Operations Research

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

Technical Note—A Note on the Equivalence of Upper Confidence Bounds and Gittins Indices for Patient Agents

Daniel Russo

To cite this article:
Daniel Russo (2021) Technical Note—A Note on the Equivalence of Upper Confidence Bounds and Gittins Indices for Patient Agents. Operations Research 69(1):273-278. https://doi.org/10.1287/opre.2020.1987

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 69, No. 1, January–February 2021, pp. 273–278
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Technical Note—A Note on the Equivalence of Upper Confidence Bounds and Gittins Indices for Patient Agents

Daniel Russo[a]

[a] Graduate School of Business, Columbia University, New York, New York 10027
Contact: djr2174@gsb.columbia.edu, https://orcid.org/0000-0001-5926-8624 (DR)

Received: March 29, 2019
Accepted: November 9, 2019
Published Online in Articles in Advance: December 24, 2020

Subject Classifications: decision analysis: sequential; dynamic programming
Area of Review: Decision Analysis

https://doi.org/10.1287/opre.2020.1987

Abstract. This note gives a short, self-contained proof of a sharp connection between Gittins indices and Bayesian upper confidence bound algorithms. I consider a Gaussian multiarmed bandit problem with discount factor γ. The Gittins index of an arm is shown to equal the γ-quantile of the posterior distribution of the arm's mean plus an error term that vanishes as $\gamma \to 1$. In this sense, for sufficiently patient agents, a Gittins index measures the highest plausible mean-reward of an arm in a manner equivalent to an upper confidence bound.

Keywords: Gittins index • upper confidence bound • multiarmed bandits

1. Introduction and Related Work

There are two separate segments of the multiarmed bandit literature. One formulates a Bayesian multiarmed bandit problem as a Markov decision process and uses tools from dynamic programming to compute or approximate the optimal policy. This literature builds on a beautiful result that shows an optimal policy selects in each period the arm with the highest Gittins index (Gittins 1974, Gittins and Jones 1979). A second segment of the literature focuses on simple heuristic algorithms—which are often easy to adapt to settings in which exact dynamic programming is computationally intractable—and studies their performance through simulation and theoretical bounds on their regret (Rusmevichientong and Tsitsiklis 2010, Kaufmann et al. 2012, Srinivas et al. 2012, Cappé et al. 2013). This literature descends from a seminal paper by Lai and Robbins (1985) that shows the asymptotic growth rate of expected regret in a frequentist model is minimized by selecting in each period the arm with greatest upper confidence bound.

A sharp relationship between upper confidence bounds and the Gittins index of a patient agent (whose discount factor is close to 1) helps to unify these two segments of the literature. This provides enormous conceptual clarity, allowing the upper confidence bounds of Lai and Robbins (1985) to be seen roughly as a generalization of and asymptotic approximation to the Gittins index. Unfortunately, such links seem to be known only to a few expert researchers. The goal of this short note is twofold. First, for Gaussian multiarmed bandit problems, it states an asymptotic equivalence between the Gittins index and a Bayesian upper confidence bound in a transparent form absent from the current literature. Second, the note gives short and elementary (if somewhat ugly) proofs that hopefully make this material accessible to a large audience of researchers.

Asymptotic links between Gittins indices and upper confidence bounds were first recognized by Chang and Lai (1987). That paper presents a sophisticated asymptotic expansion of the solution of diffusion approximations to the optimal stopping problems defining a Gittins index. Unfortunately, the analysis is highly complex and is inaccessible to most multiarmed bandit researchers. Perhaps as a result, this pioneering work appears not to be widely known or cited.[1] Hopefully, the transparent form of Theorem 1 along with its short proof will help to remedy this.

Like Chang and Lai (1987), most other closely related papers focus on developing approximations to the Gittins index with the goal of simplifying computation (Yao 2006, Chick and Gans 2009, Gutin and Farias 2016). This note was highly influenced by my reading of Gutin and Farias (2016). The upper bound on the Gittins index developed in Section 3.3 comes from analyzing their algorithm. Also related is work by Burnetas and Katehakis (2003), Kaufmann et al. (2012), and Lattimore (2016), who study the regret of Gittins index-like policies for finite-horizon undiscounted multiarmed bandit problems. See also Niño-Mora (2011) for a derivation of a finite-horizon

approximation of the Gittins index and related computational issues. This short note is distinguished from these related works in that (1) I study the Gittins index as classically defined rather than the heuristic of Niño-Mora (2011), and (2) this note is designed to develop conceptual insight through a sharp link between Gittins indices and Bayesian upper confidence bounds, rather than to develop accurate computational approximations or give a frequentist regret analysis.

