Choosing a Good Toolkit, II: Bayes-rule Based Heuristics*

Alejandro Francetich[1] and David Kreps[2]

[1]*University of Washington, Bothell*
[2]*Graduate School of Business, Stanford University*

September 3, 2019

Abstract

We study heuristics for a class of complex multi-armed bandit problems, the period-by-period choice of a set of objects or "toolkit" where the decision maker learns about the value of tools within the chosen toolkit. This paper studies heuristics that involve a decision maker who employs Bayesian inference. Analytical results are combined with simulations to gain insights into the relative performance of these heuristics. We depart from the extensive bandit-learning literature in computer science and operations research by employing the discounted-expected-reward formulation that stresses the importance of the classic exploration–exploitation tradeoff. A companion paper, Francetich and Kreps (2019), studies a variety of prior-free heuristics.

Keywords: Heuristics, multi-armed bandits, behavioral decision making

*Assistance from David Aldous, Lanier Benkard, Hans Föllmer, Gerd Gigerenzer, Michael Harrison, Guido Imbens, Daniel Russo, Edward Schlee, Benjamin Van Roy, and anonymous referees, as well as comments by seminar participants at Stanford University, Bocconi University, Princeton University, NYU, and the University of Washington Bothell, are gratefully acknowledged, as is the financial support of ERC Advanced Grand 32419 and the Stanford Graduate School of Business. This paper extends results obtained in Chapter 3 of the Ph.D. thesis of the first author.

1 Introduction

When hiring, the manager of a firm may consider hiring multiple candidates at once, at least probationally. If she is uncertain about the candidates' productivity and she can only observe their performance on the job, if hired, then the hiring problem becomes a multi-armed bandit problem where the arms are the different teams of potential employees. These arms are not independent on two grounds: (a) Learning about the productivity of a candidate might provide valuable information about others with similar qualifications and experience; and (b) even if productivity is independent across candidates, the productivity of two different teams with common members will not be. Thus, the well-known Gittins-index solution for bandits with independent arms (Gittins and Jones, 1974) does not apply.

The general problem of sequentially choosing subsets of some set when their distribution of value is unknown can be formulated as a dynamic-programming problem. However, except in very special cases, such problem is unsolvable—either analytically or numerically—when the set from which we choose is "large" (say, has four or more elements).[1] When real economic agents face problems with this structure, we posit that they employ heuristics or rules of thumb. Francetich and Kreps (2019) analyzes simple heuristics that only employ accumulated data. The present paper examines more-sophisticated heuristics that employ the decision maker's prior assessment of the problem and incorporate Bayesian updating.

Heuristics for multi-armed bandit problems have been extensively analyzed in a literature that spans computer science and operations research (CS-OR) under the rubric of *bandit learning*. Various categories of bandit problems with non-independent arms are investigated, including linear bandits, Gaussian bandits, and smooth bandits. (For an introduction to this literature focused on Thompson Sampling, see Russo et al., 2018.) The heuristics range from relatively simple to sophisticated and employ both Bayesian and classical inference methods. The paper closest to ours is Sauré and Zeevi (2013), in which the arms of the bandit are subsets of a set of products that a retailer can display to each of a finite set of sequentially-arriving customers.

Our paper departs from the bandit-learning literature in the criterion employed to assess the heuristics. The typical criterion in CS-OR is the minimization of expected undiscounted asymptotic regret. Roughly speaking, the decision-maker's regret in any period is defined as the difference between what she would receive were she clairvoyant—namely, if she knew the distribution—and what she actually receives by employing a specific heuristic. She seeks to minimize the expected value of the undiscounted sum of her period-by-period regret. This criterion biases her search among heuristics towards those that will learn the truth (or, at least, enough of the truth so that her within-period regret is eventually zero), and only if that

[1] Francetich (2018) analyzes a special continuous-time formulation where the choice set contains two elements. Even this simple formulation becomes unsolvable with four or more tools, involving higher-dimensional state spaces.

is assured does she consider the speed and cost of the learning process.[2] Instead, we employ the criterion of maximizing the expected sum of discounted rewards, which is more standard in the economics literature. As is well known for independent-arm bandit problems, as long as future rewards are discounted, there is always positive probability that the decision maker *optimally* settles for an *objectively suboptimal* arm. This means that the optimal strategy under discounted rewards gives infinite asymptotic expected regret.

The optimal strategy is practically unobtainable in our problem, but the fact that it fares poorly (relative to alternatives) under the expected asymptotic regret criterion suggests that, for a decision maker who discounts rewards, prescriptions of the bandit-learning literature must be carefully considered. The direct message of this paper is that this is so: Under discounted rewards, bandit-learning heuristics can perform badly compared to heuristics that are based on the considerations of the classic exploitation–exploration tradeoff.

Specifically, we formulate a stylized model of this type of problem and examine six Bayes-rule based heuristics, two of which—Thompson Sampling and Upper Confidence Bounds—are generally "winners" in the bandit-learning literature. A few theoretical results about the long-run behavior of these heuristics are given, which explain why Thompson Sampling and Upper Confidence Bounds do well for infinitely-patient decision makers. However, when shorter-run costs are taken into account due to discounting, we see in simulations that these long-run-excellent heuristics can fall short in comparison to heuristics that take more seriously the exploitation–exploration trade-off.

As in the companion paper, we do not claim that we have identified the ultimate list of heuristics. We motivate our heuristics either by their simplicity, desirable asymptotic properties, or performance in simulations. This leads us to the broader, less-direct message of this research: While simulations are more problem-specific than theorems, we propose that carefully examined simulation results can provide valuable insights in settings where more-formal analysis—including the identification of payoff bounds, the standard approach in CS-OR—is precluded by complexity.[3]

The rest of the paper is organized as follows. Section 2 presents the formulation of the problem. Section 3 recounts standard results on the optimal solution to the decision-maker's problem. In Section 4, we describe the six Bayes-rule based heuristics. Section 5 presents asymptotic results, while Section 6 turns to simulations. Section 7 concludes. Proofs are relegated to the appendix. A second, online appendix with supplementary material including R-language and additional data from simulations is available on the website of the first author.[4]

[2]Minimizing undiscounted expected regret goes back at least to the classic paper of Lai and Robbins (1985) and may be original to that paper.

[3]We are not the only members of the economic-theory literature to borrow from the CS-OR literature and to resort to simulation. Fudenberg and He (2018) employs simulation to compare the performance of Thompson Sampling and Upper Confidence Bounds with their proposed Player-Compatible Equilibrium in a link-formation game.

[4]https://www.uwb.edu/business/faculty/afrancetich

2 Formulation

The basic elements of the formulation are the same as in Francetich and Kreps (2019). Each date $t = 0, 1, \ldots$, a decision maker (she) chooses a subset or *toolkit* K_t from a finite set X of *tools*. The state of the world at date t, v_t, is drawn from some finite set V; the sequence $\{v_t\}$ is i.i.d. with unknown distribution μ. The decision maker's immediate net reward or payoff in period t is denoted by $W(v_t, K_t)$.

The choice of toolkit also affects how much information the decision maker collects. For each $K \subseteq X$, there is a partition $\Gamma(K)$ of V; if she chooses K_t at time t, she observes which cell of the partition $\Gamma(K_t)$ contains v_t. We assume that: (a) $\Gamma(X)$ is the finest partition, so $K_t = X$ allows the decision maker to observe v_t; (b) $\Gamma(\emptyset)$ is the coarsest partition, so $K_t = \emptyset$ produces no data in t; (c) if $K \subseteq K'$, then $\Gamma(K')$ is a (weak) refinement of $\Gamma(K)$, so K' is (weakly) more informative than K; and (d) for each K, $W(v, K)$ is constant on each cell in $\Gamma(K)$, meaning that immediate payoff is observable.

In the present paper, unlike in its companion, the decision maker exploits her prior belief over μ, which we denote by π_0. She evaluates outcomes according to the *subjective expectation*, namely the expectation under the distribution of states induced by π_0, of the normalized discounted sum of immediate payoffs with discount factor $\delta < 1$: $(1-\delta)\sum_{t=0}^{\infty} \delta^t W(v_t, K_t)$.

We posit that π_0 has finite support $\{\mu_1, \ldots, \mu_N\}$, where the hypotheses μ_n are assumed to be distinct; the prior probability that μ_n is the true distribution, $\mu_n = \mu$, is $\pi_0(\mu_n)$. We also assume that μ is in the support of the prior; in other words, that μ is one of the μ_n. Therefore, whatever outcome she observes is, almost surely, never a complete surprise. Consequently, she employs Bayesian inference; π_t denotes her posterior distribution at t (computed on the basis of data collected up to but excluding t), and $\pi_t(\mu_n)$ is the corresponding posterior probability that the true distribution is μ_n. We sometimes write π without subscripts to refer to generic beliefs (prior or posteriors).

Notation and language

The following notation is used. For each toolkit K and hypothesis μ_n, $w_n(K)$ denotes the objective expected immediate reward generated by K under μ_n: $w_n(K) = \sum_{v \in V} \mu_n(v) W(v, K)$; maximizing across toolkits yields the maximum (objective) expected reward under μ_n, $w_n^* = \max\{w_n(K) : K \subseteq X\}$, and the set of objectively optimal toolkits for μ_n, $\mathcal{K}_n^* = \{K \subseteq X : w_n(K) = w_n^*\}$. Similarly, for each K and belief π, $w(\pi, K)$ is the decision maker's subjective expected immediate reward from choosing K: $w(\pi, K) = \sum_{n=1}^{N} \pi(\mu_n) w_n(K)$; its maximum across toolkits, the highest myopic reward, is $w^*(\pi) = \max\{w(\pi, K) : K \subseteq X\}$; finally, the corresponding set of myopically optimal toolkits is $\mathcal{K}^*(\pi) = \{K \subseteq X : w(\pi, K) = w^*(\pi)\}$.

We will be making a variety of probabilistic statements concerning what happens as time passes, such as "If the decision maker uses Heuristic **H**, then with probability 1 she will learn which of her initial hypotheses about μ is the correct hypothesis." Being formal about such statements requires specifying a probability space on which the random variables of interest

(such as π_t) are defined. Note, in this regard, that the heuristic employed affects the probability measure that governs the dynamics of such random variables. However, formalizing such things makes the expositional flow quite convoluted, so we will adopt the following shorthand: **P** denotes probability statements based on the probability measure generated by the heuristic under discussion at the time and her prior beliefs; $\mathcal{N} \in \{1, \ldots, N\}$ denotes the index of the true hypothesis, so that $\mu_{\mathcal{N}} = \mu$; and $\mathbf{P}^{\mathcal{N}}$ denotes probability statements based on μ. The expectation operators **E** and $\mathbf{E}^{\mathcal{N}}$ correspond to the distributions **P** and $\mathbf{P}^{\mathcal{N}}$, respectively.

The probabilistic statement in the previous paragraph is seemingly ambiguous because it could be referring to either **P** or $\mathbf{P}^{\mathcal{N}}$. However, since **P** attaches positive (prior) probability to $\mathbf{P}^{\mathcal{N}}$, any statement that has **P** probability 1 must have $\mathbf{P}^{\mathcal{N}}$ probability 1. When we use the language that an event has probability 1 without any further qualification, it will be implicit that this means **P**-probability 1, which implies $\mathbf{P}^{\mathcal{N}}$-probability 1.

Finally, let $\mathcal{K}^*$ denote $\mathcal{K}^*_{\mathcal{N}}$, the set of objectively optimal toolkits, and let w^* denote $w^*_{\mathcal{N}}$, the value of the objectively optimal toolkits or "clairvoyance" value—the highest expected immediate reward for a decision maker who knows μ. Of course, w^* is greater than the value of the optimal strategy for our decision maker.

A special case and independence

For discussions and simulations, we employ the following special formulation introduced in Francetich and Kreps (2019): (a) The states of nature are the profiles of tool values: $v_t = (v_t(x))_{x \in X} \in (R_+)^X$, where $v_t(x)$ is the period-t value of tool x; (b) There is a vector of strictly positive "rental costs" $c = (c_x)_{x \in X}$; (c) Immediate payoff is given by the function $W^{\text{MAX}}(v, K) := \max\{v(x) : x \in K\} - \sum_{x \in K} c_x$.

In general, we make no assumption about independence of $v_t(x)$ and $v_t(x')$ for $x' \neq x$. In the context of this paper, independence can occur (or fail) at two levels. It could be that the values of the various tools are independent under each hypothesis μ_n but that the decision maker's uncertainty about μ allows for dependence in her (subjective) assessment.[5] It could also be that, even under her assessment, learning about the value of tool x provides no information about other tool. She entertains, for each tool x, hypotheses $\{\mu^x_n : n = 1, \ldots, N^x\}$ about the distribution of $v_t(x)$, and her prior π^x_0 concerning which of these hypotheses is true about tool x is independent of her priors on the hypotheses for other tools. (So, in this case, the number of full priors that she has is $N^1 \times \cdots \times N^x$.) We will refer to this second case as *full independence*. Under full independence, the toolkit chosen in a given period allows for no learning whatsoever about tools left behind on that date.

[5]Imagine that there are two tools, x and x', whose values are either 0 or 1: $v(x), v(x') \in \{0, 1\}$. The decision maker entertains two hypotheses, μ_1 and μ_2. In either case, $v(x)$ and $v(x')$ are independent. Under μ_1, we have $v(x) = 1$ with probability $3/4$ and $v(x') = 1$ with probability $1/4$; these probabilities are reversed under μ_2. Despite their (objective) independence, a high frequency of 1's for $v(x)$ suggests to the decision maker that μ_1 is likely true, which in turn leads her to expect a higher frequency of 0's for $v(x')$; vice versa for μ_2.

While independence may not be entirely natural in some applications, our simulations feature a full-independence case. This formulation stands in stark contrast to our other simulated model in which there is a great deal of dependence in the value-processes of different tools that the decision maker can exploit.

3 The optimal solution

Because of our finiteness assumptions, an optimal strategy for our decision maker is guaranteed to exist, and can (in theory) be found by either policy or value iteration. However, for all but the simplest of specifications, finding a solution is impractical: The "state space" of the corresponding dynamic programming problem consists of all (obtainable) posteriors over $\{\mu_1, \ldots, \mu_N\}$ and the so-called "curse of dimensionality" intrudes. Nonetheless, we can say some things about the solution, beyond the fact that one exists.

