Title

One and Done? Optimal Decisions from Very Few Samples

Permalink

https://escholarship.org/uc/item/1g18m250

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 31(31)

ISSN

1069-7977

Authors

Goodman, Noah
Griffiths, Thomas
Tenenbaum, Joshua
et al.

Publication Date

2009

Peer reviewed

One and Done? Optimal Decisions From Very Few Samples

Edward Vul (evul@mit.edu)
Brain and Cognitive Science, 43 Vassar St
Cambridge, MA 02139 USA

Noah D. Goodman (ndg@mit.edu)
Brain and Cognitive Science, 43 Vassar St
Cambridge, MA 02139 USA

Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Dept. of Psychology, Tolman Hall
Berkeley, CA 94720 USA

Joshua B. Tenenbaum (jbt@mit.edu)
Brain and Cognitive Science, 43 Vassar St
Cambridge, MA 02139 USA

Abstract

In many situations human behavior approximates that of a Bayesian ideal observer, suggesting that, at some level, cognition can be described as Bayesian inference. However, a number of findings have highlighted an intriguing mismatch between human behavior and that predicted by Bayesian inference: people often appear to make judgments based on a few samples from a probability distribution, rather than the full distribution. Although sample-based approximations are a common implementation of Bayesian inference, the very limited number of samples used by humans seems to be insufficient to approximate the required probability distributions. Here we consider this discrepancy in the broader framework of statistical decision theory, and ask: if people were making decisions based on samples, but samples were costly, how many samples should people use? We find that under reasonable assumptions about how long it takes to produce a sample, locally suboptimal decisions based on few samples are globally optimal. These results reconcile a large body of work showing sampling, or probability-matching, behavior with the hypothesis that human cognition is well described as Bayesian inference, and suggest promising future directions for studies of resource-constrained cognition.

Keywords: Computational modeling; Bayesian models; Process models; Sampling

Across a wide range of tasks, people seem to act in a manner consistent with optimal Bayesian models (in perception: Knill & Richards, 1996; motor action: Maloney, Trommershauser, & Landy, 2007; language: Chater & Manning, 2006; decision making: McKenzie, 1994; causal judgments: Griffiths & Tenenbaum, 2005; and concept learning: Goodman, Tenenbaum, Feldman, & Griffiths, 2008). However, despite this similarity between Bayesian ideal observers and human observers, two crucial problems remain unaddressed across these domains. First, human behavior often appears to be optimal on average, but not within individual people or individual trials: What are people doing on individual trails to produce optimal behavior in the long-run average? Second, Bayesian inference is straight-forward when considering small laboratory tasks, but intractable for large-scale problems like those that people face in the real world: How can people be carrying out generally intractable Bayesian calculations in real-world tasks? Here we will argue that both of these problems can be resolved by considering the algorithms that people may be using to approximate Bayesian inference.

The first problem is highlighted by an intriguing observation from Goodman et al. (2008) about performance in categorization tasks in which people see positive and negative exemplars of a category and are then asked to generalize any learned rules to new test items. After exposure to several category exemplars people classify new test items consistently with fully Bayesian inference, on average. This average behavior suggests that people consider many possible classification rules, update their beliefs about each one, and then classify new items by averaging the classification over all the possible rules. However, this perfectly Bayesian behavior is only evident in the average across many observers. In contrast, each individual classifies all test items in a manner consistent with only one or a few rules; which rules are considered varies from observer to observer according to the appropriate posterior probabilities (Goodman et al., 2008). Thus, it seems that an individual observer acts based on just one or a few rules sampled from the posterior distribution, and the fully Bayesian behavior only emerges when averaging many individuals, each with different sampled rules.

This sampling behavior is not limited to concept-learning tasks. In many other high-level cognitive tasks, individuals' patterns of response – and sometimes even responses on individual trials – appear to reflect just a small number of samples from the posterior predictive distribution. When predicting how long a cake will bake given that it has been in the oven for 45 minutes (Griffiths & Tenenbaum, 2006), the across-subject variance of responses is consistent with individuals guessing based on only two prior observations of cake baking times (Mozer, Pashler, & Homaei, 2008). When making estimates of esoteric quantities in the world, multiple guesses from one individual have independent error, like samples from a probability distribution (Vul & Pashler, 2008). In all of these cases (and others; e.g., Xu & Tenenbaum, 2007; Anderson, 1991; Sanborn & Griffiths, 2008), people seem to sample instead of computing the "fully Bayesian" answer.