2. Formulation and Main Result

Although the Gittins index is eventually used in multiarmed bandit problems, it is calculated by considering a modified one-armed bandit problem. Consider a single arm with uncertain quality θ. When played at time t, the arm generates a reward R_t with $R_t|\theta \sim N(\theta, \sigma_W^2)$. The posterior distribution of θ given observed rewards $R_0, \ldots R_{t-1}$ is Gaussian. We write $\theta|R_0 \ldots R_{t-1} \sim N(\mu_t, \sigma_t^2)$, where the posterior parameters evolve according to

$$\mu_t = \frac{\sigma_{t-1}^{-2}\mu_{t-1} + \sigma_W^{-2}R_{t-1}}{\sigma_{t-1}^{-2} + \sigma_W^{-2}}$$

and

$$\sigma_t^2 = \left(\frac{1}{\sigma_{t-1}^2} + \frac{1}{\sigma_W^2}\right)^{-1} = \left(\frac{1}{\sigma_0^2} + \frac{1}{t\sigma_W^2}\right)^{-1}. \tag{1}$$

To define the Gittins index, we follow the interpretation of Weber (1992). Imagine that the right to play this arm is restricted and for each play the decision maker must pay a tax λ. Alternatively, in any period the agent may choose to retire and earn a reward of 0 thereafter. This can be cast as a Markov decision process (MDP) where the state at time t is (μ_t, σ_t^2), which serves as a sufficient statistic for the agent's posterior belief. The expected reward when playing the arm at state (μ_t, σ_t^2) is μ_t. The agent's actions are simple: after the first period, given the current state of her beliefs, the agent can continue or can retire. The value function for this MDP can be written as

$$\begin{aligned} V_\gamma^\lambda(\mu, \sigma^2) &= \sup_{\tau \geq 0} \mathbb{E}\left[\sum_{t=0}^{\tau} \gamma^t(\theta - \lambda) \,|\, \mu_0 = \mu, \sigma_0^2 = \sigma^2\right] \\ &= \sup_{\tau \geq 0} \mathbb{E}\left[\sum_{t=0}^{\tau} \gamma^t(\mu_t - \lambda) \,|\, \mu_0 = \mu, \sigma_0^2 = \sigma^2\right], \end{aligned} \tag{2}$$

where the supremum is over stopping times $\tau \geq 0$ with respect to $(R_0, R_1, R_2, \ldots)$. The equality is due to the tower property of conditional expectation. (See Section A.3 in the appendix.) The Gittins index is the largest tax such that participating in this game is advantageous to the agent, written as

$$\lambda_\gamma(\mu, \sigma^2) := \sup\left\{\lambda \in \mathbb{R} \,\middle|\, V_\gamma^\lambda(\mu, \sigma^2) \geq 0\right\}. \tag{3}$$

This is interpreted sometimes as either a fair or prevailing tax.

To develop some intuition, note that for any tax $\lambda > \mu$, the agent could feasibly explore for some large number of periods and then choose to continue sampling only if posterior mean strictly exceeds the tax λ. For a very patient agent, the benefit of repeatedly playing an arm that generates rewards above the tax would dwarf the expected cost of initial exploration. As a result, the prevailing tax for the arm must be high enough that this event occurs very infrequently. The following theorem makes this intuition precise, showing that up to an error term that vanishes as $\gamma \to 1$, the Gittins index is exactly equal to the γ quantile of the $N(\mu, \sigma^2)$ prior distribution of θ. For sufficiently patient agents, a Gittins index measures the highest plausible mean-reward of an arm in a manner equivalent to a Bayesian upper confidence bound.

