Institutional Members: CEPR, NBER and Università Bocconi

WORKING PAPER SERIES

Choosing a Good Toolkit: An Essay in Behavioral Economics

Alejandro Francetich and David M. Kreps

Working Paper n. 524

This Version: September, 2014

IGIER – Università Bocconi, Via Guglielmo Röntgen 1, 20136 Milano –Italy
http://www.igier.unibocconi.it

Choosing a Good Toolkit: An Essay in Behavioral Economics

Alejandro Francetich and David M. Kreps[1]

Incomplete draft—September 2014[2]

Abstract: The problem of choosing an optimal toolkit day after day, when there is uncertainty concerning the value of different tools that can only be resolved by carrying the tools, is a multi-armed bandit problem with nonindependent arms. Accordingly, except for very simple specifications, this optimization problem cannot (practically) be solved. Decision makers facing this problem presumably resort to decision heuristics, "sensible" rules for deciding which tools to carry, based on past experience. In this paper, we examine and compare the performance of a variety of heuristics, some very simple and others inspired by the computer-science literature on these problems. Some asymptotic results are obtained, especially concerning the long-run outcomes of using the heuristics, hence these results indicate which heuristics do well when the discount factor is close to one. But our focus is on the relative performance of these heuristics for discount factors bounded away from one, which we study through simulation of the heuristics on a collection of test problems.

1. Introduction

When building models of interesting real-world phenomena, applied economic theorists are limited in what they can do; the word *tractable* is often applied as justification for models that are otherwise less than fully credible. *Tractability* is perhaps most often cited because, having posed an economic model, the model builder wants to solve for an equilibrium. But tractability can enter in a more subtle fashion: Economists stay away from situations in which they are unable to find the optimal solution to individual maximization problems that they would otherwise set for agents within their models. The inability to solve these maximization problems need not kill off all theorizing: Although it is doubtful that any economist could solve the sort of dynamic optimization problems required of agents in, say, dynamic versions of general equilibrium theory, this hasn't stopped theorists from proving that, with really smart agents who can solve those

[1] Assistance from David Aldous, Lanier Benkard, Hans Föllmer, Michael Harrison, Guido Imbens, Daniel Russo, and Benjamin Van Roy, as well as comments by seminar participants at Stanford University and Bocconi University, 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.

[2] This draft is complete as far as theoretical developments are concerned. We are still working on the simulation of test problems. See the discussion in Section 6 for our current status.

problems and with dynamically complete markets, the equilibria that emerge are Pareto efficient. But, when we want to characterize the solutions to problems posed for agents in our models, we are often limited by what we can solve.

This limitation applies to problems that otherwise are everyday problems facing real economic agents. For instance, imagine an agent who must, on a daily basis, fill up her toolkit prior to departing for work. Very concretely, think of a plumber who must decide which tools and spare parts to load on her truck. With a lot of experience, the plumber will have a pretty good idea which tools and spare parts are likely to be useful (hence included) and which not (hence excluded). But this experience may be limited for tools and parts that, in the past, have not been carried. Insofar as the decision maker, to learn how useful a particular tool might be (relative to other tools), must have some experience with that tool, she may decide to carry tool X, to learn more about it. When carrying tool X has an incremental cost, the problem becomes a multi-armed bandit problem of exploration (carry tools to learn about them) versus exploitation (don't pay the cost of a tool that, based on current information, you believe is unlikely to be of much use.)

Of course, plumbers have a wealth of experience to draw upon, theirs and that of their peers, so stocking their toolkits (or trucks) may not pose much of a problem. But, in other contexts, this issue is real and pertinent. Consider, for instance, a professional services firm. The manager of the firm must decide on whom to employ to meet the challenges that arise from day to day. Is Expert A going to be worth his salary? That can be a difficult question, if to learn how useful A will be, the manager needs to have A on staff. Or think of a sports team, say, the relief pitching staff on a baseball team. Should pitcher B be kept on the roster? How effective will B be in particular situations? How often will situations in which B is more useful than C, D, and E occur?

To give a precise formulation of the problem (as we'll see, one of many possible), imagine a decision maker (her) who must choose at each date $t = 0, 1, \ldots$ a subset of *tools* to have on hand. The universe of possible tools is a finite set X; we let $K_t \subseteq X$ denote the *toolkit* she chooses at date t. The value of each tool at date t is given by the random function $v_t : X \to R_+$. (We specify the stochastic structure of the sequence of vectors $\{v_t; t = 0, 1, \ldots\}$ two paragraphs hence.) A *gross benefit function* $U : (R_+)^X \to R_+$ is given; the decision maker's gross benefit at time t from toolkit K_t is then given by

$$U\big((v_t(x) \cdot 1_{K_t}(x))_{x \in X}\big),$$

where 1_{K_t} is the usual indicator function. That is, the overall gross value of the toolkit is a

function of the values of the various tools, *where a tool that is not in the toolkit is taken to have value zero*.[3] To have a very specific example to think about, suppose that $U(v) := \max_{x \in X} v(x)$. Hence, if the decision maker is carrying toolkit K_t at time t, her gross benefit is $\max_{x \in K_t} v_t(x)$; it is as if, at each date, only one tool out of the toolkit is used, the tool with the highest v_t-value.

The function U gives the gross benefit as a function of the v_t-values of tools in the toolkit; the decision maker must pay a price for the tools she keeps available. So, her *net benefit* (given her choice of K_t and the realized values of the tools v_t) is

$$W(v_t, K_t) := U\left((v_t(x) \cdot 1_{K_t}(x))_{x \in X}\right) - \sum_{x \in K_t} c_x,$$

where c_x can be thought of as the rental cost of tool x. The decision maker's problem is to choose her toolkits dynamically to maximize the sum of the discounted (at some discount rate $\delta \in (0, 1)$) expected values of $W(v_t, K_t)$.

The problem becomes a multi-armed bandit problem when we finish the formulation as follows:

1. We imagine that the decision maker assesses the sequence $\{v_t; t = 0, 1, \ldots\}$ as exchangeable; that is, it is i.i.d. up to an unknown distribution. Assume that the unknown distribution is drawn from a finite collection of probability distributions $\{\mu_i; i = 1, \ldots I\}$, where μ_i is a simple (finite-support) probability distribution on $(R_+)^X$. Let π^0 denote her prior assessment concerning which of μ_1 through μ_I is correct.

2. We further assume that, at date t, *the decision maker only observes the values of* $v_t(x)$ *for those* x *that are in the toolkit she selects,* K_t. Hence, in choosing her toolkit, she must weigh both the short-run benefit each tool provides and, for later purposes, the information it provides (both, in expectation) against the rental cost of the tool. That is, the "arms" in this bandit problem are the various subsets K of X.

While this is a multi-armed bandit problem, it is not of the variety of bandit problems that we know how to solve. The well-known Gittins Index solution works if and only if the "arms" of the bandit are statistically independent; learning about the distribution of returns from one arm provides no information about any other arm. In this formulation, we lose independence on two grounds. First, the formulation does not assume that the various components of the random vector v_t are independent of one another. That wouldn't be entirely natural in this

[3] The dot in $v_t(x) \cdot 1_{K_t}(x)$ may confuse you. This does not denote the dot product of two X-dimensional vectors. Rather, it is the X-dimensional vector formed by taking the product of corresponding components of each.

setting: If, say, the tools include a selection of wrenches, then learning about the value of one wrench might provide valuable information about the value of wrenches in general. And, more fundamentally, as the "arms" are *subsets* of X, learning about the distributions of $(v_t(x))_{x\in K}$ provides information about $(v_t(x))_{x\in K'}$ if K and K' have nonempty intersection.

One might hope that, while the Gittins Index cannot be applied to this problem, the literature on bandit problems gives a solution. Perhaps it does, but we have not found a general solution.[4] Solutions can be found for special cases: For instance, suppose W is linear and the values of the various tools are independent. Then the problem decomposes into a collection of simple two-arm bandits, where in each subproblem, one of the arms—the don't-carry-this-tool arm—gives a certain return of zero. Or, to take a significantly more difficult special case, Francetich (2014) solves a problem in this spirit where there are two tools, only one of which is useful at all (although the decision maker doesn't know a priori which one it is), and information about the usefulness of a tool while it is being carried arrives according to a Poisson process.

Of course, the problem is generally solvable in theory, using the methods of dynamic programming. But this is "in theory." Those methods are not practical in any but the simplest parameterizations of this problem. Nonetheless, real economic agents face problems with this structure (or variations of this structure) all the time, they make decisions, and they live with the consequences. How? Presumably, they employ heuristics or rules of thumb (or just "go with their gut").[5] [6]

Economists and economics have, we assume, an interest in decision making of all sorts in economic contexts. The problem described is an economic context, even a common economic context. But how should we approach the modeling and analysis of decision making in this context, when we have no idea what is the optimal solution? This paper suggests some approaches to this which come down to: Identify likely and/or common heuristic decision rules, and employ both deductive analysis and simulations to see how they perform, relative to one another. (We'd like to know how they perform relative to the optimal solution but, since we don't know what is the optimal solution in most cases, we can't do this.)

[4] Indeed, the sizable literature on bandit-learning, to which we will later refer, indicates that the broader community of Operations Research and Computer Science scholars have no solutions, in general.

[5] Hence the title of this paper is not *Choosing the Best Toolkit*, but instead *Choosing a Good Toolkit*.

[6] The book and, later, movie *Moneyball*, concerning the management of the Oakland A's baseball team by General Manager Billy Beane, is filled with decision making of this sort. Traditional baseball selection processes, which are denigrated in the book and movie, select players to be drafted by "how they look" rather than "how they perform"; Beane, the story goes, does better by looking at performance data. And, at one point, Beane advocates a simple linear heuristic: A line-up is judged by the sum of the on-base-percentages of its constituent parts. That may be a more sophisticated heuristic than "how the players look," and the thesis of the book and movie is that it proved to be a much more successful heuristic. But it is still, surely, a heuristic.

We emphasize *likely and/or common* as a modifier for the sort of heuristics in which we are interested; we are as interested in learning when a seemingly plausible heuristic performs poorly as we are in finding the "best performing" heuristic. A sizable literature in computer science/operations research—referred to as the bandit-learning literature—concerns decision heuristics for decision making in multi-armed bandit problems:[7] [8] These heuristics are on the high end of sophistication[9] —it is telling that they refer to their procedures as *algorithms* rather than *heuristics*—and, as such, are less likely to be employed by most economic decision makers. That said, when we get (in Section 5 of this paper) to heuristics that employ the prior assessment of the decision maker, we organize our discussion of categories of prior-based heuristics in a manner suggested by this literature.

To anticipate the obvious criticism of our research agenda: This program, when and if carried out, settles nothing. Our analytic results are weak; the simulation results will depend on the specific parameterizations that we simulate; and there is no sense in which we study the space of all tractable heuristic decision rules. But—our rebuttal of this criticism—to the extent that problems of this sort are economically common and even significant, gaining even limited knowledge of the performance of some heuristics can help us improve our understanding of both how to deal with such problems and (the limitations of) how real-life economic agents deal with them. Roughly speaking, when employing standard economic theory, economists have traded off breadth of the problems they consider against tractability of those problems, with virtually all the weight on tractability. (Hence, the "it settles nothing" criticism applies equally well to these programs.) We believe that this is the wrong weighting and that the mode of analysis here—despite its manifest flaws—is one way to achieve a better balance.

Of course, this approach belongs to behavioral economics. Much of the recent activity in behavioral economics has centered on issues of changing tastes and ambiguity; here tastes do not change and nothing is ambiguous; the issue is instead "computational complexity." This is not the first paper to explore decision heuristics. The idea that decision makers are unable to perform necessary computations required for full or *hyper*-rationality and instead behave in a *boundedly rational* fashion is generally attributed to Simon (1959), who wrote extensively on the subject (Simon 1979, 1982a, 1982b, 1997). Baumol and Quant's (1964) discussion of *rules*

[7] Even when the "arms" of the bandit are statistically independent under the prior, so that the Gittins Index can be applied, the computation of the indices is sufficiently complex so that, as a practical matter, heuristics may be employed. This literature, though, looks at quite general bandit problems.

[8] The paper in this literature that is closest to what we do here is Ryzhov and Powell (2008); this paper explicitly concerns the sort of "toolkit" bandit problem with which we deal, albeit for a somewhat different formulation.

[9] See, for instance, Russo and Van Roy (2014) for the "latest word" in sophistication.

of thumb and Radner's (1975) discussion of *satisficing* are other seminal references. More recent work includes Roth and Erev (1998), Lettau and Uhlig (1999), Rustichini (1999), and Easley and Rustichini (2005).

In multi-person settings, the complexity that prevents hyper-rational choice can arise from the fact that others are simultaneously learning and acting, hence the true environment inhabited by agents is not stationary. Agents may nonetheless attempt to "learn" about their environment using a model (necessarily incorrectly specified) that presumes stationarity; literatures that follow this pattern include learning rational expectations equilibria (Blume and Easley 1982, Bray 1982, Sargent and Marcet 1989) and learning in (repeated) games (Milgrom and Roberts 1991, Fudenberg and Kreps 1993, Kalai and Lehrer 1993, Fudenberg and Levine 1998).

The paper is organized as follows: A variety of formulations are discussed, then some limited analytical results (dealing mostly with the cases δ near zero or near one) are provided. Several simple heuristics for the problem are proposed; first we examine some that rely exclusively on the empirical data (that is, that are independent of the prior assessment of the decision maker), and then some that employ that prior assessment. Then we move to simulations, to get a sense of how well the heuristics we propose perform.

2. *Formulation*

Many of the basic elements of the problem formulation have already been given, but to reiterate:

- A decision maker (she) must, at each date $t = 0, 1, \ldots$, choose a subset K_t from a given finite set X of *tools*.
- The *values* of the tools $x \in X$ at date t are given by the random vector $v_t \in (R_+)^X$, with the xth component of v_t denoted by $v_t(x)$. The sequence of random vectors $\{v_t \in (R_+)^X; t = 0, 1, 2, \ldots\}$ is i.i.d. with (unkown to the decision maker) distribution μ^T. The decision maker begins with a (strictly positive) prior over the distribution μ^T; that is, her initial assessment is that the vector sequence has an exchangeable distribution. We assume that the support of μ^T is finite and, moreover, that the decision maker's prior assessment on μ^T is that it is drawn from a finite family of finite-support distributions $\{\mu_i; i = 1, \ldots, I\}$, with π_i^0 denoting her initial assessment that μ^T is, in fact, μ_i. We assume that the μ_i are distinct; that is, the distribution of the vector v_t under one μ_i is different from the distribution under each other μ_j. (This is w.l.o.g.; if some $\mu_i \equiv \mu_j$, the decision maker could simply combine them.) We assume that the "objectively correct" distribution μ^T is one of the μ_i.

- If the decision maker chooses toolkit $K_t \subseteq X$ at date t, her payoff in that period is

$$W(v_t, K_t) := U\big((v_t(x) \cdot 1_{K_t}(x))_{x \in X}\big) - \sum_{x \in K_t} c_x,$$

where $U : (R_+)^X \to R_+$ is a nondecreasing function, $(v_t(x) \cdot 1_{K_t}(x))_{x \in X}$ is the vector from $(R_+)^X$ whose xth component is $v_t(x)$ if $x \in K_t$ and is 0 if $x \notin K_t$,[10] and $c_x \in R_{++}$ is the (strictly positive) cost of having tool x in the date-t toolkit. The assumption that U is nondecreasing means that a higher value of a tool cannot decrease the overall value of the toolkit; combined with the application of U to the "censored" vector $(v_t(x) \cdot 1_{K_t}(x))_{x \in X}$, it means that a tool not in the toolkit contributes zero (the lowest possible value of $v_t(x)$) to U.

We will in places make assumptions about additional properties of U; the most important such assumption is that U is sub-modular. A specific functional form for U that we will use in examples is

$$U^{\text{MAX}}(v) = \max_{x \in X} v(x);$$

we also write $W^{\text{MAX}}(v, K)$ for the W function corresponding to U^{MAX}, which is

$$W^{\text{MAX}}(v, K) = \max_{x \in K} v(x) - \sum_{x \in K} c_x.$$

Note that U^{MAX} is indeed submodular as a function on R_+^X.

- The decision maker would like to choose toolkits $K_0, K_1, \ldots$ in a manner that makes the expected discounted sum of her period-by-period payoffs as large as possible, discounting with discount factor δ per period. Note our phrasing "would like"; when we finish the problem formulation in a moment, it will be in a manner that makes the problem too difficult to solve completely, so the usual formulation "the decision maker chooses her toolkits to maximize the expectation of her discounted sum of period-by-period payoffs" is, in general, impossible to fulfill.

The key element in our problem formulation is the answer to the question: In period t, if the decision maker chooses toolkit K_t, what does she learn about the realization of the vector

[10] Again, the dot in $(v_t(x) \cdot 1_{K_t}(x))_{x \in X}$ is *not* a dot product but instead the x-component-by-x-component product of $v_t(x)$ and the indicator function of the set K_t.

v_t? If, for instance, she learns the full realized value of v_t, then the problem collapses; based on what she has observed so far, she updates after each date her (current) posterior π^t over the $\{\mu_i\}$ and chooses the myopically optimal toolkit for next period. Under these informational conditions, in which her choice of toolkit has no impact on what she learns (indeed, in any formulation of information that has this feature), the problem becomes (relatively) simple and, certainly, solvable.

Our interest is instead in formulations in which her choice of K_t affects what and how much she learns about v_t. (Hence, in the usual manner of multi-armed bandit problems, she has an exploration–exploitation dilemma.) There are many ways this could happen; we will assume for this paper the following:

> *At date t, if the decision maker chooses K_t, she **observes** the vector $\big(v_t(x) \cdot 1_{K_t}(x)\big)_{x \in X}$. Or, in words, she **observes** the immediate values of the tools she carries but not those that she chose not to carry.*

This does not imply that she *learns* nothing about the values of tools that she is not carrying. We have not precluded that the values of different tools are dependent, hence (for instance) observing the value of a four-inch wrench that she is carrying may tell her something about the value of a six-inch wrench that she has not chosen, this time. The informational assumption given is about what she *observes*.

And, it should be noted, this assumption may be fairly optimistic. Consider the specific functional form U^{MAX}: The decision maker, upon seeing the value of the tools in her toolkit, employs the *one* tool that provides the greatest immediate value. This implies that she observes how valuable each tool in her toolkit would be if employed. A more realistic formulation would be that she observes an estimate of how valuable each tool in her toolkit would be and chooses the one that seems like it would be best.[11] But even our "optimistic" formulation about what she observes renders the problem too complex for solution.

One can impose on this formulation some further structural elements that might be useful. To give an example, suppose that the universe of tools X is partitioned into categories of tools; being very specific, suppose the elements of X are lawyers in a law partnership, partitioned into lawyers to litigate cases X_L, tax-law specialists X_T, contract-law specialists X_C, and so forth. Suppose in each period, the partnership must deal with a single case that requires no more than one litigator, one tax-law specialist, and so forth. This suggests both some natural functional

[11] Well, not quite. If her initial observations are not conclusive, her choice of which tool to employ would have to take into account the value of further information gained for the one tool that is employed.

forms for W and some natural constraints on the toolkits (now staff lawyers) K_t that the firm chooses at each date to have available. We will not attempt to exploit such structural elements in this paper (except for the single-tool-used structure implicit in U^{MAX}); but we do want to acknowledge the possibilities.

Some preliminary notation

Throughout, Δ denotes the simplex of probability distributions over the set $\{\mu_1, \ldots, \mu_I\}$, with π denoting a typical element of Δ and π_i denoting the probability of μ_i according to π. Of course, the decision maker's prior assessment π^0 is a member of Δ.

