Adaptivity to Smoothness in $\mathcal{X}$-armed bandits

Andrea Locatelli ANDREA.LOCATELLI@OVGU.DE
Mathematics Department, Otto-von-Guericke-Universität Magdeburg

Alexandra Carpentier ALEXANDRA.CARPENTIER@OVGU.DE
Mathematics Department, Otto-von-Guericke-Universität Magdeburg

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We study the stochastic continuum-armed bandit problem from the angle of adaptivity to *unknown regularity* of the reward function f. We prove that there exists no strategy for the cumulative regret that adapts optimally to the *smoothness* of f. We show however that such minimax optimal adaptive strategies exist if the learner is given *extra-information* about f. Finally, we complement our positive results with matching lower bounds.

Keywords: bandits with infinitely many arms, minimax rates, adaptivity, smoothness

1. Introduction

In the classical multi-armed bandit problem, an online algorithm (the *learner*) attempts to maximize its gains by sequentially allocating a portion of its budget of n pulls among a finite number of available options (arms). As the learner starts with no information about the environment it is facing, this naturally induces an exploration/exploitation trade-off. The learner needs to make sure it explores sufficiently to perform well in the future, without neglecting immediate performance entirely. In this setting, the performance of the learner can be measured by its *cumulative regret*, which is the difference between the sum of rewards it would have obtained by playing optimally (i.e. only choosing the arm with the highest expected reward), and the sum of rewards it has collected.

Continuum-armed bandit problems. In this work, we operate in a setting with infinitely many arms, which are embedded in $\mathcal{X}$ a bounded subset of $\mathbb{R}^d$, say $[0,1]^d$. Each arm $x \in \mathcal{X}$ is associated to a mean reward $f(x)$ through the reward function f. At each time t, the learner picks $X_t \in [0,1]^d$, and receives a noisy sample $Y_t = f(X_t) + \epsilon_t$ with $\mathbb{E}(Y_t) = f(X_t)$. This continuous setting is very relevant for practitioners: for example, if a company wishes to optimize the revenue associated with the price of a new product, it should consider the continuum $\mathbb{R}^+$ of possible prices. While it is known (see for example Bubeck et al. (2011b)) that in the absence of additional assumptions that link $\mathcal{X}$ and the reward function, there exists no universal algorithm that achieves sub-linear regret in this setting with infinitely many arms, under some additional structural assumptions on the reward function (such as unimodality), it is possible to optimize this price *online* to achieve non-trivial regret guarantees. When $\mathcal{X}$ is a metric space, a common assumption in the literature is to consider smooth reward functions (Agrawal (1995); Kleinberg (2004)). This *smoothness* of the reward function can either be local (Auer et al. (2007); Grill et al. (2015)) or global (Kleinberg et al. (2008); Cope (2009); Bubeck et al. (2011c); Minsker (2013)). In most of these works, the smoothness of the reward function is *known* to the learner: for example, if f such that for any $x, y \in \mathcal{X}$, we have $|f(x) - f(y)| \leq L|x-y|_\infty^\alpha$ [1], then the learner has access to L and α (see e.g. Auer et al. (2007); Bubeck et al. (2011c)). Furthermore, in this work we will use a parametrization akin to the popular Tsybakov noise condition (see e.g. Tsybakov (2004); Audibert and Tsybakov (2007)). As in Auer et al. (2007); Minsker (2013), we will assume that the volume of Δ-optimal regions decreases as $\mathcal{O}\left(\Delta^\beta\right)$ for some unknown $\beta \geq 0$. Under these assumptions, there exists strategies as e.g. HOO in Bubeck et al. (2011c)[2], that enjoy nearly optimal cumulative regret bounds of order $\tilde{\mathcal{O}}\left(n^{(\alpha+d-\alpha\beta)/(2\alpha+d-\alpha\beta)}\right)$[3], if they are tuned optimally with α. Importantly, these strategies naturally adapt to β, which controls the difficulty of the problem (with the hardest case $\beta = 0$). However, it is argued in Bubeck et al. (2011a) that this perspective is flawed, as one should instead consider strategies that can *adapt* to multiple different environments - and not strategies that are adapted to a specific environment.

Adaptivity in continuum-armed bandit. While the problem of adaptivity to unknown Lipschitz constant L (with $\alpha = 1$ known to the learner) for cumulative regret minimization has been studied in Bubeck et al. (2011a), adaptivity to unknown smoothness exponent α remains a very important open question, which, to the best of our knowledge, has only been studied in optimization. In optimization, the learner's goal is to recommend a point $x(n) \in \mathcal{X}$ such that its *simple regret* $r_n = \sup_{x\in\mathcal{X}} f(x) - f(x(n))$ is as small as possible. It has first been shown in Valko et al. (2013) (which is an extension from Munos (2011) that operates in a deterministic setting) that when $\alpha\beta = d$ i.e. if the function is *easy* to optimize[4], there exists adaptive strategies with optimal simple regret of order $\tilde{\mathcal{O}}\left(n^{-1/2}\right)$. These results were later extended in Grill et al. (2015) to the more general setting $\alpha\beta \leq d$, in which case their adaptive algorithm POO has an expected simple regret upper-bounded as $\tilde{\mathcal{O}}\left(n^{-\alpha/(2\alpha+d-\alpha\beta)}\right)$, without prior knowledge of the smoothness. This leaves open two questions. First, is this bound minimax optimal for the simple regret? And, more importantly, outside of very restrictive technical conditions on f such (e.g. self-similarity as in Minsker (2013)), is there a smoothness adaptive strategy such its cumulative regret can be upper-bounded as $\tilde{\mathcal{O}}\left(n^{(\alpha+d-\alpha\beta)/(2\alpha+d-\alpha\beta)}\right)$ for all α and β?

Adaptivity in statistics. Even though the concept of smoothness adaptive procedures is still fairly unexplored in the continuum-armed bandit setting, it has been studied extensively in the statistics literature under the name of *adaptive inference*. The first question in this field is the one of constructing estimators that adapt to the unknown model at hand (e.g. to the smoothness), i.e. adaptive estimators (see among many others Golubev (1987); Birgé and Massart (1997); Lepski and Spokoiny (1997); Tsybakov (2004)). The main takeaway is that adaptivity to unknown regularity for *estimation* is possible under most standard statistical models using model selection or aggregation techniques. These adaptive strategies were later adapted to sequential settings such as active learning by Hanneke; Koltchinskii (2010); Minsker (2012); Locatelli et al. (2017) or nonparametric optimization Grill et al. (2015), where they use a cross-validation scheme. These approaches however are not suited for cumulative regret minimization, as they typically trade-off exploitation in favor of exploration. Another fundamental question in adaptive inference is the construction of *adaptive and honest* confidence sets. Importantly, such confidence sets would naturally give rise to an upper-confidence bound type

[1] In fact, as in Bubeck et al. (2011b), we will only assume f to be *weakly-Lipschitz*, allowing us to consider $\alpha > 1$ - see Definition 1

[2] In Bubeck et al. (2011c) problems are parametrized with the *near-optimality* dimension D. Under our smoothness assumptions, these two parametrizations are equivalent with $D = \frac{d-\alpha\beta}{\alpha}$.

[3] We use the $\tilde{\mathcal{O}}$ notation to hide logarithmic factors n or δ^{-1}

[4] This assumption corresponds to the fact that the *near-optimality* dimension D from Bubeck et al. (2011c) is 0, i.e. roughly functions that have a unique maximum x^* and depart from it faster than $|x - x^*|_\infty^\alpha$.

of strategy with optimal adaptive cumulative regret guarantees. However a fundamental negative result is the non-existence of adaptive confidence sets in L_∞ for Hölder smooth functions Juditsky and Lambert-Lacroix (2003); Cai et al. (2006); Hoffmann and Nickl (2011). Interestingly, adaptive confidence sets for regression do exist under additional assumptions on the model, such as *shape constraints* (see e.g. Cai et al. (2013); Bellec (2016)).