Theorem 1. *Fix any prior mean μ and prior variance σ^2. Then,*

$$\lambda_\gamma(\mu, \sigma^2) = \mu + \Phi^{-1}(\gamma)\sigma + o(1) \quad \text{as } \gamma \to 1, \tag{4}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

Remark 1. Whereas the Gittins index is derived by considering a one-armed bandit problem, both Gittins indices and upper confidence bounds are usually applied in bandit problems with $k > 1$ arms. In that context, the Gittins index theorem shows an optimal policy plays at time t the arm $\arg\max_{i \leq k} \lambda_\gamma(\mu_{t,i}, \sigma_{t,i}^2)$ whose posterior parameters $(\mu_{t,i}, \sigma_{t,i})$ are associated with the maximal Gittins index. A Bayesian upper confidence bound algorithm plays the arm $\arg\max_{i \leq k} \mu_{t,i} + \sigma_{t,i}\Phi^{-1}(q_t)$, where the posterior quantile q_t is often treated as a tunable parameter and theory suggests values like $q_t = 1 - 1/T$ when there is a known time horizon of T. The quantile in (4) is then analogous to using the natural time horizon of $T = 1/(1-\gamma)$ for a discounted problem.

Remark 2. For readers more familiar with the upper confidence bounds of Auer et al. (2002) than the Bayesian form presented here, it is worth noting that these expressions are almost identical if an improper prior is used or an arm has been sampled a moderate number of times. More about these connections can be found in Kaufmann et al. (2012).

3. Analysis

3.1. Technical Preliminaries

3.1.1. Strict Concavity of the Square Root. The next lemma is used several times in the analysis. The idea is that because $g(x) = \sqrt{x}$ is strictly concave and $g'(x) \to 0$ as $x \to \infty$, $\sqrt{x+y} \approx \sqrt{x}$ if x is much larger than y.

Lemma 1. *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be any function satisfying $|f(x)| = o(\sqrt{x})$ as $x \to \infty$. Then,*

$$\sqrt{x + f(x)} = \sqrt{x} + o(1) \quad \text{as } x \to \infty.$$

Proof. By Taylor's theorem, there is some $\tilde{x} \in [x, x + f(x)]$ such that

$$\sqrt{x + f(x)} - \sqrt{x} = \frac{f(x)}{2\sqrt{x}} + \frac{1}{2}\left(\frac{-f(x)}{\tilde{x}^{3/2}}\right).$$

For $f(x) = o(\sqrt{x})$, both terms on the right-hand side vanish as $x \to \infty$. □

3.1.2. Gaussian Tail Behavior. Let $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ denote the PDF of the standard normal distribution. Because the PDF decays exponentially as z increases, for large values of z, tail integrals like $1 - \Phi(z) = \int_{z' > z} \phi(z')dz'$ also decay exponentially as $e^{-z^2/2}$ as $z \to \infty$. Inverting this relation suggests an asymptotic approximation of $\Phi^{-1}(\gamma) \approx \sqrt{2\log(1/(1-\gamma))}$ to the quantiles of the normal distribution. The next lemma, proved in Section A.1 in the appendix, makes this precise.

Lemma 2. *As $\gamma \to \infty$,*

$$\Phi^{-1}(\gamma) = \sqrt{2\log\left(\frac{1}{1-\gamma}\right)} + o(1).$$

The same type of saddle-point approximation shows the integral $\mathbb{E}[(Z - z)^+] = \int_{z' > z} z'\phi(z')dz'$ decays like $e^{-z^2/2}$ as $z \to \infty$. For our analysis, it is convenient to have explicit upper and lower bounds, such as those in the following lemma. The upper bound here is a standard Gaussian maximal inequality and the lower bound applies Lemma 3 in Qin et al. (2017).

Lemma 3. *For $X \sim N(0, \sigma^2)$ and $\lambda \geq \mu + 2\sigma$,*

$$\frac{\sigma^4}{\lambda^3}\phi\left(\frac{\lambda}{\sigma}\right) \leq \mathbb{E}[(X - \lambda)^+] \leq \sigma\phi\left(\frac{\lambda}{\sigma}\right).$$

3.2. Reduction to Indices for Standard Normal Distributions

With some abuse of notation, for the moment let us explicitly capture the dependence of the Gittins index on the noise variance, setting $\lambda_\gamma(\mu, \sigma^2, \tilde{\sigma}^2)$ to be the Gittins index for a bandit process with prior mean μ, prior variance σ^2, and noise variance $\tilde{\sigma}^2$. A simple standardization argument shows Gittins et al. (2011)

$$\lambda_\gamma\left(\mu, \sigma^2, \tilde{\sigma}^2\right) = \mu + \sigma\lambda_\gamma\left(0, 1, \frac{\tilde{\sigma}^2}{\sigma^2}\right).$$

Therefore, it suffices to study the Gittins index for an arm with standard normal prior and some arbitrary noise variance we denote by σ_W^2. Combining this with Lemma 2, our goal in subsequent subsections is to show $\lambda_\gamma(0,1) = \sqrt{2\log(1/(1-\gamma))} + o(1)$ as $\gamma \to 1$, where we treat $\sigma_W^2 > 0$ as an arbitrary positive constant interpreted as the noise-to-signal ratio.

