Probability in the Engineering and Informational Sciences, **34**, 2020, 406–428.
doi:10.1017/S0269964818000529

EXPLORATION–EXPLOITATION POLICIES WITH ALMOST SURE, ARBITRARILY SLOW GROWING ASYMPTOTIC REGRET

Wesley Cowan
Department of Computer Science,
Rutgers University, Piscataway, NJ 08854, USA
E-mail: cwcowan@math.rutgers.edu

Michael N. Katehakis
Department of Management Science and Information Systems,
Rutgers University, Piscataway, NJ 08854, USA
E-mail: mnk@rutgers.edu

The purpose of this paper is to provide further understanding into the structure of the sequential allocation ("stochastic multi-armed bandit") problem by establishing probability one finite horizon bounds and convergence rates for the sample regret associated with two simple classes of allocation policies. For any slowly increasing function g, subject to mild regularity constraints, we construct two policies (the g-Forcing, and the g-Inflated Sample Mean) that achieve a measure of regret of order $O(g(n))$ almost surely as $n \to \infty$, bound from above and below. Additionally, almost sure upper and lower bounds on the remainder term are established. In the constructions herein, the function g effectively controls the "exploration" of the classical "exploration/exploitation" tradeoff.

Keywords: bandits, forcing actions, inflated sample means, multi-armed, online learning, sequential allocation, upper confidence bounds

AMS Subject Classification: Primary: 62G05, secondary: 62G20

1. INTRODUCTION AND SUMMARY

The basic problem involves sampling sequentially from a finite number of $K \geqslant 2$ populations or "bandits", where each population i is specified by a sequence of real-valued i.i.d. random variables, $\{X_k^i\}_{k\geqslant 1}$, with X_k^i representing the reward received the k^{th} time population i is sampled. The distributions F_i of the X_k^i are taken to be unknown; they belong to some collection of distributions $\mathcal{F}$. We restrict $\mathcal{F}$ in two ways: The first, that each population i has some finite mean $\mu_i = \mathbb{E}[X_k^i] = \int_{-\infty}^{+\infty} x dF_i(x)$ - unknown to the controller. The purpose of this assumption is to establish for each population i the Strong Law of Large Numbers (SLLN),

$$\mathbb{P}(\lim_k \bar{X}_k^i = \mu_i) = 1. \tag{1}$$

 0269-9648/18 $25.00

Second, we assert that each population has finite variance $\sigma_i^2 = \mathrm{Var}(X_k^i) < \infty$. The purpose of this assumption is to establish for each population i the Law of the Iterated Logarithm (LIL),

$$\mathbb{P}\left(\limsup_k \frac{\bar{X}_k^i - \mu_i}{\sqrt{\ln\ln k/k}} = \sigma_i\sqrt{2}\right) = 1. \tag{2}$$

It will emerge that the important distribution properties for the populations are not the i.i.d. structure, but rather Eqs. (1), (2) alone. This allows for some relaxation of assumptions, as discussed in Section 5. In fact, the LIL (and therefore the assumption of finite variances) is only really required for the derivation of the regret remainder term bounds in the results to follow – the primary asymptotic results depend solely on the SLLN.

Let $\mu^* = \max_i \mu_i$, and define the bandit discrepancies $\{\Delta_i\}$ as $\Delta_i = \mu^* - \mu_i \geqslant 0$.

We will make the additional assumption that the optimal bandit is unique – that is, there is a unique i^* such that $\mu_{i^*} = \mu^*$.

For any adaptive policy π, let $\pi(t) = i$ indicate the event that population i is sampled at time t, and let $T_\pi^i(n) = \sum_{t=1}^n \mathbf{1}_{\pi(t)=i}$ denote the number of times i has been sampled during periods $t = 1, 2, \ldots, n$, under policy π; for convenience we define $T_\pi^i(0) = 0$ for all i, π. One is typically interested in maximizing in some well-defined sense the sum of the first n outcomes $S_\pi(n) = \sum_{i=1}^K \sum_{k=1}^{T_\pi^i(n)} X_k^i$, achieved by an adaptive policy π. To this end, we note that if the controller had complete information (i.e., knew the distributions of the X_k^i, for each i), she would at every round activate the "optimal" bandit i^*. Natural measures of the loss due to this ignorance of the distributions are the quantities below:

$$\tilde{R}_\pi(n) = n\mu^* - \sum_{i=1}^K \mu_i T_\pi^i(n) = \sum_{i=1}^K \Delta_i T_\pi^i(n), \tag{3}$$

$$R_\pi(n) = n\mu^* - \mathbb{E}\left[S_\pi(n)\right] = \sum_{i=1}^K \Delta_i \mathbb{E}\left[T_\pi^i(n)\right]. \tag{4}$$

The functions $\tilde{R}_\pi(n)$, $R_\pi(n)$ have been called in the literature pseudo-reget, and regret; for notational simplicity, their dependance on the unknown distributions is usually suppressed.

The motivation for considering minimizing alternative regret measures to $R_\pi(n)$ is that while the investigator might be pleased to know that the policy she is utilizing has minimal *expected* regret, she might reasonably be more interested in behavior of the policy on the specific sample-path she is currently exploring rather than aggregate behavior over the entire probability space. At an extreme end of this would be a result minimizing regret or pseudo-regret surely (sample-path-wise) or almost surely (with full probability), guaranteeing a sense of optimality independent of outcome. We offer an asymptotic result of this type here in Theorem 2.

Note that $\mathbb{E}[\tilde{R}_\pi(n)] = R_\pi(n)$, and "good policies" are those that achieve a small rate of increase for one of the above regret functions. Further relationships and forms of pseudo-regret are explored in Bubeck and Cesa-Bianchi [3], e.g., the "sample regret" $R'_\pi(n) = n\mu^* - S_\pi(n) = n\mu^* - \sum_{i=1}^K \sum_{k=1}^{T_\pi^i(n)} X_k^i$. We find the pseudo-reget $\tilde{R}_\pi(n) = n\mu^* - \sum_{i=1}^K \mu_i T_\pi^i(n)$ in some sense more philosophically satisfying to consider than sample regret, for the reason that – given her ignorance and the inherent randomness – the controller cannot reasonably regret the *specific reward* gained or lost from an activation of a bandit, as in $R'_\pi(n)$. She can only reasonably regret *the decision to activate that specific bandit*, which is captured by $\tilde{R}_\pi(n)$'s dependance on the $T_\pi^i(n)$s alone.

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

Thus, we are particularly interested in high probability or guaranteed (almost sure) asymptotic bounds on the growth of the pseudo-regret as $n \to \infty$. The main result of this paper is Theorem 2 which establishes, by two examples, that for any arbitrarily (slowly) increasing function $g(n)$, e.g., $g(n) = \ln \ln \ldots \ln n$, that satisfies mild regularity conditions there exist "g-good policies" π_g. The later policies are such that the following is true

$$\tilde{R}_{\pi_g}(n) = C_{\pi_g}(\{F_i\})g(n) + o(g(n)), \quad \text{as} \quad n \to \infty$$

(i.e., $\tilde{R}_{\pi_g}(n) = O(g(n))$, (a.s), as $n \to \infty$) for every set of bandit distributions $\{F_i\} \subset \mathcal{F}$, for some positive finite constant $C_{\pi_g}(\{F_i\})$.

The results presented here are intuitive, in the following way: it will be shown that for both the **g**-Forcing, and g-inflated sample means (**g-ISM**) index policies, the function g essentially sets the investigator's willingness to explore and experiment with bandits that based on available data do not currently seem to have the highest mean. Even if the controller explores very slowly (i.e., she chose a very slow growing g), *as long as she explores long enough* she will eventually develop accurate estimates of the means for each bandit, and incur very little regret (or pseudo-regret) past that point. We note here that, for the most part, we do not recommend the actual implementation or use of these policies. The cost of this guaranteed asymptotic behavior is that (depending on g and the bandit specifics), slow pseudo-regret growth is only achieved on impractically large time-scales. We find it interesting, however, that such growth can be guaranteed – independent of the specifics of the bandits! – with as weak assumptions as the SLLN. This makes these results fairly broad. Additionally, the **g**-Forcing and **g-ISM** index policies individually capture elements present in many other popular policies, and are suggestive of the almost sure asymptotical behavior of these policies. One takeaway from this is, perhaps, to emphasize that asymptotic behavior by itself is little basis for thinking of a policy as "good". As essentially any asymptotic behavior is possible (through the choice of g), any useful qualification of a policy must consider not only the asymptotic behavior, but also the timescales over which it is practically achieved.

In the remainder of the paper, we define what it means for a policy to be g-good (Definition 1), and establish the existence of g-good policies (Theorem 2) for any g satisfying mild regularity conditions. The proof is by example, through the construction of **g**-Forcing and **g-ISM** index policies that satisfy its claim. Further, bounds on the corresponding order constants of pseudo-regret growth are established for each policy (Theorems 3 and 6), as well as bounds on the asymptotic remainder terms (Theorems 5 and 5 9), bounding the remainder from both above and below. We view the proofs of the asymptotic lower bounds, as well as the derivation of the remainder terms via a sort of bootstrapping on the earlier order results, as particularly interesting.

In the attempt to generalize some of these results for the **g-ISM** index policy, an interesting effect and seeming "phase change" in the resulting dynamics was discovered. Specifically, as discussed in Remark 2, when there are multiple optimal bandits, for g of order greater than $\sqrt{n \ln \ln n}$ all optimal bandits are sampled roughly equally often, while for g of order less than $\sqrt{n \ln \ln n}$, the **g-ISM** index policy tends to fix on a single optimal bandit, sampling the other optimal bandits much more rarely in comparison.

