Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# Robust Multiarmed Bandit Problems

Michael Jong Kim, Andrew E.B. Lim

To cite this article:
Michael Jong Kim, Andrew E.B. Lim (2016) Robust Multiarmed Bandit Problems. Management Science 62(1):264-285. http://dx.doi.org/10.1287/mnsc.2015.2153

MANAGEMENT SCIENCE
Vol. 62, No. 1, January 2016, pp. 264–285
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2015.2153


# Robust Multiarmed Bandit Problems

Michael Jong Kim
Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Ontario M5S 3G8, Canada, mikekim@mie.utoronto.ca

Andrew E.B. Lim
Department of Decision Sciences and Department of Finance, NUS Business School, National University of Singapore, Singapore 119245, andrewlim@nus.edu.sg

The multiarmed bandit problem is a popular framework for studying the exploration versus exploitation trade-off. Recent applications include dynamic assortment design, Internet advertising, dynamic pricing, and the control of queues. The standard mathematical formulation for a bandit problem makes the strong assumption that the decision maker has a full characterization of the joint distribution of the rewards, and that "arms" under this distribution are independent. These assumptions are not satisfied in many applications, and the out-of-sample performance of policies that optimize a misspecified model can be poor. Motivated by these concerns, we formulate a robust bandit problem in which a decision maker accounts for distrust in the nominal model by solving a worst-case problem against an adversary ("nature") who has the ability to alter the underlying reward distribution and does so to minimize the decision maker's expected total profit. Structural properties of the optimal worst-case policy are characterized by using the robust Bellman (dynamic programming) equation, and arms are shown to be no longer independent under nature's worst-case response. One implication of this is that index policies are not optimal for the robust problem, and we propose, as an alternative, a robust version of the Gittins index. Performance bounds for the robust Gittins index are derived by using structural properties of the value function together with ideas from stochastic dynamic programming duality. We also investigate the performance of the robust Gittins index policy when applied to a Bayesian webpage design problem. In the presence of model misspecification, numerical experiments show that the robust Gittins index policy not only outperforms the classical Gittins index policy, but also substantially reduces the variability in the out-of-sample performance.

*Keywords*: bandit problems; robust control; model uncertainty; relative entropy; games against nature
*History*: Received February 28, 2013; accepted November 30, 2014, by Dimitris Bertsimas, optimization.
Published online in *Articles in Advance* August 5, 2015.

## 1. Introduction

The multiarmed bandit problem is a popular framework for studying the exploration versus exploitation trade-off. That is, how should a decision maker choose between actions that yield immediate rewards and learning-type actions that provide information about less understood alternatives and whose benefits may only come later? The trade-off originates from the well-known gambler–slot machine scenario,[1] and it has appeared in a number of applications in the operations research/management science (OR/MS) literature including dynamic assortment design (e.g., Caro and Gallien 2007, Rusmevichientong et al. 2010), Internet advertising (e.g., Babaioff et al. 2014; Scott 2010, 2013; White 2013), dynamic pricing with learning (e.g., Mersereau et al. 2009), and control of queuing systems (e.g., Niño-Mora 2012).

In the standard Markovian setting, the multiarmed bandit problem makes two central assumptions: that the decision maker has a full characterization of the dynamics for each "arm" and that "arms" are independent. These assumptions are made for the purpose of analytical tractability and lead to nice structural properties such as the optimality of an index policy (see, e.g., Bertsekas 1995, Bertsimas and Nino-Mora 1996, Gittins 1979, Tsitsiklis 1986, Whittle 1980).

The problem with these assumptions is that they not satisfied in many applications. For example, the arm dynamics may contain complex nonstationarities or depend on (possibly unknown) factors in a way that is difficult to identify and model. Arms may also be dependent because transitions of one arm may affect the dynamics of the others, but the dependence structure may be hard to specify, while calibration errors arise when there is limited or poor quality data. The following examples show that these issues arise naturally, and experiments in §6 on a Bayesian webpage design problem show that the

[1] In this scenario, a gambler tries to maximize total rewards by sequentially pulling arms from a row of slot machines (or bandits) whose reward distributions are unknown.

model misspecification can have a substantial impact on out-of-sample performance. The goal of this paper is to formulate a model of the bandit problem that accounts for model misspecification, and to determine scheduling policies that have good out-of-sample performance even if they are obtained on the basis of a misspecified nominal model.

*Dynamic Pricing and Assortment Applications.* It is known that dynamic pricing and assortment problems can be formulated as multiarmed bandit problems (see e.g., Mersereau et al. 2009, Caro and Gallien 2007, Rusmevichientong et al. 2010). The classical bandit model assumes that demand distribution is known and depends only on the current price/assortment (i.e., arm). In reality, the distribution of revenues associated with each price/assortment is misspecified, and the order in which price/assortment offerings are displayed will affect behavior at future prices/assortments in a complex manner that is difficult to model. We are interested in finding pricing/assortment policies that perform well in spite of such misspecification.

*Bandit Problems with Bayesian Learning.* A classical application for the bandit model is one where the reward for each arm is generated from a parameterized family of distributions, with parameters that are constant but unknown to the decision maker. The decision maker has a prior on the parameters for each arm, and the goal is to find a policy for selecting between arms that maximizes the expected discounted reward over an infinite horizon. Here, the state of each arm is the posterior on the unknown parameters, which is updated whenever a reward is observed, whereas priors and likelihoods are specified by the decision maker/modeler (and commonly chosen for analytical tractability as a conjugate pair). In many applications, the likelihood is not well understood or has been chosen because it is tractable and is likely to be misspecified. We are interested in finding scheduling policies that are robust to errors in the likelihood function. One example of this is a search engine's problem of optimizing over alternative webpage configurations/"arms" (layouts, designs, fonts, etc.) so as to maximize the number of customer "conversions." (Recent articles by Scott 2010, 2013 describe how Bayesian bandits are used to address this problem at Google.) With our limited understanding of how preferences are formed/the flow of ideas between users of the Internet, etc., relatively little is understood about the way in which conversions are generated and likelihood misspecification may be significant. We study such an example in §6.

### 1.1. Relevant Literature

Broadly speaking, there are two streams of research on bandit problems, one based on dynamic programming (e.g., Whittle 1980, Bertsekas 1995, Bertsimas and Nino-Mora 1996), which is the framework this paper adopts, and another based on the notion of regret (of which Lai and Robbins (1985) and Auer et al. (1995, 2002a, b) are important representatives).

Regret analysis for the bandit problem was initiated in Lai and Robbins (1985) (building on Robbins 1952), wherein optimal bounds on the growth rate of regret relative to the reward stream of the optimal arm are shown to be logarithmic, and policies that achieve them are obtained. The results in Lai and Robbins (1985) apply to specific families of reward distributions indexed by a single real parameter. Subsequent papers by Auer et al. (1995, 2002b) and specifically Auer et al. (2002a) (see also Cesa-Bianchi and Lugosi 2006) strengthen these results, showing that the optimal regret in Lai and Robbins (1985) can be achieved uniformly over time for arbitrary reward distributions that are independent across arms. Easily implementable scheduling policies achieving optimal regret are also provided.

Our paper differs from Lai and Robbins (1985), Auer et al. (1995, 2002a, b), and Cesa-Bianchi and Lugosi (2006) in a number of ways. First, our objective function is different. We focus on optimizing worst-case expected discounted performance rather than the growth rate of regret. Second, our robust model includes two features, a nominal model and an adversary who has the power to change the reward distribution from the one specified in the nominal model. One advantage of this setup is that it enables a decision maker to approximately model the system with a nominal model, which provides the opportunity to include information about problem structure (even if it is only approximately correct), and to account for errors in this approximate model through the adversary (nature). In contrast, although the main results in Auer et al. (1995, 2002a, b) and Cesa-Bianchi and Lugosi (2006) apply to any reward distribution and hence are model free (the results do assume independence across arms, however), the disadvantage is that there is little opportunity to include even approximately correct information, which may become important when the dimensionality of the problem increases (e.g., adaptations of the results in Auer et al. 2002a to the continuum bandit problem, as in Kleinberg 2004).

Research on dynamic optimization with model uncertainty has mainly focused on characterizing the optimal robust policy. As pointed out by Chan and Farias (2009), Brown et al. (2010) and Rogers (2007), solving dynamic optimization problems to optimality is difficult and one must often resort to heuristics. These papers develop methods for bounding the performance of heuristic policies relative to the optimal for non-worst-case stochastic control problems. (See

also Haugh and Lim 2012 and Ye and Zhou 2015 for related ideas.) In this regard, one contribution of this paper is that it develops systematic performance evaluation methods for worst-case robust control problems, which we apply to performance evaluation of the robust Gittins index.

We know of one other paper on robust bandit problems, which is the recent paper by Caro and Gupta (2015). We briefly highlight the differences. First, our paper adopts a different approach for expressing model uncertainty; Caro and Gupta (2015) adopts the constraint approach whereas we adopt the penalty approach. The differences between these approaches are discussed in §3.1. Second, Caro and Gupta (2015) focus on numerical algorithms for computing proposed index policies in the constraint setting. On the other hand, we spend little time on this issue because the policy we propose can be computed by using obvious extensions of algorithms for a classical bandit problem. Third, we provide theoretical upper and lower bounds on the performance of our proposed policy relative to the optimal worst-case objective value. To our knowledge, this paper is the first to provide such bounds in a worst-case setting. Despite these differences, we believe that model uncertainty is an important concern in many applications of the bandit model, and that both papers provide complementary contributions to the literature on robust stochastic optimization.

Finally, there is a relatively small, but growing, literature on bandit problems with correlated arms, i.e., where the dependency between arms is directly modeled in the nominal setting (e.g., Agrawal 1995, Brown and Smith 2013, Kleinberg 2004, Pandey et al. 2007, Ryzhov et al. 2012). When the correlation structure is known (e.g., rewards are known to be correlated according to a multivariate normal distribution), modeling this dependency and designing policies that account for it may be a reasonable approach to take. On the other hand, there are many situations where the dependence structure cannot be teased out by using historical data alone (we provide a simple illustration of this point in §6 in which we consider vector autoregressive rewards with censored observations). Furthermore, even if the correlation structure is fully known, it may still be reasonable to design heuristic policies under the (incorrect) assumption of independent arms (i.e., by only considering the marginal distributions) for the sake of tractability. This approach was taken, e.g., in the recent work of Brown and Smith (2013). In such cases, model misspecification is either unavoidable or (deliberately) introduced, and the results in this paper may be useful.

### 1.2. Summary of Contributions

In this paper, we introduce a robust bandit problem in which a decision maker accounts for his distrust in the nominal model by solving a worst-case problem against an adversary who modifies the reward distribution to maximally damage the decision maker's expected profit. One novelty of our model is that different arms may have different levels of model ambiguity, which we capture by introducing arm-dependent relative entropy terms. This allows for the situation in which the decision maker has high confidence in certain parts of the model but less confidence in others. A key motivation for studying this problem is our interest in determining scheduling policies for bandit problems that perform well when there is model uncertainty.

Structural properties of the robust bandit problem are studied by using the robust dynamic programming equations. We characterize the optimal response of nature to any scheduling policy and show, in contrast to the classical bandit problem, that arms are no longer independent under nature's optimal response. One important implication is that (easily computable) index policies, which are optimal for the classical bandit problem, are no longer optimal in the robust case, and that the substantially more demanding task of solving the dynamic programming equations must be carried out if the optimal robust policy is desired.

Motivated by these results, we propose a robust version of the Gittins index as a policy for selecting arms. Performance bounds for this policy are generated by adapting ideas from stochastic dynamic programming duality and information relaxation. Our performance analysis provides, to our knowledge, the only systematic approach to generating performance bounds for robust dynamic stochastic control problems. Numerical experiments show that the performance of the robust Gittins index can be quite close to that of the optimal robust scheduling policy (within 92% for our example).

Finally, we study the performance of the robust Gittins index policy when applied to a Bayesian webpage design problem. In this application, we develop a new data-driven method based on $k$-fold cross validation for calibrating ambiguity parameters that model the decision maker's level of trust in the nominal assumptions. Our numerical experiments show that the robust Gittins policy can outperform the classical Gittins index and substantially reduce variability in the out-of-sample-performance when the underlying model is misspecified.

### 1.3. Outline of Paper

We summarize key results for the classical bandit problem in §2. We formulate a robust version of the standard multiarmed bandit problem in §3. A robust version of the well-known Gittins index policy is proposed in §4, and performance bounds for this policy

are provided in §5. In §6, we investigate the performance of the robust Gittins index policy when applied to Bayesian webpage design problems.

## 2. A Nominal Multiarmed Bandit Problem

The classical bandit problem is defined in terms of $p$ independent risky projects and a retirement option. Retirement is defined by a payoff $M$, and each risky project $i \in \{1, \dots, p\}$ is defined by a triple $(\mathscr{X}, r^i(\cdot), \rho^i)$ where $\mathscr{X}$ is the state space for project $i$, $r^i(x^i)$ $(x^i \in \mathscr{X})$ is a state-dependent reward function, and $\rho^i(x^i, j)$ $(j \in \mathscr{X})$ is a transition probability function. Both $r^i(x^i)$ and $\rho^i(x^i, \cdot)$ are independent of the states of other projects $x^j$ $(j \neq i)$. We denote the state of project $i$ at stage $n$ by $X_n^i$ and the vector of states for each of the $p$ projects by $X_n \triangleq (X_n^1, \dots, X_n^p)$. We assume for simplicity that the state space $\mathscr{X}$ for each project is finite, though all our results extend to general state spaces. All rewards are discounted to time 0 with discount factor $\alpha \in (0, 1)$. Reward functions are assumed to be uniformly bounded.

Assumption 1. *There exists a constant $B > 0$, such that for each $i = 1, \dots, p$, reward functions $r^i\colon S \to \mathbb{R}$ satisfy*

$$\sup_{\omega \in S} |r^i(\omega)| \leq B. \tag{1}$$

At each stage $n \in \{0, 1, 2, \dots\}$ before retirement, the decision maker, with knowledge of the state of all the projects $X_n = (x^1, \dots, x^p)$, chooses between permanent retirement and one of the $p$ risky projects. If retirement is chosen, the decision maker receives a reward of $M$ and there are no future choices or rewards. If project $i \in \{1, 2, \dots, p\}$ is chosen, then reward $r^i(x^i)$ is received and the state of project $i$ changes from $X_n^i = x^i$ to a new state $X_{n+1}^i \in \mathscr{X}$ according to transition probabilities

$$\mathbb{P}(X_{n+1}^i = j \mid X_n^i = x^i) = \rho^i(x^i, j), \quad j \in \mathscr{X}$$

that are independent of the states of the other projects. The states of nonchosen projects do not change. The decision maker's objective is to maximize expected discounted reward

$$J(x_0) \triangleq \max_{\pi \in \Pi} \mathbb{E}_{\mathbb{P}}\left[\sum_{n \leq \tau} \alpha^n c(X_n, \pi_n(X_n)) \,\middle|\, X_0 = x_0\right]$$

over the class of admissible scheduling policies

$$\Pi = \{(\pi_0, \pi_1, \pi_2, \dots) \mid \pi_n\colon \mathscr{X}^p \to \{1, \dots, p\} \cup \{\text{retire}\}\},$$

where $\pi_n(x) \in \{1, \dots, p\} \cup \{\text{retire}\}$ is the decision maker's choice at stage $n$,

$$\tau \triangleq \inf\{n\colon \pi_n(X_n) = \text{retire}\}$$

is the retirement time, and the reward per stage is

$$c(x, i) = \begin{cases} r^i(x^i) & i \in \{1, \dots, p\}, \\ M & i = \text{retire}. \end{cases}$$

It can be shown that $J(x)$ is the unique solution to the dynamic programming equation

$$J(x) = \max\left\{M, \max_{i=1,\dots,p} R^i(x, J)\right\}, \tag{2}$$

where

$$R^i(x, J) \triangleq r^i(x^i) + \alpha \mathbb{E}_{\rho^i(x^i, \cdot)}[J(x^1, \dots, x^{i-1}, X^i, x^{i+1}, \dots, x^p)], \tag{3}$$

and the maximizer of the right-hand side of (2)

$$\pi_n^*(x) \triangleq \arg\max\left\{M, \max_{i=1,\dots,p} R^i(x, J)\right\}$$

characterizes the optimal project choice at stage $n$ as a function of the state of all projects $x = (x^1, \dots, x^p)$. In particular, it is optimal to choose project $i$ if $J(x) = R^i(x, J)$ and to permanently retire if $J(x) = M$.