3.3. Upper Bound on the Gittins Index

This subsection derives an upper bound on the Gittins index via an information relaxation Brown et al. (2010). We consider a decision maker who observes noiseless signals of the true arm mean θ when sampling the arm. The prevailing tax for this decision maker exceeds the prevailing tax for one who must base the decisions on noisy reward signals. As $\gamma \to 1$, this upper bound matches both a lower bound given in Lemma 5 and the posterior quantile in Lemma 2.

Lemma 4. *As $\gamma \to \infty$,*

$$\lambda_\gamma(0,1) \leq \sqrt{2\log\left(\frac{1}{1-\gamma}\right)} + o(1) \quad \text{as } \gamma \to 1.$$

Proof. To simplify notation, write $\lambda_\gamma = \lambda_\gamma(0,1)$ and note that we often use $\mathbb{E}[\theta] = 0$ to simplify expressions. Consider a decision maker who faces a one-armed bandit problem with no observation noise. For this decision maker, playing the arm once is sufficient to perfectly reveal the true arm mean θ. An optimal policy would then play the arm in every period if $\theta \geq \lambda$, and immediately retire otherwise. Of course, a Bayesian decision maker is better off basing the retirement decision on perfect knowledge of θ than on noisy signals (see e.g., DeGroot 1962). This can be verified directly in this case: the decision maker with access to noiseless observations earns

$$\begin{aligned} &-\lambda + \left(\frac{\gamma}{1-\gamma}\right)\mathbb{E}[(\theta - \lambda)^+] \\ &\quad = \mathbb{E}[\theta - \lambda] + \mathbb{E}\sum_{t=1}^{\infty} \gamma^t(\theta - \lambda)^+ \\ &\quad \geq \sup_{\tau \geq 0} \mathbb{E}\sum_{t=0}^{\infty} \gamma^t(\theta - \lambda)\mathbf{1}(\tau \geq t) = V_\gamma^\lambda(0,1). \end{aligned} \tag{5}$$

Therefore, the fair tax for the decision maker who observes noiseless signals of θ exceeds the fair tax λ_γ for one who must base a stopping decision on imperfect signals. (See also Gutin and Farias 2016 for a detailed proof.) We have the following:

$$\begin{aligned} \lambda_\gamma &:= \sup\left\{\lambda \in \mathbb{R} \,\middle|\, V_\gamma^\lambda(0,1) \geq 0\right\} \\ &\leq \sup\left\{\lambda \in \mathbb{R} \,\middle|\, \frac{\gamma}{1-\gamma}\mathbb{E}[(\theta - \lambda)^+] \geq \lambda\right\} \\ &\overset{\text{Lem.3}}{\leq} \sup\left\{\lambda \in \mathbb{R} \,\middle|\, \frac{\gamma}{1-\gamma}\phi(-\lambda) \geq \lambda\right\} \\ &= \sup\left\{\lambda \in \mathbb{R} \,\middle|\, \log\left(\frac{\gamma}{1-\gamma}\right) \geq \log\left(\frac{\lambda}{\phi(-\lambda)}\right)\right\} := \overline{\lambda}_\gamma. \end{aligned}$$

Plugging in for the normal probability density function $\phi(\cdot)$ and simplifying, we find the upper bound $\overline{\lambda}_\gamma$ on the Gittins index is defined implicitly by

$$\sqrt{2\log(\overline{\lambda}_\gamma) + \overline{\lambda}_\gamma^2} = \sqrt{2\log\left(\frac{1}{1-\gamma}\right) + 2\log\left(\gamma\sqrt{2\pi}\right)}. \quad (6)$$