Critics of the Bayesian approach (e.g., Mozer et al., 2008) have suggested that although many samples may adequately approximate Bayesian inference, behavior based on only a few samples is fundamentally inconsistent with the hypothesis that human cognition is Bayesian. Others highlight the second problem and argue that cognition cannot be Bayesian inference because exact Bayesian calculations are computationally intractable (e.g., Gigerenzer, 2008).

In this paper we will argue that acting based on a few samples can be easily reconciled with optimal Bayesian inference and may be the method by which people approximate otherwise intractable Bayesian calculations. We argue that (a) sampling behavior can be understood in terms of sensible sampling-based approaches to approximating intractable inference problems in Bayesian statistics and AI; (b) very few samples from the Bayesian posterior are often sufficient to obtain approximate predictions that are almost as good as predictions computed using the full posterior; and (c) on conservative assumptions about how much time it might cost to produce a sample from the posterior, making predictions based on very few samples (even just one), can actually be the globally optimal strategy.

Bayesian inference with samples

Bayesian probability theory prescribes a normative method to combine information and make inferences about the world. However, the claim that human cognition can be described as Bayesian inference does not imply that people are doing exact Bayesian inference.

Exact Bayesian inference amounts to fully enumerating hypothesis spaces every time beliefs are updated with new data. In any large-scale application, this is computationally intractable, so inference must be approximate. This is the case in "Bayesian" artificial intelligence and statistics, and this must apply even more in human cognition, where the real-world inferences are vastly more complex and responses are time-sensitive. The need for approximating Bayesian inference leaves two important questions. For artificial intelligence and statistics: What kinds of approximation methods work best to approximate Bayesian inference? For cognitive science and psychology: What kinds of approximation methods does the human mind use?

In the tradition of rational analysis, or Marr's computational approach (Marr, 1982), a reasonable strategy to answering the psychological question begins with good answers to the engineering question. Thus, we will explore the hypothesis that the human mind approximates Bayesian inference with some version of the algorithmic strategies that have proven best in AI and statistics, on the grounds of computational efficiency and accuracy.

In artificial intelligence and statistics, one of the most common methods for implementing Bayesian inference is with sample-based approximations. Inference by sampling is a procedure to approximate a probability distribution by repeatedly simulating a simpler stochastic process which produces alternatives from a hypothesis space according to their probability under the distribution in question. The result of any one such simulation is a sample. With sufficiently many samples, these calculations based on these approximations approach the exact calculations using the analytical probability distributions themselves[1]. Sampling methods are typically used because they are applicable to a large range of computational models and are more robust to increasing dimensionality than other approximate methods.

Many cognitively plausible sampling algorithms exist and some specific ones have been proposed to account for aspects of human behavior (Griffiths, Canini, Sanborn, & Navarro, 2007; Levy, Reali, & Griffiths, 2009; Brown & Steyvers, 2008). For our purposes, we need only assume that a person has the ability to draw samples from the hypothesis space according to the posterior probability distribution. Thus, it is reasonable to suppose that people can approximate Bayesian inference via a sampling algorithm, and evidence that humans make decisions by sampling is not in conflict with the hypothesis that the computations they are carrying out are Bayesian.

However, recent work suggests that people base their decisions on very few samples (Vul & Pashler, 2008; Goodman et al., 2008; Mozer et al., 2008) – so few that any claims of convergence to the real probability distribution do not hold. Algorithms using only two samples (Mozer et al., 2008) will have properties quite different from full Bayesian integration. This leaves us with the question: How bad are decisions based on few samples?

Bayesian and sample-based agents

To address the quality of decisions based on few samples, we will consider performance of optimal (fully Bayesian) and sample-based agents in the common scenario of choosing between two alternatives. Many experimental tasks in psychology are a variant of this problem: given everything learned, make a two-alternative forced-choice (2AFC) response. Moreover, real-world tasks often collapse onto such simple 2AFC decisions, for instance: we must decide whether to drive to the airport via the bridge or the tunnel, depending on which route is likely to have least traffic. Although this decision will be informed by prior experiences that produced intricate cognitive representations of possible traffic flow, at one particular junction these complex representations collapse onto a prediction about a binary variable and decision: Should we turn left or right?