To make notation more compact, when it is clear, we write for short $\rho^i \equiv \rho^i(x^i, \cdot)$ and

$$(x^{i-}, X^i) \equiv (x^1, \dots, x^{i-1}, X^i, x^{i+1}, \dots, x^p),$$

so that Equation (3), for example, can be written as

$$R^i(x, J) = r^i(x^i) + \alpha \mathbb{E}_{\rho^i}[J(x^{i-}, X^i)].$$

The main result for the classical bandit problem is that the optimal policy follows an index policy.

Theorem 1 (Gittins 1979, Whittle 1980). *Suppose, at an arbitrary stage, the current state vector is $x = (x^1, \dots, x^p)$. There exist index functions $g^i\colon S \to \mathbb{R}$, $i = 1, \dots, p$, such that project $i$ is selected if*

$$g^i(x^i) = \max_{j=1,\dots,p} g^j(x^j) > M.$$

*If $\max_j g^j(x^j) \leq M$, it is optimal to permanently retire from all projects.*

Using dynamic programming arguments, Whittle (1980) showed that the optimal index functions $g^i$, known as the "Gittins indices," are of the form

$$g^i(x^i) = \inf\{m\colon J^i(x^i; m) = m\}, \tag{4}$$

where $J^i(x^i; m)$ is the value function associated with the multiarmed bandit problem with a single project $i$ with retirement payment $M = m$.

Optimality of an index policy rests on several assumptions, namely, that the decision maker can fully characterize the Markovian dynamics and that projects are independent. These assumptions are typically not satisfied in practice. The objectives of this paper are to come up with policies that work well when they are violated and to develop methods of quantifying their performance.

## 3. A Robust Multiarmed Bandit Formulation

We now consider a robust version of the multiarmed bandit problem. In §3.1, we begin by discussing alternative approaches for modeling ambiguity and justifying our adoption of the penalty approach in this paper. We then formulate the robust problem as a zero-sum stochastic dynamic game in §3.2 where we introduce "nature" and its role in choosing transition probabilities that differ from the nominal, and relative entropy as a penalty on nature's choice that captures our confidence in the nominal. The associated dynamic programming equation is introduced in §3.3 and used in §3.4 to characterize nature's optimal response to the decision maker's optimal policy. For simplicity, we formulate our robust model in the setting of a finite state space. All our results extend to more general state spaces using the framework of González-Trejo et al. (2003).

### 3.1. Modeling of Ambiguity

We discuss two approaches to modeling ambiguity, the "penalty approach" and the "constraint approach." The penalty approach, which we are using, has been widely adopted in the robust control literature (e.g., Jacobson 1973; Whittle 1981, 1990, 1991; Dai Pra et al. 1996; Petersen et al. 2000), and has also been used in economics and operations research applications (e.g., Hansen and Sargent 2005, 2007, 2008; Lim and Shanthikumar 2007; Lim et al. 2012; Li and Kwon 2013). The constraint approach has been widely applied to single-stage optimization problems with parameter or moment uncertainty (e.g., Ben-Tal and Nemirovski 1998, 1999, 2000; Bertsimas and Sim 2004; El Ghaoui and Lebret 1997), but has also been extended to dynamic problems with uncertainty in the model of state dynamics (e.g., Iyengar 2005, Lim et al. 2011, Nilim and El Ghaoui 2005 and Wiesemann et al. 2013). Applications to the modeling of ambiguity-averse economic agents can be found in Epstein and Schneider (2003, 2007). A survey of stochastic optimization with model uncertainty can be found in Lim et al. (2006).

At a high level, both approaches specify a set of alternative models centered around some chosen nominal model. An adversary (nature) chooses a model from the set of alternatives to maximally damage the performance of the decision maker's policy, and the decision maker tries to optimize the worst case. Confidence in the nominal model is expressed by restricting the size of the set of alternatives. The difference between these approaches is the way in which these restrictions are imposed.

In the constraint approach, the set of alternative models is represented by a hard constraint, and confidence in the nominal model is represented by the size of this uncertainty set. For robust dynamic optimization problems, Epstein and Schneider (2003, 2007), Iyengar (2005), Nilim and El Ghaoui (2005) show that additional "rectangularity" conditions need to be satisfied by the set of alternative models in order for optimal robust policies to be dynamically consistent and obtainable by solving the associated robust dynamic programming equations.

The penalty approach expresses the decision maker's degree of confidence in the nominal model by penalizing nature's deviations from the nominal. It is a soft constraint in that the penalty appears in the objective function. A large positive penalty coefficient is chosen if the decision maker is confident about the nominal model and wants to make it costly for nature to deviate too far, and a penalty close to zero corresponds to little confidence in the nominal model and forces the decision maker to plan against a highly deviant worst case. Alternatively, the objective with penalty can be regarded as a Lagrangian relaxation of the decision maker's nominal model (see Hansen and Sargent 2008, Petersen et al. 2000, and Lim and Shanthikumar 2007 for discussion along these lines). As in the constraint approach, many penalties can be imposed (e.g., Kolmogorov distance, Hellinger distance, and relative entropy), but the resulting worst-case problem cannot be solved by using dynamic programming unless the penalty is time separable. This is one reason that relative entropy is so widely adopted in the robust control literature.

There are other advantages of the penalty approach with relative entropy. One of these is that nature's worst-case response can be computed explicitly, which is typically not the case in standard constraint models where nature's response can only be solved numerically (see, e.g., Caro and Gupta 2015 for a constraint approach applied to the robust bandit problem, and Iyengar 2005 and Nilim and El Ghaoui 2005 for the generic robust dynamic programming model in the same framework). Classifying nature's response is valuable because it gives us insight into the structure of the worst-case distribution. In this paper, this is used to characterize the structural properties of the "robust Gittins index," and to bound its worst-case performance relative to that of the optimal robust policy. More generally, entropy penalties often lead to problems with appealing structure and the associated advantages for computation and analysis.

Another advantage of our approach is that it includes robust dynamic optimization problems with learning (specifically, Bayesian bandits with ambiguity in the prior and likelihood (see §2 and §6)) as a special case. In particular, since nature's worst-case response can be characterized, the robust dynamic programming equations reduce to a recursive equation having the same state space as the nominal

Bayesian bandit problem. On the other hand, the state space for the constraint approach, even with rectangularity, is an uncountable set of posterior distributions that is generated by applying Bayes' rule to every prior–likelihood pair for every new data point (e.g., Epstein and Schneider 2003). The resulting robust dynamic programming equations are unsolvable. A Bayesian bandit approach to webpage design problems is studied in §6.

### 3.2. Robust Bandit Model

As in the classical problem, each stage involves a choice between $p$ risky projects and a permanent retirement option. Retirement leads to a one-off payment of $M$ and there are no future choices or payments. Each risky project is described by a four-tuple $(\mathscr{X}, r^i(\cdot), \rho^i, \theta^i)$ where $\mathscr{X}$ is the state space for project $i$ (assumed for simplicity to be finite), $r^i(x^i)$ $(x^i \in \mathscr{X})$ is the reward function, $\rho^i(x^i, \cdot)$ $(x^i \in \mathscr{X})$ is the nominal transition probability distribution, and $\theta^i$ is the penalty parameter for ambiguity about the model $\rho^i(x^i, \cdot)$. The nominal reward and transition distribution depend only on the state $x^i$ of project $i$. All rewards are discounted to time zero with discount factor $\alpha \in (0, 1)$. The reward $r^i(x^i)$ is assumed to be bounded.

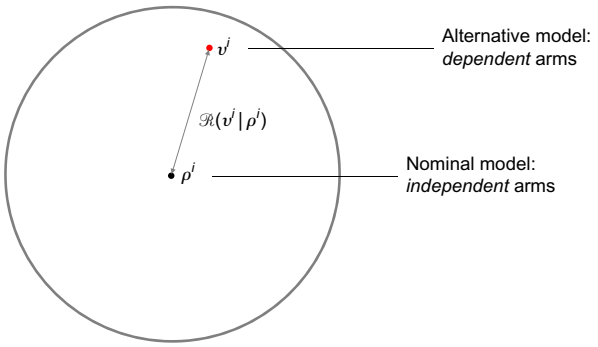
In the classical bandit problem, the choice of project $i$ when the state is $X_n = (x^1, \ldots, x^p)$ gives a reward $r^i(x^i)$, and the state of this project transitions from $X_n^i = x^i$ to $X_{n+1}^i$ according to the distribution $\rho^i(x^i, \cdot)$. In the robust problem, we distrust our model $\rho^i(x^i, \cdot)$ for state transitions and allow for the possibility that $X_{n+1}^i$ has another distribution $v^i(\cdot)$. We account for uncertainty about our specification of $\rho^i(x^i, \cdot)$ by introducing a second agent, "nature," who chooses the transition probability $v^i(\cdot)$ in an adversarial manner. We model our confidence in the nominal model by limiting the amount by which nature's choice $v^i(\cdot)$ can deviate from $\rho^i(x^i, \cdot)$ using the notion of relative entropy.

*Relative Entropy.* Our confidence in the nominal transition probability is captured through a cost for nature for deviating from the nominal $\rho^i(x^i, \cdot)$ when choosing $v^i(\cdot)$. We measure distance between nature's choice and the nominal using relative entropy $\mathscr{R}(v^i \mid \rho^i)$, which is defined as follows (see Figure 1).

Definition 1 (Relative Entropy). Let $\eta$ and $\kappa$ be probability distributions on the discrete set $\mathscr{X}$ (here, $\eta(\cdot) \equiv v^i(\cdot)$ and $\kappa(\cdot) \equiv \rho^i(x^i, \cdot)$) such that $\eta$ is absolutely continuous with respect to $\kappa$ (namely, for any state $j \in \mathscr{X}$, $\eta(j) = 0$ whenever $\kappa(j) = 0$). Relative entropy of $\eta$ with respect to $\kappa$ is given by

$$\mathscr{R}(\eta \mid \kappa) = \sum_{j \in \mathscr{X}} \eta(j) \ln \frac{\eta(j)}{\kappa(j)}. \tag{5}$$

**Figure 1** **(Color online) Distance Between Nominal and Alternative Transition Probabilities**




*Note.* To model the decision maker's fear that the nominal transition probabilities $\rho^i$ are misspecified, an adversarial agent sequentially selects alternative transition probabilities. The distance between the nominal and alternative is measured by relative entropy $\mathscr{R}(v^i \mid \rho^i)$.

If absolute continuity is not satisfied (i.e., if there is $j \in \mathscr{X}$ such that $\eta(j) > 0$ and $\kappa(j) = 0$), then $\mathscr{R}(\eta \mid \kappa) = +\infty$.

Relative entropy is convex in nature's choice $v^i(\cdot)$, nonnegative, and equals zero if and only if nature's choice equals the nominal $v^i(\cdot) = \rho^i(x^i, \cdot)$.

We model our confidence in the nominal model using a penalty term $\theta^i \mathscr{R}(v^i \mid \rho^i)$ in the objective that penalizes nature for deviating from the nominal. Specifically, a large value of $\theta^i > 0$ makes it expensive for nature to choose an alternative that is far from the nominal and corresponds to a high degree of confidence in $\rho^i$, whereas a small $\theta^i$ corresponds to less confidence in the nominal model and forces the decision maker to play against a more powerful adversary since it is less costly for nature to deviate. The choice of $\theta^i$ affects the policy that is chosen by the decision maker, and a data-driven approach to choosing these parameters will be provided in the context of a Bayesian webpage design application in §6.

*Zero-Sum Stochastic Dynamic Game.* We formulate our robust bandit problem as a zero-sum stochastic dynamic game. As a start, consider first what happens at stage $n$. If the decision maker has already retired, then there is nothing to do; there is no decision, reward, or transition. Suppose now that the decision maker has not retired and that the state of the $p$ projects $X_n = (x^1, \ldots, x^p)$ is known to both the decision maker and nature. In this case, the decision maker goes first and chooses a project $\{1, \ldots, p\}$ or retirement. If retirement is chosen (i.e., $\pi_n(X_n) =$ retire), then a payment of $M$ is received, and the game ends. Suppose however, that a risky project $i \in \{1, \ldots, p\}$ is chosen (i.e., $\pi_n(X_n) = i$). This choice is communicated to nature, who responds by choosing the probability distribution $v^i(\cdot)$ at a cost $\theta^i \mathscr{R}(v^i \mid \rho^i(x^i, \cdot))$. (Observe that nature's optimal choice can

depend on the state of all projects.) The state of project $i$ changes from $X_n^i = x^i$ to $X_{n+1}^i$ according to the distribution chosen by nature

$$P(X_{n+1}^i = j \mid X_n^i = x^i) = \upsilon^i(j), \quad j \in \mathscr{X}.$$

The state of all other projects remains unchanged. The payoff is the sum of the decision maker's reward $r^i(x^i)$ for choosing project $i$ and the penalty $\theta^i \mathscr{R}(v^i \mid \rho^i)$ on nature

$$c(x, i, \upsilon^i) = r^i(x^i) + \theta^i \mathscr{R}(\upsilon^i \mid \rho^i).$$

All rewards are discounted to time 0 with discount factor $\alpha \in (0,1)$. The decision maker and nature choose equilibrium policies to solve a zero-sum stochastic dynamic game with value function

$$V(x) \triangleq \sup_{\pi\in\Pi} \inf_{\gamma\in\Gamma} \mathbb{E}_{\mathbb{P}_{\pi,\gamma}}\left[\sum_{n\le\tau} \alpha^n c(X_n, i_n, \upsilon_n) \,\middle|\, X_0 = x\right], \quad (6)$$

where $\tau \triangleq \inf\{n\colon \pi_n(X_n) = \text{retire}\}$ is the retirement time, $i_n \triangleq \pi_n(X_n)$ is the decision maker's project choice, and $\upsilon_n \triangleq \gamma_n(X_n, i_n)$ is nature's choice of transition distribution at stage $n$. The payoff per stage is

$$c(x, i, \upsilon^i) = \begin{cases} r^i(x^i) + \theta^i \mathscr{R}(\upsilon^i \mid \rho^i) & i \in \{1, \ldots, p\}, \\ M & i = \text{retire}, \end{cases} \quad (7)$$

and the class of admissible policies for both the decision maker and nature are

$$\Pi = \{(\pi_0, \pi_1, \pi_2, \ldots) \mid \pi_n\colon \mathscr{X}^p \to \{1, \ldots, p\} \cup \{\text{retire}\}\},$$
$$\Gamma = \{(\gamma_0, \gamma_1, \gamma_2, \ldots) \mid \gamma_n\colon \mathscr{X}^p \times \{1, \ldots, p, \text{retire}\} \to \Upsilon\},$$

where

$$\Upsilon = \left\{q = (q(j), j \in \mathscr{X}) \in \mathscr{R}^{|\mathscr{X}|} \,\middle|\, q_j \ge 0, \sum_{j\in\mathscr{X}} q(j) = 1\right\}$$

is the set of probability distributions on $\mathscr{X}$.