As $\gamma \to 1$, the right-hand side tends to infinity and by Lemma 1,

$$\sqrt{2\log\left(\frac{1}{1-\gamma}\right) + 2\log\left(\gamma\sqrt{2\pi}\right)} = \sqrt{2\log\left(\frac{1}{1-\gamma}\right)} + o(1). \quad (7)$$

This implies that $\overline{\lambda}_\gamma \to \infty$ as $\gamma \to 1$. Applying Lemma 1 again shows

$$\sqrt{2\log(\overline{\lambda}_\gamma) + \overline{\lambda}_\gamma^2} = \overline{\lambda}_\gamma + o(1) \quad \text{as } \gamma \to 1. \quad (8)$$

Combining Equation (6) with (7) and (8) establishes the claim. □

3.4. Lower Bound on the Gittins Index

We construct a lower bound on the Gittins index by analyzing the fair tax for an agent who employs a suboptimal heuristic policy. This agent explores for a predetermined number of periods L. Based on the resulting signals, the agent retires if $\mu_L < \lambda$ and otherwise commits to playing the arm indefinitely. The main idea is that large L will almost perfectly reveal θ, but as $\gamma \to 1$, the cost of this initial exploration is small relative to the potential value from discovering the arm has very high quality and hence has a negligible impact on the fair tax for the game. The proof will choose L as a slowly growing function of γ, so that the lower bound constructed here matches the upper bound in Lemma 4 as $\gamma \to 1$. Specifically, a choice of $L_\gamma = \lceil \sigma_W^2 \log(1/(1-\gamma))^2 \rceil$ suffices for the proof. Note that this result matches the posterior quantile in Lemma 2, and, together with Lemma 4, completes the proof of Theorem 1.

Lemma 5. *As $\gamma \to \infty$,*

$$\lambda_\gamma(0,1) \geq \sqrt{2\log\left(\frac{1}{1-\gamma}\right)} + o(1) \quad \text{as } \gamma \to 1.$$

Proof. Consider a decision maker who faces a tax λ. Suppose the agent follows a policy of exploring for $L \in \mathbb{N}$ periods, and then either retiring if $\mu_L < \lambda$ or playing the arm for all future periods otherwise. The value of this heuristic policy is a lower bound on the optimal policy, so for all fixed $L \in \mathbb{N}$,

$$V_\gamma^\lambda(0,1) \geq -\sum_{t=0}^{L-1} \gamma^t \lambda + \frac{\gamma^L}{1-\gamma}\mathbb{E}\left[(\mu_L - \lambda)^+\right] \geq -L\lambda + \frac{\gamma^L}{1-\gamma}\mathbb{E}\left[(\mu_L - \lambda)^+\right].$$

Define $\mathcal{H}_{L-1} = (R_0, \ldots R_{L-1})$ to be the history of rewards prior to period L. The posterior mean is random due to its dependence on $\mathcal{H}_{L-1}$ and has distribution $\mu_L \sim N(\mu_0, 1-\sigma_L^2)$. Here, normality follows from the fact that μ_L is a linear combination of Gaussian observations $R_0, \ldots R_{L-1}$. We have $\mathbb{E}[\mu_L] = \mathbb{E}[\mathbb{E}[\theta|\mathcal{H}_{L-1}]] = \mu_0$ by the tower property of conditional expectation, and the variance formula follows from the law of total variance:

$$1 = \mathrm{Var}(\theta) = \mathrm{Var}(\mathbb{E}[\theta|\mathcal{H}_{L-1}]) + \mathbb{E}[\mathrm{Var}(\theta|\mathcal{H}_{L-1})] = \mathrm{Var}(\mu_L) + \sigma_L^2.$$

This implies that for any $L \in \mathbb{N}$,

$$\begin{aligned}
\lambda_\gamma &\geq \sup\left\{\lambda \in \mathbb{R} \,\middle|\, \frac{\gamma^L}{1-\gamma}\mathbb{E}\left[(\mu_L - \lambda)^+\right] \geq L\lambda\right\} \\
&\overset{\text{Lem.3}}{\geq} \sup\left\{\lambda \in \mathbb{R} \,\middle|\, \frac{\gamma^L}{1-\gamma}\frac{(1-\sigma_L^2)^2}{\lambda^3}\phi\left(\frac{\lambda}{\sqrt{1-\sigma_L^2}}\right) \geq L\lambda\right\} \\
&= \sup\left\{\lambda \in \mathbb{R} \,\middle|\, \log\left(\frac{\gamma^L}{1-\gamma}\right) + \log\left(\frac{(1-\sigma_L^2)^2}{\lambda^3}\right) \right. \\
&\qquad \left. \geq \log(L\lambda) - \log\phi\left(\frac{\lambda}{\sqrt{1-\sigma_L^2}}\right)\right\}.
\end{aligned}$$

