Structured Best Arm Identification with Fixed Confidence

Ruitong Huang RUITONG@UALBERTA.CA
Mohammad M. Ajallooeian AJALLOOE@UALBERTA.CA
Csaba Szepesvári SZEPESVA@UALBERTA.CA
Martin Müller MMUELLER@UALBERTA.CA
Department of Computing Science
University of Alberta
Edmonton, AB T6G 2E8, Canada

Editors: Steve Hanneke and Lev Reyzin

Abstract

We study the problem of identifying the best action among a set of possible options when the value of each action is given by a mapping from a number of noisy micro-observables in the so-called fixed confidence setting. Our main motivation is the application to minimax game search, which has been a major topic of interest in artificial intelligence. In this paper we introduce an abstract setting to clearly describe the essential properties of the problem. While previous work only considered a two-move-deep game tree search problem, our abstract setting can be applied to the general minimax games where the depth can be non-uniform and arbitrary, and transpositions are allowed. We introduce a new algorithm (LUCB-micro) for the abstract setting, and give its lower and upper sample complexity results. Our bounds recover some previous results, achieved in more limited settings, and also shed further light on how the structure of minimax problems influences sample complexity.

Keywords: Best Arm Identification, Monte-Carlo Tree Search, Game Tree Search, Structured Environments, Multi-Armed Bandits, Minimax Search

1. Introduction

Motivated by the problem of finding the optimal move in minimax tree search with noisy leaf evaluations, we introduce best arm identification problems with structured payoffs and micro-observables. In these problems, the learner's goal is to find the best arm when the payoff of each arm is a fixed and known function of a set of unknown values. In each round, the learner can choose one of the micro-observables to make a noisy measurement (i.e., the learner can obtain a "micro-observation"). We study these problems in the so-called fixed confidence setting.

A special case of this problem is the standard best arm identification, which has seen a flurry of activity during the last decade, e.g., (Even-Dar et al., 2006; Audibert and Bubeck, 2010; Gabillon et al., 2012; Kalyanakrishnan et al., 2012; Karnin et al., 2013; Jamieson et al., 2014; Chen and Li, 2015). Recently, Garivier et al. (2016a) considered the motivating problem mentioned above for the simplest non-trivial instance, when two players alternate for a single round each. One of their main observations is that such two-move problems can be solved more efficiently than if one considers the problem as an instance of a nested best arm identification problem. They proposed two algorithms, one for the fixed confidence setting, the other for the (asymptotic) vanishing confidence setting and provided upper bounds. An implicit (optimization-based) lower bound was also briefly sketched, together with a plan to derive an algorithm that matches it in the vanishing confidence setting.

Our main interest in this paper is to see whether the ideas of Garivier et al. (2016a) extend to more general settings, such as when the depth can be non-uniform and is in particular not limited to two, or when different move histories lead to shared states called transpositions in the language of adversarial search. While considering these extensions, we found it cleaner to introduce the abstract setting mentioned below (Section 2). The motivation here is to clearly delineate the crucial properties of the problem that our results use. For the general structured setting, in Section 3 we prove an instance dependent lower bound along the lines of Auer et al. (2002) or Garivier et al. (2016b). Mild novelty is the way our proof deals with the technical issue that best arm identification algorithms ideally stop and hence their behavior is undefined after the random stopping time). This is then specialized to the minimax game search setting (Section 4), where we show the crucial role of what we call proof sets, which are somewhat reminiscent of the so-called conspiracy sets from adversarial search (McAllester, 1988). Our lower bound matches that of Garivier et al. (2016a) in the case of two-move alternating problems. Considering again the abstract setting, we propose the new algorithm LUCB-micro (Section 5), which can be considered as a natural generalization of Maximin-LUCB of Garivier et al. (2016a) with some minor differences. Under a regularity assumption on the payoff maps, we prove that the algorithm meets the risk-requirement. We also provide a high-probability, instance-dependent upper bound on the algorithm's sample complexity (i.e., on the number of observations the algorithm takes). While this bound meets the general characteristics of existing bounds, it fails to reproduce the corresponding result of Garivier et al. (2016a). To the best of the authors' knowledge, the only comparable algorithm to study best arm identification in a full-length minimax tree search setting (which was the motivating example of our work) is FindTopWinner by Teraoka et al. (2014). This algorithm is a round-based elimination based algorithm with additional pruning steps that come from the tree structure. When we specialize our framework to the minimax game scenario and implement other necessary changes to put our work into their (ϵ, δ)-PAC setting, our upper bound is a strict improvement of theirs, e.g., in the number of samples related to the *near-optimal* micro-observables (leaves of the minimax game tree). Next, we consider the minimax setting (Section 6). First, we show that the regularity assumptions made for the abstract setting are met in this case. We also show how to efficiently compute the choices that LUCB-micro makes using a "min-max" algorithm. Finally, we strengthen our previous result so that it matches the mentioned result of Garivier et al. (2016a). After we submitted our paper, an independent work by Kaufmann and Koolen (2017) was published online, which develops a similar algorithm and essentially the same theoretical guarantee in the "min-max" setting. Our paper presents the results in a general setting, delineating the essence of the problem.

1.1. Notation

We use $\mathbb{N} = \{1, 2, \dots\}$ to denote the set of positive integers, while we let $\mathbb{R}$ denote the set of reals. For a positive integer $k \in \mathbb{N}$, we let $[k] = \{1, \dots, k\}$. For a vector $v \in \mathbb{R}^d$, we denote its i-th element by v_i; though occasionally we will also use $v(i)$ for the same purpose, i.e., we identify $\mathbb{R}^d$ and $\{f : f : [d] \to \mathbb{R}\}$ in the obvious way. We let $|v|$ denote the vector defined by $|v| = (|v_i|)_{i\in[d]}$. For two vectors $u, v \in \mathbb{R}^d$, we define $u \leqslant v$ if and only if $u_i \leqslant v_i$ for all $1 \leqslant i \leqslant d$. Further, we write $u < v$ when $u \neq v$ and $u \leqslant v$. For $B \subset [d]$, we write $u|_B$ to denote the $|B|$-dimensional vector obtained from restricting u to components with index in B: $u|_B = (u_i)_{i\in B}$. We use $\mathbf{1}_d$ to denote the d dimensional vector whose components are all equal to one. For a nonempty set B, we also use $\mathbf{1}_B$ to denote the $|B|$-dimensional all-one vector. We let $B^c = \{i \in [d] : i \notin B\}$ to denote the complementer of B (when B^c is used, the base set that the complementer is taken for should be clear from the context). The indicator function will be denoted by $\mathbb{I}\{\cdot\}$. We will use $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. For $A \subset \mathbb{R}$, $\bar{A}$ denotes its topological closure, while A° denotes its interior. Given a real value $a \in \mathbb{R}$, $a_+ = a \vee 0$ and $a_- = -(a \wedge 0)$. For a sequence $(m_0, \ldots, m_i)$ of some values and some other value m, we define $\text{join}(h, m) = (m_0, \ldots, m_i, m)$.

2. Problem setup

Fix two positive integers, L and K. A problem instance of *structured K-armed best arm identification with L micro-observations* is defined by a tuple (f, P), where $f : \mathbb{R}^L \to \mathbb{R}^K$ and $P = (P_1, \ldots, P_L)$ is an L-tuple of distributions over the reals. We let $\mu_i = \int x dP_i(x)$ denote the mean of distribution P_i. We shall denote the component functions of f by $f_1, \ldots, f_K$: $f(\mu) = (f_1(\mu), \ldots, f_K(\mu))$. The value $f_i(\mu)$ is interpreted as the payoff of arm i and we call f the *reward map*. The goal of the learner is to identify the arm with the highest payoff, which is assumed to be unique. The learner knows f, is unaware of P, and, in particular, unaware of μ. To gain information about μ, the learner can query the distributions in discrete rounds indexed by $t = 1, 2, \ldots$, in a sequential fashion. The learner is also given $\delta \in (0, 1)$, a *risk parameter* (also known as a confidence parameter). The goal of the learner is to identify the arm with the highest payoff using the least number of observations while keeping the probability of making a mistake below the specified risk level. A learner is *admissible* for a given set $\mathcal{S}$ of problem instances if *(i)* for any instance from $\mathcal{S}$, the probability of the learner misidentifying the optimal arm in the instance is below the given fixed risk factor δ; and *(ii)* the learner stops with probability one on any instance from $\mathcal{S}$. The interaction of a learner and a problem instance is shown on Fig. 1.

Input: $\delta \in (0, 1), f = (f_1, \ldots, f_K)$
for $t = 1, 2, \ldots$ **do**
  Choose $I_t \in [L]$
  Observe $Y_t \sim P_{I_t}(\cdot)$
  if Stop() **then**
    Choose $J \in [K]$, candidate optimal arm index
    $T \leftarrow t$
    return (T, J)
Admissibility: $\mathbb{P}(J \neq \text{argmax}_i f_i(\mu)) \leqslant \delta$, $\mathbb{P}(T < \infty) = 1$.

Figure 1: Interaction of a learner and a problem instance (f, P). The components of μ are $\mu_i = \int x dP_i(x)$, $i \in [L]$, and f maps $\mathbb{R}^L$ to $\mathbb{R}^K$.

Minimax games As a motivating example, consider the problem of finding the optimal move for the first player in a finite two-player minimax game. The game is finite because the game finishes in finitely many steps (by reaching one of the L possible terminating states).

The first player has K moves in the starting position. The value of each move is a function of the values $\mu \in \mathbb{R}^L$ of the L possible terminating states.

Formally, such a minimax game is described by $G = (M, H, p, \tau)$, where M is a non-empty finite set of possible moves, $H \subset \cup_{n \geqslant 0} M^n$ is a finite set of (feasible) histories of moves, the function $p : H \to \{-1, +1\}$ determines, for each feasible move history, the identity of the player on turn, and τ is a surjection that maps a subset $H_{\max} \subset H$ of move histories, the set of maximal move histories in H, to $[L]$ (in particular, note that τ may map multiple maximal histories to the same terminating state). An element h of H is maximal in H if it is not the prefix of any other history $h' \in H$, or, in other words, if it has no continuation in H. The set H has the property that if $h \in H$ then every prefix of h with positive length is also is in H. The first player's moves are given by the unit length histories in H. To minimize clutter, without the loss of generality (WLOG), we identify this set with $[K]$.

