x = 6$), while the Landscape B profile is more complicated since it features two local maxima ($x = 1$ and 10), only one of which ($x = 10$) is the global maximum. Originally appeared as Fig. 6 in [26]; reproduced with permission.

The participants' performance in a given task can be classified in terms of the growth rate of their cumulative regret, which is a measure of cumulative loss relative to the (unknown to the subject) optimal decision. As reported in [26], 70 of the 326 participants, or approximately 21%, achieved high performance, while the remainder, approximately 79%, achieved low performance. Of the 206 subjects assigned to Landscape A, 53 achieved high performance. Likewise, of the 120 subjects assigned to Landscape B, 17 achieved high performance. The high-performing subjects outperformed standard frequentist algorithms on the task, which we attribute to the subjects' having good priors about the task. Since we did not explicitly convey prior knowledge about the reward landscapes to the subjects, we postulate that they used priors developed in the course of other spatial tasks encountered in daily life. Considering the stochastic UCL algorithm as a model of the subjects' behavior, good priors correspond to good values for the parameters $\boldsymbol{\mu}_0$ and Σ_0, which quantify the subjects' intuition about the task. To learn the priors, we propose estimating them from the data. The estimated priors could then be used, e.g., to improve the performance of an automated system.

B. Fitting

In fitting the stochastic UCL model to human subject data, we seek to answer two questions. First, what distinguishes the decision-making of the subjects with high performance from those with low performance? And second, do subjects adapt their decision-making strategies to the task, i.e., the reward landscape? Our experimental design provides data from only one task per subject, so we cannot, for example, compare a single subject's performance on the different landscapes. Thus, we analyze at the population level to answer the two questions.

Each subject is classified as having high or low performance as described above. On the basis of this classification and the reward landscape, the subject is assigned to one of the four performance-landscape combined categories. We assume each subject represents an independent and identically distributed (iid) sample from the true parameter $\boldsymbol{\theta}_0$ associated with its category. We applied the estimator to data from each subject using nominal parameters $(\bar{\mu}_0, \bar{\sigma}_0^2, \lambda) = (30, 10, 0.1)$ for subjects with low performance and $(\bar{\mu}_0, \bar{\sigma}_0^2, \lambda) = (200, 10^6, 4)$ for subjects with high performance. We validated the choice of λ by performing estimation on the data from several subjects using a variety of values of λ. The optimal value of λ clearly differed between the two categories of performance but the estimates for each given performance category were fairly robust to changes in the value of λ. The fitting procedure produces a ML estimate and associated covariance matrix for each subject. By the iid assumption, it is tenable to construct a population-level parameter estimate for each of the four categories by appropriately averaging the individual subjects' estimates and covariances.

Table I presents the population-level parameter estimates, along with the mean log-likelihood values, for the four categories. The columns labeled $\hat{\theta}$ report the ML parameter estimates and those labeled σ their asymptotic standard deviations implied by (9). Recall that these parameters represent deviations from the nominal parameter values and therefore are not directly comparable between performance categories. However, comparing the magnitude of the standard deviations shows that the parameter estimates are much more precise for those categories associated with high performance. This is consistent with our findings in Section VI-D. Table II presents the ML parameter estimates transformed back into the original variables ν, μ_0, and σ_0^2; these are directly comparable.

TABLE I
PARAMETER ESTIMATES $\hat{\theta}$ AND ASSOCIATED STANDARD DEVIATIONS σ CONDITIONAL ON REGRET GROWTH ORDER AND REWARD LANDSCAPE. THE VALUES FOR HIGH PERFORMANCE ARE SIGNIFICANTLY DIFFERENT BETWEEN SURFACES AT THE 95% CONFIDENCE LEVEL (TWO-SIDED WELCH'S t-TEST [36]); OTHER COMPARISONS SHOW THAT THE PARAMETER VALUES DO NOT SIGNIFICANTLY DIFFER BETWEEN CLASSES

Low (linear, power-law) performance			
Landscape A, 153 subj.		Landscape B, 103 subj.	
Mean log likelihood: -338		Mean log likelihood: -331	
$\hat{\theta}$	σ	$\hat{\theta}$	σ
0.360	90.4	0.252	1.32
-5.22	1.27e3	-2.12	51.8
0.433	1.02e2	0.213	8.61
High (log-law) performance			
Landscape A, 53 subj.		Landscape B, 17 subj.	
Mean log likelihood: -273		Mean log likelihood: -271	
$\hat{\theta}$	σ	$\hat{\theta}$	σ
3.93e-2	1.18e-3	3.39e-2	1.04e-3
-6.86	0.226	-6.57	0.268
7.88e-7	2.34e-8	6.80e-7	2.06e-8

