Dynamic Choice over Menus

Alejandro Francetich [*†]

Department of Decision Sciences and IGIER

Bocconi University, Italy

January 29, 2014

Abstract

A decision maker faces a finite set X of alternatives and chooses subsets of X at every period $t = 0, 1, 2, \ldots$ Rewards are drawn from an unknown distribution, and the decision maker only observes the realized rewards of the alternatives in the chosen subset. The problem is a multi-armed bandit problem, where the arms are the subsets of X. These arms are not independent: Even if the rewards from individual elements are independent, those from intersecting subsets will be correlated. As we are unable to fully characterize optimal strategies in general dependent multi-armed bandit problems, I propose and analyze heuristics. Applications include hiring of teams of experts by professional-services firms.

Keywords: Multi-armed bandits, dependent arms, heuristics, prior-free
JEL Classification Numbers: C63, D81, D83

[*]Via Röntgen 1, 20136 Milan, Italy. Email address: alejandro.francetich@unibocconi.it

[†]This work is based on Chapter 3 of my doctoral dissertation, submitted to the Stanford Graduate School of Business. I am deeply indebted to David Kreps and Andrzej Skrzypacz for their support and guidance. This paper has also benefited from discussions with Pierpaolo Battigalli, Vivienne Groves, Neil Malhotra, Paul Milgrom, Davin Raiha, Ilya Segal, Joel Wiles, Robert Wilson, and seminar participants at Stanford and at Bocconi. Special thanks go to Lanier Benkard. This paper represents work in progress, and should not be quoted, reproduced, or otherwise disseminated, without permission. I gratefully acknowledge financial support from ERC advanced grant 324219. Any remaining errors and omissions are all mine.

1 Introduction

1.1 Overview

Dynamic decision problems encountered in everyday life can quickly overwhelm the ability of decision makers to formulate optimal choices. The standard technique of analysis, dynamic programming, is useful within a domain of problems. However, some important problems are too complex to be solved by this method, or for their solution to provide useful guidance in actual decision making.

Consider the class of multi-armed bandit problems. A decision maker must choose, at any date $t = 0, 1, 2, ...$, one and only one of $N \in \mathbb{N}$ actions from a set $A := \{a_1, ..., a_N\}$. Her choice of action $a_n \in A$ in period t results in a reward r_{nt}, drawn independently across time from an unknown probability distribution. The decision maker has a prior belief regarding these distributions, which she updates on the basis of interim outcomes. She discounts future rewards by $\alpha \in [0, 1)$ and chooses a sequence of actions that maximizes the expected discounted sum of rewards.

This class of problems features a basic trade-off between *exploitation* and *exploration*. Exploitation means choosing actions that give the highest expected reward under the current beliefs. If the decision maker is myopic (in the sense that she discounts future rewards entirely, or $\alpha = 0$), any optimal strategy would recommend her taking an "exploitation action" at any point. However, insofar as future rewards carry weight, the decision maker may wish to "explore." Exploration refers to choosing actions that might be inferior under the current beliefs, but that might provide information that will lead to better future choices. The forgone expected rewards from such myopically sub-optimal choices represent the cost of experimentation.

If the rewards are independent across actions as well as across time, learning about the reward distribution of action a_n has no impact on the assessment of the reward distribution of action $a_{n'}$ for $n' \neq n$. Then, it is well-known that a strategy is optimal if and only if it recommends choosing actions according to their *Gittins indices*.[1] But if the rewards are not independent, the Gittins index doesn't help; in fact, the general problem of dependent arms is an open problem.

[1]For a formal definition, see Gittins and Jones (1974). Weber (1992) provides the following intuitive characterization of the Gittins index. Imagine that there is a single action or "arm" and a charge to pulling it, as well as the option to quit (whereby the charge is avoided). The Gittins index is the value of a lump sum that makes the decision maker indifferent between giving up on the arm right away or after an optimal number of additional pulls. In the multi-armed bandit problem under independence, a strategy is optimal if and only if it recommends each period pulling an arm with highest current Gittins index, computed for each arm as if it were the only arm. Independence is key in reducing the problem with multiple arms to multiple problems, each with a single arm.

In this paper, I consider the following dynamic problem with a non-independent structure. A decision maker faces a (non-empty) finite set X of elements or "tools." Decision period $t = 0, 1, 2, ...$ is divided into two stages. In the first stage, the decision maker chooses a *subset* a_t from $Z := 2^X$, the set of all subsets of X (including the empty set). This choice is fixed for the remainder of period t. To each $x \in X$ corresponds a known, fixed cost $c_x > 0$ as well as a random period-t value, v_{xt}. The profile of period-t values, $v_t := (v_{xt})_{x \in X}$, is drawn from an unknown distribution on a finite set $V \subseteq \mathbb{R}_+^{\#X}$ at the end of the first stage. In the second stage, the decision maker observes the realized values of the tools in a_t and employs them, earning a gross payoff of $W(a_t, v_t)$ and a net payoff of $W(a_t, v_t) - \sum_{x \in a_t} c_x$. As a leading example, take W as $W(a_t, v_t) = \max_{x \in a_t} v_{xt}$. Under this formulation, the tools are (ex-post) perfectly substitutable: the decision maker employs only one of the tools in a_t, the one that provides the highest current gross benefit; she must still pay for all the selected tools.

Initial beliefs are represented by a prior. With the additional information garnered at the end of period t, the decision maker reassesses her beliefs. While she only observes v_{xt} for $x \in a_t$, she may be able to make inferences about the distribution of values $v_{x't}$ for $x' \notin a_t$ if, say, x and x' are correlated.

This stylized problem approximates many real-life situations. For instance, consider the problem of a professional-services firm, such as a consultancy or a legal partnership, employing a pool of experts to serve their clients. Imagine that only one case is handled on each day, and that this case can only be dealt with by a single staff member. The manager, after an initial interview with the client, learns the specifics of the case and assigns it to the member of her staff who is best suited for the specific job. Of course, only experts who are on staff are available for this job; hiring new experts takes time. But keeping a large staff, one that can handle all conceivable cases, is costly: experts that remain idle during any period must nonetheless be paid a wage.

I assume that value profiles are independent and identically distributed across time, with marginal distribution denoted by $\mu^{\mathcal{T}} \in \Delta(V)$, the space of probability distributions on V.[2] If the decision maker knows $\mu^{\mathcal{T}}$, her problem is simple: Having nothing to learn about the environment, her optimal strategy is to choose, in every period, any set that maximizes the expected immediate reward. Such a set will be called an *objectively* optimal bundle. With $\mu^{\mathcal{T}}$ unknown, however, exploration can be profitable, and the (*subjectively*) optimal strategy can be different from the objectively-optimal choice.

We can think of the problem as a multi-armed bandit problem in which each bundle is an arm. Alternatively, if we identify each tool with an arm, we can think of the

[2]The superscript $\mathcal{T}$ stands for "truth".

problem as a multi-choice multi-armed bandit problem; the decision maker faces a finite set of arms and may choose several of them at a time.[3] However we conceive it, the problem is a dependent multi-armed bandit problem. First, I allow the values of the different tools to be correlated under the true distribution; for instance, X might contain two different types of knives (their values being positively correlated), or sunscreen and umbrellas (whose values are negatively correlated). Second, even if the different tools are independent, the overall reward from two sets will be correlated if they intersect.

These forms of dependence preclude a complete characterization of optimal strategies in terms of index policies. In fact, the problem of characterizing optimal strategies in general dependent multi-armed-bandit problems is an open problem. Faced with this dynamic problem, what is the decision maker to do? The obvious thing is to formulate and follow *heuristics*: simple rules of behavior featuring self-correcting techniques based on past experience to improve future performance. For instance, the decision maker can follow the rule "take one x at a time for as long as its reward has been above its cost" or "try all of X for some time and then choose the subset that would have awarded the highest average return for the collected data."

The objective of this paper is to investigate the "performance" of various "classes" of heuristics. These two terms are in scare-quotes, because once we move away from optimality according to a single specific criterion, there are many different ways to measure performance, and many different heuristics that can be considered. Performance can be measured in terms of long-run properties, by comparison against the true distribution, or by comparison against the decision maker's prior. As for type of heuristics, we can (and do) investigate heuristics that ignore any prior assessment held by the decision maker (such as "keep tools whose empirical average value is above its cost"), and those that employ prior information ("retain tools about which there is large prior uncertainty").

Since there are many choices of heuristic to investigate and many (imperfect and incomplete) performance criteria that can be employed, the results I obtain do not point in any one definitive direction. However, and at the risk of oversimplifying, I find that the exploration-exploitation trade-off takes on a richer meaning in this context. Sophisticated heuristics are more likely to perform well asymptotically but usually take longer to settle. Thus, simple heuristics that have "flaws" in theory can do well when discounting is taken into account.

[3]On multi-choice multi-armed bandits, see Bergemann and Valimaki (2001). They consider a problem with countably-infinitely many ex-ante identical arms, from which the decision maker can choose up to $k \in \mathbb{N}$ each period. Having an inexhaustible supply of "untried" copies of the process, they obtain an index-type of optimal strategy. The fact that there ar infinitely many arms and that the arms are ex-ante identical are crucial for the result.

Beyond the "as if" hypothesis

Clearly, some of the most important economic decisions are dynamic and involve learning about the environment, such as decisions about savings and investment, education, and employment. It is equally clear that real-life decision makers do not formulate their decisions in the manner of economic models of these decisions; they do not anticipate all possible consequences of their decisions, nor do they solve complex dynamic programming problems that knit together lifetime consumption, education, employment, and family decisions, when they purchase groceries for tonight's meal. To close this methodological gap, economists typically rely on the "as if" hypothesis: It is not necessary that decision makers solve fantastically complex dynamic decision problems when choosing tonight's groceries; only that they act as if they do.[4]

However, to quote Radner (1975), "It is probably not good positive theory to take very seriously an assumption that anyone behaves according to a sequential strategy that maximizes an expected lifetime (or infinite horizon) utility [...] that there is no prospect of solving in the next hundred years." If we (analysts) cannot characterize the solution to a problem, how can we tell whether individuals are acting "as if"?

What we can do, and what this paper sets out to do for a particular class of problems, is to formulate "reasonable" rules of thumb that might be employed and then investigate how well they perform, learning contexts in which they fare well and contexts in which they fare poorly. This is certainly less than establishing whether the "as if" justification holds; but for these complex problems, the "as if" justification is functionally meaningless. And insofar as these type of problems are encountered in real-life economic decisions, gaining even a limited understanding of the performance of particular heuristics is a step towards improving our understanding of decision making.[5]

1.2 Related literature

While the economics literature on heuristics cannot be considered vast, we can point to several influential streams. These can be grouped by the complexity of the heuristics they propose and by the social nature of the environment they explore; that is, whether there is "learning in isolation" or co-learning.

[4]See Friedman (1994). Rust (1987) does not argue that Harold Zurcher actually solves a regenerative optimal stopping problem to design his engine-maintenance policy; yet, he finds that his actual decisions do a remarkably good job at approximating the theoretical optimal solution.

[5]Comparisons with the vast literature on heuristics for playing chess, or with the tournament results of Axelrod (1984) on the repeated prisioner's dilemma are worth mentioning. Of course, conventional wisdom about "what works" can be flawed; see, for instance, Lewis (2003) on the impact of sabermetrics on strategy in baseball.

Baumol and Quandt (1964) define "rules of thumb" as a set of rules that describe a decision procedure and that are based on objectively measurable information, that feature objectively communicable and impartial decision criteria, and that involve only simple and inexpensive calculations. Looking at a monopoly-pricing problem, they argue that such rules can be better guides to decision-making than more refined strategies when it is costly to gather information and perform involved computations.

Other early references include the work of Simon (1959, 1979, 1982a,b, 1997) and Radner (1975). Simon challenges the notion that decision makers are able to perform the necessary computations required by rationality and develops the notion of *bounded rationality*. One behavioral model of bounded-rationality is *satisficing*. Radner (1975) presents a formal model of satisficing in which a decision maker allocates effort amongst different activities searching for improvements in performance.

Lettau and Uhlig (1999) look at a repeated dynamic choice problem in which the decision maker chooses an action a out of a finite set A and, based on the unobserved realization of a state of nature s from a finite set S, observes and receives a reward. Rules of thumb are defined here as stationary strategies and are evaluated based on a numerical index called "strength." The paper looks at asymptotic properties of choosing actions according to their strength. Rustichini (1999) considers a similar problem, both under the same informational assumption and under the assumption that the realized rewards from all possible actions —not just the one actually taken— are observed. The paper compares the performances of linear and exponential adjustment rules.

An alternative approach is followed in Easley and Rustichini (2005). Focusing on the case of full observability from Rustichini (1999), they posit that the decision maker has preferences defined on actions instead of on choices. The agent's learning from experience is captured by having these preferences evolve in a Markovian fashion. They present axioms on the transition process such that actions that are optimal under the expected-utility criterion are the highest ranked in the limit.