Statistical decision theory (Berger, 1985) prescribes how information and beliefs about the world and possible rewards should be combined to define a probability distribution over possible payoffs for each available action (Maloney, 2002; Kording, 2007; Yuille & Bülthoff, 1996). An agent trying to maximize payoffs over many decisions should use these normative rules to determine the expected payoff of each action, and choose the action with the greatest expected payoff[2]. Thus, the standard for decisions in statistical decision theory is to choose the action (A^*) that will maximize expected utility under the posterior predictive distribution over possible world states (S) given prior data (D), assuming that action and state uniquely determine the utility obtained:

$$A^* = \arg\max_A \sum_S U(A;S)P(S|D). \qquad (2)$$

[1] The Monte Carlo theorem states that the expectation over a probability distribution can be approximated from samples:

$$E_{P(S)}[f(S)] \simeq \frac{1}{k}\sum_{i=1}^{k} f(S_i), \text{ when } S_i \sim P(S). \qquad (1)$$

[2] An agent might have other goals, e.g., maximizing the minimum possible payoff (i.e., extreme risk aversion); however, we will not consider situations in which such goals are likely.

If the world state is sufficiently specified, 2AFC decisions map onto a prediction about which of two actions (A_1 and A_2) will have higher expected utility – for instance: will we spend less time in traffic taking the bridge or the tunnel? Therefore, choosing among two alternatives amounts to predicting the outcome of a Bernoulli trial: will A_1 or A_2 have greater utility? Thus, $P(A^* = A_1) \equiv p$ and $P(A^* = A_2) \equiv (1-p)$, and we can simply parameterize these decisions in terms of the posterior predictive probability p. For simplicity, we will consider choosing the higher-utility action a "correct" choice, and choosing the lower-utility action an "incorrect" choice. The fully Bayesian agent will choose the more likely alternative, and will be correct p proportion of the time (we assume p is between 0.5 and 1, given symmetry).

A sample-based agent would sample possible world states, and make decisions based on those sampled world states (S_n):

$$\begin{gathered} A^* = \arg\max_A \sum_{i=1}^{k} U(A;S_i) \\ S_i \sim P(S|D), \end{gathered} \tag{3}$$

so in the case of making predictions between two alternatives one of which may be correct, the sample-based agent should choose the action corresponding to the most frequent outcome in the set of sampled world states. Thus, a sample-based agent drawing k samples will choose a particular outcome with probability q:

$$q = 1 - \Theta_{CDF}(\frac{k}{2}, p, k), \tag{4}$$

where Θ_{CDF} is the binomial cumulative density function and $k/2$ is rounded down to the nearest integer. This sample-based agent will be right with probability $qp + (1-q)(1-p)$.

Good decisions from few samples

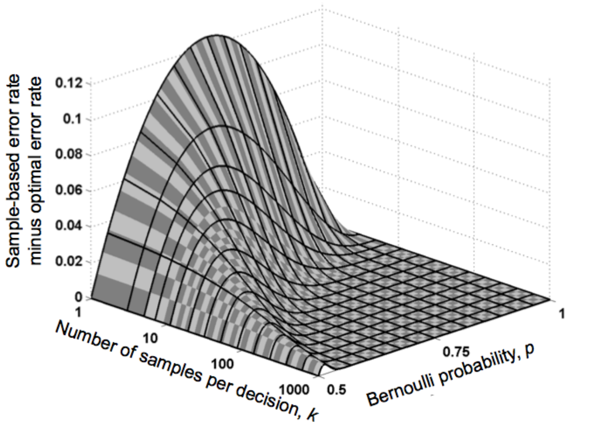

Figure 1: Increased error rate for the sample-based agent over the optimal agent as a function of the Bernoulli trial probability and the number of samples drawn for a decision (decisions based on 0 samples not shown).

So, how much worse will such 2AFC decisions be if they are based on a few samples rather than the "fully Bayesian" inference? Bernoulli estimated that more than 25,000 samples are required for "moral certainty about the true probability of a two-alternative event[3] (Stigler, 1986). Although Bernoulli's calculations were based on different derivations than those which are now accepted (Stigler, 1986), it is undeniable that *inference* based on a small number of samples differs from the "fully Bayesian" solution and will contain greater errors, but how bad are *the decisions* based on this inference?

In Figure 1 we plot the difference in error rates between the sample-based and optimal agents as a function of the underlying probability (p) and number of samples (k). When p is near 0.5, there is no use in obtaining any samples (since a perfectly informed decision will be as likely to be correct as a random guess). When p is 1 (or close), there is much to be gained from a single sample since that one sample will indicate the (nearly-deterministically correct) answer; however, subsequent samples are of little use, since the first one will provide all the gain there is to be had. Most of the benefit of large numbers of samples occurs in interim probability values (around 0.7 and lower).