For each $\pi \in \Delta$, we write $w(\pi, K)$ for $\sum_{i=0}^{I} \pi_i \mathbf{E}^i[W(v, K)]$, where $\mathbf{E}^i$ denotes expectation (over the v vector) according to the probability distribution given by μ_i. That is, $w(\pi, K)$ is the (subjective) expected current-period net payoff to the decision maker if she chooses toolkit K, when π is her assessment over which μ_i is μ^T. We let $w^*(\pi)$ denote $\max_{K \subseteq X} w(\pi, K)$, or the myopically optimal expected immediate payoff she can achieve when π is her current assessment. And we let $\mathcal{K}^*(\pi)$ denote the collection of toolkits that achieve the maximum in $w^*(\pi)$. (Since the number of toolkits is finite, the maximum is achieved: $\mathcal{K}^*(\pi)$ is never empty, although it may contain the empty set as a member.) For π that assigns probability one to μ_i, write $w_i(K)$ in place of $w(\pi, K)$, w_i^* in place of $w^*(\pi)$, and $\mathcal{K}_i^*$ in place of $\mathcal{K}^*(\pi)$.

Finally, let i^T denote the index i such that $\mu^T = \mu_i$, $\mathcal{K}^*$ denote $\mathcal{K}_{i^T}^*$, and w^* denote $w_{i^T}^*$. (Note that $i^*, \mathcal{K}^*$, and w^* are all random from the perspective of the decision maker.)

Three genericity conditions and a special case

In various propositions to follow, we sometimes assume one or more of three "genericity" conditions on the data of the problem.

Condition A. *For each $i = 1, \ldots, I$, $\mathcal{K}_i^*$ is singleton, with K_i^* denoting its single member.*

Condition B. *For every pair $i, j = 1, \ldots, I$ such that $i \neq j$, and for each $K \subseteq X$ other than $K = \emptyset$, $w_i(K) \neq w_j(K)$.*

Condition C. *For any two hypotheses μ_i and μ_j such that $i \neq j$ and for any toolkit $K \in \mathcal{K}_i^* \cup \mathcal{K}_j^*$, the distribution of $W(v, K)$ under μ_i is different from its distribution under μ_j*

Condition A is not very severe; if some $\mathcal{K}_i^*$ contains more than one toolkit, and if x is a tool in one member of $\mathcal{K}_i^*$ but not in other members of $\mathcal{K}_i^*$, then a slight perturbation in the cost c_x will pare down $\mathcal{K}_i^*$. But Conditions B and C are not as innocuous as they may at first seem; they will (generally) fail to hold in the special case of "independent tools":

In this special case, for each individual tool x, there is a finite list $\{\rho_1^x, \ldots, \rho_{I(x)}^x\}$ of possible probability laws for $v(x)$ and, for each ρ_i^x, a prior probability $\pi_i^{0,x}$ that ρ_i^x is indeed true. The number of "full hypotheses" μ_i is $\prod_{x \in X} I(x)$, where a given μ_i corresponds to a selection of one ρ_i^x for each $x \in X$, with prior probability the product of the prior probabilities of the pieces that make up μ^i. In this special case, observing $v(x)$ for a given tool x generates no information about the probability law governing $v(x')$ for $x' \neq x$; hence the tools are "independent." Of course, this doesn't make the arms of the multi-armed bandit independent, since arms of the bandit are toolkits, and two toolkits whose contents overlap are not independent in this sense.

And in this special case, if the K in Condition B or C that is in question is less than all of X, and if some $x \in X \setminus K$ has $I(x) > 1$, then the condition fails: The choice of probability law for this x gives different hypotheses μ_i and μ_j for which the corresponding distributions of $W(v, K)$ are identical, hence their expected values $w_i(K)$ and $w_j(K)$ are the same.

3. Some Theoretical Results

While the decision problem we have posed is, in general, too difficult to solve analytically (except in special and/or very simple cases), some theoretical results can be provided. None of these results are particularly original, but they are worth stating.

Proposition 1. *An optimal strategy to the problem posed exists.*

The problem as formulated is an infinite-horizon dynamic programming problem, with bounded immediate rewards and discounting. Our assumptions that the decision makers's prior has finite support, as does the distribution of values of tools under every μ_i, guarantee that, for a given prior, only countably many "states" of the problem can be reached, which moots any concerns one might have about measurability of value functions.[12] Since the set of tools is finite, a conserving strategy at every decision point must exist, which is then optimal (Kreps 2013, Proposition A6.7).

Proposition 2. *Consider the following strategy. At dates $t = 1, 2, 2^2, 2^4, \ldots$, the decision maker chooses as her toolkit all of X. At all other dates, she chooses for K_t any toolkit $K \in \mathcal{K}^*(\pi^t)$, where π^t is her (Bayesian) posterior assessment on which of the μ^i is μ^T, based on all the information she has collected up to date t.*[13] *Then, almost surely (relative to both her initial subjective assessment and the "objectively correct distribution"), her posterior assessment over the μ^i will converge to a point mass on μ^T, and (hence) she eventually chooses a toolkit from K*

[12] Francetich (2013, Theorem 3.2) extends this result.

[13] At date 0, she employs her prior.

from $\mathcal{K}^*$ for all dates t except for those t that are of the form 2^n. Hence the Cesàro averages of her per-period payoffs almost surely converge to w^*.[14]

Readers familiar with the literature on multi-armed bandits will recognize this as a standard result. If the decision maker is interested in optimizing her average (expected) reward rather than a discounted sum of rewards, then she can adopt any strategy that (a) samples every "arm" infinitely often, using the data so generated to learn (almost surely) the expected return from each arm, while (b) 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 decision maker's posteriors will almost surely converge to a point mass concentrated on μ^T. Therefore, almost surely, she eventually picks some K from $\mathcal{K}^*$.[15]

Proposition 2 provides the following immediate corollary:

Corollary to Proposition 2. *Write $u^\star(\delta, \pi^0)$ for the optimal expected value attained by the decision maker in the problem with discount factor δ, as a function of δ and the decision maker's prior π^0; that is, $u^\star(\delta, \pi^0)$ is the maximized value of the (subjective) expectation of $\sum_{t=0}^{\infty} \delta^t W(v_t, K_t)$, where we maximize over all feasible strategies for choosing toolkits by the decision maker. Then,*

$$\lim_{\delta \uparrow 1} (1-\delta) u^\star(\delta, \pi^0) = \sum_{i=1}^{I} \pi_i^0 w_i^*.$$

In words, the normalized expected discounted optimal value (as a function of δ) approaches what the decision maker expects to get if someone tells her at the outset which μ_i is μ^T, and she then chooses an optimal toolkit (for all time) given that information.

The proof is immediate: The optimal value at a given δ is at least as large as the expected value from following the strategy of Proposition 2. And, as δ approaches 1, the (normalized) expected value from using the strategy in Proposition 2 is, by Proposition 2, established to be w_i^* if μ^T is μ_i. Since $\sum_i \pi_i^0 w_i^*$ is an obvious upper bound on the decision maker's feasible expected (normalized) payoff, we have the corollary.

The Corollary tells us what happens for δ close to one. At the other end of the spectrum, when δ is close to zero, the decision maker *roughly* behaves myopically; at each date, she picks

[14] The Cesàro averages of her payoffs are the time-averages: $\sum_{t=0}^{T} W(v_t, K_t)/(\text{T+1})$.

[15] A formal proof of the proposition is given in the appendix.

whichever toolkit provides the best immediate expected return, given her current beliefs about which of the μ_i is μ^T. Note that this does not mean that she picks some (myopically optimal relative to π^0) toolkit and sticks with it forever; she may well gather information that changes her assessment of which toolkit is myopically best. And this is only a rough characterization because it doesn't account for ties in which toolkit is best; when there is a tie, ties are broken based on the information the different toolkits provide. To give a precise statement, one shows that the strategy *choose for K_t some member of $\mathcal{K}^*(\pi^t)$* is and, in current-value terms, remains, ϵ-optimal for a given $\epsilon > 0$, for all δ sufficiently close to zero.[16]

Despite these theoretical results, our interest is not in the case of δ close to one or to zero. For fixed $\delta \in (0, 1)$, a third "asymptote" concerns what happens as t approaches infinity. Readers familiar with standard (independent-arm) multi-armed bandit problems will know that the "typical" behavior in such problems is that, at some point, the decision maker stops experimenting and moves to a regime of pure exploitation, choosing one and only one option—the option that is myopically optimal—for the rest of time. This is true in our problem if Conditions A and B hold.

Proposition 3. *If Conditions A and B hold, with probability one (under either the decision-maker's subjective probability or the "objective" truth μ^T), the decision maker will, from some time on, choose the same bundle repeatedly, and this bundle will be and remain her myopic optimum. Moreover, if this bundle is not $\emptyset$, the decision maker will asymptotically learn which μ_i is the true μ^T (almost surely).*

The detailed proof is left to the appendix, but the idea is easily given. On the event where the decision maker doesn't choose the $\emptyset$ (only) past some finite date T, she must choose some one of the nonempty toolkits, call it K^0, infinitely often. Computing the long-run average (per period) net return of K^0 when chosen will a.s. converge to its expected value under the reigning distribution, which is $w_{i^T}(K^0)$. Condition B guarantees that this identifies $\mu_{i^T} = \mu^T$. And then, as the decision maker becomes more and more certain about μ^T, Condition A guarantees that she does best always to choose the toolkit that is (uniquely) myopically optimal for μ^T.

Since Condition B is somewhat unnatural (e.g., in the case of independent tools), we would like to get the conclusion of Proposition 3 under less restrictive conditions. Suppose we assume that for any two distinct toolkits K and K' and for any of the hypotheses μ_i, we have $w_i(K) \neq$

[16] The strategy of choosing the myopically best toolkit for her initial π^0 forever is, for any $\epsilon > 0$, ϵ-optimal for all sufficiently small δ, but only in terms of the initial expected value.

$w_i(K')$.[17] Then if K and K' are both selected infinitely often with positive probability, on that event the decision maker would learn (by computing averages in the fashion of the proof of Proposition 3) the values of $w_i(K)$ and $w_i(K')$, for whichever is the true μ_i. It would seem that one could then show that, as the time parameter grows, choosing whichever of K or K' gives a smaller expected immediate reward is suboptimal. But we have not tried to push through the details of this conjecture.

4. Four Prior-Free Heuristics

We divide the heuristics we present into two groups: heuristics that employ the decision maker's prior assessment π^0 and the properties of the μ_i, and those that ignore all these things and are based solely on what the decision maker observes empirically. Please note that the literal meaning of "prior free" suggests that the decision maker ignores π^0, but it does not preclude that, for instance, she immediately discards a tool that, under any hypothesis μ_i, generates value far less than its cost. We are using a very broad definition of prior free; the decision maker bases her decisions *solely* on data she observes and not on any initial structural information she might possess. In this sense, it might be more accurate to call these heuristics *naive* or even *very naive*.[18]

The four heuristics we propose have the following basic structure. The decision maker starts out with an initial toolkit K^0, which is X; that is, she carries all the tools. After some (deterministic) period of time T_1 has elapsed, she evaluates her situation on the basis of evidence so-far accumulated and drops some of the tools, leaving herself with $K^1 := K_{T_1}$. She sticks with K^1 until some second (still) deterministic time T_2, at which point she re-evaluates her situation and, perhaps, drops some further tools so that, moving forward, she carries K^2; she continues this cycle of observation and winnowing of her toolkit, with subsequent "decision dates" labelled $T_2, T_3, \ldots$ and toolkits $K^3, K^4, \ldots$.[19] Note that in these heuristics, tools are dropped from consideration and never "come back"; that is, $K^n \subseteq K^{n-1}$ for all n. Needless to say, a good case can be made for heuristics more sophisticated than these, especially where available evidence indicates that a tool that was dropped at, say, T_n because some other tool seemed superior, might later be resurrected if that other tool, per further evidence, is shown to be worse. But the four heuristics of this section are too simpleminded to consider bringing tools

[17] This is a stronger form of Condition A, which concerned uniqueness of (only) the *optimal* toolkits for each μ_i.

[18] While it may seem silly to consider these types of heuristics, it should be noted that in both the learning-rational-expectations-equilibria and learning-to-play-games literature, referenced earlier, are entirely prior-free in just this sense. So there is plenty of precedent for heuristics that are so naive.

[19] Being very pedantic: K_t is the toolkit chosen at date t, while K^n is the toolkit chosen from date T_n until $T_{n+1}-1$.

back.

Looking at heuristics with this structure, it remains to specify:

- the evaluation dates $T_1, T_2, \ldots$ (with $T_0 := 0$);
- how information is processed or evaluated; and
- the decision rule by which tools are eliminated.

The four heuristics differ in terms of the unit of analysis when it comes to evaluation: The first two evaluate individual tools, while the second two evaluate tool*kits*. Although it may seem obvious to the reader that, given the context of this problem, tool*kits* should be the unit of analysis, because of potential complementarities between tools (or the substitution of one tool for another), a particularly naive decision maker, faced with a particular tool and the question, "Should I put this tool in my kit?," might answer this question by answering the question, "What has this tool contributed in the past?" This immediately gives our first heuristic, which works *only* for the case where $U = U^{\text{MAX}}$.

The Pay-for-Itself Heuristic. *For the case $U = U^{\text{MAX}}$: At each time T_n, for each tool $x \in K^{n-1}$, let*

$$G(x) := \sum_{t=0}^{T_n-1} v_t(x) \cdot 1_{v_t(x)=\max\{v_t(x');x' \in K_t\}} - T_n c_x.$$

And let $K^n := \{x \in K^{n-1} : G(x) \geq 0\}$, for the G function computed at that date.[20] *Or, in words, we ask for each remaining tool x, has the gross value accrued from holding x justified the price that has been paid ($T_n c_x$) for keeping x in the toolkit? If so, continue to hold x. If not, drop it.*

We confess that this heuristic is something of a strawman: A little thought should convince you that it is anything but reasonable.[21] One obvious problem is that, in computing $G(x)$, we have not taken into account the possibility that, at a particular date t, $v_t(x) = v_t(x') =$

[20] The notation $G(x)$ is used as short-hand for the evaluation measure of x; this measure depends, of course, on the sequence of toolkits $\{K_t; t \leq T_n - 1\}$ as well as the realizations up to time $T_n - 1$ of the valuation vectors v_t. We suppress this dependence in our notation.

[21] Despite its flaws, which we describe momentarily, we are told by colleagues who specialize in managerial accounting that this sort of evaluation criterion is in common use. Those colleagues could not, however, provide a textbook reference that recommends this heuristic. So insofar as it is a heuristic in use, it may simply be the product of sloppy thinking.

$\max_{x'' \in K_t} v_t(x'')$, for two different tools x and x'. In such circumstances, either x or x' would be tool chosen, but not both; however, the computation of $G(x)$ gives both x and x' "credit" in all such circumstances.

We could avoid this difficulty by assuming that there is probability zero that two tools tie for "best," however this points us in the direction of much more significant weaknesses in this heuristic, which we illustrate with the following two caricature examples.

Example 4.1. Suppose $X = \{x, x'\}$; $v_t(x) = 10$ and $v_t(x') = 9$, both with certainty, under μ^T;[22] $c_x = 5$ and $c_{x'} = 3$; and $U = U^{\text{MAX}}$. When we go to compute $G(x)$ at date T_1, it is positive and equal to $5T_1$, since up to time T_1 we carry both x and x', and x is the tool employed every time. Meanwhile $G(x') = -3T_1$ at T_1; since x' has never been employed, it fails the pay-for-itself test miserably. And, hence, in implementing this heuristic, the decision maker would drop x' at T_1 and persevere with the toolkit $\{x\}$ forever more. Of course, this is the wrong decision to take: The optimal toolkit in this caricature is $\{x'\}$. The heuristic is flawed in that it pays no attention to alternatives to the "best" tool in any situation, where "best" is determined with regard to the gross payoff. This is appropriate, of course, in making the daily decision which tool in the toolkit to use; the rents have been paid and so are sunk costs. But in terms of assessing the "value" of each tool going forward, it is very much the wrong thing to do.

Example 4.2. As in Example 4.1, $X = \{x, x'\}$ and $U = U^{\text{MAX}}$. The costs are $c_x = 2$ and $c_{x'} = 3$, and the distribution of the vector $v = (v(x), v(x'))$ under μ^T is that $v = (10, 9)$ with probability 1/2 and $= (9, 10)$ with probability 1/2. Suppose T_1 is large enough so that, at every evaluation stage, the frequencies of the two possible observations of v are close to 1/2. Then $G(x) \approx 5 - 2 = 3$, while $G(x') \approx 5 - 3 = 2$. Both tools always appear to be paying for themselves, so both are always kept, for average payoffs per period of $10 - 5 = 5$. But carrying tool x alone would give average payoffs per period of $9.5 - 2 = 7.5$, which (clearly) is better. The problem is the same: The decision maker should be looking at alternatives to carrying a given tool (or set of tools) and asking, "Does this improve matters?" In this heuristic, nothing of that sort is considered.

These caricature examples show us how to proceed. If the decision maker is evaluating a particular tool x, the question she should answer is, "Would I be better off with or without x?"

The Incremental-Contribution Heuristic.[23] *At time T_n, for any subset $L \subseteq K^{n-1}$ and for any*

[22] It might be worth pointing out that, in terms of the performance of prior-free heuristics, only μ^T matters. The full range of hypotheses and the decision maker's prior over these hypotheses play no role.

[23] While the Pay-for-Itself Heuristic was defined only for the special case $U = U^{\text{MAX}}$, this heuristic can be defined

$x \in L$, define the (empirical) ***incremental contribution of*** *x* ***to*** *L by*

$$I(x, L) := \frac{1}{T_n} \sum_{t=0}^{T_n - 1} \Big[W\big(v_t, L\big) - W\big(v_t, L \setminus \{x\}\big) \Big]. \tag{4.1}$$

(The notation here suppresses the dependence of the function I on the date T_n, the history of values $\{v_t; t = 0, \ldots, T_n - 1\}$, and the toolkit K^{n-1}.) At date T_n, temporarily set $L^0 = K^{n-1}$ and compute $I(x, L^0)$ for each tool $x \in L^0$. If all tools $x \in L^0$ have $I(x, L^0) \geq 0$, then let $K^n = K^{n-1} = L^0$ and proceed to T_{n+1}. But if $I(x, L^0) < 0$ for some $x \in L^0$, then select some $x \in L^0$ which has minimal $I(x, L^0)$—denote it by x'—and let $L^1 = L^0 \setminus \{x'\}$. Calculate the values $I(x, L^1)$ for each $x \in L^1$: If $I(x, L^1) \geq 0$ for all $x \in L^1$, set $K^n = L^1$ and proceed to T_{n+1}, but if $I(x, L^1) < 0$ for some $x \in L^1$, form L^2 by dropping from L^1 some x—call it x''—that minimizes $I(x, L^1)$. Continue in this fashion, either finding some L^m for which $I(x, L^m) \geq 0$ for all $x \in L^m$, at which point K^n is set to be L^m, or dropping (one at a time) every tool originally in K^{n-1}, in which case K^n is set to be $\emptyset$.

There are several things to note about this heuristic and the way in which it has been defined:

1. Unlike the pay-for-itself heuristic, this tool-based heuristic typically involves multiple recomputations of the "value" of a tool. The idea here is: In computing $I(x, L)$, we are asking whether the decision maker is better off with x or not, where the *not-x* alternative is, the decision maker keeps everything else in L. Because the *not-x* alternative is *everything else in L*, if we drop a tool x, we keep, at least temporarily, everything else, and then recompute incremental contributions. In Example 4.1, for instance, at time T_1, the average incremental contribution of x is -4. And the average incremental contribution of x' is -3. If the heuristic called for immediate dropping of any tool with a negative incremental contribution, both tools would be dropped, even though this toolkit ($K = \emptyset$) is the worst of all four possibilities. To avoid this, the heuristic drops one tool at a time and then reconsiders whether there are other candidates for dropping.