The function $f = (f_1, \ldots, f_K)$ underlying G gives the payoffs of the first player. To define f we use the auxiliary function $V(\cdot, \mu) : H \to \mathbb{R}$ that evaluates any particular move history given the values μ assigned to terminal states. Given V, we define $f_k(\mu) = V((k), \mu)$ for any $k \in [K]$. It remains to define V: For $h \in H_{\max}$, $V(h, \mu) = \mu_{\tau(h)}$. For any other feasible history $h \in H$, $V(h, \mu) = p(h) \max\{p(h) V(h', \mu) : h' \in \mathrm{H}_{\mathrm{succ}}(h)\}$, where $\mathrm{H}_{\mathrm{succ}}(h) = \{\mathrm{join}(h, m) : m \in M\} \cap H$ is the set of *immediate successors* of h in H. Thus, when $p(h) = 1$, $V(h, \mu)$ is the maximum of the values associated with the immediate successors of h, while when $p(h) = -1$, $V(h, \mu)$ is the minimum of these values. We define $m(h, \mu)$ as the move m defining an optimal immediate successor of h given μ. Note that many of the defined functions depend on H, but the dependence is suppressed, as we will keep H fixed. One natural problem that fits our setting is a (small) game where the payoffs at the terminating states of a game are themselves randomized (e.g., at the end of a game some random hidden information such as face down cards can decide the value of the final state) or use noisy randomized evaluations as in MCTS. As explained by Garivier et al. (2016a), the setting may also shed light on how to design better Monte-Carlo Tree Search (MCTS) algorithms, which is a relatively novel class of search algorithms that proved to be highly successful in recent years (e.g., Gelly et al., 2012; Silver et al., 2016).

3. Lower bound: General setting

In this section we will prove a lower bound for the case of a fixed map f and when the set of instances is the set of all normal distributions with unit variance. We denote the corresponding set of instances by $\mathcal{S}_f^{\mathrm{norm}}$. Our results can be easily extended to the case of other sufficiently-rich families of distributions.

For the next result, assume without loss of generality that $f_1(\mu) > f_2(\mu) \geqslant \cdots \geqslant f_K(\mu)$. Fix a learner (policy) A, which maps historical observations to actions. For simplicity, we assume that A is deterministic (the extension to randomized algorithms is standard). Let $\Omega = ([L] \times \mathbb{R})^{\mathbb{N}}$ be the set of (infinite) sequences of observable-index and observation pairs so that for any $\omega = (i_1, y_1, i_2, y_2, \ldots) \in \Omega$, $t \geqslant 1$, $I_t(\omega) = i_t$ and $Y_t(\omega) = y_t$. We equip Ω with the associated Lebesgue σ-algebra $\mathcal{F}$. For an infinite sequence $\omega = (i_1, y_1, i_2, y_2, \ldots) \in \Omega$, we let $T(\omega) \in \mathbb{N} \cup \{\infty\}$ be the round index when the algorithm stops (we let $T(\omega) = \infty$ if the algorithm never stops on ω). Thus, $T : \Omega \to \mathbb{N} \cup \{\infty\}$. Similarly, define $J : \Omega \to [K + 1]$

to be the choice of the algorithm when it stops, where we define $J(\omega) = K+1$ in case $T(\omega) = \infty$.

The interaction of a problem instance (uniquely determined by μ) and the learner (uniquely determined by the associated policy A) induces a unique distribution $\mathbb{P}_{\mu,A}$ over the measurable space $(\Omega, \mathcal{F})$, where we agree that in rounds with index $t = T+1, T+2, \dots$, we specify that the algorithm chooses arm 1, while the observation distributions are modified so that the observation is deterministically set to zero. We will also use $\mathbb{E}_{\mu,A}$ to denote the expectation operator corresponding to $\mathbb{P}_{\mu,A}$.

To appease the prudent reader, let us note that our statements will always be concerned with events that are subsets of the event $\{T < \infty\}$ and as such they are not affected by how we specify the "choices" of the algorithm and the "responses" of the environment for $t > T$. Take, as an example, the expected number of steps that A takes in an environment μ, $\mathbb{E}_{\mu,A}[T]$, which we bound below. Since we bound this only in the case when the algorithm A is admissible, which implies that $\mathbb{P}_{\mu,A}(T < \infty) = 1$, we have $\mathbb{E}_{\mu,A}[T] = \mathbb{E}_{\mu,A}[T\,\mathbb{I}\{T < \infty\}]$. which shows that the behavior of $\mathbb{P}_{\mu,A}$ outside of $\mathbb{P}_{\mu,A}$ outside of $\{T < \infty\}$ is immaterial for this statement. The choices we made for $t > T$ (for the algorithm and the environment) will however be significant in that they simplify a key technical result.

To state our result, we need to introduce the set of *significant departures*, $D_\mu \subset \mathbb{R}^L$, from μ. This set contains all vectors Δ such that the best arm under $\mu + \Delta$ is not arm 1. Formally,

$$D_\mu = \{\Delta \in \mathbb{R}^L \,:\, f_1(\mu + \Delta) \leqslant \max_{i>1} f_i(\mu + \Delta)\}\,. \tag{1}$$

Theorem 1 (Lower bound) *Fix a risk parameter $\delta \in (0,1)$. Assume that A is admissible over the instance set $\mathcal{S}_f^{norm}$ at the risk level δ. Define*

$$\tau^*(\mu) = \min\left\{\sum_{i=1}^{L} n(i) \,:\, \inf_{\Delta \in D_\mu} \sum_{i=1}^{L} n(i)\Delta_i^2 \geqslant 2\log(1/(4\delta)),\, n(1), \dots, n(L) \geqslant 0\right\}. \tag{2}$$

Then, $\mathbb{E}_{\mu,A}[T] \geqslant \tau^*(\mu)$.

The proof can be shown to reproduce the result of Garivier and Kaufmann (2016) (see page 6 of their paper) when the setting is best arm identification. The proof uses standard steps (e.g., Auer et al., 2002; Kaufmann et al., 2016) and one of its main merit is its simplicity. In particular, it relies on two information theoretical results; a high-probability Pinsker inequality (Lemma 2.6 from (Tsybakov, 2008)) and a standard decomposition of divergences. The proof is given in Appendix B (all proofs omitted from the main body can be found in the appendix).

Remark 2 (Minimal significant departures ($D_\mu^{\min}$)) *From the set of significant departures one can remove all vectors d that are componentwise dominating in absolute value some other significant departure $\Delta \in D_\mu$ without effecting the lower bound. To see this, write the lower bound as $\min\{\sum_i n(i) \,:\, n \in \cap_{\Delta \in D_\mu} \Phi(\Delta)\}$, where $\Phi(\Delta) = \{n \in [0,\infty)^L \,:\, \sum_i n(i)\Delta_i^2 \geqslant 2\log(1/(4\delta))\}$. Then, if $d, \Delta \in D_\mu$ are such that $|\Delta| \leqslant |d|$ then $\Phi(\Delta) \subset \Phi(d)$. Hence, $\cap_{\Delta \in D_\mu} \Phi(\Delta) = \cap_{\Delta \in D_\mu^{\min}} \Phi(\Delta)$ where $D_\mu^{\min} = \{d \in D_\mu \,:\, \nexists \Delta \in D_\mu \text{ s.t. } |\Delta| < |d|\}$.*

4. Lower bound for minimax games

In this section we prove a corollary of the general lower bound of the previous section in the context of minimax games; the question being what role the structure of a game plays in the lower bound. For this section fix a minimax game structure $G = (M, H, p, \tau)$ (cf. Section 2). We first need some definitions:

Definition 1 (Proof sets) *Take a minimax game structure $G = (M, H, p, \tau)$ with K first moves and L terminal states. Take $j \in [K]$. A set $B \subset [L]$ is said to be sufficient for proving upper bounds on the value of move j if for any $\mu \in \mathbb{R}^L$ and $\theta \in \mathbb{R}$, $\mu|_B = \theta \mathbf{1}_B$ implies $f_j(\mu) \leqslant \theta$. Symmetrically, a set $B \subset [L]$ is said to be sufficient for proving lower bounds on the value of move j if for any $\mu \in \mathbb{R}^L$ and $\theta \in \mathbb{R}$, $\mu|_B = \theta \mathbf{1}_B$ implies $f_j(\mu) \geqslant \theta$.*

In analogy with proof sets in the minimax search algorithm of Proof Number Search (Allis, 1994; Kishimoto et al., 2012), we will call the sets satisfying the above definition upper (resp., lower) proof sets, denoted by $\mathcal{B}_j^+$ (resp., $\mathcal{B}_j^-$). In Proof Number Search, a proof set is a set of leaves with currently unknown value which, if all nodes in the set are proven, implies a proof of the root. In our case, the upper and lower proof sets establish an upper or lower bound of θ, respectively.

One can obtain upper proof sets that belong to $\mathcal{B}_j^+$ in the following way: Let H_j denote the set of histories that start with move j. Consider a non-empty $\widetilde{H} \subset H_j$ that satisfies the following properties: *(i)* if $h \in \widetilde{H}$ and $p(h) = -1$ (minimizing turn) then $|\,\mathrm{H}_{\mathrm{succ}}(h) \cap \widetilde{H}| = 1$; *(ii)* if $h \in \widetilde{H}$ and $p(h) = 1$ (maximizing turn) then $\mathrm{H}_{\mathrm{succ}}(h) \subset \widetilde{H}$. Call the set of $\widetilde{H}$ that can be obtained this way $\mathcal{H}_j^+$. From the construction of $\widetilde{H}$ we immediately get the following proposition:

Proposition 3 *Take any $\widetilde{H} \in \mathcal{H}_j^+$ as above. Then, $\tau(\widetilde{H} \cap H_{\max}) \in \mathcal{B}_j^+$.*

A similar construction and statement applies in the case of $\mathcal{B}_j^-$, resulting in the set $\mathcal{H}_j^-$. Our next result will imply that the lower bound is achieved by considering departures of a special form, related to proof sets:

Proposition 4 (Minimal significant departures for minimax games) *Without loss of generality, assume that $f_1(\mu) > \max_{j>1} f_j(\mu)$. Let*

$$S = \Big\{ \Delta \in \mathbb{R}^L \,:\, \exists 1 < j \leqslant K\,, \theta \in [f_j(\mu), f_1(\mu)], B \in \mathcal{B}_1^+, B' \in \mathcal{B}_j^- \text{ s.t.}$$
$$\Delta_i = -(\mu_i - \theta)_+,\ \forall i \in B \backslash B';\ \Delta_i = (\mu_i - \theta)_-,\ \forall i \in B' \backslash B;$$
$$\Delta_i = \theta - \mu_i,\ \forall i \in B' \cap B;\ \Delta_i = 0,\ \forall i \in (B \cup B')^c \Big\}.$$

Then, $D_\mu^{\min} \subset S \subset D_\mu$.