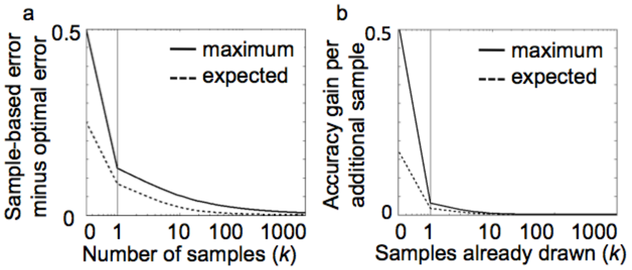

Figure 2: (a) The maximum and expected increase in error for the sample-based agent compared to the optimal agent as a function of number of samples (see text). (b) Expected and maximum gain in accuracy from an additional sample as a function of the number of samples already obtained.

Since the sample-based agent does not know what the true probability p may be for a particular decision we can consider the scenarios such an agent should expect: the average scenario (expectation over p) and the worst case scenario (maximization of the loss over p). These are displayed in Figure 2a assuming a uniform probability distribution over p. The deviation from optimal performance decreases to negligible levels with very few samples, suggesting that the sample-based agent need not have more than a few samples to approximate ideal performance. We can go further to assess just how much is gained (in terms of decreased error rate) from an additional sample (Figure 2b). Again, the vast majority of accuracy is gained with the first sample, and subsequent samples do very little to improve performance.

Thus, even though few samples will not provide a very accurate estimate of p – definitely not sufficient to have "moral

[3] Bernoulli considered *moral certainty* to be at least 1000:1 odds that the true ratio will be within $\frac{1}{50}$ of the measured ratio.

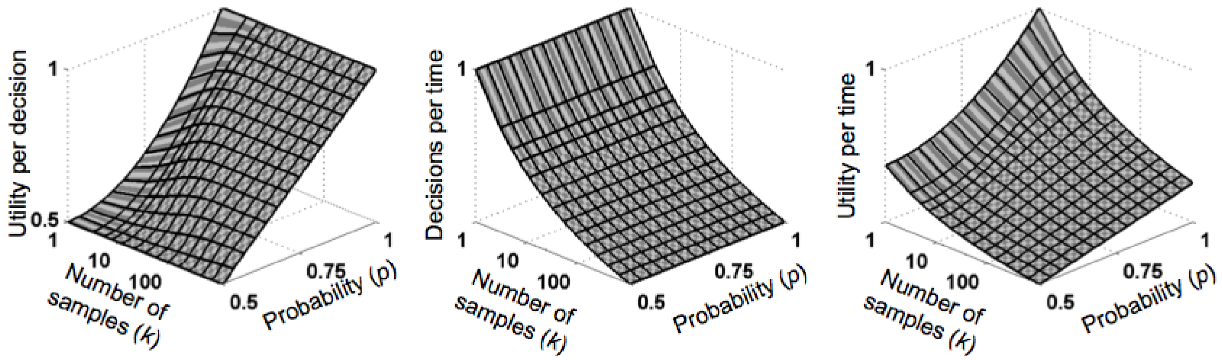

Figure 3: Expected utility per decision, the number of decisions that can be made per unit time, and the expected rate of return (utility per unit time) as a function of the Bernoulli probability and the number of samples (with an example action/sample cost ratio of 232).

certainty" – they *are sufficient to choose an action*: We do not need moral certainty to act optimally.

How many samples for a decision?

Lets take seriously the hypothesis that people make inferences based on samples. If this is the case, how many samples should people use before making a decision? For instance, how many possible arrangements of traffic across the city should we consider before deciding whether to turn left for the tunnel or right for the bridge? Considering one such possibility requires concerted thought and effort – it seems obvious that we should not pause at the intersection for several hours and enumerate all the possibilities. It also seems likely that we shouldn't just turn left or right at random without any consideration. So, how many samples should we take: how hard should we think?

Determining an optimal answer to this meta-cognitive problem requires that we specify how much a sample may "cost"? To be conservative (and for the sake of simplicity), we will assume that a sample can only cost time – it takes some amount of time to conjure up an alternate outcome, predict its value, and update a decision variable.