Learning with Extra-information. In the classical multi-armed bandit problem, this shape constrained setting was introduced in Bubeck et al. (2013). They show that if the learner is supplied with the mean reward μ^* of the best arm, and Δ the *gap* between μ^* and the second best arm's mean reward, then there exists a strategy with *bounded* regret. Recently, it was shown in Garivier et al. (2016) that only the knowledge of μ^* is necessary to achieve bounded regret. Outside of the very important and studied convexity constraint, such questions remain unexplored in our nonparametric setting, with the exception of Kleinberg et al. (2013). In this work, they consider the case where $\sup_{x\in\mathcal{X}} f(x) \approx 1$ and the noisy rewards Y_t are bounded in $[0,1]$ (i.e. the noise decays close to the maxima). Under these assumptions, they obtain faster rates for the cumulative regret in the case where f is Lipschitz. This leaves open the question whether shape constraints could facilitate adaptivity to unknown smoothness when the cumulative regret is targeted. Finally, we remark that the case $\alpha\beta = d$, which can be thought of as a shape constraint as well, has been partially treated in Bull et al. (2015) for the special class of *zooming continuous* functions (first studied in Slivkins (2011)). In this setting, Bull et al. (2015) introduced an adaptive strategy such that its expected cumulative regret is bounded as $\tilde{\mathcal{O}}\left(\sqrt{n}\right)$. However, it was shown in Grill et al. (2015) (see Appendix E therein) that the class of functions we consider here is more general than the one in Slivkins (2011); Bull et al. (2015), making these two lines of work not directly comparable. In a one-dimensional setting equivalent to ours for $\alpha\beta = 1$ but with the additional constraint that f is unimodal, Yu and Mannor (2011) and Combes and Proutiere (2014) also get an adaptive rate for the cumulative regret of order $\tilde{\mathcal{O}}\left(\sqrt{n}\right)$. Extending these results to our entire class of functions is a relevant question in this canonical setting.

1.1. Contributions and Outline

We now state our main contributions.

- Our main result Theorem 3 proves that no strategy can be optimal simultaneously over all smoothness classes for cumulative regret minimization.
- We show that under various shape constraints, adaptivity to unknown smoothness becomes possible if the learner is given this extra-information about the environment. In particular, we show that in the case $\alpha\beta = d$, there exists a smoothness adaptive strategy whose regret grows as $\tilde{\mathcal{O}}\left(\sqrt{n}\right)$ i.e. independently of α and d, without access to α.
- Finally, we show lower bounds for the simple and cumulative regret that match the known upper-bounds. Importantly, these bounds also hold in the shape-constrained settings.

In Section 2, we introduce our setting formally and show a high-probability result for a simple non-adaptive Subroutine (`SR`). In Section 3, we prove a lower-bound for the simple regret that matches the best known upper-bound for adaptive strategies (such as `POO` in Grill et al. (2015)) in the optimization setting. We then prove our main result on the non-existence of adaptive strategies for cumulative regret minimization. In Section 4, we study the shape constrained settings and introduce an adaptive Meta-Strategy, which relies on `SR` and our high-probability result of Section 2.

2. Preliminaries

2.1. Objective

We consider the d-dimensional continuum-armed bandit problem. At each time step $t = 1, 2, \dots, n$, the learner chooses $X_t \in [0,1]^d$ and receives a return (or *reward*) $Y_t = f(X_t) + \epsilon_t$. We will further assume that ϵ_t is independent from $\big((X_1, Y_1), \dots (X_{t-1}, Y_{t-1})\big)$ conditionally on X_t, and it is a zero-mean 1-sub-Gaussian[5] random variable. Finally we assume that f takes values in a bounded interval, say $[0,1]$ and we denote $M(f) \doteq \sup_{x\in[0,1]^d} f(x)$. In optimization, the objective of the learner is to recommend at the end of the game a point $x(n) \in [0,1]^d$, such that the following loss

$$r_n = M(f) - f(x(n))$$

is as small as possible, under the constraint that it can only observe n couples (X_t, Y_t) before making its recommendation. In the rest of the paper, we will refer to r_n as the *simple regret*. This objective is different from the typical bandit setting, where the cumulative regret $\widehat{R}_n = nM(f) - \sum_{t=1}^n Y_t$ is instead targeted. As a proxy for the cumulative regret, we will study the cumulative *pseudo-regret*:

$$R_n = nM(f) - \sum_{t=1}^{n} f(X_t).$$

By the tower-rule, $\mathbb{E}(Y_t) = \mathbb{E}(\mathbb{E}(Y_t|X_t)) = \mathbb{E}(f(X_t))$, and thus we have $\mathbb{E}(\widehat{R}_n) = \mathbb{E}(R_n)$, where the expectation is taken with respect to the samples collected by the strategy and its (possible) internal randomization. Our primary goal will be to design sequential exploration strategies, such that the next point to sample X_t may depend on all the previously collected samples $(X_i, Y_i)_{i<t}$, in order to optimize one of these two objectives. We note here that one can easily show that a strategy with good *cumulative regret* gives rise naturally to a strategy with good *simple regret* (for example, by choosing $x(n)$ uniformly at random over the points visited). However, the converse is obviously not true.

2.2. Assumptions

In this section, we state our assumptions on the mean reward function $f : [0,1]^d \to [0,1]$. Our first assumption characterizes the continuity, or *smoothness* of f.

Definition 1 *We say that $g : [0,1]^d \to [0,1]$ belongs to the class $\Sigma(\lambda, \alpha)$ if there exists constants $\lambda \geq 1$, $\alpha > 0$ such that for any $x, y \in [0,1]^d$:*

$$g(x) - g(y) \leq \max\{M(g) - g(x), \lambda|x - y|_\infty^\alpha\},$$

where $|z|_\infty = \max_{i\leq d} z^{(i)}$ and $z^{(i)}$ denotes the value of the i-th coordinate of the vector z, with $M(g) \doteq \sup_{x\in[0,1]^d} g(x)$.

For completeness, we also define the Hölder smoothness classes for $\alpha \in (0,1]$.

Definition 2 *We say that $g : [0,1]^d \to [0,1]$ belongs to the Hölder smoothness class $\Sigma^*(\lambda, \alpha)$ if there exists constants $\lambda \geq 1$, $0 < \alpha \leq 1$ such that for any $x, y \in [0,1]^d$:*

$$|g(x) - g(y)| \leq \lambda|x - y|_\infty^\alpha.$$

[5] We say that a random variable Z is σ-sub-Gaussian if for all $t \in \mathbb{R}$, we have $\mathbb{E}[\exp(tZ)] \leq \exp(\frac{\sigma^2 t^2}{2})$

Assumption 1 *There exists constants $\lambda \geq 1$, $\alpha > 0$ such that $f \in \Sigma(\lambda, \alpha)$.*

This assumption forbids the function f from jumping erratically close to its maximum, which would render learning extremely difficult. Indeed, for any x^* such that $f(x^*) = M(f)$, the condition simply rewrites for any $x \in [0,1]^d$:

$$M(f) - f(x) \leq \lambda |x^* - x|_\infty^\alpha.$$

For $\alpha \leq 1$, it is weaker than assuming that f belongs to the Hölder class $\Sigma^*(\lambda, \alpha)$, which is the case for example in Kleinberg (2004); Minsker (2013) (it is important to note that in Minsker (2013) a second assumption related to the notion of *self-similarity* is required to allow adaptivity to unknown smoothness α). Moreover, it allows us to consider $\alpha > 1$, without forcing the function to be constant. Our second assumption is similar to the well known *margin assumption* (also called Tsybakov noise condition) in the binary classification framework.

Assumption 2 *Let $\mathcal{X}(\Delta) \doteq \{x : M(f) - f(x) \leq \Delta\}$. There exists constants $B > 0$, $\beta \in \mathbb{R}^+$ such that $\forall \Delta > 0$:*

$$\mu(\mathcal{X}(\Delta)) = \mu\left(\{x : M(f) - f(x) \leq \Delta\}\right) \leq B\Delta^\beta,$$

where μ stands for the Lebesgue measure of a set $S \subset [0,1]^d$.