Note that the second inclusion shows that replacing D_μ by S in the definition of $\tau^*(\mu)$ would only decrease the value of $\tau^*(\mu)$, while the first inclusion shows that the value actually does not change. The following lemma, characterizing minimal departures, is essential for our proof of Proposition 4:

Lemma 5 *Take any $\mu \in \mathbb{R}^L$, $d \in D_\mu^{\min}$ and assume WLOG that $f_1(\mu) > \max_{j>1} f_j(\mu)$. Then, there exist $B \in \mathcal{B}_1^+$, $j \in \{2, \dots, K\}$ and $B' \in \mathcal{B}_j^-$ such that*

(i) $\max\{(\mu + d)_i : i \in B\} = f_1(\mu + d) = f_j(\mu + d) = \min\{(\mu + d)_i : i \in B'\}$*;*

(ii) $d_i \leqslant 0$ *if* $i \in B \backslash B'$*;* $d_i \geqslant 0$ *if* $i \in B' \backslash B$*;*

(iii) $\forall i \in B \cup B'$*, either* $(\mu + d)_i = f_1(\mu + d) = f_j(\mu + d)$ *or* $d_i = 0$*.*

Proposition 4 implies the following:

Corollary 6 *Let μ be a valuation and assume WLOG that $f_1(\mu) > \max_{j>1} f_j(\mu)$. Let $\mathcal{B}_j = \{(B, B') : B \in \mathcal{B}_1^+, B' \in \mathcal{B}_j^-\}$. Then,*

$$\tau^*(\mu) = \min_{n \in [0,\infty)^L} \Big\{ \sum_i n(i) : \min_{1<j\leqslant K, \theta \in [f_j(\mu), f_1(\mu)], (B,B') \in \mathcal{B}_j} \sum_{i \in B \backslash B'} n(i)(\mu_i - \theta)_+^2 + \sum_{i \in B' \backslash B} n(i)(\mu_i - \theta)_-^2 + \sum_{i \in B \cap B'} n(i)(\theta - \mu_i)^2 \geqslant 2 \log(\tfrac{1}{4\delta}) \Big\}\,.$$

Hence, for any algorithm A admissible over the instance set $\mathcal{S}_f^{norm}$ at the risk level δ, $\mathbb{E}_{\mu,A}[T]$ is at least as large than the right-hand side of the above display.

5. Upper bound

In this section we propose an algorithm generalizing the LUCB algorithm of Kalyanakrishnan et al. (2012) and prove a theoretical guarantee for the proposed algorithm's sample complexity under some (mild) assumptions on the structure of the reward mapping f. Our result is inspired and extends the results of Garivier et al. (2016a) (who also started from the LUCB algorithm) to the general setting proposed in this paper. In Section 6 we give a version of the algorithm presented here that is specialized to minimax games and refine the upper bound of this section, highlighting the advantages of the extra structure of minimax games.

In this section we shall assume that the distributions $(P_i)_{i \in [L]}$ are subgaussian with a common parameter, which we take to be one for simplicity:

Assumption 1 (1-Subgaussian observations) *For any $i \in [L]$, $X \sim P_i(\cdot)$,*

$$\sup_{\lambda \in \mathbb{R}} \mathbb{E}\left[\exp(\lambda(X - \mathbb{E}X) - \lambda^2/2)\right] \leqslant 1\,.$$

We will need a result for anytime confidence intervals for martingales with subgaussian increments. For stating this result, let $(\mathcal{F}_t)_{t \in \mathbb{N}}$ be a filtration over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ holding our random variables and introduce $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_{t-1}]$. This result appears as (essentially) Theorem 8 in the paper by Kaufmann et al. (2016) who also cite precursors:

Lemma 7 (Anytime subgaussian concentration) *Let $(X_t)_{t \in \mathbb{N}}$ be an $(\mathcal{F}_t)_{t \in \mathbb{N}}$-adapted 1-subgaussian, martingale difference sequence (i.e., for any $t \in \mathbb{N}$, X_t is $\mathcal{F}_t$-measurable,*

$\mathbb{E}_t X_t = 0$, *and* $\sup_\lambda \mathbb{E}_t\left[\exp(\lambda(X_t) - \lambda^2/2)\right] \leqslant 1$*). For* $t \in \mathbb{N}$*, let* $\overline{X}_t = (1/t)\sum_{s=1}^t X_s$*, while for* $t \in \mathbb{N} \cup \{0\}$ *and* $\delta \in [0,1]$ *we let*

$$\beta(t,\delta) = \log(1/\delta) + 3\log\log(1/\delta) + (3/2)(\log(\log(et)))\,.$$

Then, for any $\delta \in [0, 0.1]$*,*[1]

$$\mathbb{P}\left(\sup_{t\in\mathbb{N}} \frac{\overline{X}_t}{\sqrt{2\beta(t,\delta)/t}} > 1\right) \leqslant \delta\,.$$

For a fixed $i \in [L]$, let $N_t(i) = \sum_{s=1}^t \mathbb{I}\{I_s = i\}$ denote the number of observations taken from $P_i(\cdot)$ up to time t. Define the confidence interval $[L_t^\delta(i), U_t^\delta(i)]$ for μ_i as follows: We let

$$\widehat{\mu}_t(i) = \frac{1}{N_t(i)} \sum_{s=1}^t \mathbb{I}\{I_s = i\} Y_s\,,$$

the empirical mean of observations from $P_i(\cdot)$ to be the center of the interval (when $N_t(i) = 0$, we define $\widehat{\mu}_t(i) = 0$) and

$$L_t^\delta(i) = \max\left\{L_{t-1}^\delta(i),\, \widehat{\mu}_t(i) - \sqrt{\frac{2\beta(N_t(i), \delta/(2L))}{N_t(i)}}\right\};$$

$$U_t^\delta(i) = \min\left\{U_{t-1}^\delta(i),\, \widehat{\mu}_t(i) + \sqrt{\frac{2\beta(N_t(i), \delta/(2L))}{N_t(i)}}\right\},$$

where $\beta(t,\delta)$ is as in Lemma 7 (note that when $N_t(i) = 0$, the confidence interval is $(-\infty, +\infty)$). Let T be the index of the round when the algorithm soon to be proposed stops (or $T = \infty$ if it does not stop). Let $\xi = \cap_{t\in[T], i\in[L]}\{\mu_i \in [L_t^\delta(i), U_t^\delta(i)]\}$ be the "good" event when the proposed respective intervals before the algorithm stops *all* contain μ_i for all $i \in [L]$. One can easily verify that, regardless the choice of the algorithm (i.e., the stopping time T),

$$\mathbb{P}(\xi) \geqslant 1 - \delta\,, \tag{3}$$

$$\forall t \in \mathbb{N},\; L_t^\delta(i) \leqslant \widehat{\mu}_t(i) \leqslant U_t^\delta(i). \tag{4}$$

For $S \subset \mathbb{R}^L$ define $f(S) = \{f(s) : s \in S\}$. With this definition, let $S_t = \times_{i=1}^L [L_t^\delta(i), U_t^\delta(i)]$ then for any $j \in [K]$, $f_j(\mu) \in f_j(S_t)$ holds for any $t \geqslant 1$ on ξ. Thus, $f_j(S_t)$ is a valid, $(1-\delta)$-level confidence set for $f_j(\mu)$. For general f, these sets may have a complicated structure. Hence, we will adapt the following simplifying assumption:

Assumption 2 (Regular reward maps) *The following hold:*

(i) The mapping function f *is monotonically increasing with respect to the partial order of vectors: for any* $u, v \in \mathbb{R}^L$*,* $u \leqslant v$ *implies* $f(u) \leqslant f(v)$*;*

1. Note that $\beta(t,\delta)$ is also defined for $t = 0$. The value used is arbitrary: It plays no role in the current result. The reason we define β for $t = 0$ is because it simplifies some subsequent definitions.

(ii) For any $u, v \in \mathbb{R}^L$, $u \leqslant v$, $j \in [K]$, *the set* $D(j, u, v) = \{i \in [L] : [f_j(u), f_j(v)] \subset [u_i, v_i]\}$ *is non-empty.*

We will also let $D_t(j) = D(j, L_t^\delta, U_t^\delta)$. Note that the assumption is met when f is the reward map underlying minimax games (see the next section). The second assumption could be replaced by the following weaker assumption without essentially changing our result: with some $a > 0$, $b \in \mathbb{R}$, for any j, $u \leqslant v$, $[f_j(u), f_j(v)] \subset [au_i + b, av_i + b]$ for some $i \in [L]$. The point of this assumption is that by guaranteeing that all intervals on the micro-observables shrink, the interval on the arm-rewards will also shrink at the same rate. We expect that other ways of weakening this assumption are also possible, perhaps at the price of slightly changing the algorithm (e.g., by allowing it to use even more micro-observations per round).

At time t, let

$$B_t = \operatorname*{argmax}_{j \in [K]} f_j(L_t^\delta)\,,$$
$$C_t = \operatorname*{argmax}_{j \in [K], j \neq B_t} f_j(U_t^\delta)\,. \qquad (5)$$

Algorithm 1 LUCB-micro

> **for** $t = 1, 2, \ldots$ **do**
> Choose B_t, C_t as in Equation (5)
> Choose *any* (I_t, J_t) from $D_t(B_t) \times D_t(C_t)$
> Observe $Y_{t,1} \sim P_{I_t}(\cdot)$, $Y_{t,2} \sim P_{J_t}(\cdot)$
> Update $[L_t^\delta(I_t), U_t^\delta(I_t)]$, $[L_t^\delta(J_t), U_t^\delta(J_t)]$
> **if** Stop() **then**
> $J \leftarrow B_t$, $T \leftarrow t$
> **return** (T, J)

(B stands for candidate "best" arm, C stands for best "contender" arm). Based on the above assumption, we can now propose our algorithm, LUCB-micro (cf. Algorithm 1). Following the idea of LUCB, LUCB-micro chooses B_t and C_t in an effort to separate the highest lower bound from the best competing upper bound.[2] To decrease the width of the confidence intervals, both for B_t and C_t, a micro-observable is chosen with the help of Assumption 2(ii). This can be seen as a generalization of the choice made in Maximin-LUCB by Garivier et al. (2016a). Here, we found that the specific way Maximin-LUCB's choice is made considerably obscured the idea behind this choice, which one can perhaps attribute to that the two-move setting makes it possible to write the choice in a more-or-less direct fashion.

It remains to specify the 'Stop()' function used by our algorithm. For this, we propose the standard choice (as in LUCB):

$$\text{Stop()}: \quad f_{B_t}(L_t^\delta) \geqslant f_{C_t}(U_t^\delta). \qquad (6)$$

All statements in this section assume that the assumptions stated so far in this section hold.

The following proposition is immediate from the definition of the algorithm.

Proposition 8 (Correctness) *On the event ξ, LUCB-micro returns J correctly: $J = j^*(\mu)$.*

2. Using a lower bound departs from the choice of LUCB, which would use $f_j(\hat{\mu}_t)$ to define B_t. The reason of this departure is that we found it easier to work with a lower bound. We expect the two versions (original, our choice) to behave similarly.