2. And, for Example 4.1, the heuristic gets it right. In the first iteration at time T_1, $I(x, X) = -4$ and $I(x', X) = -3$, so tool x is dropped. And then the decision maker recomputes $I(x', \{x'\})$, which now equals 6, so it is kept.

   More generally, but still in the realm of caricatures, suppose that $U = U^{\text{MAX}}$ and v has a degenerate distribution under μ^T; that is, $v(x) = v_x$ for some strictly positive v_x

for general U. But see the discussion following: This heuristic makes the most sense when U is submodular.

with probability 1. The right thing to do, of course, is to choose some tool that maximizes $v_x - c_x$. And that is what this heuristic chooses. In any "iteration" of the computations done at date T_1, all tools except the tool with maximal v_x (of those still in contention) have $-c_x$ for their incremental contribution. Suppose x^* is among the tools that maximize $v_x - c_x$; can x^* be eliminated in any of these iterations? There are two possibilities:

a. Suppose in some iteration, v_{x^*} is the largest v of those remaining. Then, in this iteration, $I(x^*) = v_{x^*} - v_{x'} - c_{x*}$, where x' is the tool with the second largest v_x of those remaining. Tool x^* is then eliminated only if $v_{x^*} - v_{x'} - c_{x^*} \leq -c_{x'}$ (as, otherwise, x' would be eliminated instead). But, if this inequality holds, then $v_{x^*} - c_{x^*} \leq v_{x'} - c_{x'}$. Since x^* was chosen to maximize $v_x - c_x$, this means that x' is equally good as x^*, and if x^* is eliminated on this iteration, then x' remains to take its place.

b. Suppose that v_{x^*} is not one of the largest. Then, on this iteration, $I(x^*) = -c_{x^*}$. Let x' be any x of those remaining that has the largest v_x. Then, $I(x') = v_{x'} - v_{x''} - c_{x'}$, where x'' is a remaining tool with second largest v_x. Hence, $I(x') \leq v_{x'} - v_{x^*} - c_{x'}$. Now, if x^* is eliminated in this iteration, $I(x') \geq I(x^*)$, so $v_{x'} - v_{x^*} - c_{x'} \geq -c_{x^*}$, or $v_{x'} - c_{x'} \geq v_{x^*} - c_{x^*}$, and while x^* might be eliminated, x' then remains, and x' is another tool that (amongst all tools in X) maximizes $v_x - c_x$.

The point is, if there is a unique x that maximizes $v_x - c_x$, it must survive all iterations; if more than one x achieves this maximum, one of those maximizing x must survive every iteration. And, in each iteration, until a single tool remains, some tool must have a negative I, the cycle of iterations at time T_1 must result in a single x remaining (which must, therefore, be a maximizer of $v_x - c_x$), or no x remains (which happens if the maximized value of $v_x - c_x < 0$).

3. Going back to this heuristic applied to Example 4.1, we've already observed that when time T_2 rolls around, the average incremental contribution of x' is computed to be 6, and tool x' is kept. An important point about the heuristic in general is made here: In computing $I(x)$, we compute *going back to time 0* the time-average incremental gross contibution of x relative to whatever set of tools L^m remains. In terms of ease of implementation, one might think instead of keeping a running sum of the incremental contributions of x, where at date t we compute this relative to K_t. That is, in place of the summands in the definition

of $I(x)$, imagine we used

$$\Big[W\big(v_t, K_t\big) - W\big(v_t, K_t \setminus \{x\}\big)\Big]. \tag{4.2}$$

This would make life easier when it comes to implementation, but in the caricature, it could lead to the wrong answer: If, say, $3T_1 > T_2$, this alternative method of evaluating each tool would, at time T_2, lead the decision maker to compute

$$I(x') = \frac{1}{T_2}\Big[T_1 \cdot 0 + (T_2 - T_1) \cdot 9\Big] - 3 = 9 \cdot \frac{T_2 - T_1}{T_2} - 3 < 0,$$

and so she would drop tool x'.

4. The function U^{MAX} has the property that, for any vector of values v_t, the contribution of tool x to the overall gross payoff does not increase as the toolkit K gets larger or, putting it a bit more intuitively, tool x can only become more valuable as the number of tools still in consideration gets smaller, assuming that x remains in the toolkit. Hence, for $U = U^{\text{MAX}}$, computing $I(x)$ as in the definition of the heuristic, with K^{n-1}, always gives a higher net incremental value to x than we would get if instead we computed gross incremental contributions using (4.2). And, more generally, for U^{MAX}, as tools are successively eliminated at any time T^n, the tools that remain will (only) see their I (weakly) increase.

   For more general U, the construction of the heuristic makes sense *if* this property—that each tool makes a (weakly) larger incremental contribution the smaller is the set of (other) tools still in the toolkit—holds; then a tool whose I is positive at some stage (and so is not a candidate for being dropped) will not later become a candidate for dropping, unless and until we receive more evidence that the tool is not as good as was assessed based on prior evidence. In symbols, the desired property is: For any tool x, any value-of-tools vector $v \in R^X$, and any two toolkits K and K' such that $x \in K' \subseteq K$,

$$W\big(v, K\big) - W\big(v, K \setminus \{x\}\big) \leq W\big(v, K'\big) - W\big(v, K' \setminus \{x\}\big). \tag{4.3}$$

   Observe that if W, regarded as a function on $(R_+)^X$, is submodular, then (4.3) holds.

   The following example shows one way in which this heuristic is flawed:

Example 4.3. In this example, $U = U^{\text{MAX}}$ and $X = \{x, x', x''\}$. Under μ^T, there are two possible values for the vector v: $(v(x), v(x'), v(x'')) = (9, 0, 10)$ and $= (0, 9, 10)$, each with

probability 1/2. The costs of x, x', and x'' are, respectively, 3, 3, and 5. Suppose T_1 is large so that, with high probability, the decision maker observes (approximately) $T_1/2$ value vectors $(9, 0, 10)$ and $T_1/2$ vectors $(0, 9, 10)$. The net incremental contributions computed in the first iteration are approximately -3, -3, and -4, respectively. Therefore, in the first iteration of computations at T_1, assuming "representative data," tool x'' is dropped. In the next iteration (still at date T_1), the incremental contributions of x and x' are approximately 1.5 apiece; the decision maker will stick with $\{x, x'\}$, for a payoff of 3 per period. Assuming that the data she accumulates continues to be close to the underlying distribution (which, of course, is likely), she will stick with $\{x, x'\}$ forever. But this is suboptimal: The toolkit $\{x''\}$ yields a per-period payoff of 5.

The problem is evident. The incremental-contribution heuristic asks "how much is lost or gained if one tool is dropped and all the rest are maintained?"[24] The example, though, is one where the finding the best toolkit requires an answer to the question, "How much is lost or gained if several tools at once are discarded?" Rather than evaluating individual tools and their net contribution, perhaps the decision maker should choose for her unit of analysis the toolkit. This suggests the following kit-level heuristic.

The Simple Set-Based Heuristic. *At time T_n, when the set of tools carried from date T_{n-1} until date $T_n - 1$ is K^{n-1}, compute for each $L \subseteq K^{n-1}$ the value $V(L)$ defined as*

$$V(L) := \frac{1}{T_n}\left[\sum_{t=0}^{T_n-1} W\big(v_t, L\big)\right].$$

The decision maker then chooses for K^n whichever subset $L \subseteq K^{n-1}$ maximizes $V(L)$ (choosing randomly if there is a tie). Described in words, the decision maker evaluates each subset L of K^{n-1} (including K^{n-1} itself) according to the empirical average payoff L would have provided from time 0 to the present moment, and she chooses to continue with whichever subset L looks best so far.

It is worth noting that this heuristic, relative to the incremental-contribution heuristic, involves a lot of evaluations at the dates $T_1, T_2, \ldots$. If (say) X contains ℓ tools, then at the first evaluation,

[24] This problem requires at least three tools. Francetich (2013) proves that if X consists of only two tools, and if U is submodular, the incremental-contribution heuristic provides the optimal toolkit with probability approaching one as T_1 approaches infinity.

at date T_1, the decision maker is evaluating $2^\ell - 1$ toolkits. In comparison, with the incremental-contribution heuristic, at date T_1, the decision maker must evaluate ℓ tools in the first stage, then $\ell - 1$ in the second stage (still at date T_1), and so forth (until all remaining tools provide a nonnegative incremental contribution), for a maximum number of evaluations of $\ell(\ell + 1)/2$.

The obvious tension in the simple set-based heuristic (and in the incremental-contribution heuristic) involves the amount of data that must be accumulated before decisions are made. From the perspective of obtaining accurate data (data that match the distribution μ^T), one presumably wishes T_1 to be large. But large T_1 means carrying—and paying for—all the tools in X, which (we later see) will give poor overall performance for δ much less than 1. It would be preferable to discard tools quickly that, based on the evidence accumulated, provide little chance of being part of the optimal toolkit while, at the same time, delaying the decision to discard a tool for whom the evidence is not so strong. In the world of prior-free heuristics, something like the following is suggested:

Set a threshold probability $\epsilon > 0$. At date T^n (where you should think of these evaluation times coming close together, perhaps even with $T^n = n$):

1. *For each $L \subseteq K^{n-1}$, compute $V(L)$ as in the simple set-based heuristic.*
2. *For each pair of toolkits L and L', both subsets of K^{n-1}, such that $V(L) > V(L')$, conduct a paired-sample, difference-of-means test on the data sets*

$$\{W(v_t, L); t = 0, \ldots, T_n - 1\} \text{ and } \{W(v_t, L'); t = 0, \ldots, T_n - 1\},$$

where we match the data point in the first set for a specific t with the tth data point in the second set. If the critical p-value (one-sided) for this difference-of-means test is less than ϵ, say that L' is statistically dominated by L.
3. *After all such tests have been run, discard any tool x that belongs (only) to toolkits that are statistically dominated by some other toolkit.*

The idea is straightforward: A tool is discarded as soon as the evidence accumulated indicates that each toolkit to which it belongs is "unlikely" to be the best toolkit. One might worry that elements of a toolkit K'' are discarded because K'' is statistically dominated by some toolkit K', when some members of K' are discarded on similar grounds. But for some members of K' to be discarded, K' must be statistically dominated by some K, and since the binary relationship

"statistical domination" is transitive,[25] this means that K'' must be statistically dominated by some toolkit all of whose members remain after this period's discards are made.

We imagine this heuristic applied without much time between revision periods. But, as presented here (and even with the optional part 4), it involves a significant number of computations. A simpler-to-implement version of this would be to identify the toolkit L with the best performance so far and compare all other toolkits with this toolkit's performance. When it comes to simulating the heuristics, this is what we will use:

A Difference-of-Means, Set-Based Heuristic. *Set a threshold probability $\epsilon > 0$. At dates $t = 2, 3, 4, \ldots$:*

1. *Find the toolkit $L \subseteq K_{t-1}$ that gives the highest value of $V(L)$. Call this toolkit L^*. (Although it can make a difference in step 2, choose L^* arbitrarily if there are ties.)*
2. *For all $L \subseteq K_{t-1}$, perform a paired-sample, difference-of-means-test, comparing the means of $\{W(v_t, L^*); t = 0, \ldots, t-1\}$ and $\{W(v_t, L); t = 0, \ldots, t-1\}$. Say that L is **dominated** if the critical p-value for the difference is less than ϵ.*
3. *Let K_t be the union over all $L \subseteq K_{t-1}$ that are undominated in step 2.*[26]

Two desirable qualitative properties of these heuristics

The acid test of these heuristics (and others) is how well they perform on a collection of test problems, using Monte Carlo simulation. We will report some simulation results later in this paper, but before doing so and before suggesting some prior-based heuristics, we look at a couple of seemingly desirable qualitative properties that they might satisfy.

The first concerns what happens when T_1 is chosen to be very large. Having T_1 large seems a sensible parametric choice for our four heuristics as the discount factor δ approaches 1, as large T_1 means gathering a lot of information before making any decisions that eliminate tools. The information is very likely to be "accurate," in the sense that the observed empirical frequency matches μ^T and so we ask of each of the heuristics: Suppose the empirical frequency at T_1

[25] This fact does not appear in the textbooks we have consulted, but it is nonetheless true and easy to prove once you note that the sample standard deviation of the sum of two (paired) data samples is less than or equal to the sum of the sample standard deviations of the two. We are grateful to Guido Imbens for pointing this out to us.

[26] In the simulations we later run, the discreteness of our test problems will present a problem: Sample standard deviations will often be 0 (because the observations, especially early on, may be identical. If $W(v_t, L^*) - W(v_t, L)$ is constant (and strictly positive) in t when we go to apply this heuristic, we adopt the convention that L^* dominates L in the sense of step 2. More generally, the classical difference-of-means tests we implement in this heuristic are formally justified for the case of Normally distributed variates and informally by an appeal to the Central Limit Theorem. For, say, $t = 2$, with our discrete test problems, any appeal to the CLT is somewhat silly, rendering this heuristic (as implemented) unsophisticated, to say the least.

matches *precisely* μ^T and remains consistent with μ^T at each subsequent decision point (that is, for those tools that are kept). In such ideal conditions, does the heuristic lead to an optimal choice of toolkit?

For both tool-based heuristics, we already know from our examples that the answer is No.

For the simple set-based heuristic, the answer to this question is obviously yes. If the empirical frequencies of the v_t vectors precisely match μ^T, then the heuristic identifies the (or, if there are ties, an) optimal toolkit. For the difference-of-means, set-based heuristic, the answer is almost yes. It is yes if Condition A holds; then, with a large enough T_1, every toolkit but the (uniquely) optimal toolkit will be statistically dominated by the optimal tookit (for any $\epsilon > 0$, although "large enough" T_1 will depend on the choice of ϵ). If more than one toolkit is optimal under μ^T, however, the decision maker could well wind up holding the union of all such toolkits.

The second desirable property is built out of the observation that, for each of these heuristics, the sets of tools in successive toolkits are ordered by set inclusion: Tools leave the toolkit, never to return. Since there are only finitely many tools, along any sample path of observations, there is a smallest took-kit that, past some point, is carried forever, although this final toolkit might be the empty set.

Hence, along each sample path, the decision maker eventually learns the distribution of $(v(x))$ for those x in this final toolkit. While abandoning tools means that no further data about the distribution of $v(x)$'s of dropped tools may be gathered, we would at least like to know that, in the long run, the final toolkit, which we denote by $\hat{K}$,[27] is at least as good as any proper sub-kit $K' \subset \hat{K}$ (with probability one).

For the simple set-based heuristic, the answer is obviously yes. If a sub-kit K' is strictly better, the strong-law says that the empirical-frequency-based computation of its net value will (with probability one) exceed that of $\hat{K}$, and then the heuristic will move off from $\hat{K}$, contradicting the assumption that $\hat{K}$ was the final toolkit. Similar logic implies that the answer is yes for the difference-of-means heuristic, if Condition A holds. But if there is more than one optimal toolkit under μ^T, the decision maker could end with their union, which is strictly worse than any of the optimal toolkits.

For the Pay-for-Itself heuristic, the answer is no. Example 4.2 applies: The heuristic retains both x and x' (assuming data are close to the probabilities under μ^T), while either single-tool sub-kit is better than the toolkit $\{x, x'\}$.

As for the incremental-contribution heuristic, the answer is also no, for general U.

[27] Note, please, that $\hat{K}$ can be random; we know there is a final toolkit along each sample path, but that final toolkit may be different depending on the sample path.

Example 4.4. $X = \{x, x', y, y'\}$. Under μ^T, the distribution of the v vector (in the order x, x', y, y') is (10, 10, 0, 0) with probability 1/2 and (0, 0, 10, 10) with probability 1/2. Costs are $c_x = c_{x'} = c_y = c_{y'} = 3$. And

$$U\big(v(x), v(x'), v(y), v(y')\big) = \max\big\{\min\{v(x), v(x')\}, \min\{v(y), v(y')\}\big\}.$$

We assert that, under the incremental-contribution heuristic, a decision maker who has empirical frequencies that match (or come close to) μ^T will keep $K = \{x, x', y, y'\}$. The incremental contribution of x is 2; dropping x means the loss of 10 half the time, or 5 on average, more than the 3 saved in costs. Since the problem is symmetric in all four tools, the same is true of x', y, and y'. So, starting with all four tools, none is ever dropped according to this heuristic. However the per-period payoff is $10 - 12 = -2$; obviously, the empty subset is better, as are both $\{x, x'\}$ and $\{y, y'\}$.

However, this function U does not satisfy property (4.3):

Proposition 4. *Suppose that U satisfies property (4.3). Fix a (nonempty) toolkit K. If, under the distribution of v vectors given by μ^T, some proper sub-kit $K' \subset K$ gives higher per-period payoffs than does K—that is, $w_{i^T}(K') > w_{i^T}(K)$—then, for data samples that are close to matching in frequencies the distribution given by μ^T, some tool $x \in K$ must have strictly negative incremental contribution. Hence, for almost every sample path,*[28] *if $\hat{K}$ is a "final toolkit" along that sample path under the incremental-contribution heuristic,*

$$w_{i^T}(\hat{K}) \geq w_{i^T}(K') \quad \textit{for all } K' \subseteq \hat{K}.$$

We provide a detailed proof in the appendix. But the idea is easy to relate: (1) If, relative to a toolkit K and a tool $x \in K$, $I(x, K) \geq 0$ under some data set, then (using that data set) K has a weakly higher (empirical average) net reward than does $K \setminus \{x\}$: This is just a rearrangement of the inequality $I(x, K) \geq 0$. And (2) if, for a given data set, $I(x, K) \geq 0$ for every $x \in K$, then, for every sub-kit $K' \subset K$ and $x \in K'$, $I(x, K') \geq 0$. (This is where (4.3) comes in. Reducing the toolkit (weakly) improves the incremental contributions of x, along every sample path.) Hence, for a given data set, if $I(x, K) \geq 0$ for every $x \in K$, then the empirical average

[28] "Almost every" here refers both to the decision maker's initial prior and to the "objective" probability distribution on sample paths generated by μ^T.

net payoff from K, computed with the data set, is at least as large as for any proper subset of K: Starting from K, sequentially delete elements until you reach the proper sub-kit. Assertion (1) ensures that the empirical average net payoffs never increase, where (2) ensures the property needed to apply (1) repeatedly, namely that the incremental contributions that begin nonnegative remain so. And for a "final toolkit," the data sample will, by the strong law of large numbers, approach the probability distribution of μ^T.

5. Prior-Based Heuristics

The use of the decision maker's prior assessment π^0—and the full probabilistic structure implicit in the $\{\mu_i\}$—provides significantly more scope for the design of seemingly sensible heuristics. In fact, an active literature that spans the disciplines of computer science and operations research, going by the names *bandit learning* and *online optimization*, concerns a variety of categories of heuristics (generally called *algorithms* in this literature), their asymptotic characteristics, and their relative performance in test problems that often take the form of multi-armed bandits. Borrowing in part from that literature, we present here a variety of prior-based heuristics.

A. Adaptive Myopia

To set a (seeming) baseline, we begin with a heuristic that ignores the exploitation/exploration dilemma: At each date, K_t is chosen to maximize the immediate expected payoff. If information happens to arrive, it is employed; the decision maker updates her prior and chooses at date t based on her posterior. But she makes no active attempt to gain information.