Heuristics also appear in the literature on learning in games and on rational expectations. Both of these environments involve co-learning: agents learn simultaneously from their own private information, if any, and from public signals, like actions taken, realized payoffs, or market prices. In the game theory literature, heuristics play a central role in the analysis of players' learning how to play a particular game. In the rational expectations literature, their role is in explaining how equilibria can be reached.

Roth and Erev (1998) analyze experimental data on games from the point of view of *reinforcement learning*. While they study learning in games, their learning model is akin to learning in isolation: players only register their own realized payoffs and play more

often strategies that have yielded higher payoffs. The same is true for the more recent Hart and Mas-Colell (2000, 2001a,b, 2003a,b, 2004) and Hart (2005), who introduce an adaptive heuristic called *regret matching* and study its connection to correlated equilibria. Regret-matching consists in playing strategies according to how much the decision maker "regrets" not having played the strategy in the past.

Fudenberg and Kreps (1993) and Fudenberg and Levine (1998) analyze the heuristic called *fictitious play* and identify conditions under which it leads players to play according to Nash equilibria. Under fictitious play, players play best responses to their opponents' mixing according to empirical frequencies of play.[6] Fictitious play is clearly more sophisticated than reinforcement learning or regret matching; it involves learning about opponents' strategies and best-responding to these assessments.

Milgrom and Roberts (1991) propose a notion of *adaptive learning* and of *sophisticated learning*. A sequence of play is consistent with adaptive learning if it eventually consists in nearly-undominated strategies, and with sophisticated learning if players eventually play near best responses to their forecast of their opponents' strategies. In games that have a unique Nash equilibrium, convergence of play to the equilibrium is equivalent to convergence of play that is consistent with either form of learning.

While in fictitious play assessments are updated based on historical frequencies of play, sophisticated learning allows for richer sources of information. At an even higher level of sophistication is the notion of *subjective rationality*, introduced by Kalai and Lehrer (1993). Subjectively rational players play best responses to conjectures about their opponents' strategies, taking into account the continuation game.

In the rational expectations literature, Bray (1982) looks at asset trading in an economy where agents are endowed with differential information and employ least-squares regression to estimate the relationship between equilibrium prices and the information held by others. For appropriate parameter values, the market outcome converges almost surely to the rational expectations equilibrium. In the same number of the journal, Blume and Easley (1982) study an economy where agents entertain different hypotheses about how information impacts on equilibrium prices and find that, within the context of their model, learning may lead to non-rational equilibria. Sargent and Marcet (1989) consider a class of linear stochastic models where the actual law of motion is a function of the perceived law of motion. They consider learning under the (false) assumption that the perceived law of motion is the actual one and identify conditions under which least-squares learning leads to a rational expectations equilibrium.

[6] Li Calzi (1992) extends the heuristic of fictitious play to analyze learning from past experience with "similar" games.

Finally, to model the choice of production routines by firms, Nelson and Winter (1982) study "evolutionary" heuristics. Firms are deemed to be "profit-seeking" rather than profit maximizing, operating in a dynamic setting subject to unforseen changes. Firms stick to what they have been doing unless and until their performance turns bad, at which point they copy production routines of more successful rivals.

1.3 Outline of the paper

The rest of the paper is organized as follows. Section 2 presents the model. Section 3 identifies some partial, asymptotic characterizations of the optimal strategy. Section 4 presents some specific heuristics. Section 5 introduces formal criteria to evaluate these heuristics, and results in this direction. Section 6 presents simulation results for the total-perfect-substitution specification. Section 7 concludes. All proofs are in the Appendix.

2 Problem Formulation

A decision maker faces the following two-stage repeated choice problem. Time is indexed by $t = 0, 1, 2, \ldots$ A (non-empty) finite set X of "tools" is given, with cardinality $\#X$ and generic element x. To each tool $x \in X$ corresponds a cost $c_x > 0$, that is fixed and known; $c := (c_x)_{x \in X}$ denotes the profile of costs. Associated with each $x \in X$ is also a random period-t value v_{xt} that lies in some finite set $V_x \subseteq \mathbb{R}_+$, where I assume that $0 \in V_x$ (even if v_{xt} takes the value 0 with probability 0); the profile of period-t values is $v_t := (v_{xt})_{x \in X} \in V := \times_{x \in X} V_x$. Values are drawn independently over time from the "true," unknown distribution $\mu^{\mathcal{T}} \in \Delta(V)$.[7]

The decision maker is endowed with fixed prior beliefs given by $\pi_0 \in \Delta^0(\Delta(V))$, the space of simple (finite-support) probability distributions over probability distributions on V. I assume that $\mu^{\mathcal{T}}$ is in the support of π_0. This prior is updated based on interim outcomes; the period-t posterior belief is $\pi_t \in \Delta^0(\Delta(V))$.

Each period $t = 0, 1, 2, \ldots$ is divided into two stages. In the first stage, the decision maker chooses a_t from $Z := 2^X$ and pays c_x for each $x \in a_t$. This choice is irreversible for the remainder of period t. At the beginning of the second stage, the decision maker observes all the v_{xt} corresponding to $x \in a_t$ and receives a reward $\rho(a_t, v_t)$. Payoffs are discounted by $\alpha \in (0, 1)$ per period.

[7] Please note: It is the process $\{v_t : t = 0, 1, 2, \ldots\}$ that is *i.i.d.*; we do not assume independence of v_{xt} and $v_{x't}$ for two distinct tools $x, x' \in X$.

The (ex-post) immediate reward function, $\rho : Z \times V \to \mathbb{R}$, takes the form

$$\rho(a_t, v_t) := W(\nu(a_t, v_t)) - \sum_{x \in a_t} c_x,$$

where $W : V \to \mathbb{R}$ is a non-decreasing and submodular function that satisfies $W(0) = 0$, and $\nu : Z \times V \to V$ takes a subset a and a vector v and replaces the coordinates of v corresponding to elements not in a with zeros; that is, $\nu(a, v)_x := v_x$ if $x \in a$ and $\nu(a, v)_x := 0$ otherwise. The composite function $W \circ \nu$ represents (ex-post) gross immediate rewards from carrying tool set a_t when the realized value profile is v_t.

This formulation captures the fact that only realized values of carried tools can influence rewards; values of tools that are not carried are "censored" to 0. Monotonicity of W means that carrying additional tools will weakly increase gross immediate rewards. Submodularity of W captures the following idea of (ex-post) substitution: if the value of some tools increases, the marginal contribution of the remaining tools to immediate rewards cannot increase.[8] The assumption that $W(0)$ is finite is made for analytical simplicity, but it shouldn't be regarded as an inconsequential normalization. It implies that the decision maker can "get by" without tools, and might well be better off with no tools at all in some circumstances.

The specification suggested in Section 1.1 obtains when W is the max function: $W(\nu(a, v)) = \max_{x \in a} v_x$, with the convention that $\max(\emptyset) := 0$. I shall refer to this case as the case of *total perfect substitution*: Once current values are realized, the decision maker uses only one tool, the one with highest current value.[9] A more general case is the case of *perfect substitution within classes*. Given a partition $\{X_1, ..., X_n\}$ of X for some $n \leq \#X$ and a non-decreasing and submodular function $\omega : \mathbb{R}^n_+ \to \mathbb{R}$ that satisfies $\omega(0) = 0$, gross rewards are $W(\nu(a, v)) := \omega(\max_{x \in a \cap X_1} v_x, ..., \max_{x \in a \cap X_n} v_x)$. Tools fall in different "categories" and are (ex-post) perfectly substitutable within categories. For example, X might consist of tax attorneys (X_1), litigation specialists (X_2) and merger specialists (X_3); clients' cases require up to one (the best suited one) from each group.

[8] This assumption is far from innocuous; it precludes complementarities *in the employment* of tools (complementarities in the form of correlation in values is allowed).

[9] Submodularity of W is not the only possible representation of substitution. In a static problem of allocating a finite number of indivisible goods, Gul and Stacchetti (1999) introduce the gross substitutes condition and the single improvement property. Under monotonicity, these two are equivalent and imply submodularity; however, the single improvement property excludes the total-perfect-substitution specification. Let $X = \{x_1, x_2, x_3\}$, with $v_{x_1} = 1$ with probability 1 and v_{x_1}, v_{x_2} perfectly negatively correlated, each taking values 0 or 5/2; costs are given by 1, 1/2, and 1/2, respectively. The bundle $\{x_1\}$ doesn't maximize immediate rewards; yet, to improve upon it, we have to discard x_1 and add both x_2 and x_3.

Let $r : Z \times \Delta(V) \to \mathbb{R}$ be the expected immediate reward function,[10]

$$r(a, \mu) := E^{\mu}[\rho(a, v)] = \sum_{v \in V} \rho(a, v)\mu(v).$$

To avoid complications due to ties, I will assume that $r\left(a', \mu^{\mathcal{T}}\right) \neq r\left(a, \mu^{\mathcal{T}}\right)$ for $a' \neq a$.[11] The next proposition establishes that $r(\cdot, \mu)$ is submodular for each $\mu \in \Delta(V)$. In other words, the increase in expected rewards from adding a tool to a smaller set is higher than that of adding this same tool to a larger set (in the sense of set inclusion). The proof, as well as the proofs of all other results in the paper, can be found in the Appendix.

Proposition 1. *The function $r(\cdot, \mu)$ is submodular for each $\mu \in \Delta(V)$; that is, for any $\mu \in \Delta(V)$, $a \in Z$, $b \subseteq a$, and $x \in X \setminus a$, we have that $r(b \cup \{x\}, \mu) - r(b, \mu) \geq r(a \cup \{x\}, \mu) - r(a, \mu)$.*

Given any $\pi \in \Delta^0(\Delta(V))$, we can associate a distribution $\mu_\pi \in \Delta(V)$ that represents the "average" distribution induced by π. Let $\mu_0 := \mu_{\pi_0}$ denote the "average" prior belief; μ_0 gives the decision maker's assessment for the distribution of v_0 (at the beginning of period 0). Similarly, $\mu_t := \mu_{\pi_t}$ represents the posterior assessment of v_t.

A strategy σ specifies a sequence of tool subsets (or of distributions over subsets, if we allow for randomized strategies) given any possible history of choices and realized tool values. The space of strategies is denoted by Σ. Some additional notation is needed to define Σ formally. Let $\Psi := \cup_{a \in Z} \nu(a, V)$; a time-$t$ history is an element of the set $\mathcal{H}_t := \Psi^t$, with $\mathcal{H}_0 := \emptyset$, recording the history of past choices and observed realized values up to t. The space of complete histories is $\mathcal{H} := \Psi^\infty$, the space of sequences in Ψ. A strategy σ is a mapping from $\mathcal{H}$ into Z^∞ (or $\Delta(Z)^\infty$, the space of sequences of probability distributions over Z) such that, for each $t = 0, 1, 2, \ldots$, the element σ_t of the sequence is measurable under the sigma-algebra generated by $\mathcal{H}_t$.

Given tool set X, cost profile c, value-profile set V, immediate-reward function ρ, prior π_0, and discount factor $\alpha \in (0, 1)$, the decision maker's problem is to choose a strategy $\sigma \in \Sigma$ to maximize the expected discounted sum of future rewards:[12]

$$E^{\sigma, \pi_0}\left[\sum_{t=0}^{\infty} \alpha^t \rho(a_t, v_t)\right].$$

If the decision maker knows $\mu^{\mathcal{T}}$ (namely, if π_0 is a degenerate prior at $\mu^{\mathcal{T}}$), her

[10] The notation P^μ, E^μ for $\mu \in \Delta(V)$ denotes probabilities and expectations with respect to μ.

[11] This condition is generic with regards to costs, for any distribution.

[12] The notation E^{σ, π_0} denotes expectation with respect to the distribution on $\mathcal{H}$ induced by σ and π_0.

problem is simple. The objectively optimal strategy is to always choose a bundle that maximizes the expected immediate reward, accruing a payoff of:

$$w^{\mathcal{T}} := \frac{\max_{a \in Z} r\left(a, \mu^{\mathcal{T}}\right)}{1-\alpha} = \frac{r\left(a^{\mathcal{T}}, \mu^{\mathcal{T}}\right)}{1-\alpha},$$

where $a^{\mathcal{T}}$ is the objectively-optimal tool bundle. However, for even moderately rich problems, we lack the means to characterize the optimal strategy when $\mu^{\mathcal{T}}$ is unknown.

To fix ideas, consider the following example.

Example 1. The tool set is $X = \{x_0, x_1\}$. Each period $t \in \mathbb{N}$, an "opportunity" to use a tool arrives with *known* probability $\lambda^* \in (0,1)$. This arrival process is represented by:

$$s := \begin{cases} 1 & \text{with probability } \lambda^*, \\ 0 & \text{with probability } 1-\lambda^*, \end{cases}$$

where the event $s = 1$ denotes the event of arrival. Each arrival corresponds to tool x_1 with *unknown* probability $p \in [0,1]$ and to x_0 with probability $1-p$. This parameter can take two values, $\underline{p}$ and $\overline{p}$, where $0 \leq \underline{p} < \frac{1}{2} < \overline{p} \leq 1$. The "allocation" process is represented by:

$$z := \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1-p, \end{cases}$$

where the event $z = 1$ ($z = 0$) denotes the allocation of the opportunity to tool x_1 (x_0), and where z and s are conditionally independent given p. Tool values are drawn from the random variables $v_{x_0} := s(1-z)$ for tool x_0 and $v_{x_1} := sz$ for tool x_1. Table 1 presents the joint distribution of v_{x_0} and v_{x_1}, given p.