### 3.3. Value Function and Robust Dynamic Programming

The value function associated with the robust multiarmed bandit problem (6) satisfies the following dynamic programming equation.

Theorem 2. *The function $V(x)$ defined in (6) is the unique solution to*

$$V(x) = \max\left\{M, \max_{i=1,\ldots,p} H^i(x, V)\right\}, \quad (8)$$

*where*

$$H^i(x, V) = r^i(x^i) + \inf_{\upsilon^i\in\Upsilon} \mathbb{E}_{\upsilon^i}[\alpha V(x^{i-}, X^i)] + \theta^i \mathscr{R}(\upsilon^i \mid \rho^i). \quad (9)$$

*It is optimal to choose project $i$ when $X_n = (x^1, \ldots, x^p)$ if $V(x) = H^i(x, V(x))$ and to permanently retire if $V(x) = M$. The value function can be obtained as the limit*

$$V(x) = \lim_{N\to\infty} V_N(x), \quad (10)$$

*where $V_N(x)$ is the unique solution of the $N$-stage dynamic programming equation*

$$V_0(x) = \max\{M, 0\},$$
$$V_k(x) = \max\left\{M, \max_{i=1,\ldots,p} H^i(x, V_{k-1})\right\}, \quad k = 1, \ldots, N. \quad (11)$$

A proof of Theorem 2 can be found in González-Trejo et al. (2003), Theorems 4.1 and 4.2.

### 3.4. Nature's Policy

The robust dynamic programming equation given in (8) prescribes the optimal project choice for the decision maker in the presence of the adversarial agent (nature). We now discuss nature's response in more detail.

Observe first that functions $H^i$ defined in (9) can be written as

$$\begin{aligned} H^i(x, V) &= r^i(x^i) + \inf_{\upsilon^i\in\Upsilon} \mathbb{E}_{\upsilon^i}[\alpha V(x^{i-}, X^i)] + \theta^i \mathscr{R}(\upsilon^i \mid \rho^i) \\ &= r^i(x^i) + \inf_{\upsilon^i \ll \rho^i} \mathbb{E}_{\upsilon^i}[\alpha V(x^{i-}, X^i)] + \theta^i \mathscr{R}(\upsilon^i \mid \rho^i), \end{aligned} \quad (12)$$

where the second equality holds since $\mathscr{R}(\upsilon^i \mid \rho^i) = +\infty$ if $\upsilon^i$ is not absolutely continuous with respect to $\rho^i$. To solve the minimization problem (12), we appeal to the following variational formula from the theory of large deviations, which we state in the setting of discrete random variables. (See, e.g., Dupuis and Ellis 1997, Proposition 1.4.2, for a more general statement.)

Proposition 1. *Let $\mathscr{X}$ be a finite set, $Y$ be a bounded real-valued random variable on $\mathscr{X}$, let $\theta > 0$ be a positive constant, and let $\kappa$ be a probability distribution on $\mathscr{X}$. Let $\mathbb{E}_\kappa(\cdot)$ and $\mathbb{E}_\eta(\cdot)$ denote expectations with respect to probability distributions $\kappa$ and $\eta$ on $\mathscr{X}$. Then*

$$\inf_{\eta\ll\kappa}\{\mathbb{E}_\eta[Y] + \theta \mathscr{R}(\eta \mid \kappa)\} = -\theta \log \mathbb{E}_\kappa\left\{\exp\left(-\frac{Y}{\theta}\right)\right\}. \quad (13)$$

*Furthermore, the infimum in (13) is uniquely attained at probability measure*

$$\eta^*(j) = \frac{\exp(-(Y(j)/\theta))}{\mathbb{E}_\kappa\{\exp(-(Y/\theta))\}}\kappa(j). \quad (14)$$

The identity (13) can also be written as

$$\inf_{\eta\ll\kappa}\left\{\sum_{j\in\mathscr{X}} Y(j)\eta(j) + \theta \sum_{j\in\mathscr{X}} \eta(j) \ln\frac{\eta(j)}{\kappa(j)}\right\} = -\theta \log\left\{\sum_{j\in\mathscr{X}} \kappa(j) \exp\left(-\frac{Y(j)}{\theta}\right)\right\}.$$

Observe that the likelihood ratio of the optimal $\eta^*$ with respect to $\kappa$ is a random variable $d\eta^*/d\kappa$ defined on $\mathscr{X}$ where

$$\frac{d\eta^*}{d\kappa}(j) = \frac{\exp(-(Y(j)/\theta))}{\mathbb{E}_\kappa\{\exp(-(Y/\theta))\}}, \quad j \in \mathscr{X}. \tag{15}$$

Applying this to functions $H^i$ in Equation (12), we obtain

$$H^i(x, V) = r^i(x^i) - \theta^i \log \mathbb{E}_{\rho^i}\left[\exp\left(-\frac{\alpha V(x^{i-}, X^i)}{\theta^i}\right)\right]. \tag{16}$$

Representation (16) will be used to establish structural properties of the optimal policy in the next section. In addition, it follows from (14) that for each arm $i = 1, \ldots, p$, the worst-case transition probability $v^{i,*}$ chosen by nature is

$$v^{i,*}(j) = \frac{\exp(-(\alpha V(x^{i-}, j)/\theta^i))}{\mathbb{E}_{\rho^i}[\exp(-(\alpha V(x^{i-}, X^i)/\theta^i))]}\rho^i(x^i, j), \quad j \in \mathscr{X}, \tag{17}$$

which has likelihood ratio

$$\frac{dv^{i,*}}{d\rho^i}(j) = \frac{\exp(-(\alpha V(x^{i-}, j)/\theta^i))}{\mathbb{E}_{\rho^i}[\exp(-(\alpha V(x^{i-}, X^i)/\theta^i))]}, \quad j \in \mathscr{X}. \tag{18}$$

Nature's optimal response (17) is intuitive. It modifies the nominal distribution $\rho^i(x^i, \cdot)$ by increasing the probability of transition to a state with low value (i.e., $V(x^{i-}, j)$ is small) and decreasing the probability of transition to a state that has high value (i.e., $V(x^{i-}, j)$ is large).

One implication of (17) is that projects are no longer independent under the worst-case measure $v^{i,*}$. This is because $v^{i,*}$ depends on the entire state vector $x = (x^1, \ldots, x^p)$ and not just $x^i$. Since independence is a key assumption when establishing the optimality of an index policy for the classical bandit problem (Theorem 1), it follows that index policies are no longer optimal for the robust problem, and that the optimal policy can only be obtained by carrying out the substantially more difficult task of solving the dynamic programming equations (8). Consequently, we shift our focus to identifying suboptimal policies that are easy to compute and developing methods to quantify their performance in the worst-case setting.

## 4. Robust Gittins Index

Although the optimal policy for the classical bandit problem is an index policy, this is unlikely to be the case for the robust model since projects are no longer independent under nature's worst-case response. On the other hand, since the Gittins index is optimal under the nominal model, it is natural to consider a robust version of this policy, which we now define, as a suboptimal policy for the robust problem.

Consider the robust multiarmed bandit problem (6) with retirement payment $M$. We let $V^i(x^i; m)$ denote the value function for the robust single-arm bandit problem, i.e., the problem where there is only one project $i$ with the option of a retirement payment $m$. This robust single-arm value function satisfies

$$V^i(x^i) = \max\left\{M, r^i(x^i) - \theta^i \log \mathbb{E}_{\rho^i}\left[\exp\left(-\frac{\alpha V^i(X^i)}{\theta^i}\right)\right]\right\}. \tag{19}$$

Analogous to the nominal case, for arm $i$ we define the robust Gittins index $G^i(x^i)$ as

$$G^i(x^i) \triangleq \inf\{m\colon V^i(x^i; m) = m\}, \quad i = 1, \ldots, p. \tag{20}$$

The way to interpret the index $G^i(x^i)$ is as the smallest retirement payment $m$ for which the decision maker is indifferent between retiring and continuing with arm $i$.

The robust Gittins index policy is then defined as follows. The decision maker permanently retires when the index $G^i(x^i)$ of each arm $i$ falls below the specified retirement payment $M$. That is, the retirement time is given by the following stopping time

$$\tau \triangleq \inf\left\{n\colon \max_{i=1,\ldots,p} G^i(X_n^i) \le M\right\}. \tag{21}$$

At each stage $n$ before the retirement time $\tau$, the project with the highest robust Gittins index

$$i_n \triangleq \arg\max_{i=1,\ldots,p} G^i(X_n^i)$$

is chosen.

Denoting the robust Gittins index policy by $G \in \Pi$, the associated value function is

$$V_G(x_0) = \inf_{\gamma\in\Gamma} \mathbb{E}_{\mathbb{P}_{G,\gamma}}\left[\sum_{n\le\tau} \alpha^n c(X_n, i_n, v_n) \,\middle|\, x_0\right], \tag{22}$$

where $c_n(x, i, v)$ is given by (7).

The robust Gittins index does not in general maximize the objective function (22), so $V_G(x)$ is a lower bound to the robust value function $V(x)$ given in (6). One exception is the case of a one-armed problem ($p = 1$), which we state without proof.

Proposition 2. *The robust Gittins index is optimal for a robust one-armed bandit problem.*

We note that computing the robust Gittins index is no more difficult than computing the classical Gittins index—both cases involve solving one-arm optimal stopping problems.

In contrast, computing the value function $V(x)$ for the robust bandit problem involves solving the dynamic programming equation (8) defined over

a $p$-dimensional state space, which is impractical even for moderately sized problems. Similar concerns apply to $V_G(x)$, the value function for nature's stochastic control problem (22), which is defined over the same state space. The focus of the next section is developing computationally tractable methods for bounding $V_G(x)$ and $V(x)$.

## 5. Performance Evaluation

Despite its intuitive appeal, it is natural to ask whether the performance regret, namely, the difference between the value function for the robust bandit problem and the value of the robust Gittins index $(V(x) - V_G(x))/V(x)$, can be quantified.

The regret bound $(V(x) - V_G(x))/V(x)$ is difficult to compute directly since $V(x)$ and $V_G(x)$ are value functions of stochastic control problems (6) and (22), respectively. In particular, solving either problem is computationally intractable, by virtue of the curse of dimensionality, and we do not have a characterization of the optimal policies that enables us to simulate either quantity.

To overcome these issues, we sidestep the computation of $V(x)$ and $V_G(x)$ by computing upper and lower bounds $V_L(x)$ and $V_U(x)$

$$0 < V_L(x) \le V_G(x) \le V(x) \le V_U(x).$$

Observing that

$$\begin{aligned} 0 \le \frac{V(x) - V_G(x)}{V(x)} &= 1 - \frac{V_G(x)}{V(x)} \\ &\le 1 - \frac{V_L(x)}{V_U(x)} = \frac{V_U(x) - V_L(x)}{V_U(x)}, \end{aligned}$$

computing lower and upper bounds gives us an upper bound on the regret

$$0 \le \frac{V(x) - V_G(x)}{V(x)} \le \frac{V_U(x) - V_L(x)}{V_U(x)} \tag{23}$$

(where clearly smaller is better). Naturally, we are interested in bounds $V_U(x)$ and $V_L(x)$, which are easier to compute than $V(x)$ and $V_G(x)$.

The upper bound $V_U(x)$ is obtained in §5.2 by approximating nature's worst-case transition probability. The approximation will be based on a value function estimate that we will derive. Under this suboptimal response by nature, therobust problem simplifies into a classical bandit problem for which a classical Gittins index policy is optimal. The value function $V_U(x)$ of this classical problem is an upper bound, since nature is not playing optimally, and is easy to compute by simulating the classical Gittins index under the transition probability associated with weakened nature.

The lower bound $V_L(x)$ is obtained in §5.3 by adapting the idea of information relaxation (Brown et al. 2010, Haugh and Lim 2012, Rogers 2007, Ye and Zhou 2015), which relaxes the requirement that nature's actions depend only on the information available at the time of the decision. This can be viewed as a "strengthening of nature," because it gives nature the ability to make use of future state information.

### 5.1. Preliminaries

We first derive structural properties of the optimal policy that will be used to evaluate the performance of the robust Gittins index. The results presented here are robust analogs of results in Bertsekas (1995) and Tsitsiklis (1986). Proofs can be found in the appendix.

Proposition 3. *The value function $V(x)$ is a nondecreasing function in $M$. Furthermore, $V(x)$ is constant for all $M \le -B/(1-\alpha)$, and $V(x) = M$ for all $M \ge B/(1-\alpha)$, where $B > 0$ is defined in* (1).

Consider now the problem where there is only one project $i$ with the option of retirement. The value function $V^i(x^i)$ of this problem satisfies

$$V^i(x^i) = \max\left\{M, r^i(x^i) - \theta^i \log \mathbb{E}_{\rho^i}\left[\exp\left(-\frac{\alpha V^i(X^i)}{\theta^i}\right)\right]\right\} \tag{24}$$

and can be obtained as $V^i(x^i) = \lim_{N\to\infty} V_N^i(x^i)$, where

$$\begin{aligned} V_0^i(x^i) &= \max\{M, 0\}, \\ V_k^i(x^i) &= \max\left\{M, r^i(x^i) - \theta^i \log \mathbb{E}_{\rho^i}\left[\exp\left(-\frac{\alpha V_{k-1}^i(X^i)}{\theta^i}\right)\right]\right\}, \end{aligned} \tag{25}$$

for $k = 1, \ldots, N$. Similarly, we can define $V^{i-}(x^{i-})$ and $V_k^{i-}(x^{i-})$ as the value function and $k$-stage value function, respectively, associated with the problem where all projects except project $i$ can be selected, with the option of retirement. We now have the following result.

Proposition 4. *The value function satisfies the following inequality*:

$$V^{i-}(x^{i-}) \le V(x) \le V^{i-}(x^{i-}) + (V^i(x^i) - M). \tag{26}$$

Inequality (26) provides an estimate of the robust value function $V(x)$ in terms of the value function $V^{i-}(x^{i-})$ in which arm $i$ is removed and the single-arm value function $V^i(x^i)$ is in excess of the retirement payment $M$. We will see in the following two sections that the estimate of the value function suggested by inequality (26) plays a central role in evaluating the performance of the robust Gittins index.