Adaptive Myopia. *At each date t, if (based on all information the decision maker has received up to time t) the decision maker's posterior assessment is π^t, she chooses K_t arbitrarily out of $\mathcal{K}^*(\pi^t)$.*

B. Simulated annealing, ϵ-greed, and variations

In this general category of heuristic, the decision maker "mostly" chooses whichever toolkit is myopically optimal based on information received to date, but some (perhaps vanishingly small) fraction of the time she experiments with other toolkits. A simple specific version of this is the following:

Harmonic Sampling. *At each date t, K_t is selected randomly: With probability $t/(t+1)$, choose for K_t some myopically optimal toolkit $K \in \mathcal{K}^*(\pi^t)$ (arbitrarily selected if there is more than one); and with probability $1/(t+1)$, choose $K_t = X$.*

Several remarks are worth making:

1. The specific heuristic described allows for "experimentation" at every date, with vanishing probability, but with probably that vanishes slowly enough so that, almost surely, X will be chosen infinitely often. An alternative way to proceed, which has the same asymptotic outcome, is to fix in advance some set of dates $\mathcal{T} \subseteq \{0, 1, \ldots\}$ with the property that

$$\lim_{\tau \to \infty} \frac{\#[\mathcal{T} \cap \{0, \ldots, \tau\}]}{\tau} = 0,$$

   where $\#[\cdot]$ means the cardinality of the set inside the square brackets; then choose $K_t = X$ for $t \in \mathcal{T}$ and choose the myopically optimal strategy at all dates in the complement of $\mathcal{T}$. This, recall, is the approach taken in the proof of Proposition 2.

2. We have included this heuristic as an example of a prior-based heuristic. The prior and subsequent posteriors computed from the prior and available evidence are employed at those dates where a myopically optimal toolkit is chosen, to determine which toolkit is optimal. We could just as well have included variations of this heuristic as a prior-free heuristic, if "myopic optimality" is computed on the basis of empirical frequencies of the various components of the v vector, so long as we begin with at least one experimentation period.

3. In typical applications of simulated annealing to multi-armed bandit problems, when an experiment is to be conducted, each arm is chosen with probability equal to one divided by the number of arms. Because the choice of toolkit X generates all the information possible in a given period, we don't need to do this; all experiments in our heuristic involve choosing $K_t = X$. But many variations are possible where we vary the "experimental toolkit" and where we furthermore adjust the probability of experimentation with a specific toolkit to the degree of uncertainty of its value and/or the promise of value that it holds.

   The proof of Proposition 2 is easily amended to provide the following result:

Proposition 5. *If the decision maker employs the harmonic-sampling heuristic (or any alternative for which the event $\{K_t = X$ infinitely often$\}$ has probability one), π^t converges to a degenerate distribution with weight 1 on μ^T almost surely. Hence, for this specific heuristic (and any alternative for which, almost surely, $K_t = X$ infinitely often , but K_t is chosen from $\mathcal{K}^*(\pi^t)$ a proportion of the time that approaches one), the Cesàro sums of $W(v_t, K_t)$ approach w^* with probability 1.*

A variation, called the ϵ-*greedy* algorithm,[29] experiments in every period with fixed prob-

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

ability $\epsilon > 0$; i.e., the probability of experimentation in any one round does not vanish as time passes.

C. Thompson Sampling

Thompson Sampling. *For each i, fix (arbitrarily) some toolkit from $\mathcal{K}_i^*$, which we denote K_i^*.*[30] *Then, at each date t, choose K_t randomly: With probability π_i^t, choose K_i^*.*

This heuristic was originally proposed in Thompson (1933) and, therefore, is almost certainly the seminal heuristic of its general type.

An obvious variation on this heuristic is to take into account the possible gains from each $K \in \mathcal{K}_i^*$: Choose from $\mathcal{K}_i^*$ at date t with probability proportional to $\pi_i^t \times (w_i^* + a)$, for some positive constant a.[31]

The story that motivates this heuristic may seem a bit forced, but for what it is worth: At date t, the decision maker believes that $\mu^T = \mu_i$ with probability π_i^t. So she "simulates" which hypothesis is true, selecting μ_i with its probability of being true, and then selects a best toolkit according to the outcome of this simulation.

In general, Thompson sampling may not lead the decision maker to the truth. There are two ways this can happen. First, suppose some toolkit K is optimal for both μ_i and μ_j and, moreover, the distribution of $\{v_t(x); x \in K\}$ is the same under μ_i and μ_j. (Some tool not in K might satisfy our assumption that the overall distributions of v under different hypotheses are different.) If μ_i and μ_j are the only two hypotheses, the decision maker always chooses K and, of course, she is unable to learn which hypothesis is true. Second, suppose there are two tools, $X = \{x, x'\}$, and two possibilties for μ^T: Under μ_1, $v_t = (v_t(x), v_t(x')) = (7, 6)$ or (3,4), each with probability one-half, while under μ_2, $v_t = (7, 4)$ or $(3, 6)$, each with probability one-half. The cost of each tool is 5. Suppose $U = U^{\text{MAX}}$. Then both $\{x\}$ and $\{x'\}$ are optimal under both hypotheses. If the decision maker implements Thompson sampling by choosing $\{x\}$ to be K_1^* and $\{x'\}$ to be K_2^*, then she never learns anything about μ_1 versus μ_2, because the distribution of net per-period payoffs from $\{x\}$ is the same under both hypotheses, and similarly for $\{x'\}$. (The distribution of payoffs from $\{x\}$ is different from that of payoffs from $\{x'\}$ under either hypothesis, however.)

Of course, in neither example does she care whether she learns the truth. So we still have the possibility of proving for Thompson sampling the second part of Proposition 5. But to get

[30] Fixing one K_i^* from each $\mathcal{K}_i^*$ simplifies the proof of Proposition 6, and so we do so. But we do not believe it is necessary.

[31] The constant a is included to deal with the possibility that $K_i^* = \emptyset$ for some i.

the first part, we need to rule out both sorts of examples.

Proposition 6. *If Condition C holds and the decision maker employs the Thompson-sampling heuristic (or its variation), then her posterior assessments π^t converge to a degenerate distribution on μ^T almost surely. And even if Condition C does not hold, under Thompson sampling (or the variation), the Cesàro sums of $W(v_t, K_t)$ approach w^* with probability one.*

The proof is provided in the appendix.

D. Upper-Confidence-Bound Heuristics

Slightly bridled optimism. *Fix some $\epsilon > 0$. Fix, for each i, some $K \in \mathcal{K}_i^*$, denoting this choice by K_i^*.*[32] *At time t with posterior π^t, let*

$$m^*(\pi^t;\epsilon) := max\{w_i^*; i = 1,\ldots,I, \pi_i^t > \epsilon\} \quad and \quad I^*(\pi^t;\epsilon) := \{i : \pi_i^t > \epsilon, w_i^* = m^*(\pi^t;\epsilon)\}.$$

For K_t, choose any member of K_i^* for any $i \in I^*(\pi^t;\epsilon)$.

In words, consider all the hypotheses μ_i that, per the current posterior, have probability more than ϵ. (Call such hypotheses "plausible.") Choose for the current toolkit any toolkit that is myopically optimal for the plausible hypothesis that, if true, would give the highest per-period net expected reward. This is the optimism part of the heuristic; the decision maker goes with the *most optimistic* (plausible) scenario available, and sticks with that scenario until information received suggests that some other (plausible) scenario gives a chance of doing better, which could happen either because data indicate that the plausible hypothesis on which basis the toolkit was selected is no longer plausible, or because some previously implausible hypothesis becomes (by virtue of the data) plausible. "Slightly bridled" refers to the the plausibility restriction: There must be a reasonable chance of the scenario, per information so far gathered (and the prior).

If it is unclear why we call this an upper-confidence-bound (UCB) heuristic, consider the following prior-free heuristic (which is more typical of UCB heuristics/algorithms in the literature.)

A Prior-free UCB Heuristic. *Fix some integer $T > 0$, and choose $K_t = X$ for all $t < T$. At time T, compute for each toolkit K the sample average of the net payoff it would have provided, denoted $m(K,t)$, and the sample standard deviation of those payoffs, $s(K,t)$, where the data*

[32] As in the case of Thompson sampling, this simplifies the proof of Proposition 7, although we believe Proposition 7 remains true without this.

sample includes all periods up to t in which K or a superset of K was the chosen toolkit. And, for some fixed parameter $\alpha > 0$, in period t choose K_t as that toolkit that maximizes $m(K,t) + \alpha s(K,t)$.

This heuristic is an "upper-confidence-bound heuristic" in the following sense: Depending on the value of α, the closed interval $\left[m(K,t) - \alpha s(K,t), m(K,t) + \alpha s(K,t)\right]$ is, in the usual fashion, a classical-statistics confidence interval for the true average payoff that K will generate.[33] So, after an initial period of data collection, the heuristic chooses that toolkit the upper bound of whose confidence interval is largest.

Slightly-bridled optimism, as defined, can be equivalently recast as follows, which comes closer to the spirit of the prior-free UCB heuristic.[34] (a) At time t and for the current posterior π^t, call μ_i a plausible hypothesis if $\pi_i^t \geq \epsilon$. (b) For each toolkit $K \subseteq X$ (and, implicitly, for the given posterior π^t), define UCB(K) := max$\{w(\pi, K); \pi$ puts weight only on plausible hypotheses $\mu_i\}$.[35] (c) For K_t, choose a toolkit that maximizes the current values of UCB(K).

If ϵ is fixed, then (of course) this heuristic has a chance of, in the long run, winding up with the "wrong" toolkit: Suppose, for instance, there are two tools, x and x', two hypotheses μ_1 and μ_2, and $U = U^{\text{MAX}}$. Suppose $\{x\}$ is the optimal toolkit under μ_1 and $\{x'\}$ is optimal under μ_2. Suppose that $v(x)$ has the same degenerate distribution under both μ_1 and μ_2, so choosing $\{x\}$ provides no information. And suppose that $\pi_2^0 < \epsilon$. Then, $\{x\}$ is chosen and continues to be chosen—the posteriors are all identically π^0—despite the fact that there is prior (and posterior) probability π_2^0 that this is the "wrong" toolkit.

And even if we wind up with the "right" toolkit, in some cases the decision maker can fail to learn the truth. Consider the prior example but where $\{x\}$ is the best toolkit under both hypotheses. The decision maker starts (and ends) with the right toolkit, but she fails to learn whether μ^T is μ_1 or μ_2.[36] (And this can happen even though the distribution of $v(x')$ is different under the two hypotheses about μ^T.) So, on both these grounds, we cannot duplicate the conclusions of Propositions 5 or 6 for this heuristic. But, we can get something close.

---

[33] A slightly more sophisticated version of this heuristic would have α depend on the number of samples on which basis $m(K,t)$ and $s(K,t)$ have been computed.

[34] Benjamin Van Roy, who has been very helpful in acquainting us with the literature on bandit learning, suggested in particular that we recast our heuristic in this fashion.

[35] But while closer in spirit, this is not quite the same. A different UCB-style heuristic, which is very similar in spirit to the prior-free version, is to evaluate, for each K, the full distribution of net returns it would generate under the current posterior and then find the value that falls at, say, the 95th percentile of that distribution. Choose for K_t whichever K has the largest 95th percentile value.

[36] We are *not* asserting that she is at all troubled about this.

Proposition 7. *Fix a model and, in particular, fix some prior π^0 over the (fixed) set of hypotheses $\{\mu_1, \ldots, \mu_I\}$. Let $P(\epsilon)$ denote the probability (with respect to the decision maker's prior) of the event*

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

computed if the decision maker employs the slightly-bridled-optimism heuristic with threshold probability ϵ.[37] *Then $\lim_{\epsilon\downarrow 0} P(\epsilon) = 1$.*

The proof is given in the appendix.

In the heuristics so far described, the discount factor δ is irrelevant to the decision maker's behavior. So, while Propositions 5, 6, and 7 can be interpreted as saying that these are "good" heuristics for problems with δ close to one (so that, asymptotically, all that matters is behavior in the tail field of events), they say little about how well these heuristics do in terms of initial expected discounted values, discounted at some fixed $\delta < 1$. To put it most starkly, if $\delta = 0$, we know the answer, and none of these three heuristics (necessarily) comes anywhere close. By continuity, the same is true for δ in some neighborhood of zero.

Moreover, both Thompson sampling and slightly bridled optimism restrict the decision maker to the use of toolkits that are optimal for some one of the μ_i. From the perspective of getting a good expected discounted value for a specific δ, this restriction can be ill-considered. Two reasons why are provided in the following two simple examples.

Example 5.1. $X = \{x, x', y\}$, $c_x = c_{x'} = 5$ and $c_y = 3.9$. There are two possibilities for μ^T. Under the first, μ_1, $v = (v_x, v_{x'}, v_y) = (14, 10, 12)$ with probability 0.9 and $(10, 14, 12)$ with probability 0.1, while under μ_2, $v = (14, 10, 12)$ with probability 0.1 and $(10, 14, 12)$ with probability 0.9. The U function is U^{MAX}. It is easy to compute that $\{x\}$ is optimal under μ_1 and $\{x'\}$ is optimal under μ_2. Suppose the decision maker begins with $\pi^0 = (0.5, 0.5)$. The (strict) myopic best toolkit is $K = \{y\}$, so for $\delta = 0$, that is the optimal strategy. (Nothing is learned, so the decision maker never moves from this toolkit.) By continuity of value functions in δ, this will remain the optimal strategy for δ close to zero. For δ sufficiently high, the optimal strategy is to choose $\{x\}$ when $\pi_1^t \geq 0.5$ and to choose $\{x'\}$ when $\pi_1^t \leq 0.5$. (At $\pi_1^t = 0.5$,

[37] Note that as we shift ϵ, the random variables π^t and K^t, viewed as (random) functions on the state space, change, since they depend on both the realization of the basic $\{v_t\}$ random process and decisions made by the decision maker. Hence, it might be better to write $\pi^t(\epsilon)$ and $K^t(\epsilon)$.

both $\{x\}$ and $\{x'\}$ are optimal. And for the crucial value of δ that separates the two regimes, so is $\{y\}$.)

This example is simple enough so that the fully optimal solution (for all values of δ and priors π^0) can be computed via value iteration. The optimal strategy is time homogeneous; that is, the (optimal) choice of toolkit at time t depends only on δ and π_1^t (the posterior probability that μ_1 is μ^T): For $\delta \geq 0.7641$ (approximately), the decision maker should choose either $\{x\}$ or $\{x'\}$ for all π^t, with $\{x\}$ chosen if $\pi_1^t \geq 0.5$. But for $\delta \leq 0.7641$, there is a range of posterior values (centered at 0.5 and larger the smaller is δ) for which it is optimal to choose $\{y\}$. See Figure 1. It is perhaps worth observing that even if the decision maker begins with a prior π_1^0 that makes, say, $\{x\}$ the optimal initial toolkit, if $\delta < 0.7641$, it is possible that a "bad draw" ($v_1(x) = 10$) leads to a posterior such that the optimal choice for K_1 is $\{y\}$, at which point nothing more is ever learned, and $\{y\}$ continues (forever) to be the optimal choice.

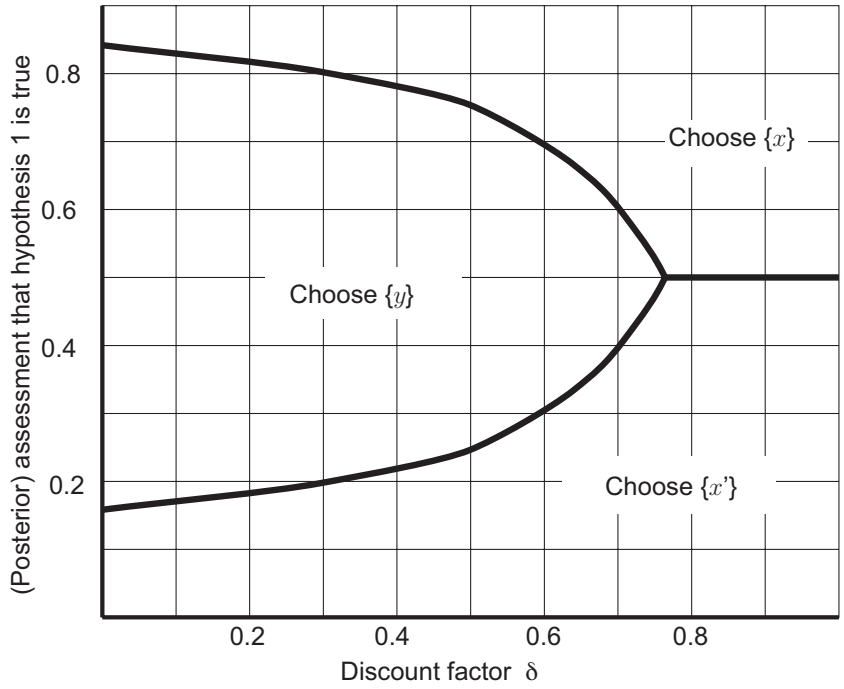

Figure 1. The optimal strategy for Example 5.1

The point here should be obvious: In the example, toolkit $\{y\}$ is, for some discount factors and prior assessments, a good compromise toolkit, even though it is not part of the optimal toolkit for either hypothesis. Thompson sampling and slightly bridled optimism give no consideration to such a toolkit.

Example 5.2. $X = \{x, x', x''\}$, $c_x = c_{x'} = 5$, and $c_{x''} = 1.2$. U is U^{MAX}. There are two possibilities for μ^T. Under the first, μ_1, $v = (v_x, v_{x'}, v_{x''}) = (8, 1, 0.9)$ with probability 0.9 and $= (8, 15, 1.1)$ with probability 0.1; under μ_2, $v = (8, 1, 0.9)$ with probability 0.1 and $= (8, 15, 1.1)$ with probability 0.9. The optimal toolkit under μ_1 is $\{x\}$, and $\{x'\}$ is optimal under μ_2. Hence, x'' is never part of a toolkit chosen by a decision maker using either the Thompson-sampling or the slightly-bridled-optimism heuristic.

Indeed, x'' is never chosen by a decision maker for its immediate (myopic) value as a tool, it is part of no toolkit in any $\mathcal{K}^*(\pi)$ for any π. By itself (for the toolkit $\{x''\}$), it generates negative immediate net payoffs; and in a toolkit with either x or x', its value is always less than the value of the other available tool(s).

Yet, in the fully optimal solution to the problem (which we can derive by computational methods, because the problem is so simple), $\{x''\}$ is sometimes part of the optimal toolkit. As in Example 5.1, the optimal choice of toolkit depends on the discount factor δ and the posterior π_1^t. For $\delta \leq 0.416$ (approximately), an increasing, real-valued function $\pi^*(\delta)$ with $\pi^*(0) = 0.5$ divides the space into two: For $\pi_1^t \leq \pi^*(\delta)$, choose $\{x'\}$, while for $\pi_1^t \geq \pi^*(\delta)$, choose $\{x\}$. For $\delta \geq 0.416$, the optimal choice of toolkit based on π^t is: choose $\{x'\}$ if $\pi_1^t \leq 68/112$; choose $\{x, x''\}$ if $68/112 \leq \pi_1^t \leq \pi^{**}(\delta)$ for some function $\pi^{**}(\delta)$; and choose $\{x\}$ for $\pi_1^t \geq \pi^{**}(\delta)$.