Proposition 1. *(a) Consider the following strategy: At dates t of the form $t = 2^k$ for $k = 0, 1, 2, \ldots$, the decision maker chooses $K_t = X$; at all other dates, she chooses $K_t \in \mathcal{K}^*(\pi_t)$ (employing π_0 at $t = 0$). Then, with probability one, π_t converges to a point mass on $\mu_{\mathcal{N}}$ and she eventually chooses $K_t \in \mathcal{K}^*$ for all dates t except (perhaps) for those of the form $t = 2^k$.[6] Hence, the Cesàro averages of her per-period payoffs almost surely converge to w^*. (b) Write $u^\star(\delta, \pi_0)$ for the value of the problem as a function of δ and π_0; that is, $u^\star(\delta, \pi_0)$ is the maximized value of the subjective expectation of $(1-\delta)\sum_{t=0}^{\infty} \delta^t W(v_t, K_t)$, where we maximize over all feasible strategies. Then,*

$$\lim_{\delta \uparrow 1} u^\star(\delta, \pi_0) = w^*.$$

(c) For a fixed $\delta < 1$, let π_t be the decision maker's posterior at time t when she employs her optimal strategy. Then, $\{\pi_t : t = 0, 1, \ldots\}$ converges with probability 1. If we denote this a.s. limit by π^∞, the decision maker eventually chooses only toolkits $K_t \in \mathcal{K}^*(\pi^\infty)$.

Readers familiar with the literature on multi-armed bandits will recognize part (a) as a standard result. If the decision maker is interested in optimizing her average (expected) reward rather than the discounted sum of rewards, she can adopt any strategy that: 1) samples every "arm" infinitely often, using the data so generated to learn (almost surely) the expected return from each arm, while 2) choosing whichever arm is myopically optimal based on information gathered so far a proportion of the time that approaches one. In our specific problem, choosing the toolkit X infinitely often generates all the information needed; the posterior will almost surely converge to a point mass on $\mu_{\mathcal{N}}$.[7] As π_t converges

[6] "Eventually" means "for all dates $t \geq t^*$ for some date t^*."

[7] There is a subtle issue here. DeFinetti's Theorem tells us that π_t based *only* on outcomes at dates of the form $t = 2^k$ converges to a point mass on μ. But, perhaps, the outcomes at all other dates—and there are many more of them—might lead her astray. Lemmas 3 and 4 in the appendix show that this will almost surely not happen.

to a point mass on μ, eventually we will have $w(\pi_t, K) < w(\pi_t, K')$ for all $K \notin \mathcal{K}^*$ and $K' \in \mathcal{K}^*$. Therefore, almost surely, she eventually picks $K_t \in \mathcal{K}^*$ on a frequency of dates that approaches 1.

For part (b): w^* is an obvious upper bound on the value of the problem, namely the maximum feasible expected (normalized discounted) payoff for the decision maker. The optimal strategies (for each δ) must do at least as well as the strategy in (a), whose expected value approaches w^* (as $\delta \uparrow 1$).

Finally, for part (c): Convergence of the posterior beliefs involves a simple invocation of the martingale convergence theorem. Then, use the argument employed for part (a).

4 Bayes-rule-based Heuristics

We examine and compare six Bayes-rule-based heuristics. In all cases, each date, the decision maker uses whatever information has come her way in the past to update her beliefs and make a choice. Whenever a heuristic recommends choosing from a set that contains multiple toolkits, we assume that the decision maker picks arbitrarily.[8]

Adaptive Myopia (AM). *At each date t, choose a myopically optimal toolkit: $K_t \in \mathcal{K}^*(\pi_t)$. (At $t = 0$, choose some $K_0 \in \mathcal{K}^*(\pi_0)$.)*

Harmonic Sampling (HS). *At each date t, K_t is selected randomly: With probability $t/(t+1)$, choose some $K_t \in \mathcal{K}^*(\pi_t)$; and with probability $1/(t+1)$, choose $K_t = X$.*

A decision maker employing AM ignores the exploration-exploitation tradeoff, choosing instead to maximize her myopic payoff. This does not imply that she never learns: She uses whatever information comes her way by updating her beliefs. However, any information that she gets arrives serendipitously. HS "fixes" this by choosing X as the toolkit at random times, where those times are arranged so that X is sampled infinitely often with probability one but with vanishing frequency. Of course, many similar fixes are available; for instance, X could be chosen at an infinite deterministic set of dates, but with vanishing frequency, in the spirit of Proposition 1(a). The names "simulated annealing" and "ϵ-greedy" are associated in the literature with strategies in the spirit of HS (ϵ-greedy, however, with a small but non-vanishing probability of experimentation at each date).[9]

The next two heuristics are very popular in the CS-OR literature, for reasons that will become clear. The first was originally proposed in Thompson (1933) and, therefore, is almost certainly the seminal heuristic of its general type.

[8] Whenever the decision maker must make arbitrary choices to break ties, the tie-break rule she employs at a given period cannot involve information that she does not yet possess. This issue is further discussed in Section 5, under the rubric of "predictability," and in the appendix; see also the appendix of Francetich and Kreps (2019).

[9] See, for instance, Tokic and Palm (2011).

Thompson Sampling (TS). *For each μ_n, fix (arbitrarily) some toolkit $K_n^* \in \mathcal{K}_n^*$.*[10] *At each date t, choose $K_t = K_n^*$ with probability $\pi_t(\mu_n)$.*

The story that seemingly motivates TS may seem a bit forced, but for what it is worth: At date t, the decision maker believes that $\mathcal{N} = n$ with probability $\pi_t(\mu_n)$. So she "simulates" which hypothesis is true, selecting μ_n with its probability of being true, and then selects a best toolkit according to the outcome of this simulation. (We'll provide a better rationale for this heuristic in a few paragraphs.)

The fourth heuristic that we study is an upper-confidence-bound heuristic. To motivate it, imagine that, at the beginning of time, the decision maker had the power to pick which of the hypotheses $\mu_1, \ldots, \mu_N$ will serve as μ. She would compare the (objective) expected values of the optimal toolkits under each hypothesis and set as μ the one that attains the highest of these values. Then, she would proceed to choose the corresponding optimal toolkit daily. In symbols, she would set $\mu = \mu_n$ for n such that $w_n^* = \max\{w_1^*, w_2^*, \ldots, w_N^*\}$ and then choose $K_t \in \mathcal{K}_n^*$ for every t. In reality, of course, she does not have this power. Nonetheless, she might be "optimistic" and think that nature has chosen μ this way, in her favor. Such optimism, however, cannot (reasonably) be sustained for hypotheses against which enough evidence has been accumulated.

This optimism is the idea behind the next heuristic. Given some $\epsilon > 0$, let $\mathcal{P}_\epsilon(\pi) \subseteq \{1, \ldots, N\}$ be the set of hypotheses whose probability, as assessed by π, is greater than ϵ: $\mathcal{P}_\epsilon(\pi) = \{n : \pi(\mu_n) > \epsilon\}$. In words, $\mathcal{P}_\epsilon(\pi)$ is the set of hypotheses that are "sufficiently plausible" under the decision maker's belief π given the "plausibility criterion" ϵ.

The Bayesian Upper-Confidence-Bound Heuristics (BUCB). *Fix some $\epsilon > 0$ and some $K_n^* \in \mathcal{K}_n^*$ for each μ_n. At time t, choose $K_t = K_n^*$ for n such that $w_n^* = \max\{w_m^* : m \in \mathcal{P}_\epsilon(\pi_t)\}$.*

BUCB is a "Bayesian" version of the UCB heuristic analyzed, for instance, in Francetich and Kreps (2019). In the reinforcement-learning style of UCB, the decision maker chooses among all toolkits the one that provides the "best plausible" return measured as the right-hand endpoint of a classical confidence interval of that toolkit's payoff. BUCB employs the same logic adapted to this Bayesian setting. The decision maker rules out as implausible any hypothesis μ_n whose posterior probability is less than ϵ and computes the value of the optimal toolkits under each of the remaining plausible hypothesis; then, she chooses the toolkit with the highest best-but-still-plausible return.

TS and BUCB are popular in the CS-OR literature because, with some caveats in the case of BUCB, they are guaranteed to get to the truth, eventually. (We provide precise statements of these results in the next section.) The same is true of HS, but TS and BUCB get to the truth more efficiently (more quickly and at smaller cost) than does HS. Therefore, for the objective of minimizing undiscounted asymptotic regret, they fare well.

[10] Fixing one K_n^* from each $\mathcal{K}_n^*$ simplifies the proof of Proposition 3; however, we do not believe it is necessary.

The reason they do well, at least in our simulations, is somewhat orthogonal to how we have motivated them. The key in each case is that the decision maker limits herself to choosing toolkits that are optimal for some one of the hypotheses μ_n. This limited selection of toolkits is particularly useful in *dis*confirming hypotheses: Roughly speaking, if K_n^* is chosen infinitely often and $\mathcal{N} = n$, then K_n^* would yield w_n^* on average (when chosen); *if it does not do so*, the decision maker learns that $\mathcal{N} \neq n$. This is reminiscent of Sauré and Zeevi (2013), who provide conditions under which the retailer can easily and quickly rule out certain products as suboptimal.

In the case of BUCB, where the decision maker chooses optimistically among plausible K_n^*, either her optimism about n is eventually rewarded or she learns that her optimism was misplaced; μ_n becomes implausible, and she tries the next-most-optimistic K_m^*. The only way this can fail is if $\mu_{\mathcal{N}}$ is deemed implausible from the start (and evidence does not lead it to become plausible) or if, through bad luck, it comes to seem implausible; the prior probability of both of these events goes to zero as ϵ goes to zero. In the case of TS, the argument is roughly that, if $\mathcal{N} \neq n$, then the posterior probability on μ_n must fall to zero—otherwise, K_n^* would be chosen infinitely often, which would show that $\mathcal{N} \neq n$—and so K_n^* is rarely chosen. This argument does not quite work because of the possibility of ties, but what is true is that the decision maker learns the truth *insofar as she cares about it* with probability one and, as she learns about it, she is choosing toolkits that are more likely to be good choices for her with higher probability. (Again, formal statements are provided in the next section.)

In both TS and BUCB, however, decisions made early on hold the possibility of being disastrously wrong. Neither heuristic considers how K_n^* does if $\mathcal{N} = m$ for some $m \neq n$, namely how the optimal toolkit under one hypothesis fares under a different one. This is perhaps most obvious for BUCB. Suppose there are two hypotheses, μ_1 and μ_2, each with prior probability $1/2$; w_1^* is slightly greater than w_2^*, so BUCB (with $\epsilon < 1/2$) recommends choosing K_1^* until the posterior probability on μ_1 falls below ϵ, if it ever does. Now, suppose that $w_1(K_2^*)$, the expected reward from K_2^* under μ_1, is much larger than $w_2(K_1^*)$. Prudence would suggest starting with K_2^*; starting with K_1^* will produce terrible results under μ_2, which has with prior probability $1/2$, while starting with K_2^* if $\mathcal{N} = 1$ is much less bad.[11] This suggests the following variation where we search among optimal toolkits K_n^* but with a view to maximizing the subjective expected immediate payoff they yield, $w(\pi, K_n^*)$.

The Alternative Bayesian Upper-Confidence-Bound Heuristics (BUCBx). *Fix some $\epsilon > 0$ and some $K_n^* \in \mathcal{K}_n^*$ for each μ_n. At time t, choose $K_t = K_n^*$ for n such that $w(\pi_t, K_n^*) = \max\{w(\pi_t, K_m^*) : m \in \mathcal{P}_\epsilon(\pi_t)\}$.*

In words, choose among the objectively optimal toolkits corresponding to sufficiently plausible hypothesis the one that yields the highest myopic payoff. We will see in the next

[11] For TS, in this specific case, the choice of K_1^* about half the time gives poor results if $\mathcal{N} = 2$; a better strategy would be to choose K_2^* with probability, say, 0.9 at the start, since it performs well even if $\mathcal{N} = 1$.

section that this modification *partially* hurts the long-run performance of the heuristic vis-à-vis BUCB. But for δ bounded away from one, it can (and, in our simulations, it does) improve performance, at least for some test problems.

For both BUCB and BUCBx as defined, the possibility exists that, at some point, $\pi_t(\mu_n) \leq \epsilon$ for all of the hypotheses μ_n. Were such a situation to arise, we could (by convention) set $K_t = \emptyset$. Alternatively, as what we do in our simulations, we could fall back on TS, having the decision maker choose $K_t = K_n^*$ with probability $\pi_t(\mu_n)$. We can always prevent this issue altogether by setting $\epsilon < 1/N$, as we do for the theoretical results of next section.

The five prior-based heuristics discussed so far are insensitive to the discount factor δ. That is, the value of δ has no impact on the choice of toolkit at time t; all that matters is the sequence of toolkits previously chosen and the results they generated. Except for AM, information is (more or less) consciously sought; at least, the decision maker is making some concession to exploration versus exploitation. But the *value* of information, which is greater the closer δ is to 1, does not factor into the tradeoff. For HS, BUCB, and BUCBx, we could try to incorporate a δ-based tradeoff into the story: For HS, increase the probability that X is chosen at any stage as δ increases; for BUCB and BUCBx, decrease ϵ—that is, demand more evidence before deeming a hypothesis to be implausible—as δ increases. In fact, we could do something similar for TS, letting the probability that K_n^* is chosen at time t under π_t be $\phi_\delta(\pi_t(\mu_n))/\left[\sum_{n'} \phi_\delta \pi_t(\mu_{n'}))\right]$ for some family of concave functions $\{\phi_\delta(\cdot) : \delta \in (0,1)\}$ where $\delta' > \delta$ implies that $\phi_{\delta'}(\cdot)$ is a concave transformation of $\phi_\delta(\cdot)$.

However, consider the following possibility. A tool $x \in X$ is such that, for all $v \in V$ and all toolkits K such that $x \notin K$, $W(v, K \cup \{x\}) < W(v, K)$. For instance, in the case $W = W^{\mathrm{MAX}}$, suppose $v(x) < c_x$ for all v; no matter what v is, tool x is worth less than it costs. Nonetheless, the cost of carrying x is small and x is highly informative about v: $v(x) \neq v'(x)$ for all $v \neq v'$, so that x tells the decision maker exactly which v has occurred at a low cost. Then, especially for δ close to 1, the decision maker may wish to add this "worse-than-useless-for-immediate-purposes" tool to her toolkit, at least early on, for its informative value. Yet, except for HS, tool x will never be chosen by one of our heuristics. The direct value-of-information calculation that is an important part of the optimal strategy is in no way a part of AM (of course), TS, BUCB, or BUCBx. Our final heuristic has the decision maker doing limited value-of-information calculations, calculations that would lead her rather naturally to include such a tool in her early toolkits.