Thus, if a given sample is free (costs 0 time), then we should take infinitely many samples, and make the best decision possible every time. If a sample costs 1 unit of time, and the *action time* (the time that it would take us to act once we have chosen to do so) is also 1 unit of time, then we should take zero samples we should guess randomly. To make this peculiar result intuitive, lets be concrete: if we have 100 seconds, and the action time is fixed to be 1 second, then we can make 100 random decisions, which will be right 50% of the time, thus giving us an expected reward of $50. If taking a single sample to improve our decision will cost an additional second per decision, then if we take one sample per decision, each decision will take 2 seconds, and we could make at most 50 of them. It is impossible for the expected reward from this strategy to be greater than guessing randomly, since even if 100% of the decisions are correct, only $50 will be gained. Moreover, since 100% accuracy based on one sample is extremely unlikely (this could only arise in a completely deterministic prediction task), substantially less reward should be expected. Thus, if obtaining a sample takes as long as the action, and we do not get punished for an incorrect answer, we should draw zero samples per decision and make as many random decisions as we can. More generally, we can parameterize how much a sample "costs" as the ratio between the time required to make an action and the time required to obtain one sample (action/sample ratio) – intuitively, a measure of how many samples it would take to double the time spent on a decision.

The expected accuracy for a sample-based agent (previous section) gives us the expected utility per decision as a function of k (the number of samples) and p (the Bernoulli trial probability; Figure 3a), and the utility function. We consider two utility functions for the 2AFC case: *no punishment* – correct: gain 1; incorrect lose 0) and *symmetric* – correct: gain 1; incorrect: lose 1. Given one particular action/sample time ratio, we can compute the number of decisions made per unit time (Figure 3b). Multiplying these two functions together yields the expected utility per unit time (Figure 3c).

Since p is unknown to the agent, an ideal k must be chosen by taking the expectation over p. This marginalization (assuming a uniform distribution over p) for many different action/sample time ratios is displayed in Figure 4. It is clear that as samples become cheaper, one is best advised to take more of them converging to the limit of infinitely many samples when the samples are free (the action/sample time ratio is infinity).

In Figure 5 we plot the optimal number of samples as a function of the action/sample time ratio. Remarkably, for ratios less than 10, one is best advised to make decisions based on only one sample if the utility function is symmetric. Moreover, with no punishment for incorrect answers, the action/sample time ratio must be 2 or greater before taking any samples (making a guess thats at all educated, rather than fully random) becomes a prudent course of action. Thus, under some assumptions about how much it costs to think, making guesses based on very few samples (e.g., one) is the best course of action: Making many locally suboptimal decisions quickly is the globally optimal strategy.

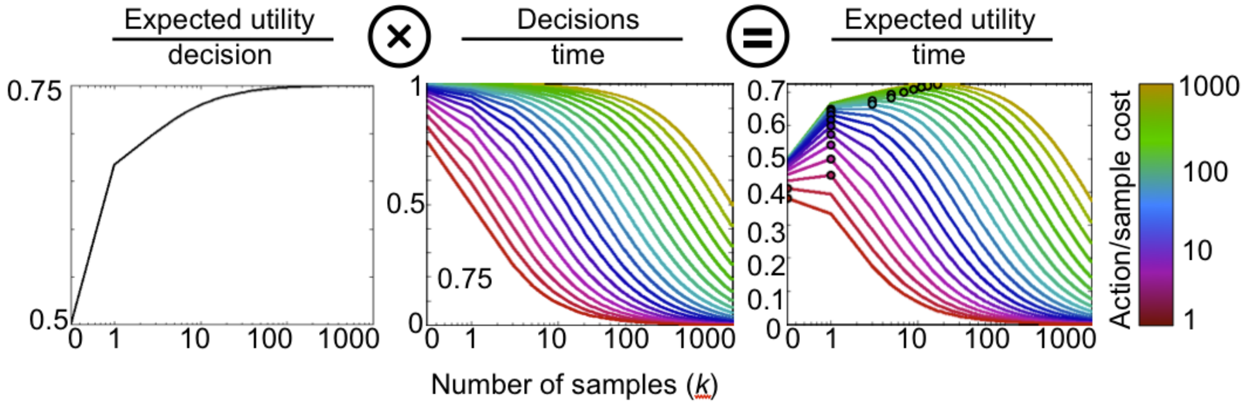

Figure 4: Expected utility per decision, number of decisions per unit time, and expected utility per unit time (rate of return) as a function of the number of samples and action/sample cost ratios. Action/sample cost ratios are logarithmically spaced between 1 (red) and 1000 (yellow). In the last graph the solid circles indicate the optimal number of samples at that action/sample cost ratio. (The utility function used for this figure contained no punishment for an incorrect choice and +1 for a correct choice.)