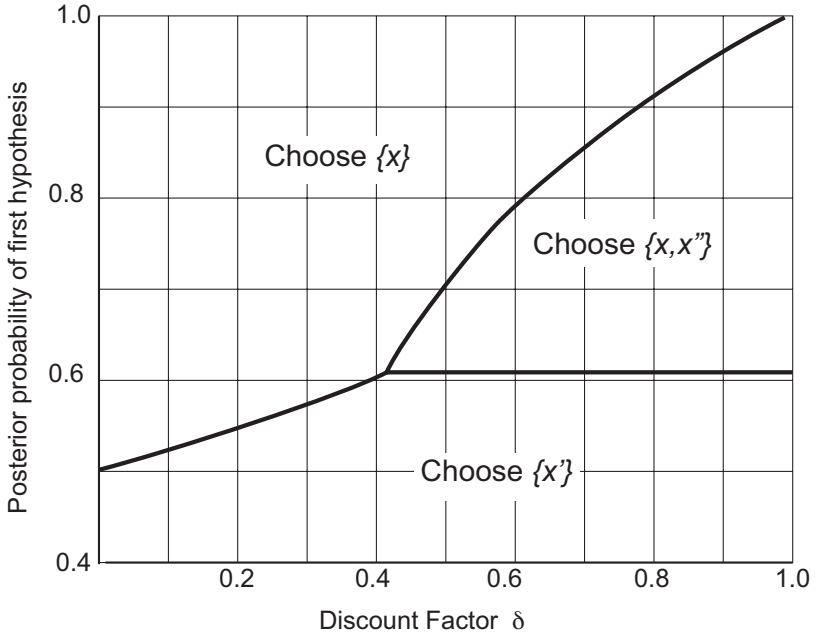

Figure 2. The optimal strategy for Example 5.2

The intuition here is straightforward: If $\pi_1^t \leq 0.5$, the myopically optimal toolkit is $\{x'\}$. Since this toolkit also provides as much or more information than any alternative, this is the obvious choice. For $\pi_1^t > 0.5$, the myopically optimal toolkit is $\{x\}$, but this provides no information; therefore, for π_1^t low enough (but greater than 0.5), it makes sense to choose a toolkit that provides information; moreover, the greater is δ, the more this information is worth, so the greater is the range of π_1^t for which it is best to choose a toolkit that provides information. There is no point in carrying both x' and x'', since they provide the same information and x' dominates x'' for immediate purposes. But what about $\{x'\}$ versus $\{x''\}$ versus $\{x, x'\}$ versus $\{x, x''\}$? These all provide exactly the same information, so their relative values are determined by the immediate net payoffs they generate. (That is, in Bellman's equation, the values of optimal continuation from all four are identical.) At this point, it is easy to compute that, between the four, the best immediate expected value comes from $\{x'\}$ if $\pi_1^t \leq 68/112$ and from $\{x, x''\}$ if $\pi_1^t \geq 68/112$. Finally, there is the cutoff value given by $\pi^*(\delta)$ for $\delta \leq 0.416$ and $\pi^{**}(\delta)$ for $\delta \geq 0.416$, where the immediate cost of information balances the future value from having the information; these cutoff values can only be derived numerically.[38]

So why is x'' sometimes chosen (along with x)? Because it is a cheaper way to get the same information as is obtained from $\{x'\}$. And, if the decision maker is choosing to get that information anyway, for $\pi_1^t \geq 68/112$, $\{x, x''\}$ is informationally equivalent to $\{x'\}$ and is a better toolkit in terms of immediate net expected payoffs. That is, *x'' may seem a "useless" tool in the sense that, on the job, it will never be used. But, it is a cheap source of information*. The first part of this—that it is useless in terms of immediate needs and so would not be chosen by a decision maker who knows μ^T—is why it is part of no toolkit from any $\mathcal{K}_i^*$ and so is ignored by both Thompson sampling and slightly bridled optimism. But the second part is why, when the decision maker is trying to learn which μ_i is the truth, it may be a very useful tool to put into her toolkit.

E. Approximate-dynamic-programming heuristics

These considerations lead to our final category of prior-based heuristics, based on the literature on approximate dynamic programming:[39] The reason we look for heuristics is because the dynamic programming problem that would give the full solution to our problem is too difficult to solve. Approximate dynamic programming suggests ways in which the methods of dynamic programming might be employed "partially" in such situations.

[38] It is more accurate to say that *we* are unable to derive them except via computation. Note that $\delta = 0.416$ (approximately) is where $\pi^*(\delta) = \pi^{**}(\delta) = 68/112$.

[39] See, for instance, Bertsekas (2012).

Consider, for instance, the following: Define

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

This is the best the decision maker can do when her assessment 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*. Define iteratively, for $m = 1, 2, \ldots,$

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

where $\tilde{\pi}'$ represents the random posterior the decision maker will assess on the basis of her prior π and information she receives in a single period from choosing K. Or, in words, $u^m(\pi)$ is the optimal 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. And let $K^{*m}(\pi)$ be any toolkit that achieves the maximum in the definition of $u^m(\pi)$. (If there is a tie for the best toolkit, an arbitrary selection should be made.)

By standard results in dynamic programming for this sort of problem (bounded per-period rewards, discounted with $\delta < 1$), we know that $u^m(\pi)$ converges, as $m \to \infty$, to the optimal value function. And (as long as the selection made in the event of ties is made consistently) the $K^{*m}(\pi)$ "settle down" to the optimal toolkit as a function of the posterior π. So, if we could carry out these calculations for large m, there would be no point to this paper. But, instead, 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 myopia-shortly heuristic. *Pick a (relatively) small positive integer m. At time t, when the decision maker's posterior is π^t, choose for K_t the toolkit $K^{*m}(\pi^t)$.*

By a relatively small m, we mean: $m = 1$ or, perhaps, 2.[40]

It may be instructive to see what this heuristic generates (in terms of strategy) for our examples. The myopia-shortly heuristic, because it uses an "underestimate" of the value of information, will tend to favor strategies that expend fewer resources on obtaining information than is optimal. In Example 5.1, for instance, the region of posteriors (for a given δ) for which $\{y\}$ is optimal should grow. This effect should be diminished the smaller is δ; the less the

[40] In the terminology of approximate dynamic programming, this heuristic for $m = 1$ is a *rollout* strategy.

future matters, the less an underestimate of the value of information should matter. So what do we get in Example 5.1? In Table 1, we provide the $\{x'\}$ to $\{y\}$ cutoff-posterior values for the optimal strategy and for the strategies derived from myopia-shortly with $m = 1$ and $m = 2$, for $\delta = 0.95, 0.9, 0.85, \ldots, 0.5$. (The cutoff posteriors between using $\{y\}$ and $\{x\}$ are symmetric.) It is clear from the numbers that $m = 1$ provides a fairly good approximation to the optimal solution, and $m = 2$ provides an extremely good approximation.

Discount factor	Optimal cutoff posterior	Cutoff posterior for m = 2	Cutoff posterior for m = 1
0.5	0.2452	0.245	0.245
0.55	0.2673	0.265	0.255
0.6	0.2983	0.295	0.283
0.65	0.3383	0.335	0.295
0.7	0.3934	0.385	0.315
0.75	0.472	0.465	0.345
0.8	0.5	0.5	0.385
0.85	0.5	0.5	0.475
0.9	0.5	0.5	0.5
0.95	0.5	0.5	0.5

Table 1. Comparing the optimal strategy in Example 5.1 with the strategies derived by the myopia-shortly heuristic for $m = 1, 2$.

And, doing a similar analysis for Example 5.2, yields the following: Since the cutoff posterior between choosing $\{x'\}$ and $\{x'', x\}$ is based entirely on immediate payoff considerations (the information content of the two toolkits is the same), this cutoff is 68/112 for all m. The cutoff that is sensitive to the level m is the cutoff-posterior between choosing $\{x'', x\}$ and $\{x\}$ alone; this is where the value of information comes in. Table 2 provides the cutoffs for the optimal strategy and for myopia shortly, for $m = 1, 2$, and for the same set of discount factors as in Table 1.[41] As with Example 5.1, we see that myopia-shortly for $m = 1$ provides a fairly good approximation to the optimal strategy, and $m = 2$ provides a very good approximation. Even for $\delta = 0.95$, the range of posteriors for which the optimal strategy and myopia shortly for $m = 2$ disagree is $\pi_1^t \in (0.971, 0.983)$. And, of course, over that range, the cost of a "mistake" is apt to be small, since the value functions are, of course, very close to one another near the critical cutoff levels.

[41] The numbers in Tables 1 and 2 for $n = 2$ are derived numerically. Subject to roundoff in the numerical procedures, they are accurate to three decimal places. For $m = 1$, the cutoffs can be derived in closed form: For Example 5.1, the formula is $(50 - 5\delta)/(320 - 248\delta)$ for $\delta \leq 35/44$ and $\min\{0.5, (25\text{-}16\delta)/(160(1 - \delta))\}$ for $\delta \geq 35/44$. For Example 5.2, the upper cutoff is $0.9 - 3(1 - \delta)/(14\delta)$.

Discount factor	Optimal cutoff posterior	Cutoff posterior for m = 2	Cutoff posterior for m = 1
0.5	0.706	0.704	0.686
0.55	0.751	0.749	0.725
0.6	0.791	0.788	0.757
0.65	0.825	0.822	0.785
0.7	0.856	0.853	0.808
0.75	0.884	0.88	0.829
0.8	0.91	0.905	0.846
0.85	0.934	0.928	0.862
0.9	0.962	0.952	0.876
0.95	0.983	0.971	0.889

Table 2. Comparing the optimal strategy in Example 5.2 with the strategies derived by the myopia-shortly heuristic for $m = 1, 2$.

The examples might lead the reader to be overly impressed with how well myopia-shortly seems to do. But Examples 5.1 and 5.2 are, to some extent, "cooked' to have this happen. For one thing, the information imparted by different toolkits is simple: One tool provides no information at all, and the other tools all provide the same information. And, for a second thing, once the decision maker commits to getting information, the information in a single draw is fairly decisive: From a prior of $\pi_1^0 = 0.5$, the first posterior is either 0.1 or 0.9. Suppose that we keep the qualitative structure of Example 5.2, but instead of such decisive information, the distribution of v under μ_1 gives two values with (respective) probabilities 0.6 and 0.4, with the reverse probabilities under μ_2. Then, one informative signal isn't going to shift the prior by much. To flesh this out:

Example 5.3. This example has the same structure as Example 5.2, but with the parameters changed: $X = \{x, x', x''\}$, $c_x = c_{x'} = 5, c_{x''} = 0.1$. Under μ_1, $v = (v(x), v(x'), v(x'')) = (15, 10, 0.9)$ with probability 0.6 and $= (15, 20, 1.1)$ with probability 0.4; under μ_2, the support of v is the same, and the probabilities switch to 0.4 and 0.6, respectively. Therefore, $K_1^* = \{x\}$ and $K_2^* = \{x'\}$. U is U^{MAX}.

The fully optimal strategy is depicted in the manner of Figures 1 and 2 in Figure 3. The value of π_1 at which the decision maker shifts from $\{x'\}$ to $\{x, x'\}$, assuming she wishes to gather information at all, is $\pi_1 = 11/20$. And Table 3 gives the upper cutoff values for myopia shortly, $m = 1, 2$.

While the myopia-shortly-derived strategies come close to matching the optimal strategies for discount factors $\delta \leq 0.7$ or so, they do less and less well as δ approaches 1. In fact, for any

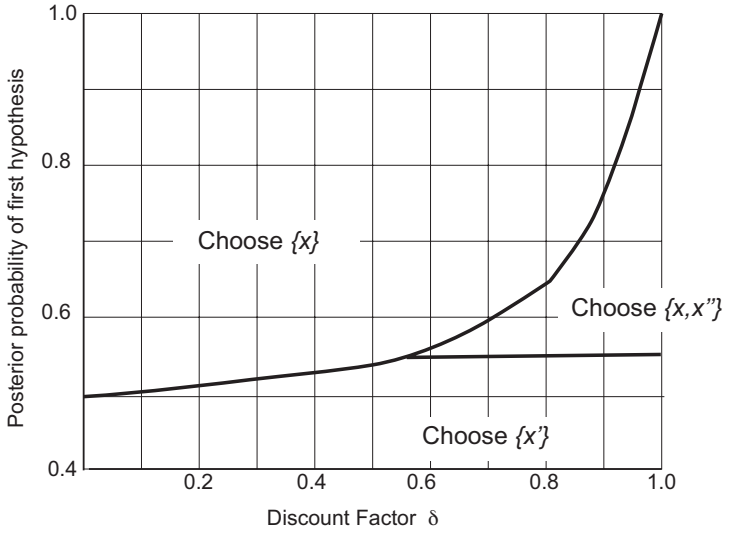

Figure 3. The optimal strategy for Example 5.3

Discount factor	Optimal cutoff posterior	Cutoff posterior for m = 2	Cutoff posterior for m = 1
0.5	0.536*	0.536*	0.533*
0.55	0.543*	0.542*	0.538*
0.6	0.551	0.549*	0.543*
0.65	0.572	0.567	0.548*
0.7	0.595	0.586	0.558
0.75	0.618	0.604	0.567
0.8	0.644	0.622	0.575
0.85	0.693	0.639	0.582
0.9	0.758	0.658	0.589
0.95	0.863	0.677	0.595

Table 3. Cutoff values for Example 5.3, for myopia-shortly, $m = 1, 2$. (When the cutoff value is less than 0.55, it marks the cutoff between using $\{x'\}$ and $\{x\}$. These are marked with asterisks. Values greater than 0.55 are the cutoff values between using $\{x, x''\}$ and $\{x\}$; for all these parameters, the strategy changes from $\{x'\}$ to $\{x, x''\}$ at $\pi_1^t = 0.55$.)

fixed prior π^0, the optimal strategy for δ close enough to 1 has $K_0 = \{x, x''\}$,[42] But for $m = 1$ and for any prior $\pi_1^0 > 0.6$, and for $m = 2$ and any prior $\pi_1^0 > 36/52$, the decision maker is told by the heuristic to choose $K_0 = \{x\}$: The decision maker will not spend any resources on gathering information if there is no chance that her posterior will land her in a region where her final choice (that is, the once-and-for-all choice of a myopically opimal toolkit) is different from what she would choose right now. And, in this example, she cannot reach a posterior

[42] We know this from the corollary to Proposition 2.

$\pi_1^t \leq 0.5$ —where she would switch from $\{x\}$ to $\{x'\}$ —in one step if her prior is above 0.6 or in two steps if her prior is above $36/52$.

Hence, when information arrives "slowly," but δ is close to one, myopia-shortly for small m is in danger of doing poorly. Compare with our other prior-based heuristics, all of which are bound to do fairly well when δ is close to one.

To be more systematic in comparing the performance of these different heuristics, we should be comparing their performance quantitatively, on test problems. So we turn at last to simulations.

6. *Simulations*

To get an idea of how well these heuristics perform relative to one another, we resort to simulations. All of the examples we will simulate are of a variety that might be called "small support of v models." We let $\mathcal{V}$ be the union of the supports of the v_t vectors under the various hypotheses; this sort of model is characterized by $\mathcal{V}$ being a relatively small set. The various μ_i, then, are distinguished by their different probability distributions over $\mathcal{V}$. We present the basic data of such examples in tables such as Table 4a, using Example 5.1 as our example.

Recall that in Example 5.1, there are three tools, x, x', and y, and two hypotheses, μ_1 and μ_2. The set $\mathcal{V}$ has two elements, $(14, 10, 12)$ and $(10, 14, 12)$. Under μ_1, the probability that $v_t = (14, 10, 12)$ is 0.9; under μ_2, this probability is 0.1. The costs of the three tools are, respectively, $5, 5,$ and 3.9. The prior probabilities of the two hypotheses are $\pi_1^0 = 0.5$ and $\pi_2^0 = 0.5$ Compare these numbers with Table 4a, and the format we employ for "small $\mathcal{V}$ models" should be apparent.

	x	*x'*	*y*	*μ1*	*μ2*
v1	14	10	12	0.9	0.1
v2	10	14	12	0.1	0.9
prior π				0.5	0.5
cost	5	5	3.9		

Table 4a. The Data for Example 5.1.

The form of the function U and the discount factor δ are not provided in Figure 4a. In all our examples, U will be U^{MAX}. For discount rates, we show results for $\delta = 0.7, 0.8,$ and 0.9, to see how the relative performance of the heuristics changes as the discount rate changes.

In all our simulations of these small $\mathcal{V}$ models, we simulate out to $t = 64$. Note that $0.9^{65} = 0.00106$; with the largest discount factor we investigate, stopping at $t = 64$ means that we "miss" around one-tenth of one percent of the total weight given to outcomes. (For $\delta = 0.8$, we are missing 5×10^{-7}, which is truly insignificant.) We simulate for a variable number of

trials—this number will be reported with the data—where, on each trial, we simulate a selection of which hypothesis is μ^T and then see how the decision maker fares under the various heuristics as far as date $t = 64$.

We simulate nine different heuristics: (1) Pay-for-itself (PfI); (2) Incremental-Contribution (IC); (3) Simple Set Based (SSB); (4) Different-of-Means, Set-Based (DoM); (5) Adaptive Myopia (AdM) (6) Harmonic Sampling (HAR); (7) Thompson Sampling (THO); (8) Upper-Confidence-Bound (UCB); and (9) Myopia-Shortly (MyS). Note that some of these heuristics require parametric specification: In the first three, one must specify the dates at which evaluations take place. For DoM, the threshold critical probability must be specified; recall that for DoM, we always take $T_n = 1 + n$.[43] In UCB, the threshold probability ϵ must be specified. In MyS, the horizon m must be specified; we found that implementing Myopia-Shortly for anything larger than $m = 1$ very difficult (and time consuming) to do. Hence, in all cases, MyS refers to the Myopia-Shortly heuristic with $m = 1$.

With regard to the first four heuristics, we always take $T_n = Ln$ for some single parameter L.[44] (We expect that these heuristics will "want" smaller L for smaller δ, something we investigate later.)

Hence, in reporting results, we follow the sort of data supplied in Table 4a with a second set of "heuristics' parameters" as in Table 4b. This gives L, the threshold mean difference in DoM, and the threshold probability ϵ for UCB. We also give two benchmarks for comparison with out simulation results: The *Static Myopic Value* is the per-period payoff that can be achieved by choosing in each period a toolkit that is optimal for the initial prior π^0.[45] And the *Clairvoyance Value* is the average payoff that would be received (per period) if the decision maker, prior to the selection of K_0, was told which hypothesis μ_i is μ^T, where the averaging is with respect to the decision maker's initial prior.

K (PfI, IC, SSB)	4
threshold critical probability (DoM)	0.05
threshold probability in UCB	0.05
Static Myopic Value	8.1
Clairvoyance Value	8.6

Table 4b. Heuristics' Parameters and Benchmarks, Example 5.1

And, following all the model data and the heuristics parameters, we present the simulation results.

[43] We reiterate the point made in fn.26: For very small T, the "logic" of using a classical difference-of-mean test is very strained in our test problems, because the Normal variates assumption on with such tests are built is a very poor approximation to the actual situation.

[44] Since we begin with $T = 0$, this means we gather L periods of data and do evaluations, gather L more and reevaluate, and so forth.

[45] In comparison, Adaptive Myopia chooses toolkits adapting to any information that arrives.

The most important data are the average performances, measured by normalized discounted sums of payoffs, of each of the nineheuristics. By normalized discounted sums of payoffs, we mean $(1-\delta)$ times the simple discounted sum of payoffs, averaged over the trials.[46] To get a sense of how much variation there are in these summary statistics, we also provide the sample standard deviations in these normalized discounted sums of payoffs. Note that we have these results for each of the nine heuristics, and for each of three discound rates. Hence the data appear as in Table 4c, which gives the results for our simulations of Example 4.1, where we simulated for 1000 trials. The average values of the normalized discounted sum of payoffs are provided in each cell of the left-hand matrix, with the sample standard deviations on the right.[47] Because this problem is simple enough so that we can compute the fully optimal strategy, we provide the simulated results of applying the optimal strategy in a penultimate row. (As we move to more complex examples, this row will be missing.)