Let $u^0(\pi)$ the highest discounted payoff the decision maker can attain when her belief is π and *she must choose one toolkit that she will employ for the rest of time, no matter what (more) she may learn from it*:

$$u^0(\pi) = \frac{w^*(\pi)}{1-\delta}.$$

Define iteratively, for $m = 1, 2, \ldots$,

$$u^m(\pi) = \max_{K \subseteq X} \left\{ w(\pi, K) + \mathbf{E}^\pi [\delta u^{m-1}(\widetilde{\pi})] \right\},$$

where $\widetilde{\pi}$ represents the random posterior the decision maker will assess based on π and the information she receives from choosing K. In words, $u^m(\pi)$ is the value function derived from solving the *m-step finite-horizon dynamic programming problem*, where the last decision taken is to choose whichever toolkit is myopically optimal given the (then-held) posterior assessment and to stick with that choice for the rest of time. We denote by $\mathcal{K}^{*m}(\pi)$ the set of toolkits that attain the maximum for $u^m(\pi)$: $\mathcal{K}^{*m}(\pi) = \{K \subseteq X : w(\pi, K) + \mathbf{E}^{\pi}[\delta u^{m-1}(\widetilde{\pi})] = u^m(\pi)\}$.

By standard results in dynamic programming for this sort of problem (bounded per-period reward, discounting), we know that $u^m(\pi)$ converges to the value function of the infinite-horizon problem as $m \to \infty$, and (as long as ties are broken consistently) $K^{*m}(\pi) \in \mathcal{K}^{*m}(\pi)$ "settles down" to an optimal toolkit given π. So, if we could perform the calculations for large m, there would be no point to this paper. However, as m grows large, for most problems of this sort, the computations become too many and too complex. So how about carrying out these computations for small m and doing what is recommended?

The Approximate-Dynamic-Programming (ADP) heuristics.[12] *Pick a (relatively) small positive integer m. At time t, choose $K_t = K^{*m}(\pi_t)$.*

By relatively small m, we mean $m = 1$ or, perhaps, 2. Since this heuristic depends on the value of m, we use ADP1 for ADP with $m = 1$ and ADP2 for ADP with $m = 2$. As a practical matter, for even a moderate size problem (four tools, four v-vectors, six hypotheses), implementing and simulating ADP2 is outside our capabilities. When we turn to simulations, we'll always be working with ADP1, which is equivalent to the Knowledge Gradient (KG) algorithm for online learning under discounting (Powell and Ryzhov, 2012; Ryzhov et al., 2019).[13]

5 Asymptotic Behavior of the Heuristics

We can say little about the performance of these heuristics for δ bounded away from one without resorting to simulations of test problems. But we can provide some theoretical results concerning how they behave "in the long run," which translates into results about how they perform for δ close to one.

We state our asymptotic results in the fashion of the *Strong Law of Large Numbers* (SLLN): The probability of some event is 1, or close to 1, or approaches 1. As previously discussed, this is with reference to both the decision maker's prior *and* the true distribution, at least insofar as the decision maker attaches positive probability to the latter. Our proofs involve a

[12] The term *approximate dynamic programming* is due to Bertsekas, who has written extensively on this sort of heuristic for decision making. See, for instance, Bertsekas (2012).

[13] A different heuristic based on value-of-information computations is Information Directed Sampling (IDS; Russo and Van Roy, 2016). In an undiscounted setting, under IDS, the toolkit recommended each period is the one that minimizes squared regret normalized by the reduction in entropy of the posterior distribution of the objectively optimal toolkit.

technical restriction on how the decision maker randomizes, as in both HS and TS, and on how she breaks ties: Such choices must be "predictable" in the sense of employing only pre-t data at period t. (The appendix provides formal details.)

There is no guarantee that AM or ADP1 (or, for that matter, ADPm for any fixed m) will find the objectively optimal toolkit with probability one, regardless of the value of δ. Consider, for instance, the following simple specification: $W = W^{\text{MAX}}$; there are two tools, $X = \{x, x'\}$, and two v vectors, $V = \{v_1, v_2\} = \{(5, 10), (5, 1)\}$, where $v_1 = (5, 10)$ means $v_1(x) = 5$ and $v_1(x') = 10$; there are two hypotheses, $\mu_1(v_1) = 0.1$ and $\mu_1(v_2) = 0.9$, and $\mu_2(v_1) = 0.9$ and $\mu_2(v_2) = 0.1$; and the costs of the tools are 1 apiece. The values of $w_n(K)$ are shown in Table 1.

Suppose that $\pi_0(\mu_1)$ is 0.9. A decision maker who employs AM computes her immediate expected reward as 0 from $\emptyset$, 4 from $\{x\}$, $0.9 \times 0.9 + 8.1 \times 0.1 = 1.62$ from $\{x'\}$, and $0.9 \times 3.5 + 0.1 \times 7.5 = 3.9$ from $\{x, x'\}$, so she chooses $K_0 = \{x\}$. In so doing, she learns nothing when she sees $v_0(x) = 5$ and, therefore, at time 1, she will (again) choose $\{x\}$, and so on. Hence, with probability 0.1, she is forever picking an objectively suboptimal toolkit.

For a decision maker who employs ADP1, we need a prior probability $\pi_0(\mu_1)$ that is much closer to 1. Suppose that $\pi_0(\mu_1) = 0.99$. Then, a decision maker who chooses $\emptyset$ or $\{x\}$ for K_0 learns nothing; her posterior will equal her prior. But if she chooses either $\{x'\}$ or $\{x, x'\}$, she either observes $v_0(x') = 1$ or $v_0(x') = 10$. In the first case, her posterior is $\pi_1(\mu_1) > 0.9988$, while in the second case, she assesses $\pi_1(\mu_1) > 0.9166$. But, even in the second case, if she must then make a once-and-for-all choice of toolkit, she chooses $\{x\}$. Hence, in employing ADP1, her optimal choice at time 0, and therefore at every time, is $\{x\}$ *no matter how close* δ *is to 1*. This means that, with this prior, ADP1 fails to generate the objectively optimal choice for all time with probability 0.01.

Of course, the failure to find the objectively optimal toolkit with positive probability need not be subjectively suboptimal for fixed $\delta < 1$. It is easy to prove that, for a fixed $\delta < 1$ and a prior $\pi_0(\mu_1)$ sufficiently close to 1, the optimal solution to the problem is to stick with $\{x\}$ forever. However, one can also show that for a given prior $\pi_0(\mu_1) < 1$, this does not happen if δ is close enough to 1. (This is, essentially, a corollary to Proposition 1(b).) This example illustrates that, for ADPm for some fixed m, there are priors $\pi_0(\mu_1) < 1$ such that one gets stuck with the "wrong" toolkit with probability $1 - \pi_0(\mu_1)$ regardless of how close δ is to 1. And while the numbers in this example are, perhaps, extreme, it is not difficult to have more complex examples in which the presence of a tool which is good-all-purpose in application but which provides no information about other, more speculative

	∅	{x}	{x'}	{x,x'}
μ_1	0	4	0.9	3.5
μ_2	0	4	8.1	7.5

Table 1. Values of $w_n(K)$. For each of the four toolkits, the expected value under the two hypotheses are given.

tools, produces this phenomenon for ADP1 under reasonable parameter values. As we will see in the simulations, a similar phenomenon can arise when the myopically optimal toolkit produces some information but does not change the assessments about one or more tools not contained in it.

None of this can happen with HS:

Proposition 2. *If heuristic HS is employed, then (a) the decision maker learns the value of $\mathcal{N}$ and (b) the Cesàro sum of her payoffs converges to w^*, both with probability 1. Hence, as $\delta \to 1$, the decision maker's reward under HS approaches w^*.*

For TS, there is no guarantee that the decision maker will learn $\mathcal{N}$. But the outcome for her (asymptotically) is just as good as if she did.

Proposition 3. *If the decision maker employs TS, the Cesàro sums of $W(v_t, K_t)$ approach w^* with probability one. Hence, as $\delta \to 1$, the decision maker's reward under TS approaches w^*.*

The stories for BUCB and BUCBx are more complicated. Recall that, for these heuristics, we had the decision maker fix some $K_n^* \in \mathcal{K}_n^*$. In the first part of the next proposition, we introduce the following additional condition: For any two hypotheses μ_n and μ_m, if the (random) immediate reward from the optimal toolkit under μ_n, $W(v_t, K_n^*)$, has the same distribution under μ_n as it does under μ_m, then K_n^* is also objectively optimal under μ_m.

Proposition 4. *(a). Suppose that the following condition holds: For each pair of hypotheses μ_n and μ_m, and for K_n^*, if $W(v_t, K_n^*)$ has the same distribution under μ_n as it does under μ_m, then $K_n^* \in \mathcal{K}_m^*$. Then, if the decision maker employs either BUCB or BUCBx with $\epsilon < 1/N$, the Cesàro sums of $W(v_t, K_t)$ approach w^* with probability one. Hence, as $\delta \to 1$, the decision maker's reward under BUCB or BUCBx approaches w^*. (b) If the condition in part (a) does not hold, and if the decision maker employs BUCB with $\epsilon < 1/N$, define P_ϵ as the probability assessed by the decision maker ex ante that the limit of the Cesaro sums of her rewards will converge to the objectively optimal value:*

$$P_\epsilon = \mathbf{P}\left(\left\{\lim_{T\to\infty} \frac{1}{T+1}\sum_{t=0}^{T} W(v_t, K_t) = w^*\right\}\right).$$

Then, $\lim_{\epsilon\downarrow 0} P_\epsilon = 1$. Hence, as $\delta \to 1$, by selecting ϵ that goes to zero as $\delta \to 1$, the decision maker's expected payoff approaches w^*.

Part (b) of Proposition 4 only works for BUCB; it is quite definitely false for BUCBx. It is easiest to explain why (and provide a concrete example) after going through the proof, so we leave this for the appendix.

It is worth pointing out that our assumption that $\Gamma(K')$ is a weak coarsening of $\Gamma(K)$ when $K' \subseteq K$ is unnecessary for these propositions; we need (only) that $\Gamma(X)$ is the discrete partition of V for Proposition 2 and that, for each π, the decision maker is capable of computing $\mathcal{K}^*(\pi)$.

Propositions 3 and 4 begin to indicate why TS and BUCB are popular in the CS-OR literature where the criterion is minimization of asymptotic regret, although they are not dispositive on this score. Proposition 2 says that HS gets to the full truth with probability 1, a stronger result than Propositions 3 and 4, yet HS provides infinite expected asymptotic regret as long as there is positive probability that X is not the optimal toolkit: While HS eventually learns the truth, it does not adapt its "exploration" to what it learns and instead "explores," choosing X, infinitely often. TS and BUCB get to as much of the truth as is needed (in the case of BUCB, if the condition in Proposition 4(a) holds) and do so efficiently; the CS-OR literature— for instance, Russo et al./ (2018) for TS—derives bounds (in a somewhat different setting) on their speed of convergence.

The proofs in the appendix indicate why TS and BUCB are efficient in our setting. Both of them are effective in ruling out hypotheses that are not true. If K_n^* is chosen "frequently" and μ_n is not true, this will be revealed: Employing Bayes rule drives $\pi_t(\mu_n)$ to zero. In the case of TS, this means that K_n^* is chosen "infrequently" for any $n \neq \mathcal{N}$; in the case of BUCB, when $\pi_t(\mu_n)$ falls below ϵ, K_n^* is no longer chosen (unless $K_n^* = K_m^*$ for some $m \neq n$). Note in this regard that while TS (roughly speaking) is designed to disprove all hypotheses that are not the truth (or, more accurately, that give rewards less than those given by the truth), BUCB only disproves those that are false and that, if true, would provide better rewards than does the truth. As we will see in our simulations, this makes BUCB "faster."

6 Simulations

The results of Section 5 indicate long-run strengths of HS, TS, and (subject to a qualification) BUCB and BUCBx, which in turn implies that they will be close to optimal as $\delta \to 1$. But this is no guarantee with how they do relative to one another or relative to AM or ADP for a particular $\delta < 1$. To see how they do, we see no alternative to setting test problems and simulating the different heuristics. Of course, the choice of test problem can color the results derived from simulations. But the simulations—if the test problems are simple enough to allow interpretation of the results—can shed some light. So, we report on some simulations on test problems that, while too complex to solve fully, are still simple enough that yield some insight into the strengths and weaknesses of the heuristics.

We simulate twelve different scenarios: two "models" $\times$ two cost levels $\times$ three discount factors. The basic structure of these scenarios is the same as in the simulations in Francetich and Kreps (2019), although the models in the present paper specify and make use of the decision maker's prior. Immediate payoff is $W^{\text{MAX}}(v_t, K_t) = \max\{v_t(x) : x \in K_t\} - \sum_{x \in X} c_x$.

There are four tools; on a given day, each tool is either "useful" or "not useful." If tool 1 is useful, its value is $v_t(1) = 15.1$; otherwise, $v_t(1) = 1.1$. Similarly, $v_t(2) = 15.2$ or 1.2, $v_t(3) = 15.3$ or 1.3, and $v_t(4) = 15$ or 1. All tools cost the same, either 2 or 3 apiece. The discount factors we consider are $\delta = 0.9, 0.96$, and 0.99.

In the first model, Model 1, one and only one of the four tools is useful on any given day. Thus, V consists of four vectors: $(15.1, 1.2, 1.3, 1)$—tool 1 is useful; $(1.1, 15.2, 1.3, 1)$—tool 2 is useful; $(1.1, 1.2, 15.3, 1)$—tool 3 is useful; and $(1.1, 1.2, 1.3, 15)$—tool 4 is useful. The decision maker entertains six hypotheses, which are listed in Table 2(a) along with their prior probability. Under each hypothesis, the probability of being useful on any given day is 0.4 for each of two of the tools and only 0.1 for each of the other two tools. We use the term "good" to refer to a tool that has a higher probability of being useful. The hypotheses are labelled ν_{XY}, where X and Y denote the two good tools. For instance, under hypothesis ν_{14}, tools 1 and 4 are good while tools 2 and 3 are not good (or bad); the decision maker assesses probability 0.18 that this hypothesis is true. Table 2(b) provides the (objectively) optimal toolkit and its expected value under each hypothesis; panel (c) gives the myopically optimal toolkit under the prior and its expected value for the two different cost levels, as well as the clairvoyance value, w^*.