While each time at most one tool can be used, and one of the tools is more frequently "useful" than the other, having both tools in the tool belt can help the decision maker

$v_0 v_1$	0	1	Marginal
0	$1-\lambda^*$	$\lambda^* p$	$1-\lambda^*+\lambda^* p$
1	$\lambda^*(1-p)$	0	$\lambda^*(1-p)$
Marginal	$1-\lambda^* p$	$\lambda^* p$	1

Table 1: Joint distribution of tool values, given p.

distinguish between failure of arrival for a given tool from failure of arrival altogether. On the other hand, tool values are negatively correlated given p, so experience on a single tool provides information about the other tool: A "good streak" on tool x_0 is bad news about x_1, and viceversa.

3 Optimality and Asymptotic Results

The problem as formulated is a discounted dynamic programming problem with bounded per-period rewards. The techniques of dynamic programming — producing optimal strategies by looking for conserving or unimprovable strategies, finding value functions by value iteration, and computing optimal strategies by policy iteration — are available to us, in theory.[13] Employing these techniques in practice involves formulating infinitely many beliefs and correctly anticipating the impact of all possible choices, both on immediate rewards and on future assessments and choices.

We can identify the prior π_0 with a finite set of distributions $\{\mu^k : k = 1, ..., K_0\}$ (its support) and a vector $(\pi_0(k))_{k=1,...,K_0}$ in the simplex of dimension $K_0 - 1$ (where $K_0 \in \mathbb{N}$). By assumption, $\mu^{\mathcal{T}} = \mu^k$ for some k; w.l.o.g., all distributions in the support are different: $\mu^k \neq \mu^{k'}$ for $k \neq k'$. At each date t, the decision maker observes the components of the realization of v_t that correspond to elements in a_t; alternatively, she observes a cell $p_t(v_t)$ of a partition $P_t := \{p_t(v) : v \in V\}$ of V, where $p_t(v_t)$ is the level set of $\nu(a_t, \cdot)$ at $\nu(a_t, v_t)$. The collection of all partitions is denoted by $\mathcal{P}$. If $(\pi_t(k))_{k=1,...,K_0}$ is the period-t prior and the decision maker observes $p_t \in P_t \in \mathcal{P}$, Bayes' rule gives:

$$\pi_{t+1}(k|p_t) = \frac{\pi_t(k)\mu^k(p_t)}{\sum_{\kappa=1}^{K_0} \pi_t(\kappa)\mu^\kappa(p_t)}.$$

The state space is $\mathcal{D} := \left\{\pi \in \Delta^0(\Delta(V)) : supp(\pi) \subseteq supp(\pi_0)\right\}$, where $supp(\pi)$ is the support of π, and can be represented as the unit simplex of dimension $K_0 - 1$. The Bayesian map is $\tau : \mathcal{D} \times Z \to \Delta^0(\mathcal{D})$; $\tau(\pi, a)$ denotes the distribution over posteriors of π based on the observations from the possible partitions corresponding to a.

Given a (stationary) strategy $\sigma \in \Sigma$, the *value function of* σ, $w^\sigma : \mathcal{D} \to \mathbb{R}$, is:

$$w^\sigma(\pi) := E^{\sigma,\pi}\left[\sum_{t=0}^{\infty} \alpha^t \rho(a_t, v_t)\right];$$

the *optimal-value function* of the problem, $w^* : \mathcal{D} \to \mathbb{R}$, is $w^*(\pi) := \sup_{\sigma\in\Sigma} w^\sigma(\pi)$. With

[13] For details, see Bertsekas (2005) or Appendix 6 in Kreps (2013).

$\mathcal{D}$ as the state space and τ as the transition map, the *Bellman equation* is:

$$w^*(\pi) = \max_{a \in Z} \left\{ r(a, \mu_\pi) + \alpha E^{\tau(\pi,a)} \left[w^*(\pi') \right] \right\}.$$

A (stationary) strategy $\sigma^* \in \Sigma$ is optimal if $w^{\sigma^*}(\pi) = w^*(\pi)$ for all $\pi \in \mathcal{D}$.

Proposition 2. *There exists an optimal (stationary) strategy.*

By Proposition 2, w^* can be attained. The difficulty lies in characterizing strategies $\sigma^* \in \Sigma$ that attain w^*.

Example 1 (continuation). We can identify $\mathcal{D}$ with the interval $[0,1]$, where $\pi \in [0,1]$ represents the belief that $p = \overline{p}$. Define $\mu(\pi) := \pi\overline{p} + (1-\pi)\underline{p}$. If the bundle selected at t is $\{x_0\}$, the posterior that $p = \overline{p}$ after observing $v_{x_0 t} = 0$ and $v_{x_0 t} = 1$ are, respectively,

$$\pi^{x_0,0}(\pi) := \frac{\pi(1 - \lambda^*(1-\overline{p}))}{1 - \lambda^*(1 - \mu(\pi))}, \quad \pi^{x_0,1}(\pi) := \frac{\pi(1-\overline{p})}{1-\mu(\pi)}.$$

The corresponding formulas for bundle $\{x_1\}$ are,

$$\pi^{x_1,0}(\pi) := \frac{\pi(1 - \lambda^*\overline{p})}{1 - \lambda^*\mu(\pi)}, \quad \pi^{x_1,1}(\pi) := \frac{\pi\overline{p}}{\mu(\pi)}.$$

Finally, consider having both tools at period t. The event $(v_{x_0 t}, v_{x_1 t}) = (0,0)$ is uninformative about p, so DM learns nothing new. If $(v_{x_0 t}, v_{x_1 t}) = (1,0)$, the posterior is:

$$\pi^{X,(1,0)}(\pi) := \frac{\pi(1-\overline{p})}{1-\mu(\pi)} = \pi^{x_0,1}(\pi).$$

Finally, if $(v_{x_0 t}, v_{x_1 t}) = (0,1)$, we get:

$$\pi^{X,(0,1)}(\pi) := \frac{\pi\overline{p}}{\mu(\pi)} = \pi^{x_1,1}(\pi).$$

Since arrivals can correspond to at most one arm, these last two posteriors are the same as those derived from the single "successful" tool. The informational advantage of having both tools will appear in the probabilities that lead to these posteriors. The Bellman

equation is:

$$
\begin{aligned}
w^*(\pi) = \max \{ & \delta w^*(\pi), \lambda^* (1 - \mu(\pi)) - c_{x_0} \\
& + \delta \left[\lambda^* (1 - \mu(\pi)) w^* \left(\frac{\pi(1-\overline{p})}{1-\mu(\pi)} \right) + (1 - \lambda^* (1 - \mu(\pi))) w^* \left(\frac{\pi(1-\lambda^*(1-\overline{p}))}{1-\lambda^*(1-\mu(\pi))} \right) \right], \\
& \lambda^* \mu(\pi) - c_{x_1} + \delta \left[\lambda^* \mu(\pi) w^* \left(\frac{\pi \overline{p}}{\mu(\pi)} \right) + (1 - \lambda^* \mu(\pi)) w^* \left(\frac{\pi(1-\lambda^* \overline{p})}{1-\lambda^* \mu(\pi)} \right) \right], \\
& \lambda^* - c_{x_0} - c_{x_1} + \delta \left[\lambda^* \mu(\pi) w^* \left(\frac{\pi \overline{p}}{\mu(\pi)} \right) + \lambda^* (1 - \mu(\pi)) w^* \left(\frac{\pi(1-\overline{p})}{1-\mu(\pi)} \right) + (1 - \lambda^*) w^*(\pi) \right] \Bigg\}.
\end{aligned}
$$

Figure 1 presents the myopic and optimal strategies when $\lambda^* = \frac{4}{5}$, $\underline{p} = \frac{1}{10}$, $\overline{p} = \frac{9}{10}$, $\delta = \frac{9}{10}$, and $c_{x_0} = c_{x_1} = \frac{1}{3}$. Under the myopic strategy, DM chooses both tools when she is sufficiently unsure about which tool is more likely the "more-frequently valuable" one. Under the optimal strategy, she carries both tools more often, as doing this provides valuable information.

The standard approach offers partial, asymptotic, characterizations; I state these as Propositions 3 to 5. Proposition 3 says that, under the optimal strategy, exploration eventually gives way to exploitation. As the decision maker gathers more and more information, the impact of further information on her assessment fades.

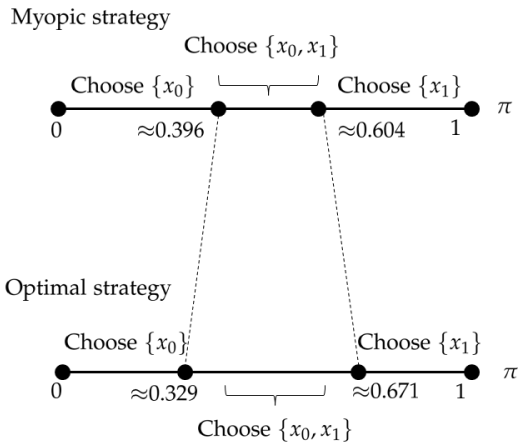

Figure 1: Myopic and Optimal strategy; $\lambda^* = \frac{4}{5}$, $\underline{p} = \frac{1}{10}$, $\overline{p} = \frac{9}{10}$, $\delta = \frac{9}{10}$, $c_{x_0} = c_{x_1} = \frac{1}{3}$.

Proposition 3. *Let $w^{\mathcal{M}} : \mathcal{D} \to \mathbb{R}$ be the myopic-value function:*

$$w^{\mathcal{M}}(\pi) := \frac{\max_{a \in Z} r(a, \mu_\pi)}{1 - \alpha}.$$

Then, $\left| w^*(\pi_t) - w^{\mathcal{M}}(\pi_t) \right| \xrightarrow{a.s.} 0$.

Proposition 4 says that, as the decision maker becomes increasingly patient, she can approach her full-information payoff arbitrarily closely. A strategy that achieves this is to follow the myopically optimal strategy given current beliefs most of the time, while trying all tools infinitely often, but with vanishing frequency. Formally, for each $n \in \mathbb{N}$, consider the strategy that selects all of X every period 2^n and follows the myopically optimal strategy every other period. Under this strategy, the decision maker eventually learns the true distribution $\mu^{\mathcal{T}}$ almost surely; by choosing myopically according to her posterior, and experimenting less and less often, she is eventually following the objectively-optimal strategy.

Of course, it can take an arbitrarily long time for the decision maker to learn the true distribution with arbitrary precision. But if she is infinitely patient, this cost of initial exploration is outweighed by the (distant) future rewards of "getting it right."

Proposition 4. *Let $r^{\mathcal{T}} := \max_{a \in Z} r\left(a, \mu^{\mathcal{T}}\right)$ be the average value of the problem for a decision maker who knows $\mu^{\mathcal{T}}$, and let $w^*_\alpha : \mathcal{D} \to \mathbb{R}$ be the optimal-average-value function,*

$$w^*_\alpha(\pi) := (1 - \alpha) \max_{\sigma \in \Sigma} w^\sigma(\pi).$$

For any $\pi \in \mathcal{D}$, $\lim_{\alpha \to 1} w^*_\alpha(\pi) = r^{\mathcal{T}}$.

A sufficiently impatient decision maker, on the other hand, can do best not exploring at all. If the weight on future rewards is low enough, the rewards to exploring are too low to be worthwhile. Proposition 5 establishes that myopic behavior is approximately optimal for a sufficiently impatient decision maker. Formally, a (stationary) strategy $\sigma \in \Sigma$ is *ϵ-optimal* for some $\epsilon > 0$ if $w^\sigma(\pi) > w^*(\pi) - \epsilon$ for all $\pi \in \mathcal{D}$.

Proposition 5. *Given $\pi \in \mathcal{D}$, let $\sigma^{\mathcal{M}}(\pi) \in \arg\max_{a \in Z} r(a, \mu_\pi)$ denote the myopic strategy. For any $\epsilon > 0$ there exists an $\alpha_\epsilon \in (0, 1)$ such that $\sigma^{\mathcal{M}}$ is ϵ-optimal if $\alpha \in (0, \alpha_\epsilon)$.*

4 Some Heuristics

The standard approach doesn't take us much further than partial characterizations and approximation results except in very simple examples. Therefore, I turn to *heuristics*. Formally, a heuristic is a strategy, an element of Σ. However, following Baumol and Quandt (1964), I reserve the term for strategies that are "simple" to formulate and follow and that involve revisions based on past performance.