	Average performance			Standard deviations		
δ	0.7	0.8	0.9	0.7	0.8	0.9
PfI	2.085	3.503	5.584	0.243	0.344	0.374
IC	2.095	3.504	5.547	0.113	0.162	0.210
SSB	2.095	3.504	5.547	0.113	0.162	0.210
DoM	3.659	4.849	6.305	1.250	1.422	1.386
AdM	8.100	8.100	8.091	0	0	0.000
HAR	4.161	5.137	6.379	1.191	1.034	0.723
THO	7.893	8.087	8.307	0.810	0.628	0.402
UCB	7.661	7.906	8.201	0.996	0.752	0.458
MyS	8.1	8.1	8.344	0	0	0.384
Optimal Value	8.1	8.168	8.37			

Table 4c. Average performance of the heuristics, for $\pi^0 = (0.5, 0.5)$, Example 5.1, 1000 trials.

While the data in Table 4c provide the "answer" to the question, *How well do the various heuristics do, measured by the decision maker's objective function, on this specific problem?*,

[46] (1) In case it is not obvious: Normalizing in this fashion means that the normalized discounted sum of payoffs is set on the scale of the (unnormalized) payoffs in each period. If, for instance, the decision maker chooses toolkits that give a constant payoff of 9 (say) in each period, then the *normalized*, discounted sum of her payoffs will be 9.

(2) To reiterate, for $\delta = 0.9$, because we stop at $t = 64$, we are missing around 0.1% of the total value. If you look at the simulation results for Example 5.1 as presented in Table 4c, you see average performance results for AdM of 8.1 for $\delta = 0.7$ and 0.8, and 8.091 for $\delta = 0.9$. In fact, AdM for this problem chooses toolkit $\{y\}$ in all time periods for this problem—this toolkit yields no information, so the prior never changes—and toolkit $\{y\}$ gives a net payoff of 8.1 in each period. Hence, the results for $\delta = 0.9$ would also be 8.1 if we went out enough time periods (beyond $t = 64$), so that the "missing weight" was smaller. In reporting later results, we "round up" results reported for $\delta = 0.9$ to get an even number, for any heuristic that chooses in a manner that gives the same, certain payoff in each period.

[47] Note that some of the sample standard deviations are zero. For MyS at $\delta = 0.5$ and 0.7, in particular, this happens because the heuristic calls for the selection of $\{y\}$, which provides no information, in all periods (as does the optimal strategy for these discount rates).

there are other data we can and sometimes do report, to help gain insights into the data provided in Table 4c. In particular, Table 4c indicates that the four prior-free heuristics do significantly worse than the four prior-based heuristics, and more so the smaller is δ. This happens because the prior-free heuristics take a while to "get started": PfI, IC, and SSB all carry $K_t = X$ for $t = 0, 1, 2,$ and 3, and, for small δ, the first three periods can represent a large share of the whole payoff. DoM, with its first revision at $t = 2$, does better in this regard. (And note that, as implemented, HAR begins with $K_0 = X$ with probability one.)

Table 4d gives us data relevant to this point. It reports the sample means and standard deviations for "snapshots" of how the heuristics performed at times $t = 1, 4, 8, 16, 32,$ and 64. Note that all the heuristics except for MyS make choices of K_t in ways that have nothing to do with δ, so we can report snapshot results for these heuristics in a single line; the choice of K_t for MyS does depend on δ, so we have three lines in Table 4d for MyS, one for each value of δ.

	Average payoffs in period						Standard deviations					
	t=1	*t=4*	*t=8*	*t=16*	*t=32*	*t=64*	*t=1*	*t=4*	*t=8*	*t=16*	*t=32*	*t=64*
PfI	0.100	8.336	8.545	8.546	8.584	8.548	0	1.580	1.308	1.273	1.222	1.267
IC	0.100	8.382	8.458	8.382	8.434	8.370	0	1.057	0.950	1.057	0.985	1.073
SSB	0.100	8.382	8.458	8.382	8.434	8.370	0	1.057	0.950	1.057	0.985	1.073
DoM	0.100	7.685	7.929	8.069	8.401	8.458	0	2.292	1.840	1.703	1.436	1.365
AdM	8.100	8.100	8.100	8.100	8.100	8.100	0	0	0	0	0	0
HAR	4.158	7.179	7.665	7.981	8.276	8.455	4.213	3.270	2.863	2.366	1.965	1.523
THO	7.964	8.496	8.628	8.564	8.592	8.556	1.753	1.328	1.162	1.247	1.211	1.257
UCB	6.984	8.492	8.644	8.560	8.592	8.556	2.001	1.333	1.140	1.252	1.211	1.257
MyS δ=0.7	8.100	8.100	8.100	8.100	8.100	8.100	0	0	0	0	0	0
MyS δ=0.8	8.100	8.100	8.100	8.100	8.100	8.100	0	0	0	0	0	0
MyS δ=0.9	8.248	8.492	8.644	8.564	8.592	8.556	1.564	1.333	1.140	1.247	1.211	1.257

Table 4d. Snapshot performance of the heuristics, for $\pi_1^0 = 0.5$, Example 5.1.

It is worth noting that the column for $t = 64$ can be used to answer the question, Does the heuristic get to the "right" toolkit by $t = 64$? More generally, the full table tells us (roughly) how long it takes for a heuristic to get to the right toolkit (if it does). Recall that the clairvoyance value in this example is 8.6; if the mean performance at $t = 64$ is close to this value, then we know that (in most of the iterations of our simulation, at least) the heuristic got there. So, for instance, by $t = 64$, all of the heuristics except for AdM and MyS for $\delta = 0.7$ and 0.8 seem to be "getting to the right answer," most of the time. There are a couple of fine points to make about this:

1. "Getting it right" is *not* the same thing as being optimal. While MyS for $\delta = 0.7$ and 0.8 and AdM are furthest from "getting it right" by $t = 64$, in fact they are making precisely the choices that the optimal strategy would make for these discount rates, namely to always choose the uninformative toolkit $\{y\}$. That is, the optimal strategy for this problem and those values of δ would not "get it right," either.

2. Note that THO, UCB, and MyS for $\delta = 0.9$ have average payoffs in period 8 that *exceed* the clairvoyance value of 8.6. Bear in mind that these are simulation results; even if, in the 1000 trials, in 500 trials μ^T was μ_1 (and in the other 500 it was μ_2, in those cases where $\mu^T = \mu_1$ (so that, presumably, the heuristics are leading the decision maker to choose $\{x\}$), more than 90% of the 500 may have drawn $v_x(8) = 14$, giving a net payoff 9 from $\{x\}$ at $t = 8$.

Tables 5 and 6 present similar data for Examples 5.2 and 5.3, respectively. We continue to see results similar to those for Example 5.1: THO, UCB, and MyS vie for the top spot. The prior-free heuristics generally do worse than the prior-based heuristics, but this is largely due to their poor performance early on. We know from the propositions that HAR, THO, and UCB will all "get it right" in the end,[48] while there is no guarantee of this for MyS; in fact, simple calculations show that, regardless of δ, in these examples there is positive probability that, when $\mu^T = \mu_2$, MyS will at some point recommend the uninformative $\{x\}$, which traps the decision maker.[49] There are no guarantees that the four prior-free heuristics will get it right in the end and, in fact, since they offer no way back once a tool is dropped, there must be positive probability that the decision maker winds up with the wrong tool kit. But the simulation results suggest that this doesn't happen very often, for these problems.

The performance of Adaptive Myopia (AdM) for these two examples should be regarded with caution. At the outset, in each example, the toolkits $\{x\}$ and $\{x'\}$ tie for the myopic optimum. The way the simulation program was written chooses $\{x\}$, which is (of course) uninformative. Had $\{x'\}$ been chosen instead, the performance of AdM would look significantly better. One might think of a hybrid form of AdM that avoids this issue. But it is worth observing that if, say, $c_x = 4.9999$ instead of 5, AdM performs as indicated, while if $c_x = 5.00001$, AdM does much better. The performance of AdM, and also UCB (and, to a lesser extent, MyS), is *not* continuous in the parameters, even if discontinuities in information flow are not at issue.[50]

To readers of this version of the paper: *We are posting an incomplete version of this paper because co-author Francetich is on the job market, and this work represents both an important part of his Ph.D. thesis from 2013 and a significant portion of the work he has done during his first post-doctoral year. We are currently engaged in creating, simulating, and analyzing more complex test problems. The discussion to follow will give you an indication of what we have found so far, as well as where we believe this research is going.*

Tables 7 and 8 present the results of simulation of two more-complex problems.

---

[48] For UCB, this isn't certain for a given threshold probability.

[49] In Example 5.3, recommends $K_0 = \{x'\}$ and continued use of $\{x'\}$, as long as the decision maker has at least as many 20's as 10's. But if she ever sees more 10's than 20's—which happens with probability one if $\mu^T = \mu_1$ and with strictly positive probability (strictly less than one) if $\mu^T = \mu_2$—MyS tells her to choose $\{x\}$.

[50] Of course, discontinuities in information flows, caused by small changes in the v-vectors, can have discontinuous impact on any of the prior-based heuristics.

Basic data:

	Tools		Hypotheses		
	x	*x′*	*y*	*m1*	*m2*
v1	8	1	0.9	0.9	0.1
v2	8	15	1.1	0.1	0.9
prior π				0.5	0.5
cost	5	5	1.2		

Heuristics' Parameter and Benchmarks:

K (Pfl, IC, SSB)	4
reshold critical probability (DoM)	0.05
threshold probability in UCB	0.05
Static Myopic Value	3
Clairvoyance Value	5.8

Performance levels---1000 iterations

	Average performance			Standard deviation		
δ	0.7	0.8	0.9	0.7	0.8	0.9
Pfl	1.512	2.415	3.726	2.954	2.931	2.928
IC	1.512	2.415	3.726	2.954	2.931	2.928
SSB	1.518	2.424	3.735	2.942	2.897	2.851
DoM	2.525	3.311	4.298	3.100	3.082	3.005
AdM	3	3	3	0	0	0
HAR	2.668	3.265	4.058	3.082	2.985	2.860
THO	4.059	4.379	4.817	3.441	3.269	3.070
UCB	4.098	4.525	5.033	4.632	4.115	3.517
MyS	4.539	4.805	5.076	4.077	3.675	3.245
Optimal Value	4.644	4.925	5.31			

Snapshots of average payoffs (and standard deviations)---1000 iterations:

	Average payoffs in period						Standard deviations					
	t=1	*t=4*	*t=8*	*t=16*	*t=32*	*t=64*	*t=1*	*t=4*	*t=8*	*t=16*	*t=32*	*t=64*
Pfl	0.293	5.408	5.638	5.670	5.750	5.520	3.502	4.361	4.168	4.109	4.026	4.164
IC	0.293	5.408	5.638	5.670	5.750	5.520	3.502	4.361	4.168	4.109	4.026	4.164
SSB	0.293	5.464	5.674	5.688	5.720	5.514	3.502	4.084	3.950	3.940	3.913	4.044
DoM	0.293	4.572	5.413	5.745	5.789	5.559	3.502	4.565	4.214	4.037	3.983	4.125
AdM	3	3	3	3	3	3	0	0	0	0	0	0
HAR	2.757	4.163	4.907	5.230	5.479	5.403	4.681	4.291	4.121	4.106	3.992	4.080
THO	3.861	4.778	5.289	5.492	5.716	5.541	4.872	4.514	4.352	4.244	4.008	4.139
UCB	2.986	5.415	5.723	5.758	5.828	5.604	7.003	4.328	4.021	3.979	3.917	4.069
MyS δ=0.7	5.296	5.163	5.408	5.485	5.513	5.331	4.297	4.105	3.885	3.829	3.811	3.925
MyS δ=0.8	5.296	5.163	5.408	5.485	5.513	5.331	4.297	4.105	3.885	3.829	3.811	3.925
MyS δ=0.9	5.296	5.163	5.408	5.485	5.513	5.331	4.297	4.105	3.885	3.829	3.811	3.925

Table 5. Simulation Results for Example 5.2.

Basic data:

	Tools		Hypotheses		
	x	*x′*	*x″*	*m1*	*m2*
v1	15	10	0.9	0.6	0.4
v2	15	20	1.1	0.4	0.6
prior π				0.5	0.5
cost	5	5	0.1		

Heuristics' Parameters and Benchmarks:

K (Pfl, IC, SSB)	4
hold critical probability (DoM)	0.05
threshold probability in UCB	0.05
Static Myopic Value	10
Clairvoyance Value	10.5

Performance levels---1000 iterations

	Average performance			Standard deviation		
δ	0.7	0.8	0.9	0.7	0.8	0.9
Pfl	7.776	8.004	8.360	1.340	1.342	1.416
IC	8.131	8.586	9.224	1.191	1.088	0.998
SSB	8.114	8.571	9.222	1.153	0.989	0.803
DoM	8.195	8.457	8.823	1.388	1.355	1.320
AdM	10	10	10	0	0	0
HAR	8.790	9.089	9.463	1.241	1.042	0.786
THO	10.116	10.112	10.101	1.518	1.252	0.948
UCB	10.130	10.121	10.105	2.261	1.886	1.437
MyS	10.15	10.16	10.16	1.942	1.504	1.02
Optimal Value	10.067	10.09	10.152			

Snapshots of average payoffs (and standard deviations)---1000 iterations:

	Average payoffs in period						Standard deviations					
	t=1	*t=4*	*t=8*	*t=16*	*t=32*	*t=64*	*t=1*	*t=4*	*t=8*	*t=16*	*t=32*	*t=64*
Pfl	7.485	8.895	8.775	8.855	9.330	9.345	2.500	3.757	3.833	3.870	4.089	4.039
IC	7.485	10.530	9.990	10.110	10.370	10.230	2.500	4.115	4.149	4.148	4.133	4.143
SSB	7.485	10.215	10.075	10.095	10.335	10.195	2.500	2.947	2.954	2.954	2.936	2.949
DoM	7.485	9.031	8.999	9.303	9.864	10.060	2.500	3.391	3.459	3.530	3.602	3.614
AdM	10	10	10	10	10	10	0	0	0	0	0	0
HAR	8.898	9.675	9.819	9.904	10.246	10.193	3.285	2.807	2.631	2.615	2.422	2.345
THO	9.995	10.370	10.165	10.080	10.405	10.530	3.666	3.539	3.444	3.586	3.482	3.540
UCB	10.170	10.450	10.115	10.050	10.500	10.615	5.000	4.982	4.974	4.870	4.233	3.682
MyS δ=0.7	10.140	10.265	10.095	10.020	10.315	10.155	3.563	3.157	2.815	2.551	2.254	2.158
MyS δ=0.8	10.140	10.265	10.095	10.020	10.315	10.155	3.563	3.157	2.815	2.551	2.254	2.158
MyS δ=0.9	10.140	10.265	10.095	10.020	10.315	10.155	3.563	3.157	2.815	2.551	2.254	2.158

Table 6. Simulation Results for Example 5.3.

Basic data:

	Tools					Hypotheses		
	1	*2*	*3*	*4*	*5*	*m1*	*m2*	*m3*
v1	25	0	14	6	9	0.4	0.25	0.1
v2	0	25	6	14	9	0.4	0.25	0.1
v3	0	0	14	6	9	0.1	0.25	0.1
v4	0	0	6	14	9	0.1	0.25	0.7
prior π						0.85	0.05	0.1
cost	6	6	4	4	2			

Performance levels---1000 iterations

	Average performance			Standard deviation		
δ	0.7	0.8	0.9	0.7	0.8	0.9
Pfl	1.557	2.679	4.292	2.812	2.612	2.414
IC	1.712	2.942	4.695	2.568	2.153	1.682
SSB	1.753	2.998	4.750	2.598	2.207	1.764
DoM	0.982	1.446	2.283	2.651	2.408	2.161
AdM	7.189	7.195	7.216	2.481	1.982	1.434
HAR	3.707	4.581	5.739	2.873	2.494	1.924
THO	6.647	6.823	7.153	4.958	4.060	2.869
UCB	6.658	6.860	7.178	2.609	2.352	1.966
MyS	7.315	7.259	7.500	2.883	2.739	2.061

Heuristics' Parameters and Benchmarks:

K (Pfl, IC, SSB)	4
critical probability (DoM)	0.05
threshold probability in UCB	0.05
Static Myopic Value	7.127
Clairvoyance Value	7.99

Snapshots of average payoffs (and standard deviations)---1000 iterations:

	Average payoffs in period						Standard deviations					
	t=1	*t=4*	*t=8*	*t=16*	*t=32*	*t=64*	*t=1*	*t=4*	*t=8*	*t=16*	*t=32*	*t=64*
Pfl	0.063	6.624	6.374	6.718	6.021	6.775	4.869	9.847	10.127	10.182	10.351	10.030
IC	0.063	7.123	7.176	7.472	6.748	6.940	4.869	7.302	7.494	7.110	7.118	6.776
SSB	0.063	7.357	7.132	7.559	6.800	6.978	4.869	7.787	7.824	7.665	7.597	7.283
DoM	0.063	1.828	2.708	3.826	4.218	5.132	4.869	7.620	7.890	8.186	8.325	8.154
AdM	7.344	7.262	7.261	7.315	7.340	7.413	3.925	5.288	5.356	5.269	5.164	5.06
HAR	4.100	6.325	6.775	7.359	7.317	8.288	6.640	8.041	8.529	8.353	8.731	8.167
THO	6.625	7.145	6.915	7.783	7.611	8.332	10.132	9.820	10.060	9.421	9.567	8.872
UCB	6.479	7.262	7.158	7.617	7.531	8.316	7.982	8.005	8.876	9.295	9.651	8.913
MyS δ=0.7	7.574	7.592	7.378	7.512	7.451	7.822	6.462	6.613	6.995	6.771	6.839	6.43
MyS δ=0.8	7.229	7.718	7.386	7.863	7.607	8.016	6.335	7.471	8.243	7.772	7.998	7.561
MyS δ=0.9	7.229	7.880	7.536	7.915	7.484	8.155	6.335	7.792	8.719	8.381	8.753	8.106

Table 7. Simulation Results for Example 6.1, a Three-hypothesis Example.