This assumption naturally captures the difficulty of finding the maxima of f: if β is close to 0, there is no restriction on the Lebesgue measure of the Δ-optimal set - on the other hand, if β is large, there are less potentially optimal regions in the space, and we hope that a good algorithm will take advantage of this to focus on these regions more closely, by discarding the many sub-optimal regions quicker.
Intuitively, the smoother f is around one of its maxima x^*, the harder it is for it to "take-off" from x^*, and thus higher values for β are geometrically impossible. The following proposition (its proof is in Appendix A.1) formalizes this intuition, and characterizes the interplay between the different parameters of the problem, α, β and d.

Proposition 1 *If f is such that Assumptions 1 and 2 are satisfied for $\alpha > 0, \beta \in \mathbb{R}^+$, then $\alpha\beta \leq d$.*

In the rest of the paper, we will fix $B > 0$ as well and $\lambda = 1$. This can be relaxed to $\lambda \geq 1$ or a known upper bound on λ, such as $\log(n)$ for n large enough, being known to the learner. We make this choice as our goal in the present work is to fundamentally understand adaptivity with respect to the smoothness α.

Definition 3 *We say that $f \in \mathcal{P}(\alpha, \beta) \doteq \mathcal{P}(\lambda, \alpha, \beta, B, [0,1]^d)$ if f is such that Assumptions 1 and 2 are satisfied for $\alpha > 0, \beta \geq 0$.*

2.3. A simple strategy for known smoothness

The main building block on which our adaptive results are built is a non-adaptive Subroutine (SR), which takes α as input and operates on the dyadic partition of $[0,1]^d$. Importantly, our results depend on bounds that hold with high-probability, whereas to the best of our knowledge, the analysis of the HOO in Bubeck et al. (2011c) yields results in expectation. For completeness, we introduce and analyze this simple Subroutine. The strategy, its description and analysis can be found in the Appendix A.2. We now state our main result for this non-adaptive Subroutine.

Proposition 2 *Let $n \in \mathbb{N}^*$. The Subroutine run on a problem characterized by $f \in \mathcal{P}(\alpha, \beta)$ with input parameters α, n and $0 < \delta < e^{-1}$ is such that with probability at least $1 - 4\delta$:*

- $\mathcal{X}(0) \subset \mathcal{A}_{L+1} \subset \mathcal{X}\left(C\left(\frac{n}{\log(\frac{n}{\delta})}\right)^{-\alpha/(2\alpha+d-\alpha\beta)}\right)$*, where $C > 0$ does not depend on n, δ.*
- *For any recommendation, $x(n) \in \mathcal{A}_{L+1}$, we have: $M(f) - f(x(n)) \leq C\left(\frac{n}{\log(\frac{n}{\delta})}\right)^{-\alpha/(2\alpha+d-\alpha\beta)}$*
- *For all $T \leq n$, we have $R_T \leq D \log(\frac{n}{\delta})^{\alpha/(2\alpha+d-\alpha\beta)} T^{(\alpha+d-\alpha\beta)(2\alpha+d-\alpha\beta)}$, where $D > 0$ is a constant that does not depend on T, n, δ, α.*

The proof of this result can be found in Appendix A.3. The second conclusion of Proposition 2 is a direct implication of the first conclusion, and shows that with high-probability, as we recover an entire level set of optimal size, recommending *any* point in the active set $\mathcal{A}_{L+1}$ leads to optimal simple regret. This will prove handy for adaptivity to unknown smoothness for the simple regret objective. The third conclusion will be used in Section 4, where we show that if the learner is provided with extra-information, adaptivity to unknown smoothness is possible for cumulative regret.

3. Adaptivity to unknown smoothness in optimization and regret minimization

In this section, we explore the problem of adaptivity to *unknown* smoothness α for both the simple regret and cumulative regret objectives. We show that for optimization, adaptivity is possible without sacrificing minimax optimality: there exists an agnostic strategy that performs almost as well as the optimal strategy that has access to the smoothness. For cumulative regret, we show that there exists no adaptive minimax optimal strategy.

3.1. Adaptivity for optimization

We start by proving a lower bound on the simple regret over the class of functions $\mathcal{P}(\alpha, \beta)$, which holds even for strategies that have access to both α and β.

Theorem 1 (Lower bound on simple regret) *Fix $d \in \mathbb{N}^*$. Let $\alpha > 0$ and $\beta \geq 0$ such that $\alpha\beta \leq d$. For n large enough, for any strategy that samples at most n noisy function evaluations and returns a (possibly randomized) recommendation $x(n)$, there exists $f \in P(\alpha, \beta)$, where $M(f)$ is fixed and known to the learner, such that:*

$$\mathbb{E}[r_n] \geq C n^{-\alpha/(2\alpha+d-\alpha\beta)},$$

where $C > 0$ is a constant that does not depend on n, and the expectation is taken with respect to both the noise in the sampling process and the possible randomization of the strategy.

The proof of this result can be found in Appendix B.1. It shows that even over a set of functions that all belong to *known* class $\mathcal{P}(\alpha, \beta)$, this is the best possible convergence rate for the simple regret that one can hope for. An important takeaway from the proof of this result is that it also holds in the easier setting where $M(f)$ the maximum of f is known to the learner. A direct corollary of this result is a lower bound on the cumulative regret for any strategy.

Corollary 1 (Lower bound on cumulative regret) *Fix $d \in \mathbb{N}^*$. Let $\alpha > 0$ and $\beta \geq 0$ such that $\alpha\beta \leq d$. For n large enough, any strategy with access to at most n noisy function evaluations suffers a cumulative regret such that:*

$$\sup_{f \in \mathcal{P}(\alpha,\beta)} \mathbb{E}[R_n] \geq C n^{(\alpha+d-\alpha\beta)/(2\alpha+d-\alpha\beta)},$$

where $C > 0$ is a constant that does not depend on n, and the expectation is taken with respect to both the noise in the sampling process and the possible randomization of the strategy.

This result follows directly from Theorem 1, by remarking that any strategy with a good cumulative regret in expectation can output a recommendation $x(n)$ such that $\mathbb{E}[r_n] \leq \frac{\mathbb{E}[R_n]}{n}$ (see Section 3 in Bubeck et al. (2011b)). Therefore, any strategy with a cumulative regret that's strictly smaller than the rate in Corollary 1 would have an associated simple regret in contradiction with Theorem 1.

We now exhibit *adaptive* strategies that are minimax optimal (up to log factors) for the simple regret. Importantly, these strategies perform almost as well as the best strategies that have access to α and β.

Theorem 2 (Adaptive upper-bound for simple regret) *Let $n \in \mathbb{N}^*$. Assume that $\alpha \in [1/\log(n), \log(n)]$ and $\beta \geq 0$ such that $\alpha\beta \leq d$, both unknown to the learner. There exists adaptive strategies such that for any $f \in \mathcal{P}(\alpha, \beta)$ with maximum $M(f)$:*

$$M(f) - \mathbb{E}[f(x(n))] \leq C \left(\frac{\log^p(n)}{n}\right)^{\alpha/(2\alpha+d-\alpha\beta)},$$

where $C > 0$ is a constant that does not depend on n and p is a universal constant.

In order to match the rate in Theorem 1 for the simple regret, a natural strategy is to aggregate different recommendations output by a non-adaptive (i.e. that takes the smoothness α as input) strategy, run with a diversity of smoothness parameters. We exhibit two such strategies that rely on this scheme.

Strategy 1 (Cross-validation): Grill et al. (2015) introduces a strategy (`POO`) that adapts to unknown smoothness for the simple regret. It launches several `HOO`(i) (Bubeck et al. (2011c)) instances in parallel according to a logarithmic schedule over the smoothness parameters α_i (indexing the instances). The final recommendation of the Meta-Strategy is made by first choosing the instance `HOO`(i^*) with the best average empirical performance. The final recommendation is then drawn uniformly at random over the points $\{X_{i^*}(t)\}_t$ visited by `HOO`(i^*). An important technical remark is that the fastest attainable rate in this setting is $\mathcal{O}\left(1/\sqrt{n}\right)$, which is is of the same order as the stochastic error induced by the final cross-validation scheme. For this strategy, we have $p = 2$ in Theorem 2.