TABLE II
PARAMETER ESTIMATES ν, μ_0, σ_0^2 AND ASSOCIATED STANDARD DEVIATIONS σ CONDITIONAL ON REGRET GROWTH ORDER AND REWARD LANDSCAPE

Low (linear, power-law) performance			
Landscape A, 153 subj.		Landscape B, 103 subj.	
Parameter	Value	Parameter	Value
ν	2.78	ν	3.97
μ_0	15.5	μ_0	21.6
σ_0^2	4.54	σ_0^2	5.42
High (log-law) performance			
Landcape A, 53 subj.		Landscape B, 17 subj.	
Parameter	Value	Parameter	Value
ν	25.5	ν	29.5
μ_0	25.3	μ_0	6.08
σ_0^2	3.32e5	σ_0^2	3.35e5

Table II allows us to answer our first question about the differences between subjects with different levels of performance. The parameter values clearly differ more between levels of performance rather than between landscapes. Between levels of performance the parameters that differ the most are the decision noise parameter ν and the prior uncertainty σ_0^2. Larger values of ν are associated with more random decision-making, while larger values of σ_0^2 represent greater uncertainty about the rewards which is associated with placing a higher value on information. Both of these factors tend to encourage exploration, and the values of both ν and σ_0^2 are much greater for subjects with high performance than those with low performance. Thus, for both landscapes, the high-performing subjects explore more than the low-performing ones, which presumably helps them discover the regions of high rewards. Furthermore, the subjects with high performance use correlated priors which allow them to quickly explore large regions of the reward surface.

We can compare the quality of the model fits by comparing the mean log-likelihood values across categories provided on Table I. Again, we see starker differences between levels of performance than between landscapes. Between landscapes, the fits are approximately equal in quality, while between performance levels there is substantial difference, equal to an approximate doubling of the fitted model's predictive power.

Table I allows us to answer our second question about the degree to which subjects match their strategies to the task. We focus on comparing the parameters across landscape conditions for each of the performance categories separately. For low-performing subjects, comparing the relative magnitudes of the parameter estimates and their standard deviations suggests that there is no significant difference between the two landscape conditions. The two-sided Welch's t-test [36] confirms that the difference in the parameter estimates is statistically insignificant. For high-performing subjects, the parameter estimates are much more precise, and the two-sided Welch's t-test confirms that the difference in the parameter estimates is statistically significant at the 95% confidence level. In other words, the fitting procedure is able to distinguish that the high-performing subjects have strategies that are matched to the landscape.

C. Discussion

The results of the fitting exercise demonstrate an estimator for a model of human decision-making behavior. The estimator allows one to quantify a human subject's intuition in a statistically powerful way. We observe that the model fits are of higher quality for subjects with high performance. This suggests that the stochastic UCL model is better suited to the decision-making behavior of subjects who are experts at the task; a different model may be more appropriate for lower performing subjects. We also observe that subjects with high performance seem to have effective priors: these priors have low certainty (large values of σ_0^2), but exploit correlation in the rewards due to the smoothness of the reward landscapes by using positive values of the length scale parameter λ. When such correlation structures exist, they can be exploited to greatly improve performance [31], as our human subjects appear to have done. The estimator provides a way to learn effective priors from a human operator. In the absence of a correlation structure, the above fitting process can still be applied by setting $\lambda = 0$, although convergence of the estimator will be slower, requiring longer series of choice data than those studied here.

By analyzing data from a human subject experiment, we have shown the effectiveness of the linearization procedure for extending the estimator to a model with a nonlinear objective function. The known asymptotic properties of the estimator allowed us to perform tests for statistical significance and find differences in behavior.

VIII. Conclusion

Motivated by the parameter estimation problem for decision-making models, we studied the ML parameter estimation problem for softmax decision-making models with linear objective functions. Such models occur frequently in the neuroscience and machine learning literatures. We derived conditions under which the ML estimator converges on the correct parameter values, characterized the estimator's asymptotic distribution, and showed how to use this distribution to formulate confidence intervals for the parameter estimates.

The estimator convergence results we state in Theorem 1 are specific to the case where the objective function is linear in the unknown parameters. However, we showed how the estimation procedure can be extended to nonlinear objective functions by linearizing about a nominal point in parameter space. We showed that we could estimate the true value of the parameters of the stochastic UCL decision-making algorithm developed in [26]. The amount of data required to perform useful estimation depends on the region of parameter space, with parameters representing priors that strongly influence behavior being easier to estimate. For example, small variances σ_0^2 represent strong beliefs and large correlation length scales λ represent highly structured beliefs.