In Model 2, the tools are fully independent. Under each hypothesis, whether a tool is useful on any given day is independent of the usefulness of the other tools. Therefore, V consists of 16 vectors: $(1.1, 1.2, 1.3, 1)$—no tool is useful; $(15.1, 1.2, 1.3, 1)$—only tool 1 is useful; $(15.1, 15.2, 1.3, 1)$—tools 1 and 2 are useful; and so on. Similarly, each tool is either good or not independently of the goodness of other tools, so there are 16 hypotheses; we label them μ_{WXYZ}, where the subscript can have up to four characters and indicates which tools are good. For instance, under μ_{13}, tools 1 and 3 are good; under μ_{1234}, all four tools are good. We use μ_0 to denote the hypothesis where none of the tools are good.

Full independence means that the model is specified by three probabilities for each tool: the probability that a tool is useful on any given day if it is good; the probability that the tool is useful on any given day if it is not good; and the decision maker's prior probability that the tool is good. Table 3(a) gives the numerical values we use. Table 3(b) provides, for each of the 16 hypotheses, the prior probability and both the corresponding objectively optimal kit and its expected value for costs 2 and 3, respectively. Table 3(c) provides "initial conditions" and the clairvoyance value w^*.

In both models, the optimal toolkit for each hypothesis is unique. But while Model 1 satisfies the condition in Proposition 4(a), Model 2 does not: For instance, the toolkit $\{1, 2\}$ is optimal for μ_{12} (under either cost level) and the immediate reward from $\{1, 2\}$ has the same distribution under μ_{12} as under μ_{123} and μ_{1234}, yet $\{1, 2\}$ is not optimal under either of the latter two hypotheses.

We henceforth use the term *Problem X-Y* where $X = 1, 2$ represents the model and $Y = 2, 3$ indicates the tools' cost.

	Hypotheses					
v	v_{12}	v_{13}	v_{14}	v_{23}	v_{24}	v_{34}
(15.1,1.2,1.3,1)	0.4	0.4	0.4	0.1	0.1	0.1
(1.1,15.2,1.3,1)	0.4	0.1	0.1	0.4	0.4	0.1
(1.1,1.2,15.3,1)	0.1	0.4	0.1	0.4	0.1	0.4
(1.1,1.2,1.3,15)	0.1	0.1	0.4	0.1	0.4	0.4
π_0	0.25	0.22	0.18	0.15	0.12	0.08

(a) Probabilities

v	v_{12}	v_{13}	v_{14}	v_{23}	v_{24}	v_{34}
Optimal toolkit	{1,2}	{1,3}	{1,4}	{2,3}	{2,4}	{3,4}
Value if cost = 2	8.36	8.42	8.26	8.46	8.32	8.38
Value if cost = 3	6.36	6.42	6.26	6.46	6.32	6.38

(b) Optimal toolkits and their value

	Cost = 2	Cost = 3
Myopically optimal toolkit at π_0	{1,2,3,4}	{1,2,3}
Myopic value at π_0	7.1512	4.1734
Clairvoyance value w^*	8.367	6.367

(c) Initial conditions

Table 2. Model 1. (See text for explanation.)

	Value		Prob. of useful		Prior that
Tools	if useful	if not useful	if good	if not good	tool is good
1	15.1	1.1	0.5	0.2	0.51
2	15.2	1.2	0.495	0.205	0.505
3	15.3	1.3	0.49	0.21	0.5
4	15	1	0.485	0.215	0.495

(a) Tool values and probabilities

		Optimal toolkit		Value	
Hypothesis	π_o	for cost = 2	for cost = 3	for cost = 2	for cost = 3
μ_{1234}	0.06374	{1,3}	{1,3}	7.679	5.679
μ_{123}	0.06503	{1,3}	{1,3}	7.679	5.679
μ_{124}	0.06374	{1,2}	{1,2}	7.63975	5.63975
μ_{12}	0.06503	{1,2}	{1,2}	7.63975	5.63975
μ_{134}	0.06248	{1,3}	{1,3}	7.679	5.679
μ_{13}	0.06374	{1,3}	{1,3}	7.679	5.679
μ_{14}	0.06248	{1,4}	{1,4}	7.47075	5.47075
μ_1	0.06374	{1}	{1}	6.1	5.1
μ_{234}	0.06124	{2,3}	{2,3}	7.669055	5.669055
μ_{23}	0.06248	{2,3}	{2,3}	7.669055	5.669055
μ_{24}	0.06124	{2,4}	{2,4}	7.509965	5.509965
μ_2	0.06248	{2}	{2}	6.13	5.13
μ_{34}	0.06003	{3,4}	{3,4}	7.548695	5.548695
μ_3	0.06124	{3}	{3}	6.16	5.16
μ_4	0.06003	{4}	{4}	5.79	4.79
μ_0	0.06124	{3,4}	{3}	2.566945	1.24

(b) Optimal toolkits and their value

	Cost = 2	Cost = 3
Myopically optimal toolkit at π_o	{1,2,3}	{2,3}
Myopic value at π_o	5.42893	3.37535
Clairvoyance value w^*	6.92458	5.2133

(c) Initial conditions

Table 3. Model 2. (See text for explanation.)

Simulation protocols

The basic simulation protocols are similar to those followed in Francetich and Kreps (2019). Each of the twelve scenarios (2 models $\times$ 2 cost levels $\times$ 3 values of δ) was simulated separately, for 10,000 iterations each. On each iteration, a hypothesis was chosen according to the decision-maker's prior and then the sequence $\{v_t\}$ was simulated out to a horizon T that depended on the value of δ: For $\delta = 0.9$, we chose $T = 64$; for $\delta = 0.96$, $T = 115$; and for $\delta = 0.99$, $T = 450$. These horizons are such that we lose around 0.1% of the total value for $\delta = 0.9$ and around 1% for $\delta = 0.96$ and 0.99 due to the truncation; we went out "further" in terms of total value for $\delta = 0.9$ because the heuristics were still making a lot of choices at $T = 64$, while they had largely "settled down" at $T = 115$. On each iteration, each of the six heuristics was implemented alongside two benchmarks:

- The *Simple Set Based Heuristic* (SSB) from Francetich and Kreps (2019). In this heuristic, a strictly positive integer τ (a parameter) is set. The decision maker chooses $K_t = X$ for $t = 0, \ldots, \tau - 1$. For $t \geq \tau$, she computes for each toolkit K the average payoff it *would have generated* and chooses for K_t whichever toolkit has the highest average. The term *would have generated* is italicized because, if K_t is chosen at date t, the decision maker is able to compute what her reward would have been had she chosen any kit $K \subseteq K_t$, and these counterfactual rewards are included in her averages.

- The second benchmark, called *Ideal*, imagines that the decision maker, at date t, has seen the full history $\{v_s : s = 0, 1, \ldots, t-1\}$ and forms a Bayesian posterior based on (this) complete information. At each date, she chooses a toolkit that maximizes her myopic payoff given her complete-information posterior. Obviously, this heuristic is infeasible, but it sets an upper bound on how well any decision rule can do—including the too-hard-to-compute optimal strategy.

The simulations were conducted in R, on an iMac. Simulating out to $T = 450$ was time-consuming: Running 10,000 iterations out to $T = 450$ for Model 2 took a bit over 48 hours to run. This is largely because of ADP; if ADP is omitted from the program (so only AM, HS, TS, BUCB, BUCBx, SSB, and Ideal are simulated), 10,000 iterations took approximately 7.2 hours.

The parameters ϵ in BUCB and BUCBx and τ in SSB

BUCB and BUCBx require specifying the parameter ϵ that determines when a hypothesis is deemed to be implausible. The performance of these heuristics can depend significantly on the value of this parameter. Similarly, our benchmark prior-free SSB is parameterized by τ, the number of periods that pass before the decision maker begins to make real choices. Hence, as a preliminary step, we simulated BUCB, BUCBx, and SSB in all four problems for a variety of values of ϵ and τ to find the (approximate) values of these parameters associated with the best performance, which we then used in our main simulations. These preliminary

simulations consisted of 5000 iterations apiece and yielded the values displayed in Table 4. (The online appendix provides full results of these preliminary simulations.) It is noteworthy that, in general, the optimal values of ϵ are relatively large while the optimal values of τ are relatively small. As seen more generally in Francetich and Kreps (2019), "fast and sloppy" decision making is generally superior to "slow and careful."

For Model 1 (at both cost levels), the best value of ϵ for BUCB is in all cases is at least 0.25, which is the prior probability of the most likely hypothesis. Hence, no hypothesis qualifies as plausible at the outset—and perhaps at other times. BUCB (and BUCBx) are implemented in the simulations to default to TS in such circumstances.

Simulation results

Having fixed the parameter values for SSB, BUCB, and BUCBx, we proceeded to the main simulations. A variety of statistics were collected concerning the simulation results. We present here a selection of those results; full results of the simulations (and the R scripts employed) can be found in the online appendix. Interpretation follows presentation of the results.

Overall performance

Table 5 provides the overall performance results for the four problems. For instance, in the 10,000 iterations of Problem 1-2 with $\delta = 0.99$, ADP averaged a normalized discounted sum of payoffs (out to $T = 451$) of 8.1255, while AM averaged 8.129; for $\delta = 0.9$, the numbers are 7.5705 for ADP and 7.5853 for AM. In Problem 1-3, we have 6.0148 and 5.7599, respectively, for $\delta = 0.99$, and 4.7562 and 4.6784 for $\delta = 0.9$. (The corresponding sample variances are provided in the online appendix.)

	Problem 1-2			Problem 1-3			Problem 2-2			Problem 2-3		
	δ = 0.9	δ = 0.96	δ = 0.99	δ = 0.9	δ = 0.96	δ = 0.99	δ = 0.9	δ = 0.96	δ = 0.99	δ = 0.9	δ = 0.96	δ = 0.99
τ for SSB	9	9	10	3	6	8	3	4	4	2	3	5
ε for BUCB	0.25	0.325	0.275	0.3	0.3	0.3	0.05	0.05	0.05	0.0625	0.05	0.05
ε for BUCBx	0.15	0.15	0.1	0.025	0.1	0.1	0.025	0.025	0.025	0.05	0.05	0.05

Table 4. Optimal parameter values for BUCB and BUCBx.

	Problem 1-2			Problem 1-3			Problem 2-2			Problem 2-3		
	δ = 0.9	δ = 0.96	δ = 0.99	δ = 0.9	δ = 0.96	δ = 0.99	δ = 0.9	δ = 0.96	δ = 0.99	δ = 0.9	δ = 0.96	δ = 0.99
SSB	7.4120	7.6329	7.9536	4.3629	4.9839	5.7777	5.5246	5.8120	6.1806	3.2896	3.8042	4.4294
AM	7.5853	7.8473	8.1290	4.6784	5.2215	5.7599	6.0677	6.2677	6.4438	4.0451	4.3195	4.5944
HS	7.5530	7.8156	8.1085	4.5494	5.1764	5.8435	5.8786	6.1828	6.4740	3.4473	4.0606	4.5684
TS	5.9465	6.8383	7.7929	3.9395	4.8674	5.8095	5.4678	5.8338	6.3284	3.7268	4.0836	4.6111
BUCB	6.3211	7.1059	7.8840	4.3628	5.1376	5.9174	5.8319	6.1073	6.5025	3.8557	4.2088	4.7381
BUCBx	6.4460	7.1819	7.8995	4.4194	5.1808	5.9231	6.0167	6.2273	6.4270	4.0451	4.3190	4.5946
ADP	7.5705	7.8325	8.1255	4.7562	5.4105	6.0148	6.0511	6.2779	6.5454	4.1017	4.3781	4.7151
Ideal	7.6000	7.8622	8.1373	4.9175	5.5255	6.0590	6.2787	6.5111	6.7157	4.4166	4.7466	5.0100

Table 5. Basic results for four test problem-cost pairs.

Table 6 provides, for each simulation and pair of heuristics, the one-sided, paired-sample difference-of-means t statistic.[14] A positive entry means that the corresponding row heuristic outperformed the column heuristic; a negative entry means the column heuristic performed better. Most of these t-statistics are quite large; the heuristics perform differently in most cases, and the sample size is 10,000. Of course, standard measures of the "significance" of these t-statistics apply only if you have a single paired-comparison in mind, a priori. But, in most cases, these numbers are so large that we can conclude that the performance differences are, for the most part, quite significant statistically.

To help unpack the information in Tables 5 and 6, Table 7 provides for each problem and δ the rank ordering of the heuristics by overall performance. Some of the differences in overall performance are statistically insignificant, even with 10,000 trials; we indicate this by putting in parentheses "ranks" that are have paired-comparison t's of 2.5 or less. For instance, for Problem 2-2, $\delta = 0.9$, AM is in second place to Ideal, but it is not significatively different (measured in this fashion) from the rank 3 heuristic, ADP; so, the corresponding entry in Table 7 is 2 (3). ADP ranked 3rd but is not that different from the rank 2 heuristic (AM) or the rank 4 heuristic (BUCBx), so the corresponding entry is 3 (2,4).