We mention in closing that a class of policies of interest in the classical bandit problem are the so-called "project-by-project retirement" (PPR) policies (Bertsekas 1995). Specifically, a policy is of the PPR type if there exist index functions $g^i\colon S \to \mathbb{R}$, $i = 1, \ldots, p$ such that at any stage $n$ and any state $x = (x^1, \ldots, x^p)$, the policy

1. permanently retires project $i$, if $g^i(x^i) \le M$; and
2. continues with some project, if $\min_j g^j(x^j) > M$.

One implication of Proposition 4 is that there exists a PPR policy that is optimal for the robust bandit problem.

### 5.2. Upper Bound (Weakening Nature)

We obtain the upper bound $V_U(x)$ in (23) by approximating nature's worst-case response $\nu^{i,*}$ given in Equation (18). The resulting policy is not optimal for nature, and it follows that the optimal scheduling policy, when playing against this suboptimal adversary, has a value function that has an upper bound $V_U(x) \ge V(x)$.

To characterize $V_U(x)$, recall from (18) that nature's worst-case distribution for arm $i$ has the form

$$\frac{d\nu^{i,*}}{d\rho^i}(j) = \frac{\exp(-(\alpha V(x^{i-}, j)/\theta^i))}{\mathbb{E}_{\rho^i}[\exp(-(\alpha V(x^{i-}, X^i)/\theta^i))]}, \quad j \in \mathcal{X}. \tag{27}$$

The value function estimate (26) of Proposition 4 suggests we approximate the value function as

$$V(x) \approx V^{i-}(x^{i-}) + (V^i(j) - M).$$

It now follows that nature's optimal response (27) can be approximated as

$$\begin{aligned}
\frac{d\nu^{i,*}}{d\rho^i}(j) &= \frac{\exp(-(\alpha V(x^{i-}, j)/\theta^i))}{\mathbb{E}_{\rho^i}[\exp(-(\alpha V(x^{i-}, X^i)/\theta^i))]} \\
&\approx \frac{\exp(-(\alpha(V^{i-}(x^{i-}) + V^i(j) - M)/\theta^i))}{\mathbb{E}_{\rho^i}[\exp(-(\alpha(V^{i-}(x^{i-}) + V^i(X^i) - M)/\theta^i))]} \\
&= \frac{\exp(-(\alpha(V^{i-}(x^{i-}) - M)/\theta^i))}{\exp(-(\alpha(V^{i-}(x^{i-}) - M)/\theta^i))} \\
&\quad \cdot \frac{\exp(-(\alpha V^i(j)/\theta^i))}{\mathbb{E}_{\rho^i}[\exp(-(\alpha V^i(X^i)/\theta^i))]} \\
&= \frac{\exp(-(\alpha V^i(j)/\theta^i))}{\mathbb{E}_{\rho^i}[\exp(-(\alpha V^i(X^i)/\theta^i))]} \\
&=: \frac{d\eta^{i,*}}{d\rho^i}(j), \quad j \in \mathcal{X}.
\end{aligned} \tag{28}$$

The transition probabilities for nature implied by (28) are

$$\begin{aligned}
&\mathbb{P}(X^i_{n+1} = j \mid X^i_n = x^i) \\
&\quad \triangleq \eta^{i,*}(x^i, j) = \frac{\exp(-(\alpha V^i(j)/\theta^i))}{\mathbb{E}_{\rho^i}[\exp(-(\alpha V^i(X^i_{n+1})/\theta^i))]}\rho^i(x^i, j), \\
&\qquad\qquad j \in \mathcal{X}.
\end{aligned} \tag{29}$$

Note that the approximation $\eta^{i,*}(x^i, \cdot)$ to nature's worst-case response has the same intuitive properties as $\nu^{i,*}(\cdot)$; it modifies the nominal transition probability $\rho^i(x^i, \cdot)$ by increasing (decreasing) the probability of transition to a state with low value (high value).

In contrast to the worst-case measure $\nu^{i,*}(\cdot)$ defined by (17), which depends on the state of all arms, $\eta^{i,*}(x^i, \cdot)$ only depends on the state of arm $i$. Likewise, the reward for choosing arm $i$

$$\tilde{r}^{i,*}(x^i) \triangleq r^i(x^i) + \theta^i \mathcal{R}(\eta^{i,*} \mid \nu^{i,*}) \tag{30}$$

only depends on the state of arm $i$. That is, the robust problem with the suboptimal response $\eta^{i,*}$ reduces to a classical bandit problem with transitions (29) and rewards (30), and it follows from Theorem 1 that the Gittins index policy is optimal. The value function $V_U(x)$ for this classical problem is an upper bound for $V(x)$. Since transition probabilities $\eta^{i,*}$ are easy to generate, the upper bound $V_U(x)$ can be estimated by using simulation, which we will illustrate in §5.4.

### 5.3. Lower Bound (Strengthening Nature)

We now obtain the lower bound $V_L(x)$ in (23) by adapting the idea of information relaxation (Brown et al. 2010, Haugh and Lim 2012, Rogers 2007, Ye and Zhou 2015). Here, we relax the requirement that nature's actions depend only on the information available at the time of the decision and allow nature to select transition probabilities that depend on future realizations of the state. Clearly, the expected reward $V_L(x)$ generated by the robust Gittins index policy against this stronger adversary is less than the expected reward $V_G(x)$, defined in (22), that it generates when playing against the usual (and less powerful) nonanticipative adversary, so $V_L(x) \le V_G(x)$.

**5.3.1. Road Map.** Before we get into the details of computing the lower bound $V_L(x)$, we provide a high-level road map of the basic reasoning and key quantities that make up this subsection. The road map consists of three parts:

(i) *Nature's Control Problem.* We first observe that, under the robust Gittins index policy $G$, nature's sequential responses can be viewed as a stochastic control problem in which nature's actions are transition probabilities. Consequently, we can interpret $V_G(x)$ as the value function of nature's stochastic control problem. To make this interpretation absolutely clear, we rename the function $V_G(x)$ as $W(q, x)$, reminding the reader that this is the value function for nature's problem as opposed to the decision maker's problem. The first step for computing the bound is reformulating nature's problem such that its decisions are likelihood ratios instead of transition probabilities.

(ii) *Information Relaxation.* We next observe that, if nature's actions are allowed to make use of future state information, we can obtain an easy to compute lower bound $V_L^0(x)$. Although it is possible to stop here (because we have found a lower bound to $V_G(x)$), such a lower bound is typically loose.

(iii) *Information Penalty.* The goal of the final part is to tighten the lower bound $V_L^0(x)$. To do this, we still endow nature with knowledge of future project choices and state transitions, but we now introduce a penalty function $\lambda(h, \mathbf{y})$ that penalizes nature whenever this information is used to make a decision. We will show that solving this penalty version of the problem produces a lower bound $V_L^\lambda(x)$ that can be substantially tighter than the lower bound $V_L^0(x)$.

With the road map set, we now work through the mathematical details for each of the three parts.

#### 5.3.2. Computing the Lower Bound

(i) *Nature's Control Problem.* Recall that the worst-case expected discounted reward associated with the robust Gittins index policy is obtained by solving a stochastic control problem for nature given by

$$V_G(x_0) = \inf_{\gamma\in\Gamma} \mathbb{E}_{\mathbb{P}_{G,\gamma}}\left[\sum_{n\le\tau} \alpha^n c(X_n, i_n, \upsilon_n) \,\middle|\, x_0\right], \quad (31)$$

where $\tau$ is the retirement time (21) associated with the robust Gittins index.

As described in §3.4, when the state is $X_n = (x^1, \ldots, x^p)$ and project $i$ is chosen by the decision maker, nature's optimal response is a probability distribution $\upsilon_n(\cdot)$ for the transition of the state $X_n^i$ of project $i$

$$\mathcal{P}(X_{n+1}^i = j \mid X_n^i = x^i) = \upsilon_n(j).$$

Observe that there is a one-to-one relationship between probability distributions $\upsilon_n(\cdot)$ that are absolutely continuous with respect to $\rho^i(x^i, \cdot)$ and likelihood ratios/random variables $y_n = (y_n(j), j \in \mathcal{X})$ satisfying

$$\mathbb{E}_{\rho^i(x^i,\cdot)}[y_n] = \sum_{j\in\mathcal{X}} y_n(j)\rho(x^i, j) = 1$$

through the relationship

$$y_n(j) \equiv \frac{d\upsilon_n}{d\rho^i}(j) = \frac{\upsilon_n(j)}{\rho^i(x^i, j)}, \quad j \in \mathcal{X},$$

or equivalently

$$\upsilon_n(j) = y_n(j)\rho^i(x^i, j), \quad j \in \mathcal{X}. \quad (32)$$

It follows that optimizing over transition probabilities $\gamma = (\upsilon_0, \upsilon_1, \upsilon_2, \ldots) \in \Gamma$ in (31) is equivalent to optimizing over likelihood ratios $y = (y_0, y_1, y_2, \ldots)$, where nature's choice $y_n$ at stage $n$ can depend on the state of all projects as well as the project $(X_n, i_n)$ chosen at time $n$; i.e.,

$$y_n \in \left\{ y = (y(j), j \in \mathcal{X}) \,\middle|\, y\colon \mathcal{X}^p \times \{1, \ldots, p\} \to \mathbb{R}^{|\mathcal{X}|}, \right.$$
$$\left. y(j) \ge 0, \text{ and } \sum_{j\in\mathcal{X}} y(j)\rho(x^i, j) = 1 \right\}.$$

Put simply, nature's problem is equivalent to selecting a sequence of likelihood ratios $y = (y_0, y_1, y_2, \ldots)$ that minimizes the decision maker's total profits. Therefore, if we denote the set of all such sequences by $\mathcal{Y}$, nature's control problem can be equivalently formulated as follows:

$$W(q_0, x_0) \triangleq \inf_{\mathbf{y}\in\mathcal{Y}} \mathbb{E}_{\mathbb{P}_G}\left[\sum_{n\le\tau-1} \alpha^n q_n c(X_n, i_n, y_n) + \alpha^\tau q_\tau M \,\middle|\, x_0\right]$$
$$= V_G(x_0), \quad (33)$$

where $q_n$ satisfies

$$q_0 = 1,$$
$$q_n = \prod_{k=0}^{n-1} y_k(X_{k+1}^{i_k}), \quad n \ge 1$$

and the reward terms are

$$c(X_n, i_n, y_n) = r^{i_n}(X_n^{i_n}) + \theta^{i_n} \sum_{j\in\mathcal{X}} y_n(j)\log(y_n(j))\rho^{i_n}(X_n^{i_n}, j).$$

Note that we have intentionally renamed the function $V_G(x)$ as $W(q, x)$ to remind the reader that this is the value function for nature's problem as opposed to the decision maker's problem.

Observe that the probability measure $\mathbb{P}_G$ in (33) is the nominal model for the state process $\{X_n, n \ge 0\}$ under the robust Gittins index policy, and it does not depend on nature's policy $\mathbf{y}$. This implies that states can be generated under the robust Gittins index by using the nominal transition probabilities, which turns out to be useful when generating a lower bound for (33).

We now write out the dynamic programming equations for problem (33). At any arbitrary stage, suppose the current state is $x = (x^1, \ldots, x^p)$ and project $i$ has the highest Gittins index; i.e., $G^i(x^i) = \max_j G^j(x^j)$. Then, for all $x = (x^1, \ldots, x^p)$ such that $G^i(x^i) > M$, it follows by standard dynamic programming arguments that $W(q, x)$ satisfies the recursion

$$W(q, x) = \inf_y \left\{ q \cdot \left( r^i(x^i) + \theta^i \sum_{j\in\mathcal{X}} y(j)\log(y(j))\rho^i(x^i, j) \right) \right.$$
$$\left. + \alpha \sum_{j\in\mathcal{X}} W(q \cdot y(j), x^{i-}, j)\rho^i(x^i, j) \right\}, \quad (34)$$

and for all $x = (x^1, \ldots, x^p)$ such that $G^i(x^i) = M$,

$$W(q, x) = q \cdot M.$$

The infimum in (34) is taken over all nonnegative random variables $y = (y(j), j \in \mathcal{X})$ defined on $\mathcal{X}$ satisfying

$$\mathbb{E}_{\rho^i(x^i,\cdot)}[y] = \sum_{j\in\mathcal{X}} y(j)\rho^i(x^i, j) = 1.$$

(ii) *Information Relaxation.* The goal now is to find an easy-to-compute lower bound to (33). The key idea is to relax the constraint $y \in \mathcal{Y}$ that restricts nature's decisions to nonanticipative policies. This relaxation leads to a lower bound since nature's decisions can now depend on future state realizations and project choices.

We begin by formalizing the notion that nature can use future state and project information to make its decisions. Specifically, let $\mathcal{G}_n = \sigma(X_0, i_0, \ldots, X_n, i_n)$ be the sigma-algebra generated by the history of state–project pairs up until stage $n$ (intuitively, the information available at stage $n$) and let $\mathcal{G}_\infty = \bigvee_{n=0}^{\infty} \mathcal{G}_n$ be the smallest sigma-algebra containing all the $\mathcal{G}_n$ (intuitively, the complete record of state–project pairs until retirement). We relax the nonanticipation constraint that $y_n$ be $\mathcal{G}_n$-measurable in (33) and allow $y_n$ to be $\mathcal{G}_\infty$-measurable. Denoting by $\bar{\mathcal{Y}}$ the set of all policies $\mathbf{y} = (y_n\colon n \geq 0)$ such that $(y_n(j), j \in \mathcal{X})$ is $\mathcal{G}_\infty$-measurable and satisfies

$$\sum_{j\in\mathcal{X}} y_n(j)\rho^{i_n}(x_n^{i_n}, j) = 1,$$

the relaxed problem is

$$\begin{aligned} V_L^0(x_0) &\triangleq \inf_{\mathbf{y}\in\bar{\mathcal{Y}}} \mathbb{E}_{\mathbb{P}_G}[P(\mathbf{y}) \mid q_0, x_0] \\ &= \mathbb{E}_{\mathbb{P}_G}\left[\inf_{\mathbf{y}\in\bar{\mathcal{Y}}} P(\mathbf{y}) \,\middle|\, q_0, x_0\right], \end{aligned} \tag{35}$$

where

$$P(\mathbf{y}) \triangleq \sum_{n=0}^{\tau-1} \alpha^n q_n c(X_n, i_n, y_n) + \alpha^\tau q_\tau M.$$

The infimum moves inside the expectation because the admissible set $\bar{\mathcal{Y}}$ allows nature to optimize by using knowledge of the entire sample path of states and projects until retirement.

This is a lower bound to (34) since

$$\begin{aligned} W(q_0, x_0) &= \inf_{\mathbf{y}\in\mathcal{Y}} \mathbb{E}_{\mathbb{P}_G}[P(\mathbf{y}) \mid q_0, x_0] \\ &\geq \inf_{\mathbf{y}\in\bar{\mathcal{Y}}} \mathbb{E}_{\mathbb{P}_G}[P(\mathbf{y}) \mid q_0, x_0] \\ &= V_L^0(x_0), \end{aligned} \tag{36}$$

where the inequality follows because $\mathcal{Y} \subseteq \bar{\mathcal{Y}}$. That is, nature is given more information when the class of admissible policies enlarges from $\mathcal{Y}$ to $\bar{\mathcal{Y}}$.