Strategy 2 (Nested Aggregation): The first conclusion of Proposition 2 shows that our Subroutine recovers with high-probability an *entire level-set* of optimal size. As the smoothness classes $\Sigma(1, \alpha)$ are nested for increasing values of α, this allows us to use directly the nested aggregation scheme (Algorithm 1) in Locatelli et al. (2017) by splitting the budget among several `SR` instances indexed by smoothness parameters α_i over a grid that covers the range $[1/\lfloor\log(n)\rfloor, \lfloor\log(n)\rfloor]$. Importantly, the final recommendation $x(n)$ output by this nested aggregation procedure comes with high-probability guarantees which is an improvement over `POO`.

A common caveat of these adaptive strategies is that their exploration of the space crucially depends on a covering of the possible smoothness parameters. This is necessary to ensure that there is a Subroutine run with a smoothness parameter which is very close to the true smoothness of the function. However, Subroutines (either our Subroutine 2 or `HOO`) run with smoothness parameters $\alpha_i \ll \alpha$ incur a high-regret as they explore at a too small scale, while subroutines run with $\alpha_i > \alpha$ come with no regret guarantee. As the budget is split equally among the Subroutines run in parallel, the total cumulative regret of these adaptive exploration strategies cannot be bounded and is provably sub-optimal. This naturally leads to the following question: is there an adaptive strategy that enjoys a minimax optimal cumulative regret over classes $\mathcal{P}(\alpha, \beta)$?

3.2. Impossibility result for cumulative regret

In this section, we answer the previous question negatively, and show that designing an adaptive strategy with minimax optimal cumulative regret is a hopeless quest. We first state this result in a general theorem and then instantiate it in multiple settings to show its implications.

Theorem 3 *Fix $\gamma \geq \alpha > 0$ and $\beta \geq 0$ such that $\gamma\beta \leq d$. Consider a strategy such that for any $f \in \mathcal{P}(\gamma, \beta)$, we have $\mathbb{E}[R_n] \leq R_{\gamma,\beta}(n)$ with $R_{\gamma,\beta}(n)^{(2\alpha+d-\alpha\beta)/(\alpha+d-\alpha\beta)} \leq 0.008n$. Then this strategy is also such that:*

$$\sup_{f\in\mathcal{P}(\alpha,\beta)} \mathbb{E}[R_n] \geq 0.008nR_{\gamma,\beta}(n)^{-\alpha/(\alpha+d-\alpha\beta)},$$

where the expectations are taken with respect to the strategy and the samples collected.

The proof of this result can be found in Appendix A.3 and uses the same techniques as in the proof of Theorem 1, but with the following twists: the value of the maximum across the set of problems we consider is not fixed, nor is the value of the smoothness, which can be either be α or γ, depending on the presence of a rough peak of smoothness α. This construction forces any strategy into an exploration exploitation dilemma parametrized by $R_{\gamma,\beta}(n)$.

Theorem 3 can be understood in the following way: for any strategy, performing at a certain rate $R_{\gamma,\beta}(n)$ uniformly over all problems in a subclass $\mathcal{P}(\gamma, \beta) \subset \mathcal{P}(\alpha, \beta)$ comes with a price: on at least one problem that belongs to the class $\mathcal{P}(\alpha, \beta)$, it has to suffer an expected regret that depends inversely on $R_{\gamma,\beta}(n)$. This directly leads to our claim that adaptivity to the smoothness for the cumulative regret objective is impossible. Consider strategies such that $R_{\gamma,\beta}(n) \leq \mathcal{O}\left(n^{1-\gamma/(2\gamma+d-\gamma\beta)+\epsilon}\right)$ for any $\epsilon > 0$ (we showed in Proposition 2 that such strategies exist). Then its regret over the class $\mathcal{P}(\alpha, \beta)$ is necessarily lower bounded as $\mathcal{O}\left(n^{1-\alpha/(2\alpha+d-\alpha\beta)+\nu}\right)$, where $\nu = \left(\frac{\alpha+d-\alpha\beta}{2\alpha+d-\alpha\beta} - \frac{\gamma+d-\gamma\beta}{2\gamma+d-\gamma\beta} - \epsilon\right)\frac{\alpha}{\alpha+d-\alpha\beta}$. As soon as $\alpha < \gamma$, we have $\nu > 0$ for ϵ small enough, which implies that the strategy considered is strictly sub-optimal over the class $\mathcal{P}(\alpha, \beta)$. We remark that by plugging $\alpha = \gamma$ in Theorem 3, we recover the lower-bound of Corollary 1. We now illustrate our impossibility result in a very simple one-dimensional setting with $\beta = 1$.

Example. Fix $\gamma = 1$ and $\alpha = 1/2$, as well as $d = 1$ and $\beta = 1$. The minimax optimal rate for the cumulative regret over $\mathcal{P}(1,1)$ is of order $\mathcal{O}\left(\sqrt{n}\right)$. One can easily check that the minimax optimal rate for the class $\mathcal{P}(1/2,1)$ is of order $\mathcal{O}\left(n^{2/3}\right)$. The previous Theorem tells us that any strategy that achieves a regret of order $\mathcal{O}\left(n^{1/2}\right)$ over $\mathcal{P}(1,1)$ incurs a regret of order at least $\mathcal{O}\left(n^{3/4}\right)$ on a problem in $\mathcal{P}(1/2,1)$, which is strictly sub-optimal.

3.3. Discussion

This result shows that for the problem of adaptivity to unknown smoothness, there exists a fundamental difference between optimization and cumulative regret minimization. In optimization, adaptivity to unknown smoothness is possible (at the price of a logarithmic factor), while Theorem 3 rules out the existence of strategies that are minimax optimal simultaneously for two smoothness classes. This fundamental difference is related to the adaptive inference paradox in statistics: while adaptive estimation is usually possible, adaptive and honest confidence sets usually do not exist over standard models Cai et al. (2006); Hoffmann and Nickl (2011). The problem of simple regret minimization is akin to adaptive estimation, as it is a pure exploration problem. Model selection techniques (as e.g. cross validation or Lepski's methods) can be safely employed to aggregate the output of several Subroutines run in parallel and corresponding to different values of α, enabling thus adaptivity to α. In a sense, there is no price to pay if one over-explores, which is akin to over-smoothing in adaptive estimation. On the other hand, the problem of cumulative regret minimization requires a careful trade-off between exploration and exploitation. Since this trade-off should depend on the unknown α *exactly*, this leaves no room for over-exploration. This bears strong similarities with model testing and adaptive uncertainty quantification, i.e. the problem of constructing adaptive and honest confidence sets, and as such it is not possible to adapt to the smoothness for the problem of cumulative regret minimization. This is particularly interesting in light of Bubeck et al. (2011b), where it is remarked that any strategy with good cumulative regret naturally gives rise to a strategy with good simple regret. We show here that in this adaptive setting, the minimax optimal attainable rates are not identical (up to a factor n). The proof of this result crucially depends on the fact that the value of the maximum over the class of functions we consider is not fixed and depends on the smoothness of f, which forces any strategy into an exploration and exploitation dilemma. We also remark here that β is fixed in our construction: this shows that even for known β, minimax optimal adaptive strategies over the classes $\cup_{\alpha>0}\mathcal{P}(\alpha,\beta)$ do not exist, and the intrinsic difficulty in the problem of adaptivity is tied to the unknown smoothness. Interestingly, despite β being fixed, the minimax rate itself is not fixed as it depends on the smoothness which can take values α and γ. Finally, we remark that this rate is tight in the sense that there exists a strategy that takes $R_{\gamma,\beta}(n)$ and α, γ, β as inputs and incurs the regret on $\mathcal{P}(\alpha,\beta)$ and $\mathcal{P}(\gamma,\beta)$ prescribed by Theorem 3. This strategy is simply to use $R_{\gamma,\beta}(n)^{(2\alpha+d-\alpha\beta)/(\alpha+d-\alpha\beta)}$ samples with $\texttt{SR}(\alpha)$, and afterwards to play $\texttt{SR}(\gamma)$ within the confidence set output by $\texttt{SR}(\alpha)$.

