Motor Learning by Sequential Sampling of Actions

Adrian Haith and John Krakauer, *Dept. Neurology, Johns Hopkins University, Baltimore, MD, USA*

Existing theories of learning, from state space models of adaptation to reinforcement learning (RL), almost unanimously frame motor learning as a process of optimization: identify and exploit the actions that minimize costs and/or maximize rewards. However, human and animal behavior often stands in stark contrast to this principle. For instance, animals free to choose one of several actions that earn rewards of differing magnitude do not identify and then exclusively select the most rewarding action. Rather, they tend to select each action with a frequency that depends on the relative amount of reward it provides – known as the "matching law" [1]. RL theories resort to using artificial "soft-max" action selection rules to rectify this mismatch between theory and empirical observation. A similar situation arises in the execution of individual movements. Movement kinematics vary considerably from trial-to-trial and this variability is largely attributable to central planning mechanisms [2,3] rather than imperfect execution. Both types of variability are often attributed to 'exploration' of alternative actions; a further elaboration of an often already complex learning process. Here we propose an alternative and much simpler model of action selection and learning. **Our core assumption is that the goal of behavior is not to identify and exploit the very best actions, but rather to select actions according to a probability distribution in which more rewarding and/or less costly actions are more likely**. In other words, the goal of behavior is simply to perform well most of the time. We show that this simple premise leads to a simple algorithm for action selection and a rich set of behavioral predictions, including the emergence of exponential learning curves and a relationship between variability and rate of learning.

Suppose a motor task is characterized by a cost function J defined over potential actions u. We assume that the goal of behavior is to select each possible action u with a probability directly related to its cost $p(u) \propto e^{-\beta J(u)}$. Here β is a free parameter that controls the tightness of the distribution relative to the cost function. If the cost is quadratic this leads to a Gaussian distribution. We assume that the sole goal of the subject is to sample a series of actions $u_1, u_2, \ldots$ from this distribution. This goal is complicated by the fact that knowing the absolute probability of each action depends on knowing the value of all possible actions one could take (in order to normalize the probability distribution). Fortunately, a well-known algorithm, Markov chain Monte Carlo (MCMC) [4] (Figure 1), allows this problem to be overcome. The idea behind MCMC sampling is to construct a random sequence of samples (like a random walk) with the property that the asymptotic distribution of these samples matches the target distribution. Critically, this can be achieved without needing to know the absolute probability or cost of an action relative to all other actions. To generate each new action, a new *candidate* action is randomly picked from a local distribution around the previous action, and then either *accepted* or *discarded* based on its value relative to the last accepted action. The sole requirements of this algorithm, therefore, are that the value of an action can be observed by executing it, and that the value of the last-accepted action/value pair be held in memory to compare to the value of the new candidate action.

This algorithm almost trivially leads to a prediction of matching behavior in discrete-action operant conditioning tasks, since achieving matching is basically the premise of the algorithm. However, it does so far more parsimoniously than value-based RL models. More interestingly, this algorithm also gives rise to a theory of learning in the form of the early trials in which actions gradually converge on the target distribution from a weak initial action (Fig. 2A). Applying this algorithm to model behavior in a simple motor learning task (adaptation to a visuomotor rotation [5]) leads to surprisingly rich predictions. In particular, it leads to exponential learning curves (Fig. 2B) and a correlation between error size and amount of learning (Fig. 2D) – typically considered to be hallmarks of supervised learning models such as state space or Kalman filter models of learning. Individual learning curves generated via MCMC show a profound qualitative resemblance to actual subject data [5] (Fig. 2C). Our theory also predicts many features of learning that are unexplained by existing models, including the fact that increased variability is associated with faster learning [6,7] (Fig. 3), and random walk behavior across trials in task-irrelevant dimensions [8].

In summary, sampling-based learning offers a plausible and parsimonious model of learning across many domains. Although a sampling-based mechanism clearly cannot account for all aspects of learning (it cannot, for instance account for model-based contributions to action selection [9-11]), the simplicity and effectiveness of sampling-based learning makes it a computational principle that may well contribute to learning across many different learning systems that are currently modeled with far more complex algorithms. If this proves to be the case, there may be significant implications for how we view the various neural mechanisms that support these behaviors.