2. RELATED LITERATURE

Robbins [17] first analyzed the problem of maximizing asymptotically the expected value of the sum $S_\pi(n)$, for the case of Bernoulli bandit and $K = 2$ using only the assumption

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

of the SLLN for $\mathcal{F}$. He constructed a policy, π_R, which aside from forced choices (when "time" coincided with two predetermined sparse sequences of integers) it always chose the arm having the current best winning percentage, and showed that with probability one, as $n \to \infty$, $S_{\pi_R}(n)/n \to \mu^*$. From this he was able to claim, using the uniformly integrability property for the case of Bernoulli bandits that

$$R_{\pi_R}(n) = o(n), \quad \text{as} \quad n \to \infty. \tag{5}$$

Lai and Robbins [13] considered the case in which the collection of distributions $\mathcal{F}$ to consist of univariate density functions $f(x;\theta_i)$ with respect to some measure ν_i, where $f(.;.)$ is known and the unknown scalar parameter θ_i is in some known set Θ. Let $\mu_i = \mu(\theta_i) = \mathbb{E}[X_1^i]$, $\mu^* = \max_i\{\mu(\theta_i)\} = \mu(\theta^*)$, $\Delta_i(\theta_i) = \mu(\theta^*) - \mu(\theta_i)$, and let $\mathbb{I}(\theta||\theta') = \int_{-\infty}^{\infty} \ln((f(x;\theta))/(f(x;\theta')))f(x;\theta)\,\mathrm{d}\nu(x)$ denote the Kullback–Leibler divergence between $f(x;\theta)$ and $f(x;\theta')$. They established, under mild regularity conditions ((1.6), (1.7) and (1.9) therein), that if one requires a policy to have a regret that increases at slower than linear rate:

$$R_\pi(n) = o(n^\alpha), \quad \forall \alpha > 0, \quad \text{as} \quad n \to \infty, \quad \forall\{\theta_i\} \subset \Theta, \tag{6}$$

then π must sample among populations in such a way that its regret satisfies

$$\liminf_n \frac{R_\pi(n)}{\ln n} \geqslant M_{\mathrm{LR}}(\theta_1, \ldots, \theta_K), \quad \forall\{\theta_i\} \subset \Theta, \tag{7}$$

where

$$M_{\mathrm{LR}}(\theta_1, \ldots, \theta_K) = \sum_{i:\mu(\theta_i)\neq\mu^*} \Delta_i(\theta_i)/\mathbb{I}(\theta_i||\theta^*).$$

Burnetas and Katehakis [4] extended and simplified the above work for the case in which the collection of distribution $\mathcal{F}$ is specified by a known function $f(x;\underline{\theta}_i)$ that may depend on an unknown vector parameter $\underline{\theta}_i \in \underline{\Theta}_i$, as follows. Let $\underline{\underline{\theta}} := (\underline{\theta}_1, \ldots, \underline{\theta}_K) \in \underline{\underline{\Theta}} = \underline{\Theta}_1 \times \cdots \times \underline{\Theta}_K$, $\mu^* = \mu(\underline{\underline{\theta}}^*) = \max_i\{\mu(\underline{\theta}_i)\}$, $\Delta_i(\underline{\theta}_i) = \mu^* - \mu(\underline{\theta}_i)$. They showed, under certain regularity conditions (part 1 of Theorem 1, therein) that if a policy satisfied Eq. 6, $\forall \underline{\underline{\theta}} \in \underline{\underline{\Theta}}$, then it must sample among populations in such a way that its regret satisfies:

$$\liminf_n \frac{R_\pi(n)}{\ln n} \geqslant M_{\mathrm{BK}}(\underline{\underline{\theta}}), \quad \forall \underline{\underline{\theta}} \in \underline{\underline{\Theta}}, \tag{8}$$

where

$$M_{\mathrm{BK}}(\underline{\underline{\theta}}) = \sum_{i\in B(\underline{\underline{\theta}})} \Delta_i(\underline{\theta}_i)/\inf_{\underline{\theta}'_i}\{\mathbb{I}(\underline{\theta}_i, \underline{\theta}'_i) \ : \ \mu(\underline{\theta}'_i) > \mu(\underline{\underline{\theta}}^*)\}. \tag{9}$$

Further, under certain regularity conditions (cf. conditions "A1–A3" therein) regarding the estimates $\hat{\underline{\theta}}_i = \hat{\underline{\theta}}_i^n(X_1^i, \ldots, X_{T_\pi(n)}^i)$ of the parameters $\underline{\theta}_i$, $f(.;.)$ and $\underline{\Theta}_i$, they showed that policies which, after taking some small number of samples from each population, always choose the population $\pi^0(n)$ with the largest value of the population-dependent index:

$$u_i(\hat{\underline{\theta}}_i^n) = \sup_{\underline{\theta}'_i \in \underline{\Theta}_i} \left\{ \mu(\underline{\theta}'_i) \ : \ \mathbb{I}(\hat{\underline{\theta}}_i^n, \underline{\theta}'_i) < \frac{\ln n + o(\ln n)}{T_{\pi^0}^i(n)} \right\}. \tag{10}$$

are asymptotically efficient (or optimal), i.e.,

$$\limsup_n \frac{R_{\pi^0}(n)}{\ln n} \leqslant M_{\mathrm{BK}}(\underline{\theta}_1, \ldots, \underline{\theta}_K), \quad \forall \underline{\underline{\theta}} \in \underline{\underline{\Theta}}. \tag{11}$$

The index policy π^0 above was a simplification of a UCB type policy first introduced in Lai and Robbins [13] that utilized forced actions. Policies that satisfy the requirements of

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

Eq. 5, Eq. 6, and Eq. 11 were respectively called ***uniformly consistent*** (UC), ***uniformly fast convergent*** (UF), and ***uniformly maximal convergence rate*** (UM) or simply ***asymptotically optimal*** (or asymptotically efficient). The lower bound of Eq. 9 provides a baseline for comparison of the quality of policies and together with Eq. 11 and Eq. 8 provide an alternative way to state the asymptotic optimality of a policy π^0 as:

$$R_{\pi^0}(n) = M_{\mathrm{BK}}(\underline{\underline{\theta}}) \ln n + o(\ln n), \qquad \forall \underline{\underline{\theta}} \in \underline{\underline{\Theta}}. \tag{12}$$

Policies that achieve this minimal asymptotic growth rate have been derived for specific parametric models in Lai and Robbins [13], Burnetas and Katehakis [4], Honda and Takemura [10–12], Cowan et al. [8], and references therein. In general, it is not always easy to obtain such optimal polices; thus, policies that satisfy the less strict requirement of Eq. 6, $\forall \underline{\underline{\theta}} \in \underline{\underline{\Theta}}$, have been constructed, cf. Auer et al. [2], Audibert et al. [1], Bubeck and Cesa-Bianchi [3], and references therein. Such policies usually bound the regret as follows:

$$R_{\pi}(n) \leqslant M^0(\underline{\underline{\theta}}) \ln n + M^1(\underline{\underline{\theta}}), \qquad \text{for all } n \text{ and all } \underline{\underline{\theta}}, \tag{13}$$

where $M^0(\underline{\underline{\theta}})$ is, often much, bigger than $M_{\mathrm{BK}}(\underline{\underline{\theta}})$, for all $\underline{\underline{\theta}}$.

The results presented herein can seem surprising, and it may appear to contradict (at least for $g(n) = \ln n$) the classical lower bound $M_{\mathrm{BK}}(\underline{\underline{\theta}})$ of $R_{\pi}(n)/\ln n$ for UF policies π. For example, if we take F_i to be the normal distribution with unknown mean μ_i and unknown variance σ_i^2, we have for any UF policy π:

$$\lim_n \frac{\mathbb{E}[\tilde{R}_{\pi}(n)]}{\ln n} \geqslant \mathbb{M}_{\mathrm{BK}}(\underline{\mu}, \underline{\sigma}^2) = \sum_{i:\mu_i \neq \mu^*} \frac{2\Delta_i}{\ln(1 + ((\Delta_i^2)/(\sigma_i^2)))} \ .$$

On the other hand, we establish in the sequel that:

$$\begin{aligned} \lim_n \frac{\tilde{R}_{\pi_g^F}(n)}{g(n)} &= C_{\pi_g^F}(\{F_i\}) = \sum_{i:\mu_i \neq \mu^*} \Delta_i \quad \text{(a.s.)}, \\ \lim_n \frac{\tilde{R}_{\pi_g^O}(n)}{g(n)} &= C_{\pi_g^O}(\{F_i\}) = K - 1 \quad \text{(a.s.)}. \end{aligned} \tag{14}$$

However, no such contradiction exists: $M_{\mathrm{BK}}(\underline{\underline{\theta}})$ limits the $\lim_n \mathbb{E}[\tilde{R}_{\pi}(n)]/\ln n$ of a UF policy from below. In such contexts that π_g^F or π_g^O are UF, if such contexts exist, the above constants will be bounded from below by $M_{\mathrm{BK}}(\underline{\underline{\theta}})$. In such contexts that π_g^F or π_g^O are not UF, the bound does not apply. In such instances, we may in fact conclude from the results presented herein, and standard results relating modes of convergence, that for the policies constructed here, for $g(n) = O(\ln n)$, the sequences of random variables $\tilde{R}_{\pi_g^F}(n)/g(n)$, $\tilde{R}_{\pi_g^O}(n)/g(n)$ are not uniformly integrable. An example as to how this can occur is given via the proof of Theorem 2 of Cowan et al. [8], where with a non-trivial probability, non-representative initial sampling of each bandit biases expected future activations of sub-optimal bandits super-logarithmically. This effect does not influence the long-term almost sure behavior of these policies. For other significant related recent work, we refer to Garivier et al. [9], Lattimore [14], Ortner [16], Orabona and Pál [15], Cowan and Katehakis [5–7].

3. MAIN THEOREMS

We characterize a policy by the rate of growth of its pseudo-regret function $\tilde{R}_{\pi}(n)$ with n in the following way.

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

DEFINITION 1: *For a function $g(n)$, a policy π is g-good if for every set of bandit distributions $\{F_i\} \subset \mathcal{F}$, there exists a constant $C_\pi(\{F_i\}) < \infty$ such that*

$$\limsup_n \frac{\tilde{R}_\pi(n)}{g(n)} \leqslant C_\pi(\{F_i\}) \ (a.s) \quad as\ n \to \infty. \tag{15}$$

Remark 1: Essentially, a policy is g-good if $\tilde{R}_\pi(n) \leqslant O(g(n))$ (a.s), $n \to \infty$. Trivially, policies exist that are n-good (i.e., $\tilde{R}_\pi(n) \leqslant O(n)$ (a.s.)), for example, any policy that samples all populations at constant rate $1/K$.

We next state the following theorem:

THEOREM 2: *For g, an unbounded, positive, increasing, concave, differentiable, sub-linear function, there exist g-good policies.*

The proof of this theorem is given by example with Theorems 3 and 6, which demonstrate two g-good policies: the **g**-Forcing and the **g-ISM** index policies.

We note that in the sequel it will be assumed that any g considered is an *unbounded, positive, increasing, concave, differentiable, sub-linear* function.

3.1. The Class of g-Forcing Policies

Let g be as hypothesized in Theorem 2. We define a **g**-Forcing policy π_g^F in the following way:

g-Forcing policy:

A policy π_g^F that first samples each bandit once, then for $t \geqslant K$,

$$\pi_g^F(t+1) = \begin{cases} \arg\max_i \bar{X}^i_{T^i_{\pi_g^F}(t)} & \text{if } \min_i T^i_{\pi_g^F}(t) \geqslant g(t), \\ \arg\min_i T^i_{\pi_g^F}(t) & \text{else.} \end{cases} \tag{16}$$

Briefly, at any time t, if any population has been sampled fewer than $g(t)$ times, sample it. Otherwise, sample from the population with the current highest sample mean. Ties are broken either uniformly at random, or at the discretion of the investigator. In this way, g can be seen as determining the rate of exploration of currently sub-optimal bandits. This can be viewed as a variant on the policy π_R considered in Robbins [17].

It is convenient to define the following constant,

$$S_\Delta = \sum_{i:\mu_i \neq \mu^*} \Delta_i. \tag{17}$$

The value S_Δ in some sense represents the pseudo-regret incurred each time the sub-optimal bandits are all activated once. The next result states that **g**-Forcing policies satisfy the conditions of Theorem 2.