The variety of heuristics that we could employ for this problem is wide. I look at specific heuristics that share the following characteristics:

1. The decision maker starts by employing all tools in X for the first $T_1 \in \mathbb{N}$ periods.

2. At the end of date T_1, she undertakes the first evaluation and revision, switching — possibly randomly — to a set $X_1 \subseteq X$. The chosen set X_1 is employed until (the end of) date $T_2 \in \mathbb{N}, T_2 > T_1$, at which point she performs another revision.

3. The process is continued, for a specified increasing sequence of revision times $\{T_m\}_{m=1}^{+\infty}$. The bundle recommended at period t, a_t, is X_m for $T_m < t \leq T_{m+1}$.

See Figure 2. To identify specific heuristics that conform to these general characteristics, it remains to specify the sequence of revision dates and the revision criteria.

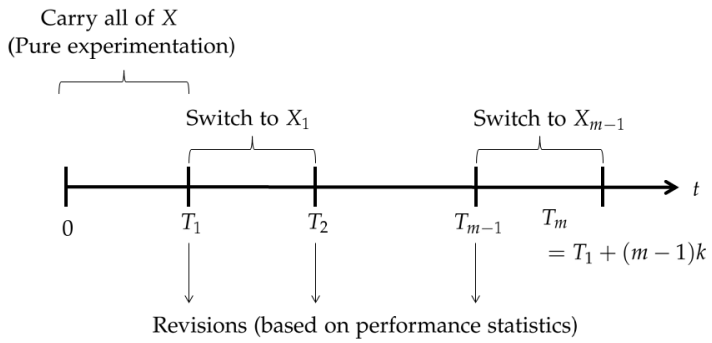

Figure 2: Structure of heuristics

For the professional-services firm, we can think of the "tool values" as the monetary value to the clients of the experts' reports. The total-perfect-substitution specification for profits represents the case where experts are perfectly substitutable in providing recommendations to the clients. All experts employed must be paid, but only the best of the reports produced is sold.

Recruitment heuristics that conform to the structure in Figure 2 recommend initially hiring the whole pool of experts, for a fixed probation period. During this period, employees keep a record of the performance of the candidates. This record is employed to evaluate the candidates; some of them are laid off. The remaining candidates are hired under short-term contracts, and performances are reassessed.

The heuristics discussed in what follows are *prior-free heuristics*: Their revision criteria ignore the decision maker's prior. They are all that is available to a decision maker who is unwilling either to specify or to employ a prior assessment. When she follows prior-free heuristics, the decision maker is a *naive empiricist*. Heuristics that employ the decision maker's prior, *prior-based heuristics*, are to be explored in a companion paper.

Prior-free heuristics are universally applicable; they recommend the same course of action to any decision maker, regardless of her prior beliefs (or lack thereof). Prior-based heuristics must be tailored to the beliefs of the decision maker, but they allow us to incorporate an accurate measure of the value of information into decision making.

Tool-based heuristics

I begin with four heuristics that have the individual tool as unit of analysis. These heuristics are adapted from what I perceive is common practice in cost accounting, where individual facilities/processes/products/factors of production are evaluated according to whether they "pay for themselves" in terms of their historical contribution.[14]

The first of the heuristics below makes sense only under total perfect substitution. The rest of the heuristics have a wider applicability; accordingly, they will be introduced in terms of an arbitrary submodular function W.

Total-contribution heuristic. *At each T_m and for each $x \in X_{m-1}$, compute:*

$$v^{TC}_{T_m}(x; X_{m-1}) := \frac{1}{T_m} \sum_{\tau \leq T_m} v_{x\tau} s_x(X_{m-1}, v_\tau),$$

[14]Conversations with members of the Stanford GSB faculty who specialize in managerial accounting suggest that this is common practice. It would be interesting to look into how much such rules resemble evaluations in college admissions, or hiring decisions of assistant faculty.

where:

$$s_x(a,v) := \begin{cases} 1 & x \in a \ \text{ and } \ v_x > v_{x'} \ \text{ for all other } \ x' \in a, \\ 0 & x \notin a \ \text{ or } \ v_{x'} > v_x \ \text{ for some other } \ x' \in a; \end{cases}$$

set $X_0 := X$, $s_x(\{x\}, v) := 1$ *and* $s_x(\emptyset, v) := 0$ *for all* $x \in X$. *That is,* $s_x(a,v) = 1$ *if* x *is the tool employed from bundle* a *when the realized value-profile is* v.[15] *That is, for each* x *carried so far, compute its (average) total gross contribution. Set* $X_m = \left\{ x \in X_{m-1} : v^{TC}_{T_m}(x; X_{m-1}) \geq c_x \right\}$; *discard all tools whose total gross contribution to the latest bundle falls short of their cost.*

A professional-services firm following this heuristic to hire experts will evaluate their candidates comparing the average revenues each of them generated for the firm, the average value of their reports that were *actually sold*, against their wages. Candidates whose wage is higher than the revenues created are dismissed.

While this heuristic has a certain intuitive appeal, it has some obvious defects. It only gives credit to a tool if used, while the decision to use a tool depends only on values, not costs. Hence, a tool whose value is dominant will prevail, if its cost is not too high relative to its benefits, however more expensive than others it may be. Moreover, the heuristic doesn't recognize that smaller subsets of the selected tools might do better. It gives too much credit to "star" experts, but can also lead to a sub-optimally large staff.

Example 2. Let $X = \{x, x'\}$, with $\mu^{\mathcal{T}}$ prescribing that $v_x = 100$ and $v_{x'} = 99$ with probability 1. Let $c_x = 10$ and $c_{x'} = 1$. The objectively-optimal strategy is to carry x' alone. But this tool will never be used if accompanied by x. Hence, the total-contribution heuristic leads the decision maker to keep x instead of x'.

Example 3. Suppose that we have two tools, $X = \{x, x'\}$, and that $\mu^{\mathcal{T}}$ assigns probability 0.5 to the vector $v = (v_x, v_{x'}) = (100, 98)$ and probability 0.5 to $v = (98, 100)$. Costs are $c_x = 10$ and $c_{x'} = 11$. With a representative sample of value profiles, experience will indicate that both tools pay for themselves, and the heuristic will recommend keeping both. However, the decision maker is better off keeping tool x only.

These examples suggest that, among other things, we give too much credit to a tool by crediting it with the total value produced when it attains the highest gross value. For instance, a firm following this heuristic completely overlooks the work of experts whose analysis is below that of star employees, even if it is only marginally so and yet they command a much lower wage.

[15] For notational simplicity, I ignore ties. The analysis can be modified to accommodate ties.

Given the full set of tools we are carrying, we might do better to evaluate tools based on the *incremental* value they contribute to the set. Consider the following evaluation measure, the *incremental contribution* of tool $x \in a$ to the bundle a:

$$v_T^{IC}(x;a) := \frac{1}{T}\sum_{\tau \leq T}\left[W(\nu(a,v_t)) - W(\nu(a \setminus \{x\},v_t))\right].$$

In words, we look at the difference between the average gross contribution of the current bundle and that of the bundle with the tool in question removed.

In both examples, the incremental contributions of both tools (to the full set X) are below the costs. If we were to discard all tools whose incremental contributions fall short of their cost, the decision maker would be left (suboptimally) empty-handed. The incremental contribution is useful for evaluating one tool at a time, assuming the other tools will be kept. Therefore, evaluations must be performed "sequentially."

Incremental-contribution heuristic. *At each time T_m, all tools in X_{m-1} are evaluated according to $v_{T_m}^{IC}(\cdot; X_{m-1})$. If any tool provides an incremental contribution below its cost, the one with the least incremental contribution net of cost is dropped. (Any convenient tie-breaking rule can be used if there are ties.) If tool $x^1 \in X_{T_{m-1}}$ is removed, the remaining tools are re-evaluated according to $v_{T_m}^{IC}(\cdot; X_{m-1} \setminus \{x^1\})$. The process is repeated until all remaining tools, if any, have incremental contributions that cover their costs; X_m is the set of "surviving" tools.*

If the professional-services firm follows this second heuristic, it will evaluate each candidate comparing the difference between the firm's actual average revenues and the average revenues that would have obtained without this candidate, against her wage. Those candidates who command a wage that is higher than their contribution to average revenues are laid off.

In Examples 2 and 3, this heuristic gives the "right" answer (under a representative sample). In fact, it leads to the objectively-optimal bundle, under a representative sample, in all two-tool problems;[16] but it can fail if there are more than two tools.

Example 4. Suppose that there are three tools, $X = \{x, x', x''\}$, with $v_x = 0.9, v_{x'} = 1$, and $v_{x''} = 0$ with probability 0.5, while $v_x = 0.9, v_{x'} = 0$ and $v_{x''} = 1$ also with probability 0.5 under $\mu^{\mathcal{T}}$. Costs are $c_x = 0.8, c_{x'} = 0.6$, and $c_{x''} = 0.7$. The objectively-optimal bundle is the singleton $\{x\}$. In terms of (gross) values, however, the pair x', x'' dominates x. Tool x gets no credit for contributing in the first T_1 rounds. While all tools have a negative incremental contribution, that of x is the lowest. As a result, x is the first tool to be

[16] See Claim 1 in the Appendix.

dropped. Tools x'' and x' are dropped next, in that order. The decision maker is left empty-handed, and the objectively-optimal tool is the first one discarded.

These heuristics ignore the variability of tool values; they focus on mean rewards. The next heuristic takes this additional information into account. Incremental contributions are augmented with standard deviations, to bias evaluations in favor of tools for which information is imprecise. In terms of the hiring example, the firm evaluates contributions to revenues taking into account the variability of revenues.

Augmented-incremental-contribution heuristic. *Given a weight parameter $\theta > 0$, tool $x \in X_{m-1}$ is evaluated at time T_m according to $v^{IC}_{T_m}(x; X_{m-1}) + \theta \frac{\hat{\sigma}_{xT_m}}{\sqrt{T_m}}$, where $\hat{\sigma}_{xT_m}$ is the sample standard deviation of v_x.*[17] *The revision criterion is the same as that of the incremental-contribution heuristic.*

The rationale for augmenting the performance statistic of a tool with the sample standard deviation of its value is to incorporate a measure of the uncertainty regarding the tool. Without invoking the prior, sample variance is the natural choice for a measure of uncertainty. Of course, this is a rough measure. It conflates two forms of uncertainty: variability of tool values under the true distribution, and variability of the unknown distribution according to the prior. While we want to experiment longer with tools that have larger uncertainty of the second type only, we can't separate the two under prior-free heuristics.

Bundle-based heuristics

While the incremental-contribution heuristic takes into account the impact, on any given tool, of the other tools that are available, it can be "unfair" to individual tools. Example 3 makes this clear. When we compare x against the bundle $\{x', x''\}$, it takes both (collectively) to beat x. However, we don't take into consideration the cost of this doubleton when we compute the incremental contribution of each tool. Therefore, I turn to heuristics that have bundles as unit of analysis.

I begin by introducing a general-purpose bundle-based evaluation statistic. Under a given heuristic, a given bundle a may be contained in the recommended a_t for some t but not for all t; we only have data on all of the tools in a for those times t such that

[17] If $\sigma > 0$ is the standard deviation of a random variable X, the standard deviation from the sample average of a random sample from X consisting of n observations is $\frac{\sigma}{\sqrt{n}}$. With a diffuse prior, sample moments are also Bayesian posterior moments.

$a \subseteq a_t$. Moreover, the different a_t may not be nested. For $a \in Z$ and $T = 0, 1, \ldots$, let:

$$V_T^{\mathcal{B}}(a) := \frac{\sum_{t=0}^{T} W(\nu(a, v_t)) I\left(a \subseteq a_t\right)}{\sum_{t=0}^{T} I\left(a \subseteq a_t\right)},$$

where $I\left(a \subseteq a_t\right)$ equals 1 if $a \subseteq a_t$ and 0 otherwise. The first bundle-based heuristic involves choosing X_m as the subset of X_{m-1} that has had the highest average $V_{T_m}^{\mathcal{B}}$. This is what the decision maker would optimally do if X_{m-1} where the entire set and if the empirical distribution where the true distribution.

Simple bundle-based heuristic. *At date T_m, set X_m to be the subset a of X_{m-1} that maximizes $V_{T_m}^{\mathcal{B}}(a) - \sum_{x \in a} c_x$ (breaking ties in any convenient fashion).*

Augmented simple bundle-based heuristic. *Given $\theta > 0$, at date T_m, set X_m to be the subset a of X_{m-1} that maximizes $V_{T_m}^{\mathcal{B}}(a) + \theta \frac{\hat{\sigma}_{aT_m}}{\sqrt{T_m}} - \sum_{x \in a} c_x$, where $\hat{\sigma}_{aT_m}$ is the sample standard deviation of $W(\nu(a, v))$ (breaking ties in any convenient fashion).*

If the firm follows these heuristics to hire experts, it will evaluate candidates in teams rather than individually. The team whose contract is renewed is the team that generates the highest average profits.