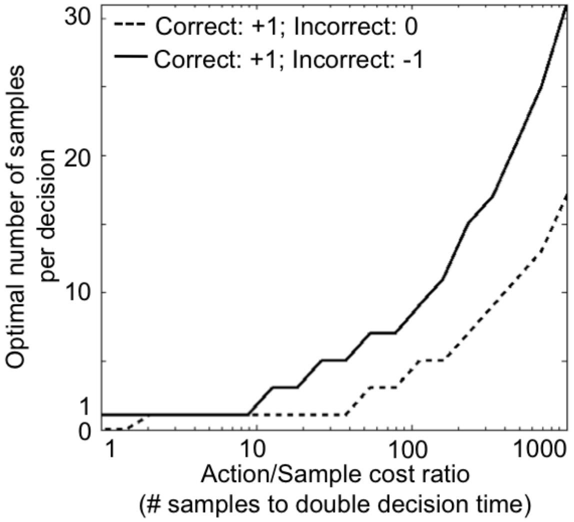

Figure 5: The optimal number of samples as a function of the action/sample time-cost ratio for each of two utility functions (symmetric – correct: +1, incorrect: -1; and no punishment for incorrect answers – correct: +1, incorrect: 0).

Discussion

We began with the observation that, on average, people tend to act consistently with ideal Bayesian inference, integrating information to optimally build models of the world; however, locally, they appear to be systematically suboptimal, acting based on a very limited number of samples. This has been used to argue that people are not “fully Bayesian” (Mozer et al., 2008). Instead, we have argued that sample-based approximations are a powerful method for implementing approximate Bayesian inference. Although with few samples, sample-based inferences will deviate from *exact* Bayesian inference, we showed that for two-alternative forced-choice tasks, a decision based on a very small set of samples is nearly as good as an optimal decision based on a full probability distribution. Moreover, we showed that given reasonable assumptions about the time it takes to produce an exact sample, a policy of making decisions based on very few samples (even just one) is globally optimal, maximizing long-run utility.

In this paper we considered sample-based decisions about predictions of variables that had not been previously observed – predictions computed through Bayesian inference over latent variables and structures in the world. However, a large prior literature on “probability matching” (Herrnstein, 1961; Vulkan, 2000) has studied a very similar phenomenon in a simpler task. In probability matching, subjects predict the outcome of a trial based on the relative frequencies with which that outcome has been observed in the past. Thus, subjects have direct evidence of the probability that lever A or lever B should be pulled, but they do not seem to maximize; instead, they “probability match” and choose levers with a frequency proportional to the probability of reward. On our account, this literal “probability matching” behavior amounts to making decisions based on one sample, while decisions based on more samples would correspond to Luce choice decisions with an exponent greater than 1 (Luce, 1959).

“Probability matching” to previously observed frequencies is naturally interpreted as sampling prior events from memory. This interpretation is consistent with recent work suggesting that people make decisions based on samples from memory. Stewart, Chater, and Brown (2006) suggested that a policy of making decisions through binary preference judgments among alternatives sampled from memory can account for an assortment of human judgment and decision-making errors. Similarly, Schneider, Oppenheimer, and Detre (2007) suggest that votes from sampled orientations in multidimensional preference space can account for violations of coherent normative utility judgments. A promising direction for future research would be to relate models suggesting that samples are drawn from cognitive processes such as memory, to models like those we have described in our paper, in which samples are drawn from probability distributions reflecting ideal inferences about the world.

How much might a sample "cost"? A relevant measure of sample cost in multiple-trial experiments is the ratio between the time it takes to make an action and go on to the next trial and the time required to draw a sample to inform a decision about that action. Ratios near 10 seem quite reasonable: most experimental trials last a few seconds, and it can arguably cost a few hundred milliseconds to consider a hypothesis. Of course, this is speculation. However, in general, it seems to us that in most experimental tasks, the benefits gained from a better decision are relatively small compared to the costs of spending a very long time thinking. So, if thinking amounts to sampling possible alternatives before making a decision, it should not be surprising that people regularly seem to use so few samples.