Basic data:

	Tools					Hypotheses			
	1	*2*	*3*	*4*	*5*	*m1*	*m2*	*m3*	*m4*
v1	25	0	14	6	9	0.4	0.25	0.15	0.2
v2	0	35	6	14	9	0.4	0.25	0.15	0.2
v3	0	0	14	6	9	0.1	0.25	0.1	0.5
v4	0	0	6	14	9	0.1	0.25	0.6	0.1
prior						0.4	0.2	0.2	0.2
cost	6	6	4	4	2				

Performance levels---1000 iterations

	Average performance			Standard deviation		
δ	0.7	0.8	0.9	0.7	0.8	0.9
Pfl	2.376	3.275	4.653	4.714	4.382	4.167
IC	2.823	4.017	5.785	4.524	3.970	3.350
SSB	2.796	3.973	5.708	4.487	3.915	3.300
DoM	1.962	2.397	3.339	4.158	3.672	3.253
AdM	8.316	8.316	8.369	5.321	4.437	3.464
HAR	4.802	5.647	6.782	4.630	4.012	3.285
THO	6.818	6.982	7.390	5.452	4.637	3.771
UCB	5.646	6.021	6.819	7.725	6.502	5.007
MyS	8.438	8.511	8.708	4.897	4.088	3.332

Heuristics' Parameters:

K (Pfl, IC, SSB)	4
critical probability (DoM)	0.05
threshold probability in UCB	0.05
Static Myopic Value	8.46
Clairvoyance Value	9.7

Snapshots of average payoffs (and standard deviations)---1000 iterations:

	Average payoffs in period						Standard deviations					
	t=1	*t=4*	*t=8*	*t=16*	*t=32*	*t=64*	*t=1*	*t=4*	*t=8*	*t=16*	*t=32*	*t=64*
Pfl	0.763	6.086	6.391	6.382	7.187	6.942	8.7154	12.539	11.823	11.643	11.184	11.020
IC	0.763	8.134	8.251	8.059	8.417	8.494	8.7154	10.418	9.711	9.697	9.029	8.732
SSB	0.763	7.953	8.242	7.871	8.204	8.364	8.7154	9.979	9.376	9.298	8.751	8.598
DoM	0.763	2.612	3.340	4.553	6.441	7.027	8.7154	9.610	9.417	9.585	9.595	9.652
AdM	7.968	8.299	8.381	8.256	8.633	8.711	11.522	9.6442	8.7515	8.3699	8.0433	8.0348
HAR	4.631	7.068	7.608	8.096	8.736	9.084	10.107	10.102	9.981	9.821	9.437	9.274
THO	6.026	7.183	7.438	7.622	8.636	9.257	11.250	11.502	10.927	10.601	10.352	10.089
UCB	4.505	6.353	7.042	7.785	8.867	9.259	15.305	12.976	12.062	10.635	9.296	9.769
MyS δ=0.7	7.990	8.750	8.592	8.358	9.134	9.224	9.815	10.081	9.9197	9.6415	9.3776	9.1672
MyS δ=0.8	7.990	8.679	8.634	8.476	9.071	9.251	9.815	9.8318	9.7092	9.7503	9.6718	9.4987
MyS δ=0.9	7.684	8.789	8.597	8.640	9.319	9.370	10.968	10.501	10.007	9.998	9.688	9.549

Table 8. Simulation Results for Example 6.2, a Four-hypothesis Example.

1. The test problem depicted in Table 7, labelled Example 6.1, has three hypotheses (the different μ's); the problem in Table 8, labelled Example 6.2, has four. In our simulation program, *as currently written*, four hypotheses is the most we can accomodate: For purposes of running MyS, before we begin to iterate in the simulation, we compute best myopic-choice values for a grid of posteriors with fineness 0.01. That is, we find for every posterior of the form $(0.1n_1, 0.1n_2, \ldots)$ the value of the (myopically) optimal kit, where n_i ranges between 0 and 100 and i indexes the various μ's. Since the probabilities must sum to 1, this means that with I hypotheses, we need around 100^{I-1} optimizations conducted. For $I = 3$, this is 10,000 optimizations; for $I = 4$, one million. We haven't dared to try $I = 5$ on our local desktops. (None of the other heuristics present much of a problem for large I.)

   We use these values in MyS as follows: Given a "postion," which is a current posterior, we compute for each kit and each of the possible v vectors what the next posterior will be. (Since different kits have different informational content, we must do this for each kit separately.) And for the continuation value in the MyS calculation, we round each of these "post-posteriors" down to the nearest 0.01 in all but the last component (which, of course, is rounded up).

   This, of course, is for the purpose of finding, given the current posterior, the "best" kit under MyS. Note that in these more complex examples, MyS seems to be emerging as the best of our heuristics, and increasingly so the more complex is the test problem. We need to do a lot more simulations, but our working conjecture is that this will be true, unless we construct a test problem with the objective of making one of the other heuristics look good. The point here is that *insofar as MyS looks better, perhaps a cruder form of MyS—say, where we compute continuation values on a grid of fineness 0.1 and linearly interpolate to get approximations to the "true" MyS continuation value—will continue to produce recommendations that, when implemented, do nearly as well as MyS for a finer grid*. This is clearly a conjecture worth pursuing.

2. In Table 7, AdM does slightly better than do THO and UCB; in Table 8, AdM does signicantly better. (A back-of-the-envelope comparison of the mean performance levels, assuming an equal sample standard deviations of 3.3, gives a Student's t of 6.63 in comparing the performance of AdM and THO, THO being the closer to AdM of the two. Of course, a paired sample test of the difference in means would be better, and we will conduct this. But with a t of 6.63, it is pretty clear that AdM is beating the socks off of both THO and UCB. And, recall, AdM is the heuristic based on a philosophy of, Just ignore the exploitation/exploration dilemma and go for exploitation.

   Now, it could be that although AdM is not "consciously" trying to gain information, in this test problem, the information is coming in, nonetheless. So we redo the analysis of this problem but with a prior assessment of $(0.15, 0.25, 0.25, 0.35)$. For this prior, the initial best tool-kit is $\{x_5\}$, which provides no information. Hence, for this initial assessment, AdM is providing no information at all. The deck, in other words, is stacked against AdM doing

well.

The results are shown in Table 9.

Basic data:

	Tools					Hypotheses			
	1	*2*	*3*	*4*	*5*	*m1*	*m2*	*m3*	*m4*
v1	25	0	14	6	9	0.4	0.25	0.15	0.2
v2	0	35	6	14	9	0.4	0.25	0.15	0.2
v3	0	0	14	6	9	0.1	0.25	0.1	0.6
v4	0	0	6	14	9	0.1	0.25	0.6	0.1
prior						0.15	0.25	0.25	0.35
cost	6	6	4	4	2				

Heuristics' Parameters:

K (Pfl, IC, SSB)	4
critical probability (DoM)	0.05
threshold probability in UCB	0.05
Static Myopic Value	7
Clairvoyance Value	8.34

Performance levels---1000 iterations

	Average performance			Standard deviation		
δ	0.7	0.8	0.9	0.7	0.8	0.9
Pfl	0.645	1.670	3.203	4.082	3.493	3.077
IC	1.170	2.552	4.557	4.108	3.436	2.655
SSB	1.193	2.588	4.598	4.056	3.348	2.551
DoM	0.558	1.232	2.453	3.755	3.286	2.937
AdM	7.000	7.000	6.993	0	0	0
HAR	3.233	4.245	5.564	4.194	3.500	2.651
THO	5.914	6.045	6.372	3.865	3.153	2.369
UCB	3.195	4.029	5.282	6.267	4.827	3.342
MyS	7.248	7.387	7.538	4.371	3.515	2.605

Snapshots of average payoffs (and standard deviations)---1000 iterations:

	Average payoffs in period						Standard deviations					
	t=1	*t=4*	*t=8*	*t=16*	*t=32*	*t=64*	*t=1*	*t=4*	*t=8*	*t=16*	*t=32*	*t=64*
Pfl	-1.234	4.784	5.188	5.245	5.608	5.945	8.372	11.434	10.773	10.306	9.584	9.371
IC	-1.234	7.164	7.260	7.212	7.369	7.547	8.372	9.037	8.525	8.008	6.998	6.787
SSB	-1.234	7.242	7.399	7.296	7.375	7.575	8.372	8.694	8.170	7.829	6.812	6.700
DoM	-1.234	1.772	3.452	4.398	5.711	6.885	8.372	8.885	8.928	8.713	8.181	8.081
AdM	7.000	7.000	7.000	7.000	7.000	7.000	0	0	0	0	0	0
HAR	2.623	5.888	7.234	7.381	7.450	7.783	9.676	8.515	8.224	7.763	6.834	6.917
THO	5.518	5.833	6.697	7.114	6.888	8.173	8.636	8.666	8.487	8.826	8.469	8.343
UCB	0.870	5.119	6.648	6.875	7.405	8.284	15.010	10.470	8.940	7.699	6.971	7.759
MyS δ=0.7	6.432	7.476	7.793	7.677	7.522	7.859	8.800	7.573	7.309	6.822	6.172	6.258
MyS δ=0.8	6.432	7.489	7.916	7.771	7.481	7.754	8.800	7.600	7.454	7.068	6.274	6.323
MyS δ=0.9	6.432	7.513	7.891	7.740	7.581	7.988	8.800	8.080	7.672	7.239	6.582	6.515

Table 9. Simulation Results for Example 6.2 with a Different Prior

Once again, AdM is beating THO and UCB. And this time, it isn't because AdM is getting lucky in the information it accumulates. So, we hypothesize, it isn't that AdM is good, but that, for these parameters, THO and UCB are bad. They are bad in the sense that they are taking too long to get the information, given the discount rates. The snapshots confirm our earlier theoretical results; THO and UCB are getting the information eventually: They are outperforming the other heuristics (including MyS and AdM) in the later periods.[51] So, perhaps, the way to improve their performance is to "speed them up." In particular, it might make sense to set the threshhold probability in UCB somewhat higher. We reran the test problem of Table 9, but with a threshold probability for UCB of 0.1, and the average performance of UCB improved from 5.28 (in Table 9) to 5.9. (We will continue to explore this issue.)

3. Going back to Table 7, the snapshots indicate that the prior-free heuristics IC and SSB are doing somewhat worse at dates 32 and 64 than they are doing at dates 16 and, perhaps,

[51] The theory suggests that HAR will do likewise, eventually. But "eventually" for HAR, for this example, is later than it is for THO and UCB.

8. A potential weakness of these heuristics is that they give "no way back" once a tool is discarded. So what we may be seeing in this simulation (to be investigated) may be an instance of this weakness, where, at least some of the time, toolkits are shrinking too much.

The reader can no doubt see other hypotheses worthy of investigation via these simulations. So, at this point, all we can do is say, Stay tuned.

References

Baumol, William, and Richard Quandt (1964), "Rules of Thumb and Optimally Imperfect Decisions," *American Economic Review*, Vol. 54, 23-46.

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

Blume, Larry, and David Easley (1982), "Learning to be Rational," *Journal of Economic Theory*, Vol. 26, 340-51.

Bray, Margaret (1982), "Learning, Estimation, and the Stability of Rational Expectations," *Journal of Economic Theory*," Vol. 26, 318-39.

Easley, David, and A. Rustichini (2005), "Optimal Guessing: Choice in Complex Environments," *Journal of Economic Theory*, Vol. 124, 1-21.

Francetich, Alejandro (2013), *Contributions to Microeconomic Theory,* Stanford CA: Ph.D. thesis, Graduate School of Business, Stanford University.

Francetich, Alejandro (2014), "Managing Multiple Research Projects," manuscript, Department of Decision Sciences and IGIER, Bocconi University.

Francetich, Alejandro, and David M. Kreps (2014), "Bayesian Inference Does Not Lead You Astray . . . On Average," mimeo, forthcoming in *Economics Letters*.

Fudenberg, Drew, and David M. Kreps (1993), "Learning Mixed Equilibria," *Games and Economic Behavior*, Vol. 5, 320-67.

Fudenberg, Drew, and David Levine (1998), *The Theory of Learning in Games*, Cambridge, MA: MIT Press. Neveu, J. (1975), *Discrete Parameter Martingales*, Amsterdam: North-Holland.

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

Kalai, Ehud, and Ehud Lehrer (1993), "Rational Learning Leads to Nash Equilibrium," *Econometrica*, Vol. 61, 1019-45.

Kreps, David (2013), *Microeconomic Foundations: I. Choice and Competitive Markets*, Prince-

ton: Princeton University Press.

Lettau, M., and Harold Uhlig (1999), "Rules of Thumb Versus Dynamic Programming," *American Economic Review*, Vol. 89, 148-71.

Lewis, Michael (2003), *Moneyball: The Art of Winning and Unfair Game*, New York: W. W. Norton & Co.

Milgrom, Paul, and D. John Roberts (1991), "Adaptive and Sophisticated Learning in Normal Form Games," *Games and Economic Behavior*, Vol. 3, 82-100.

Radner, Roy (1975), "Satisficing," *Journal of Mathematical Economics*, Vol. 2, 253-62.

Roth, Alvin E., and Ido Erev (1998), "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria," *American Economic Review*, Vol. 88, 848-81.

Russo, Daniel, and Benjamin Van Roy (2014), "Learning to Optimize via Information-Directed Sampling," Cornell University Library: arXiv:1403.5556.

Rustichini, A. (1999), "Optimal Properties of Stimulus-Response Learning Models," *Games and Economic Behavior*, Vol. 29, 244-73.

Sargent, Tom, and Albert Marcet (1989), "Convergence of Least Squares Learning Mechanisms in Self-Referential Linear Stochastic Models," *Journal of Economic Theory*, Vol. 48, 337-68.

Simon, Herbert (1959), "Theories of Decision-Making in Economics and Behavioral Science," *American Economic Review*. Vol. 49, 253-83.

Simon, Herbert (1979), "Rational Decision Making in Business Organizations," *American Economic Review*, Vol. 69, 493-513.

Simon, Herbert (1982a), *Models of Bounded Rationality, Vol. 1: Economic Analysis and Public Policy*, Cambridge, MA: MIT Press.

Simon, Herbert (1982b), *Models of Bounded Rationality, Vol. 2: Behavioral Economics and Business Organization*, Cambridge, MA: MIT Press.

Simon, Herbert (1997), *Models of Bounded Rationality, Vol. 3: Empirically Grounded Economic Reason*, Cambridge, MA: MIT Press.

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; Palm, Günther (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, pp.335–346, ISBN978-3-642-24455-1.

Appendix

In the proofs of some of the propositions, we employ a particular instantiation of the decision maker's probability model. The state space is, at a minimum, $\Omega := \{1, \ldots, I\} \times ((R_+)^X)^{\{0,1,\ldots\}}$, with typical element $\omega = (i, v_0, v_1, \ldots)$. Write Ω^i as the subset of Ω consisting of points whose first component is i, and assign probability to Ω so that the probability of Ω^i is π_i and, conditional on being in Ω^i, the probability distribution on the sequence $\{v_t\}$ renders these vectors i.i.d. with distribution given by μ_i. Of course, the event $\mu^T = \mu_i$ is, in this state space, just Ω^i. When dealing with harmonic sampling and Thompson sampling, in which the decision maker chooses K_t randomly, the "date t" component v_t is supplemented by a uniform-$[0, 1]$ random variate, each such independent of all other sources of randomness, which can be used to affect whatever randomized choices are needed.

We denote by $\mathbf{P}$ the probability of various events defined on Ω, according to the decision maker's subjective prior assessment, while $\mathbf{P}^i$ will denote probability conditional on the event $\Omega^i = \{\omega : \mu^T = \mu_i\}$. We denote expectation with respect to $\mathbf{P}$ by $\mathbf{E}$, and expectation with respect to $\mathbf{P}^i$ by $\mathbf{E}^i$. As long as $\mathbf{P}(\Omega^i)$ (which is π_i^0) is strictly positive, a maintained hypothesis, any statement that is true almost surely with respect to $\mathbf{P}$ is true almost surely with respect to $\mathbf{P}^i$. Conversely, any statement that is a.s. true with respect to each $\mathbf{P}^i$ is true $\mathbf{P}$-a.s.

In what follows, we examine a number of random variables and random processes defined on this state space. For instance, we often deal with the decision maker's posterior assessment π^t, with ith component π_i^t. Note that the value of such random variables depends on ω, of course, but also on the decision rule or heuristic the decision maker employs for choosing her toolkits as a function of things she has observed.

The σ-field generated by all information available to the decision maker after v_t is realized (and she observes as much of this as is provided given her choice of K_t) will be denoted by F_{t+1}. Hence F_0 is the trivial σ-field, and π^t is F_t-measurable. But v_t, or rather those parts of v_t that she observes, is only F_{t+1} measurable. Of course, the filtration $\{F_t\}$ is affected by her choice of heuristic. Note that where she chooses K_t randomly, the choice of K_t is incorporated in F_{t+1} but, typically, not in F_t.

But, regardless of her heuristic or decision rule, relative to her subjective probability assessment, her sequence of posteriors $\{\pi^t\}$ forms a vector martingale relative to the filtration $\{F_t\}$. (This well-known result is a consequence of the law of iterated expectations.) Since it is bounded below by 0 and above by 1, it converges to some π^∞ $\mathbf{P}$-a.s.: Letting F_∞ be the total of all information she possesses, π^∞ closes the martingale of posterior assessments.

There are two more general results that we will need.

Lemma 1. *For each $i = 1, \ldots, I$, the two (one-dimensional) stochastic processes $\{\pi_i^t; t = 0, \ldots, \infty\}$ and $\{ln(\pi_i^t); t = 0, \ldots, \infty\}$ are submartingales (for the filtration $\{F_t\}$) under the probability measure $\mathbf{P}^i$.* (See Francetich and Kreps, 2014.)

Lemma 2. *If* $\{\zeta_t; t = 0, 1, \ldots\}$ *is a martingale with uniformly bounded increments, then*

$$\lim_{t\to\infty} \frac{1}{t}\zeta_t \;=\; 0$$

almost surely.[52]

The proof of Proposition 2

Let G_∞ be the sigma-field generated by the data received at dates $1, 2, 2^2, \ldots$ only. (These are the dates, recall, when the decision maker chooses $K_t = X$.) Of course, G_∞ is a sub-sigma-field of F_∞. Hence

$$\mathbf{P}\left[\{\mu^T = \mu_i\}\middle|G_\infty\right] = \mathbf{E}\left[\mathbf{P}[\{\mu^T = \mu_i\}|F_\infty]\middle|G_\infty\right] = E\left[\pi_i^\infty\middle|G_\infty\right].$$

But the nature of the information in G_∞, combined with our assumption that the μ_i distributions are all distinct, ensures that, almost surely,

$$\mathbf{P}\left[\{\mu^T = \mu_i\}\middle|G_\infty\right] = \begin{cases} 1, & \text{on } \Omega^i, \text{ and} \\ 0, & \text{on the complement of } \Omega^i. \end{cases}$$

Since the values of π_i^∞ must lie between zero and one, it must be that (a.s.)

$$\pi_i^\infty = \begin{cases} 1, & \text{on } \Omega^i, \text{ and} \\ 0, & \text{on the complement of } \Omega^i. \end{cases}$$

Hence,

$$\lim_{t\to\infty} \pi_i^t = \begin{cases} 1, & \text{on } \Omega^i, \text{ and} \\ 0, & \text{on the complement of } \Omega^i \end{cases}, \quad \mathbf{P}-\text{a.s.}$$

This ensures that along any sample path for which this convergence happens, if $K \notin \mathcal{K}_i^*$, then a time must come when (outside of dates of the form 2^n) K is no longer chosen; that is, along almost every sample path belonging to Ω^i, for all large t not of the form 2^n, $K_t \in \mathcal{K}_i^*$.

Write

$$\frac{1}{T+1}\sum_{t=0}^{T} W(v_t, K_t) = \frac{1}{T+1}\sum_{t=0}^{T}\left[W(v_t, K_t) - w(\pi^t, K_t)\right] + \frac{1}{T+1}\sum_{t=0}^{T} w(\pi^t, K_t).$$

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

To complete the proof, we must show that the limit (in T) of the left-hand side is w^*, almost surely. We will do this by showing the limit of the first summation on the right-hand side is zero, almost surely, while the limit of the second summation is w^*, also almost surely.

For the first summation, note that $\mathbf{E}[W(v_t, K_t)|F_t] = w(\pi^t, K_t)$. Therefore, if we define $\zeta_T = \sum_{t=0}^{T} \left[W(v_t, K_t) - w(\pi^t, K_t)\right]$, $\{\zeta_t, F_{t+1}\}$ forms a bounded-increments martingale under $\mathbf{P}$. Apply Lemma 2.