Even though adaptivity to the unknown smoothness for cumulative regret minimization is impossible in general, an interesting open problem is to find natural conditions under which adaptivity becomes possible, which we explore in the next section. This course of research was also taken in the problem of constructing adaptive and honest confidence sets, and while they mostly do not exist in all generality, it is well known that under some specific shape constraints, they exist Cai et al. (2013); Bellec (2016). We refer to these settings as learning with *extra-information*. First, we will show that adaptivity is possible over the subclass $\cup_{\alpha>0}\mathcal{P}(\alpha,\beta,M(f))$ where $M(f)$ denotes the *fixed* value of f at its maxima. Next, we will show that adaptivity is possible over classes $\cup_{\alpha>0}\mathcal{P}(\alpha, \beta(\alpha))$ for $\beta(\alpha) = (2r-1)/r + d/\alpha$ for some fixed $r \in [0, 1/2]$.

4. Learning in the presence of extra-information

In this section, we investigate two settings where the learner is given *extra-information* and show that adaptivity to unknown smoothness is possible for the cumulative regret. We explore two conditions: the case where $M(f)$ the value at the maxima is known to the learner and the *known rate* setting, which we describe later. To solve these problems, we introduce meta-strategies which act on a set of subroutines (Subroutine 2, SR) initialized with different smoothness parameters. Specifically, different runs of Subroutine 2 are kept active in parallel, and at each round the Meta-Strategy decides *online* to further allocate a fraction $\sqrt{n}$ of the total budget n to Subroutines that exhibit good early performances, in a sense we shall make clear later. Each time a Subroutine is given a fraction of the budget to perform new function evaluations, learning resumes for this Subroutine where it was halted: we stress here that the information acquired by Subroutines is never thrown.

Known $M(f)$ setting. At the beginning of the game, the learner is given $M(f)$ the value of f at its maxima, allowing for more efficient exploitation. In light of our the proof of Theorem 3 (which does not cover this setting), we see intuitively that the exploitation exploration dilemma leading to the impossibility result arose from the two different values that $M(f)$ could take in our class of functions. Here, as soon as the strategy has identified a region where f is close in value to $M(f)$, it can exploit aggressively and keep track on-the-fly of the regret it incurs. By being aware of its own performance, the learner can adjust its exploration/exploitation trade-off optimally.

Known rate setting. The learner is provided with extra-information $R^*(n,\delta)$ that we call the *rate*. $R^*(n,\delta)$ is a high-probability bound on the pseudo-regret of one of the Subroutines used by the Meta-Strategy, had it been run in isolation with a budget n of function evaluations. Although it is more general, this covers the canonical case $\alpha\beta = d$. A similar setup was explored in the recent work Agarwal et al. (2017), where they come up with a meta-strategy to aggregate bandit algorithms that also works under adversarial settings.

4.1. Description of the Meta-Strategy

We first describe the initialization phase of the Meta-Strategy and notations, and then explain how it operates in each setting.

Initialization: The Meta-Strategy has three parameters: the maximum budget n, which we assume for simplicity to be of the form m^2 for some $m \in \mathbb{N}^*$, and a confidence parameter δ, as well as an extra-information parameter $M(f)$ or $R^*(n,\delta)$. It uses multiple instances of Subroutine 2, which are run in parallel with smoothness parameters α_i over the grid $\{i/\lfloor\log(n)\rfloor^2\}$ with $i \in \{1, ..., \lfloor\log(n)\rfloor^3\}$. First, each Subroutine is initialized with a smoothness parameter α_i, a confidence parameter $\delta_0 = \delta/\lfloor\log(n)\rfloor^3$, and we refer to this Subroutine as $\mathtt{SR}(i)$. $T_i(T)$ is the number of function evaluations performed by $\mathtt{SR}(i)$ from time $t = 1$ to T. Each time $\mathtt{SR}(i)$ performs a function evaluation in a point $X_i(t)$ (where $X_i(t)$ for $t \leq T_i(T)$ corresponds to the t-th function evaluation performed by $\mathtt{SR}(i)$) it receives $Y_i(t)$, which is passed to the Meta-Strategy. In both settings, the Meta-Strategy updates the quantity $\widehat{S}_T(i) = \sum_{t=1}^{T_i(t)} Y_i(t)$ each time $\mathtt{SR}(i)$ performs new function

Algorithm 1 Extra-information Meta-Strategy

```
Initialization
Input: $n$, $\delta$, $M(f)$ or $R^*(n,\delta)$ and SR
$\delta_0 = \frac{\delta}{\lfloor\log(n)\rfloor^2}$, $T = 0$
for $i = 1, ..., \lfloor\log(n)\rfloor^3$ do
    $\alpha_i = \frac{i}{\lfloor\log(n)\rfloor^3}$
    Initialize SR($i$) with $\delta_0$, $n$, $\alpha_i$
    $T_i(T) = 0$, $\widehat{S}_T(i) = 0$
end for
Case 1 ($M(f)$ known):
while $T < n$ do
    $k = \arg\min_i \left[T_i(T)M(f) - \widehat{S}_T(i)\ \right]$
    Perform $\sqrt{n}$ function evaluations with SR($k$)
    $T_k(T) = T_k(T) + \sqrt{n}$, $T = T + \sqrt{n}$
    $\widehat{S}_T(k) = \sum_{t=1}^{T_k(T)} Y_k(t)$
end while

Case 2 ($R^*$ known):
$\mathcal{A}_1 = \{1, ..., \lfloor\log(n)\rfloor^3\}$ (set of active SR($i$))
$T = |\mathcal{A}_1|\sqrt{n}$, $N = 1$ (round)
while $T < n$ do
    for $i \in \mathcal{A}_N$ do
        Perform $\sqrt{n}$ function evaluations with SR($i$)
        $T_i(T) = N\sqrt{n}$
        $\widehat{S}_T(i) = \sum_{t=1}^{T_i(T)} Y_i(t)$
    end for
    $k = \arg\max_{i\in\mathcal{A}_N} \widehat{S}_T(i)$
    $\mathcal{A}_{N+1} = \mathcal{A}_N$
    for $i \in \mathcal{A}_N$ do
        if $\widehat{S}_T(k) - \widehat{S}_T(i) > R^*(n,\delta) + \sqrt{T_i(t)}\log(n\lfloor\log(n)\rfloor^3/\delta)$ then
            Eliminate SR($i$), $\mathcal{A}_{N+1} = \mathcal{A}_{N+1} \setminus \{i\}$
        end if
    end for
    $N = N + 1$, $T = T + |\mathcal{A}_N|\sqrt{n}$
end while
Spend rest of the budget with SR($i$) for $i \in \mathcal{A}_N$
```

evaluations. We will also consider the empirical regret $\widehat{R}_T(i) = T_i(T)M(f) - \widehat{S}_T(i)$.

Case 1 ($M(f)$ known): The Meta-Strategy is called with parameter $M(f) = \max_{x\in\mathcal{X}} f(x)$. After the initialization, the Meta-Strategy operates in rounds of length $\sqrt{n}$. At the beginning of each round at time $T = u\sqrt{n}$ for some $u \in \{0, ..., \sqrt{n}\}$, the next batch of $\sqrt{n}$ function evaluations are allocated to the Subroutine which has accumulated the smallest empirical regret up to time T. More precisely, the index $k = \arg\min_i \widehat{R}_T(i)$ is chosen, and SR(k) resumes its learning where it was halted, performing $\sqrt{n}$ more function evaluations. The number of samples allocated to SR(k) and its empirical regret $\widehat{R}_T(k)$ are then updated. As the heuristic is to allocate new samples to the Subroutine that has currently incurred the smallest regret, this ensures that the regret incurred by each of the Subroutines grows at the same rate and is of the same order at time n. Therefore, we expect the Meta-Strategy to perform almost as well as the best Subroutine it has access to, up to a multiplicative factor that depends on the total number of Subroutines.