THEOREM 3: *For a policy π_g^F as in (16), π_g^F is g-good, and*

$$\mathbb{P}\left(\lim_n \frac{\tilde{R}_{\pi_g^F}(n)}{g(n)} = S_\Delta\right) = 1. \tag{18}$$

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

The above theorem can be strengthened in the following way, bounding the asymptotic remainder terms almost surely:

THEOREM 4: *For a policy π_g^F as in (16), the following are true:*

$$\mathbb{P}\left(\limsup_n(\tilde{R}_{\pi_g^F}(n) - S_\Delta g(n)) \leqslant S_\Delta\right) = 1, \tag{19}$$

and

$$\mathbb{P}\left(\liminf_n(\tilde{R}_{\pi_g^F}(n) - S_\Delta g(n)) \geqslant 0\right) = 1. \tag{20}$$

PROOF OF THEOREMS 3 AND 4: Theorems 3, 4 follow immediately from the following proposition, the proof of which is given in Appendix 5:

PROPOSITION 5: *For policy π_g^F as in (16), the following is true: For every $\epsilon > 0$, almost surely there exists a $N_\epsilon < \infty$ such that, for all $n \geqslant N_\epsilon$,*

$$g(n)S_\Delta - \epsilon \leqslant \tilde{R}_{\pi_g^F}(n) \leqslant \lceil g(n)\rceil S_\Delta. \tag{21}$$

Using the above relation to bound first the limits as $n \to \infty$ of $\tilde{R}_{\pi_g^F}(n)/g(n)$, then $\tilde{R}_{\pi_g^F}(n) - S_\Delta g(n)$ (observing that $\lceil g(n)\rceil - g(n) \leqslant 1$), give the desired results. ■

Proposition 5 is considerably stronger than Theorems 3, 4. However, it somewhat obscures the true nature of what is going on: for sufficiently large n, almost surely, sub-optimal bandits ($i : \mu_i \neq \mu^*$) are *only* activated during the "forcing" phase of the policy, when some activations are below g. As a result, since g increases slowly (e.g. is sub-linearly), for large n, $T^i_{\pi_g^F}(n) = \lceil g(n)\rceil$ – except for a discrepancy that occurs, for a brief stretch ($< K$) of activations, whenever g surpasses the next integer threshold. At this point, the policy raises the activations of each sub-optimal bandit, restoring the previous equality. Hence, in fact, equality holds in Proposition 5 ($\tilde{R}_{\pi_g^F}(n) = \lceil g(n)\rceil S_\Delta$) for most large n. Discrepancy occurs increasingly rarely with n, based on the hypotheses on g. If, additionally, the controller specifies a deterministic scheme for tie-breaking, pseudo-regret may be determined explicitly for all sufficiently large n. Leaving ties to the discretion of the controller, Proposition 5 is as strong as a statement can be made.

3.2. The Class of g-ISM-Index Policies

In this section, we consider an index policy related to the classical "UCB" index policies. Let g be as hypothesized. For each i, define an *index* on $(j,k) \in \mathbb{Z}_+^2$,

$$u_i(j,k) = \bar{X}^i_k + \frac{g(j)}{k}. \tag{22}$$

g-ISM index policy:
A policy π_g^O that first samples each bandit once, then for $t \geqslant K$,

$$\pi_g^O(t+1) = \arg\max_i u_i(t, T^i_{\pi_g^O}(t)) = \arg\max_i \left(\bar{X}^i_{T^i_{\pi_g^O}(t)} + \frac{g(t)}{T^i_{\pi_g^O}(t)}\right). \tag{23}$$

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

Briefly, at any time, the sample means of each bandit are "inflated" by the $g(t)/T^i_{\pi^O_g}(t)$ term, and the policy always activates the bandit with the largest inflated sample mean. When unsampled, a bandit's inflated sample mean increases essentially at rate g, hence g drives the rate of exploration of current sub-optimal bandits. While this policy is inspired by more traditional "Upper Confidence Bound" policies, we refer to this as an Inflated Sample Mean policy, as it has no deliberate connection to confidence bounds.

More general index policies of this type could also be considered, for instance based on an index $\bar{X}^i_k + H_i(g(j)/k)$ where H_i is some positive, increasing function of its argument. This is more in line with the common UCB policies, which frequently have inflation terms of the form $O(\sqrt{\ln n/T^i_\pi(n)})$ (though this is hardly necessary, c.f. Cowan et al. [8]) with $\ln n$ serving the "exploration-driving" role of g. However, introducing this extra H_i function does not influence the order of the growth of pseudo-regret, it simply changes the relevant order constants, at the cost of complicating the analysis.

Theorem 6 below shows that a **g-ISM** index policy satisfies the conditions of Theorem 2, and gives the minimal order constant $C_{\pi^O_g}$ for this policy.

THEOREM 6: *For a policy π^O_g as in (23), if the optimal bandit is unique,*

$$\mathbb{P}\left(\lim_n \frac{\tilde{R}_{\pi^O_g}(n)}{g(n)} = K - 1\right) = 1. \tag{24}$$

The proof of this theorem depends on the following propositions, the proofs of which are given in Appendix 5. Interestingly, these results (and therefore Theorem 6) depend only on the assumption of the SLLN, not the LIL.

PROPOSITION 7: *For each sub-optimal i, $\forall \epsilon \in (0, \Delta_i/2)$, $\exists$ (a.s.) a finite constant C^i_ϵ such that for $n \geqslant K$,*

$$T^i_{\pi^O_g}(n) \leqslant \frac{g(n)}{\Delta_i - 2\epsilon} + C^i_\epsilon. \tag{25}$$

PROPOSITION 8: *For each sub-optimal $i \neq i^*$, $\forall \epsilon \in (0, \min_{j \neq i^*} \Delta_j/2)$, $\exists$ (a.s.) some finite N' such that for $n \geqslant N'$,*

$$\frac{g(n)}{(1+\epsilon)(\Delta_i + 2\epsilon) + 2\epsilon} \leqslant T^i_{\pi^O_g}(n). \tag{26}$$

PROOF OF THEOREM 6: For each sub-optimal bandit i, as an application of Propositions 7 and 8, taking the limit of $T^i_{\pi^O_g}(n)/g(n)$ first as $n \to \infty$, then as $\epsilon \to 0$, gives $\lim_n T^i_{\pi^O_g}(n)/g(n) = 1/\Delta_i$, almost surely. The theorem then follows similarly, from the definition of pseudo-regret, Eq. (3). ■

Remark 2: In the case that the optimal bandit is not unique, it happens that Proposition 7 still holds. It can be shown then that π^O_g remains g-good in this case, and it has a limiting order constant of at most $K - K^*$ (K^* as the number of optimal bandits). We leave as an open question, however, that of producing a Proposition 8 type lower bound and the verification of $K - K^*$ as the minimal order constant. The proof of Proposition 8 for $K^* = 1$ depends on establishing a lower bound on the activations of the unique optimal bandit: in short, at time n, since the sub-optimal bandits are activated at most $O(g(n))$ times (which holds independent of K^*), it follows from its uniqueness that the optimal bandit is activated

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

at least $n - O(g(n))$ times. If, however, $K^* > 1$ and the optimal bandit is not unique, while the optimal bandits must have been activated at least $n - O(g(n))$ *in total* at time n, the distribution of these activations among the optimal bandits is hard to pin down. Simple simulations seem to indicate a sort of "phase change", in that for g of order greater than $\sqrt{n \ln \ln n}$, all optimal bandits are sampled roughly equally often, while for g of order less than $\sqrt{n \ln \ln n}$, the policy tends to fix on a single optimal bandit, sampling the other optimal bandits much more rarely in comparison.

We offer the following as a potential explanation of this observed effect (and justification of the difficult to observe $\ln \ln n$ term): Let us hypothesize, for the moment, that under any circumstances, the optimal bandits are activated linearly with time, that is for any optimal i^*, $T^{i^*}_{\pi^O_g}(n) = O(n)$, with the order coefficient depending on the specifics of that bandit. Under policy π^O_g, activations are governed by a comparison of indices. We consider then the fluctuations in value of the two terms of the index, the sample mean $\bar{X}^{i^*}_{T^{i^*}_{\pi^O_g}(n)}$ and the inflation term $g(n)/T^{i^*}_{\pi^O_g}(n)$. Under the assumption, the optimal bandits are activated linearly, and reasonable assumptions on the bandit distributions (to grant the LIL), the fluctuations in the sample mean over time will be of order $O(\sqrt{\ln \ln n / n})$. The fluctuations in the inflation term will be of order $O(g(n)/n)$. It would seem to follow then that for g of order less than $O(\sqrt{n \ln \ln n})$, when comparing indices of optimal bandits, the sample mean is the dominant contribution to the index, while for g of order greater than $O(\sqrt{n \ln \ln n})$, the inflation term is the dominant contribution to the index. When the inflation term dominates, among the optimal bandits an "activate according to the largest index" policy essentially reduces to a "activate according to the smallest number of activations" policy, which leads to equalization and all optimal bandits being activated roughly equally often. When the sample mean dominates, among the optimal bandits an "activate according to the largest index" policy essentially reduces to an "activate according to the highest sample mean" or "play the winner" policy, which leads to the policy fixing on certain bandits for long periods.

This explanation would additionally suggest that on one side of the phase change, when the inflation term dominates, the only properties of the optimal bandits that matter for the dynamics of the problem are their means, that they all have the optimal mean μ^*. But on the other side of the phase change, when the sample mean dominates, other properties such as the variances $\{\sigma_i^2\}$ influence the dynamics, through the LIL. However, at this point in time, this remains, while interesting, speculative.

Based on the above results, we have the following result: For each $i \neq i^*$, $\forall \epsilon > 0$, $\exists$ (a.s.) some finite N_ϵ such that for $n \geqslant N_\epsilon$,

$$\frac{1-\epsilon}{\Delta_i} g(n) \leqslant T^i_{\pi^O_g}(n) \leqslant \frac{1+\epsilon}{\Delta_i} g(n). \tag{27}$$

Similarly, for the optimal bandit i^*,

$$n - (1+\epsilon) \sum_{i \neq i^*} \frac{1}{\Delta_i} g(n) \leqslant T^{i^*}_{\pi^O_g}(n) \leqslant n - (1-\epsilon) \sum_{i \neq i^*} \frac{1}{\Delta_i} g(n). \tag{28}$$

It follows trivially from these that each bandit is activated infinitely often, i.e., almost surely $\{T^i_{\pi^O_g}(n)\}_{n \geqslant 1}$ is equivalent to the sequence $\{0, 1, \ldots\}$, through with some (finite) stretches

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

of term repetition. It follows then, applying the LIL that

$$\mathbb{P}\left(\limsup_n \pm \frac{\bar{X}^i_{T^i_{\pi_g^O}(n)} - \mu_i}{\sqrt{\ln\ln T^i_{\pi_g^O}(n)/T^i_{\pi_g^O}(n)}} = \sigma_i\sqrt{2}\right) = 1. \tag{29}$$

This provides greater control over the sample mean of each bandit than what the SLLN alone allows, and allows the results of the previous asymptotic results to be strengthened, as in the following theorem.