We then fit the stochastic UCL model to data from a human subject experiment. The estimates show a statistically significant difference in behavior between subjects who exhibit good performance in similar but different tasks. Quantifying this difference is of interest for both the science of decision-making and also for the development of automation. The estimator developed in this paper, as applied to the stochastic UCL model, provides a tool for quantifying human decision-making behavior in multi-armed bandit problems. This tool will facilitate the principled development of human-centered automation systems.

References

[1] D. Böhning, "Multinomial logistic regression algorithm," *Ann. Inst. Statist. Math.*, vol. 44, no. 1, pp. 197–200, 1992.

[2] C. G. Broyden, "The convergence of a class of double-rank minimization algorithms," *IMA J. Appl. Math.*, vol. 6, no. 1, pp. 76–90, 1970.

[3] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?," *Perspectives on Psychological Sci.*, vol. 6, no. 1, pp. 3–5, 2011.

[4] J. D. Cohen, S. M. McClure, and A. J. Yu, "Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration," *Philosph. Trans. Roy. Soc. B: Bio. Sci.*, vol. 362, no. 1481, pp. 933–942, 2007.

[5] J. A. Cooper, M. A. Gorlick, T. Denny, D. A. Worthy, C. G. Beevers, and W. T. Maddox, "Training attention improves decision making in individuals with elevated self-reported depressive symptoms," *Cognitive, Affective Behavioral Neurosci.*, vol. 14, no. 2, pp. 729–741, June, 2014.

[6] N. D. Daw, "Trial-by-trial data analysis using computational models," *Decision Making, Affect, and Learning: Attention and Performance XXIII*, vol. 23, pp. 3–38, 2011.

[7] N. D. Daw, J. P. O'Docherty, P. Dayan, B. Seymour, and R. J. Dolan, "Cortical substrates for exploratory decisions in humans," *Nature*, vol. 441, no. 7095, pp. 876–879, 2006.

[8] R. Fletcher, "A new approach to variable metric algorithms," *The Comput. J.*, vol. 13, no. 3, pp. 317–322, 1970.

[9] K. Gimpel and N. A. Smith, "Softmax-margin training for structured log-linear models," Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-LTI-10-008, 2010.

[10] J. Gläscher, N. Daw, P. Dayan, and J. P. O'Doherty, "States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning," *Neuron*, vol. 66, no. 4, pp. 585–595, 2010.

[11] A. S. Goldberger, *A Course in Econometrics*. Cambridge, MA, USA: Harvard Univ. Press, 1991.

[12] D. Goldfarb, "A family of variable-metric methods derived by variational means," *Math. Comput.*, vol. 24, no. 109, pp. 23–26, 1970.

[13] E. Kaufmann, O. Cappé, and A. Garivier, "On Bayesian upper confidence bounds for bandit problems," in *Proc. Int. Conf. Artif. Intell. Statist.*, La Palma, Canary Islands, Spain, Apr. 2012, pp. 592–600.

[14] S. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[15] B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 957–968, Jun. 2005.

[16] B. Lau and P. W. Glimcher, "Dynamic response-by-response models of matching behavior in rhesus monkeys," *J. Experimental Anal. Behavior*, vol. 84, no. 3, pp. 555–579, Nov. 2005.

[17] D. McFadden, “Conditional logit analysis of qualitative choice behavior,” in *Frontiers Econometrics*, P. Zarembka, Ed. New York: Academic Press, 1974, pp. 105–142.

[18] D. Mitra, F. Romeo, and A. Sangiovanni-Vincentelli, “Convergence and finite-time behavior of simulated annealing,” *Adv. Appl. Probability*, vol. 18, no. 3, pp. 747–771, 1986.

[19] P. R. Montague, B. King-Casas, and J. D. Cohen, “Imaging valuation models in human choice,” *Annu. Rev. Neurosci.*, vol. 29, pp. 417–448, 2006.

[20] M. R. Nassar and J. I. Gold, “A healthy fear of the unknown: Perspectives on the interpretation of parameter fits from computational models in neuroscience,” *PLoS Comput. Bio.*, vol. 9, no. 4, 2013, e1003015.

[21] A. Nedic, D. Tomlin, P. Holmes, D. A. Prentice, and J. D. Cohen, “A decision task in a social context: Human experiments, models, and analyses of behavioral data,” *Proc. IEEE*, vol. 100, no. 3, pp. 713–733, 2012.

[22] W. K. Newey and D. McFadden, “Large sample estimation and hypothesis testing,” in *Handbook of Econometrics*, R. F. Engle and D. L. McFadden, Eds. Philadelphia, PA, USA: Elsevier, 1994, vol. 4, ch. 36, pp. 2111–2245.