Let T denote the round index when LUCB-micro stops[3] and define $c = \frac{f_1(\mu)+f_2(\mu)}{2}$ and $\Delta = f_1(\mu) - f_2(\mu)$, where we assumed that $f_1(\mu) > f_2(\mu) \geqslant \max_{j\geqslant 2} f_j(\mu)$. The main result of this section is a high-probability bound on T, which we present next. The following lemma is the key to the proof:

Lemma 9 *Let $t < T$. Then, on ξ, there exists $J \in \{B_t, C_t\}$ such that $c \in [f_J(L_t^\delta), f_J(U_t^\delta)]$ and $f_J(U_t^\delta) - f_J(L_t^\delta) \geqslant \Delta/2$.*

The proof follows standard steps (e.g., Garivier et al. 2016a). In particular, the above lemma implies that if $T > t$ then for $J \in \{B_t, C_t\}$, $c \in [f_J(L_t^\delta), f_J(U_t^\delta)]$ and $f_J(U_t^\delta) - f_J(L_t^\delta) \geqslant \Delta/2$. This in turn implies that for $i \in \{I_t, J_t\}$, $N_t(i)$ cannot be too large.

Theorem 10 (LUCB-micro upper bound) *Let*

$$H(\mu) = \sum_{i\in[L]} \left\{ \frac{1}{(c-\mu_i)^2} \bigwedge \frac{1}{(\Delta/2)^2} \right\}, \quad t^*(\mu) = \min\{t \in \mathbb{N} \,:\, 1 + 8H(\mu)\beta(t, \delta/(2L)) \leqslant t\}\,.$$

Then, for $\delta \leqslant 0.1$, on the event ξ, the stopping time T of LUCB-micro satisfies $T \leqslant t^*(\mu)$.

Note that $\beta(t,\delta) \propto \log\log t$ and thus $t^*(\mu)$ is well-defined. Furthermore, letting $c_\delta = \log(2L/\delta) + 3\log\log(2L/\delta)$, for δ sufficiently small and $H(\mu)$ sufficiently large, elementary calculations give

$$t^*(\mu) \leqslant 16H(\mu)c_\delta + 16H(\mu)\log\log(8H(\mu)c_\delta)\,.$$

Remark 11 *The constant $H(\mu)$ acts as a hardness measure of the problem. Theorem 10 can be applied to the best arm identification problem in the multi-armed bandits setting, as it is a special case of our problem setup. Compared to state-of-the-art results available for this setting, our bound is looser in several ways: We lose on the constant factor multiplying $H(\mu)$ (Kalyanakrishnan et al., 2012; Jamieson and Nowak, 2014; Jamieson et al., 2014; Kaufmann et al., 2016), we also lose an additive term of $H(\mu)\log\log(L)$ (Chen and Li, 2015). We also lose $\log(L)$ terms on the suboptimal arms (Simchowitz et al., 2017). Comparing with the only result available in the two-move minimax tree setting, due to Garivier et al. (2016a), our bound is looser than their Theorem 1. This motivates the refinement of this result to the minimax setting, which is done in the next section, and where we recover the mentioned result of Garivier et al. (2016a). On the positive side, our result is more generally applicable than any of the mentioned results. It remains an interesting sequence of challenges to prove an upper bound for this or some other algorithm which would match the mentioned state-of-the-art results, when the general setting is specialized.*

6. Best move identification in minimax games

3. The number of observations, or number of rounds as per Fig. 1, taken by LUCB-micro until it stops is $2T$.

In this section we will show upper bounds on the number of observations LUCB-micro takes in the case of minimax game problems. We still assume that the micro-observations are subgaussian (Assumption 1) and the optimal arm is unique. To apply our result, this leaves us with showing that the payoff function in the minimax game satisfies the regularity assumption (Assumption 2).

Algorithm 2 MinMax.

> **Inputs:** $h \in H$, $u, v \in \mathbb{R}^L$.
> **if** $h \in H_{\max}$ **then**
>   **return** h
> **else if** $p(h) = -1$ **then**
> $h \leftarrow \text{join}(h, m(h, u))$
> **else if** $p(h) = 1$ **then**
> $h \leftarrow \text{join}(h, m(h, v))$
> **return** MinMax(h, u, v)

Fix a minimax game structure $G = (M, H, p, \tau)$. We first show that Property (i) of Assumption 2 holds. This follows easily from the following lemma, which can be proven by induction based on "distance from the terminating states".

Lemma 12 *For any $h \in H$ and $u, v \in [0,1]^L$ such that $u \leqslant v$, $V(h,u) \leqslant V(h,v)$.*

From this result we immediately get the following corollary:

Corollary 13 *For $u, v \in [0,1]^L$ such that $u \leqslant v$, $f(u) \leqslant f(v)$, hence Assumption 2 (i) holds.*

For $j \in [K]$, $u, v \in \mathbb{R}^L$, $u \leqslant v$, per Property (ii) of Assumption 2, we need to show that the sets $D(j,u,v)$ are nonempty. For a history $h = (m_1, m_2, \ldots, m_\ell) \in H$ and $1 \leqslant k \leqslant \ell$, we denote its length-$k$ prefix $(m_1, \ldots, m_k)$ by h_k. We give an algorithmic demonstration, which also shows how to efficiently pick an element of these sets. The resulting algorithm is called MinMax (cf. Algorithm 2). We define MinMax in a recursive fashion: For each nonmaximal history the algorithm extends the history by adding the move which is optimal for u for minimizing moves, while it extends it by adding the optimal move for v for maximizing moves, and then it calls itself with the new history. The algorithm returns when its input is a maximal history. To show that $\tau(\text{MinMax}(h,u,v)) \in D(j,u,v)$ we have the following result:

Lemma 14 *Fix $u, v \in \mathbb{R}^L$, $u \leqslant v$, and $j \in [K]$. Let $h = \text{MinMax}((j))$ and in particular let $h = (m_1 = j, m_2, \ldots, m_\ell)$. Then, for all $1 \leqslant k < \ell$,*

$$[V(h_k, u), V(h_k, v)] \subset [V(h_{k+1}, u), V(h_{k+1}, v)],$$

where h_k is the length-k prefix of h.

We immediately get that $i = \tau(h)$ is an element of $D(j,u,v)$:

Corollary 15 *For j, u, v, h as in the previous result, setting $i = \tau(h)$, $[f_j(v), f_j(v)] \subset [u_i, v_i)]$, hence $i \in D(j,u,v) \neq \varnothing$.*

With this, we have shown that all the assumptions needed by Theorem 10 are satisfied, and in particular, we can use $I_t = \text{MinMax}(B_t, L_t^\delta, U_t^\delta)$ and $J_t = \text{MinMax}(C_t, L_t^\delta, U_t^\delta)$. We call the resulting algorithm LUCBMinMax. Then, Theorem 10 gives:

Corollary 16 *Let ξ, c, Δ, $H(\mu), t^*$ be as in Section 5. If T is the stopping time of* LUCB-MinMax *running on a minimax game search problem then $T \leqslant t^*$.*

When applied to a minimax game, as defined in Section 2, the upper bound of Corollary 16 is loose and can be further improved as shown in the result below. To state this result we need some further notation. Given a set of reals S, define the "span" of S as $\text{span}(S) = \max_{u,v \in S} u - v$. For a path $h \in H$ that connects some move in $[K]$ and some move in $[L]$: $h = (m_1, \ldots, m_\ell)$ with some $\ell \geqslant 0$, $m_1 \in [K]$ and $m_\ell \in [L]$. Finally, for $i \in [L]$ such that there is a unique path $h \in H$ satisfying $\tau(h) = i$, define $\mathbb{V}(i, \mu) = \{V(h_k, \mu) : h = (m_1, \ldots, m_\ell = i), m_1 \in [K], 1 \leqslant k < \ell\}$. Let $\mathbb{V}(i, \mu)$ be an empty set if there is multiple $h \in H$ such that $\tau(h) = i$.

Theorem 17 (LUCBMinMax on MinMax Trees) *Let*

$$H(\mu) = \sum_{i \in [L]} \min\{\frac{1}{\text{span}(\mathbb{V}(i, \mu) \cup \{c, \mu_i\})^2}, \frac{4}{\Delta^2}\},$$
$$t^*(\mu) = \min\{t \in \mathbb{N} : 1 + 8H(\mu)\beta(t, \delta/(2L)) \leqslant t\}.$$

Then, on ξ, the stopping time T of LUCBMinMax *satisfies $T \leqslant t^*(\mu)$.*

Remark 18 *Note that this result recovers Theorem 1 of Garivier et al. (2016a). To see this note that for every leaf (i, j) (as numbered in their paper), $\mu_{i,1} \in \mathbb{V}((i, j), \mu)$. Also note that $\mu_{i,1} \leqslant \mu_{i,j}$, thus $|c - \mu_{i,j}| \leqslant \max\{|c - \mu_{i,1}|, |\mu_{i,j} - \mu_{i,1}|\}$. Therefore,* $\text{span}(\{\mu_{i,1}, \mu_{i,j}, c\}) = \max\{|\mu_{i,1} - c|, |\mu_{i,j} - \mu_{i,1}|\}$.

7. Discussion and Conclusions

The gap between the lower bound and the upper bound There is a substantial gap between the lower and the upper bound. Besides the gaps that already exist in the multi-armed bandit setting and which have been mentioned before, there exists a substantial gap: In particular, it is not hard to show that in regular minimax game trees with a fixed branching factor of κ and depth d, the upper bound scales with $O(\kappa^d)$ while the lower bound scales with $O(\kappa^{d/2})$. One potential is to improve the lower bound so as to consider adversarial perturbations of the values assigned to the leaf nodes: That is, after the algorithm is fixed, an adversary can perturb the values of μ to maximize the lower bound. Simchowitz et al. (2017) introduce an interesting technique for proving lower bounds of this form and they demonstrate nontrivial improvements in the multi-armed bandit setting.

Does the algorithm need to explore all leaves? The hardness measure $H(\mu)$ is rooted in a uniform bound that suggests that all the leaves must potentially be pulled, which may not hold for some particular structure. In particular, the algorithm may be able to benefit from the specific structure of f, saving explorations on some leaves. We present one example when f is a minimax game tree, as in Figure 2. Assume that $\mu_{1,i} = \mu^* >> \mu^{**} = \mu_{j,i}$ for $1 \leqslant i \leqslant K$ and $2 \leqslant j \leqslant K$. A reasonable algorithm would sample each arm once, then discover that the others arms are much less than the sampled leaf under arm 1. Then the algorithm will continue to explore the other leaves of arm 1, and decide arm 1 to be the

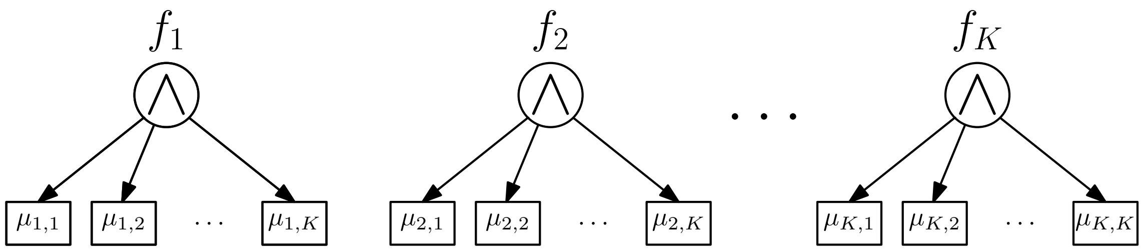

Figure 2: A 2-layer minimax tree. The symbol '$\wedge$' in the node denotes that it is a "min" node.

best arm. This behavior is also in agreement with our lower bound, where the resulting constraints are:

$$N_{1,i}(\mu^* - \mu^{**})^2 \geqslant 2\log(1/4\delta); \quad \sum_{i=1}^{K} N_{j,i}(\mu^* - \mu^{**})^2 \geqslant 2\log(1/4\delta) \quad \forall j \neq 1,$$

which implies $N_{1,i} \geqslant 1$ and $\sum_{i=1}^{K} N_{j,i} \geqslant 1$ for $j \neq 1$ if $\mu^* - \mu^{**}$ is large enough. As we can see from this example, $(K-1)^2$ (out of K^2) leaves need no exploration at all. On the other hand, although we don't have a tight upper bound, our algorithm in practice manages to explore the remaining $K-1$ leaves under arm 1 for the next $K-1$ rounds, and then make the right decision.