THEOREM 9: *For a policy π_g^O as in (23), then the following are true:*

a) if $g(n) = o(n/\ln\ln n)$,

$$\mathbb{P}\left(\limsup_n \frac{\tilde{R}_{\pi_g^O}(n) - (K-1)g(n)}{\sqrt{g(n)\ln\ln g(n)}} \leqslant 2\sqrt{2}\sum_{i\neq i^*}\frac{\sigma_i}{\sqrt{\Delta_i}}\right) = 1, \tag{30}$$

b) if $g(n) = o(n^{2/3})$,

$$\mathbb{P}\left(\liminf_n \frac{\tilde{R}_{\pi_g^O}(n) - (K-1)g(n)}{\sqrt{g(n)\ln\ln g(n)}} \geqslant -3\sqrt{2}\sum_{i\neq i^*}\frac{\sigma_i}{\sqrt{\Delta_i}}\right) = 1. \tag{31}$$

In short, we have that for a **g-ISM** index policy π_g^O,

$$\tilde{R}_{\pi_g^O}(n) = (K-1)g(n) + O(\sqrt{g(n)\ln\ln g(n)}).$$

It should be observed that, unlike previous results, this theorem is somewhat restrictive in its allowed g. However, since the focus is traditionally on logarithmic regret, i.e., $g(n) = O(\ln n)$, it is clear that the above restrictions are nothing serious.

This theorem follows trivially from the following refinements of Propositions 7, 8, and the definition of pseudo-regret, Eq. (3). Their proofs are given in Appendix C.

PROPOSITION 10: *If $g(n) = o(n/\ln\ln n)$, for each sub-optimal $i \neq i^*$, the following holds almost surely:*

$$\limsup_n \frac{\Delta_i T^i_{\pi_g^O}(n) - g(n)}{\sqrt{g(n)\ln\ln g(n)}} \leqslant \frac{2\sigma_i\sqrt{2}}{\sqrt{\Delta_i}}. \tag{32}$$

PROPOSITION 11: *If $g(n) = o(n^{2/3})$, for each sub-optimal $i \neq i^*$, the following holds almost surely:*

$$\liminf_n \frac{\Delta_i T^i_{\pi_g^O}(n) - g(n)}{\sqrt{g(n)\ln\ln g(n)}} \geqslant -\frac{3\sigma_i\sqrt{2}}{\sqrt{\Delta_i}}. \tag{33}$$

Again, we leave as an open problem that of extending these results to the case of non-unique optimal bandits.

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

4. COMPARISON BETWEEN POLICIES

We have established two policies, **g**-Forcing and **g-ISM** index, that each achieve $O(g(n))$ pseudo-regret, almost surely. The question of which policy is "better" is not necessarily well posed. For one thing, the asymptotic pseudo-regret growth of either policy can be improved by picking a slower g. In this sense, there is certainly no "optimal" policy as there will always be a slower g. For a fixed g, however, the question of which policy is better becomes context specific: for some bandit distributions, the order constant of the **g**-Forcing policy, $S_\Delta = \sum_{i:\mu_i \neq \mu^*} \Delta_i$, will be smaller than the order constant of the **g-ISM** index policy, $K-1$; for some bandit distributions, the comparison will go the other way.

In terms of the results presented here, the pseudo-regret of the **g**-Forcing policy is much more tightly controlled, Proposition 5 bounding the fluctuations in pseudo-regret around $S_\Delta g(n)$ by at most a constant – indeed, at most S_Δ. The bounds on the **g-ISM** index policy however are $O(\sqrt{g(n)\ln\ln g(n)})$. But, this additional control of the **g**-Forcing policy comes at a cost. It follows from the proof of Proposition 5 that for sub-optimal i, for all large n,

$$T^i_{\pi^F_g}(n) \approx g(n). \tag{34}$$

However, for the **g-ISM** index policy, following the proof of 6, for all sub-optimal i, and large n,

$$T^i_{\pi^O_g}(n) \approx \frac{g(n)}{\Delta_i}. \tag{35}$$

It is clear from this that the **g**-Forcing policy is in some sense the more democratic of the two, sampling all sub-optimal bandits equally, regardless of quality. The **g-ISM** index policy is the more meritocratic, sampling sub-optimal bandits more rarely the farther they are from the optimum. This has the effect of boosting the sampling of bandits near the optimum, but this effect is somewhat counterbalanced as they contribute less to the pseudo-regret.

5. RELAXING ASSUMPTIONS: NON-I.I.D. BANDITS

The assumption that the results from each bandit are i.i.d. is fairly standard – the problem is generally phrased as a matter of knowledge discovery about a set of unknown distributions, through the use of repeated measurements. However, it is interesting to observe that this assumption actually plays no part in the results and proofs present in this paper. The sole distributional property that mattered for establishing the policies as g-good was the assumption that for each bandit there existed some finite μ_i such that $\bar{X}^i_k \to \mu_i$ almost surely with k (though the LIL was utilized to great effect in bounding the remainder terms). In fact, the expected values of the individual X^i_j need not be μ_i, nor must the X^i_k be independent of each other for a given i. Further, it is never necessary that the bandits themselves be independent of each other! In that regard, the results herein are actually quite general statements about minimizing pseudo-regret under arbitrary multidimensional stochastic processes that satisfy that strong large number law-type requirement.

However, a word of caution is due: removing the restrictions on $\{X^i_k\}_{k\geqslant 1}$ in this way, while not influencing the proofs of the results presented here, does somewhat call into question the definition of "pseudo-regret" as given in Eq. (3). The individual sample means freed, it is not necessarily reasonable to define a finite horizon pseudo-regret, $\tilde{R}_\pi(n)$, in terms of the infinite horizon means, $\{\mu_i\}$. For instance, it is no longer necessarily true that the optimal, complete knowledge policy on any finite horizon is simply to activate a bandit with infinite horizon mean μ^* at every point. A more applicable definition of pseudo-regret

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

would have to take into account what is reasonable to know or measure about the state of each bandit in finite time.

Acknowledgements

We acknowledge support for this work from the National Science Foundation NSF grants CMMI-14-50743, and CMMI-16-62629.

References

1. Audibert, J-Y, Munos, R., & Szepesvári, C. (2009). Exploration - exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science* 410: 1876–1902.
2. Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning* 47: 235–256.
3. Bubeck, S. & Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*.
4. Burnetas, A.N. & Katehakis, M.N. (1996). Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics* 17: 122–142.
5. Cowan, W. & Katehakis, M.N. (2015a). An Asymptotically Optimal Policy for Uniform Bandits of Unknown Support. *arXiv preprint: arXiv:1505.01918*.
6. Cowan, W. & Katehakis, M.N. (2015b). Multi-armed bandits under general depreciation and commitment. *Probability in the Engineering and Informational Sciences* 29(1): 51–76.
7. Cowan, W. & Katehakis, M.N. (2015c). Asymptotically Optimal Sequential Experimentation Under Generalized Ranking. *arXiv preprint arXiv:1510.02041*.
8. Cowan, W., Honda, J., & Katehakis, M.N. (2018). Normal bandits of unknown means and variances. *Journal of Machine Learning Research* 18(154): 1–28.
9. Garivier, A., Ménard, P., & Stoltz, G. (2018). Explore first, exploit next: the true shape of regret in bandit problems. *Mathematics of Operations Research*. doi: 10.1287/moor.2017.0928.
10. Honda, J. & Takemura, A (2010) An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, 67–79, Citeseer.
11. Honda, J. & Takemura, A. (2011). An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning* 85: 361–391.
12. Honda, J. & Takemura, A. (2013) Optimality of Thompson sampling for Gaussian bandits depends on priors. *arXiv preprint arXiv:1311.1894*.
13. Lai, T.L. & Robbins, H.E. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6: 4–2.
14. Lattimore, L. (2018). Refining the confidence level for optimistic bandit strategies. *Journal of Machine Learning Research* 19: 765–796.
15. Orabona, F., D. Pál (2016) Open problem: Parameter-free and scale-free online algorithms. *Conference on Learning Theory*, 1659–1664.
16. Ortner, R. (2018). Regret Bounds for Reinforcement Learning via Markov Chain Concentration. *arXiv preprint arXiv:1808.01813*.
17. Robbins, H.E. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Monthly* 58: 527–536.

APPENDIX A. PROOF OF PROPOSITION 5

PROOF: To prove Proposition 5, it will suffice to show the following: For all $i : \mu_i \neq \mu^*$ and all $\delta > 0$, $\exists$ (a.s.) a finite time $T_\delta < \infty$ such that,

$$g(t) - 2\delta \leqslant T^i_{\pi^F_g}(t) \leqslant \lceil g(t) \rceil \quad \forall t \geqslant T_\delta. \tag{A.1}$$

Theorem 5 follows from this result and Eq. (3), with the appropriate choice of δ.

Without loss of generality, we may restrict ourselves to $\delta < 1/2$.

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

As a preliminary step: Based on the properties of g, if K is the total number of bandits, there exists a finite, not random, time t_δ such that , the following is true:

$$g(t+K) < g(t) + \delta, \quad \forall\, t \geqslant t_\delta. \tag{A.2}$$

This follows from the observation that $g(t+K) \leqslant g(t) + g'(t)K$, and that $g'(t) \to 0$.

When implementing a **g**-Forcing policy π_g^F (hereafter referenced simply as π), there are essentially two alternating phases (or modes) of the policy: "catch up" and "play the winner". During "catch up", some number of bandits have fewer than g activations (the sub-g bandits), and they are activated until all bandits have at least g activations. During "play the winner", each bandit has at least g activations, and the bandit with the current greatest sample mean is activated. These phases can be seen as governed by the function $\Delta(t) = g(t) - \min_i T_\pi^i(t)$ so that when $\Delta(t) > 0$, the policy is in "catch up" mode, when $\Delta(t) \leqslant 0$, the policy is in "play the winner" mode.

Having activated bandits according to policy π up to time t_δ, suppose that $\Delta(t_\delta) > 0$, hence the policy enters or is in a period of "catch up". Let $d(= d(t_\delta))$ be the number of sub-g bandits at time t_δ. Because g is increasing, and there are d sub-g bandits at time t_δ, it will take at least d "catch up" activations before the policy enters a period of "play the winner" ($\Delta \leqslant 0$). Consider activating bandits according to policy π for d activations. Note, $d \leqslant K$, so from Ineq. A.2 and increasing property of g we have: $g(t_\delta + d) < g(t_\delta) + \delta$. Additionally, $\min_i T_\pi^i(t_\delta + d) \geqslant \min_i T_\pi^i(t_\delta) + 1$, as every bandit realizing the minimum activations will have been activated at least once. It follows that

$$\begin{aligned} \Delta(t_\delta + d) &= g(t_\delta + d) - \min_i T_\pi^i(t_\delta + d) \\ &< g(t_\delta) + \delta - \min_i T_\pi^i(t_\delta) - 1 \\ &= \Delta(t_\delta) - (1 - \delta). \end{aligned} \tag{A.3}$$

Hence, after a period of d activations from time t_δ, the spread Δ has decreased by at least $1 - \delta$. Repeating this argument, based on the number of sub-g bandits (if any) at time $t_\delta + d$, it is clear that eventually – *in finite time* – a time $T_\Delta < \infty$ is reached such that $\Delta(T_\Delta) \leqslant 0$. At this point, all bandits have been activated at least g times, and the policy enters a period of "play the winner". We observe the loose, but sample-path-wise, bound that,

$$T_\Delta \leqslant t_\delta + K\frac{(\Delta(t_\delta))^+}{1-\delta} \leqslant t_\delta + K\frac{g(t_\delta)}{1-\delta} < \infty, \tag{A.4}$$

since $\Delta(t) \leqslant g(t)$ always, and at every step the number of sub-g bandits is at most K. Observe that if in fact $\Delta(t_\delta) \leqslant 0$, then we may take $T_\Delta = t_\delta$.