$\delta = 0.9$

		SSB	AM	HS	TS	BUCB	BUCBx	ADP
	AM	25.2						
	HS	22.1	-11.1					
	TS	-79.9	-94.5	-91.8				
Problem 1-2	BUCB	-59.3	-73.9	-71.1	17.5			
	BUCBx	-51.0	-65.1	-62.5	23.8	6.6		
	ADP	24.8	-4.0	4.1	91.9	71.5	62.5	
	Ideal	27.9	7.1	14.3	95.2	74.8	65.9	9.3
	AM	22.2						
	HS	15.8	-12.5					
	TS	-22.9	-38.1	-33.2				
Problem 1-3	BUCB	0.0	-19.2	-11.3	19.7			
	BUCBx	3.1	-17.5	-7.9	22.8	3.1		
	ADP	30.6	7.4	17.7	44.4	23.6	20.9	
	Ideal	43.9	31.7	35.1	53.3	34.8	34.0	21.2
	AM	38.3						
	HS	30.0	-16.0					
	TS	-3.3	-36.0	-25.5				
Problem 2-2	BUCB	18.7	-15.1	-3.1	22.1			
	BUCBx	32.2	-3.8	10.1	31.6	10.7		
	ADP	41.8	-1.2	15.7	36.0	13.9	2.3	
	Ideal	58.9	20.5	40.8	51.1	32.0	20.5	20.1
	AM	44.4						
	HS	11.6	-39.2					
	TS	24.1	-17.3	16.5				
Problem 2-3	BUCB	32.3	-10.4	25.3	8.1			
	BUCBx	44.4	-0.8	39.2	17.3	10.4		
	ADP	49.7	3.3	45.1	21.2	14.5	3.3	
	Ideal	77.9	29.1	82.2	41.4	35.6	29.1	21.8

$\delta = 0.96$

		SSB	AM	HS	TS	BUCB	BUCBx	ADP
	AM	45.7						
	HS	40.1	-15.6					
	TS	-70.8	-97.7	-93.2				
Problem 1-2	BUCB	-47.2	-73.9	-69.2	21.5			
	BUCBx	-41.0	-67.8	-63.1	28.6	7.5		
	ADP	45.3	-5.3	5.4	93.5	69.4	63.3	
	Ideal	50.6	8.7	19.2	99.0	75.2	69.2	12.8
	AM	23.8						
	HS	22.5	-6.6					
	TS	-10.5	-29.9	-28.0				
Problem 1-3	BUCB	14.2	-8.1	-3.9	22.4			
	BUCBx	18.3	-4.3	0.4	26.4	4.2		
	ADP	60.2	23.7	31.8	53.5	29.4	25.0	
	Ideal	72.5	44.4	52.6	66.7	44.6	41.5	26.5
	AM	44.8						
	HS	41.1	-10.3					
	TS	1.9	-39.3	-33.2				
Problem 2-2	BUCB	26.9	-15.4	-7.5	26.4			
	BUCBx	38.6	-4.4	4.9	34.8	10.7		
	ADP	53.2	1.2	13.0	43.3	17.3	5.4	
	Ideal	79.2	32.3	49.2	68.4	45.6	33.8	34.7
	AM	42.6						
	HS	24.6	-25.2					
	TS	22.3	-18.7	2.0				
Problem 2-3	BUCB	33.9	-8.9	13.6	12.7			
	BUCBx	42.5	-1.6	25.1	18.7	8.9		
	ADP	54.5	6.0	37.0	26.1	16.2	6.1	
	Ideal	93.5	46.3	87.0	63.9	55.6	46.2	47.1

$\delta = 0.99$

		SSB	AM	HS	TS	BUCB	BUCBx	ADP
	AM	87.8						
	HS	76.3	-25.1					
	TS	-39.7	-94.0	-85.9				
Problem 1-2	BUCB	-17.7	-72.7	-64.2	21.7			
	BUCBx	-13.7	-68.0	-59.8	25.7	4.2		
	ADP	89.2	-3.2	13.8	89.6	67.8	63.4	
	Ideal	95.6	11.7	28.7	96.3	74.8	70.2	13.4
	AM	-2.2						
	HS	11.5	13.2					
	TS	6.3	6.3	-6.1				
Problem 1-3	BUCB	28.7	21.1	14.3	25.8			
	BUCBx	29.6	22.1	15.4	27.3	1.6		
	ADP	62.6	35.5	37.1	58.6	30.8	29.4	
	Ideal	71.9	42.2	47.8	74.2	47.6	47.2	31.3
	AM	32.7						
	HS	40.3	4.9					
	TS	19.3	-16.5	-23.4				
Problem 2-2	BUCB	43.2	8.7	4.8	31.0			
	BUCBx	29.1	-2.4	-6.9	13.5	-10.4		
	ADP	51.6	17.2	14.5	36.6	7.9	18.5	
	Ideal	81.2	46.6	50.7	77.6	48.4	45.5	41.5
	AM	22.1						
	HS	20.3	-4.3					
	TS	25.0	2.3	6.4				
Problem 2-3	BUCB	44.7	20.6	26.8	24.1			
	BUCBx	22.2	0.2	4.3	-2.3	-20.6		
	ADP	44.8	20.9	29.0	16.3	-3.9	20.8	
	Ideal	99.1	74.1	88.6	77.7	59.8	74.1	67.1

Table 6. Paired-sample t-statistics for pairs of heuristics in each test problem, for each δ.

[14] Let x_{Hk} be the net reward produced by heuristic H on iteration $k = 1, \ldots, 10,000$. Table 6 provides the t statistic for the one-sided "mean = 0" test for $\{x_{Ak} - x_{Bk} : k = 1, \ldots, 10,000\}$, where A is the row heuristic and B, the column one. Due to the variation arising from drawing different hypotheses each iteration, this test is considerably more powerful than a difference-of-means test based on the samples $\{x_{Ak}\}$ versus $\{x_{Bk}\}$.

	Problem 1-2			Problem 1-3			Problem 2-2			Problem 2-3		
	δ = 0.9	δ = 0.96	δ = 0.99	δ = 0.9	δ = 0.96	δ = 0.99	δ = 0.9	δ = 0.96	δ = 0.99	δ = 0.9	δ = 0.96	δ = 0.99
SSB	5	5	5	6 (7)	7	7 (8)	7	8 (7)	8	8	8	8
AM	2	2	2	3	3	8 (7)	2 (3)	3 (2)	5 (6)	3 (4)	3 (4)	6 (4, 5)
HS	4	4	4	4	5	5	5	5	4	7	7	7
TS	8	8	8	8	8	6	8	7 (8)	7	6	6	4 (5, 6)
BUCB	7	7	7	7 (6)	6	4 (3)	6	6	3	5	5	2
BUCBx	6	6	6	5	4	3 (4)	4 (3)	4	6 (5)	4 (3)	4 (3)	5 (4, 6)
ADP	3	3	3	2	2	2	3 (2, 4)	2 (3)	2	2	2	3
Ideal	1	1	1	1	1	1	1	1	1	1	1	1

Table 7. Rank orders of the heuristics' overall performances. (See text for explanation.)

Evolution of average payoffs

Tables 8, 9, and 10 provide data on the average payoffs for each heuristic at different points in time. Table 8 gives results for $t = 0$ and supplies both the initially chosen toolkit and the average payoff it yields. The initially chosen toolkit for TS is random, as is the initially chosen toolkit in Model 1 for BUCB (which, as previously noted, defaults to TS, as no hypothesis has prior probability exceeding ϵ). Tables 9 and 10 provide average payoffs for $t = 1, 4, 8, 16, 32$, and 450, for the simulations for $\delta = 0.99$. (Of all the simulations, only ADP is directly sensitive to the value of δ, although changes in the parameter values ϵ and τ cause SSB, BUCB, and BUCBx to behave differently for different δ.)

Probability of hitting the objectively optimal tookit by $t = 450$

Table 11 provides the percentage of iterations, for $\delta = 0.99$, in which the toolkit selected by each heuristic at $t = 450$, K_{450}, is (objectively) optimal for the hypothesis drawn at the start of each iteration. For Problems 1-2 and 1-3, the heuristics are likely to have reached the optimal toolkit by $t = 450$. These rates are much lower for Problems 2-2 and 2-3. This begs the question: Is there a specific factor causing these low rates? For both Problems 2-2 and 2-3, Table 12 provides the hit-rate data stratified by the number of good tools in each iteration. For instance, 2,530 out of 10,000 iterations for Problem 2-2 had hypotheses with three good tools (μ_{123}, μ_{124}, μ_{134}, or μ_{234}); ADP reached an optimal toolkit in 45.8% of them.

	Problem 1-2		Problem 1-3		Problem 2-2		Problem 2-3	
	K_0	Value of K_0	K_0	Value of K_0	K_0	Value of K_0	K_0	Value of K_0
SSB & HS	{1,2,3,4}	7.1512	{1,2,3,4}	3.1512	{1,2,3,4}	4.7317	{1,2,3,4}	0.7317
AM & Ideal	{1,2,3,4}	7.1512	{1,2,3}	3.2914	{1,2,3}	5.2489	{2,3}	3.3754
TS	random	4.3558	random	2.3558	random	5.0094	random	3.2505
BUCB	random	4.3558	random	2.3558	{1,3}	5.3664	{1,3}	3.3664
BUCBx	{1,2}	4.8845	{1,2}	2.8845	{2,3}	5.3754	{2,3}	3.3745
ADP δ = 0.9	{1,2,3,4}	7.1512	{1,2,3}	3.2914	{1,2,4}	5.3215	{1,2}	3.3025
ADP δ = 0.96	{1,2,3,4}	7.1512	{1,2,3}	3.2914	{1,2,4}	5.3215	{1,2,3}	2.3241
ADP δ = 0.99	{1,2,3,4}	7.1512	{1,2,3}	3.2914	{1,2,3,4}	4.7317	{1,2,3,4}	0.7317

Table 8. Expected payoffs at $t = 0$.

	Problem 1-2			Problem 1-3			Problem 2-2			Problem 2-3		
	t = 1	t = 4	t = 8	t = 1	t = 4	t = 8	t = 1	t = 4	t = 8	t = 1	t = 4	t = 8
SSB	7.1518	7.1511	7.1522	3.1512	3.1511	5.2621	4.6827	5.7089	5.8785	0.7488	0.7363	4.2515
AM	7.1518	7.3495	7.6298	3.5878	4.5103	4.9404	5.7613	5.9489	6.1820	3.7033	3.9742	4.2790
HS	7.1518	7.3019	7.5864	3.3585	4.2730	4.8205	5.2359	5.7516	6.1553	2.3448	3.4972	4.0679
TS	4.6735	5.3996	6.1310	2.6183	3.5850	4.1497	5.1190	5.1810	5.5802	3.4026	3.5219	3.7685
BUCB	5.1117	6.0569	6.6366	3.2016	4.0858	4.6384	5.5487	5.5177	5.8901	3.5237	3.5007	3.8589
BUCBx	5.1629	6.0999	6.6615	3.3417	4.1076	4.6649	5.6996	5.9264	6.1486	3.7033	3.9742	4.2790
ADP	7.1518	7.3332	7.5894	3.5556	4.4192	5.1074	5.1012	5.5825	6.1771	2.3859	3.6889	4.3252
Ideal	7.1518	7.3495	7.6335	3.5878	4.7127	5.3258	5.7715	6.1469	6.4559	3.9080	4.2758	4.7088

Table 9. Average payoffs at $t = 1, 4$, and 8, for $\delta = 0.99$.

	Problem 1-2			Problem 1-3			Problem 2-2			Problem 2-3		
	t = 16	t = 32	t = 450	t = 16	t = 32	t = 450	t = 16	t = 32	t = 450	t = 16	t = 32	t = 450
SSB	7.8040	7.9719	8.1939	5.4198	5.8593	6.3524	5.9583	6.0517	6.6320	4.3942	4.3994	4.9921
AM	7.9729	8.2678	8.2079	5.5486	5.9246	6.0669	6.3327	6.4984	6.5994	4.4152	4.5546	4.8806
HS	7.9625	8.2346	8.2047	5.4850	5.9575	6.3345	6.3575	6.5556	6.6773	4.4764	4.6173	4.8981
TS	7.1450	8.0105	8.2079	5.1043	5.9687	6.3903	6.0008	6.2108	6.8426	4.1652	4.3915	5.1323
BUCB	7.3344	8.0522	8.2079	5.4692	6.0692	6.3903	6.2049	6.4593	6.8191	4.2359	4.5113	5.1370
BUCBx	7.4765	8.0803	8.2079	5.4906	6.0799	6.3903	6.3577	6.5787	6.5714	4.4152	4.5555	4.8875
ADP	7.9705	8.2611	8.2079	5.7758	6.2173	6.3903	6.3876	6.6732	6.7587	4.7613	4.8338	5.0007
Ideal	8.0487	8.2857	8.2079	5.9278	6.2756	6.3903	6.6298	6.7809	6.8769	5.0201	4.9803	5.1799

Table 10. Average performance at $t = 16, 32$, and 450.

	SSB	AM	HS	TS	BUCB	BUCBx	ADP	Ideal
Problem 1-2	97.34%	100%	99.78%	100%	100%	100%	100%	100%
Problem 1-3	98.44%	88.70%	97.76%	100%	100%	100%	100%	100%
Problem 2-2	59.54%	64.57%	71.99%	95.82%	92.60%	63.86%	75.49%	100%
Problem 2-3	48.09%	31.84%	49.09%	95.77%	92.43%	31.86%	50.51%	100%

Table 11. Percentage of iterations where K_{450} is an objectively optimal toolkit, for $\delta = 0.99$.

No. of good tools	No. of iterations	SSB	AM	HS	TS	BUCB	BUCBx	ADP	Ideal
4	634	20.5%	20.8%	24.4%	96.8%	84.2%	19.9%	26.7%	100%
3	2530	34.2%	34%	39.1%	96.4%	90.4%	32.8%	45.8%	100%
2	3750	74.4%	64.2%	79.9%	95.6%	90.6%	65.4%	84.1%	100%
1	2458	84.7%	98.8%	99%	94.9%	98.3%	95.7%	99.2%	100%
0	628	14%	100%	99.7%	97%	100%	100%	100%	100%

Problem 2-2

No. of good tools	No. of iterations	SSB	AM	HS	TS	BUCB	BUCBx	ADP	Ideal
4	635	16.4%	1.7%	18.1%	97.5%	84.4%	1.7%	17.8%	100%
3	2539	24.8%	3%	23.9%	96.5%	89.8%	3%	26.2%	100%
2	3755	40.7%	3.4%	31.7%	95.7%	91.5%	3.4%	33.1%	100%
1	2461	95.4%	95.8%	97.2%	94.4%	96.8%	95.9%	98.3%	100%
0	610	33%	100%	99%	97%	100%	100%	100%	100%

Problem 2-3

Table 12. Percentage of iterations where K_{450} is an objectively optimal toolkit for Problems 2-2 and 2-3 with $\delta = 0.99$, stratified by the number of good tools in the reigning hypothesis.

Speed of reaching "final toolkit"

Table 13 provides statistics about the distribution of times by which each heuristic reaches its "final toolkit" for the $\delta = 0.99$ simulations. Specifically, we record the earliest time by which 25%, 50%, 75%, 90%, and 100% of the 10,000 trials have reached a toolkit from which they do not depart subsequently: $\min\{t : K_t = K_{t+i} \text{ for all } i = 0, \ldots, T-t\}$. For instance, for ADP for Problem 1-2 and $\delta = 0.99$, we have that $K_t = K_8$ for $t = 9, 10, \ldots, 450$ on at least 2,500 iterations, and $K_t = K_{105}$ for all $t = 106, \ldots, 450$, for all 10,000 iterations. The entries 450+ indicate that, for at least one iteration of the simulation, K_{450} was different from K_{449}.