[23] A. Y. Ng and S. J. Russell, “Algorithms for inverse reinforcement learning,” in *Proc. Int. Conf. Mach. Learn.*, 2000, pp. 663–670.

[24] V. Ramanujam and H. Balakrishnan, “Estimation of maximum-likelihood discrete-choice models of the runway configuration selection process,” in *Proc. Amer. Control Conf.*, 2011, pp. 2160–2167.

[25] P. Reverdy, “Human-inspired algorithms for search: a framework for human-machine multi-armed bandit problems,” Ph.D. dissertation, Dept. Mech. Aerosp. Eng., Princeton Univ., Princeton, NJ, USA, 2014.

[26] P. Reverdy, V. Srivastava, and N. E. Leonard, “Modeling human decision-making in generalized Gaussian multi-armed bandits,” *Proc. IEEE*, vol. 102, no. 4, pp. 544–571, 2014.

[27] H. Robbins, “Some aspects of the sequential design of experiments,” *Bull. Amer. Math. Soc.*, vol. 58, pp. 527–535, 1952.

[28] S. Russell, “Learning agents for uncertain environments,” in *Proc. 11th ACM Annu. Conf. Comput. Learn. Theory*, 1998, pp. 101–103.

[29] K. Samejima, Y. Ueda, K. Doya, and M. Kimura, “Representation of action-specific reward values in the striatum,” *Science*, vol. 310, no. 5752, pp. 1337–1340, 2005.

[30] D. F. Shanno, “Conditioning of quasi-Newton methods for function minimization,” *Math. Comput.*, vol. 24, no. 111, pp. 647–656, 1970.

[31] V. Srivastava, P. Reverdy, and N. E. Leonard, “Correlated multiarmed bandit problem: Bayesian algorithms and regret analysis,” arXiv:1507.01160v2, 2015.

[32] A. R. Stewart, M. Cao, A. Nedic, D. Tomlin, and N. E. Leonard, “Towards human-robot teams: Model-based analysis of human decision making in two-alternative choice tasks with social feedback,” *Proc. IEEE*, vol. 100, no. 3, pp. 751–775, 2012.

[33] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*. Cambridge, MA, USA: MIT Press, 1998.

[34] The Mathworks, Inc., Fminunc 2015. [Online]. Available: http://www.mathworks.com/help/optim/ug/fminunc.html

[35] C. J. C. H. Watkins and P. Dayan, “Q-learning,” *Mach. Learn.*, vol. 8, no. 3–4, pp. 279–292, 1992.

[36] B. L. Welch, “The generalization of “Student’s” problem when several different population variances are involved,” *Biometrika*, vol. 34, no. 1–2, pp. 28–35, 1947.

[37] R. C. Wilson, A. Geana, J. M. White, E. A. Ludvig, and J. D. Cohen, “Humans use directed and random exploration to solve the explore-exploit dilemma,” *J. Experimental Psychology: Gen.*, vol. 143, no. 6, pp. 2074–2081, 2014.

[38] R. C. Wilson and Y. Niv, “Is model fitting necessary for model-based fMRI?,” in *Proc. Multi-Disciplinary Conf. Reinforcement Learn. Decision Making*, 2013, p. S41.

Paul Reverdy, (M’14) received the B.S. degree in engineering physics and the B.A. degree in applied mathematics from the University of California, Berkeley, Berkeley, CA, USA, in 2007, and the M.A. and Ph.D. degrees in mechanical and aerospace engineering from Princeton University, Princeton, NJ, USA, in 2011 and 2014, respectively.

From 2007 to 2009, he worked as a Research Assistant at the Federal Reserve Board of Governors, Washington, DC, USA. He is currently a Postdoctoral Fellow with the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA. His research interests are in the areas of control and robotics with current interests in human and automated decision making, machine learning, engineering design, and navigation.

Naomi Ehrich Leonard (F’07) received the B.S.E. degree in mechanical engineering from Princeton University, Princeton, NJ, USA, in 1985, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, MD, USA, in 1991 and 1994, respectively.

From 1985 to 1989, she worked as an Engineer in the electric power industry. She is the Edwin S. Wilsey Professor of Mechanical and Aerospace Engineering and Director of the Council on Science and Technology at Princeton University. She is also an associated faculty member of Princeton University’s Program in Applied and Computational Mathematics. Her research and teaching are in control and dynamical systems with current interests in coordinated control for multi-agent systems, mobile sensor networks, collective animal behavior, and human decision-making dynamics.