Having entered a period of $\Delta \leqslant 0$ or "play the winner" at time T_Δ, let $t \geqslant T_\Delta$ such that $\Delta(t) \leqslant 0$ but $\Delta(t+1) > 0$. That is, in the transition from time t to $t+1$, g surpasses the number of activations of some bandits and the policy enters a period of "catch up". At such a point, we have the following relations:

$$\min_i T_\pi^i(t+1) < g(t+1) < g(t) + \delta \leqslant \min_i T_\pi^i(t) + \delta. \tag{A.5}$$

The first inequality is simply that $\Delta(t+1) > 0$, the second following since $t \geqslant t_\delta$, and the last since $\Delta(t) \leqslant 0$. However, since the T_π^i are integer valued and non-decreasing, the above yields

$$\min_i T_\pi^i(t+1) = \min_i T_\pi^i(t). \tag{A.6}$$

Combining Eqs. (A.5), (A.6) yields the important relation that $\Delta(t+1) < \delta$. Note additionally,

$$g(t+1) < g(t) + \delta \leqslant \min_i T_\pi^i(t) + \delta < \min_i T_\pi^i(t+1) + 1. \tag{A.7}$$

Again noting the T_π^i are integer valued, this implies that while there are sub-g bandits at time $t+1$, the only sub-g bandits are those that realize the minimum number of activations

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

$\min_i T_\pi^i(t+1)$. All other bandits have activations strictly greater than g. Let the number of sub-g bandits at time $t+1$ again be denoted $d = d(t+1)$. For $d' < d$ ($\leqslant K$) additional activations under π, in the "catch up" phase, we have that $\min_i T_\pi^i(t+1+d') = \min_i T_\pi^i(t+1)$ and $g(t+1+d') < g(t+1) + \delta$. Hence, $\Delta(t+1+d') < \Delta(t+1) + \delta < 2\delta$. For d additional activations after time $t+1$, each sub-g bandit has been activated once, raising the minimum number of activations by 1: $\min_i T_\pi^i(t+1+d) = \min_i T_\pi^i(t+1) + 1$. Additionally, $g(t+1+d) < g(t+1) + \delta$, hence $\Delta(t+1+d) < \Delta(t+1) - \delta < 0$.

We see therefore that after T_Δ, at any point at which Δ becomes positive after being at most zero, it is at most 2δ for a finite time – the "catch up" phase – before becoming negative. Hence it follows, that for $t \geqslant T_\Delta$, $\Delta(t) \leqslant 2\delta$, or for each i

$$g(t) - 2\delta \leqslant T_\pi^i(t). \tag{A.8}$$

Note, this is true for all i. This acts as justification for the description of g as the "forcing function", as the policy forces all activations to grow at least at g asymptotically.

Since g is unbounded and increasing, all populations are sampled infinitely often over time. Taking the strong law of large numbers to hold, for every $\epsilon > 0$ and each i, there exists almost surely some finite N_ϵ^i such that $\bar{X}_k^i \in [\mu_i - \epsilon, \mu_i + \epsilon]$ for all $k \geqslant N_\epsilon^i$. It is worth noting here that while such a N_ϵ^i exists, it is random and unknowable to the investigator. Because of the properties of g, we may define a finite $T_\epsilon^i > T_\Delta$ such that $N_\epsilon^i \leqslant g(T_\epsilon^i) - 2\delta$. By Eq. (A.8), we have that for all $t \geqslant T_\epsilon^i$,

$$\bar{X}^i_{T_\pi^i(t)} \in [\mu_i - \epsilon, \mu_i + \epsilon]. \tag{A.9}$$

Hence we have for each population, for every $\epsilon > 0$, there exists almost surely a finite random time $T_\epsilon = \max_i T_\epsilon^i < \infty$ past which the sample mean is trapped within the $\mu_i \pm \epsilon$ interval.

Fix ϵ sufficiently small, so as to distinguish μ^* from the other means (i.e., $[\mu^* - \epsilon, \mu^* + \epsilon] \cap [\mu_i - \epsilon, \mu_i + \epsilon] = \emptyset$ for all $i : \mu_i \neq \mu^*$). By the previous observations, we have therefore that for all $t \geqslant T_\epsilon$, for all sub-optimal i and any optimal i^*,

$$\bar{X}^{i^*}_{T_\pi^{i^*}(t)} > \bar{X}^i_{T_\pi^i(t)}. \tag{A.10}$$

In short, almost surely there exists a finite time T_ϵ past which the sample means of sub-optimal bandits are always inferior to the sample mean of any optimal bandit.

By the structure of the policy π, for all $t \geqslant T_\epsilon$, sub-optimal populations are only activated during the g-forced "catch up" periods. If at time T_ϵ, the number of times a sub-optimal bandit i has been activated is greater than g – for instance due to it, at some point, having the largest sample mean during a "play the winner" period – that population will not be sampled again until g has increased to overcome this temporary excess. As g is increasing and unbounded, this must occur in finite time. Once this occurs, as observed previously, g can only exceed T_π^i by at most 2δ before bandit i is again activated, raising T_π^i above g once more. As this "catch up" is the only time bandit i is activated, and $\delta < 1/2$, it follows that there exists some finite time $\tilde{T}_\epsilon^i > T_\epsilon$ such that for $t \geqslant \tilde{T}_\epsilon^i$, $T_\pi^i(t) \leqslant \lceil g(t) \rceil$. Taking $T_\delta = \max_{i:\mu_i \neq \mu^*} \tilde{T}_\epsilon^i$, and noting that $t_\delta \leqslant T_\Delta \leqslant T_\epsilon \leqslant T_\delta < \infty$, we have that for $t \geqslant T_\delta$, for all sub-optimal i,

$$g(t) - 2\delta \leqslant T_\pi^i(t) \leqslant \lceil g(t) \rceil. \tag{A.11}$$

■

APPENDIX B. PROOFS OF PROPOSITIONS 7 AND 8

In this section, π refers to a **g-ISM** index policy as in Eq. 23. The results to follow depend on the following lemma.

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

LEMMA 12: *Under the assumption of Eq. (1), for each i, and for any $\epsilon > 0$, the inequality:*

$$u_i(j,k) < \mu_i - \epsilon$$

holds for only finitely many (j,k)-pairs, almost surely.

PROOF: As an application of the strong law, almost surely there is some finite N_ϵ^i such that $\bar{X}_k^i > \mu - \epsilon/2$, for all $k \geqslant N_\epsilon^i$. For such k, as g is positive, $u_i(j,k) = \bar{X}_k^i + g(j)/k \geqslant \mu_i - \epsilon$, for all j. For any $k < N_\epsilon^i$, the relation $u_i(j,k) = \bar{X}_k^i + g(j)/k < \mu_i - \epsilon$ may be true only for finitely many j since g is increasing. ■

Proof of Proposition 7: For $i \neq i^*$, we define the following quantities. Taking $\epsilon > 0$, and $2\epsilon < \mu^* - \mu_i$, and $n \geqslant K$,

$$\begin{aligned} n_1^i(n,\epsilon) &= \sum_{t=N}^{n} \mathbf{1}\{\pi(t+1) = i, u_i(t, T_\pi^i(t)) \geqslant \mu^* - \epsilon, \bar{X}^i_{T_\pi^i(t)} \leqslant \mu^i + \epsilon\} \\ n_2^i(n,\epsilon) &= \sum_{t=N}^{n} \mathbf{1}\{\pi(t+1) = i, u_i(t, T_\pi^i(t)) \geqslant \mu^* - \epsilon, \bar{X}^i_{T_\pi^i(t)} > \mu^i + \epsilon\} \\ n_3^i(n,\epsilon) &= \sum_{t=N}^{n} \mathbf{1}\{\pi(t+1) = i, u_i(t, T_\pi^i(t)) < \mu^* - \epsilon\}. \end{aligned} \tag{B.1}$$

Hence we have the following relationship,

$$T_\pi^i(n+1) = 1 + \sum_{t=N}^{n} \mathbf{1}\{\pi(t+1) = i\} = 1 + n_1^i(n,\epsilon) + n_2^i(n,\epsilon) + n_3^i(n,\epsilon). \tag{B.2}$$

The proof proceeds via a pointwise bound on each of the three terms. For the first term,

$$\begin{aligned} n_1^i(n,\epsilon) &\leqslant \sum_{t=N}^{n} \mathbf{1}\{\pi(t+1) = i, \mu^i + \epsilon + g(t)/T_\pi^i(t) \geqslant \mu^* - \epsilon\} \\ &= \sum_{t=N}^{n} \mathbf{1}\{\pi(t+1) = i, g(t)/((\mu^* - \mu_i) - 2\epsilon) \geqslant T_\pi^i(t)\} \\ &\leqslant \sum_{t=N}^{n} \mathbf{1}\{\pi(t+1) = i, g(n)/((\mu^* - \mu_i) - 2\epsilon) \geqslant T_\pi^i(t)\} \\ &\leqslant \frac{g(n)}{(\mu^* - \mu_i) - 2\epsilon} + 1. \end{aligned} \tag{B.3}$$

The last inequality comes from viewing $T_\pi^i(t)$ as a sum of $\mathbf{1}\{\pi(t+1) = i\}$ indicators, and seeing that the condition on it bounds the number of non-zero terms in this sum.

For the second term,

$$\begin{aligned} n_2^i(n,\epsilon) &\leqslant \sum_{t=N}^{n} \mathbf{1}\{\pi(t+1) = i, \bar{X}^i_{T_\pi^i(t)} > \mu^i + \epsilon\} \\ &= \sum_{t=N}^{n} \sum_{k=1}^{t} \mathbf{1}\{\pi(t+1) = i, \bar{X}_k^i > \mu^i + \epsilon, T_\pi^i(t) = k\} \end{aligned}$$

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

$$
\begin{aligned}
&= \sum_{t=N}^{n} \sum_{k=1}^{t} \mathbf{1}\{\pi(t+1) = i, T_\pi^i(t) = k\} \mathbf{1}\{\bar{X}_k^i > \mu^i + \epsilon\} \\
&\leqslant \sum_{k=1}^{n} \mathbf{1}\{\bar{X}_k^i > \mu^i + \epsilon\} \sum_{t=k}^{n} \mathbf{1}\{\pi(t+1) = i, T_\pi^i(t) = k\} \\
&\leqslant \sum_{k=1}^{n} \mathbf{1}\{\bar{X}_k^i > \mu^i + \epsilon\}. \qquad \textbf{(B.4)}
\end{aligned}
$$

The last inequality holds as, for a given k, $\{\pi(t+1) = i, T_\pi^i(t) = k\}$ may be true for only one t. Taking it one step further, we have

$$
n_2^i(n, \epsilon) \leqslant \sum_{k=1}^{\infty} \mathbf{1}\{\bar{X}_k^i > \mu^i + \epsilon\}, \qquad \textbf{(B.5)}
$$

and since the strong law of large numbers is taken to hold, we have therefore that $n_2^i(n)$ is almost surely bound by a finite constant, for all $n \geqslant K$.