We estimate the lower bound $V_L^0(x_0)$ using Monte Carlo simulation, namely, by sampling multiple realizations of the random variable

$$\inf_{\mathbf{y}\in\bar{\mathcal{Y}}} P(\mathbf{y}) \tag{37}$$

and averaging over them. Each sample of (37) is obtained by generating states and projects $(X_0, i_0)$, $(X_1, i_1)$, $(X_2, i_2), \ldots, (X_\tau, i_\tau)$ until retirement, using the nominal model $\rho^i$ and the robust Gittins policy, and solving the associated optimization problem for its optimal value. We will see in §5.4 that the lower bound $V_L^0(x_0)$ can be loose, which is unsurprising because the ability to use future state information gives nature a substantial advantage that can be used to significantly lower the expected total profit. In the next section, we show how ideas from the literature on information relaxation (Brown et al. 2010, Rogers 2007) can be used to improve this bound.

(iii) *Information Penalty.* The lower bound $V_L^0(x_0)$ can be uninformative because unrestricted use of future information when optimizing $P(\mathbf{y})$ can give nature substantially more power than when nature is restricted to adapted policies. We counteract this effect by using ideas from Brown et al. (2010) and Rogers (2007), where nature's power is "controlled" by imposing a penalty whenever it uses future information to make a decision. The penalized problem can be regarded as a dual relaxation of nature's nonanticipative constraint $\mathbf{y} \in \mathcal{Y}$ in (33), and it leads to an improved lower bound if the penalty is well chosen.

For any function $h\colon \{1, \ldots, p\} \times S \to \mathbb{R}$, we define the penalty function

$$\begin{aligned} \lambda(h, \mathbf{y}) \triangleq \sum_{n=0}^{\tau-1} \alpha^{n+1} q_n \Bigg( & y_n(x_{n+1}^{i_n}) h(i_n, x_{n+1}^{i_n}) \\ & - \sum_{j=1}^{N} y_n(j) h(i_n, j) \rho^{i_n}(x_n^{i_n}, j) \Bigg) \end{aligned} \tag{38}$$

and the associated penalized problem

$$\inf_{\mathbf{y}\in\bar{\mathcal{Y}}} \{P(\mathbf{y}) - \lambda(h, \mathbf{y})\} \tag{39}$$

in which this penalty is included in the objective function. Intuitively, the penalty function (38) attempts to eliminate the (total discounted) benefits for nature when likelihood sequences $\mathbf{y}$ are selected from the enlarged admissible set $\bar{\mathcal{Y}}$ as opposed to the nonanticipative set $\mathcal{Y}$. Of course, for this penalty structure to actually work, one must check that for $\mathbf{y} \in \mathcal{Y}$ the penalty reduces to zero in expectation. The following

result shows that this is indeed the case; consequently, this penalty function can be used to compute a lower bound of $V_G(x_0) = W(q_0, x_0)$.

Lemma 1 (Weak Duality). *For any* $\mathbf{y} \in \mathcal{Y}$, *the penalty function* $\lambda(h, \mathbf{y})$ *satisfies*

$$\mathbb{E}_{\mathbb{P}_G}[\lambda(h, \mathbf{y}) \mid q_0, x_0] = 0.$$

*Furthermore,*

$$V_L^\lambda(x_0) \triangleq \mathbb{E}_{\mathbb{P}_G}\left[\inf_{\mathbf{y}\in\bar{\mathcal{Y}}}\{P(\mathbf{y}) - \lambda(h, \mathbf{y})\} \,\middle|\, q_0, x_0\right] \le W(q_0, x_0). \tag{40}$$

Proof. See the appendix. □

Proposition 1 shows that any choice of the function $h(i_n, j)$ delivers a lower bound to $W(q_0, x_0)$. The challenge now is to specify $h(i_n, j)$ to make this lower bound as large as possible. In this regard, observe first that choosing $h \equiv 0$ corresponds to a penalty function $\lambda \equiv 0$ and the lower bound $V_L^\lambda(x_0) = V_L^0(x_0)$, given in (35). On the other hand, the discussion in Brown et al. (2010) explains that the function $\lambda(h, \mathbf{y})$ can be interpreted as a penalty on nature for using future information, and that a natural choice is $h(i, j) = V^i(j)$, the value function of the robust single-armed bandit problem. Bounds associated with this choice will be studied in §5.4.

A key observation is that the lower bound $V_L^\lambda(x_0)$ in (40) is in a form that makes it convenient to estimate using simulation. In particular, we can write

$$V_L^\lambda(x_0) = \mathbb{E}_{\mathbb{P}_G}[Y]$$

where the random variable $Y \equiv Y(X_0, i_0, \ldots, X_\tau, i_\tau)$ is obtained by solving

$$\begin{aligned} Y(X_n, i_n, n \le \tau) &\triangleq \inf_{\mathbf{y}\in\bar{\mathcal{Y}}}\{P(\mathbf{y}) - \lambda(h, \mathbf{y})\} \\ &= \inf_{\mathbf{y}\in\bar{\mathcal{Y}}}\Bigg\{\sum_{n=0}^{\tau-1}\alpha^n q_n c(X_n, i_n, y_n) + \alpha^\tau q_\tau M \\ &\quad - \sum_{n=0}^{\tau-1}\alpha^{n+1} q_n\Bigg(y_n(X_{n+1}^{i_n})h(i_n, X_{n+1}^{i_n}) \\ &\quad - \sum_{j\in\mathcal{X}} y_n(j)h(i_n, j)\rho^{i_n}(X_n^{i_n}, j)\Bigg)\Bigg\} \end{aligned} \tag{41}$$

subject to

$$q_{n+1} = \prod_{k=0}^{n-1} y_k(x_{k+1}^{i_k}), \quad n \ge 1,$$
$$q_0 = 1.$$

It follows that $V_L^\lambda(x_0)$ can be estimated by using Monte Carlo simulation. Specifically, we simulate $k = 1, \ldots, K$ project–state sample paths until retirement, $(x_0^{(k)}, i_0^{(k)}, \ldots, x_{\tau_k}^{(k)}, i_{\tau_k}^{(k)})$ using the nominal model $\rho^i$ $(i = 1, \ldots, p)$ and the robust Gittins index, and we compute $Y^{(k)} \equiv Y^{(k)}(x_n^{(k)}, i_n^{(k)}; n \le \tau_k)$ for each such sample path by solving (41). An estimate of $V_L^\lambda(x_0)$ is the sample average

$$V_L^\lambda(x_0) \approx \frac{1}{K}\sum_{k=1}^K Y^{(k)} = \frac{1}{K}\sum_{k=1}^K Y^{(k)}(x_n^{(k)}, i_n^{(k)}; n \le \tau_k). \tag{42}$$

We now discuss how the optimization problem (41) can be solved. Assume first that a sample path of projects and states $(x_n, i_n, n \le \tau)$ has been generated. Conditional on this sequence, (41) is a deterministic optimal control problem with state $q_n \ge 0$ and control $y_n$. Specifically, we define the value function at stage $k$ when the state is $q$ as

$$\begin{aligned} W_k^\lambda(q) = \inf_{\mathbf{y}\in\bar{\mathcal{Y}}}\Bigg\{&\sum_{n=k}^{\tau-1}\alpha^n q_n c(x_n, i_n, y_n) + \alpha^\tau q_\tau M \\ &- \sum_{n=k}^{\tau-1}\alpha^{n+1} q_n(y_n(x_{n+1}^{i_n})h(i_n, x_{n+1}^{i_n}) \\ &- \sum_{j=1}^{N} y_n(j)h(i_n, j)\rho^{i_n}(x_n^{i_n}, j))\Bigg\} \end{aligned} \tag{43}$$

subject to

$$q_{n+1} = q_n y(x_{n+1}^{i_n}), \quad n \ge k,$$
$$q_k = q.$$

Observe that we have written the constraint in (41) as a recursive equation for the state $q_n$. Clearly, the optimal objective value for (41) is given by $Y((x_n, i_n), n \le \tau) = W_0^\lambda(1)$. (To ease notation, we suppress the dependence of $W_k^\lambda(q)$ on $(x_n, i_n, n \le \tau)$.)

It is clear from (43) and the multiplicative structure of the state equation for $q_n$ that $W_k^\lambda(q)$ is linear in $q$; namely, $W_k^\lambda(q) = qW_k^\lambda(1)$. (Alternatively, this can be shown inductively by using the associated dynamic programming equations.) This simplifies the computation of $W_k^\lambda(q)$ since the value function is now completely characterized by $W_k^\lambda(1)$, which is a function of time $k$. The following result gives a recursive equation for computing $Y(x_n, i_n, n = 0, \ldots, \tau)$.

Theorem 3. *Suppose we are given a sequence of state–project pairs* $(x_0, i_0, \ldots, x_\tau, i_\tau)$ *and a retirement time* $\tau$. *The value function of the optimal control problem* (43) *satisfies*

$$W_n^\lambda(q) = qW_n^\lambda(1) = qf_n^\lambda,$$

*where*

$$\begin{aligned} f_n^\lambda = r^{i_n}(x_n^{i_n}) - \theta^{i_n}\ln\Bigg\{&\sum_{j\ne x_{n+1}^i}\rho^{i_n}(x_n^{i_n}, j)\exp\left(-\frac{\alpha}{\theta^{i_n}}h(i_n, j)\right) \\ &+ \rho^{i_n}(x_n^{i_n}, x_{n+1}^{i_n})\exp\Bigg(-\frac{\alpha}{\theta^{i_n}} \\ &\cdot\left[\frac{f_{n+1}^\lambda - h(i_n, x_{n+1}^{i_n})(1-\rho^{i_n}(x_n^{i_n}, x_{n+1}^{i_n}))}{\rho^{i_n}(x_n^{i_n}, x_{n+1}^{i_n})}\right]\Bigg)\Bigg\} \end{aligned} \tag{44}$$

$$f_\tau^\lambda = M.$$

*The optimal objective value, conditional on* $(x_0, i_0, \ldots, x_\tau, i_\tau)$*, is*

$$Y(x_n, i_n, n=0,\ldots,\tau) = W_0^\lambda(1) = f_0^\lambda.$$

*Nature's worst-case response* $y^* = (y_0^*, \ldots, y_\tau^*)$ *is given by*

$$y_n^*(j) = \beta \exp\left(-\frac{\alpha}{\theta^{i_n}} h(i_n, j)\right), \quad j \neq x_{n+1}^{i_n},$$

$$y_n^*(x_{n+1}^{i_n}) \tag{45}$$
$$= \beta \exp\left(-\frac{\alpha}{\theta^{i_n}}\left[\frac{f_{n+1}^\lambda - h(i_n, x_{n+1}^{i_n})(1-\rho^{i_n}(x_n^{i_n}, x_{n+1}^{i_n}))}{\rho^{i_n}(x_n^{i_n}, x_{n+1}^{i_n})}\right]\right),$$

*where the normalizing constant*

$$\beta = \left\{ \sum_{j \neq x_{n+1}^{i_n}} \rho^{i_n}(x_n^{i_n}, j) \exp\left(-\frac{\alpha}{\theta^{i_n}} h(i_n, j)\right) + \rho^{i_n}(x_n^{i_n}, x_{n+1}^{i_n}) \cdot \exp\left(-\frac{\alpha}{\theta^{i_n}}\left[\frac{f_{n+1}^\lambda - h(i_n, x_{n+1}^{i_n})(1-\rho^{i_n}(x_n^{i_n}, x_{n+1}^{i_n}))}{\rho^{i_n}(x_n^{i_n}, x_{n+1}^{i_n})}\right]\right)\right\}^{-1}$$

*is chosen such that* $\mathbb{E}_{\rho^{i_n}(x_n^{i_n}, \cdot)}[y_n^*] = 1$.

The proof of Theorem 3 is given in the appendix. We also highlight that the method presented in this section does not only apply to the robust Gittins index policy; it is a generic method that can be applied to any policy of interest.

### 5.4. Numerical Example

We now provide an example that illustrates how the upper bound $V_U(x)$ of §5.2 and the lower bound $V_L^\lambda(x)$ of §5.3 can be used to quantify the performance of the robust Gittins index policy. In the calculation of the lower bound, we use $h(i, j) = V^i(j)$, the value function of the robust one-armed problem, to compute the penalty.

Suppose the decision maker can choose between $p = 2$ projects. The state space of each project $S = \{1, 2, 3\}$, and their associated transition matrices and reward vectors are given by

$$\rho^1 = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.7 & 0.1 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}, \quad \rho^2 = \begin{bmatrix} 0.4 & 0.5 & 0.1 \\ 0.5 & 0.4 & 0.1 \\ 0.0 & 0.0 & 1.0 \end{bmatrix},$$
$$r^1 = \begin{bmatrix} 10 & 5 & 1 \end{bmatrix}, \quad r^2 = \begin{bmatrix} 6 & 8 & 2 \end{bmatrix}.$$

Note that we have chosen transition matrices $\rho^1$ and $\rho^2$ such that state $x = 3$ is an absorbing state. This is simply to ensure that permanently retiring from all projects is a viable option. The retirement reward is $M = 20$, discount factor is $\alpha = 0.8$, and penalty parameters are $\theta^1 = 2$ and $\theta^2 = 4$. We choose the initial state vector $(X_0^1, X_0^2) = (1, 1)$.

**Table 1** **Single-Arm Value Function $V^i(x)$ and Gittins Index $G^i(x)$**

| $x$ | 1 | 2 | 3 | $x$ | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| $V^1(x)$ | 29.96 | 23.64 | 20.00 | $G^1(x)$ | 50.00 | 27.10 | 5.00 |
| $V^2(x)$ | 27.10 | 28.99 | 20.00 | $G^2(x)$ | 33.37 | 40.00 | 10.00 |

**Figure 2** **(Color online) Single-Arm Value Function $V^1(1)$ vs. Retirement Reward $M$**


Using these model parameters, we first compute the single-arm value function $V^i(x)$ and robust Gittins index $G^i(x)$ for $i = 1, 2$ and $x = 1, 2, 3$, in Table 1.[2]

We plot the graph of $V^1(1)$ for different retirement rewards $M$ in Figure 2. From Lemma 3, we know that $V^1(1)$ is a nondecreasing function in $M$, and that $V^1(1)$ is constant for all $M \leq -B/(1-\alpha) = -50$, and $V^1(1) = M$ for all $M \geq B/(1-\alpha) = 50$. For $M = 20$, the graph in Figure 2 corroborates the values $V^1(1) = 29.96$ and $G^1(1) = \inf\{m\colon V^1(1) = m\} = 50$, as reported in Table 1.