Revisions under these heuristics mimic the optimal strategy under no learning, with sample averages standing in for expectations. The non-augmented version picks out the objectively-optimal bundle with probability approaching 1, under $\mu^{\mathcal{T}}$, as T_1 goes to infinity. However, for any fixed T_1, it can lead the decision maker astray if she is unlucky and gets a non-representative sample of tool values.

Given any two actions a, a', the *regret for action a' based on action a* is a measure of the difference in the payoff the decision maker would have enjoyed had she taken action a' whenever she took action a in the past. We propose the following heuristics based on the notion of "regret matching" from Hart and Mas-Colell (2000).

Regret-matching heuristic. *At period T_m, compute the vector of regrets for each $a \subseteq X_{m-1}$*

$$R_{T_m} := \left[\max \left\{ V_{T_m}^{\mathcal{B}}(a) - V_{T_m}^{\mathcal{B}}\left(X_{m-1}\right) + \sum_{x \in X_{m-1} \setminus a} c_x, 0 \right\} \right]_{a \subseteq X_{m-1}}.$$

Draw X_m from Z according to the probability distribution obtained by normalizing the vector of regrets, provided it is non-null, and keep the same bundle otherwise.[18]

[18] We can allow for *inertia*, namely a positive probability of sticking to the previous choice even if there

Augmented regret-matching heuristic. *Given $\theta > 0$, at date T_m, compute augmented regrets:*

$$R^{\theta}_{T_m} := \left[\max \left\{ V^{\mathcal{B}}_{T_m}(a) + \theta \frac{\hat{\sigma}_{aT_m}}{\sqrt{T_m}} - V^{\mathcal{B}}_{T_m}\left(X_{m-1}\right) - \theta \frac{\hat{\sigma}_{X_{m-1}T_m}}{\sqrt{T_m}} + \sum_{x \in X_{m-1} \setminus a} c_x, 0 \right\} \right]_{a \subseteq X_{m-1}}.$$

Carry on revisions as under the regret-matching heuristic.

When the firm follows regret-matching heuristics, it looks at all possible teams that can be formed out their current pool of experts. Recontracting for the next term is random, with higher priority given to the teams that would have performed better than the current pool.

Example 5. Consider a single tool with value $v = 0$ with probability 0.99 and $v = 51$ with probability 0.01, and with rental cost equal to $\frac{1}{2}$. The objectively-optimal bundle is the non-empty bundle. However, the likelihood is high that, for a small T_1, the observed sample will feature only zeros and the tool will be discarded.

The problem in Example 5 would be averted if the tool were picked back up often enough. A simple way to "bring back" tossed tools while preserving the accuracy of long-run recommendations is to combine myopic choices with experimentation. The next heuristic is inspired in the numerical-optimization method of simulated annealing, which combines traditional optimization methods with random exploration.

Simulated-annealing heuristic. *Fix a parameter $p \in (0, 1]$. For period T_m, let $a^*_{T_m}$ be the maximizer of $V^{\mathcal{B}}_{T_m}(a) - \sum_{x \in a} c_x$ out of all the possible subsets of X_{m-1} (using any convenient tie-breaking rule, if needed). Then,*

$$X_m := \begin{cases} a' & \text{for each non-empty } a' \in Z, \text{ each with probability } p / \left[m 2^{(\#X-1)} \right], \\ a^*_{T_m} & \text{with probability } 1 - p/m. \end{cases}$$

The randomization in this rule is independent across revision periods, and independent of all other sources of uncertainty in the problem.

The overall probability of "experimenting," p/m, is chosen so that every bundle is chosen infinitely often almost surely (by the second Borel-Cantelli lemma), but at the same time the frequency of experimenting goes to zero. The parameter p represents a weight on experimentation: the higher the value of p, the more frequent the choice

is positive regret for other alternatives.

will be at random, giving all tools a chance. Note that, under the simulated-annealing heuristic, choices need not be nested.

In this case, the firm usually evaluates its employees as in the simple bundle-based heuristic. Occasionally, instead of making decisions based on performance records, it selects a team at random for an additional term or probation period. This team may include candidates previously laid off.

5 Evaluating Heuristics: Formal Criteria

These heuristics are not, of course, the only possibilities. Moreover, some of their flaws have been pointed out. How should heuristics be evaluated? The obvious answer is, They should be evaluated relative to how they compare to the optimal strategy. But this is a difficult criterion to implement: We don't know how to characterize the optimal strategy (in general). Instead, I ask in this section whether a heuristic makes the "right" choice when the decision maker has accumulated enough information to make the right choice. Or at least, whether it makes a choice upon which there is no obvious improvement.

When provided with "enough" information to choose the objectively-optimal bundle, a "good" heuristic ought to make that choice. This notion can be operationalized in two ways: for what happens as the initial revision time T_1 goes to infinity, and for decisions that are made in the long run. If a heuristic is very likely to choose the objectively-optimal bundle most of the time, provided that initial exploration is sustained for a sufficiently long time, I shall say that the heuristic is *asymptotically accurate*.

Definition 1. *A family of heuristics indexed by initial revision date $T_1 \in \mathbb{N}$ is **asymptotically accurate** if it selects the objectively-optimal bundle a fraction of the time that approaches 1 with probability approaching 1, as T_1 goes to infinity. In symbols,*

$$P^{\mu^{\mathcal{T}}}\left(\lim_{T\to\infty}\frac{\sum_{t=0}^{T} I\left(a_t = a^{\mathcal{T}}\right)}{T} \text{ exists and equals } 1\right) \underset{T_1\to+\infty}{\longrightarrow} 1,$$

where a_t is the (random) bundle chosen at time t under H.

Since such an indexed family of heuristics differ only on the initial revision date, I shall refer to asymptotic accuracy as a property of heuristics. While asymptotic accuracy is (in principle) desirable, carrying all tools for a long period of time is likely to be far from optimal for decision makers with discount factors bounded away from 1.

However, to the extent that the decision maker eventually settles on some bundle, she will accumulate (as time goes by) accurate information about that bundle and about all its sub-bundles. A good heuristic should not have the decision maker choosing bundles that contain better sub-bundles a non-vanishing fraction of the time. This gives the second formal criterion, which calls for a preliminary definition.

Definition 2. *A bundle $a \in Z$ is **internally stable relative to** $\mu \in \Delta(V)$ if, for all $a' \subseteq a$, $r(a, \mu) \geq r(a', \mu)$. Fixing $\mu^{\mathcal{T}}$, $a \in Z$ is **internally stable** if it is internally stable relative to $\mu^{\mathcal{T}}$.*

Definition 3. *A heuristic H satisfies the **no-better-subsets** criterion if, for every bundle $a \in Z$ that is not internally stable,*

$$P^{\mu^{\mathcal{T}}}\left(\limsup_{T\to\infty} \frac{\sum_{t=0}^{T} I(a_t = a)}{T} > 0\right) = 0.$$

Let $\mathcal{S}(X, \mu)$ be the collection of subsets of X that are internally stable relative to $\mu \in \Delta(V)$. The next proposition shows that, by submodularity of W, this collection satisfies the property of "downward closure": Any subset of a bundle that is internally stable relative to μ is also internally stable relative to μ.

Proposition 6. *For any $\mu \in \Delta(V)$, if $a \in Z$ is internally stable relative to μ, then so is any $a' \subseteq a$: that is to say, if $a \in \mathcal{S}(X, \mu)$, and $a' \subseteq a$, then $a' \in \mathcal{S}(X, \mu)$.*

While not the ultimate test of a heuristic, these criteria capture desirable properties. The examples presented in Section 4 show that the total-contribution heuristic fails on both counts, while the incremental-contribution and the regret-matching heuristics are not asymptotically accurate. The positive results are given below. Proposition 7 refers to the incremental-contribution heuristic.[19] Proposition 8 summarizes the formal properties of the simple bundle-based and the simulated annealing heuristics.

Proposition 7. *The incremental-contribution heuristic satisfies the no-better-subsets criterion.*

[19] A question that is left for future research is whether we can identify a list of desirable properties, introduced as axioms, that pins down a small class of heuristics.

Proposition 8. *The simple and the simulated-annealing heuristics are asymptotically accurate and satisfy the no-better-subsets criterion. Moreover, the simulated-annealing heuristic chooses the objectively-optimal bundle under $\mu^{\mathcal{T}}$ a fraction of time that approaches 1 with probability 1.*[20]

The next proposition presents the formal results corresponding to the regret-matching heuristic. The proposition establishes that it satisfies the no-better-subsets criterion, as well as a weak form of asymptotic accuracy: It selects the objectively-optimal bundle a fraction of the time that doesn't vanish with probability approaching 1, as T_1 goes to infinity.[21]

Proposition 9. *The regret-matching heuristic satisfies the no-better-subsets criterion. Moreover, it selects the objectively-optimal bundle a fraction of the time that doesn't vanish with probability approaching 1, as T_1 goes to infinity. In symbols,*

$$P^{\mu^{\mathcal{T}}}\left(\lim_{T\to\infty}\frac{\sum_{t=0}^{T} I\left(a_t = a^{\mathcal{T}}\right)}{T} \quad \textit{exists and is positive}\right) \underset{T_1\to+\infty}{\longrightarrow} 1.$$

where a_t is the (random) bundle chosen at time t.

The additional term in the augmented versions of the heuristics converges to 0 since sample variances go to 0, by the Strong Law of Large Numbers. Hence, Propositions 7 to 9 apply to the augmented counterparts as well.

6 Evaluating Heuristics: Simulations

While formal criteria provide some guidance, they represent long-run considerations, involving asymptotically infinite amounts of data. While this is relevant for infinitely patient decision makers, we don't know what is optimal for discount factors bounded away from 1. Hence, heuristics must be assessed empirically. There are (at least) two ways to interpret this: What heuristics are descriptive of what people do in practice? More normatively, how do different heuristics perform in specific test problems?

I address the second question by performing computer simulations:

[20] While the proof of Proposition 7 invokes submodularity of W, Proposition 8 is true for any gross-reward function W.

[21] The first part of this claim is true provided that there is no inertia in the heuristic (see footnote 18). With inertia, the regret-matching heuristic satisfies a weak form of the no-better-subsets criterion: It recommends switching away from bundles that are not internally stable with positive probability.

1. A test problem is fixed. Test problems consist of a specification of the sets X and V, the function W, and the parameters c, α, and $\mu^{\mathcal{T}}$.

2. A time horizon, T, and the number of runs, B, are set.

3. The revision date T_1 is set, with $T_m = T_1 + (m-1)k$ for a given k; the parameters p and θ are chosen.[22]

From each simulation, I collect a number of performance measures, including the average normalized utility and the frequency of choice of each bundle in the last period under each heuristic. Tool values are drawn from a normal population with unknown mean but known variance denoted by Σ_0.[23] The prior on the vector μ of means is a normal distribution, with mean μ_0 and variance S_0. Tables 2 and 3 provide some results for a four-tool test problem, under the total-perfect-substitution specification. Expected rewards are the expected first-order statistic of the tools "sampled." The formula for this expectation is taken from Afonja (1972). The cost vector is $c = (0.5, 0.5, 0.5, 0.5)$, and the discount factor is $\alpha = 0.99$. The prior mean is $\mu_0 = (4, 4, 4, 4)$. We have $T = 100$, $T_1 = k = 10$, and $p = \theta = 1$. The program was run $B = 10000$ times.[24] The data in Table 2 was generated specifying the following variance-covariance matrices:

$$S_0 = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}; \quad \Sigma_0 = \begin{bmatrix} 4 & -3.96 & 0 & 0 \\ -3.96 & 4 & 0 & 0 \\ 0 & 0 & 4 & -3.96 \\ 0 & 0 & -3.96 & 4 \end{bmatrix}.$$

For Table 3, these are:

$$S_0 = \begin{bmatrix} 0.01 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 10 \end{bmatrix}; \quad \Sigma_0 = \begin{bmatrix} 1 & -0.99 & 0 & 0 \\ -0.99 & 1 & 0 & 0 \\ 0 & 0 & 1 & -0.99 \\ 0 & 0 & -0.99 & 1 \end{bmatrix}.$$

In the specification of Table 2, the decision maker is equally uncertain about the mean values of the tools; in Table 3, she is fairly certain about the mean values tools x_1 and

[22] Something that remains to be done is to see how the heuristics perform as these parameters are varied; especially, if they are set endogenously.

[23] The normal specification, of course, violates the finite-support assumption made earlier to simplify the analysis. However, the simulations are based on "discretizations" of the normal distribution, as they were performed via computer.

[24] Details of the simulations, including the MATLAB routines employed, are available upon request.

x_2, while she is substantially uncertain about the means of tools x_3 and x_4. As for tool values, tools x_1 and x_2 are almost perfectly negatively correlated, as are x_3 and x_4. These two blocks are independent of each other. Under total-perfect substitution and given the costs, these blocks are more frequently profitable than as singletons.