As for the second summation, look (only) along sample paths for which the posteriors converge. We know that along each such sample path that is in Ω^i, $K_t \in \mathcal{K}_i^*$ eventually (in t), for all $t \neq 2^n$. But then $w(\pi^t, K)$ converges (except for $t = 2^n$) to w_i^*. That is, for almost every sample path, except for times $t = 2^n$, $w(\pi^t, K_t)$ converges to w^*. Taking Cesàro sums wipes out the effect of the terms in the sum for times $t = 2^n$ (the terms are uniformly bounded), finishing the proof. ∎

Concerning Proposition 3

There are two parts to the proof of the proposition. First, we show that, for each i, there exists $\epsilon > 0$ such that, if the decision maker's posterior assessment π^t puts weight $1 - \epsilon$ or more on $\mu^T = \mu_i$, then the decision maker will optimally choose the single bundle $K_i^* \in \mathcal{K}_i^*$ at date t; Condition A ensures that $\mathcal{K}_i^*$ has but one bundle. Then, we show that if any bundle K^0 is selected infinitely often, Condition B ensures that the decision maker's posterior will converge to the "truth"; that is,

$$\lim_{t\to\infty} \pi_i^t = \begin{cases} 1, & \text{on } \Omega^i, \text{ and} \\ 0, & \text{on the complement of } \Omega^i. \end{cases}$$

These two together give the desired result.

For the first part, let B be an upper bound on the absolute value of the function w. Then, in terms of future values, the best the decision maker can do relative to the worst that can happen to her is bounded by $2B/(1-\delta)$. Suppose she reaches a point where her posterior assigns probability $1 - \epsilon$ to μ_i being the truth, and (per Condition A) suppose that the (uniquely) best toolkit K_i^* for μ_i is $\gamma > 0$ better under μ_i than is the second best toolkit under μ_i. Then, by choosing any toolkit other than K_i^*, she gives up an immediate expected return of at least $(1-\epsilon)\gamma - 2\epsilon B$, for a future gain that is bounded above by $2\delta B\epsilon/(1-\delta)$. (Her expected continuation value if she chooses any other toolkit is bounded above by what she would get if she learns the true state with certainty, and her expected continuation value if she chooses K_i^* is bounded below by what she gets if she chooses K_t^* for the rest of time. The bound given is their difference, discounted by one period.) Hence, the net gain from choosing a bundle other than K_t^* is bounded above by

$$\frac{2\delta B\epsilon}{1-\delta} + 2\epsilon B - (1-\epsilon)\gamma = \epsilon\left[\frac{2B + \gamma(1-\delta)}{1-\delta}\right] - \gamma.$$

Hence, if

$$\epsilon < \frac{\gamma(1-\delta)}{2B + \gamma(1-\delta)} ,$$

she is worse off for choosing something other than K_i^*.

And for the second part, we repeat the argument given for Proposition 2, except now, if K^0 is chosen infinitely often, we let τ_n be the nth time that K^0 is chosen and we let G_∞ be the σ-field generated by $\{W(v_{\tau_n}, K^0); n = 1, 2, \ldots\}$. On the subspace Ω_i, the successive averages

$$\frac{1}{n}\sum_{m=1}^{n} W(v_{\tau_m}, K^0)$$

converge almost surely to $w(\mu_i, K^0)$, and these are (by Condition B) different in different Ω^i, so the limiting average identifies a.s. (conditional on G_∞) which μ_i is the true μ^T (that is, to which Ω^i the state ω belongs). Hence, on the event where K^0 is chosen infinitely often, the decision maker's posterior converges to a point mass on the truth and, by the first paragraph, it must be that K^0 is the optimal bundle for whichever is the true μ^i and, moreover, that K^0 must eventually be the only toolkit that is chosen. ∎

Proof of Proposition 4

For $\mathbf{P}$-almost-every sample path of the $\{v_t\}$ process, for each $K \subseteq X$,

$$\lim_{T\to\infty} \frac{1}{T+1}\sum_{t=0}^{T} W(v_t, K) \;=\; w(\mu^T, K).$$

Note that this is about a fixed K and not K_t, so this is a direct consequence of DeFinetti's Theorem. Discard from Ω any sample paths for which this is not true for any K.

Hence, for *every* sample path for which some toolkit $\hat{K}$ is the final toolkit, for every $K \subseteq \hat{K}$, the empirically observed averages of $W(v_t, K)$ converge to $w(\mu^T, K)$. Now suppose that for some sample path along which $\hat{K}$ is the final toolkit and for some $\check{K} \subseteq \hat{K}$, $w(\mu^T, \check{K}) > w(\mu^T, \hat{K})$. Then along this sample path, for all T sufficiently large

$$\frac{1}{T}\sum_{t=0}^{T-1} W(v_t, \check{K}) > \frac{1}{T}\sum_{t=0}^{T-1} W(v_t, \hat{K}). \qquad (A.1)$$

In particular, this will be true for all T_n, for sufficiently large n. But it must also be true that, for each $x \in \hat{K}$,

$$I(x, \hat{K}) = \frac{1}{T_n} \sum_{t=0}^{T_n - 1} \left[W\big(v_t, \hat{K}\big) - W\big(v_t, \hat{K} \setminus \{x\}\big) \right] \geq 0. \qquad (A.2)$$

Let $x_1, x_2, \ldots, x_m$ be an enumeration of $\hat{K} \setminus \check{K}$, and temporarily let $K(i) = \hat{K} \setminus \{x_1, \ldots, x_i\}$, for $i = 1, \ldots, m$, where we use $K(0)$ for $\hat{K}$. Property (4.3) implies that for $i = 1, \ldots, m-1$ and for all v,

$$W\big(v, K(0)\big) - W\big(v, K(0) \setminus \{x_{i+1}\}\big) \leq W\big(v, K(i)\big) - W\big(v, K(i) \setminus \{x_{i+1}\}\big),$$

so that (A.2) implies that, for all sufficiently large n and for $i = 1, \ldots, m-1$,

$$\frac{1}{T_n} \sum_{t=0}^{T_n - 1} \left[W\big(v_t, K(i-1)\big) - W\big(v_t, K(i)\big) \right] \geq 0. \qquad (A.3i)$$

Sum up the inequalities $(A.3i)$ for $i = 1, \ldots, m$, and you get

$$\frac{1}{T_n} \sum_{t=0}^{T_n - 1} \left[W\big(v_t, K(0)\big) - W\big(v_t, K(m)\big) \right] \geq 0,$$

which can be rewritten

$$\frac{1}{T_n} \sum_{t=0}^{T_n - 1} W\big(v_t, \hat{K}\big) \geq \frac{1}{T_n} \sum_{t=0}^{T_n - 1} W\big(v_t, \check{K}\big),$$

which contradicts (A.1) ∎

Concerning Proposition 5.

The proof of Proposition 5 for any heuristic that chooses $K_t = X$ for infinitely many t (even if the times t for which this is so are randomly determined) requires only very slight modifications from the proof of Proposition 2 given earlier, so we leave this to the reader.

Proof of Proposition 6.

Recall that F_t is the σ-field generated by all information received up to and including the observation of (any observed parts of) v_{t-1}, so that π^t is F_t-measurable. In Thompson

sampling, K_t is chosen randomly (prior to the realization of v_t) based on π^t; and we need to extend the notation accordingly: Let G_t be F_t augmented by the choice of K_t, so that F_0 is the trivial σ-field, G_0 refines F_0, F_1 refines G_0, and so forth.

Discard from Ω the null-set of sample paths for which the decision maker's posteriors π^t do not converge. Recall that π^∞ denotes the limit of these posteriors and, as a random variable, closes the martingale of posterior assessments.

We would like to conclude that $\pi_i^\infty = 1_{\Omega^i}$, but (of course) the examples given in the body of the paper shows that this is not true in general; Proposition 6 asserts that this is true if Condition C holds, but not (necessarily) otherwise.

Suppose that, for some i, $\pi_i^\infty \neq 1_{\Omega^i}$. Since $0 \leq \pi_i^\infty \leq 1$, this implies that $\int_{\Omega^i} \pi_i^\infty(\omega)\mathbf{P}(d\omega) \leq \mathbf{P}(\Omega^i) = \pi_i^0$, and since $\mathbf{E}[\pi_i^\infty] = \pi_i^0$, this implies that for some $j \neq i$ (where $j \in \{1, \ldots, I\}$), $\mathbf{P}\{\mu^T = \mu_j \text{ and } \pi_i^\infty > 0\} > 0$. (There may be many such j.) Fix some j so that this is so. Since, on the event $\{\mu^T = \mu_j \text{ and } \pi_i^\infty > 0\}$, K_i^* is chosen with strictly positive probability bounded away from zero for all dates sufficiently large (large enough so that $\pi_i^t > \pi_i^\infty/2$, say), K_i^* will be chosen infinitely often for a.e. $\omega \in \mathbf{P}\{\mu^T = \mu_j \text{ and } \pi_i^\infty > 0\}$. By the sort of argument given in the proof of Proposition 2, then, the decision maker asymptotically learns the full distribution of generated by K_i^* on this event. Hence, on this event, the distribution of $W(v, K_i^*)$ must be the same under μ_j as under μ_i, for otherwise (on this event, which is a subevent of Ω^j), she would asymptotically come to realize that μ^T is *not* μ_i, and π_i^t would asymptotically approach zero. This would contradict Condition C; if Condition C holds, $\pi_i^\infty = 1_{\Omega^i}$.

And what if Assumption C does not hold? Suppose that, for some ℓ, $\pi_\ell^\infty \neq 1_{\Omega^\ell}$ (so Assumption C cannot hold). Then we know that there must be at least one $j \neq \ell$ such that $\mathbf{P}\{\mu^T = \mu_j \text{ and } \pi_\ell^\infty > 0\} > 0$.

Define a binary relation $\prec$ on $\{1, \ldots, I\}$:

$$i \prec j \text{ if } \mathbf{P}\{\mu^T = \mu_j \text{ and } \pi_i^\infty > 0\} > 0.$$

Note that we allow $j = i$ in this definition. And, in fact, it must be true that $i \prec i$: Lemma 1 given at the start of the Appendix states that, with respect to $\mathbf{P}^i$, $\{\pi_i^t; t = 0, 1, \ldots\}$ is a closed submartingale. Therefore, $\mathbf{E}^i[\pi_i^\infty] \geq \pi_i^0 > 0$, and so π_i^∞ must be strictly positive with positive probability on Ω^i.

Let $\bar{\prec}$ be the transitive closure of $\prec$. For each i, let $I(i) = \{j : i \bar{\prec} j\}$ and let $\Lambda^i = \cup_{j \in I(i)} \Omega^j$. (Note that $i \in I(i)$.) We assert that for each i,

$$\sum_{j \in I(i)} \mathbf{E}[\pi_j^\infty \cdot 1_{\Lambda^i}] = \sum_{j \in I(i)} \pi_j^0. \tag{A.4}$$

It is of course true that $\mathbf{E}[\pi_j^\infty] = \pi_j^0$, since π^∞ closes the martingale of posterior assessments. The point is that for all $j \in I(i)$, $\pi_j^\infty = 0$ on the complement of Λ^i, so omitting the complement of Λ^i in the integrals on the left-hand side loses nothing.

Interchange the summation and the integral on the left-hand side of (A.4):

$$\mathbf{E}\left[1_{\Lambda^i} \sum_{j \in I(i)} \pi_j^\infty\right] = \sum_{j \in I(i)} \pi_j^0.$$

Since $\mathbf{E}[1_{\Lambda^i}] = \sum_{j \in I(i)} \pi_j^0$, this implies that, for every i,

$$\sum_{j \in I(i)} \pi_j^\infty = 1 \text{ a.e. on } \Lambda^i. \tag{A.5}$$

Now go back to any ℓ for which $\pi_\ell^\infty \neq 1_{\Omega^\ell}$, and take any $i \neq \ell$ such that $\ell \prec i$. Apply (A.5) for this specific i. Since $\pi_\ell^\infty > 0$ with positive probability on $\Omega^i \subseteq \Lambda^i$, we conclude that $\ell \in I(i)$. Hence, for every $i \neq \ell$ such that $\ell \prec i$, there is a chain $\ell = \ell^0 \prec i = \ell^1 \prec \ell^2 \prec \ldots \prec \ell^m = \ell$.

Consider any pair i and j, $i \neq j$, such that $i \prec j$. We know from the last part of the argument where we assumed that Condition C holds that the distribution of $W(v_t, K_i^*)$ under μ_i must be the same as under μ_j. This implies that $w_i^* = w_j(K_i^*)$ and, of course, $w_j(K_i^*) \leq w_j^*$. Applying this to the cycle created last paragraph, we have

$$w_\ell^* = w_j(K_\ell^*) \leq w_j^* \leq w_{\ell^1}^* \leq \ldots \leq w_i^*,$$

and so we conclude all the weak inequalities must equalities. To summarize this part of the argument:

If $i \prec j$ or, equivalently, if $\mathbf{P}\{\mu^T = \mu_j$ and $\pi_i^\infty > 0\} > 0$, then $w_i^* = w_j(K_i^*) = w_j^*$.

For the remainder of the proof, we suppose that μ^T is, in fact, μ_i: That is, we show what happens on Ω^i, for arbitrary i.

Define random variables $Y_j(t) := 1_{\{K_t = K_j^*\}} W(v_t, K_j^*)$ and $Y(t) := \sum_{j=1}^I Y_j(t)$. 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 $Y_j(t)$, assuming that they exist, so we'll look at those.

The limit of the Cesàro sums of the $Y_j(t)$, or $\lim_{T\to\infty}\left[\sum_{t=0}^T Y_j(t)/(T+1)\right]$, is

$$\lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \left[Y_j(t) - \pi_j^t w_i(K_j^*)\right] \quad + \quad \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \pi_j^t w_i(K_j^*), \tag{A.6}$$

assuming both limits exist.

And, under $\mathbf{P}^i$, they do (almost surely): To begin, compute

$$\begin{aligned}
&\mathbf{E}^i\left[Y_j(t) - \pi_j^t w_i(K_j^*)\middle| F_t\right] = \\
&\mathbf{E}^i\left[1_{\{K_t=K_j^*\}} W(v_t, K_j^*) - \pi_j^t w_i(K_j^*)\middle| F_t\right] = \\
&\mathbf{E}^i\left[\mathbf{E}^i\left[1_{\{K_t=K_j^*\}} W(v_t, K_j^*)|G_t\right]\middle| F_t\right] - \pi_j^t w_i(K_j^*),
\end{aligned} \tag{A.7i}$$

because π^t is F_t-measurable and $w_i(K_j^*)$ is a (deterministic) scalar. Moreover, the event $\{K_t = K_j^*\}$ is G_t-measurable, so the string of equalities in (A.7i) continues

$$= \mathbf{E}^i\left[1_{\{K_t=K_j^*\}} \mathbf{E}^i\left[W(v_t, K_j^*)\middle| G_t\right]\middle| F_t\right] - \pi_j^t w_i(K_j^*). \tag{A.7ii}$$

Under Thompson sampling and on the event Ω^i, the value of v_t is independent of all information in G_t and is distributed according to μ_i, so $\mathbf{E}^i[W(v_t, K_j^*)|G_t] = w_i(K_j^*)$. And $\mathbf{E}^i[1_{K_t=K_j^*}|F_t] = \pi_j^t$. Hence, the expression in (A.7ii) is 0. But this implies that if we let $\zeta_T = \sum_{t=0}^T [Y_j(t) - \pi_j^t w_i(K_j^*)]$, $\{\zeta_T\}$ is a martingale with bounded increments with respect to $\mathbf{P}^i$. Hence, Lemma 2 at the start of the Appendix ensures that the limit of the Cesàro sums of $\{\zeta_T\}$ is almost surely 0 (under $\mathbf{P}^i$).

Hence, we are left with

$$\lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \pi_j^t w_i(K_j^*).$$

We know that, for every sample path, the sequence π_j^t converges to π_j^∞. So along each sample path, this Cesàro limit is just $\pi_j^\infty w_i(K_j^*)$.

If $\pi_j^\infty = 0$, this is zero. If $\pi_j^\infty > 0$, then we know from our earlier argument that (for almost every sample path) $w_i(K_j^*) = w_i^*$. And so, when we recompose the sum of these Cesàro sums of the $Y_j(t)$ to find the limit of the Cesàro sums of the $Y(t)$, we get

$$\lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} Y(t) = \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \sum_{j=1}^{I} Y_j(t) = \sum_{j=1}^{I} \pi_j^\infty w_i(K_j^*) = w_i^*.$$

■

Proof of Proposition 7.

(Having been very careful with the details in the proof of Proposition 6, we are a bit more discursive in this proof.)

Fix the ϵ hurdle rate and some i. (Assume that $\epsilon < 1/I$, so the decision maker never finds herself in a situation in which all hypotheses have been deemed to be implausible.) We will discuss what happens on Ω^i or, equvalently, under $\mathbf{P}^i$.

We know that the posteriors π^t converge to some π^∞, $\mathbf{P}^i$-a.s.. Divide sample paths in Ω^i into two events, those for which $\pi_i^\infty > \epsilon$, and those for which $\pi_i^\infty \leq \epsilon$.

On the first of these events, we assert that the limit of the Cesàro sums of payoffs converges to $w^* = w_i^*$, $\mathbf{P}^i$-a.s. (That is, the probability of this event under $\mathbf{P}^i$ is the same as the probability of this event intersected with the event where the limit of the Cesàro sums of payoffs converges to w_i^*.) Once we have shown this, we need only show that the probability of the complementary event is bounded above by some estimate that goes to zero as ϵ goes to zero. Since the latter step is easy, we do it first: Lemma 1 tells us that $\{\ln(\pi_i^t); t = 0, 1, \ldots, \infty\}$ is a submartingale under $\mathbf{P}^i$. Hence,

$$\mathbf{E}^i\big[\ln(\pi_i^\infty)\big] \geq \ln(\pi_i^0).$$

The integrand is bounded above by zero, so the integral over any subset of Ω^i must satisfy the same inequality. Therefore,

$$\mathbf{E}^i\left[\ln(\pi_i^\infty)1_{\{\pi_i^\infty \leq \epsilon\}}\right] \geq \ln(\pi_i^0).$$

But an obvious upper bound on the left-hand side integral is $\mathbf{P}^i\{\pi_i^\infty \leq \epsilon\} \times \ln(\epsilon)$, and so we have

$$\mathbf{P}^i\big\{\pi_i^\infty \leq \epsilon\big\} \leq \frac{\ln(\pi_i^0)}{\ln(\epsilon)},$$

which has limit 0 as $\epsilon \downarrow 0$.

Now consider the event $\{\pi_i^\infty > \epsilon\}$. The only way in which the decision maker could fail to be choosing K_i^* eventually (for all t beyond a certain point) is if, for some $j \neq i$, she is choosing K_j^* infinitely often. Suppose she is choosing K_j^* infinitely often. For this to be true, it must be that

1. $\pi_j^\infty \geq \epsilon$, for otherwise, past some point in time, μ_j will forever after be deemed implausible and K_j^* will not be a candidate for K_t, and
2. $w_j^* \geq w_i^*$, for otherwise, once π_i^t is greater than ϵ and remains there forever after, K_j^* will not be selected as K_i^* offers a better (optimistic) prospect.

Now if K_j^* is selected infinitely often, the decision maker sees infinitely many draws of the random variable $W(v_t, K_j^*)$. By computing sample means, she learns $w_i(K_j^*)$, $\mathbf{P}^i$-almost surely. It can't be that $w_i(K_j^*) < w_i^* \leq w_j^*$, for the data would then tell her that μ^T is not μ_j, and π_j^t would have to approach zero. Since $w_i(K_j^*) \leq w_i^*$, the only possibility is that

$w_i(K_j^*) = w_i^*$ (and is equal to w_j^*, since otherwise the decision maker would recognize the $\mu^T \neq \mu_j$).

But this says that any K_j^* that is selected infinitely often produces, under $\mathbf{P}^i$, the same expected per period return as w_i^*. The argument used in the proof of Proposition 6 is then easy adapted to show that the Cesàro sums of payoffs must have limit w_i^*. ∎