To compute the upper bound $V_U(1, 1)$ defined in §5.2, we next compute the (approximated) worst-case transition probabilities $\eta^{i,*}$ defined in (28) and the "rewards" $\tilde{r}^i$ defined in (30).

$$\eta^{1,*} = \begin{bmatrix} 0.11 & 0.17 & 0.72 \\ 0.01 & 0.61 & 0.38 \\ 0.00 & 0.00 & 1.00 \end{bmatrix}, \quad \eta^{2,*} = \begin{bmatrix} 0.34 & 0.30 & 0.36 \\ 0.42 & 0.23 & 0.35 \\ 0.00 & 0.00 & 1.00 \end{bmatrix},$$
$$\tilde{r}^1 = \begin{bmatrix} 12.61 & 5.75 & 1.00 \end{bmatrix}, \quad \tilde{r}^2 = \begin{bmatrix} 7.00 & 8.94 & 2.00 \end{bmatrix}.$$

Following the procedures outlined in §§5.2 and 5.3, we are now ready to compute the upper bound $V_U(1, 1)$, and lower bounds $V_L^\lambda(1, 1)$ (with penalty) and $V_L^0(1, 1)$ (without penalty), in a simulation with 100,000 trials. The means, mean standard errors (MSE), and computational times (in seconds) for the three quantities are summarized in Table 2.

For this particular example, Table 2 shows that the lower bound $V_L^\lambda(1, 1)$ is significantly tighter than the lower bound $V_L^0(1, 1)$ obtained when no penalty is

[2] The single-arm value functions $V^i(x)$ were computed by using the value iteration given in (25). The Gittins indices $G^i(x) = \inf\{m\colon V^i(x) = m\}$ were then found through a simple numerical search in $m$.

**Table 2 Performance Bounds of the Robust Gittins Policy**

| | Mean | MSE | Times (sec.) |
|---|---|---|---|
| UB | 35.2896 | 0.0119 | 7.2696 |
| LB (penalty) | 32.4674 | 0.0030 | 214.0646 |
| LB (no penalty) | 15.0808 | 0.0057 | 179.5728 |

used. In particular, the value function $V_G(1,1)$ associated with the robust Gittins index policy is no less than

$$V_U(1,1) - V_L(1,1) = 35.2896 - 32.4674 = 2.8222,$$

from the true value function $V(1,1)$. Put another way, the percentage difference

$$0 \le \frac{V(1,1) - V_G(1,1)}{V(1,1)} \le \frac{V_U(1,1) - V_L^{\lambda}(1,1)}{V_U(1,1)} = 0.080$$

(see (23)) implies that the robust Gittins index policy will obtain a value $V_G(1,1)$ that is at least 92% of the true value $V(1,1)$. It is also important to note that the mean standard errors (MSE) are quite small, suggesting that the mean estimates are precise. All computations in Table 2 were done on an Intel Core i3-2310M CPU at 2.10 GHz with 2.00 GB RAM.

## 6. Application to Bayesian Webpage Design Problems

Online experimentation and optimization is increasingly important in web design and online marketing applications (see, e.g., Scott 2010, 2013; and White 2013). A simple but important example of such a problem is that of webpage design optimization, where a designer proposes several webpage design alternatives, each "differing in font, image choice, layout, and other design elements" (Scott 2010, p. 640), and would like to determine the one that is optimal according to an objective designed by the website owner. Finding the optimal webpage typically involves experimentation, and a number of online services for doing this are available.

Customer preferences can change rapidly, so periods of exploration to assess the profitability of different alternatives need to be properly balanced with periods of exploitation where the design that currently shows the most promise is displayed. It is therefore a common feature of these applications to formulate this problem as a bandit problem where "arms" correspond to design alternatives and the "reward," depending on the application, could be the net value of "checkouts" or "sales" or the number of views of a particular news article, during a predefined time duration $\delta$. Although bandit models naturally capture this exploration–exploitation trade-off, the dynamics of customer preferences are complex and not well understood, and the assumptions of the classical bandit problem are unlikely to hold. It thus follows that the robust bandit problem, which explicitly accounts for violations of the classical bandit assumptions, is a natural framework to address this problem.

With these considerations in mind, we use simulation to compare the performance of the robust and classical Gittins policies under progressively serious violations of the assumptions of the classical bandit problem. One striking observation is that the robust policy generates a higher expected profit with substantially lower downside risk relative to the classical Gittins policy, even under relatively mild violations of the classical assumptions. The reduction in downside risk suggests that the robust policy has the potential for improved out-of-sample performance when the assumptions of the classical bandit model do not hold, and we plan to test this conjecture in future work.

As a final note, one methodological contribution of this section is the proposal of a new data-driven method based on $k$-fold cross validation for calibrating ambiguity parameters that model the decision maker's level of trust in the nominal assumptions.

### 6.1. Nominal Bayesian Model

In the Bayesian model, there are $p$ alternative designs, and the nominal model assumes that rewards $d_n^i$ for design $i \in \{1, \ldots, p\}$ are generated independently and identically distributed (iid) from a distribution $f_{\eta_i}^i(\cdot)$ that depends on an unknown parameter $\eta_i$. The nominal model assumes a prior distribution $x_0^i(\cdot)$ on the parameter $\eta_i$, which, upon observing a new reward $d_1^i$, is updated to a posterior $x_1^i(\cdot)$ using Bayes' rule.

For analytical tractability, conjugate prior–likelihood pairs $(x_0^i(\cdot), f_{\eta_i}^i(\cdot))$ are commonly chosen for the reward distribution of the model. For example, we adopt a gamma-exponential conjugate pair, which consists of an exponentially distributed likelihood

$$f_{\eta_i}^i(d) = \eta_i e^{-\eta_i d}, \tag{46}$$

and a gamma$(\alpha_i, \beta_i)$ distributed prior

$$x_0^i(\eta) = \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} \eta^{-\alpha_i - 1} e^{-\eta \beta_i}, \tag{47}$$

in the example below. This nominal model family is convenient because after a reward $d_1^i = d$ is observed, the posterior $x_1^i(\eta)$ is a gamma$(\alpha_i + 1, \beta_i + d)$ distribution, so we can identify the prior (i.e., system state) for arm $i$, $x_0^i(\cdot)$, with the parameter pair $(\alpha_i, \beta_i)$ instead of the entire probability distribution. We are concerned about the impact of the likelihood $f_{\eta_i}^i(d)$ being misspecified on the performance of the classical and robust Gittins indices. We emphasize that

convenient prior–likelihood pairs are often chosen, even if they are known to be misspecified, because of the substantial computational advantages that they bring, and that one important advantage of our robust model is that it allows the decision maker to retain the convenience of this modeling assumption while accounting for the fact that it is wrong.

### 6.2. Calibration of Ambiguity Parameters

Our formulation of the robust bandit problem, and the associated robust Gittins index, depends on the choice of the ambiguity parameters $\theta = (\theta^1, \ldots, \theta^p)$, and it is clearly desirable that these parameters be chosen to "optimize" out-of-sample performance of the policy that we compute. We now provide an approach for doing this that only uses (limited) historical rewards that we have for each of the projects. The method we propose is an adaption of $k$-fold cross validation (Tibshirani 1996).

Suppose for each design $i = 1, \ldots, p$ that the decision maker has historical rewards $\mathcal{D}_i = \{d_1^i, \ldots, d_{n_i}^i\}$. First, values for the ambiguity parameters $\theta = (\theta^1, \ldots, \theta^p)$ are fixed. Next, each data set $\mathcal{D}_i = \{d_1^i, \ldots, d_{n_i}^i\}$, $i = 1, \ldots, p$, is (randomly) partitioned into $k$ subsets of approximately equal size. One of the subsets is selected as the validation set $\mathcal{V}_i$, and the remaining $k-1$ subsets constitute the training set $\mathcal{T}_i$. For ease of notation, we relabel the elements of the validation and training sets as $\mathcal{V}_i = \{v_1^i, \ldots, v_{\nu_i}^i\}$ and $\mathcal{T}_i = \{t_1^i, \ldots, t_{\tau_i}^i\}$, respectively.

Using the training sets $\mathcal{T}_i$, $i = 1, \ldots, p$, the posterior is updated by using Bayes' rule under the nominal model. In the case of the gamma-exponential model family (46) and (47), the posterior for design $i$ is

$$x_{\tau_i}^i = \left(\alpha_i + \tau_i, \beta_i + \sum_{j=1}^{\tau_i} t_j^i\right). \tag{48}$$

After the posteriors have been updated by using the training data, we evaluate the performance of the robust Gittins index policy using the validation data. To do this, we compute a bootstrap estimate of the expected total discounted reward under the robust Gittins index policy by sampling with replacement from the validation sets $\mathcal{V}_1, \ldots, \mathcal{V}_p$ (see, e.g., Efron (1979)). We denote this estimate as $\psi_1(\theta)$.

This process is repeated $k-1$ additional times, each time selecting a different subset for the validation set, which produces $k-1$ additional bootstrap estimates $\psi_2(\theta), \ldots, \psi_k(\theta)$. The average over these $k$ repeats given by

$$\psi_{\text{ave}}(\theta) = \frac{1}{k}\sum_{l=1}^{k} \psi_l(\theta) \tag{49}$$

is then computed. The number $\psi_{\text{ave}}(\theta)$ is an estimate of the performance of the robust Gittins index for the ambiguity parameters $\theta = (\theta^1, \ldots, \theta^p)$.

The final values selected for the ambiguity parameters $\theta^* = (\theta^{1,*}, \ldots, \theta^{p,*})$ are those that maximize this performance metric; i.e.,

$$\theta^* = \arg\max_{\theta} \psi_{\text{ave}}(\theta). \tag{50}$$

### 6.3. Simulation Experiments

Suppose at each stage the decision maker can choose from one of $p = 3$ arms. The decision maker models the reward distributions according to the gamma-exponential conjugate pair described in the previous subsection. We will consider the case in which the decision maker initially does not know which arm has the highest mean before data has been collected. To do this, we choose equal (unit) values for the shape and rate parameters of the gamma prior (see Table 3), which is natural choice.

We consider three cases for the underlying reward distributions: (1) correctly specified independent arms, (2) misspecified independent arms, and (3) misspecified dependent arms. In all the experiments, a discount factor of $\alpha = 0.55$ was used.

*Case* 1: *Correctly Specified Independent Arms*. We first consider the case in which the decision maker's nominal assumptions are correct. That is, the true underlying reward distributions follow independent exponential distribution (one for each arm), with mean parameters given in Table 4.

We now describe the experiment. Each iteration of the experiment begins by sampling 10 historical rewards $\{d_1^i, \ldots, d_{10}^i\}$ from the exponential distributions given in Table 4. With this data, we next use the $k$-fold cross validation with a partition parameter of $k = 2$, as outlined in §6.2, to determine the optimal ambiguity parameter values $\theta^* = (\theta^{1,*}, \theta^{2,*}, \theta^{3,*})$ for this particular data set. Fixing these values, the out-of-sample performance of the classical and robust Gittins index policies were then computed by using Monte Carlo simulation, with rewards generated from the true data-generating model from Table 4. This gives

**Table 3** **Nominal Gamma Prior Parameters**

| Arm $i$ | $\alpha_i$ (shape) | $\beta_i$ (rate) |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 1 | 1 |
| 3 | 1 | 1 |

**Table 4** **Underlying Exponential Parameters (Data-Generating Model)**

| Arm $i$ | Mean ($\lambda_i$) |
|---|---|
| 1 | 120 |
| 2 | 60 |
| 3 | 80 |

**Figure 3** (Color online) Case 1: Robust vs. Nominal Gittins Index Policy Performance

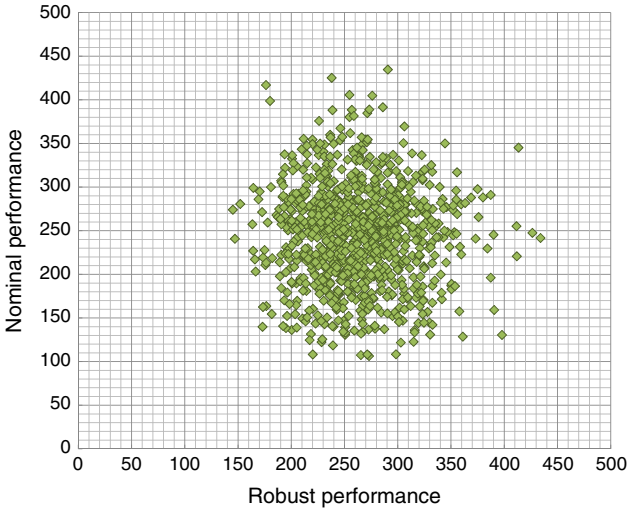



**Table 5** Case 1: Out-of-Sample Performance

| Policy | Average performance | Sample standard deviation |
|---|---|---|
| Robust Gittins index policy | 263.06 | 44.41 |
| Nominal Gittins index policy | 244.59 | 55.47 |

an estimate of the expected discounted reward for the classical and robust Gittins policies using the data sets $\{d_1^i, \ldots, d_{10}^i\}$ generated at the start of the experiment. This experiment was repeated 1,000 times (with a newly generated data set for each iteration), and a scatter plot of the rewards is shown in Figure 3. The aggregate results are reported in Table 5.

It is interesting to note that, although the performance of the two policies are comparable, the robust policy reduces the out-of-sample sample standard deviation of rewards by approximately 20%. More generally, rewards generated by the robust Gittins index in other tests (not reported) were typically less variable, but with comparable or better expected value, than those generated by the classical Gittins index policy. This finding is even more pronounced in the presence of model misspecification, as we will see in the following two cases.

*Case* 2: *Misspecified Independent Arms*. We next consider the case in which the underlying reward distributions are misspecified, but still remain independent across arms. In particular, the true underlying reward distributions now follow independent truncated normal distributions (one for each arm), with mean and standard deviation parameters given in Table 6.

That is, if arm $i = 1, 2, 3$ is chosen, i.i.d. rewards $d_n^i$ are generated from $\max\{0, \mathcal{N}(\mu_i, \sigma_i)\}$, where $\mathcal{N}(\mu_i, \sigma_i)$ is a normal distribution with mean $\mu_i$ and standard deviation $\sigma_i$. Obviously, one could choose a far more complex underlying model (e.g., mixture distributions, nonstationarities, etc.), but even in this simple controlled setting we are able to show the impact of model family misspecification. The same experiment described above was run 1,000 times, and a scatter plot of the results is shown in Figure 4, with aggregate results in Table 7.

**Table 6** Underlying Normal Parameters (Data-Generating Model)

| Arm $i$ | Mean ($\mu_i$) | Standard deviation ($\sigma_i$) | Coeff. of variation |
|---|---|---|---|
| 1 | 120 | 10 | 0.0833 |
| 2 | 60 | 100 | 1.6667 |
| 3 | 80 | 80 | 1.0000 |

**Figure 4** (Color online) Case 2: Robust vs. Nominal Gittins Index Policy Performance

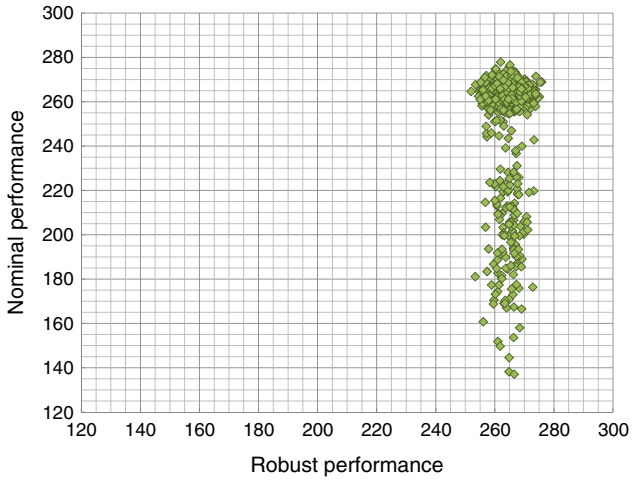



Table 7 shows that the robust Gittins index policy provides a modest 3.43% improvement over the nominal Gittins index policy. What is more striking is substantial reduction in variability and downside risk in the out-of-sample performance when using the robust Gittins index policy, as shown in Figure 4. In a follow-up paper, we are developing a theoretical explanation as to why the robust policy leads to a reduction in variability in out-of-sample performance.