For the third term, note that from the structure of the policy, a population is only sampled if it has the maximal current index. Hence, if $\pi(t+1) = i$, it must be true that $u_{i^*}(t, T_\pi^{i^*}(t)) \leqslant u_i(t, T_\pi^i(t))$. Hence we have the bound,

$$
\begin{aligned}
n_3^i(n, \epsilon) &\leqslant \sum_{t=N}^{n} \mathbf{1}\{\pi(t+1) = i, u_{i^*}(t, T_\pi^{i^*}(t)) < \mu^* - \epsilon\} \\
&\leqslant \sum_{t=N}^{n} \mathbf{1}\{u_{i^*}(t, T_\pi^{i^*}(t)) < \mu^* - \epsilon\} \qquad \textbf{(B.6)} \\
&\leqslant \sum_{t=N}^{\infty} \mathbf{1}\{u_{i^*}(t, T_\pi^{i^*}(t)) < \mu^* - \epsilon\}.
\end{aligned}
$$

From the prior observation about the form of the index, Lemma 12, we have that $u_{i^*}(t, T_\pi^{i^*}(t)) < \mu^* - \epsilon$ is true for only finitely many t, almost surely. Hence, from the above bound, $n_3^i(n)$ is almost surely bound by a finite constant, for all $n \geqslant K$.

Combining the above results bounding n_1^i, n_2^i, n_3^i with Eq. (B.2), and observing too that $T_\pi^i(n) \leqslant T_\pi^i(n+1)$, we have that almost surely there exists some finite C_ϵ^i such that for all $n \geqslant K$,

$$
T_\pi^i(n) \leqslant \frac{g(n)}{(\mu^* - \mu_i) - 2\epsilon} + C_\epsilon^i. \qquad \textbf{(B.7)}
$$

■

Proof of Proposition 8: Define a constant $P_\Delta = \sum_{i \neq i^*} 1/(\mu^* - \mu_i)$. Taking $\epsilon < \min_{j \neq i^*}(\mu^* - \mu_j)/2$, we may apply Proposition 7 to yield for each $i \neq i^*$, $\exists$ (a.s.) a finite N_ϵ^i such that $T_\pi^i(n) \leqslant (1+\epsilon)g(n)/(\mu^* - \mu_i)$ for all $n \geqslant N_\epsilon^i$. Taking $N_\epsilon = \max_{i \neq i^*} N_\epsilon^i$, summing over these relations and taking $n \geqslant N_\epsilon$,

$$
\sum_{i \neq i^*} T_\pi^i(n) \leqslant (1+\epsilon) g(n) P_\Delta. \qquad \textbf{(B.8)}
$$

The sum above equals the number of activations of sub-optimal bandits up to and including time n. As the total number of bandit activations up to time n is n, we have from the above that $T_\pi^{i^*}(n) \geqslant n - O(g(n))$.

Trivially from this, the optimal bandit i^* is activated infinitely often, approaching full density of activations as n increases.

Given this linear lower bound on $T_\pi^{i^*}$, it follows that $u_{i^*}(n, T_\pi^{i^*}(n))$ converges to μ^*, almost surely. Hence, almost surely there exists a finite $\tilde{N}_\epsilon$ such that for $n \geqslant \tilde{N}_\epsilon$, $u_{i^*}(n, T_\pi^{i^*}(n)) \leqslant \mu^* + \epsilon$.

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

As under this policy a bandit is only activated when it has the maximal index, it follows that infinitely often (on the activations of i^*), the indices of all sub-optimal bandits are at most $\mu^* + \epsilon$. Given the structure of the indices, it follows that these sub-optimal bandits must be activated infinitely often as well. Hence, almost surely, $T_\pi^i(n)$ increases without bound, for all i. Applying the strong law here, since there are finitely many bandits being considered, $\exists$ (a.s.) a finite "ϵ-trapping time", $\tilde{N}_\epsilon^{\text{trap}}$, such that

$$\bar{X}^i_{T_\pi^i(n)} \in [\mu_i - \epsilon, \mu_i + \epsilon], \quad \forall n \geqslant \tilde{N}_\epsilon^{\text{trap}} \quad \textit{and } \forall i.$$

Let $\{n_k\}_{k \geqslant 0}$ be the infinite sequence of times at which bandit i^* has the current optimal index (and hence is activated next). For a given $i \neq i^*$, we have that for all sufficiently large k ($n_k \geqslant \tilde{N}_\epsilon^{\text{trap}}$),

$$\begin{aligned}
\max_{n_k \leqslant n \leqslant n_{k+1}} u_i(n, T_\pi^i(n)) &\leqslant (\mu_i + \epsilon) + \frac{g(n_{k+1})}{T_\pi^i(n_k)} \\
&= (\mu_i + \epsilon) + \frac{g(n_{k+1})}{g(n_k)} \frac{g(n_k)}{T_\pi^i(n_k)} \\
&= (\mu_i + \epsilon) + \frac{g(n_{k+1})}{g(n_k)} (u_i(n_k, T_\pi^i(n_k)) - \bar{X}^i_{T_\pi^i(n_k)}) \\
&\leqslant (\mu_i + \epsilon) + \frac{g(n_{k+1})}{g(n_k)} (u_i(n_k, T_\pi^i(n_k)) - (\mu_i - \epsilon)).
\end{aligned} \tag{B.9}$$

Additionally, however, at time n_k bandit i^* has the largest index. For sufficiently large k ($n_k \geqslant \tilde{N}_\epsilon$), this index must be at most $\mu^* + \epsilon$. Hence for $n_k > \max(\tilde{N}_\epsilon, \tilde{N}_\epsilon^{\text{trap}})$, for $i \neq i^*$ we have that $u_i(n_k, T_\pi^i(n_k)) \leqslant u_{i^*}(n_k, T_\pi^{i^*}(n_k)) \leqslant \mu^* + \epsilon$, and

$$\begin{aligned}
\max_{n_k \leqslant n \leqslant n_{k+1}} u_i(n, T_\pi^i(n)) &\leqslant (\mu_i + \epsilon) + \frac{g(n_{k+1})}{g(n_k)} ((\mu^* + \epsilon) - (\mu_i - \epsilon)) \\
&= (\mu_i + \epsilon) + \frac{g(n_{k+1})}{g(n_k)} (\mu^* - \mu_i + 2\epsilon).
\end{aligned} \tag{B.10}$$

Since we took g to be concave, $g(n_{k+1}) \leqslant g(n_k) + (n_{k+1} - n_k) g'(n_k)$. The difference $n_{k+1} - n_k - 1$ is the number of sub-optimal bandit activations between the k and $k+1$-th activations of bandit i^*. This is bound from above by the *total* number of sub-optimal activations prior to time n_{k+1}, which by Eq. (B.8) is at most $(1+\epsilon) g(n_{k+1}) P_\Delta$ for all $n_{k+1} \geqslant N_\epsilon$. Hence,

$$g(n_{k+1}) \leqslant g(n_k) + ((1+\epsilon) g(n_{k+1}) P_\Delta + 1) g'(n_k). \tag{B.11}$$

As $g' \to 0$, for all sufficiently large k, we have that $(1+\epsilon) P_\Delta g'(n_k) < 1$ and

$$\frac{g(n_{k+1})}{g(n_k)} \leqslant \frac{1 + ((g'(n_k))/g(n_k)))}{1 - (1+\epsilon) P_\Delta g'(n_k)}. \tag{B.12}$$

As g is taken to be increasing, and g' is taken to limit to 0, we have from the above that there is some finite $\tilde{N}_\epsilon^g$ such that for all sufficiently large k ($n_k \geqslant N_\epsilon^g$), $g(n_{k+1})/g(n_k) \leqslant 1 + \epsilon$. Hence, for $n_k \geqslant \max(N_\epsilon, \tilde{N}_\epsilon, \tilde{N}_\epsilon^{\text{trap}}, \tilde{N}_\epsilon^g)$,

$$\max_{n_k \leqslant n \leqslant n_{k+1}} u_i(n, T_\pi^i(n)) \leqslant (\mu_i + \epsilon) + (1+\epsilon)(\mu^* - \mu_i + 2\epsilon). \tag{B.13}$$

Let $N_\epsilon^K = \min\{n_k : n_k > \max(N_\epsilon, \tilde{N}_\epsilon, \tilde{N}_\epsilon^{\text{trap}}, \tilde{N}_\epsilon^g)\} < \infty$. As the upper bound above no longer depends on k, we have that for $n \geqslant N_\epsilon^K$,

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

$$u_i(n, T_\pi^i(n)) \leqslant (\mu_i + \epsilon) + (1+\epsilon)(\mu^* - \mu_i + 2\epsilon). \tag{B.14}$$

Observing that $\bar{X}^i_{T^i_\pi(n)} \geqslant \mu_i - \epsilon$, the above yields $\mu_i - \epsilon + g(n)/T^i_\pi(n) \leqslant (\mu_i + \epsilon) + (1+\epsilon)(\mu^* - \mu_i + 2\epsilon)$, or

$$\frac{g(n)}{(1+\epsilon)(\mu^* - \mu_i + 2\epsilon) + 2\epsilon} \leqslant T^i_\pi(n). \tag{B.15}$$

■

APPENDIX C. PROOFS OF PROPOSITIONS 10 AND 11

We present the following preliminary bounds to aid in the proofs of Propositions 10, 11. In this section, π is taken to be an **g-ISM** index policy as in Eq. 23. Additionally, it is convenient to define

$$P_\Delta = \sum_{i \neq i^*} \frac{1}{\mu^* - \mu_i}. \tag{C.1}$$

It follows from Propositions 7, 8 that for any $\epsilon > 0$, $\exists$ (a.s.) some finite N_ϵ such that for $n \geqslant N_\epsilon$, the following holds: for $i \neq i^*$,

$$\frac{1-\epsilon}{\mu^* - \mu_i} g(n) \leqslant T^i_\pi(n) \leqslant \frac{1+\epsilon}{\mu^* - \mu_i} g(n). \tag{C.2}$$

And similarly, for the optimal bandit,

$$n - (1+\epsilon)P_\Delta g(n) \leqslant T^{i^*}_\pi(n) \leqslant n - (1-\epsilon)P_\Delta g(n). \tag{C.3}$$

To simplify the case for the optimal bandit, slightly, it also holds that for all sufficiently large n, $T^{i^*}_\pi(n) \geqslant n/2$. We will also observe here, as an aside, that for some finite $\tilde{N}_\epsilon$,

$$(1-\epsilon)/(\mu^* - \mu_i) g(n) > 6, \quad \text{for all } n \geqslant \tilde{N}_\epsilon, \quad \text{and } i \neq i^*.$$

As each bandit is activated infinitely often, T^i_π increases without bound with n, and hence we may apply the Law of the Iterated Logarithm in the following way: There exists a finite time N'_ϵ such that for $n \geqslant N'_\epsilon$, for each bandit i,

$$|\bar{X}^i_{T^i_\pi(n)} - \mu_i| \leqslant \sigma_i \sqrt{2}(1+\epsilon)\sqrt{\frac{\ln \ln T^i_\pi(n)}{T^i_\pi(n)}}. \tag{C.4}$$

However, since $\sqrt{\ln \ln x / x}$ is decreasing for all $x \geqslant 6$, we may apply the above bounds to have that, for $n \geqslant \max(N_\epsilon, N'_\epsilon, \tilde{N}_\epsilon, 12)$, for $i \neq i^*$,

$$|\bar{X}^i_{T^i_\pi(n)} - \mu_i| \leqslant \sigma_i \sqrt{2}(1+\epsilon)\sqrt{\frac{\ln \ln (((1-\epsilon)/(\mu^* - \mu_i)) g(n))}{((1-\epsilon)/(\mu^* - \mu_i)) g(n)}}, \tag{C.5}$$

and for the optimal bandit,

$$|\bar{X}^{i^*}_{T^{i^*}_\pi(n)} - \mu^*| \leqslant \sigma_{i^*} \sqrt{2}(1+\epsilon)\sqrt{\frac{\ln \ln (n/2)}{n/2}}. \tag{C.6}$$

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

Proof of Proposition 10: Let $1 > \epsilon > 0$. For $i \neq i^*$, let

$$h_i(t) = \sigma_i \sqrt{2(\mu^* - \mu_i)} \frac{(1+\epsilon)^2}{\sqrt{1-\epsilon}} \sqrt{\frac{\ln \ln g(t)}{g(t)}}. \tag{C.7}$$

Observe that $h_i \to 0$ from above as $t \to \infty$. Note that there exists a $T_\epsilon < \infty$ such that for $t \geqslant T_\epsilon$, $g(t)/(\mu^* - \mu_i - 2h_i(t))$ is increasing. The proof proceeds analogously to the proof of Proposition 7, utilizing the improved iterated logarithm bounds above.