The description just given is not completely accurate for HS and TS. In each of these heuristics, there is always a small probability that $K_t = X$ is chosen strictly for informational purposes long after the decision maker has gathered enough information to conclude that she is "done." Hence, for HS and TS, at each date t we computed the myopically optimal toolkit under the posteriors that HS and TS had generated, and in constructing Table 13, we recorded the last date at which that myopically optimal toolkit changed.

How many tools are in the myopically optimal toolkit

Table 14 provides data not collected in our main simulations but that helps interpret what we see. It answers the following question: Suppose a posterior probability on the simplex of hypotheses is chosen "at random" and the myopically optimal toolkit for that posterior is computed (for a given problem). How many tools will that toolkit contain?

The idea is that the more tools such a myopically optimal toolkit contains, the higher the amount of information that will flow "serendipitously" to AM and even to ADP, hence the more likely it is that they will come to learn the truth even as they focus more on exploitation. Table 14 provides the answer for each problem, with the following complication: In Model 2, posteriors are always "independent"; the probability that one tool is good is independent of the status of all other tools. Hence, while for Problems 1-2 and 1-3 we searched over all

Percentile	SSB	AM	HS	TS	BUCB	BUCBx	ADP	Ideal
25	54	7	7	10	8	8	8	8
50	122	11	11	16	14	14	11	11
75	243	18	18	25	24	23	18	18
90	374	28	28	37	36	25	27	26
100	450+	118	118	144	117	163	105	92

Problem 1-2

SSB	AM	HS	TS	BUCB	BUCBx	ADP	Ideal
14	2	2	4	3	3	2	2
34	6	7	10	8	8	6	6
77	14	16	19	17	15	13	13
165	24	30	31	29	29	22	21
450+	134	441	126	95	114	106	98

Problem 1-3

Percentile	SSB	AM	HS	TS	BUCB	BUCBx	ADP	Ideal
25	44	5	5	15	9	5	4	8
50	134	15	17	44	24	15	12	19
75	317	29	49	100	44	28	24	35
90	426	49	147	215	66	45	40	55
100	450+	450+	449	449	225	182	180	174

Problem 2-2

SSB	AM	HS	TS	BUCB	BUCBx	ADP	Ideal
33	2	4	17	11	2	3	8
100	5	15	37	24	5	9	19
245	17	64	79	42	17	17	35
388	46	206	169	61	46	33	54
450+	242	450+	450+	216	224	165	212

Problem 2-3

Table 13. Speed of reaching the final toolkit.

	Size of myopically optimal toolkit				
	0	1	2	3	4
Problem 1-2	0%	0%	0.22%	9.59%	90.18%
Problem 1-3	0%	0.025%	30.19%	56.96%	12.83%
Problem 2-2	0%	0.697%	98.54%	0.78%	0%
Problem 2-3	0%	42.9%	57.11%	0%	0%

Table 14. Fraction of the simplex of posteriors over hypotheses for which different size toolkits are myopically optimal. (See text for explanation.)

possible posteriors on the simplex of posteriors, for Problems 2-2 and 2-3 we searched over all possible fully-independent posteriors. In each case, we sampled 5000 posteriors.

Interpretation of the simulation results

The data contained in Tables 5 through 13 lead to the following conclusions. In all cases, the qualifier "At least, for these test problems, . . .," should be appended:

1. ADP is a strong contender across the board, finishing second or third. But since the highest-ranking heuristic, Ideal, is infeasible, this really means that ADP ranks first or second in all twelve simulations.

2. AM does very well in Problem 1-2 and fairly well in the other problems for $\delta = 0.9$ and 0.96. Its very strong performance in Problem 1-2 is perhaps best explained by Table 14; for Problem 1-2, the myopically optimal toolkit is the informationally rich kit $\{1, 2, 3, 4\}$ for 90% of possible posteriors. Of course, the posteriors that AM passes through are not randomly selected, but when AM settles on a posterior in one of the corners of the simplex of posteriors, where smaller toolkits are myopically optimal, it is because evidence it received serendipitously led it to what is likely to be an "objectively optimal" corner of the simplex. On the other hand, its relatively poorer performance for the other problems and, in particular, for $\delta = 0.99$, reflects the fact that it receives less information serendipitously. We see evidence of this in Table 11 as well, where the probability that AM reaches the objectively optimal toolkit falls dramatically as we move from Problem 1-2 through Problems 1-3, 2-2, and 2-3. Indeed, by Problem 2-3, AM performs worse than SSB along this metric.

3. HS can be thought of as AM with a nod towards exploration that eventually, with probability 1, leads to the truth. Thought of in that fashion, HS's exploration seems extremely clumsy. First, it trails AM significantly in all twelve simulations (Table 7). While it is guaranteed to get to the truth, in Problems 2-2 and 2-3, it failed to do so by $t = 450$ in approximately 38% and 51% of the iterations, respectively.[15]

[15]The data in Table 13 concerning HS should be read with care. For Problem 2-3, Table 13 says that HS reached its final destination for 90% of the trials by $t = 206$. But since this final destination is objectively wrong in 50% of the trials and, we know, it will get to the objectively correct toolkit eventually with probability 1, in at least 40% of the trials, it has been stuck at a toolkit from $t = 206$ to $t = 450$ that, eventually, it must abandon.

4. In these problems, BUCB and TS do well at what has made them popular in the CS-OR literature: They get to an objectively optimal toolkit most of the time (Table 11). (Of course, we know that, eventually, TS will get there with probability one.) Per Table 13, BUCB gets there faster, in general, than TS does (more for Model 2 than Model 1). But their relatively poor performance early on (Tables 8 and 9) relegates their overall performance to mediocre or worse for $\delta = 0.9$ and 0.96. Even for $\delta = 0.99$, TS merits mediocre in Problem 2-3. In contrast, BUCBx, with its concession to exploitation (in comparison to BUCB), outperforms TS in all twelve cases and outperforms BUCB in all cases except Problem 2-3 for $\delta = 0.99$. BUCBx reaches the objectively optimal toolkit by $t = 450$ in all trials for Model 1 (in theory, it must do so eventually with probability 1), but for Model 2, its performance in this regard is on par with AM and worse than the other four Bayes-based heuristics.

5. Table 11 shows that, for Model 1, all of the heuristics find the objectively optimal toolkit by $t = 450$ all the time except for SSB and HS, and in most iterations for those two. But for Model 2, only TS, BUCB, and (of course) Ideal get there in most iterations. Tables 12 and 13 together indicate what is going wrong. For AM, the decision maker is choosing a myopically optimal toolkit. Per Table 13, it is always a one- or two-element toolkit for Problem 2-3; for Problem 2-2, it is a one- or two-element toolkit for over 90% of the possible posteriors. For BUCBx, she chooses a myopically optimal kit from among those that are optimal for some hypothesis, but, as Table 3 shows, that means she is choosing from among *all* the two- and one-element toolkits. If, say, all four tools are good, it is likely that evidence will accumulate that the tools in the myopically optimal toolkit will confirm their goodness, which in most cases will only strengthen the decision-maker's conviction that the selected toolkit is myopically optimal. In other words, for Model 2, the only way to confirm the goodness of a tool is to choose it. If it is not chosen, it never gets a chance to show that it is good.

7 Concluding remarks

It is hardly surprising (at least, ex post) that BUCB and TS do poorly relative to ADP for smaller δ—and in some cases they do very poorly—while they are relatively much better for $\delta = 0.99$. They are designed not for the traditional exploration–exploitation tradeoff of discounted multi-armed bandit problems, but instead to find the objectively optimal toolkit and to do so relatively efficiently. It is worth observing, however, that in special cases, they can perform poorly even at the task for which they are designed. Suppose that we add to Model 2 a fifth tool whose value $v_t(5)$ is never greater than one, so it is never employed if the toolkit includes at least one other tool. However, this fifth tool has a low cost c_5 and is informationally rich: There is a one-to-one map from $(v_t(1), v_t(2), v_t(3), v_t(4))$ to $v_t(5)$. (Of course, this is inconsistent with "full independence" of the five tools.) BUCB and TS, as well

as BUCBx and AM, never choose a toolkit containing this fifth tool since it is never a member of any toolkit from $\mathcal{K}_n^*$ for any n (nor is it ever a member of any myopically optimal toolkit). But, if c_5 is close to zero, it is likely to be chosen by ADP in conjunction with the myopically optimal toolkit; the only case in which it would not be chosen by ADP for sufficiently small c_5 is if no realization of v_t yields a posterior that affects which toolkit is myopically optimal in $t+1$. Thus, with this tool in existence, ADP is likely to provide a payoff equal to the payoff from using Ideal less c_5.[16]

One might argue that such useless-in-application-but-informationally-rich tools are unlikely to exist. At the same time, this illustrates in stark fashion that, of all our heuristics, only HS and ADP are designed with the classic exploration-exploitation tradeoff in mind; and while HS is, relatively speaking, very slow to get to the truth, ADP does seem (among the heuristics we explored, in the context of our test problems) to be the all-purpose winner if we ignore how computationally intensive it is. This suggests that one should seek direct measures of the informational richness of different tools or toolkits, to be able to fashion heuristics that more directly balance exploration and exploitation. For an example of this in CS-OR, see Russo and Van Roy (2016).

Note that the heuristics investigated here depend for their (good) level of performance on the decision maker having a reasonably accurate model of her situation. We cooked the accuracy of the decision maker's model into our results by drawing the value distribution from $\{\mu_1, \ldots, \mu_N\}$ and by evaluating performance relative to her prior. Suppose, in contrast, that the decision maker has a good structural model but her prior on the various hypotheses is badly flawed. Bayesian updating of such a prior could take a long time to correct the flaws. In this respect, one can make a case for less sophisticated heuristics, such as SSB, that depend less on the decision maker's hypotheses. SSB is at best a mediocre contender in the context of this paper and how we evaluated the performance of the heuristics. But compared to the Bayes-based heuristics, SSB is robust in its mediocrity, while the former are superior for accurate models but may be disastrous for a decision maker with a badly misspecified prior.

To close, we take the following broad view of what this paper teaches us. Cognitive psychologists have studied heuristics and, in particular, the qualities perceived as important to being a good heuristic. Gigerenzer and Todd (2000) emphasize *ecological rationality*, or how the structure of the environment matches the structure of the heuristic. Our results, if nothing else, confirm that, in the context of choosing a toolkit, ecological rationality is paramount. In particular, what works well in a low δ environment can be quite different from what works well when δ is close to one. What works well when information flows freely (as in Model 1) can be quite different from what works well where there are few if any serendipitous informational gains (as in Model 2). To take this to more real-life examples,

[16]This fifth tool could be even more informationally valuable, if, for instance, $v_t(5)$ under hypothesis n has a different support than under hypothesis m for every pair of hypotheses $n \neq m$.

consider the difference in heuristics for choosing players for a baseball team—where the use of individual past-performance data is prevalent—from choosing "components" for a winning basketball team, where how well the players fit together becomes much more of an issue. Or consider the research into staffing policies done in the Stanford Project on Emerging Companies (Hannan and Baron, 2002): Organizations aiming for a strong and internally focused organization culture choose new employees based on entirely different (heuristic) criteria than do organizations that seek breakthrough technologies.

Gigerenzer and Todd (2000) proposes two criteria for assessing heuristics: *coherence*, which concerns the internal, logical coherence of the judgments involved in a heuristic, and *performance*, which relates to how the heuristic fares in real-world environments. Roughly speaking, we see our theoretical results to be largely concerned with coherence, while the simulation results are about performance, albeit performance in a simulated world, not the real thing.

Viewed from the lens of mainstream economics, we suspect that readers will have found the theoretical results to be more satisfying and convincing than the conclusions we draw based on our simulations. The theoretical results are general, logical propositions; in contrast, our simulation-based conclusions are drawn from an extremely limited set of simulations. This takes us back to our less-direct point: One can and should learn from both sorts of results (as well as from field-based empirical results). In particular, economic theorists have a tendency to prove propositions about what happens for dynamic phenomena *in the limit*—for instance, as discount factors go to one—because *in the limit* is where one finds tools to prove the propositions, such as the Strong Law of Large Numbers. In the terminology of Gigerenzer and Todd (2000), proposition-proving is most often about *coherence.*

However, to understand what our propositions tell us about the real world—that is, about *performance*—one should understand, in practical terms, how close δ must be to 1 for a given level of approximation. It is good to know that TS and HS will lead to the truth, eventually. But this may be misleading without some sense of what, practically speaking, *eventually* is. Where that sense can only be derived from simulation, simulation should be employed.

It would be even better to complement what this paper does with experimental or field research about how individuals act when facing these sorts of problems. While waiting for such work, we hope the reader is convinced (as are we) that a *combination* of theorem-proving and simulation improves understanding when theorem-proving alone cannot take us far.

References

Bertsekas, Dimitri P. (2012), *Dynamic Programming and Optimal Control, Vol. II: Approximate Dynamic Programming,* 4th edition, Nashua, NH: Athena Scientific.

Doob, J. L. (1936), "Note on Probability," *Annuals of Mathematics*, Vol. 37, 363-67.

Francetich, Alejandro (2018), "Efficient Multi-Agent Experimentation and Multi-Choice Bandits," *Economics Bulletin*, Vol. 38, No. 4, October, A163.

Francetich, Alejandro, and David M. Kreps (2014), "Bayesian Inference Does Not Lead You Astray...On Average," *Economics Letters*, Vol. 125, Issue 3, December, 444-446.

Francetich, Alejandro, and David M. Kreps (2019), "Choosing a Good Toolkit, I: Prior-free Heuristics," mimeo.

Fudenberg, Drew, and Kevin He (2018), "Player-Compatible Equilibrium," mimeo.

Gigerenzer, Gerd, and Peter M. Todd (2000), *Simple Heuristics That Make Us Smart*, Oxford: Oxford University Press.

Gittins, J., and D. Jones (1974), "A Dynamic Allocation Index for the Sequential Design of Experiments," *Progress in Statistics*, Amsterdam: North-Holland, 241-66.

Gittins, J. (1979). "Bandit Processes and Dynamic Allocation Indices," *Journal of the Royal Statistical Society,* Series B, Vol. 41, 148-77.

Kallenberg, Olav (1988), "Spreading and Predictable Sampling in Exchangeable Sequences and Processes," *The Annals of Probability*, Vol. 16, 508-34.