*Case* 3: *Misspecified Dependent Arms*. We last consider the case of misspecification in the underlying reward distributions with dependent arms. In particular, the true underlying reward distributions follow an $r$-lag vectorautoregressive (VAR($r$)) model. That is,

**Table 7** Case 2: Out-of-Sample Performance

| Policy | Average performance | Sample standard deviation |
|---|---|---|
| Robust Gittins index policy | 264.49 | 3.82 |
| Nominal Gittins index policy | 255.42 | 23.91 |

the demand vector $D_n = (d_n^1, d_n^2, d_n^3) \in \mathbb{R}^3$ at stage $n$ satisfies

$$D_n = \sum_{l=1}^{r} \Phi_l \cdot D_{n-l} + \epsilon_n, \tag{51}$$

where $\Phi_l \in \mathbb{R}^{3\times 3}$, $l = 1, \ldots, r$, are the coefficient matrices, and $\epsilon_n \in \mathbb{R}^3$ are i.i.d. error terms that follow the same truncated normal distributions given in Table 6.

It is important to point out that in a bandit setting, if the underlying data-generating mechanism is a vector time series as in (51), model misspecification is unavoidable and therefore a genuine concern. This is because even if the decision maker knows the underlying model structure (which is not typically the case), at each stage only a single component of the time series vector $D_n = (d_n^1, d_n^2, d_n^3)$ is observable. Consequently, the decision maker can at best only model the marginal reward distributions, because learning the correlation structure is not possible.

For simplicity, we will consider a VAR(1) model, with the coefficient matrix

$$\Phi_1 = \begin{pmatrix} -0.5 & 0.5 & 0 \\ 0 & -0.5 & 0.5 \\ 0.5 & 0 & -0.5 \end{pmatrix}. \tag{52}$$

Again, the same experiment described above was run 1,000 times, and a scatter plot of the results is shown in Figure 5, with aggregate results in Table 8.

Here, we again see that the robust Gittins index policy provides a modest improvement over the nominal Gittins index policy in terms of expected out-of-sample performance (8.11%), while providing a significant improvement in terms of variability and downside-risk reductions. It is interesting to note that the benefits of the robust Gittins index policy seem to be most pronounced when moving from the correctly specified independent setting (Case 1) to the misspecified independent setting (Case 2). This seems to suggest that even slight model misspecification should not be taken lightly.

**Figure 5** (Color online) Case 3: Robust vs. Nominal Gittins Index Policy Performance

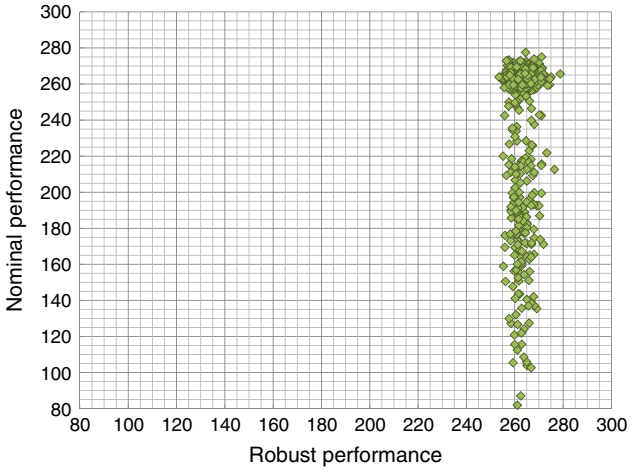



**Table 8** Case 3: Out-of-Sample Performance

| Policy | Average performance | Sample standard deviation |
|---|---|---|
| Robust Gittins index policy | 263.67 | 3.90 |
| Nominal Gittins index policy | 242.28 | 41.04 |

### 6.4. Discussion

Our experiments show that the robust Gittins policy can substantially outperform the classical Gittins policy, even with relatively mild violations of the nominal assumptions (Case 2), with the main performance advantage being a reward distribution with substantially smaller downside risk and higher expected reward. The relative improvement of the robust policy over the nominal increases as the number of assumption violations increases (Case 3).

There are limitations in our simulation study that we would like to acknowledge. First, although our example considers the simplest web design optimization problem, there is increasing interest in whether performance can be further improved by incorporating personalized information about customers and "customizing" the pages that are shown. Although problems of this nature are important, they bring substantial computational challenges since the dimension of each state is now higher. We have chosen to ignore these issues in our simulation study because we wish to clearly evaluate the impact of misspecification and the value of accounting for robustness through the robust Gittins policy. Studying higher-dimensional problems would require heuristics, which would confound our understanding of whether differences in performance between the robust and Gittins policy are due to the approximations or our method of accounting for robustness, and we defer study of these problems to the future.

Second, it would also be interesting to compare the performance of the robust and classical Gittins policies in a live experiment, where responses to designs come from actual customers. In this paper, we limit ourselves to a simulation study because it gives us control over the data-generating process and allows us to evaluate the impact of progressive violations of the assumptions of the classical bandit model. Although the performance of the robust Gittins policy in our experiments—in particular, the higher expected reward at reduced downside risk—gives us reason to be hopeful that its out-of-sample performance will be superior to that of the classical Gittins policy, live

experiments are needed to test this hypothesis. We plan to pursue this direction of work in the future in the context of a more comprehensive study.

## 7. Concluding Remarks

In this paper, we studied a robust formulation of the multiarmed bandit problem. The model allows the decision maker to account for different levels of distrust for different parts of the model, and the goal is to find the scheduling policy that optimizes the worst-case expected reward when model uncertainty is modeled by using the notion of relative entropy. Analysis of the robust version of the dynamic programming equation establishes the optimality of a so-called project-by-project retirement (PPR) policy. However, it is also shown that projects are no longer independent under nature's worst-case response, even if they are independent under the nominal model, implying that index policies are no longer optimal for the robust problem. This leads us to study a robust version of the classical Gittins index policy. Estimates of the robust value function and the notion of information relaxation are used to derive performance bounds that suggest that the robust Gittins policy is close to optimal. We also compare the performance of the robust Gittins policy to that of the classical Gittins index policy for a misspecified Bayesian webpage design problem. Our experiment suggests that the robust policy can achieve improved expected profit together with a reduction in its variability when compared to the classical Gittins index.

### Acknowledgments

The authors acknowledge feedback from the participants at the "Optimization Under Uncertainty" Workshop (2013) hosted by the Institute for Mathematical Sciences at the National University of Singapore, and the applied probability cluster at the INFORMS Annual Meeting (2013). The authors also thank Huaning Cai, Jeremy Chen, Shea Chen, Dan Iancu, Daniel Kuhn, Gah-Yi Vahn, Tong Wang, Wolfram Wiesemann, the two referees, the associate editor, and the department editor for their helpful comments and feedback. A special thanks goes to Poomos Wimonkittiwat, whose unassuming comment devastated an earlier version of the paper; its fix takes the current form of §5.3. This research was partially supported by a National Sciences and Engineering Research Council of Canada Postdoctoral Fellowship (NSERC PDF; to M. J. Kim) and the National Science Foundation [Grants CMMI-1031637 and CMMI-1201085]. The opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

### Appendix A. Proof of Proposition 3

The proof follows mathematical induction. The base case $V_0(x) = \max\{M, 0\}$, clearly satisfies the properties. Assume now that for some $k$, $V_k(x)$ satisfies the properties. We show that

$$V_{k+1}(x) = \max\Big\{M, \max_{i=1,\dots,p} H^i(x, V_k)\Big\}$$

also satisfies the properties. Since it is assumed that $V_k(x)$ is nondecreasing in $M$, it follows immediately that $V_{k+1}(x)$ is also nondecreasing in $M$. Assume now that $M \le -B/(1-\alpha)$. Then, by the induction hypothesis, $V_k(x) = \beta_k$ for some constant $\beta_k$. Then,

$$V_{k+1}(x) = \max\Big\{M, \alpha\beta_k + \max_{i=1,\dots,p} r^i(x^i)\Big\}.$$

Since

$$\begin{aligned}\alpha\beta_k + \max_{i=1,\dots,p} r^i(x^i) &\ge \alpha\beta_k - B\\ &\ge \alpha\beta_k - M(1-\alpha)\\ &\ge \alpha M - M(1-\alpha)\\ &= M,\end{aligned}$$

it follows that $V_{k+1}(x) = \alpha\beta_k + \max_i r^i(x^i)$, which is constant. Finally, suppose $M \ge B/(1-\alpha)$. Then, by the induction hypothesis, $V_k(x) = M$, and

$$V_{k+1}(x) = \max\Big\{M, \alpha M + \max_{i=1,\dots,p} r^i(x^i)\Big\}.$$

Since

$$\alpha M + \max_{i=1,\dots,p} r^i(x^i) \le \alpha M + B \le \alpha M + M(1-\alpha) = M,$$

it follows that $V_{k+1}(x) = M$. Thus, it follows by mathematical induction that $V_k(x)$ satisfies the properties for all $k$. Since Theorem 2 implies that the value function can be obtained as the limit $V(x) = \lim_{k\to\infty} V_k(x)$, it follows that $V(x)$ also satisfies the properties, which completes the proof. □

### Appendix B. Proof of Proposition 4

To show the existence of an optimal PPR policy, it suffices to prove the following inequality. For any $x = (x^1, \dots, x^p)$ and $i = 1, \dots, p$,

$$V^{i-}(x^{i-}) \le V(x) \le V^{i-}(x^{i-}) + (V^i(x^i) - M). \tag{B1}$$

The first inequality in (B1) clearly holds. The proof of the second inequality follows mathematical induction. The base case $V_0(x) = V_0^{i-}(x) = V_0^i(x) = \max\{M, 0\}$, clearly satisfies the inequality. Assume now that, for some $k$,

$$V_k(x) \le V_k^{i-}(x^{i-}) + (V_k^i(x^i) - M). \tag{B2}$$

We wish to show that the inequality holds for $V_{k+1}(x)$. Applying the above induction hypothesis, it follows that

$$\begin{aligned}&V_{k+1}(x)\\ &= \max\left\{M, \max_{j=1,\dots,p} r^j(x^j) - \theta^j \log \mathbb{E}_{\rho^j}\left[\exp\left(-\frac{\alpha V_k(x^{j-}, X^j)}{\theta^j}\right)\right]\right\}\\ &\le \max\Big\{M, r^i(x^i) + \alpha(V_k^{i-}(x^{i-}) - M)\\ &\qquad - \theta^i \log \mathbb{E}_{\rho^i}\left[\exp\left(-\frac{\alpha V_k^i(X^i)}{\theta^i}\right)\right],\end{aligned}$$

$$\max_{j\neq i}\left\{r^j(x^j)+\alpha(V_k^i(x^i)-M)\right. \tag{B3}$$
$$\left.-\theta^j\log\mathbb{E}_{\rho^j}\left[\exp\left(-\frac{\alpha V_k^{i-}(x^{j-},X^j)}{\theta^j}\right)\right]\right\}\Bigg\}$$
$$\le\max\left\{\max\left\{M,\, r^i(x^i)-\theta^i\log\mathbb{E}_{\rho^i}\left[\exp\left(-\frac{\alpha V_k^i(X^i)}{\theta^i}\right)\right]\right\}\right.$$
$$+\alpha(V_k^{i-}(x^{i-})-M),$$
$$\left.\max\left\{M,\max_{j\neq i}r^j(x^j)-\theta^j\log\mathbb{E}_{\rho^j}\left[\exp\left(-\frac{\alpha V_k^{i-}(x^{j-},X^j)}{\theta^j}\right)\right]\right\}\right.$$
$$\left.+\alpha(V_k^i(x^i)-M)\right\},$$

where the second inequality follows because of $V_k^i(x^i)$, $V_k^{i-}(x^{i-})\ge M$. Finally, recognizing that

$$V_{k+1}^i(x^i)=\max\left\{M,\, r^i(x^i)-\theta^i\log\mathbb{E}_{\rho^i}\left[\exp\left(-\frac{\alpha V_k^i(X^i)}{\theta^i}\right)\right]\right\},$$
$$V_{k+1}^{i-}(x^{i-})=\max\left\{M,\max_{j\neq i}r^j(x^j)\right.$$
$$\left.-\theta^j\log\mathbb{E}_{\rho^j}\left[\exp\left(-\frac{\alpha V_k^{i-}(x^{j-},X^j)}{\theta^j}\right)\right]\right\},$$

we see that the right-hand side of the above inequality (B3) is equal to

$$\begin{aligned}&\max\{V_{k+1}^i(x^i)+\alpha(V_k^{i-}(x^{i-})-M),\, V_{k+1}^{i-}(x^{i-})+\alpha(V_k^i(x^i)-M)\}\\&\le\max\{V_{k+1}^i(x^i)+\alpha(V_{k+1}^{i-}(x^{i-})-M),\, V_{k+1}^{i-}(x^{i-})\\&\qquad+\alpha(V_{k+1}^i(x^i)-M)\}\\&\le\max\{V_{k+1}^i(x^i)+V_{k+1}^{i-}(x^{i-})-M,\, V_{k+1}^{i-}(x^{i-})+V_{k+1}^i(x^i)-M\}\\&=V_{k+1}^{i-}(x^{i-})+(V_{k+1}^i(x^i)-M).\end{aligned}$$

Thus, it follows by mathematical induction that $V_k(x)$ satisfies inequality (B2) for all $k$. Since Theorem 2 implies that the value function can be obtained as the limit $V(x)=\lim_{k\to\infty}V_k(x)$, it follows that $V(x)$ satisfies inequality (B1), which completes the proof. □

## Appendix C. Proof of Lemma 1

We first show that for any $\mathbf{y}\in\mathcal{Y}$, the penalty function $\lambda(h,\mathbf{y})$ satisfies

$$\mathbb{E}_{\mathbb{P}_G}[\lambda(h,\mathbf{y})\mid q_0,x_0]=0.$$

For fixed constant $k\ge 0$, define

$$\lambda^k(h,\mathbf{y})\triangleq\sum_{n=0}^{(\tau-1)\wedge k}\alpha^{n+1}q_n\left(y_n(x_{n+1}^{i_n})h(i_n,x_{n+1}^{i_n})\right.$$
$$\left.-\sum_{j=1}^{N}y(j)h(i_n,j)\rho^{i_n}(x_n^{i_n},j)\right),$$

where $\wedge$ is the "min" operator. Using straightforward conditioning arguments, it is easy to show that for each $k\ge 0$,

$$\mathbb{E}_{\mathbb{P}_G}[\lambda^k(h,\mathbf{y})\mid q_0,x_0]=0.$$

Then, by Assumption 1, the bounded convergence theorem implies

$$\mathbb{E}_{\mathbb{P}_G}[\lambda(h,\mathbf{y})\mid q_0,x_0]=\lim_{k\to\infty}\mathbb{E}_{\mathbb{P}_G}[\lambda^k(h,\mathbf{y})\mid q_0,x_0]=0.$$