Now, choose $L_\gamma = \lceil \sigma^2 \log(\frac{1}{1-\gamma})^2 \rceil$, which tends slowly to infinity as $\gamma \to 1$, and set $\underline{\lambda}_\gamma$ to be the lower bound corresponding to the choice of $L = L_\gamma$. Plugging in for the normal PDF and simplifying, we find $\underline{\lambda}_\gamma$ is defined implicitly by

$$\sqrt{4\log\left(\underline{\lambda}_\gamma\right) + \frac{\underline{\lambda}_\gamma^2}{2\left(1-\sigma_{L_\gamma}^2\right)}} = \sqrt{\log\left(\frac{1}{1-\gamma}\right) + h(\gamma)}, \quad (9)$$

where $h(\gamma) := -\log(L_\gamma) + L_\gamma \log(\gamma) + 2\log(1-\sigma_{L_\gamma}^2) + \log(\sqrt{2\pi})$. We want to focus on the dominant terms on each side of Equation (9), which are $\underline{\lambda}_\gamma^2/2(1-\sigma_{L_\gamma}^2)$ and $\log(\frac{1}{1-\gamma})$. The next result shows the $h(\gamma)$ term has an asymptotically negligible influence.

Lemma 6. *As $\gamma \to \infty$, $h(\gamma) = o(\sqrt{\log(1/\gamma)})$ as $\gamma \to 1$.*

Together with Lemma 1, this shows $\sqrt{\log(1/\gamma) + h(\gamma)} = \sqrt{\log(1/\gamma)} + o(1)$ as $\gamma \to 1$. Hence, the solution $\underline{\lambda}_\gamma$ to

Equation (9) must also tend to ∞ as $\gamma \to 1$. Then, again by Lemma 1,

$$\sqrt{4\log\left(\underline{\lambda}_\gamma\right) + \frac{\underline{\lambda}_\gamma^2}{2\left(1-\sigma_{L_\gamma}^2\right)}} = \frac{\underline{\lambda}_\gamma}{\sqrt{2\left(1-\sigma_{L_\gamma}^2\right)}} + o(1). \quad (10)$$

Combining Equations (9) and (10) gives

$$\underline{\lambda}_\gamma = \sqrt{2\left(1-\sigma_{L_\gamma}^2\right)\log\left(\frac{1}{1-\gamma}\right)} + o(1). \quad (11)$$

The only remaining subtlety is the term $(1-\sigma_{L_\gamma}^2)$, which appears here since after L_γ measurements, the agent still has some remaining uncertainty about the value of θ. From Equation (1) for posterior variance, $\sigma_{L_\gamma}^2 \le \sigma_W^2/L_\gamma$. Plugging in for $L_\gamma = \lceil \sigma_W^2 \log(1/(1-\gamma))^2 \rceil$ gives $\sigma_{L_\gamma}^2 \log(1/(1-\gamma)) \le 1$. Plugging this into (11) gives

$$\begin{aligned}\underline{\lambda}_\gamma &\ge \sqrt{2\log\left(\frac{1}{1-\gamma}\right) - 2} + o(1)\\ &= \sqrt{2\log\left(\frac{1}{1-\gamma}\right)} + o(1). \quad \square\end{aligned}$$

4. Limitations and Open Problems

Although this note shows an equivalence between a Gittins index and a Bayesian upper confidence bound, it should be stressed that this equivalence is asymptotic as the effective time horizon of the problem grows. In particular, the Gittins index carefully captures the value of exploration given the time horizon of the problem and the variance of reward noise. Upper confidence bound algorithms do not and can engage in wasteful exploration if there is significant observation noise relative to the problem's time horizon.

One natural open direction is to extend Theorem 1 and its proof to single parameter exponential family distributions. Another question is whether extensions of the analysis can yield appropriate uniform or functional limit theorems analogous to Theorem 1. This is important to providing frequentist regret analysis of Gittins index algorithms or Bayesian regret analysis of upper confidence bound approximations. See Chang and Lai (1987) and Lattimore (2016).