Some observations suggested by these tables, consistent with other simulations not reported here, are the following:

- Larger sets are generally chosen less frequently than they are objectively optimal. In particular, singletons are chosen more frequently than it is optimal.
- Consistent with the first observation, the frequency of reaching the objectively-optimal bundle is higher when singletons are objectively optimal.
- The relative performance of all the heuristics is consistent across simulations. In particular, incremental-contribution heuristics, simple bundle-based heuristics, and the simulated-annealing heuristic are the ones that perform the best in terms of average normalized payoff.

Bundle	Optimal	TC	IC	Aug. IC	Simple	Aug. Simple	Regret	Aug. Reg.	SA
$\emptyset$	0	0	0	0	0	0	0	0	0
$\{x_1\}$	0.0027	0.0082	0.0891	0.0887	0.1062	0.0177	0.1153	0.1107	0.1364
$\{x_2\}$	0.0031	0.0058	0.0834	0.0821	0.1014	0.0146	0.1136	0.1043	0.1352
$\{x_4\}$	0.0025	0.0068	0.0834	0.0837	0.0991	0.0142	0.1101	0.1012	0.1335
$\{x_5\}$	0.0037	0.0069	0.0949	0.0948	0.112	0.0171	0.1225	0.1161	0.1433
$\{x_1, x_2\}$	0.289	0.1143	0.2187	0.2228	0.2035	0.2577	0.1685	0.1888	0.1406
$\{x_1, x_3\}$	0.029	0.0331	0.0505	0.0487	0.0392	0.0499	0.0496	0.0475	0.0431
$\{x_1, x_4\}$	0.0364	0.0378	0.0556	0.0536	0.0433	0.0561	0.0522	0.049	0.0442
$\{x_2, x_3\}$	0.0332	0.0336	0.051	0.0491	0.0406	0.0516	0.0467	0.0461	0.0425
$\{x_2, x_4\}$	0.0308	0.0349	0.05	0.0486	0.0403	0.0491	0.0455	0.044	0.0426
$\{x_3, x_4\}$	0.2906	0.1127	0.2227	0.2275	0.2141	0.2681	0.1758	0.1913	0.1385
$\{x_1, x_2, x_3\}$	0.0643	0.122	0.0001	0	0	0.0483	0	0.0003	0
$\{x_1, x_2, x_4\}$	0.0754	0.1249	0.0004	0.0003	0.0001	0.0539	0.0001	0.0004	0.0001
$\{x_1, x_3, x_4\}$	0.0741	0.1295	0.0001	0.0001	0	0.053	0.0001	0.0002	0
$\{x_2, x_3, x_4\}$	0.0652	0.1245	0.0001	0	0.0001	0.0481	0	0.0001	0
$\{x_1, x_2, x_3, x_4\}$	0	0.105	0	0	0	0.0006	0	0	0

(a) Frequency of optimality and of choices

Statistic	TC	IC	Aug. IC	Simple	Aug. Simple	Regret	Aug. Reg	SA
Utility	4.9067	4.974	4.977	5.1827	4.9624	4.923	4.9016	4.8411
Frequency of optimal	0.3265	0.4224	0.4272	0.3915	0.5899	0.34	0.3379	0.2632
Frequency of too few tools	0.0839	0.4246	0.4219	0.4793	0.1361	0.5166	0.4884	0.5822
Frequency of IS	0.9374	1	1	1	0.9999	1	1	1

(b) Average normalized payoff and choice statistics

Table 2: Summary of Simulations results. Panel 2a shows the frequency, across the 10000 runs, with which each bundle is objectively optimal (the column labelled "Optimal"), and ultimately selected by each heuristic. Panel 2b presents, for each heuristic, the average normalized payoff, the frequency of leading to the objectively-optimal bundle, to a proper subset of the latter, and to an internally-stable bundle.

Bundle	Optimal	TC	IC	Aug. IC	Simple	Aug. Simple	Regret	Aug. Reg.	SA
$\emptyset$	0	0	0	0	0	0	0	0	0
$\{x_1\}$	0	0.0018	0.0844	0.0833	0.1018	0.0258	0.1137	0.1102	0.1241
$\{x_2\}$	0	0.002	0.0833	0.0829	0.103	0.0215	0.1153	0.1185	0.13
$\{x_3\}$	0.3294	0.1919	0.3434	0.3458	0.3593	0.3174	0.3537	0.3437	0.3479
$\{x_4\}$	0.3194	0.1848	0.334	0.3369	0.3489	0.3092	0.3454	0.3342	0.3412
$\{x_1, x_2\}$	0.2927	0.2108	0.1297	0.126	0.0747	0.2412	0.0598	0.0787	0.0494
$\{x_1, x_3\}$	0	0.0217	0.0008	0.0007	0.0001	0.006	0.0001	0.0011	0.0005
$\{x_1, x_4\}$	0	0.0215	0.0003	0.0003	0	0.0075	0.0002	0.0003	0.0002
$\{x_2, x_3\}$	0	0.0215	0.0004	0.0004	0.0003	0.0073	0.0002	0.0006	0.0001
$\{x_2, x_4\}$	0	0.0208	0.0007	0.0007	0.0002	0.0051	0.0003	0.0006	0.0003
$\{x_3, x_4\}$	0.0585	0.1492	0.023	0.023	0.0117	0.059	0.0113	0.0121	0.0063
$\{x_1, x_2, x_3\}$	0	0.0726	0	0	0	0	0	0	0
$\{x_1, x_2, x_4\}$	0	0.0725	0	0	0	0	0	0	0
$\{x_1, x_3, x_4\}$	0	0.0089	0	0	0	0	0	0	0
$\{x_2, x_3, x_4\}$	0	0.0106	0	0	0	0	0	0	0
$\{x_1, x_2, x_3, x_4\}$	0	0.0094	0	0	0	0	0	0	0

(a) Frequency of optimality and of choices

Statistic	TC	IC	Aug. IC	Simple	Aug. Simple	Regret	Aug. Reg	SA
Utility	5.1875	5.2645	5.265	5.4014	5.2843	5.2457	5.1976	5.188
Frequency of optimal	0.6294	0.7598	0.7607	0.7185	0.8677	0.6936	0.6782	0.6508
Frequency of too few tools	0.004	0.1936	0.1933	0.2486	0.0599	0.2609	0.2456	0.2698
Frequency of IS	0.9881	1	1	1	1	1	1	1

(b) Average normalized payoff and choice statistics

Table 3: Summary of Simulations results. Panel 3a shows the frequency, across the 10000 runs, with which each bundle is objectively optimal (the column labelled "Optimal"), and ultimately selected by each heuristic. Panel 3b presents, for each heuristic, the average normalized payoff, the frequency of leading to the objectively-optimal bundle, to a proper subset of the latter, and to an internally-stable bundle.

Table 4 reflects the bias towards singletons by comparing the frequency with which singletons are ultimately chosen by the heuristics against the frequency with which they are objectively optimal in the simulations. To obtain a clearer picture on the comparative performance of the heuristics in terms of average normalized payoffs, Tables 5 and 6 present tests of the hypotheses that the mean normalized payoff of each heuristic is equal to the expected payoff from the objectively-optimal bundle (Panels 5a and 6a), and equal to each other (Panels 5b and 6b). All tests are asymptotic tests at the 0.01 level.

The simulation results suggest that the best of these heuristics are the incremental-contribution and the simple bundle-based heuristics. As Example 4 indicates, however, the former can perform poorly in other test problems. This point is revisited in Table 7. There are 3 tools, and the maximum of the values of two of them are (almost surely) greater than that of the third. Costs are chosen so that this third tool is the objectively optimal choice, yet it has the lowest negative incremental contribution to the full set. If the firm follows the simple bundle-based heuristic, it will favor teams whose best report within the team has been, on average, the best across all teams. This heuristic can help identify synergies: The best expert in a team may not be as good on her own.

Singleton bias?	TC	IC	Aug. IC	Simple	Aug. Simple	Regret	Aug. Reg	SA
$\{x_1\}$	9091.988	31273.79	31196.66	34460.96	13422.5868	36098.89	35279.8629	39740.04
$\{x_2\}$	7637.165	30162.67	29905.46	33594.16	12171.3629	35797.39	34122.2657	39537.37
$\{x_3\}$	8273.674	30162.69	30221.84	33165.17	12001.082	35172.27	33553.3538	39249.47
$\{x_4\}$	8334.567	32378.85	32360	35513.63	13189.0123	37361.25	36240.2715	40896.47

(a) Statistics for the singleton-bias test for Table 2. All differences are significant.

Singleton bias?	TC	IC	Aug. IC	Simple	Aug. Simple	Regret	Aug. Regret	SA
$\{x_1\}$	4246.252	30359.64	30143.05	33672.61	16272.867	35815.26	35190.2928	37638.9
$\{x_2\}$	—	30143.05	30064.03	33881.14	14822.3512	36098.98	36662.8567	38653.63
$\{x_3\}$	48720.16	72308	72693.26	74875.77	68179.4557	73967.05	72356.1122	73031
$\{x_4\}$	47601.66	70806.46	71268.58	73188.99	66892.4339	72629.2	70838.2956	71955.72

(b) Statistics for the singleton-bias test for Table 3. All reported differences are significant.

Table 4: Singleton bias. Panel 4a shows the results of testing the hypothesis that the mean frequency of leading to each singleton is higher than the frequency with which said singleton is objectively optimal, using the data summarized in Table 2. Panel 4b presents the corresponding result for the data from Table 3. All tests are asymptotic tests at the 0.01 level.

Optimal?	TC	IC	Aug. IC	Simple	Aug. Simple	Regret	Aug. Reg	SA
Test	-64.2336***	-54.3236***	-54.0946***	-25.7671***	-58.8076***	-58.8028***	-60.9478***	-66.0625***

(a) Are mean normalized payoffs equal to the full-information value?

Equal means?	IC	Aug. IC	Simple	Aug. Simple	Regret	Aug. Regret	SA
TC	-23.6482***	-24.7394***	-40.6635***	-32.7027***	-4.9799***	1.4576	16.7787***
IC	—	-5.1537***	-31.6233***	4.6138***	23.9087***	22.3931***	35.5299***
Aug. IC	—	—	-31.224***	5.8661***	25.284***	23.3749***	36.4369***
Simple	—	—	—	33.0349***	38.5633***	39.7126***	47.0884***
Aug. Simple	—	—	—	—	13.01***	18.4648***	32.1495***
Regret	—	—	—	—	—	6.4365***	22.1164***
Aug. Regret	—	—	—	—	—	—	15.3894***

(b) Are mean normalized payoffs equal to each other?

Table 5: Summary of Tests for the results in Table 2. Panel 5a shows the results of testing the hypothesis that the mean normalized payoff is equal to the objectively-optimal payoff. Panel 5b shows the paired-sample tests of equal means. Values above the main diagonal are test statistics. All tests are asymptotic tests at the 0.01 level.

Despite its theoretical flaws, the relatively good performance of the incremental-contribution heuristic in the main test problems suggests that it might be an effective heuristic based on computational grounds. If the decision maker is carrying $n \leq \#X$ tools, the total-contribution heuristic requires her to compute n evaluations at the next evaluation date, and the incremental-contribution heuristic, at most $n(n+1)/2$. Bundle-based heuristics involve 2^n evaluations. If the cost of computational burden is taken into account, in a discounted formulation, the decision maker would like to make decisions about which tools to drop fairly quickly (albeit sensibly), to avoid their cost. Presumably,

Optimal?	TC	IC	Aug. IC	Simple	Aug. Simple	Regret	Aug. Reg	SA
Test	-26.0516***	-22.3355***	-22.315***	-14.9427***	-21.5853***	-23.1079***	-25.5703***	-25.9391***

(a) Are mean normalized payoffs equal to the full-information value?

Equal means?	IC	Aug. IC	Simple	Aug. Simple	Regret	Aug. Regret	SA
TC	-35.3365***	-35.4003***	-47.6419***	-54.8681***	-24.6869***	-4.0793***	-0.1814
IC	—	-1.3714	-33.1041***	-16.6174***	18.8594***	35.2627***	39.6881***
Aug. IC	—	—	-33.0093***	-15.8811***	20.1441***	35.4302***	40.0163***
Simple	—	—	—	28.1959***	37.459***	46.7016***	48.644***
Aug. Simple	—	—	—	—	26.5034***	45.6723***	48.8977***
Regret	—	—	—	—	—	24.3909***	30.3816***
Aug. Regret	—	—	—	—	—	—	4.0988***

(b) Are mean normalized payoffs equal to each other?

Table 6: Summary of Tests for the results in Table 3. Panel 6a shows the results of testing the hypothesis that the mean normalized payoff is equal to the objectively-optimal payoff. Panel 6b shows the paired-sample tests of equal means. Values above the main diagonal are test statistics. All tests are asymptotic tests at the 0.01 level.

she would also like to lighten the computational burden: If the decision maker follows bundle-based heuristics, every tool she drops cuts the number of future evaluations by half!