For $n \geqslant T_\epsilon$, define the following functions:

$$\begin{aligned}
\tilde{n}_1^i(n) &= \sum_{t=T_\epsilon}^{n} \mathbf{1}\{\pi(t+1) = i, u_i(t, T_\pi^i(t)) \geqslant \mu^* - h_i(t), \bar{X}^i_{T_\pi^i(t)} \leqslant \mu_i + h_i(t)\} \\
\tilde{n}_2^i(n) &= \sum_{t=T_\epsilon}^{n} \mathbf{1}\{\pi(t+1) = i, u_i(t, T_\pi^i(t)) \geqslant \mu^* - h_i(t), \bar{X}^i_{T_\pi^i(t)} > \mu_i + h_i(t)\} \\
\tilde{n}_3^i(n) &= \sum_{t=T_\epsilon}^{n} \mathbf{1}\{\pi(t+1) = i, u_i(t, T_\pi^i(t)) < \mu^* - h_i(t)\}.
\end{aligned} \tag{C.8}$$

Hence, we have the following relationship, that for $n \geqslant T_\epsilon$,

$$T_\pi^i(n) \leqslant T_\epsilon + 1 + \tilde{n}_1^i(n) + \tilde{n}_2^i(n) + \tilde{n}_3^i(n). \tag{C.9}$$

The proof proceeds as in the proof of Proposition 7, bounding each of the three terms. For the first,

$$\begin{aligned}
\tilde{n}_1^i(n) &\leqslant \sum_{t=T_\epsilon}^{n} \mathbf{1}\{\pi(t+1) = i, \mu_i + h_i(t) + g(t)/T_\pi^i(t) \geqslant \mu^* - h_i(t)\} \\
&= \sum_{t=T_\epsilon}^{n} \mathbf{1}\{\pi(t+1) = i, g(t)/((\mu^* - \mu_i) - 2h_i(t)) \geqslant T_\pi^i(t)\} \\
&\leqslant \sum_{t=T_\epsilon}^{n} \mathbf{1}\{\pi(t+1) = i, g(n)/((\mu^* - \mu_i) - 2h_i(n)) \geqslant T_\pi^i(t)\} \\
&\leqslant \frac{g(n)}{(\mu^* - \mu_i) - 2h_i(n)} + 1.
\end{aligned} \tag{C.10}$$

As before, the last inequality comes from viewing $T_\pi^i(t)$ as a sum of $\mathbf{1}\{\pi(t+1) = i\}$ indicators, and seeing that the condition on it bounds the number of non-zero terms in this sum. It is also important to observe here that we are explicitly in a regime in which $g(t)/((\mu^* - \mu_i) - 2h_i(t))$ is an increasing function with t.

For the second term,

$$\begin{aligned}
\tilde{n}_2^i(n) &\leqslant \sum_{t=T_\epsilon}^{n} \mathbf{1}\{\pi(t+1) = i, \bar{X}^i_{T_\pi^i(t)} > \mu_i + h_i(t)\} \\
&\leqslant \sum_{t=T_\epsilon}^{n} \mathbf{1}\{\bar{X}^i_{T_\pi^i(t)} - \mu_i > h_i(t)\} \\
&\leqslant \sum_{t=T_\epsilon}^{n} \mathbf{1}\left\{\sigma_i \sqrt{2}(1+\epsilon) \sqrt{\frac{\ln \ln(((1-\epsilon)/(\mu^* - \mu_i))g(t))}{((1-\epsilon)/(\mu^* - \mu_i))g(t)}} > h_i(t)\right\}.
\end{aligned} \tag{C.11}$$

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

The last inequality holds, by the iterated logarithm bound in Eq. (C.5). Taking it one step further, we have

$$\tilde{n}_2^i(n) \leqslant \sum_{t=T_\epsilon}^{\infty} \mathbf{1}\left\{ \frac{\sigma_i\sqrt{2}(1+\epsilon)}{h_i(t)} \sqrt{\frac{\ln\ln(((1-\epsilon)/(\mu^*-\mu_i))g(t))}{((1-\epsilon)/(\mu^*-\mu_i))g(t)}} > 1 \right\}. \tag{C.12}$$

Note that as

$$\lim_t \frac{\sigma_i\sqrt{2}(1+\epsilon)}{h_i(t)} \sqrt{\frac{\ln\ln(((1-\epsilon)/(\mu^*-\mu_i))g(t))}{((1-\epsilon)/(\mu^*-\mu_i))g(t)}} = \frac{1}{1+\epsilon} < 1, \tag{C.13}$$

the event indicated in the above sum bounding $\tilde{n}_2^i(n)$ may occur only finitely many times, almost surely. Hence, $\tilde{n}_2^i(n)$ is almost surely bound by a finite constant, for all $n \geqslant T_\epsilon$.

For the third term, as before, by the structure of the policy, a population is only sampled if it has the maximal current index. Hence, if $\pi(t+1) = i$, it must be true that $u_{i^*}(t, T_\pi^{i^*}(t)) \leqslant u_i(t, T_\pi^i(t))$. It follows that

$$\begin{aligned}
\tilde{n}_3^i(n) &\leqslant \sum_{t=T_\epsilon}^{n} \mathbf{1}\{\pi(t+1) = i, u_{i^*}(t, T_\pi^{i^*}(t)) < \mu^* - h_i(t)\} \\
&\leqslant \sum_{t=T_\epsilon}^{n} \mathbf{1}\{u_{i^*}(t, T_\pi^{i^*}(t)) < \mu^* - h_i(t)\} \\
&= \sum_{t=T_\epsilon}^{n} \mathbf{1}\left\{ \bar{X}^{i^*}_{T_\pi^{i^*}(t)} + \frac{g(t)}{T_\pi^{i^*}(t)} < \mu^* - h_i(t) \right\} \\
&\leqslant \sum_{t=T_\epsilon}^{n} \mathbf{1}\left\{ -\sigma_{i^*}\sqrt{2}(1+\epsilon)\sqrt{\frac{\ln\ln(t/2)}{t/2}} + \frac{g(t)}{T_\pi^{i^*}(t)} < -h_i(t) \right\},
\end{aligned} \tag{C.14}$$

the last equation coming from the iterated logarithm bound for the optimal bandit, Eq. (C.6). As a final simplification,

$$\tilde{n}_3^i(n) \leqslant \sum_{t=T_\epsilon}^{\infty} \mathbf{1}\left\{ -\sigma_{i^*}\sqrt{2}(1+\epsilon)\sqrt{\frac{\ln\ln(t/2)}{t/2}} < -h_i(t) \right\}. \tag{C.15}$$

If $g(n) = o(n/\ln\ln n)$, it is easy to verify that the indicated event in the above sum can only occur for finitely many t. Hence, by the above, there is a finite constant bounding $\tilde{n}_3^i(n)$ for all $n \geqslant T_\epsilon$.

Combining the above results, there is a finite constant D_i^ϵ such that for all $n \geqslant T_\epsilon$,

$$T_\pi^i(n) \leqslant \frac{g(n)}{(\mu^* - \mu_i) - 2h_i(n)} + D_i^\epsilon. \tag{C.16}$$

We have from this that

$$(\mu^* - \mu_i)T_\pi^i(n) - g(n) \leqslant g(n)\frac{2h_i(n)}{(\mu^* - \mu_i) - 2h_i(n)} + (\mu^* - \mu_i)D_i^\epsilon. \tag{C.17}$$

For a fixed $\epsilon > 0$, the above yields (taking the limit, given the choice of $h_i(n)$),

$$\limsup_n \frac{(\mu^* - \mu_i)T_\pi^i(n) - g(n)}{\sqrt{g(n)\ln\ln g(n)}} \leqslant \frac{2\sigma_i\sqrt{2}(1+\epsilon)^2}{\sqrt{\mu^* - \mu_i}\sqrt{1-\epsilon}}. \tag{C.18}$$

As the above holds for all $\epsilon > 0$, this yields, almost surely,

$$\limsup_n \frac{(\mu^* - \mu_i)T_\pi^i(n) - g(n)}{\sqrt{g(n)\ln\ln g(n)}} \leqslant \frac{2\sigma_i\sqrt{2}}{\sqrt{\mu^* - \mu_i}}. \tag{C.19}$$

■

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

Proof of Proposition 11: Let $\epsilon \in (0,1)$. Recall from the proof of Proposition 8 the infinite sequence $\{n_k\}_{k\geqslant 0}$ of times at which the index of the optimal bandit i^* is maximal. For notational convenience, we will write $u_i(n) = u_i(n, T_\pi^i(n))$, and for $i \neq i^*$, we define

$$U_k^i = \max_{n_k \leqslant n \leqslant n_{k+1}} u_i(n), \tag{C.20}$$

and

$$M_k^i = \max_{n_k \leqslant n \leqslant n_{k+1}} \bar{X}^i_{T_\pi^i(n)}. \tag{C.21}$$

We have the following relations,

$$\begin{aligned} U_k^i &\leqslant \left(\max_{n_k \leqslant n' \leqslant n_{k+1}} \bar{X}^i_{T_\pi^i(n')} \right) + \frac{g(n_{k+1})}{T_\pi^i(n_k)} \\ &= M_k^i + \frac{g(n_{k+1})}{g(n_k)} \frac{g(n_k)}{T_\pi^i(n_k)} \\ &= M_k^i + \frac{g(n_{k+1})}{g(n_k)} (u_i(n_k) - \bar{X}^i_{T_\pi^i(n_k)}) \\ &\leqslant M_k^i + \frac{g(n_{k+1})}{g(n_k)} (u_{i^*}(n_k) - \bar{X}^i_{T_\pi^i(n_k)}). \end{aligned} \tag{C.22}$$