In general, we would expect that a problem with a feed forward neural network structure is easier than that of a tree structure, as the share of the leaves provides more information and thus save the exploration. This is illustrated on Fig. 3, where an optimal arm can be identified solely based on the network structure, thus the algorithm requires **0** samples for all possible μ. Note that our lower bound does not fail, as we have $D_\mu = \varnothing$ here.

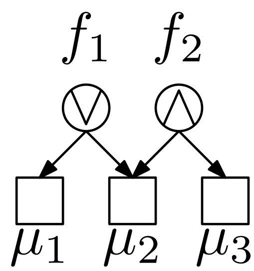

Figure 3: An example that needs no exploration. The symbol '$\wedge$' (resp., '$\vee$') in the node denotes that it is a "min" (resp., "max") node.

Acknowledgments

References

Louis Victor Allis. *Searching for solutions in games and artificial intelligence*. Ponsen & Looijen, 1994.

Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *23rd Annual Conference on Learning Theory*, COLT 2010, pages 13–p, 2010.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 32(1):48–77, 2002.

Lijie Chen and Jian Li. On the optimal sample complexity for best arm identification. *arXiv preprint arXiv:1511.03774*, 2015.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(Jun):1079–1105, 2006.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *26th Annual Conference on Neural Information Processing Systems*, NIPS 2016, pages 3212–3220. Curran Associates Inc., 2012.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *COLT*, pages 998–1027, 2016.

Aurélien Garivier, Emilie Kaufmann, and Wouter M. Koolen. Maximin action identification: A new bandit framework for games. In *29th Annual Conference on Learning Theory*, COLT 2016, pages 1028–1050, 2016a.

Aurélien Garivier, Emilie Kaufmann, and Tor Lattimore. On explore-then-commit strategies. *arXiv preprint arXiv:1605.08988*, 2016b.

Sylvain Gelly, Levente Kocsis, Marc Schoenauer, Michèle Sebag, David Silver, Csaba Szepesvári, and Olivier Teytaud. The grand challenge of computer Go: Monte Carlo tree search and extensions. *Communications of ACM*, 55(3):106–113, 2012.

Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *Information Sciences and Systems (CISS), 2014 48th Annual Conference on*, pages 1–6. IEEE, 2014.

Kevin G. Jamieson, Matthew Malloy, Robert D. Nowak, and Sébastien Bubeck. lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *27th Annual Conference on Learning Theory*, volume 35 of *COLT 2014*, pages 423–439, 2014.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning*, ICML 2012, pages 655–662, 2012.

Zohar Shay Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of The 30th International Conference on Machine Learning*, volume 28 of *ICML 2013*, pages 1238–1246, 2013.

Emilie Kaufmann and Wouter Koolen. Monte-carlo tree search by best arm identification. *arXiv preprint arXiv:1706.02986*, 2017.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17 (1):1–42, 2016.

Akihiro Kishimoto, Mark H.M. Winands, Martin Müller, and Jahn-Takeshi Saito. Game-tree search using proof numbers: The first twenty years. *ICGA Journal*, 35(3):131–156, 2012.

David Allen McAllester. Conspiracy numbers for min-max search. *Artificial Intelligence*, 35 (3):287–310, 1988.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Max Simchowitz, Kevin G. Jamieson, and Benjamin Recht. The simulator: Understanding adaptive sampling in the moderate-confidence regime. abs/1702.05186, 2017.

Kazuki Teraoka, Kohei Hatano, and Eiji Takimoto. Efficient sampling method for monte carlo tree search problem. *IEICE TRANSACTIONS on Information and Systems*, 97(3): 392–398, 2014.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

Appendix A. The Uniqueness Assumption

Recall that throughout the paper, by definition, we have the following assumption:

Assumption 3 *The instance is such that $j^*(\mu) = \operatorname{argmax}_j f_j(\mu)$ is unique.*

We state this assumption explicitly here, so that we can refer to it easily throughout the appendix.

Appendix B. Proofs for Section 3

Here we prove Theorem 1, which is restated for the convenience of the reader:

Theorem 1 (Lower bound) *Fix a risk parameter $\delta \in (0,1)$. Assume that A is admissible over the instance set $\mathcal{S}_f^{norm}$ at the risk level δ. Define*

$$\tau^*(\mu) = \min\left\{ \sum_{i=1}^{L} n(i) \,:\, \inf_{\Delta \in D_\mu} \sum_{i=1}^{L} n(i)\Delta_i^2 \geqslant 2\log(1/(4\delta)),\, n(1),\ldots,n(L) \geqslant 0 \right\}. \quad (2)$$

Then, $\mathbb{E}_{\mu,A}[T] \geqslant \tau^*(\mu)$.

We start with the two information theoretic results mentioned in the main body of the text. To state these results, let $\mathrm{D}(P,Q)$ denote the Kullback–Leibler (KL) divergence of two distributions P and Q. Recall that this is $\int \log(\frac{dP}{dQ})dQ$ when P is absolutely continuous with respect to Q and is infinite otherwise. For the next result, let $N(i) = \sum_{t=1}^{T} \mathbb{I}\{I_t = i\}$ denote the number of times an observation on the micro-observable with index $i \in [L]$ before time T.

Lemma 19 (Divergence decomposition) *For any $\mu, \mu' \in \mathbb{R}^L$ it holds that*

$$\mathrm{D}(\mathbb{P}_{\mu,A}, \mathbb{P}_{\mu',A}) = \frac{1}{2} \sum_{i=1}^{L} \mathbb{E}_{\mu,A}[N(i)]\, (\mu_i - \mu'_i)^2 \,. \tag{7}$$

Note that $\frac{1}{2}(\mu_i - \mu'_i)^2$ on the right-hand side is the KL divergence between the normal distributions with means μ_i and μ'_i and both having a unit variance. The result, naturally, holds for other distributions, as well. This is the result that relies strongly on that we forced the same observations and same observation-choices for $t > T$. In particular, this is what makes the left-hand side of (7) finite! The proof is standard and hence is omitted.

Lemma 20 (High probability Pinsker, e.g., Lemma 2.6 from (Tsybakov, 2008)) *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$ and let $E \in \mathcal{F}$ be an arbitrary event. Then,*

$$P(E) + Q(E^c) \geqslant \frac{1}{2} \exp(-\mathrm{D}(P,Q)) \,.$$

Proof [of Theorem 1] WLOG, we may assume that D°_μ is non-empty. Pick any $\Delta \in D^\circ_\mu$ and let $\mu' = \mu + \Delta$. Let $E = \{J \neq 1\}$. Since A is admissible, $\mathbb{P}_{\mu,A}(E) \leqslant \delta$. Further, since $\Delta \in D^\circ_\mu$, 1 is not an optimal arm in μ'. Hence, again by the admissibility of A, $\mathbb{P}_{\mu',A}(E^c) \leqslant \delta$. Therefore, by Lemma 20,

$$2\delta \geqslant \mathbb{P}_{\mu,A}(E) + \mathbb{P}_{\mu',A}(E^c) \geqslant \frac{1}{2} \exp(-\mathrm{D}(\mathbb{P}_{\mu,A}, \mathbb{P}_{\mu',A})) \,.$$

Now, plugging in (7) of Lemma 19 and reordering we get

$$\log(1/(4\delta)) \leqslant \mathrm{D}(\mathbb{P}_{\mu,A}, \mathbb{P}_{\mu',A}) = \frac{1}{2} \sum_{i=1}^{L} \mathbb{E}_{\mu,A}[N(i)]\, (\mu_i - \mu'_i)^2 \,.$$

The result follows by continuity, after noting that $T = \sum_{i=1}^{L} N(i)$, that $\Delta \in D^\circ_\mu$ was arbitrary. ■

Appendix C. Proofs for Section 4

We start with the following lemma:

Lemma 21 *Pick any $\mu \in \mathbb{R}^L$, $j \in [K]$. Then,*

$$\forall B \in \mathcal{B}^+_j, f_j(\mu) \leqslant \max\{\mu_i \,:\, i \in B\}; \quad \forall B \in \mathcal{B}^-_j, f_j(\mu) \geqslant \min\{\mu_i \,:\, i \in B\}.$$

Proof Fix any $\mu \in \mathbb{R}^L$, $j \in [K]$, $B \in \mathcal{B}_j^+$. Let $u = \max\{\mu_i : i \in B\}$. We want to show that $f_j(\mu) \leqslant u$. Define $\mu' \geqslant \mu$ such that $\mu'_i = \mu_i$ if $i \notin B$, and $\mu'_i = u$ otherwise. As noted earlier (cf. Corollary 13), f_j is monotonically increasing. Hence, $f_j(\mu) \leqslant f_j(\mu') \leqslant u$, where the last inequality follows because $B \in \mathcal{B}_j^+$. The proof concerning $\mathcal{B}_j^-$ is analogous and is left to the reader. ■

Lemma 5 *Take any $\mu \in \mathbb{R}^L$, $d \in D_\mu^{\min}$ and assume WLOG that $f_1(\mu) > \max_{j>1} f_j(\mu)$. Then, there exist $B \in \mathcal{B}_1^+$, $j \in \{2, \ldots, K\}$ and $B' \in \mathcal{B}_j^-$ such that*

(i) $\max\{(\mu + d)_i : i \in B\} = f_1(\mu + d) = f_j(\mu + d) = \min\{(\mu + d)_i : i \in B'\}$*;*

(ii) $d_i \leqslant 0$ *if* $i \in B \backslash B'$*;* $d_i \geqslant 0$ *if* $i \in B' \backslash B$*;*

(iii) $\forall i \in B \cup B'$*, either* $(\mu + d)_i = f_1(\mu + d) = f_j(\mu + d)$ *or* $d_i = 0$*.*

Proof Note that since $d \in D_\mu$, $f_1(\mu + d) \leqslant f_j(\mu + d)$ for some $j \neq 1$. Fix such an index j.