Lai, T. D., and Herbert Robbins, "Asymptotically Efficient Adaptive Allocation Rules," *Advances in Applied Mathematics,* Vol. 6, 4-22.

Nevey, Jacques (1975), *Discrete-parameter Martingales,* Volume 10 of North-Holland mathematical library, Elsevier.

Powell, Warren, and Ilya Ryzhov (2012), "Optimal Learning and Approximate Dynamic Programming," in Lewis, Frank, and Derong Liu (2012), *Approximate Dynamic Programming, First Edition*, John Wiley & Sons, Inc.

Powell, Warren, Ilya Ryzhov, and Peter Frazier (2019), "The Knowledge Gradient Algorithm for a General Class of Online Learning Problems," *Operations Research*, Vol. 60 (1), 180-195.

Russo, Daniel, and Benjamin Van Roy (2016), "Learning to Optimize via Information-Directed Sampling," Cornell University Library: arXiv:1403.5556.

Russo, Daniel, David Tse, and Benjamin Van Roy (2016), "Algorithm Design and Regret Analysis for Online Optimization with Time Preference," mimeo.

Sauré, Denis, and Assaf Zeevi (2013), "Optimal Dynamic Assortment Planning with Demand Learning," *Manufacturing and Service Operations Management*, Vol. 15, 387-404.

Thompson, William R. (1933) "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples." *Biometrika*, **25**(3-4):285-294.

Tokic, Michel, and Günther Palm (2011), "Value-Difference Based Exploration: Adaptive

Control Between Epsilon-Greedy and Softmax," KI 2011: Advances in Artificial Intelligence, Lecture Notes in Computer Science 7006, Springer-Verlag, 335-346, ISBN978-3-642-24455-1.

A Appendix: Proofs

We use a standard probability space $(\Omega, \mathcal{F}, P)$ on which we define a random variable $\mathcal{N}$, with support $\{1, \ldots, N\}$, and processes $\{v_t\}$ and $\{u_t\}$. The process $\{u_t\}$ is i.i.d. with uniform distribution on $[0, 1]$ and is independent of everything else; for each $n = 1, \ldots, N$, the probability of the event $\{I = n\}$ is $\pi_0(\mu_n)$; and the $\{v_t\}$ are conditionally i.i.d, conditional on the value of $\mathcal{N}$, with distribution μ_n if $\mathcal{N} = n$. We construct the interlaced filtration $\mathcal{F}_0 \subseteq \mathcal{G}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{G}_1 \subseteq \ldots \mathcal{F}_\infty$, as follows: $\mathcal{F}_0$ is trivial; $\mathcal{G}_0$ is the sigma-field generated by u_0; for $t \geq 1$, $\mathcal{F}_t$ is generated by $(u_0, v_0, \ldots, u_{t-1}, v_{t-1})$ and $\mathcal{G}_t$ is $\mathcal{F}_t$ augmented by u_t. The role of the uniform random variables u_t is explained momentarily.

On this space are defined the random variables $\{K_t, \pi_t, w_t : t = 0, 1, \ldots\}$ representing, for each date t, the choice of toolkit (K_t), the decision maker's posterior (π_t, with π_0 given), and her reward ($w_t = W(v_t, K_t)$). These random variables depend on the "overall state of nature" $\omega = (u_0, v_0, u_1, v_1, \ldots)$ as well as on the heuristic the decision maker employs; we suppress this dependence from notation. (Each of our results involves only one heuristic at a time; when proving the result for a given heuristic, we assume that the random variables are defined relative to this heuristic.)

Of course, when the decision maker chooses K_t, she has (in general) less information than $\mathcal{G}_t$; she knows $\{u_0, u_1, \ldots, u_t\}$, but she will not (necessarily) have observed v_0 through v_{t-1}. Fixing (implicitly) the heuristic and defining $\mathcal{H}_0$ as $\mathcal{G}_0$ (which she is assumed to know as she goes to choose K_0), we can inductively define the information she has when she chooses K_t by the sigma-field $\mathcal{H}_t$: For $t \geq 1$, $\mathcal{H}_t$ is $\mathcal{H}_{t-1}$ augmented by u_t and the information gained from the cell of $\Gamma(K_{t-1})$ that contains v_{t-1}.

Formally, we require that K_t is $\mathcal{H}_t$ measurable. That is, if the choice of K_t is determined by more than the information available just after time $t-1$—namely, by $\mathcal{H}_{1-1}$ augmented by the information gained from K_{t-1} and the information provided by the cell of $\Gamma(K_{t-1})$ that contains v_{t-1}—it must be determined by this information *and* u_t. The role of u_t is to serve as a feasible randomization device or tie-break rule: We suppose that the decision maker, insofar as she randomizes her choices, bases those choices at time t on the value of u_t.

We use five technical lemmas:

Lemma 1. *If $\{\zeta_t : t = 0, 1, \ldots\}$ is a martingale with uniformly bounded increments, $\lim_{t\to\infty} \zeta_t/t = 0$ P-a.s.*[17]

Lemma 2. *Suppose $\{X_t : t = 0, 1, \ldots\}$ is a process adapted to the filtration $\{\mathcal{F}_t : t = 0, 1, \ldots\}$, all*

[17] See Neveu, (1975, Proposition VII-2-4).

defined on $(\Omega, \mathcal{F}, P)$, where the conditional distribution of X_t, conditional on $\mathcal{F}_{t-1}$, is given by some fixed distribution function F that is also the marginal distribution function of each X_t. (Think of X_t as taking values in some R^k.) Let $\{\chi_t : t = 0, 1, \ldots\}$ be a process defined on the same probability space such that each χ_t equals either 0 or 1, χ_0 is constant, and χ_t is $\mathcal{F}_{t-1}$-measurable.[18] *Fix a measurable set A in the range of the X_t's, and let p be the (marginal) probability that any $X_t \in A$. Write $\mathcal{F}_\infty$ for the meet (limit) of the $\mathcal{F}_t$, and write $B = \{\omega \in \Omega : \sum_{t=0}^{\infty} \chi_t(\omega) = \infty\}$. Then,*

$$\lim_{t\to\infty} \frac{\sum_{s=0}^{t} \chi_s \mathbf{1}_{\{X_s \in A\}}}{\sum_{s=0}^{t} \chi_s} = p \quad \text{on } B, \text{ P-a.s.},$$

where $\mathbf{1}_{\{.\}}$ is the usual indicator function. If the support of each X_t is bounded,

$$\lim_{t\to\infty} \frac{\sum_{s=0}^{t} \chi_s X_s}{\sum_{s=0}^{t} \chi_s} = \int x F(\mathrm{d}x) \quad \text{on } B, \text{ P-a.s.}$$

Think of the case where $\{X_t\}$ is an i.i.d. sequence and $\mathcal{F}_t$ is the σ-field generated by $\{X_s : s \leq t\}$. A statistician is keeping track of how many times X_t is in A (or of the Cesàro sums of the X_t) but with the following complication: She only includes in her sample some of the X_t. Specifically, X_t is included if $\chi_t = 1$ and is not included if $\chi_t = 0$. The decision whether to include X_t is made based on information received *prior to* time t; that is, $\chi_t \in \mathcal{F}_{t-1}$. Please note that, for a given sample path, nothing guarantees that χ_t will be 1 infinitely often; there are sample paths for which the statistician only selects finitely many of the X_t to include in her sample. But on the event, call it B, where she samples infinitely many of the X_t, the usual conclusions of the SLLN (both for the Cesàro sums and for the frequency that the selected X_t lie in some given set A) hold: They converge to what they "should" converge almost surely on B. (Of course, the result is trivial if $P(B) = 0$; it is only meaningful when $P(B) > 0$.)

This is a classic result, which Kallenberg (1988) attributes to Doob (1936). It is easily proved using Lemma 1: For each ω, given $n = 1, 2, \ldots$, let $T_n(\omega) := \min\{t : \sum_{s=0}^{t} \chi_s(\omega) = n\}$, where we let $T_n(\omega) = \infty$ if $\sum_{s=0}^{\infty} \chi_s(\omega) < n$. Define

$$\zeta_n := \begin{cases} \mathbf{1}_{\{X_{T_n} \in A\}} - p & \text{if } T_n < \infty, \\ 0 & \text{if } T_n = \infty. \end{cases}$$

Then, $\{\zeta_1 + \ldots + \zeta_n : \mathcal{F}_{T_n}\}$ is a bounded-increments martingale (where $\mathcal{F}_\infty$ is the meet of the $\mathcal{F}_t$). Apply Lemma 1.[19]

Lemma 3. *Fix a probability space $(\Omega, \mathcal{F}, P)$, sub-σ-fields $\mathcal{G}$ and $\mathcal{G}'$ where $\mathcal{G}'$ is a refinement of $\mathcal{G}$,*

[18] In the terminology of stochastic processes, χ_t is predictable.

[19] We are grateful to Hans Foellmer for the reference to Neveu (1975) and to David Aldous for the references to Kallenberg (1988) and Doob (1936).

and an event $A \in \mathcal{F}$. Let $B = \{\omega \in \Omega : P(A|\mathcal{G})(\omega) = 1\}$ (for any convenient version of $P(A|\mathcal{G})$). Then, up to a P-null set, $P(A|\mathcal{G}') = 1$ on B (and the same is true with 0 replacing 1).

In words: If, based on the information provided by $\mathcal{G}$, a decision maker (or statistician) concludes that A has definitely happened (or definitely not happened), then giving her more information will (a.s.) not change that judgment. The proof is a trivial application of the definition of conditional probability: Since we have said "up to a P-null set," there is nothing to prove if $P(B) = 0$. So assume that $P(B) > 0$. Now $P(A|\mathcal{G})$ is $E[\mathbf{1}_A|\mathcal{G}]$ and, since $B \in \mathcal{G}$,

$$\int_B \mathbf{1}_A dP = \int_B E[\mathbf{1}_A|\mathcal{G}]dP = \int_B P(A|\mathcal{G})dP = P(B),$$

where the last equality holds since $P(A|\mathcal{G}) = 1$ on B. But since $B \in \mathcal{G} \subseteq \mathcal{G}'$, B is $\mathcal{G}'$ measurable, and:

$$P(B) = \int_B \mathbf{1}_A dP = \int_B P(A|\mathcal{G}')dP.$$

Since conditional probabilities are bounded above by 1, the integrand in the last integral must be 1 (a.s.) on B. (And similarly for 0.)

Lemma 4. *Fix a probability space $(\Omega, \mathcal{F}, P)$ and a filtration of sub-σ-fields $\{\mathcal{F}_t : t = 1, 2, ...\}$. For any event $A \in \mathcal{F}$, $\{P(A|\mathcal{F}_t) : t = 1, \ldots\}$ is a martingale and, being bounded, converges almost surely. Moreover, the a.s. limit of $P(A|\mathcal{F}_t)$ is $P(A|\mathcal{F}_\infty)$, where $\mathcal{F}_\infty$ is the meet of all the $\mathcal{F}_t$.*

This is entirely standard.

Lemma 5. *Fix a probability space $(\Omega, \mathcal{F}, P)$, a filtration of sub-σ-fields $\{\mathcal{F}_t : t = 1, 2, ...\}$, and an event $A \in \mathcal{F}$ such that $P(A) > 0$. Write P^A for P conditional on A: $P^A(B) = P(A \cap B)/P(A)$ for all $B \in \mathcal{F}$. Then, the stochastic processes $\{P(A|\mathcal{F}_t) : t = 1, 2, \ldots\}$ and $\{\ln(P(A|\mathcal{F}_t) : t = 1, 2, \ldots\}$ are both submartingales (for the filtration $\{\mathcal{F}_t\}$) under P^A.*

Of the five lemmas, only Lemma 5 is not a standard result (as far as we know), and perhaps needs some explanation. We are thinking of a decision maker or statistician receiving a sequence of informative signals whose content is specified by the $\mathcal{F}_t$. This individual is interested in particular in the conditional probability that event A is true, where A has strictly positive prior probability. Of course, under her prior, her successive conditional probabilities that A is true form a martingale; that is Lemma 4. What Lemma 5 says is that, if we look at the process of posterior probabilities that A is true, not under her prior but instead under the conditional probability distribution $P(\cdot|A)$, then this process and the process formed by the log of her posteriors are submartingales. Please note, we are not talking about $\{P(A|A, \mathcal{F}_t) : t - 1, 2, \ldots\}$, but about the process $\{P(A|\mathcal{F}_t) : t = 1, 2, \ldots\}$ under the probability law $P(\cdot|A)$. For this result (which we believe is originally attributable to Turing), see Francetich and Kreps (2014).

Since there are only finitely many μ_n, N applications of Lemma 4 tell us that, whatever the decision maker is doing, her sequence of posteriors $\{\pi_t(\mu_n) : t = 0, 1, \ldots\}$ converges for each n with probability 1. Let π_∞ be the limit posterior; we have $\pi_\infty(\mu_n) = \mathbf{P}(\mathcal{N} = n|\mathcal{H}_\infty)$, where $\mathcal{H}_\infty$ is the σ-field generated by all the information she receives.

Proof of Proposition 1. At dates $t = 2^k$ for $k = 0, 1, \ldots$, the decision maker's chooses $K_t = X$; since the marginal distributions of the vector v_t are different under the different μ_n, it is an immediate consequence of Lemma 2 that, if she ignores information received on dates other than those of the form $t = 2^k$, she will learn the true value of $\mathcal{N}$ P-a.s. Since the totality of information she learns is a refinement of what she learns on those dates only, Lemma 3 tells us that, conditional on all the information she does receive, she learns $\mathcal{N}$ a.s. Then, Lemma 4 tells us that, when we supplement the information she receives on other dates, her posterior converges to $\pi_\infty(\mu_n) = 1$ on the event $\{\mathcal{N} = n\}$ and to $\pi_\infty(\mu_n) = 0$ in its complement. Since at times t not of the form 2^k she chooses a myopically optimal toolkit given π_t, on $\{\mathcal{N} = n\}$, past some point in time, at those times she must (a.s.) be choosing a toolkit that is optimal under $\mu_\mathcal{N}$. (Any toolkit that is not optimal against $\mu_\mathcal{N}$ is, eventually, not going to be myopically optimal.) This, together with a simple application of the SLLN concerning the rewards she receives (once she has settled on toolkits that are optimal given $\mu_\mathcal{N}$), gives the result. ■

Proof of Proposition 2. This is a simple variation on the proof of Proposition 1. The decision maker chooses $K_t = X$ infinitely often with probability one. On that (a.s.) event, the v_t observed reveal $\mathcal{N}$ with probability 1. Data she receives at other dates cannot change her posterior assessments that $\mathcal{N} = n$, for each n; they converge to 1 if $\mathcal{N} = n$ and 0 if not. Hence, on dates when she chooses myopically, she is eventually choosing objectively optimal toolkits. As $t \to \infty$, the rewards from these choices converge to w^* with probability 1. ■

Concerning the proofs of Propositions 3 and 4, we give a detailed proof of Proposition 3, and then use the ideas in this proof to give a more discursive proof of Proposition 4.