For n such that $n_k \leqslant n \leqslant n_{k+1}$, trivially $u_i(n) \leqslant U_k^i$. It follows that

$$\frac{g(n)}{T_\pi^i(n)} \leqslant (M_k^i - \bar{X}^i_{T_\pi^i(n)}) + \frac{g(n_{k+1})}{g(n_k)} (u_{i^*}(n_k) - \bar{X}^i_{T_\pi^i(n_k)}). \tag{C.23}$$

Defining the following terms for space,

$$\begin{aligned} A_{n,k} &= (M_k^i - \bar{X}^i_{T_\pi^i(n)}), \\ B_k &= \frac{g(n_{k+1})}{g(n_k)} u_{i^*}(n_k) - \mu^*, \\ C_k &= \frac{g(n_{k+1})}{g(n_k)} \bar{X}^i_{T_\pi^i(n_k)} - \mu_i, \\ \Delta(n) &= g(n) - (\mu^* - \mu_i) T_\pi^i(n). \end{aligned} \tag{C.24}$$

The above relation may be rearranged to yield

$$\Delta(n)/T_\pi^i(n) \leqslant A_{n,k} + B_k - C_k. \tag{C.25}$$

We may apply the iterated logarithm bounds of Eq. (C.5), to yield a finite K_A such that for $k \geqslant K_A$,

$$A_{n,k} \leqslant 2\sigma_i \sqrt{2}(1+\epsilon) \sqrt{\frac{\ln\ln(((1-\epsilon)/(\mu^* - \mu_i)) g(n_k))}{((1-\epsilon)/(\mu^* - \mu_i)) g(n_k)}}. \tag{C.26}$$

Similarly, there is a finite K_B such that for $k \geqslant K_B$, observing that for sufficiently large k, $T_\pi^{i^*}(n_k) \geqslant n_k/2$,

$$B_k \leqslant \frac{g(n_{k+1})}{g(n_k)} \left(\mu^* + \sigma_{i^*} \sqrt{2}(1+\epsilon) \sqrt{\frac{\ln\ln(n_k/2)}{n_k/2}} + \frac{g(n_k)}{n_k/2} \right) - \mu^*. \tag{C.27}$$

And finally, there is a finite K_C such that for $k \geqslant K_C$,

$$C_k \geqslant \frac{g(n_{k+1})}{g(n_k)} \left(\mu_i - \sigma_i \sqrt{2}(1+\epsilon) \sqrt{\frac{\ln\ln(((1-\epsilon)/(\mu^* - \mu_i)) g(n_k))}{((1-\epsilon)/(\mu^* - \mu_i)) g(n_k)}} \right) - \mu_i. \tag{C.28}$$

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

Rearranging terms for space again, for $k \geqslant \max(K_A, K_B, K_C)$ we have

$$\Delta(n)/T_\pi^i(n) \leqslant A_{n,k} + B_k - C_k \leqslant \tilde{A}_k + \tilde{B}_k + \tilde{C}_k + \tilde{D}_k, \tag{C.29}$$

where

$$\begin{aligned}
\tilde{A}_k &= (\mu^* - \mu_i)\left(\frac{g(n_{k+1})}{g(n_k)} - 1\right) \\
\tilde{B}_k &= \sigma_i\sqrt{2}(1+\epsilon)\left(2 + \frac{g(n_{k+1})}{g(n_k)}\right)\sqrt{\frac{\ln\ln(((1-\epsilon)/(\mu^* - \mu_i))g(n_k))}{((1-\epsilon)/(\mu^* - \mu_i))g(n_k)}} \\
\tilde{C}_k &= \sigma_{i^*}\sqrt{2}(1+\epsilon)\frac{g(n_{k+1})}{g(n_k)}\sqrt{\frac{\ln\ln(n_k/2)}{n_k/2}} \\
\tilde{D}_k &= \frac{g(n_{k+1})}{g(n_k)}\frac{g(n_k)}{n_k/2}.
\end{aligned} \tag{C.30}$$

Noting that each of the above are positive, we have from Eq. (C.29),

$$\frac{\Delta(n)}{\sqrt{g(n)\ln\ln g(n)}} \leqslant \frac{(\tilde{A}_k + \tilde{B}_k + \tilde{C}_k + \tilde{D}_k)T_\pi^i(n)}{\sqrt{g(n)\ln\ln g(n)}}. \tag{C.31}$$

Note that, applying Eq. (C.2) in this case, we have some finite K_ϵ such that for $k \geqslant K_\epsilon$,

$$T_\pi^i(n) \leqslant T_\pi^i(n_{k+1}) \leqslant \frac{1+\epsilon}{\mu^* - \mu_i} g(n_{k+1}). \tag{C.32}$$

Recall from the proof of Proposition 8 that there is a finite K'_ϵ such that for $k \geqslant K'_\epsilon$, $g(n_{k+1}) \leqslant (1+\epsilon)g(n_k)$. Noting too that $g(n_k) \leqslant g(n)$, we have that for $k \geqslant \max(K_\epsilon, K'_\epsilon)$,

$$\frac{\Delta(n)}{\sqrt{g(n)\ln\ln g(n)}} \leqslant \frac{(\tilde{A}_k + \tilde{B}_k + \tilde{C}_k + \tilde{D}_k)}{\sqrt{g(n_k)\ln\ln g(n_k)}}\frac{(1+\epsilon)^2}{(\mu^* - \mu_i)} g(n_k). \tag{C.33}$$

We have

$$\begin{aligned}
\frac{\tilde{D}_k g(n_k)}{\sqrt{g(n_k)\ln\ln g(n_k)}} &= \frac{g(n_{k+1})}{g(n_k)}\frac{g(n_k)}{n_k/2}\frac{g(n_k)}{\sqrt{g(n_k)\ln\ln g(n_k)}} \\
&\leqslant 2(1+\epsilon)\frac{g(n_k)^{3/2}}{n_k\sqrt{\ln\ln g(n_k)}} \\
&= o(1).
\end{aligned} \tag{C.34}$$

The last relationship follows, taking $g(n) = o(n^{2/3})$.

We have

$$\begin{aligned}
\frac{\tilde{C}_k g(n_k)}{\sqrt{g(n_k)\ln\ln g(n_k)}} &= 2\sigma_{i^*}(1+\epsilon)\frac{g(n_{k+1})}{g(n_k)}\sqrt{\frac{\ln\ln(n_k/2)}{n_k}\frac{g(n_k)}{\ln\ln g(n_k)}} \\
&\leqslant 2\sigma_{i^*}(1+\epsilon)^2\sqrt{\frac{\ln\ln(n_k/2)}{n_k}\frac{g(n_k)}{\ln\ln g(n_k)}} \\
&= o(1).
\end{aligned} \tag{C.35}$$

The last relationship follows, taking $g(n) = o(n/\ln\ln n)$.

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press

We have

$$\begin{aligned}\frac{\tilde{B}_k g(n_k)}{\sqrt{g(n_k)\ln\ln g(n_k)}} &= \sigma_i\sqrt{2}(1+\epsilon)\left(2+\frac{g(n_{k+1})}{g(n_k)}\right)\sqrt{\frac{\ln\ln(((1-\epsilon)/(\mu^*-\mu_i))g(n_k))}{((1-\epsilon)/(\mu^*-\mu_i))g(n_k)}}\sqrt{\frac{g(n_k)}{\ln\ln g(n_k)}} \\ &\leqslant \frac{\sigma_i\sqrt{2}(1+\epsilon)(3+\epsilon)}{\sqrt{((1-\epsilon)/(\mu^*-\mu_i))}}\sqrt{\frac{\ln\ln(((1-\epsilon)/(\mu^*-\mu_i))g(n_k))}{\ln\ln g(n_k)}} \\ &= \frac{\sigma_i\sqrt{2}(1+\epsilon)(3+\epsilon)}{\sqrt{((1-\epsilon)/(\mu^*-\mu_i))}}(1+o(1)).\end{aligned} \tag{C.36}$$

The last relationship follows, taking the $\{n_k\}_{k\geqslant 0}$ as infinite and unbounded, and g as increasing and unbounded.

We have

$$\frac{\tilde{A}_k g(n_k)}{\sqrt{g(n_k)\ln\ln g(n_k)}} = (\mu^*-\mu_i)\left(\frac{g(n_{k+1})}{g(n_k)}-1\right)\sqrt{\frac{g(n_k)}{\ln\ln g(n_k)}}. \tag{C.37}$$

Let $\delta > 1$ by fixed. We use the bound here that for all positive $x \leqslant 1-1/\delta$, $1/(1-x) \leqslant 1+\delta x$. Applying Eq. (B.12), we have for sufficiently large k,

$$\begin{aligned}\frac{g(n_{k+1})}{g(n_k)} - 1 &\leqslant \frac{1+((g'(n_k))/(g(n_k)))}{1-(1+\epsilon)P_\Delta g'(n_k)} - 1 \\ &\leqslant \left(1+\frac{g'(n_k)}{g(n_k)}\right)(1+\delta(1+\epsilon)P_\Delta g'(n_k)) - 1 \\ &= g'(n_k)(\delta(1+\epsilon)P_\Delta + o(1)).\end{aligned} \tag{C.38}$$

The last relationship follows, as $g' \to 0$ and $g \to \infty$ with n_k. Applying this to the above bound,

$$\begin{aligned}\frac{\tilde{A}_k g(n_k)}{\sqrt{g(n_k)\ln\ln g(n_k)}} &\leqslant (\mu^*-\mu_i)(\delta(1+\epsilon)P_\Delta + o(1))g'(n_k)\sqrt{\frac{g(n_k)}{\ln\ln g(n_k)}} \\ &= o(1).\end{aligned} \tag{C.39}$$

The last relationship follows, taking $g(n) = o(n^{2/3})$.

Applying all of the above to the bound in Eq. (C.33), this yields

$$\frac{\Delta(n)}{\sqrt{g(n)\ln\ln g(n)}} \leqslant \left(\frac{\sigma_i\sqrt{2}(1+\epsilon)(3+\epsilon)}{\sqrt{((1-\epsilon)/(\mu^*-\mu_i))}}(1-o(1))+o(1)\right)\frac{(1+\epsilon)^2}{(\mu^*-\mu_i)}, \tag{C.40}$$

or

$$\limsup_n \frac{\Delta(n)}{\sqrt{g(n)\ln\ln g(n)}} \leqslant \left(\frac{\sigma_i\sqrt{2}(1+\epsilon)(3+\epsilon)}{\sqrt{((1-\epsilon)/(\mu^*-\mu_i))}}\right)\frac{(1+\epsilon)^2}{(\mu^*-\mu_i)}. \tag{C.41}$$

Taking the limit as $\epsilon \to 0$ completes the proof,

$$\limsup_n \frac{g(n)-(\mu^*-\mu_i)T_\pi^i(n)}{\sqrt{g(n)\ln\ln g(n)}} \leqslant \frac{3\sigma_i\sqrt{2}}{\sqrt{\mu^*-\mu_i}}. \tag{C.42}$$

■

https://doi.org/10.1017/S0269964818000529 Published online by Cambridge University Press