To construct B and B', we will pick $\widetilde{H}_1 \in \mathcal{H}_1^+$, $\widetilde{H}_j \in \mathcal{H}_j^-$ and set $B = \tau(\widetilde{H}_1 \cap H_{\max})$ and $B' = \tau(\widetilde{H}_j \cap H_{\max})$. By the construction of $\mathcal{H}_1^+$, to pick $\widetilde{H}_1$ it suffices to specify the unique successor h' in $\widetilde{H}_1$ of any history $h \in \widetilde{H}_1$ with $p(h) = -1$. For this, we let $h' \in H$ be the successor for which $V(h', \mu + d) = V(h, \mu + d)$. Similarly, by the construction of $\mathcal{H}_j^-$, to pick $\widetilde{H}_j$ it suffices to specify the unique successor h' in $\widetilde{H}_j$ of any history $h \in \widetilde{H}_j$ with $p(h) = +1$. Again, we let $h' \in H$ be the successor for which $V(h', \mu + d) = V(h, \mu + d)$. Note that by Proposition 3, $B \in \mathcal{B}_1^+$ and $B' \in \mathcal{B}_j^-$.

Let us now turn to the proof of (i). We start by showing that

$$f_1(\mu + d) = \max\{(\mu + d)_i : i \in B\} \,. \tag{8}$$

To show this, we first prove that

$$V(h, \mu + d) \leqslant f_1(\mu + d) \quad \forall h \in \widetilde{H}_1 \,. \tag{9}$$

The proof uses induction based on the length of histories in $\widetilde{H}_1$.

There is only one history of length 1 (base case): $h = (1)$. By the definition of f_1, $V(h, \mu + d) = f_1(\mu + d)$. Now, assume that the statement holds for all histories up to length $c \geqslant 1$. Take any $h \in \widetilde{H}_1$ of length $c + 1$. Let $h' \in \widetilde{H}_1$ be the unique immediate predecessor of h: $h \in \mathrm{H}_{\mathrm{succ}}(h')$. This is well-defined thanks to the definition of H and the construction of $\widetilde{H}_1$. If $p(h') = -1$ then, by the definition of $\widetilde{H}_1$, $V(h, \mu + d) = V(h', \mu + d)$. By the induction hypothesis, $V(h', \mu + d) \leqslant f_1(\mu + d)$, implying $V(h, \mu + d) \leqslant f_1(\mu + d)$. On the other hand, if $p(h') = +1$ then $f_1(\mu + d) \geqslant V(h', \mu + d) = \max\{V(\tilde{h}, \mu + d), \tilde{h} \in \mathrm{H}_{\mathrm{succ}}(h')\} \geqslant V(h, \mu + d)$, finishing the induction. Hence, we have proven (9).

Now, we claim that there exists $h^* \in \widetilde{H}_1 \cap H_{\max}$ such that $V(h^*, \mu + d) = f_1(\mu + d)$. This, together with (9) implies (8).

We construct h^* in a sequential process. For this, we will choose a sequence of moves $m_1, \ldots, m_k$ such that $(m_1, \ldots, m_i) \in H_{\max} \cap \widetilde{H}_1$ and $V((m_1, \ldots, m_i), \mu + d) = f_1(\mu + d)$ for any $1 \leqslant i \leqslant k$. In a nutshell, this sequence is an "optimal sequence of moves" that starts with move 1, which is also known as a principal variation for the game under move 1. In details, the construction is as follows: To start, we choose $m_1 = 1$. Then $V((m_1), \mu+d) = f_1(\mu+d)$, by the definition of f_1. Assume that for some $i \geqslant 1$, we already chose $(m_1, \ldots, m_i)$ so that $V((m_1, \ldots, m_i), \mu + d) = f_1(\mu + d)$ holds. If $h \doteq (m_1, \ldots, m_i) \in H_{\max}$, we let $k = i$ and we are done. Otherwise, let $m_{i+1} = m(h, \mu+d)$ (this is the "optimal move" at h under valuation $\mu + d$). Thus, $V(\text{join}(h, m_{i+1}), \mu + d) = V(h, \mu + d) = f_1(\mu)$. Further, by the construction of $\widetilde{H}_1$, $\text{join}(h, m_{i+1}) \in \widetilde{H}_1$. Since all histories in H are bounded in length, the process ends after some k moves for some finite k, at which point we are done proving our statement.

To recap, so far we have proved (8). An entirely analogous proof (left to the reader) shows that also $f_j(\mu + d) = \min\{(\mu + d)_i \,:\, i \in B'\}$.

We now prove that $f_1(\mu+d) = f_j(\mu+d)$, finishing the proof of (i). Assume to the contrary that $f_1(\mu+d) < f_j(\mu+d)$. Consider the map $g : \alpha \mapsto f_1(\mu+\alpha d) - f_j(\mu+\alpha d)$ on the interval $\alpha \in [0, 1]$. Note that g is continuous, $g(0) > 0 > g(1)$. Hence, by the intermediate value theorem, there exists $\alpha \in (0, 1)$ such that $g(\alpha) = 0$. Note that $f_1(\mu + \alpha d) = f_j(\mu + \alpha d)$. Hence, $\alpha d \in D_\mu$. Since $\alpha|d| < |d|$, $d \in D_\mu^{\min}$ cannot hold, a contradiction. Hence, $f_1(\mu+d) = f_j(\mu + d)$.

Let us now turn to the proof of (ii). We prove that $d_i \leqslant 0$ holds for all $i \in B\backslash B'$. (The statement concerning elements of $B'\backslash B$ follows similarly, the details are left to the reader.) For the proof, assume to the contrary of the desired statement that there exists some $i \in B\backslash B'$ such that $d_i > 0$. Let $d' \in \mathbb{R}^L$ be such that $d'_k = d_k$ for $j \neq i$, and $d'_i = 0$. Thus, $d' < d$. By Corollary 13, $f_1(\mu + d') \leqslant f_1(\mu + d) \leqslant f_j(\mu + d) = \min\{(\mu + d)_k \,:\, k \in B'\} = \min\{(\mu + d')_k \,:\, k \in B'\} \leqslant f_j(\mu + d')$, where the last equality is due to $i \notin B'$ (hence, $(\mu + d)|_{B'} = (\mu + d')|_{B'}$) while the last inequality follows from Lemma 21. This implies that $d' \in D_\mu$. This together with $|d'| < |d|$ contradicts $d \in D_\mu^{\min}$. Thus, (ii) holds.

It remains to prove (iii). For this pick $i \in B$. Since $f_1(\mu + d) = f_j(\mu + d)$ has already been established, it suffices to show that either $(\mu + d)_i = f_1(\mu + d)$ or $d_i = 0$. (The case when $i \in B'$ is symmetric and is left to the reader.) If $i \in B \cap B'$ then by (i), $(\mu + d)_i \leqslant \max_{k\in B}(\mu + d)_k = f_1(\mu + d) = f_j(\mu + d) = \min_{k\in B'}(\mu + d)_k \leqslant (\mu + d)_i$, showing that $(\mu + d)_i = f_1(\mu + d) = f_j(\mu + d)$. Hence, assume that $i \notin B \cap B'$. If $d_i = 0$ or $(\mu + d)_i = f_1(\mu + d)$ then we are done. Otherwise, by (ii), $d_i < 0$ and by (i), $(\mu + d)_i < \max_{k\in B}(\mu + d)_k = f_1(\mu + d)$. Let $\epsilon = f_1(\mu + d) - (\mu + d)_i$. Note that $\epsilon > 0$. Define $d' \in \mathbb{R}^L$ so that $d'_k = d_k$ if $k \neq i$ and let $d'_i = -(d_i + \epsilon)_-$. That is, d_i is shifted up towards zero by a positive amount so that it never crosses zero. Then, $|d'| < |d|$. Note also that $\mu_i + d'_i = \mu_i + \min(d_i + \epsilon, 0) \leqslant \mu_i + d_i + \epsilon = f_1(\mu + d) = \max_{k\in B}(\mu + d)_k$. Hence, $\max_{k\in B}(\mu + d')_k = \max_{k\in B}(\mu + d)_k = f_1(\mu + d)$ and thus by Lemma 21, $f_1(\mu + d') \leqslant \max_{k\in B}(\mu+d)'_k = f_1(\mu+d)$. By (i), $f_1(\mu+d) = f_j(\mu+d) = \min_{k\in B'}(\mu+d)_k$. By the definition of d' (thanks to $i \notin B'$) and Lemma 21, $\min_{k\in B'}(\mu + d)_k = \min_{k\in B'}(\mu + d')_k \leqslant f_j(\mu + d')$. Putting together the inequalities, we get $f_1(\mu + d') \leqslant f_j(\mu + d')$. Hence, $d' \in D_\mu$. However, this and $|d'| < |d|$ contradict $d \in D_\mu^{\min}$, finishing the proof of (iii). ∎

Lemma 22 *Given any $\mu \in \mathbb{R}^L$ and any $\theta \in \mathbb{R}$, define μ' as follows:*

$$\mu'_i = \begin{cases} \theta, & i \in \mathcal{I}; \\ \mu_i, & \textit{otherwise}, \end{cases}$$

where $\mathcal{I} \subset \{i \,:\, \mu_i \geqslant \theta\}$. Then, $f_j(\mu') \geqslant \min\{\,\theta, f_j(\mu)\,\}$ for any $j \in [K]$.

Proof Fix $j \in [K]$. We prove $V(h, \mu') \geqslant \min\{\theta,\, V(h, \mu)\}$ for $h \in H$ by induction based on how close a history h is to being a maximal history. Note that this suffices to prove the statement thanks to $f_j(\mu') = V((j), \mu') \geqslant \min\{\theta,\, V((j), \mu)\} = \min\{\theta,\, f_j(\mu)\}$.