Acknowledgments

Much of this short note was written as material for a doctoral course taught at Northwestern University in Spring 2017. The author thanks the students in that course for their questions and feedback.

Appendix: Omitted Technical Details

A.1. Proof of Lemma 2

Proof. We use the following standard bounds on the normal CDF Gordon (1941): for all $z \ge 0$,

$$\left(\frac{z}{1+z^2}\right)\phi(z) \le 1-\Phi(z) \le \left(\frac{1}{z}\right)\phi(z).$$

We can use this to upper bound $\Phi^{-1}(\gamma)$ as follows:

$$\begin{aligned}\Phi^{-1}(\gamma) &= \inf\{z \in \mathbb{R} \mid \Phi(z) \ge 1-\gamma\}\\ &\le \inf\left\{z \in \mathbb{R} \mid \left(\frac{1}{z}\right)\phi(z) \ge 1-\gamma\right\}\\ &= \inf\left\{z \in \mathbb{R} \mid \log\left(\frac{\phi(z)}{z}\right) \ge \log(1-\gamma)\right\} := \bar{z}_\gamma.\end{aligned}$$

Plugging in for the normal PDF ϕ and simplifying, we find that $\bar{z}_\gamma$ is defined implicitly by

$$\sqrt{\bar{z}_\gamma^2 + 2\log\left(\bar{z}_\gamma\sqrt{2\pi}\right)} = \sqrt{2\log\left(\frac{1}{1-\gamma}\right)}. \quad (A.1)$$

As $\gamma \to 1$, the right-hand side of (A.1) tends to ∞, so it must be that $\bar{z}_\gamma \to \infty$. But since $\log(\bar{z}_\gamma\sqrt{2\pi}) = o(\sqrt{\bar{z}_\gamma})$ as $\gamma \to 1$, applying Lemma 1 gives $\sqrt{\bar{z}_\gamma^2 + 2\log(\bar{z}_\gamma\sqrt{2\pi})} = \bar{z}_\gamma + o(1)$. We conclude

$$\bar{z}_\gamma = \sqrt{2\log\left(\frac{1}{1-\gamma}\right)} + o(1) \quad \text{as } \gamma \to 1.$$

The proof of the lower bound follows the same steps and is omitted. $\square$

A.2. Proof of Lemma 2

We show $h(\gamma) = o(\sqrt{\log(1/\gamma)})$ as $\gamma \to 1$. We evaluate each term in the expression $h(\gamma) := -\log(L_\gamma) + L_\gamma\log(\gamma) + 2\log(1-\sigma_{L_\gamma}^2) + \log(\sqrt{2\pi})$. Since $\log(\gamma) = -(1-\gamma) + o(1-\gamma)$ as $\gamma \to 1$, we have $L_\gamma\log(\gamma) \to 0$. In addition, $2\log(1-\sigma_{L_\gamma}^2) \to 0$ since by (1), $\sigma_{L_\gamma}^2 \le \sigma_W^2/L_\gamma \to 0$ as $\gamma \to 1$. Finally, $\log(L_\gamma) = 2\log(\sigma) + 2\log\log(\frac{1}{1-\gamma}) = o(\sqrt{\log(\frac{1}{1-\gamma})})$.

A.3. Further Justification for Equation 2

Equation (2) relies on Doob's optional sampling theorem. Here we note the technical conditions ensuring this applies. Let $\mathcal{H}_t$ denote the sigma algebra generated by $R_0, \ldots, R_{t-1}$ and let τ be any stopping time with respect to $\{\mathcal{H}_t : t \in 0, 1, \ldots\}$. Define the martingale $M = \{M_n : n = 0, 1, \ldots\}$ by

$$M_n = \sum_{t=0}^{n} \gamma^t(\theta - \mathbb{E}[\theta \mid \mathcal{H}_{t-1}]).$$