We know that, based on the formal criteria, tool-based heuristics can perform poorly in problems designed to make them fail. We also know that the simple bundle-based heuristic involves either waiting a long time to make choices or risking dismissal of valuable tools. The simulated-annealing heuristic brings prematurely-dismissed tools back with positive probability, and eventually "gets it right"; but it may take arbitrarily long to do so. What is needed is a heuristic that dismisses tools that are quickly seen to be of little value, and that brings back some previously-discarded tools if new information casts their past performance under a new, more favorable, light.

7 Remaining Work and (Interim) Conclusion

The findings in Section 6 are suggestive, but it remains to find a more systematic way to present them. Moreover, the simulations run so far present a puzzle. Tools with high variance are favored over tools with low variance, but tools with high variability of mean values are not favored over those with low variability.[25]

[25]I am currently engaged in investigating and comparing examples whose structure conforms to the example on top of page 32, but where the variance on the mean values of tools 3 and 4 (the entries 10 in S_0) takes on values 10, 7, 4, and 1, while the variability of the values of these two tools (given the mean) ranges from small to large. This should help clarify the effects of uncertainty about the mean and variability in tool value given the mean. I hope to be able to report these results in a revised version, shortly.

Statistic	TC	IC	Aug. IC	Simple	Aug. Simple	Regret	Aug. Reg	SA
Utility	0.4859	0.6011	0.6952	0.7511	0.5546	0.6302	0.5448	0.5726

(a) Average normalized payoff

Equal means?	IC	Aug. IC	Simple	Aug. Simple	Regret	Aug. Regret	SA
TC	-51.8169***	-89.7851***	-58.5797***	-34.5073***	-62.735***	-18.9049***	-30.7858***
IC	—	-63.8172***	-34.4158***	34.144***	-18.7807***	19.1004***	10.9606***
Aug. IC	—	—	-12.9636***	88.4437***	49.8922***	51.929***	48.5816***
Simple	—	—	—	45.6977***	27.7663***	42.8854***	38.6231***
Aug. Simple	—	—	—	—	-55.3583***	3.4145***	-7.2581***
Regret	—	—	—	—	—	29.1639***	22.2147***
Aug. Regret	—	—	—	—	—	—	-8.3117***

(b) Are mean normalized payoffs equal to each other?

Table 7: Poor performance of the incremental-contribution heuristic. There are three tools, x_1, x_2 and x_3. The values of tool x_3 are drawn from a normal distribution with mean 2 and variance 1. Given an independent gaussian noise ϵ, the values of tools x_1 and x_2 are, respectively, $v_{x_1} = v_{x_3} + \epsilon$ and $v_{x_2} = v_{x_3} - \epsilon$. Tool costs are $1 + \frac{1}{2\sqrt{\pi}}$ for tools x_1 and x_2, and 1 for tool x_3. Under this specification, the objectively-optimal bundle is the singleton $\{x_3\}$. However, under a representative sample, the incremental contribution of x_3 to the full set is negative and the lowest of all the tools.

Beyond these empirical findings, the essence of multi-armed-bandit problems is the exploration-exploitation trade-off. Exploration is valuable when there is uncertainty about the distribution, and one expects any but the most naive of decision makers to understand this, and to given tools whose mean value is more uncertain a "greater chance" to prove themselves. To put this sort of consideration into the story, we need to turn to prior-based heuristics, which is a direction of ongoing work.

Appendix

Lemma 1. *The lattices $(Z, \subseteq)$ and $\left(\{0,1\}^{\#X}, \geq\right)$ are order isomorphic.*

Proof. Define the function $\iota : Z \to \{0,1\}^{\#X}$ as $\iota(a) := (I(x_i \in a))_{i \in \{1,...,\#X\}}$. Pick any $y \in \{0,1\}^{\#X}$ and define $a_y := \{x_{y_i} : y_i = 1\}$; we have that $\iota(a_y) = y$. Moreover, for any $a, b \in Z$, $\iota(a) = \iota(b)$ if and only if $I(x_i \in a) = I(x_i \in b)$ for all $i \in \{1, ..., \#X\}$, which is equivalent to $a = b$. Thus, ι is a bijection. Moreover, $\iota(a) \geq \iota(b)$ if and only if $(I(x_i \in a))_{i \in \{1,...,\#X\}} \geq (I(x_i \in b))_{i \in \{1,...,\#X\}}$, or $b \subseteq a$. Thus, ι is order-preserving. □

Lemma 2. *If V is a lattice and W is submodular, $\rho(\cdot, v)$ is submodular for each $v \in V$.*

Proof. By Lemma 1, we can represent ν as the operation $* : \{0,1\}^{\#X} \times V \to V$ given by $a * v := (a_x v_x)_{x \in X}$; similarly, we can write the cost of bundle a as $c'a$. Fix $v \in V$ and

$a, a' \in \{0,1\}^{\#X}$. Since $V \subseteq \mathbb{R}_+^{\#X}$, we have $(a \vee a') * v = (a * v) \vee (a' * v)$ and $(a \wedge a') * v = (a * v) \wedge (a' * v)$. By submodularity of W, submodularity of $W(\nu(\cdot, v))$ follows:

$$\begin{aligned}
&W(\iota(a \vee a') * v) + W(\iota(a \wedge a') * v) \\
&= W((\iota(a) * v) \vee (\iota(a') * v)) + W((\iota(a) * v) \wedge (\iota(a') * v)) \\
&\leq W(\iota(a) * v)) + W((\iota(a') * v)).
\end{aligned}$$

As the sum of a submodular and a linear function is submodular, the result follows. □

Proof of Proposition 1. The result follows from Corollary 2.6.2. in Topkis (1998), by Lemma 2 and invoking $\left(V, 2^V\right)$ as the measurable space. □

Proof of Proposition 2. Since V is finite, the set of posteriors reachable from π_0 is countable. Restricting attention to countably many posteriors, measurability issues do not arise. Since (a) every conserving strategy is optimal, and (b) there are only finitely many available actions at each stage, we know that a conserving strategy exists.[26] □

Lemma 3. *If the decision maker chooses the full set X on period 2^n for $n \in N$ and follows the myopic strategy at all other dates, her posterior converges to a point mass at $\mu^{\mathcal{T}}$ almost surely.*

Proof. (*Argument suggested by Lanier Benkard*.) Under the proposed strategy, the partition will be complete (each cell will be a singleton) infinitely often. For any $k' \neq k$ such that $\mu^k, \mu^{k'} \in supp(\pi_{t+1})$:

$$\begin{aligned}
\frac{\pi_{t+1}(k|p_0, ..., p_t)}{\pi_{t+1}(k'|p_0, ..., p_t)} &= \frac{\pi_0(k)\mu^k(p_0)...\mu^k(p_t)}{\pi_0(k')\mu^{k'}(p_0)...\mu^{k'}(p_t)}, \\
\ln\left(\frac{\pi_{t+1}(k|p_0, ..., p_t)}{\pi_{t+1}(k'|p_0, ..., p_t)}\right) &= \ln\left(\frac{\pi_0(k)}{\pi_0(k')}\right) + \sum_{s=0}^{t}\left[\ln\left(\frac{\mu^k(p_s)}{\mu^{k'}(p_s)}\right)\right].
\end{aligned}$$

Let k^* index the true distribution; $\mu^{k^*} = \mu^{\mathcal{T}}$. For k^* and any other $k \neq k^*$,

$$\sum_{s=0}^{t} \ln\left(\frac{\mu^{\mathcal{T}}(p_s)}{\mu^k(p_s)}\right) = \sum_{s=0}^{t}\sum_{P \in \mathcal{P}}\left[\ln\left(\frac{\mu^{\mathcal{T}}(p_s)}{\mu^k(p_s)}\right)\right] I(P = P_s).$$

Now, as $t \to \infty$, there are some $P \in \mathcal{P}$ such that $P = P_s$ infinitely often and there are

[26] We can also establish existence by the methods of Blackwell (1965) and Maitra (1968). In particular, $r(a, \cdot)$ is Lipschitz continuous and $\tau(\cdot, a)$ is weakly continuous, for each $a \in Z$.

some others for which $P = P_s$ finitely often. In either case, for any $P \in \mathcal{P}$, we have:[27]

$$E^{\mu^{\mathcal{T}}}\left[\ln\left(\frac{\mu^{\mathcal{T}}(p)}{\mu^{k}(p)}\right)\right] = \sum_{v\in V}\mu^{\mathcal{T}}(p(v))\ln\left(\mu^{\mathcal{T}}(p)\right) - \sum_{v\in V}\mu^{\mathcal{T}}(p(v))\ln\left(\mu^{k}(p)\right) > 0.$$

Since the complete partition appears infinitely often, this strict inequality implies that the sum diverges to $+\infty$ almost surely. It follows that the ratio of posteriors diverges to $+\infty$ almost surely, which establishes the result. □

Proof of Proposition 3. Let σ^* be an optimal stationary strategy. Then,

$$\begin{aligned} w^*(\pi_t) &= r(\sigma^*(\mu_t),\mu_t) + \alpha E^{\tau(\pi_t,\sigma^*(\pi_t))}\left[w^*(\pi_{t+1})\right] \\ &\leq (1-\alpha)w^{\mathcal{M}}(\pi_t) + \alpha E^{\tau(\pi_t,\sigma^*(\pi_t))}\left[w^*(\pi_{t+1})\right]. \end{aligned}$$

Subtracting $\alpha w^*(\pi_t)$ from both sides and rearranging, we get:

$$w^*(\pi_t) - w^{\mathcal{M}}(\pi_t) \leq \frac{\alpha}{1-\alpha}\left[E^{\tau(\pi_t,\sigma^*(\pi_t))}\left[w^*(\pi_{t+1})\right] - w^*(\pi_t)\right].$$

Moreover, by definition, $w^*(\pi_t) - w^{\mathcal{M}}(\pi_t) \geq 0$. Hence,

$$E^{\tau(\pi_t,\sigma^*(\pi_t))}\left[w^*(\pi_{t+1})\right] - w^*(\pi_t) \geq \frac{1-\alpha}{\alpha}\left[w^*(\pi_t) - w^{\mathcal{M}}(\pi_t)\right] \geq 0.$$

It follows that the process $\{w^*(\pi_t)\}_{t\in\mathbb{N}}$ is a uniformly bounded sub-martingale, so there exists some bounded random variable $\widetilde{w}$ such that $w^*(\pi_t) \xrightarrow{a.s.} \widetilde{w}$, $w^*(\pi_t) \overset{a.s.}{\leq} E^{\tau(\pi_t,\sigma^*(\pi_t))}[\widetilde{w}|h_t]$ and $E^{\tau(\pi_t,\sigma^*(\pi_t))}[\widetilde{w}|h_t] \xrightarrow{a.s.} \widetilde{w}$. Therefore, $w^*(\pi_t) - w^{\mathcal{M}}(\pi_t)$ is almost-surely bounded above by:

$$\frac{\alpha}{1-\alpha}\left[E\left\{E^{\tau(\pi_{t+1},\sigma^*(\pi_{t+1}))}[\widetilde{w}|h_{t+1}]\middle| h_t\right\} - w^*(\pi_t)\right] = \frac{\alpha}{1-\alpha}\left[E^{\tau(\pi_t,\sigma^*(\pi_t))}[\widetilde{w}|h_t] - w^*(\pi_t)\right];$$

this upper bound converges to 0 almost surely, as desired. □

Proof of Proposition 4. Fix $\pi \in \mathcal{D}$; by Lemma 3, the value function of the proposed strategy satisfies $r^{\mathcal{T}} \geq w^*_\alpha(\pi) \geq \left(1-\alpha^{2^{N+1}}\right)(-B) + \alpha^{2^{N+1}}\left[r\left(a^{\mathcal{M}}(\mu_{2^N}),\mu^{\mathcal{T}}\right) - \frac{1}{N}\right]$ for each $N \in \mathbb{N}$, where $B > 0$ is a bound on ρ and $a^{\mathcal{M}}(\mu_\pi)$ is the myopic choice for μ_π. Now, we have $r^{\mathcal{T}} \geq \limsup_{\alpha\to 1} w^*_\alpha(\pi) \geq \liminf_{\alpha\to 1} w^*_\alpha(\pi) \geq r\left(a^{\mathcal{M}}(\mu_{2^N}),\mu^{\mathcal{T}}\right) - \frac{1}{N}$ for all $N \in \mathbb{N}$. Again by Lemma 3, we have that $\left|r\left(a^{\mathcal{M}}(\mu_{2^N}),\mu^{\mathcal{T}}\right) - r^{\mathcal{T}}\right|$ converges to 0 almost surely. Hence, $r^{\mathcal{T}} \geq \limsup_{\alpha\to 1} w^*_\alpha(\pi) \geq \liminf_{\alpha\to 1} w^*_\alpha(\pi) \geq r^{\mathcal{T}}$. □

[27] If β is an interior point in an M-dimensional simplex for $M \in \mathbb{N}$, the function $y \mapsto \sum_{m=1}^{M}\beta_m \ln(y_m)$ (defined on the interior of the simplex) is uniquely maximized at β.