A

Algorithm: Action selection by MCMC

Maintain current action u_c and cost $J(u_c)$

1. Sample new action locally around current action

$$Q(u|u_c) = N(u_c, \sigma^2)$$

2. Execute action to measure cost $J(u)$
3. Accept or reject new sample based on relative improvement in cost

$$p(\text{accept}) = \frac{e^{-J(u)}}{e^{-J(u)}+e^{-J(u_c)}}$$

(Metropolis-Hastings rule)

4. If accepted, update current action $u \rightarrow u_c$

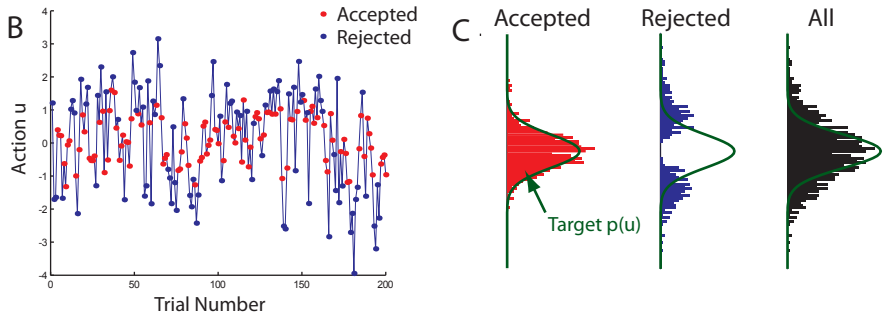


Fig 1. A) Algorithm for sequential sampling of actions based on Metropolis-Hastings MCMC. Cost function is a quadratic centered on 0. B) Example of a sequence of actions generated by this sampling procedure under a quadratic cost (Gaussian distribution). C) Histograms of 'accepted', 'rejected', and both combined. Note that while the MCMC procedure only guarantees that 'accepted' trials will be distributed according to the target distribution, rejected trials do not severely compromise the overall distribution of actions

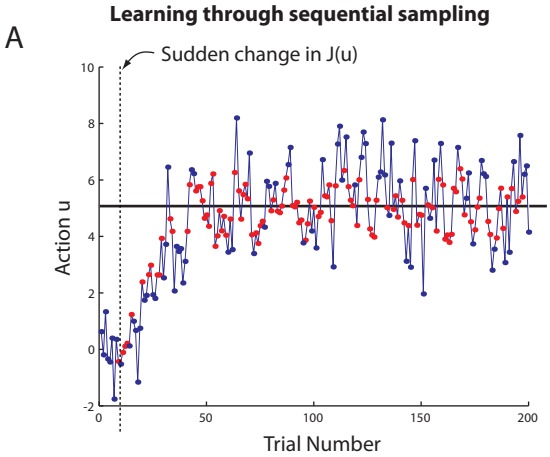
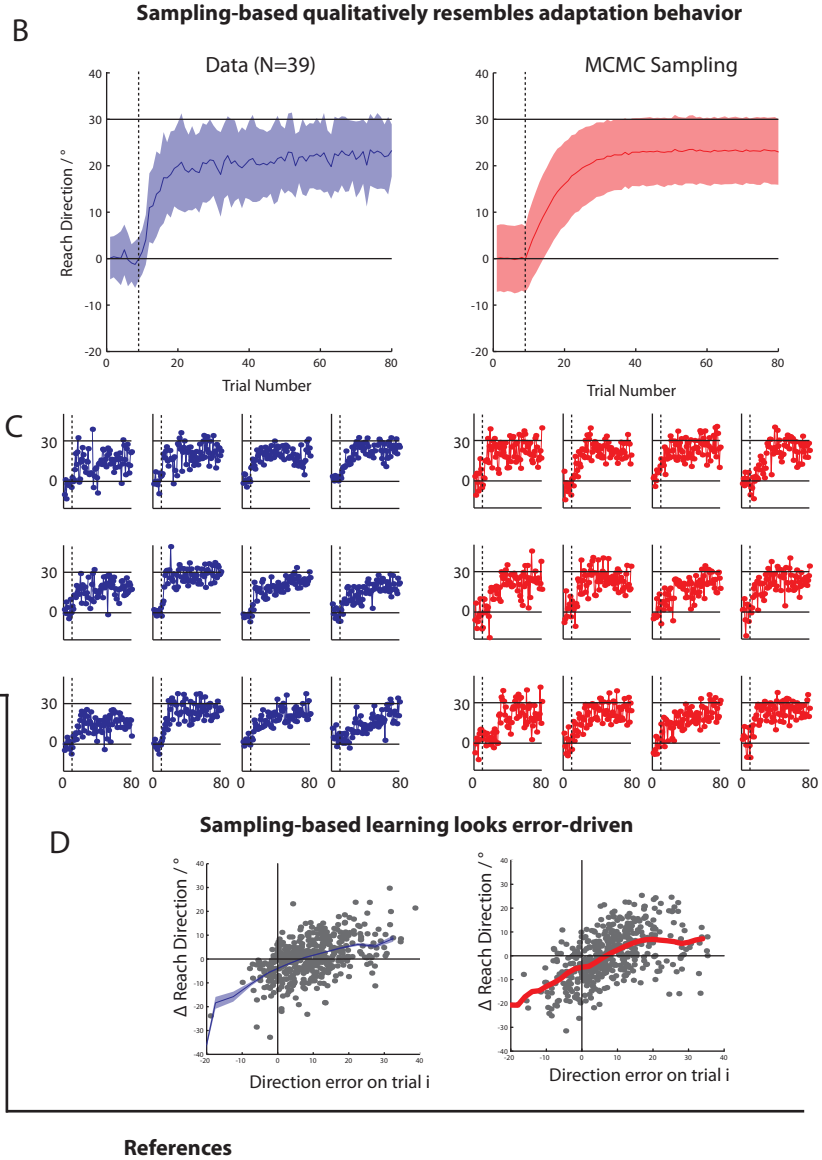