Define function c so that $c(h) = 0$ if $h \in H_{\max}$, and $c(h) = 1 + \max\{c(h') : h' \in \mathrm{H}_{\mathrm{succ}}(h)\}$ otherwise.
Base case: If $h \in H_{\max}$, then $V(h, \mu') = \mu'_i \in \{\mu_i, \theta\} \geqslant \min\{\theta, \mu_i\} = \min\{\theta, V(h, \mu)\}$ for some $i \in [L]$.
Induction step: Assume that for any $h \in H$ such that $c(h) \leqslant c$, $V(h, \mu') \geqslant \min\{\theta,\, V(h, \mu)\}$. Given h such that $c(h) = c + 1$, if $p(h) = 1$,

$$\begin{aligned} V(h, \mu') &= \max\{V(h', \mu') : h' \in \mathrm{H}_{\mathrm{succ}}(h)\} \\ &\geqslant \max\{\min\{\theta, V(h', \mu)\} : h' \in \mathrm{H}_{\mathrm{succ}}(h)\} \quad \text{(by induction)} \\ &\overset{\text{(a)}}{\geqslant} \min\{\theta,\, V(h'^{*}, \mu)\} \\ &= \min\{\theta,\, V(h, \mu)\}, \end{aligned}$$

where in (a), h'^{*} is the optimal h' such that $V(h'^{*}, \mu) = V(h, \mu)$. If $p(h) = -1$,

$$\begin{aligned} V(h, \mu') &= \min\{V(h', \mu') : h' \in \mathrm{H}_{\mathrm{succ}}(h)\} \\ &\geqslant \min\{\min\{\theta,\, V(h', \mu)\} : h' \in \mathrm{H}_{\mathrm{succ}}(h)\} \\ &\overset{\text{(b)}}{\geqslant} \min\{\theta,\, \min\{V(h', \mu) : h' \in \mathrm{H}_{\mathrm{succ}}(h)\}\} \\ &\geqslant \min\{\theta, V(h, \mu)\}. \end{aligned}$$

Here (b) holds because for any $h' \in \mathrm{H}_{\mathrm{succ}}(h)$, $V(h', \mu) \geqslant \min\{V(h', \mu) : h' \in \mathrm{H}_{\mathrm{succ}}(h)\}$, thus $\min\{\theta\,,\, V(h', \mu)\} \geqslant \min\{\theta,\, \min\{V(h', \mu) : h' \in \mathrm{H}_{\mathrm{succ}}(h)\}\}$. ■

With this, we are ready to prove Proposition 4, which we repeat here for the reader's convenience:

Proposition 4 (Minimal significant departures for minimax games) *Without loss of generality, assume that $f_1(\mu) > \max_{j>1} f_j(\mu)$. Let*

$$\begin{aligned} S = \Big\{ \Delta \in \mathbb{R}^L \,:\, &\exists 1 < j \leqslant K\,, \theta \in [f_j(\mu), f_1(\mu)], B \in \mathcal{B}_1^+, B' \in \mathcal{B}_j^- \;\; s.t. \\ &\Delta_i = -(\mu_i - \theta)_+,\; \forall i \in B \backslash B';\; \Delta_i = (\mu_i - \theta)_-,\; \forall i \in B' \backslash B; \\ &\Delta_i = \theta - \mu_i,\; \forall i \in B' \cap B;\; \Delta_i = 0,\; \forall i \in (B \cup B')^c \Big\}. \end{aligned}$$

Then, $D_\mu^{\min} \subset S \subset D_\mu$.

Proof First we prove $D_\mu^{\min} \subset S$. For this take any $d \in D_\mu^{\min}$. Since $d \in D_\mu$, by Lemma 5, for some $j > 1$, $f_1(\mu + d) = f_j(\mu + d)$. WLOG assume $j = 2$. We will prove that:

$$\exists B \in \mathcal{B}_1^+,\; B' \in \mathcal{B}_2^- \;\text{s.t.}\; \forall i \in (B \cup B')^c,\; d_i = 0. \tag{10}$$

By Lemma 5, there exist $B \in \mathcal{B}_1^+$ and $B' \in \mathcal{B}_2^-$ such that

$$\max\{(\mu + d)_i \,:\, i \in B\} = f_1(\mu + d) = f_2(\mu + d) = \min\{(\mu + d)_i \,:\, i \in B'\};$$

Take these sets and pick some $i \in (B \cup B')^c$. If $d_i = 0$, we are done. Otherwise, let $d'_k = d_k$ for all $k \neq i$ and let $d'_i = 0$. Then, $|d'| < |d|$. By Lemma 21 and Lemma 5 (i),

$$\begin{aligned} f_1(\mu + d') &\leqslant \max\{(\mu + d')_j \,:\, j \in B\} = \max\{(\mu + d)_j \,:\, j \in B\} \\ &= f_1(\mu + d) \\ &\leqslant f_2(\mu + d) \\ &= \min\{(\mu + d)_j,\, j \in B'\} = \min\{(\mu + d')_j \,:\, j \in B'\} \\ &\leqslant f_2(\mu + d'). \end{aligned}$$

Thus $d' \in D_\mu$, which contradicts that $d \in D_\mu^{\min}$, establishing (10). Also we have $f_1(\mu + d) = f_2(\mu + d)$.

Let $\theta = f_1(\mu + d) = f_2(\mu + d)$. For $i \in B \backslash B'$, by Lemma 5 (ii) and (iii), $d_i = -(\mu_i - \theta)_+$. Similarly, $d_i = (\mu_i - \theta)_-$ for $i \in B' \backslash B$. Note that for $i \in B \cap B'$,

$$\begin{aligned} (\mu + d)_i &\leqslant \max\{(\mu + d)_i \,:\, i \in B \cap B'\} \\ &\leqslant \max\{(\mu + d)_i \,:\, i \in B\} \\ &= f_1(\mu + d) = \theta = f_2(\mu + d) \\ &= \min\{(\mu + d)_i \,:\, i \in B'\} \\ &\leqslant \min\{(\mu + d)_i \,:\, i \in B' \cap B\} \\ &\leqslant (\mu + d)_i\,. \end{aligned}$$

Thus, $(\mu + d)_i = \theta$, and therefore $d_i = \theta - \mu_i$. It remains to prove $\theta \in [f_2(\mu), f_1(\mu)]$. We prove this by contradiction. Assume that $\theta < f_2(\mu)$. Define d' as follows:

$$d'_i = \begin{cases} -(\mu_i - f_2(\mu))_+\,, & \text{if } i \in B\,; \\ 0\,, & \text{otherwise.} \end{cases}$$

We will prove the following claims:

(i) $|d'| < |d|$;

(ii) $f_1(\mu + d') \leqslant f_2(\mu)$;

(iii) $f_2(\mu + d') \geqslant f_2(\mu)$.

Altogether these contradict $d \in D_\mu^{\min}$.

To show (i), note that for $i \in B^c$ or $i \in B$ such that $\mu_i \leqslant f_2(\mu)$, $|d'_i| = 0 \leqslant |d_i|$. Assume $i \in B$ such that $\mu_i > f_2(\mu) > \theta$. Then $0 > d'_i = f_2(\mu) - \mu_i > \theta - \mu_i = -(\mu_i - \theta)_+ = d_i$, thus $|d'_i| < |d_i|$. Therefore $|d'| < |d|$, proving (i).

For (ii), note that for $i \in B$, $\mu_i + d'_i = \mu_i - (\mu_i - f_2(\mu))_+ \leqslant f_2(\mu)$, thus $\max_{i \in B} \mu_i + d'_i \leqslant f_2(\mu)$. By Lemma 21, we also have $f_1(\mu + d') \leqslant \max_{\in B} \mu_+ d'_i$, which together with the previous inequality implies (ii).

Lastly, for proving (iii) define $\mathcal{I} = \{i \in B \,:\, \mu_i \geqslant f_2(\mu)\}$. Then $\mu' := \mu + d'$ can be rewritten as

$$\mu'_i = \begin{cases} f_2(\mu)\,, & \text{if } i \in \mathcal{I}\,; \\ \mu_i\,, & \text{otherwise.} \end{cases}$$

By Lemma 22, $f_2(\mu + d') \geqslant f_2(\mu)$, showing (iii) .

The inequality $\theta \leqslant f_1(\mu)$ can also be proved using analogous ideas. Therefore, $\theta \in [f_2(\mu), f_1(\mu)]$. Combining all the previous statements leads to the conclusion $D_\mu^{\min} \subset S$.

Let us now prove that $S \subset D_\mu$. Take any element $\Delta \in S$. Let $j \in [K]$, $B \in \mathcal{B}_1^+$ and $B' \in \mathcal{B}_j^-$ as in the definition of S. WLOG assume that $j = 2$. Let $\mu' = \mu + \Delta$. It suffices to show that $f_1(\mu') \leqslant \theta$ and $f_2(\mu') \geqslant \theta$. We show $f_1(\mu') \leqslant \theta$, leaving the proof of the other relationship to the reader (the proof is entirely analogous to the one presented). By Lemma 21, it suffices to show that $\max\{\mu'_i : i \in B\} \leqslant \theta$. When $i \in B \backslash B'$, $\Delta_i = -(\mu_i - \theta)_+$. Thus, $\mu'_i = \mu_i - \max(\mu_i - \theta, 0) = \mu_i + \min(\theta - \mu_i, 0) \leqslant \mu_i + \theta - \mu_i \leqslant \theta$. If $i \in B \cap B'$, $\mu'_i = \mu_i + (\theta - \mu_i) = \theta$, thus finishing the proof. ■

Appendix D. Proofs for Section 5

We start with the correctness result:

Proposition 8 (Correctness) *On the event ξ, LUCB-micro returns J correctly: $J = j^*(\mu)$.*

Proof Assume to the contrary that $J \neq j^*(\mu)$. WLOG let $j^*(\mu) = 1$. By Assumption 2(i) the definition of ξ and that of J, C_T, the stopping rule, $f_J(\mu) \geqslant f_J(L_T^\delta) \geqslant f_{C_T}(U_T^\delta) \geqslant f_1(U_T^\delta) \geqslant f_1(\mu)$. This contradicts Assumption 3. ■

For proving the sample complexity bound, we consider the following result:

Lemma 9 *Let $t < T$. Then, on ξ, there exists $J \in \{B_t, C_t\}$ such that $c \in [f_J(L_t^\delta), f_J(U_t^\delta)]$ and $f_J(U_t^\delta) - f_J(L_t^\delta) \geqslant \Delta/2$.*

Proof We first prove that $c \in \mathcal{I} \doteq \cup_{j \in \{B_t, C_t\}}[f_j(L_t^\delta), f_j(U_t^\delta)]$. For this, it suffices to show that it does not hold that $c \in \mathcal{I}^c$ where $\mathcal{I}^c = \mathbb{R} \backslash \mathcal{I}$ is the complementer of $\mathcal{I}$. Now, $c \in \mathcal{I}^c$ holds iff at least one of the four conditions hold: *(i)* $f_{B_t}(L_t^\delta) > c$ and $f_{C_t}(L_t^\delta) > c$; *(ii)* $f_{B_t}(U_t^\delta) < c$ and $f_{C_t}(U_t^\delta) < c$; *(iii)* $f_{B_t}(U_t^\delta) < c$ and $f_{C_t}(L_t^\delta) > c$; *(iv)* $f_{B_t}(L_t^\delta) > c$ and $f_{C_t}(U_t^\delta) < c$. Consider the following:

- Case (i) implies that $f_{B_t}(\mu) \geqslant f_{B_t}(L_t^\delta) > c$ and similarly $f_{C_t}(\mu) > c$. Thus there are two arms with payoff greater than c, which contradicts Assumption 3.
- Case (ii) implies that no arm has payoff above c, which contradicts the definition of c.
- Case (iii) Then $f_{C_t}(L_t^\delta) > c > f_{B_t}(U_t^\delta) \geqslant f_{B_t}(L_t^\delta)$, which contradicts the definition of B_t.
- Case (iv) If this is true, then by definition the algorithm has stopped, hence $t \nless T$.