We should emphasize that we are not arguing that all human actions and decisions are based on very few samples. The evidence for sampling-based decisions arises when people make a decision or a choice based on what they think is likely to be true (Which example is in the concept? How long will this event last? How many airports are there in the US?). In other situations people appear to integrate over the posterior, or to take many more samples, such as when people make graded inductive judgments (How similar is A to B? How likely is it that X has property P given that Y does? How likely do you think that F causes G?). It is interesting to consider why there may be a difference between these sorts of decisions.

Under reasonable two-alternative choice scenarios, people are best advised to make decisions based on few samples (future work will extend this to n-AFC and continuous choice decisions). This captures a very sensible intuition: when we are deciding whether to turn left or right at an intersection, we should not enumerate every possible map of the world. We do not need "moral certainty" about the probability that left or right will lead to the fastest route to our destination – we just need to make a decision. We must implicitly weigh the benefits of improving our decision by thinking for a longer period of time against the cost of spending more time and effort deliberating. Intuition suggests that we do this in the real world: we think harder before deciding whether to go north or south on an interstate (where a wrong decision can lead to a detour of many miles), than when we are looking for a house (where the wrong decision will have minimal cost). Empirical evidence confirms this: when the stakes are high, people start maximizing instead of "probability matching" (Shanks, Tunney, & McCarthy, 2002). Nonetheless, it seems that in simple circumstances, deliberating is rarely the prudent course of action – for the most part, making quick, locally suboptimal, decisions is the globally optimal policy: one and done.

Acknowledgments. This work was supported by ONR MURI (JBT, PI: Bavelier); Russell Sage, NDSEG, NSF Dissertation grant (EV).

References

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

Brown, S., & Steyvers, M. (2008). Detecting and predicting changes. *Cognitive Psychology*, *58*, 49–67.

Chater, N., & Manning, C. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*(7), 335–344.

Gigerenzer, G. (2008). *Rationality for mortals: How people cope with uncertainty*. Oxford University Press, USA.

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Griffiths, T., Canini, K., Sanborn, A., & Navarro, D. (2007). Unifying rational models of categorization via the hierarchical dirichlet process. In *Proceedings of the 29th annual conference of the cognitive science society* (pp. 323–328).

Griffiths, T., & Tenenbaum, J. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354 – 384.

Griffiths, T., & Tenenbaum, J. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.

Herrnstein, R. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, *4*(3), 267.

Knill, D., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge: Cambridge University Press.

Kording, K. (2007). Decision theory: What should the nervous system do? *Science*, *318*(5850), 606.

Levy, R., Reali, F., & Griffiths, T. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In *Advances in neural information processing systems* (Vol. 21).

Luce, R. (1959). *Individual choice behavior*. New York, NY: Wiley.

Maloney, L. (2002). Statistical decision theory and biological vision. In D. Heyer & R. Mausfield (Eds.), *Perception and the physical world* (pp. 145–189). New York, NY: Wiley.

Maloney, L., Trommershauser, J., & Landy, M. (2007). Questions without words: A comparison between decision making under risk and movement planning under risk. In W. Gray (Ed.), *Integrated models of cognitive systems* (pp. 297–313). New York, NY: Oxford University Press.

Marr, D. (1982). *Vision*. Cambridge: MIT Press.

McKenzie, C. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and bayesian inference. *Cognitive Psychology*, *26*, 209–239.

Mozer, M., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science*, *32*, 1133-1147.

Sanborn, A., & Griffiths, T. (2008). Markov Chain Monte Carlo with People. In *Advances in neural information processing systems* (Vol. 20). MIT Press.

Schneider, A., Oppenheimer, D., & Detre, G. (2007). Application of Voting Geometry to Multialternative Choice. In *Proceedings of the 29th annual conference of the cognitive science society* (pp. 635–640).

Shanks, D., Tunney, R., & McCarthy, J. (2002). A re-examination of probability matching and rational choice. *J. Behav. Dec. Making*, *15*, 233–250.

Stewart, N., Chater, N., & Brown, G. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26.

Stigler, S. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge: Harvard University Press.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*(1), 101–118.

Xu, F., & Tenenbaum, J. (2007). Sensitivity to sampling in bayesian word learning. *Developmental Science*, *10*(3), 288–297.

Yuille, A., & Bülthoff, H. (1996). Bayesian decision theory and psychophysics. In D. Knill & W. RIchards (Eds.), *Perception as bayesian inference* (pp. 123–161).