For each fixed n, $\mathbb{E}[M_n] = 0$. Equation (2) states that $\mathbb{E}[M_\tau] = 0$. (To compare, recall the definition $\mu_t = \mathbb{E}[\theta \mid \mathcal{H}_{t-1}]$). This follows by Doob's optional sampling theorem since M is a

uniformly integrable martingale. To show M is uniformly integrable, it suffices to show it is bounded in L^2. We have

$$\begin{aligned}\sup_n \mathbb{E}\left[M_n^2\right] &= \sup_n \sum_{t=0}^{n} \gamma^t \mathbb{E}\left[(\theta - \mathbb{E}[\theta \mid \mathcal{H}_{t-1}])^2\right] \\ &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathrm{Var}(\theta \mid \mathcal{H}_{t-1})] \\ &\le \sum_{t=0}^{\infty} \gamma^t \mathrm{Var}(\theta) < \infty,\end{aligned}$$

where the inequality $\mathbb{E}[\mathrm{Var}(\theta \mid \mathcal{H}_{t-1})] \le \mathrm{Var}(\theta)$ is standard and follows from Jensen's inequality for conditional expectations.

Endnote

[1] According to Google Scholar, Chang and Lai (1987) was cited only once in 2018, whereas Lai and Robbins (1985) was cited well over 200 times.

References

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

Brown DB, Smith JE, Sun P (2010) Information relaxations and duality in stochastic dynamic programs. *Oper. Res.* 58(4-part-1):785–801.

Burnetas AN, Katehakis MN (2003) Asymptotic Bayes analysis for the finite-horizon one-armed-bandit problem. *Probab. Engrg. Inform. Sci.* 17(1):53–82.

Cappé O, Garivier A, Maillard OA, Munos R, Stoltz G (2013) Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Ann. Statist.* 41(3):1516–1541.

Chang F, Lai TL (1987) Optimal stopping and dynamic allocation. *Adv. Appl. Probab.* 19(4):829–853.

Chick SE, Gans N (2009) Economic analysis of simulation selection problems. *Management Sci.* 55(3):421–437.

DeGroot MH (1962) Uncertainty, information, and sequential experiments. *Ann. Math. Statist.* 33(2):404–419.

Gittins JC, Jones DM (1974) A dynamic allocation index for the sequential design of experiments. Gani J, Sarkadi K, Vineze I, eds. *Progress in Statistics*. (North-Holland, Amsterdam), 241–266.

Gittins J, Jones D (1979) A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika* 66(3):561–565.

Gittins J, Glazebrook K, Weber R (2011) *Multi-Armed Bandit Allocation Indices* (John Wiley & Sons, UK).

Gordon RD (1941) Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *Ann. Math. Statist.* 12(3):364–366.

Gutin E, Farias V (2016) Optimistic Gittins indices. Lee DD, Sugipama M, Luxburg UV, Guyon I, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 29 (Curran Associates, Red Hook, NY), 3153–3161.

Kaufmann E, Cappé O, Garivier A (2012) On Bayesian upper confidence bounds for bandit problems. *Proc. Conf. Artificial Intelligence Statist.* (PLMR, *La Palma, Canary Islands*), 592–600.

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

Lattimore T (2016) Regret analysis of the finite-horizon Gittins index strategy for multi-armed bandits. Feldman V, Rakhlin A, Shamir O, eds. *Proc. Conf. Learn. Theory*, (PLMR, *New York*)1214–1245.

Niño-Mora J (2011) Computing a classic index for finite-horizon bandits. *INFORMS J. Comput.* 23(2):254–267.

Qin C, Klabjan D, Russo D (2017) Improving the expected improvement algorithm. Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Red Hook, NY.), 5381–5391.

Rusmevichientong P, Tsitsiklis J (2010) Linearly parameterized bandits. *Math. Oper. Res.* 35(2):395–411.

Srinivas N, Krause A, Kakade S, Seeger M (2012) Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Trans. Inform. Theory* 58(5):3250–3265.

Weber R (1992) On the Gittins index for multiarmed bandits. *Ann. Appl. Probab.* 2(4):1024–1033.

Yao YC (2006) Some results on the Gittins index for a normal reward process. *Time Series and Related Topics: In Memory of Ching-Zong Wei*, Institute of Mathematical Statistics Lecture Notes – Monograph Series, vol. 52 (IMS, Beachwood, OH), 284–294.

Daniel Russo is an assistant professor in the Decision, Risk, and Operations Division of Columbia Business School. His research lies at the intersection of statistical machine learning and sequential decision-making, and contributes the fields of online optimization and reinforcement learning.