The weak duality result follows since

$$\begin{aligned}W(q_0,x_0)&=\inf_{\mathbf{y}\in\mathcal{Y}}\mathbb{E}_{\mathbb{P}_G}[P(\mathbf{y})\mid q_0,x_0]\\&=\inf_{\mathbf{y}\in\mathcal{Y}}\mathbb{E}_{\mathbb{P}_G}[P(\mathbf{y})-\lambda(h,\mathbf{y})\mid q_0,x_0]\\&\ge\inf_{\mathbf{y}\in\bar{\mathcal{Y}}}\mathbb{E}_{\mathbb{P}_G}[P(\mathbf{y})-\lambda(h,\mathbf{y})\mid q_0,x_0]\\&=\mathbb{E}_{\mathbb{P}_G}\left[\inf_{\mathbf{y}\in\bar{\mathcal{Y}}}P(\mathbf{y})-\lambda(h,\mathbf{y})\,\middle|\,q_0,x_0\right]\\&=:V_L^\lambda(x_0),\end{aligned}$$

which completes the proof. □

## Appendix D. Proof of Theorem 3

The dynamic programming equations for the determinist problem (43) are

$$W_n^\lambda(q)=\inf_y\left\{q\cdot(r^{i_n}(x_n^{i_n})+\theta^{i_n}\sum_{j\in\mathcal{X}}y(j)\log(y(j))\rho^{i_n}(x_n^{i_n},j))\right.$$
$$-\alpha q\cdot(y(x_{n+1}^{i_n})h(i_n,x_{n+1}^{i_n})$$
$$-\sum_{j\in\mathcal{X}}y(j)h(i_n,j)\rho^{i_n}(x_n^{i_n},j))$$
$$\left.+\alpha W_{n+1}^\lambda(q\cdot y(x_{n+1}^{i_n}))\right\}, \tag{D1}$$
$$W_\tau^\lambda(q_\tau)=q_\tau\cdot M,$$

where at each stage $n=\tau-1,\dots,0$, $y=\{y(j),\,j\in\mathcal{X}\}$ is constrained to being nonnegative and to satisfy

$$\mathbb{E}_{\rho^{i_n}(x_n^{i_n},\cdot)}[y_n]=\sum_{j\in\mathcal{X}}y(j)\rho^{i_n}(x_n^{i_n},j)=1. \tag{D2}$$

Substituting $W_n^\lambda(q)=qf_n^\lambda$, we obtain

$$f_n^\lambda=\inf_y\left\{r^{i_n}(x_n^{i_n})+\theta^{i_n}\sum_{j\in\mathcal{X}}y(j)\log(y(j))\rho^{i_n}(x_n^{i_n},j)\right.$$
$$-\alpha(y(x_{n+1}^{i_n})h(i_n,x_{n+1}^{i_n})$$
$$\left.-\sum_{j\in\mathcal{X}}y(j)h(i_n,j)\rho^{i_n}(x_n^{i_n},j))+\alpha y(x_{n+1}^{i_n})f_{n+1}^\lambda\right\}, \tag{D3}$$
$$f_\tau^\lambda=M.$$

The optimization problem in (D3) is

$$\min_y\Phi(y),$$

subject to the nonnegativity constraint $y(j)\ge 0$ for all $j\in\mathcal{X}$ and (D2), where

$$\Phi(y)=r^{i_n}(x_n^{i_n})+\theta^{i_n}\sum_{j\in\mathcal{X}}y(j)\log(y(j))\rho^{i_n}(x_n^{i_n},j)$$
$$-\alpha\left(y(x_{n+1}^{i_n})h(i_n,x_{n+1}^{i_n})-\sum_{j\in\mathcal{X}}y(j)h(i_n,j)\rho^{i_n}(x_n^{i_n},j)\right)$$
$$+\alpha y(x_{n+1}^{i_n})f_{n+1}^\lambda.$$

Observing that $\Phi(y)$ is convex, we can compute the optimal $y$ using standard first-order methods. Specifically, the Lagrangian is given by

$$\Lambda(y,\mu)=\Phi(y)+\mu\left(\sum_{j\in\mathcal{X}}y(j)\rho^{i_n}(x_n^{i_n},j)-1\right),$$

where $\mu \in \mathbb{R}$ is the Lagrangian multiplier. Observing that

$$\frac{\partial \Lambda}{\partial y(j)}(y, \mu) = \theta^{i_n}\rho^{i_n}(x_n^{i_n}, j)\log y(j) + \theta^{i_n}\rho^{i_n}(x_n^{i_n}, j) + \alpha h(i_n, j)\rho^{i_n}(x_n^{i_n}, j) + \mu\rho^{i_n}(x_n^{i_n}, j), \quad j \neq x_{n+1}^{i_n},$$

$$\frac{\partial \Lambda}{\partial y(x_{n+1}^{i_n})}(y, \mu) = \theta^{i_n}\rho^{i_n}(x_n^{i_n}, x_{n+1}^{i_n})\log y(x_{n+1}^{i_n}) + \theta^{i_n}\rho^{i_n}(x_n^{i_n}, x_{n+1}^{i_n}) + \alpha h(i_n, x_{n+1}^{i_n})\rho^{i_n}(x_n^{i_n}, x_{n+1}^{i_n}) - \alpha h(i_n, x_{n+1}^{i_n}) + \alpha f_{n+1}^{\lambda} + \mu\rho^{i_n}(x_n^{i_n}, x_{n+1}^{i_n}),$$

we see that the unique stationary points are

$$y^*(j) = \exp\left(-\frac{\alpha}{\theta^{i_n}}h(i_n, j) - \frac{\mu}{\theta^{i_n}} - 1\right)$$

$$y^*(x_{n+1}^{i_n}) = \exp\left(-\frac{\alpha}{\theta^{i_n}}\left[\frac{f_{n+1}^{\lambda} - h(i_n, x_{n+1}^{i_n})(1 - \rho^{i_n}(x_n^{i_n}, x_{n+1}^{i_n}))}{\rho^{i_n}(x_n^{i_n}, x_{n+1}^{i_n})}\right] - \frac{\mu}{\theta^{i_n}} - 1\right).$$

Choosing $\mu$ to satisfy the constraint (D2) gives the expression (45) for nature's optimal choice at stage $n$. Substituting back into the objective function $\Phi(y)$ gives the recursion (44). □

## References


Agrawal R (1995) The continuum-armed bandit problem. *SIAM J. Control Optim.* 33(6):1926–1951.

Auer P, Cesa-Bianchi N, Fischer P (2002a) Finite-time analysis of the multiarmed bandit problem. *Machine Learn.* 47(2–3): 235–256.

Auer P, Cesa-Bianchi N, Freund Y, Schapire RE (1995) Gambling in a rigged casino: The adversarial multiarmed bandit problem. *Proc. 6th Annual IEEE Sympos. Foundations Comput. Sci.* (IEEE Computer Society Press, Washington, DC), 322–331.

Auer P, Cesa-Bianchi N, Freund Y, Schapire RE (2002b) The nonstochastic multiarmed bandit problem. *SIAM J. Comput.* 32(1):48–77.

Babaioff M, Sharma Y, Slivkins A (2014) Characterizing truthful multi-armed bandit mechanisms. *SIAM J. Comput.* 43(1): 194–230.

Ben-Tal A, Nemirovski A (1998) Robust convex optimization. *Math. Oper. Res.* 23(4):769–805.

Ben-Tal A, Nemirovski A (1999) Robust solutions to uncertain programs. *Oper. Res. Lett.* 25(1):1–13.

Ben-Tal A, Nemirovski A (2000) Robust solutions of linear programming problems contaminated with uncertain data. *Math. Programming* 88(3):411–424.

Bertsekas DP (1995) *Dynamic Programming and Optimal Control Volume II* (Athena Scientific, Belmont, MA).

Bertsimas D, Nino-Mora J (1996) Conservation laws, extended polymatroids and multiarmed bandit problems: A polyhedral approach to indexable systems. *Math. Oper. Res.* 21(2):257–306.

Bertsimas D, Sim M (2004) The price of robustness. *Oper. Res.* 52(1):35–53.

Brown DB, Smith JE (2013) Optimal sequential exploration: Bandits, clairvoyants, and wildcats. *Oper. Res.* 61(3):644–665.

Brown DB, Smith JE, Sun P (2010) Information relaxations and duality in stochastic dynamic programs. *Oper. Res.* 58(4): 785–801.

Caro F, Gallien J (2007) Dynamic assortment with demand learning for seasonal consumer goods. *Management Sci.* 53(2):276–292.

Caro F, Gupta AD (2015) Robust control of the multi-armed bandit problem. *Ann. Oper. Res.*, ePub ahead of print August 21, http://link.springer.com/article/10.1007%2Fs10479-015-1965-7.

Cesa-Bianchi N, Lugosi G (2006) *Prediction, Learning, and Games* (Cambridge University Press, Cambridge, UK).

Chan CW, Farias VF (2009) Stochastic depletion problems: Effective myopic policies for a class of dynamic optimization problems. *Math. Oper. Res.* 34(2):333–350.

Dai Pra P, Meneghini L, Runggaldier WJ (1996) Connections between stochastic control and dynamic games. *Math. Control Signals Systems* 9(4):303–326.

Dupuis P, Ellis RS (1997) *A Weak Convergence Approach to the Theory of Large Deviations* (Wiley, New York).

Efron B (1979) Bootstrap methods: Another look at the jackknife. *Ann. Statist.* 7(1):1–235.

El Ghaoui L, Lebret H (1997) Robust solutions to least-square problems to uncertain data matrices. *SIAM J. Matrix Anal. Appl.* 18(4):1035–1064.

Epstein LG, Schneider M (2003) Recursive multiple-priors. *J. Econom. Theory* 113(1):1–31.

Epstein LG, Schneider M (2007) Learning under ambiguity. *Rev. Econom. Stud.* 74(4):1275–1303.

Gittins JC (1979) Bandit processes and dynamic allocation indices. *J. Royal Statist. Soc. Ser. B* 41(2):148–164.

González-Trejo JI, Hernández-Lerma O, Hoyos-Reyes LF (2003) Minimax control of discrete-time stochastic systems. *SIAM J. Control Optim.* 41(5):1626–1659.

Hansen LP, Sargent TJ (2005) Robust estimation and control under commitment. *J. Econom. Theory* 124(2):258–301.

Hansen LP, Sargent TJ (2007) Robust estimation and control without commitment. *J. Econom. Theory* 136(1):1–27.

Hansen LP, Sargent TJ (2008) *Robustness* (Princeton University Press, Princeton, NJ).

Haugh MB, Lim AEB (2012) Linear-quadratic control and information relaxations. *Oper. Res. Lett.* 40(6):521–528.

Iyengar GN (2005) Robust dynamic programming. *Math. Oper. Res.* 30(2):257–280.

Jacobson DH (1973) Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Trans. Automatic Control* 18(2): 124–131.

Kleinberg R (2004) Nearly tight bounds for the continuum-armed bandit problem. Jordan MI, LeCun Y, Solla SA, eds.*Advances in Neural Information Processing Systems 17* (MIT Press, Cambridge, MA), 697–704.

Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* 6(1):4–22.

Li JY, Kwon RH (2013) Portfolio selection under model uncertainty: A penalized moment-based optimization approach. *J. Global Optim.* 56(1):131–164.

Lim AEB, Shanthikumar JG (2007) Relative entropy, exponential utility, and robust dynamic pricing. *Oper. Res.* 55(2):198–214.

Lim AEB, Shanthikumar JG, Shen ZJ Max (2006) Model uncertainty, robust optimization, and learning. Johnson MP, Norman B, Secomandi N, eds. *2006 TutORials Oper. Res.* (INFORMS, Catonsville, MD), 66–94.

Lim AEB, Shanthikumar JG, Vahn G-Y (2012) Robust portfolio choice with learning in the framework of regret: Single-period case. *Management Sci.* 58(9):1732–1746.

Lim AEB, Shanthikumar JG, Watewai T (2011). Robust asset allocation with benchmarked objectives. *Math. Finance* 21(4): 643–679.

Mersereau AJ, Rusmevichientong P, Tsitsiklis JN (2009) A structured multiarmed bandit problem and the greedy policy. *IEEE Trans. Automatic Control* 54(12):2787–2802.

Nilim A, El Ghaoui L (2005) Robust control of Markov decision processes with uncertain transition matrices. *Oper. Res.* 53(5): 780–798.

Niño-Mora J (2012) Towards minimum loss job routing to parallel heterogeneous multiserver queues via index policies. *Eur. J. Oper. Res.* 220(3):705–715.

Pandey S, Chakrabarti D, Agarwal D (2007) Multi-armed bandit problems with dependent arms. *Proc. 24th Internat. Conf. Machine Learn.* (ACM, New York), 721–728.

Petersen IR, James MR, Dupuis P (2000) Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Trans. Automatic Control* 45(3):398–412.

Robbins H (1952) Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* 55:527–535.

Rogers LCG (2007) Pathwise stochastic optimal control. *SIAM J. Control Optim.* 46(3):1116–1132.

Rusmevichientong P, Shen ZJM, Shmoys DB (2010) Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Oper. Res.* 58(6):1666–1680.

Ryzhov IO, Powell WB, Frazier PI (2012) The knowledge gradient algorithm for a general class of online learning problems. *Oper. Res.* 60(1):180–195.

Scott SL (2010) A modern Bayesian look at the multi-armed bandit. *Appl. Stochastic Models Bus. Indust.* 26(6):639–658.

Scott SL (2013) Multi-armed bandit experiments. Accessed August 15, 2013, https://support.google.com/analytics/answer/2844870?hl=en.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc. Ser. B* 58(1):267–288.

Tsitsiklis JN (1986) A lemma on the multi-armed bandit problem. *IEEE Trans. Automatic Control* 31(6):576–577.

White JM (2013) *Bandit Algorithms for Website Optimization: Developing, Deploying, Debugging* (O'Reilly Media, Sebastopol, CA).

Whittle P (1980) Multi-armed bandits and the Gittins index. *J. Royal Statist. Soc. Ser. B* 42(2):143–149.

Whittle P (1981) Risk-sensitive linear/quadratic/Gaussian control. *Adv. Appl. Probab.* 13(4):764–777.

Whittle P (1990) A risk-sensitive maximum principle. *Systems Control Lett.* 15(3):183–192.

Whittle P (1991) A risk-sensitive maximum principle: The case of imperfect state observations. *IEEE Trans. Automatic Control* 36(7):793–801.

Wiesemann W, Kuhn D, Rustem B (2013) Robust Markov decision processes. *Math. Oper. Res.* 38(1):153–183.

Ye F, Zhou E (2015) Information relaxation and dual formulation of controlled Markov diffusions. *IEEE Trans. Automatic Control* Forthcoming.