Case 2 (R^* known): Here, the Meta-Strategy is called with parameter $R^*(n,\delta)$. It proceeds in rounds and performs a *successive elimination* of the Subroutines. At round N, we call $\mathcal{A}_N$ the set of active Subroutines, with $\mathcal{A}_1 = \{1, ..., \lfloor\log(n)\rfloor^3\}$. The rate $R^*(n,\delta)$ is such that there exists $i^* \in \mathcal{A}_1$ for which for all $T \in \{\sqrt{n}, ..., n\}$ we have: $TM(f) - \sum_{t=1}^{T} f(X_{i^*}(t)) \leq R^*(n,\delta)$ with probability at least $1-\delta$. For any $i \in \mathcal{A}_N$, the Meta-Strategy allocates $\sqrt{n}$ function evaluations to be performed by SR(i), and the Meta-Strategy updates: $\widehat{S}_T(i) = \sum_{t=1}^{T_i(T)} Y_i(t)$. At the end of a round, the Meta-Strategy keeps computes the index $k = \arg\max_{i\in\mathcal{A}_T} \widehat{S}_T(i)$ of the best performing (active) Subroutine. Any active SR(i) that meets the following condition is *eliminated*:

$$\widehat{S}_T(k) - \widehat{S}_T(i) > R^*(n,\delta) + 2\sqrt{T_i(t)}\log(n\lfloor\log(n)\rfloor^3/\delta).$$

Heuristically, the Meta-Strategy uses $\texttt{SR}(k)$ as a pivot to eliminate the remaining active Subroutines, as the samples collected by $\texttt{SR}(k)$ cannot be too far $M(f)$, and this difference depends on $R^*(n, \delta)$. This extra-information allows the Meta-Strategy to eliminate Subroutines that perform poorly at the optimal rate. It is important to note that this cannot be done in the general setting, as this rate depends on both α and β, which are unknown to the learner.

4.2. Main Results for the Meta-Strategy

We now state our main *adaptive* results for these shape-constrained settings.

Theorem 4 *Fix $\alpha \in [0.5\sqrt{d/\log(n)}, \lfloor \log(n) \rfloor]$ and $\beta \geq 0$ such that $\alpha\beta \leq d$, with both parameters unknown to the learner. For any $f \in \mathcal{P}(\alpha, \beta)$ such that f takes value $M(f)$ at its maxima, the Meta-Strategy 1 run with budget n, confidence parameter $\delta = 1/\sqrt{n}$ and $M(f)$ is such that its regret is bounded as:*

$$\mathbb{E}(R_n) \leq C \log^p(n) n^{1-\alpha/(2\alpha+d-\alpha\beta)},$$

where the expectation is taken with respect to the samples, $C > 0$ and p do not depend on n.

This matches (up to log factors) the minimax optimal rate for the class of functions $f \in \mathcal{P}(\alpha, \beta)$ with $M(f)$ fixed that we proved in Corollary 1.

Theorem 5 *Fix α, β as in Theorem 4. For any $f \in \mathcal{P}(\alpha, \beta)$, the Meta-Strategy 1 run with budget n, confidence parameter δ and access to the parameter $R^*(n, \delta)$ is such that with probability at least $1 - 2\delta$, its pseudo-regret is bounded as:*

$$R_n \leq \lfloor \log(n) \rfloor^3 \left(2R^*(n, \delta) + 8\sqrt{n} \log\left(\frac{n \lfloor \log(n) \rfloor^3}{\delta} \right) + \sqrt{n} \right),$$

where the expectation is taken with respect to the samples.

By Lemma 3 in the Appendix, which bounds the best attainable rate attainable by the Subroutines run smoothness parameters α_i over a grid of step-size $\lfloor \log(n) \rfloor^2$, we know that there exists $\texttt{SR}(i^*)$ such that with probability at least $1 - \delta$, its pseudo-regret is such that $R_n(i^*) \leq C \log^p\left(\frac{n}{\delta}\right) n^{1-\alpha/(2\alpha+d-\alpha\beta)}$ with $p \leq 1$ and where $C > 0$ does not depend on n and δ. This naturally leads to the following Corollary:

Corollary 2 *Fix α, β as in Theorem 4. Let $r = \frac{\alpha+d-\alpha\beta}{2\alpha+d-\alpha\beta}$ be known to the learner, without direct access to α nor β. Then for any $f \in \mathcal{P}(\alpha, \beta)$, the Meta-Strategy 1 run with budget n, confidence parameter $\delta = n^{-1/2}$ and $R^*(n) = \log^2(n) n^r$ is such that for n large enough its expected pseudo-regret is upper-bounded as:*

$$\mathbb{E}[R_n] \leq \lfloor \log(n) \rfloor^3 \left(2\log^2(n) n^{1-\alpha/(2\alpha+d-\alpha\beta)} + 8\sqrt{n} \log\left(n^{3/2} \lfloor \log(n) \rfloor^3 \right) + \sqrt{n} \right),$$

where the expectation is taken with respect to the samples.

This matches the minimax optimal rate (up to log factors) for the cumulative regret that we proved in Corollary 1. In particular, if $\alpha\beta = d$, then our Meta-Strategy run with budget n, confidence parameter $\delta = n^{-1/2}$ and $R^*(n) = \log^2(n)\sqrt{n}$, is such that its expected pseudo-regret is of order $\tilde{\mathcal{O}}\left(\sqrt{n}\right)$. This extends the result of Bull et al. (2015) to our setting and interestingly, we also recover a result of Yu and Mannor (2011) (Theorem 4.2 and Assumption 3.2) and Combes and Proutiere (2014) (Proposition 1 and Assumption 2) in the one-dimensional unimodal continuum-armed bandit setting, but *without assuming unimodality*.

Acknowledgments

The authors would like to thank Samory Kpotufe for many useful discussions. This work is partly funded by the Deutsche Foschungsgemeinschaft (DFG, German Research Foundation) on the Emmy Noether grant for the project MuSyAD (CA 1488/1-1) 'anomaly detection in complex systems', on the GK 2297 MathCoRe on "Mathematical Complexity Reduction" – 314838170, GRK 2297 MathCoRe, and on the SFB 1294 Data Assimilation on "Data Assimilation — The seamless integration of data and models", Project A03.

References

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38, 2017.

R Agrawal. The continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, 33: 1926–1951, 1995.

Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *International Conference on Computational Learning Theory*, pages 454–468. Springer, 2007.

Pierre C Bellec. Adaptive confidence sets in shape restricted regression. *arXiv preprint arXiv:1601.05766*, 2016.

Lucien Birgé and Pascal Massart. From model selection to adaptive estimation. In *Festschrift for lucien le cam*, pages 55–87. Springer, 1997.

S. Bubeck, G. Stoltz, and J. Yu. Lipschitz bandits without the lipschitz constant. In *Algorithmic Learning Theory*, pages 144–158. Springer, 2011a.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011b.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvari. X-armed bandits. *Journal of Machine Learning Research*, 12(May):1655–1695, 2011c.

Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, pages 122–134, 2013.

Adam D Bull et al. Adaptive-treed bandits. *Bernoulli*, 21(4):2289–2307, 2015.

T Tony Cai, Mark G Low, et al. Adaptive confidence balls. *The Annals of Statistics*, 34(1):202–228, 2006.

T Tony Cai, Mark G Low, Yin Xia, et al. Adaptive confidence intervals for regression functions under shape constraints. *The Annals of Statistics*, 41(2):722–750, 2013.

Richard Combes and Alexandre Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *International Conference on Machine Learning*, pages 521–529, 2014.

Eric Cope. Regret and convergence bounds for immediate-reward reinforcement learning with continuous action spaces. *IEEE Transactions on Automatic Control*, 54(6):1243–1253, 2009.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *arXiv preprint arXiv:1602.07182*, 2016.