Thus, we see that $c \in \mathcal{I}^c$ cannot hold and hence $c \in [f_J(L_t^\delta), f_J(U_t^\delta)]$ for either $J = B_t$ or $J = C_t$, proving the first part. Next, note that for any $j \in [L]$, $|c - f_j(\mu)| \geqslant \frac{\Delta}{2}$. Hence, also $|c - f_J(\mu)| \geqslant \frac{\Delta}{2}$. Also note that $f_J(\mu) \in [f_J(L_t^\delta), f_J(U_t^\delta)]$. Thus, $f_J(U_t^\delta) - f_J(L_t^\delta) \geqslant |c - f_J(\mu)| \geqslant \frac{\Delta}{2}$. ■

We can now prove Theorem 10:

Theorem 10 (LUCB-micro upper bound) *Let*

$$H(\mu) = \sum_{i \in [L]} \left\{ \frac{1}{(c-\mu_i)^2} \bigwedge \frac{1}{(\Delta/2)^2} \right\}, \quad t^*(\mu) = \min\{t \in \mathbb{N} : 1 + 8H(\mu)\beta(t, \delta/(2L)) \leqslant t\}.$$

Then, for $\delta \leqslant 0.1$, on the event ξ, the stopping time T of LUCB-micro satisfies $T \leqslant t^*(\mu)$.

Proof Let τ be a fixed deterministic integer. Now, on ξ,

$$\begin{aligned}
\min(T,\tau) &\leqslant 1 + \sum_{t=1}^{\tau} \mathbb{I}\{t < T\} \\
&\overset{\text{(a)}}{\leqslant} 1 + \sum_{t=1}^{\tau} \mathbb{I}\left\{\exists J \in \{B_t, C_t\} \text{ s.t. } c \in [f_J(L_t^\delta), f_J(U_t^\delta)] \text{ and } f_J(U_t^\delta) - f_J(L_t^\delta) \geqslant \Delta/2\right\} \\
&\overset{\text{(b)}}{\leqslant} 1 + \sum_{t=1}^{\tau} \mathbb{I}\left\{\exists I \in \{I_t, J_t\} \text{ s.t. } c \in [L_t^\delta(I), U_t^\delta(I)] \text{ and } U_t^\delta(I) - L_t^\delta(I) \geqslant \Delta/2\right\} \\
&\leqslant 1 + \sum_{t=1}^{\tau} \sum_{i \in [L]} \mathbb{I}\{i \in \{I_t, J_t\}\} \, \mathbb{I}\left\{c \in [L_t^\delta(i), U_t^\delta(i)] \text{ and } U_t^\delta(i) - L_t^\delta(i) \geqslant \Delta/2\right\} \\
&\overset{\text{(c)}}{\leqslant} 1 + \sum_{t=1}^{\tau} \sum_{i \in [L]} \mathbb{I}\{i \in \{I_t, J_t\}\} \, \mathbb{I}\left\{N_t(i) \leqslant 8\beta(N_t(i), \delta/(2L)) \left(\frac{1}{(c-\mu_i)^2} \wedge \frac{1}{(\Delta/2)^2}\right)\right\} \\
&\overset{\text{(d)}}{\leqslant} 1 + \sum_{i \in [L]} \sum_{t=1}^{\tau} \mathbb{I}\{i \in \{I_t, J_t\}\} \, \mathbb{I}\left\{N_t(i) \leqslant 8\beta(\tau, \delta/(2L)) \left(\frac{1}{(c-\mu_i)^2} \wedge \frac{1}{(\Delta/2)^2}\right)\right\} \\
&\leqslant 1 + \sum_{i \in [L]} 8\beta(\tau, \delta/(2L)) \left(\frac{1}{(c-\mu_i)^2} \wedge \frac{1}{(\Delta/2)^2}\right) \\
&= 1 + 8H(\mu)\beta(\tau, \delta/(2L)).
\end{aligned}$$

Here, (a) holds by the first part of Lemma 9, (b) holds by Assumption 2(ii), (c) holds by the definition of β, (d) holds because $\beta(\cdot, \delta/(2L))$ is increasing. Picking any τ such that $8H(\mu)\beta(\tau, \delta/(2L)) \leqslant \tau - 1$, we have $\min(T,\tau) \leqslant \tau$, showing that $T \leqslant \min(T,\tau) \leqslant \tau$. ■

Appendix E. Proofs for Section 6

Lemma 12 *For any $h \in H$ and $u, v \in [0,1]^L$ such that $u \leqslant v$, $V(h,u) \leqslant V(h,v)$.*

Proof We prove the result by induction based on how close a history h is to being a maximal history. As in an earlier proof, for $h \in H$, we let $c(h) = 0$ if $h \in H_{\max}$ and otherwise we let $c(h) = 1 + \max\{c(h') : h' \in \mathrm{H}_{\text{succ}}(h)\}$, where recall that $\mathrm{H}_{\text{succ}}(h)$ denotes the set of immediate successors of $h \in H$ in H.

Base case: If $c(h) = 0$ (i.e., $h \in H_{\max}$), then $V(h,u) = u_{\tau(h)} \leqslant v_{\tau(h)} = V(h,v)$.

Induction step: Assuming that for all the $h' \in H$ with $c(h) \leqslant c$ with some $c \geqslant 0$ it holds that $V(h', u) \leqslant V(h', v)$. Take $h \in H$ such that $c(h) = c + 1$. WLOG assume that $p(h) = 1$. We have:

$$\begin{aligned} V(h, u) &= V(\mathrm{join}(h, m(h, u)), u) \\ &\leqslant V(\mathrm{join}(h, m(h, u)), v) \leqslant V(\mathrm{join}(h, m(h, v)), v) \\ &= V(h, v)\,, \end{aligned}$$

where the first and the last equalities are by definition, the first inequality is by the induction hypothesis, and the second inequality is due to the definition of $m(h, v)$. ■

Lemma 14 *Fix $u, v \in \mathbb{R}^L$, $u \leqslant v$, and $j \in [K]$. Let $h = \mathrm{MinMax}((j))$ and in particular let $h = (m_1 = j, m_2, \ldots, m_\ell)$. Then, for all $1 \leqslant k < \ell$,*

$$[V(h_k, u), V(h_k, v)] \subset [V(h_{k+1}, u), V(h_{k+1}, v)]\,,$$

where h_k is the length-k prefix of h.

Proof Fix $0 \leqslant k < \ell$ and $u \leqslant v$. WLOG assume that $p(k) = 1$. By the definition of $V(h, \mu)$ and $m(h, \mu)$,

$$V(h, v) = \max\{V(h', v) \,:\, h' \in \mathrm{H}_{\mathrm{succ}}(h)\} = V(\mathrm{join}(h, m(h, v)), v)\,.$$

Hence, by the definition of MinMax and the above identity, $V(h_k, v) = V(h_{k+1}, v)$. Further, $V(h_k, u) = \max\{V(h', u) \,:\, h' \in \mathrm{H}_{\mathrm{succ}}(h)\} \geqslant V(h_{k+1}, u)$. Thus,

$$V(h_k, v) \leqslant V(h_{k+1}, v) \text{ and } V(h_k, u) \geqslant V(h_{k+1}, u)\,,$$

finishing the proof. ■

Theorem 17 (LUCBMinMax on MinMax Trees) *Let*

$$\begin{aligned} H(\mu) &= \sum_{i \in [L]} \min\{\frac{1}{\mathrm{span}(\mathbb{V}(i, \mu) \cup \{c, \mu_i\})^2}, \frac{4}{\Delta^2}\}, \\ t^*(\mu) &= \min\{t \in \mathbb{N} \,:\, 1 + 8H(\mu)\beta(t, \delta/(2L)) \leqslant t\}\,. \end{aligned}$$

Then, on ξ, the stopping time T of LUCBMinMax *satisfies $T \leqslant t^*(\mu)$.*

Proof Recall that $I_t = \tau(\mathrm{MinMax}(B_t, L_t^\delta, U_t^\delta))$ and $J_t = \tau(\mathrm{MinMax}(C_t, L_t^\delta, U_t^\delta))\}$. Assume that ξ holds. We prove that $\mathbb{V}(I_t, \mu) \subset [L_t^\delta(I_t), U_t^\delta(I_t)]$ and $\mathbb{V}(J_t, \mu) \subset [L_t^\delta(J_t), U_t^\delta(J_t)]$ hold. The rest of the proof is similar to that of Theorem 10.

Consider I_t. The proof for J_t works the same way and is hence omitted. If there is multiple path $h \in H$ such that $\tau(h) = I_t$, then $\mathbb{V}(I_t, \mu) = \varnothing \subset [L_t^\delta(I_t), U_t^\delta(I_t)]$. Otherwise, let $h \in H$ be the unique path. Since I_t is pulled, $h = \mathrm{MinMax}(m)$ for some $m \in M$. Note that Lemma 14 implies that $[V(h_k, L_t^\delta), V(h_k, U_t^\delta)] \subset [V(h_{k+1}, L_t^\delta), V(h_{k+1}, U_t^\delta)]$. Thus it is sufficient to prove that for $1 \leqslant k < \ell$, $V(h_k, \mu) \in [V(h_k, L_t^\delta), V(h_k, U_t^\delta)]$. However, this follows by Lemma 12 and because on the event ξ, $L_t^\delta \leqslant \mu \leqslant U_t^\delta$ holds.

Now let $S(i) = \mathbb{V}(i, \mu) \cup \{c, \mu_i\}$. Fix $t < T$. By the above result and by Lemma 9, for one of $J = B_t$ or $J = C_t$, if $I = \mathrm{MinMax}(J, L_t^\delta, U_t^\delta)$ then $S(I) \subset [L_t^\delta(I), U_t^\delta(I)]$, which implies that $U_t^\delta(I) - L_t^\delta(I) \geqslant \mathrm{span}(S(I))$. Therefore,

$$\begin{aligned}
\min(T, \tau) &\leqslant 1 + \sum_{t=1}^{\tau} \mathbb{I}\{t < T\} \\
&\leqslant 1 + \sum_{t=1}^{\tau} \mathbb{I}\left\{\exists I \in \{I_t, J_t\} \text{ s.t. } U_t^\delta(I) - L_t^\delta(I) \geqslant \mathrm{span}(S(I))\right\} \\
&\leqslant 1 + \sum_{t=1}^{\tau} \sum_{i \in [L]} \mathbb{I}\{i \in \{I_t, J_t\}\} \, \mathbb{I}\left\{N_t(i) \leqslant \frac{8\beta(N_t(i), \delta/(2L))}{\mathrm{span}(S(i))^2}\right\} \\
&\leqslant 1 + \sum_{i \in [L]} \sum_{t=1}^{\tau} \mathbb{I}\{i \in \{I_t, J_t\}\} \, \mathbb{I}\left\{N_t(i) \leqslant \frac{8\beta(\tau, \delta/(2L))}{\mathrm{span}(S(i))^2}\right\} \\
&\leqslant 1 + \sum_{i \in [L]} \frac{8\beta(\tau, \delta/(2L))}{\mathrm{span}(S(i))^2} = 1 + 8H(\mu)\beta(\tau, \delta/(2L))
\end{aligned}$$

■