Proof of Proposition 5. If the discount factor is $\alpha \in (0,1)$, the optimal-value function is bounded by $\frac{B}{1-\alpha}$, where $B \in \mathbb{R}$ is a bound on ρ. For any strategy σ and given any $\pi \in \mathcal{D}$,

$$E^{\sigma,\pi}\left(\sum_{t=0}^{\infty}\alpha^t\rho(a_t,v_t)\right) \leq \min\left\{\frac{B}{1-\alpha}, w^{\mathcal{M}}(\pi)(1-\alpha)+\frac{\alpha B}{1-\alpha}\right\}.$$

Thus, $w^*(\pi) \leq w^{\mathcal{M}}(\pi)(1-\alpha)+\frac{\alpha B}{1-\alpha}$. We also have:

$$w^*(\pi) \geq w^{\sigma^{\mathcal{M}}}(\pi) = w^{\mathcal{M}}(\pi)(1-\alpha)+\alpha E^{\tau\left(\pi_t,\sigma^{\mathcal{M}}(\pi_t)\right)}\left[w^{\sigma^{\mathcal{M}}}(\pi')\right] \geq w^{\mathcal{M}}(\pi)(1-\alpha)-\frac{\alpha K}{1-\alpha}.$$

Hence, $w^*(\pi)-w^{\mathcal{M}}(\pi) \leq \alpha w^{\mathcal{M}}(\pi)+\frac{2\alpha B}{1-\alpha} \leq \frac{3\alpha B}{1-\alpha}$. Let $\alpha_\epsilon := \frac{\epsilon}{3B+\epsilon}$; clearly, $\alpha_\epsilon \in (0,1)$. For $\alpha \in (0,\alpha_\epsilon)$, $\frac{3\alpha B}{1-\alpha} < \frac{3\alpha_\epsilon B}{1-\alpha_\epsilon} = \epsilon$ and so $w^{\sigma^{\mathcal{M}}}(\pi) \geq w^*(\pi)-\frac{3\alpha B}{1-\alpha} > w^*(\pi)-\epsilon$. As the choice of ϵ and α_ϵ is independent of π, it follows that $\sigma^{\mathcal{M}}$ is ϵ-optimal. □

Claim 1. *Let $X := \{x, x'\}$ and assume the decision maker has accurate data, so that $v_T^{IC}(x;a) = r\left(a,\mu^{\mathcal{T}}\right) - r\left(a\setminus\{x\},\mu^{\mathcal{T}}\right)$. Then, the incremental-contribution heuristic picks out $a^{\mu^{\mathcal{T}}}$.*

Proof. If X is objectively optimal, we have $r\left(\{x,x'\},\mu^{\mathcal{T}}\right) > r\left(\{x\},\mu^{\mathcal{T}}\right)$, $r\left(\{x,x'\},\mu^{\mathcal{T}}\right) > r\left(\{x'\},\mu^{\mathcal{T}}\right)$; hence, $v_T^{IC}(x;X), v_T^{IC}(x';X) > 0$: the heuristic recommends keeping both tools. If $\emptyset$ is objectively optimal, $0 = r\left(\emptyset,\mu^{\mathcal{T}}\right) > r\left(\{x\},\mu^{\mathcal{T}}\right)$ and $0 = r\left(\emptyset,\mu^{\mathcal{T}}\right) > r\left(\{x'\},\mu^{\mathcal{T}}\right)$; it follows that $v_T^{IC}(x;X) \leq v_T^{IC}(x;\{x\}) = r\left(\{x\},\mu^{\mathcal{T}}\right) < 0$, and similarly for x': the heuristic recommends discarding both tools. Finally, consider the case with $\{x\}$ the objectively-optimal bundle; the argument for $\{x'\}$ is identical. We have $r\left(\{x\},\mu^{\mathcal{T}}\right) > r\left(\{x'\},\mu^{\mathcal{T}}\right)$, $r\left(\{x\},\mu^{\mathcal{T}}\right) > r\left(X,\mu^{\mathcal{T}}\right)$, and $r\left(\{x\},\mu^{\mathcal{T}}\right) > 0$. Thus, $v_T^{IC}(x;X) > 0 > v_T^{IC}(x';X)$; x' is discarded first, while x is kept: $v_T^{IC}(x;\{x\}) = r\left(\{x\},\mu^{\mathcal{T}}\right) > 0$. □

Proof of Proposition 6. Assume that $a \subseteq X$ is internally stable relative to μ but that there is some set $b \subseteq a$ with a more profitable subset $c \subseteq b$: $r(c,\mu) > r(b,\mu)$. Take the set $d := (a\setminus b)\cup c$. Now, by Proposition 1, the function $r(\cdot,\mu)$ is submodular. Since $a = b\cup d$ and $c = b\cap d$, it follows that $r(a,\mu)+r(c,\mu) = r(b\cup d,\mu)+r(b\cap d,\mu) \leq r(b,\mu)+r(d,\mu)$, or $r(a,\mu)-r(d,\mu) \leq r(b,\mu)-r(c,\mu)$. Thus, $r(c,\mu) > r(b,\mu)$ implies that $r(d,\mu) > r(a,\mu)$, contradicting internal stability of a. □

For $a \in Z$ and $x \in a$, let:

$$v_T^{IC}(x;a) = \frac{1}{T}\sum_{\tau\leq T}\left[W(\nu(a,v_t)) - W(\nu(a\setminus\{x\},v_t))\right],$$
$$\bar{v}^{IC}(x;a) := r\left(a,\mu^{\mathcal{T}}\right) - r\left(a\setminus\{x\},\mu^{\mathcal{T}}\right);$$

these are, respectively, the sample and population incremental contribution of tool $x \in a$ to bundle a.

Lemma 4. *For $x \in a' \subseteq a \in Z$, $\bar{v}^{IC}(x;a') \geq \bar{v}^{IC}(x;a)$.*

Proof. The result is an immediate consequence of Proposition 2. □

Lemma 5. *For $x \in a' \subseteq a \in Z$ and $T \in \mathbb{N}$, $v_T^{IC}(x;a') \geq v_T^{IC}(x;a)$.*

Proof. The result follows by submodularity of $W(\nu(\cdot, v))$ for any $v \in V$. □

Lemma 6. *Let $\{a_t\}_{t=1}^{+\infty}$ be the (random) outcome of implementing the incremental-contribution heuristic. Since choices are nested, we can define $a^* := \cap_{t=1}^{+\infty} a_t$. For each $x \in a^*$, we have $\lim_{T\to+\infty} v_T^{IC}(x;a^*) = \bar{v}^{IC}(x,a^*) + c_x$ almost surely.*

Proof. Follows by the Strong Law of Large Numbers. □

Lemma 7. *Bundle $a \in Z$ is internally stable if and only if $\bar{v}^{IC}(x,a) \geq 0$ for all $x \in a$.*

Proof. If a is internally stable, for any $x \in a$, it follows that $0 \leq r\left(a, \mu^{\mathcal{T}}\right) - r\left(a \setminus \{x\}, \mu^{\mathcal{T}}\right) = \bar{v}^{IC}(x,a)$. Conversely, assume there is some $x^0 \in a$ such that $\bar{v}^{IC}\left(x^0, a\right) < 0$. Then, the proper subset $a \setminus \{x^0\}$ of a yields a strictly higher immediate reward. □

Proof of Proposition 7. Under the incremental-contribution heuristic, choices are nested. Denote the limit set by a^*. By Lemma 6, $a^* = \left\{x \in X : \bar{v}^{IC}(x,a^*) \geq 0\right\}$. The conclusion follows by Lemma 7. □

Proof of Proposition 8. The proposition follows by the Strong Law of Large Numbers and a variation on the proof of Lemma 3. □

Proof of Proposition 9. The first part of the proposition follows from the fact that every bundle that is not internally stable has a proper subset with positive regret. Similarly, there is non-negative regret regarding the objectively-optimal bundle (with respect to the whole of X). The second part of the proposition follows from this observation. □

References

Afonja, B. (1972). The moments of the maximum of correlated normal and t-variates. *Journal of the Royal Statistical Society, Series B (Methodological)* 34(2):251–262.

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books, Inc.

Baumol, W. and Quandt, R. (1964). Rules of thumb and optimally imperfect decisions. *The American Economic Review* 54(2):23–46.

Bergemann, D. and Valimaki, J. (2001). Stationary multi-choice bandit problems. *Journal of Economic Dynamics and Control* 25:1585–1594.

Bertsekas, D. (2005). *Dynamic Programming and Optimal Control*, vol. 1. 3rd ed. Athena Scientific.

Blackwell, D. (1965). Discounted dynamic programming. *The Annals of Mathematical Statistics* 36(1):226–235.

Blume, L. and Easley, D. (1982). Learning to be rational. *Journal of Economic Theory* 26:340–351.

Bray, M. (1982). Learning, estimation, and the stability of rational expectations. *Journal of Economic Theory* 26:318–339.

Easley, D. and Rustichini, A. (2005). Optimal guessing: Choice in complex environments. *Journal of Economic Theory* 124:1–21.

Friedman, M. (1994). The methodology of positive economics. In: D. Hausman (ed.) *The Philosophy of Economics: An Antology*, 2nd ed., chap. 9. Cambridge University Press.

Fudenberg, D. and Kreps, D. (1993). Learning mixed equilibria. *Games and Economic Behavior* 5:320–367.

Fudenberg, D. and Levine, D. (1998). *The Theory of Learning in Games*. MIT Press.

Gittins, J. and Jones, D. (1974). A dynamic allocation index for the sequential design of experiments. In: *Progress in Statistics*, pp. 241–266. Amsterdam: North Holland.

Gul, F. and Stacchetti, E. (1999). Walrasian equilibrium with gross substitutes. *Walrasian Equilibrium with Gross Substitutes* 87:95–124.

Hart, S. (2005). Adaptive heuristics. *Econometrica* 73(5):1401–1430.

Hart, S. and Mas-Colell, A. (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica* 68(5):1127–1150.

Hart, S. and Mas-Colell, A. (2001a). A general class of adaptive strategies. *Journal of Economic Theory* 98:26–54.

Hart, S. and Mas-Colell, A. (2001b). A reinforcement procedure leading to correlated equilibrium. In: G. Debreu et al. (eds.) *Economics Essays: A Festschrift for Werner Hildenbrand*. Springer-Verlag.

Hart, S. and Mas-Colell, A. (2003a). Regret-based dynamics. *Games and Economic Behavior* 45:375–394.

Hart, S. and Mas-Colell, A. (2003b). Uncoupled dynamics do not lead to nash equilibrium. *The American Economic Review* 93:1830–1836.

Hart, S. and Mas-Colell, A. (2004). Stochastic uncoupled dynamics and nash equilibria. Mimeo, DP-371, Center for Rationality, the Hebrew University of Jerusalem.

Kalai, E. and Lehrer, E. (1993). Rational learning leads to nash equilibrium. *Econometrica* 61(5):1019–1045.

Kreps, D. (2013). *Microeconomic Foundations: I. Choice and Competitive Markets*. Princeton University Press.

Lettau, M. and Uhlig, H. (1999). Rules of thumb versus dynamic programming. *The American Economic Review* 89(1):148–174.

Lewis, M. (2003). *Moneyball: The Art of Winning an Unfair Game*. W. W. Norton and Company Inc.

Li Calzi, M. (1992). Playing games by similarities. Ph.D. thesis, Stanford Graduate School of Business.

Maitra, A. (1968). Discounted dynamic programming on compact metric spaces. *Sankhya: The Indian Journal of Statistics, Series A (1961-2000)* 30(2):211–216.

Milgrom, P. and Roberts, J. (1991). Adaptive and sophisticated learning in normal form games. *Games and Economic Behavior* 3:82–100.

Nelson, R. and Winter, S. (1982). *An Evolutionary Theory of Economic Change*. The Belknap Press of Harvard University Press.

Radner, R. (1975). Satisficing. *Journal of Mathematical Economics* 2:253–262.

Roth, A. and Erev, I. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review* 88(4):848–881.

Rust, J. (1987). Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica* 55(5):999–1033.

Rustichini, A. (1999). Optimal properties of stimulus-response learning models. *Games and Economic Behavior* 29:244–273.

Sargent, T. and Marcet, A. (1989). Convergence of least squares learning mechanisms in self-referential linear stochastic models. *Journal of Economic Theory* 48:337–368.

Simon, H. (1959). Theories of decision-making in economics and behavioral science. *The American Economic Review* 49(3):253–283.

Simon, H. (1979). Rational decision making in business organizations. *The American Economic Review* 69(4):493–513.

Simon, H. (1982a). *Models of Bounded Rationality*, vol. 1: Economic Analysis and Public Policy. The MIT Press.

Simon, H. (1982b). *Models of Bounded Rationality*, vol. 2: Behavioral Economics and Business Organization. The MIT Press.

Simon, H. (1997). *Models of Bounded Rationality*, vol. 3: Empirically Grounded Economic Reason. The MIT Press.

Topkis, D. (1998). *Supermodularity and Complementarity*. Princeton University Press.

Weber, R. (1992). On the gittins index for multiarmed bandits. *The Annals of Applied Probability* 2(4).