Georgii Ksenofontovich Golubev. Adaptive asymptotically minimax estimators of smooth signals. *Problemy Peredachi Informatsii*, 23(1):57–67, 1987.

Jean-Bastien Grill, Michal Valko, and Rémi Munos. Black-box optimization of noisy functions with unknown smoothness. In *Advances in Neural Information Processing Systems*, pages 667–675, 2015.

S. Hanneke. Adaptive rates of convergence in active learning. *Proceedings of the 22nd Annual Conference on Learning Theory, COLT 2009*, pages 353–364.

Marc Hoffmann and Richard Nickl. On adaptive inference and confidence bands. *The Annals of Statistics*, pages 2383–2409, 2011.

Anatoli Juditsky and Sophie Lambert-Lacroix. Nonparametric confidence set estimation. 2003.

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

Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, pages 697–704. MIT Press, 2004.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690. ACM, 2008.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *arXiv preprint arXiv:1312.1277*, 2013.

V. Koltchinskii. Rademacher complexities and bounding the excess risk of active learning. *Journal of Machine Learning Research*, 11:2457–2485, 2010.

Oleg V Lepski and VG Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, pages 2512–2546, 1997.

Andrea Locatelli, Alexandra Carpentier, and Samory Kpotufe. Adaptivity to noise parameters in nonparametric active learning. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1383–1416, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR. URL http://proceedings.mlr.press/v65/locatelli-andrea17a.html.

Stanislav Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13 (Jan):67–90, 2012.

Stanislav Minsker. Estimation of extreme values and associated level sets of a regression function via selective sampling. In *Conference on Learning Theory*, pages 105–121, 2013.

Rémi Munos. Optimistic Optimization of Deterministic Functions without the Knowledge of its Smoothness. In *Advances in Neural Information Processing Systems*, 2011.

Aleksandrs Slivkins. Multi-armed bandits on implicit metric spaces. In *Advances in Neural Information Processing Systems*, pages 1602–1610, 2011.

Alexandre B Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, pages 135–166, 2004.

Alexandre B Tsybakov. Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats, 2009.

Michal Valko, Alexandra Carpentier, and Rémi Munos. Stochastic simultaneous optimistic optimization. In *International Conference on Machine Learning*, pages 19–27, 2013.

Jia Yuan Yu and Shie Mannor. Unimodal bandits. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 41–48. Omnipress, 2011.

Appendix A. Proofs of Section 2

A.1. Proof of Proposition 1

Proof Consider x^* such that $f(x^*) = M(f)$ and the L_∞-ball of radius r centered in x^*, $r \in (0, 1]$. By smoothness of f around x^*, for any x such that $|x - x^*|_\infty \leq r$, we have:

$$|f(x) - M(f)| \leq \lambda r^\alpha,$$

which brings $\mu(\mathcal{X}(\lambda r^\alpha)) \geq r^d$. On the other hand, by Assumption 2, we have $\mu(\mathcal{X}(\lambda r^\alpha)) \leq B\lambda^\beta r^{\alpha\beta}$. Combining both conditions, we have for all $r \in (0, 1]$:

$$\frac{1}{B\lambda^\beta} \leq r^{\alpha\beta - d}.$$

As this has to hold true for all $r \in (0, 1]$, considering $r_l = 2^{-l}$ yields $\alpha\beta \leq d$.

■

A.2. Non-adaptive Subroutine

We first define a dyadic hierarchical partitioning of $[0, 1]^d$, on which our strategy bases its exploration of the space.

Algorithm 2 Non-adaptive Subroutine (SR)

```
Input: $n$, $\delta$, $\alpha$
Initialization: $t = 2^d t_{1,\alpha}$, $l = 1$, $\mathcal{A}_1 \doteq G_1$ (active space), $\forall l' > 1, \mathcal{A}_{l'} \doteq \emptyset$
while $t \leq n$ do
  $\widehat{M}_l = 0$
  for each active cell $C \in \mathcal{A}_l$ do
    Perform $t_{l,\alpha}$ function evaluations in $x_C$ the center of $C$
    $\widehat{f}(x_C) = \frac{1}{t_{l,\alpha}} \sum_{i=1}^{t_{l,\alpha}} Y_{C,i}$
    $\widehat{M}_l = \max(\widehat{M}_l, \widehat{f}(x_C))$
  end for
  for each active cell $C \in \mathcal{A}_l$ do
    if $\left\{\widehat{M}_l - \widehat{f}(x_C) \leq B_{l,\alpha}\right\}$ then
      $\mathcal{A}_{l+1} = \mathcal{A}_{l+1} \cup \{C' \in G_{l+1} \cap C\}$ // keep all children $C'$ of $C$ active
    end if
  end for
  Increase depth to $l = l + 1$, and set $t = t + |\mathcal{A}_l| \cdot t_{l,\alpha}$
end while
$L = l - 1$          // the final completed depth
Sample any $x \in \mathcal{A}_{L+1}$ until budget expires
Output: $\mathcal{A}_{L+1}$ // return active set after final depth $L$
```

Definition 4 *We write G_l for the regular dyadic grid on the unit cube of mesh size 2^{-l}. It defines naturally a partition of the unit cube in 2^{ld} smaller cubes, or cells $C \in G_l$ with volume 2^{-ld} and edge length 2^{-l}. We have $[0,1]^d = \bigcup_{C \in G_l} C$ and $C \cap C' = \emptyset$ if $C \neq C'$, with $C, C' \in G_l^2$. We define x_C as the center of $C \in G_l$, i.e. the barycenter of C.*
We write $r_l \doteq \max_{x,y \in C} |x - y|_\infty = 2^{-l}$ for the diameter of cells $C \in G_l$.

The Subroutine takes as input parameter α the smoothness parameters, n the maximum sampling budget, and δ a confidence parameter. In order to find the maxima of f, it refines a dyadic partition of the space, starting with 2^d hypercubes to sample from, and zooming in on regions that are close (in function value) to the optima. At depth l, the active cells in $\mathcal{A}_l$ are sampled $t_{l,\alpha} \doteq 0.5 \log(1/\delta_l) b_{l,\alpha}^{-2}$ times, where $b_{l,\alpha} \doteq r_l^\alpha$ and $\delta_l \doteq \delta 2^{-l(d+1)}$. After collecting $t_{l,\alpha}$ noisy evaluations $(Y_{C,i})_{i \leq t_{l,\alpha}}$, it computes a simple average to estimate $f(x_C)$:

$$\widehat{f}(x_C) = \frac{1}{t_{l,\alpha}} \sum_{i=1}^{t_{l,\alpha}} Y_{C,i}.$$

Once all the cells at depth l have been sampled, the Subroutine computes a current estimate of the maximum $\widehat{M}_l = \max_{C \in \mathcal{A}_l} \widehat{f}(x_C)$. Then, for each cell C in the active set $\mathcal{A}_l$, it compares $\widehat{M}_l - \widehat{f}(x_C)$ with $B_{l,\alpha} = 2\big(\sqrt{\frac{\log(1/\delta_l)}{2t_{l,\alpha}}} + b_{l,\alpha}\big)$, where we set $t_{l,\alpha}$ such that the variance term is of the same magnitude as the bias term $b_{l,\alpha}$. If $\widehat{M}_l - \widehat{f}(x_C) \geq B_{l,\alpha}$, this cell is *eliminated*, as the Subroutine rules it unlikely that there exists $x \in C$ such that $f(x) = M(f)$. On the other hand, if $\widehat{M}_l - \widehat{f}(x_C)$ is smaller than $B_{l,\alpha}$, then C is kept active, and all its children $\{C' : C \cap G_{l+1}\}$ are added to $\mathcal{A}_{l+1}$. This process is repeated until the budget is not sufficient to sample all the cells that are still active at depth $L+1$, and the Subroutine returns $\mathcal{A}_{L+1}$ the last active set, and the recommendation $x(n)$ can be any point chosen in $\mathcal{A}_{L+1}$.