Proof of Proposition 3. Suppose that, for some $n \in \{1, \ldots, N\}$, $\pi_\infty(\mu_n) \neq \mathbf{1}_{\{\mathcal{N}=n\}}$ $\mathbf{P}$-a.s. Since $0 \leq \pi_\infty(\mu_n) \leq 1$, this implies that $\int_{\{\mathcal{N}=n\}} \pi_\infty(\mu_n)(\omega)\mathbf{P}(d\omega) < \mathbf{P}\big(\{\mathcal{N} = n\}\big) = \pi_0(\mu_n)$; and since $\mathbf{E}[\pi_\infty(\mu_n)] = \pi_0(\mu_n)$, we must have $\mathbf{P}\{\mathcal{N} = n' \text{ and } \pi_\infty(\mu_n) > 0\} > 0$ for some $n' \neq n$. There may be many such n'; fix one. On the event $\{\mathcal{N} = n' \text{ and } \pi_\infty(\mu_n) > 0\}$, K_n^* is eventually chosen with probability bounded away from zero; thus, K_n^* will be chosen infinitely often for $\mathbf{P}$-almost all $\omega \in \{\mathcal{N} = n' \text{ and } \pi_\infty(\mu_n) > 0\}$. Using Lemmas 2 and 3, then, the decision maker asymptotically learns the full distribution of payoffs generated by K_n^* on this event. Hence (on this event), the distribution of $W(v, K_n^*)$ must be the same under $\mu_{n'}$ as under $\mu_\mathcal{N}$, for otherwise $\pi_t(\mu_n)$ would asymptotically approach zero.

Define a binary relation $\prec$ on $\{1,\ldots,N\}$ as follows:

$$n \prec n' \text{ if } \mathbf{P}\{\mathcal{N} = n' \text{ and } \pi_\infty(\mu_n) > 0\} > 0.$$

Note that we allow $n' = n$ in this definition. In fact, it must be true that $n \prec n$: Lemma 5 states that, with respect to $\mathbf{P}[\cdot|\mathcal{N} = n]$, which we abbreviate as $\mathbf{P}^n$, $\{\pi_t(\mu_n) : t = 0, 1, \ldots\}$ is a closed submartingale. Therefore, $\mathbf{E}[\pi_\infty(\mu_n)|\mathcal{N} = n] \geq \pi_0(\mu_n) > 0$, and so $\pi_\infty(\mu_n)$ must be strictly positive with positive probability on $\{\mathcal{N} = n\}$.

Let $\overset{*}{\prec}$ be the transitive closure of $\prec$. For each n, let $I(n) = \{n' : n \overset{*}{\prec} n'\}$ and let $\Lambda^n = \cup_{n' \in I(n)}\{\mathcal{N} = n'\}$. (Note that $n \in I(n)$.) We assert that, for each n,

$$\sum_{n' \in I(n)} \mathbf{E}[\pi_\infty(\mu_{n'}) \cdot 1_{\Lambda^n}] = \sum_{n' \in I(n)} \pi_0(\mu_{n'}). \tag{A.1}$$

It is of course true that $\mathbf{E}[\pi_\infty(\mu_{n'})] = \pi_{0n'}$, since π^∞ closes the martingale of posteriors. The point is that for all $n' \in I(n)$, $\pi_\infty(\mu_{n'}) = 0$ on the complement of Λ^n, so omitting the complement of Λ^n in the integrals on the left-hand side loses nothing. Interchange the summation and the integral on the left-hand side of (A.1):

$$\mathbf{E}\left[\mathbf{1}_{\Lambda^n} \sum_{n' \in I(n)} \pi_\infty(\mu_{n'})\right] = \sum_{n' \in I(n)} \pi_0(\mu_{n'}).$$

Since $\mathbf{E}[1_{\Lambda^n}] = \sum_{n' \in I(n)} \pi_0(\mu_{n'})$, this implies that, for every n,

$$\sum_{n' \in I(n)} \pi_\infty(\mu_{n'}) = 1 \text{ } \mathbf{P}\text{-a.s. on } \Lambda^n.$$

Now, go back to any m for which $\pi_\infty(\mu_m) \neq 1_{\{\mathcal{N}=m\}}$, and take any $n \neq m$ such that $m \prec n$. Apply (A.1) for this specific n. Since $\pi_\infty(\mu_m) > 0$ with positive probability on $\{\mathcal{N} = n\} \subseteq \Lambda^n$, we conclude that $m \in I(n)$. Hence, for every $n \neq m$ such that $m \prec n$, there is a chain $m = m^0 \prec n = m^1 \prec m^2 \prec \ldots \prec m^\ell = m$.

Consider any pair n and n', $n \neq n'$, such that $n \prec n'$. We know from the first part of this proof that the distribution of $W(v_t, K_n^*)$ under μ_n must be the same as under $\mu_{n'}$. This implies that $w_n^* = w_{n'}(K_n^*)$ and, of course, $w_{n'}(K_n^*) \leq w_{n'}^*$. Applying this to the cycle created last paragraph, we conclude that this weak inequality must be an equality. That is:

If $n \prec n'$, then $w_n^* = w_{n'}(K_n^*) = w_{n'}^*$.

For the remainder of the proof, we fix an arbitrary n and show what happens on the event $\{\mathcal{N} = n\}$. Define random variables $Y_m(t) := 1_{\{K_t = K_m^*\}} W(v_t, K_m^*)$ and $Y(t) := \sum_{m=1}^N Y_m(t)$. That is, $Y(t)$ is the decision-maker's actual net payoff in period t. The limit of the Cesàro sums of the $Y(t)$ (in which we are interested) is the sum of the limits of the Cesàro sums

of the $Y_m(t)$, assuming that these limits exist. The limit of the Cesàro sums of the $Y_m(t)$, $\lim_{T\to\infty}\left[\sum_{t=0}^{T} Y_m(t)/(T+1)\right]$, is:

$$\lim_{T\to\infty}\frac{1}{T+1}\sum_{t=0}^{T}\left[Y_m(t)-\pi_t(\mu_m)w_n(K_m^*)\right] \;+\; \lim_{T\to\infty}\frac{1}{T+1}\sum_{t=0}^{T}\pi_t(\mu_m)w_n(K_m^*), \tag{A.2}$$

assuming both limits exist. Under $\mathbf{P}^n$, they do (almost surely). To prove this, compute:

$$\mathbf{E}^n\left[Y_m(t)-\pi_t(\mu_m)w_n(K_n^*)\middle|\mathcal{F}_t\right]=\mathbf{E}^n\left[\mathbf{E}^n\left[\mathbf{1}_{\{K_t=K_m^*\}}W(v_t,K_m^*)|\mathcal{G}_t\right]\middle|\,\mathcal{F}_t\right]-\pi_t(\mu_m)w_n(K_m^*);$$

the equality holds because π_t is $\mathcal{F}_t$-measurable and $w_n(K_m^*)$ is deterministic. Moreover, the event $\{K_t=K_m^*\}$ is $\mathcal{G}_t$-measurable, so this last equality continues:

$$=\mathbf{E}^n\left[\mathbf{1}_{\{K_t=K_m^*\}}\mathbf{E}^n\left[W(v_t,K_m^*)\middle|\mathcal{G}_t\right]\middle|\,\mathcal{F}_t\right]-\pi_t(\mu_m)w_n(K_m^*). \tag{A.3}$$

Under TS and on the event $\{\mathcal{N}=n\}$, the value of v_t is independent of all information in $\mathcal{G}_t$ and is distributed according to μ_n, so $\mathbf{E}^n[W(v_t,K_m^*)|\mathcal{G}_t]=w_n(K_m^*)$ and $\mathbf{E}^n[1_{K_t=K_m^*}|\mathcal{F}_t]=\pi_t(\mu_m)$. Hence, the expression in (A.3) is 0, which implies that the process $\{\zeta_T\}$ defined by $\zeta_T=\sum_{t=0}^{T}[Y_m(t)-\pi_t(\mu_m)w_n(K_m^*)]$ is a martingale with bounded increments with respect to $\mathbf{P}^n$. Lemma 1 ensures that the limit of the Cesàro sums of $\{\zeta_T\}$ is 0 ($\mathbf{P}^n$-a.s.). So, we are left in (A.2) with:

$$\lim_{T\to\infty}\frac{1}{T+1}\sum_{t=0}^{T}\pi_t(\mu_m)w_n(K_m^*).$$

We know that, for every sample path, the sequence $\pi_t(\mu_m)$ converges to $\pi_\infty(\mu_m)$. So, along each sample path, this Cesàro limit is just $\pi_\infty(\mu_m)w_n(K_m^*)$. If $\pi_\infty(\mu_m)=0$, this is zero. If $\pi_\infty(\mu_m)>0$, then we know from our earlier argument that (for almost every sample path) $w_n(K_m^*)=w_n^*$. Therefore, when we recompose the sum of these Cesàro sums of the $Y_m(t)$ to find the limit of the Cesàro sums of the $Y(t)$, we get:

$$\lim_{T\to\infty}\frac{1}{T+1}\sum_{t=0}^{T}Y(t)=\lim_{T\to\infty}\frac{1}{T+1}\sum_{t=0}^{T}\sum_{m=1}^{N}Y_m(t)=\sum_{j=1}^{I}\pi_\infty(\mu_n)w_n(K_m^*)=w_n^*.$$

This concludes the proof. ■

Proof of Proposition 4. Part (a): Suppose that the decision maker is employing either BUCB or BUCBx with threshold ϵ and she chooses K_n^* infinitely often; then, it must be that $\pi_\infty(\mu_n)\geq\epsilon$. If $\{\mathcal{N}=\ell\}$ for some ℓ such that $W(v_t,K_n^*)$ has a different distribution under μ_n than under μ_ℓ , then (applying Lemmas 2 and 3 yet again) choosing K_n^* infinitely often would reveal that $\{\mathcal{N}\neq n\}$ and so $\pi_\infty(\mu_n)$ would have to be 0, a contradiction. Thus, if K_n^* is chosen infinitely often, $w_n^*=w(K_n^*,\mu_m)$ for whichever $m=\mathcal{N}$ and the SLLN implies that the average payoff on those dates must converge to w_m^*. It follows that, on the event $\{\mathcal{N}=m\}$, any n such that

K_n^* is chosen infinitely often grants the decision maker a limiting average payoff of w_m^* and makes the Cesàro sums of her payoffs converge (a.s.) w_m^*.

Part (b): Fix n. Consider the event $\{\mathcal{N} = n \text{ and } \pi_\infty(\mu_n) \leq \epsilon\}$. Lemma 5 tells us that $\{\ln(\pi_t(\mu_n)) : t = 0, 1, \ldots, \infty\}$ is a submartingale under $\mathbf{P}^n$, so $\mathbf{E}^n\big[\ln(\pi_\infty(\mu_n))\big] \geq \ln(\pi_0(\mu_n))$. Since the integrand is bounded above by zero, $\mathbf{E}^n\left[\ln(\pi_\infty(\mu_n))\mathbf{1}_{\{\pi_\infty(\mu_n)\leq\epsilon\}}\right] \geq \ln(\pi_0(\mu_n))$. An obvious upper bound on the latter integral is $\mathbf{P}^n\{\pi_\infty(\mu_n) \leq \epsilon\} \times \ln(\epsilon)$, so we have:

$$\mathbf{P}^n\big\{\pi_\infty(\mu_n) \leq \epsilon\big\} \leq \frac{\ln(\pi_0(\mu_n))}{\ln(\epsilon)},$$

which converges to 0 as $\epsilon \downarrow 0$. Now, consider the event $\{\mathcal{N} = n \text{ and } \pi_\infty(\mu_n) > \epsilon\}$. The only way in which the decision maker could fail to be choosing K_n^* eventually is if, for some $m \neq n$, she is choosing some other K_m^* infinitely often. For this to be true, it must be that:

1. $\pi_\infty(\mu_m) \geq \epsilon$, for otherwise, past some point in time, μ_m will forever after be deemed implausible and K_m^* will not be a candidate for K_t;

2. $w_m^* \geq w_n^*$, for otherwise, once $\pi_t(\mu_n)$ is greater than ϵ and remains there forever after, K_m^* will not be selected as K_n^* offers a better plausible prospect.

Now, if K_m^* is selected infinitely often, the decision maker learns $w_n(K_m^*)$ $\mathbf{P}^n$-almost surely. It must be that $w_n(K_m^*) \geq w_n^*$; otherwise, the data would tell her that $\mathcal{N} \neq m$ and $\pi_t(\mu_m)$ would approach zero. Since $w_n(K_m^*) \leq w_n^*$, the only possibility is that $w_n(K_m^*) = w_n^*$. The argument in the proof of Proposition 3 is then easily adapted to show that the Cesàro sums of payoffs must have limit w_n^*. ■

The reason why part (b) fails for BUCBx is that argument in point 2 falls apart. Imagine there are two tools, x and y, with values $v = (v(x), v(y)) = (10, 12)$ or $(10, 1)$; there are two hypotheses, μ_1 and μ_2; the probability that $v = (10, 12)$ is 0.9 under μ_1 and 0.1 under μ_2; the prior is $\pi_0(\mu_1) = \pi_0(\mu_1) = 0.5$; her reward is $W = W^{\text{MAX}}$; and $c_1 = c_2 = 7$. We have $K_1^* = \{y\}$ and $K_2^* = \{x\}$. BUCB (with $\epsilon < 0.5$) starts with $K_0 = \{y\}$. If μ_1 is true, chances are that she eventually learns this (the probability goes to 1 as $\epsilon \to 0$); if μ_2 is true, $\pi_t(\mu_1)$ eventually falls to where she shifts to $\{x\}$, which is the right choice in this situation. But if she employs BUCBx, the best K_n^* in terms of immediate payoff is $\{x\}$. Choosing $K_0 = \{x\}$, her posterior freezes. If μ_2 is true, this is the wrong long-run choice.