Fig 2. A) MCMC naturally gives rise to learning-like behavior if the initial action sampled is far from the optimum. B) Comparison between MCMC model and data from subjects adapting reaching movements to a 30 deg visuomotor rotation (mean +/- std). C) Individual learning episodes generated via MCMC show a strong qualitative resemblance to individual subject behavior. D) Trial-by-trial changes in reach direction are correlated with error size - often considered a hallmark of error-driven learning.



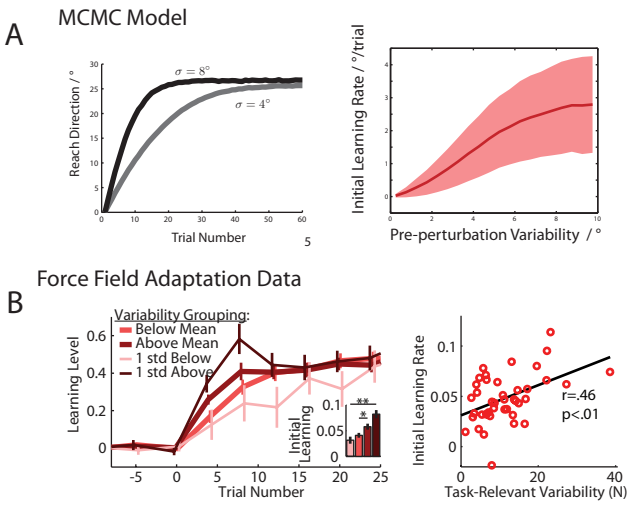


Fig 3. A) MCMC predicts a relationship between baseline variability and learning rate, based on the width of the proposal distribution. B) Data from Wu et al.[6] demonstrating this relationship in force field learning tasks.

References

1 Baum, J Experimental anal behav 22, 231-242 (1974).
2 Churchland et al., Neuron 52, 1085-1096 (2006).
3 Chaisanguanthum et al., J Neurosci 34, 12071-12080 (2014).
4 Chib and Greenberg, The American Statistician 49, 327-335 (1995).
5 Kitago et al., Frontiers in human neuroscience 7, 307 (2013).
6 Wu et al., Nature neuroscience 17, 312-321 (2014).
7 Fernandez-Ruiz et al., Behavioural brain research 219, 8-14 (2011).
8 van Beers et al., Journal of neurophysiology 109, 969-977 (2013).
9 Daw et al., Nature neuroscience 8, 1704-1711 (2005).
10 Dayan, Neural networks 22, 213-219 (2009).
11 Haith and Krakauer Progress in Motor Control VII 782, 1-21 (2013).