A.3. Proof of Proposition 2

Let us write in this proof in order to simplify the notations

$$t_l = t_{l,\alpha}, \quad b_l = b_{l,\alpha}, \quad B_l = B_{l,\alpha} \quad \text{and} \quad N_l = |\mathcal{A}_l|.$$

Step 1: A favorable event.
Consider a cell C of depth l. We define the event:

$$\xi_{C,l} = \Big\{ |t_l^{-1} \sum_{i=1}^{t_l} Y_{C,i} - f(x_C)| \leq \sqrt{\frac{\log(1/\delta_l)}{2t_l}} \Big\},$$

where the $(Y_{C,i})_{i \leq t_l}$ are samples collected in C at point x_C if C if the algorithm samples in cell C. We remind that

$$\widehat{f}(x_C) = \frac{1}{t_l} \sum_{i=1}^{t_l} Y_{C,i}.$$

As $Y_{C,i} = f(x_C) + \epsilon_i$ where $\{\epsilon_i\}_{i \leq n}$ are zero-mean 1-sub-Gaussian independent random variables, we know from Hoeffding's concentration inequality that $\mathbb{P}(\xi_{C,l}) \geq 1 - 2\delta_l$.

We now consider

$$\xi = \Big\{ \bigcap_{l \in \mathbb{N}^*, C \in G_l} \xi_{C,l} \Big\},$$

the intersection of events such that for all depths l and any cell $C \in G_l$, the previous event holds true. Note that at depth l there are 2^{ld} such events. A simple union bound yields $\mathbb{P}(\xi) \geq 1 - \sum_l 2^{ld}\delta_l \geq 1 - 4\delta$ as we have set $\delta_l = \delta 2^{-l(d+1)}$.

On the event ξ, for any $l \in \mathbb{N}^*$, as we have set $t_l = \frac{\log(1/\delta_l)}{2b_l^2}$, plugging this in the bound implies that for each cell $C \in G_l$ that has been sampled t_l times we have:

$$|\widehat{f}(x_C) - f(x_C)| \leq b_l. \tag{1}$$

Note that by Assumption 1, b_l is such that for any $x \in C$, where $C \in G_l$, we have:

$$|f(x) - f(x_C)| \leq \max\{M(f) - f(x_C), b_l\}. \tag{2}$$

Step 2: No mistakes.
For $l \in \mathbb{N}^*$, let us consider $C \in G_l$ such that $\exists x^* \in C$, $x^* \in \mathcal{X}(0)$ i.e. $f(x^*) = M(f)$. Let us assume that $C \in \mathcal{A}_l$. Then on ξ:

$$\begin{aligned} \widehat{M_l} \geq \widehat{f}(x_C) &\geq f(x_C) - b_l \\ &\geq f(x^*) - 2b_l \\ &\geq M(f) - 2b_l \end{aligned} \tag{3}$$

Moreover, we have:

$$\widehat{M_l} \leq M(f) + b_l \tag{4}$$

Equation (4) yields:

$$\begin{aligned}\widehat{M}_l - \widehat{f}(x_C) &\leq M(f) + b_l - (M(f) - 2b_l) \\ &\leq 3b_l < 4b_l = B_l\end{aligned}$$

This shows that on ξ any cell $C \in \mathcal{A}_l$ that contains a global optimum x^* is never eliminated by the algorithm at depth l, and all its children are added to $\mathcal{A}_{l+1}$. As at depth $l = 1$, all cells are active, by induction we have $\forall l \geq 1$:

$$\{\mathcal{X}(0) \cap G_l\} \subset \mathcal{A}_l \tag{5}$$

Step 3: A maximum gap.

Now consider an active cell at depth l: $C \in \mathcal{A}_l$ such that all its children are added to $\mathcal{A}_{l+1}$. If this cell is kept active at depth $l+1$, then it is such that:

$$\widehat{M}_l - \widehat{f}(x_C) \leq B_l = 4b_l.$$

By Equations (3) and (1), we know that on ξ:

$$\widehat{M}_l - \widehat{f}(x_C) \geq M(f) - 2b_l - (f(x_C) + b_l),$$

which brings that all cells kept active are such that:

$$M(f) - f(x_C) \leq 7b_l$$

By Equation (2), we know that $\forall x \in C : f(x_C) - f(x) \leq \max\{M(f) - f(x), b_l\} \leq 7b_l$, where we upper bound using the previous equation. This rewrites:

$$M(f) - f(x) \leq 7b_l + M(f) - f(x_C),$$

which implies that for any x in C kept active at depth $l+1$:

$$M(f) - f(x) \leq 14b_l, \tag{6}$$

which implies:

$$\mathcal{A}_{l+1} \subset \mathcal{X}(14b_l) \tag{7}$$

Step 4: A bounded number of active cells.

By Assumption 2, we know that $\mu(\mathcal{X}(14b_l)) \leq B14^\beta b_l^\beta$. As each cell of depth l has an L_∞-volume of r_l^d, this allows us to bound the number of remaining active cells N_{l+1} on ξ for $l \geq 1$:

$$\begin{aligned}N_{l+1} &\leq B14^\beta b_l^\beta r_{l+1}^{-d} \\ &\leq 2^{\alpha\beta} B(14)^\beta r_{l+1}^{\alpha\beta - d}\end{aligned} \tag{8}$$

Define $B' = \max(1, B)(14)^\beta$, then $N_l \leq 2^d B' r_l^{\alpha\beta - d}$ for all $l \geq 1$.

Step 5: A minimum depth.

We first bound L the maximal depth by above naively. Notice that t_L itself has to be smaller than n, otherwise the budget is insufficient to sample a single active times t_L times, and the algorithm stops. This yields $L \leq \frac{1}{2\alpha} \log_2(2n)$, which brings the following bound:

$$\log(1/\delta_L) = \log(2^{L(d+1)}/\delta) \leq \frac{d+1}{2\alpha} \log(\frac{2n}{\delta}) \tag{9}$$

As we sample each active cell at depth l a number $t_l = \frac{\log(1/\delta_l)}{2b_l^2}$ times, we can now upper bound the total number of samples that the algorithm needs to reach depth L:

$$\begin{aligned}
\sum_{l=1}^{L} t_l N_l &\leq 2^d B' \sum_{l=1}^{L} \frac{\log(1/\delta_l)}{2r_l^{2\alpha}} r_l^{\alpha\beta-d} \\
&\leq \frac{1}{2} 2^d B' \log(1/\delta_L) \sum_{l=1}^{L} r_l^{\alpha\beta-d-2\alpha} \\
&\leq \frac{1}{2} 2^d B' \log(1/\delta_L) \sum_{l=1}^{L} 2^{l(2\alpha+d-\alpha\beta)} \\
&\leq \frac{1}{2} 2^d B' \log(1/\delta_L) \frac{2^{L(2\alpha+d-\alpha\beta)}}{2^{2\alpha+d-\alpha\beta}-1} \\
&\leq 2^d B' \log(1/\delta_L) \frac{2^{L(2\alpha+d-\alpha\beta)}}{2\alpha+d-\alpha\beta},
\end{aligned}$$

where we use $2^c - 1 \geq c/2$ for any $c \in \mathbb{R}^+$in the last line. Combined with Equation (9), this yields:

$$\sum_{l=1}^{L} t_l N_l \leq 2^d B'(d+1) \log\left(\frac{2n}{\delta}\right) \frac{2^{L(2\alpha+d-\alpha\beta)}}{2\alpha(2\alpha+d-\alpha\beta)}. \tag{10}$$

This implies that for any $T \leq n$, after T function evaluations, the following depth $L(T)$ is reached:

$$L(T) \geq \frac{1}{2\alpha+d-\alpha\beta} \log_2\Big(\frac{2\alpha(2\alpha+d-\alpha\beta)T}{D\log(\frac{2n}{\delta})}\Big), \tag{11}$$

where $D = 2^d B'(d+1)$

Step 6: Conclusion.

Using Equation (11) with $T = n$, we can now ready to bound the simple regret r_n with high probability, as we have on ξ by Equation (7)

$$\mathcal{A}_{L+1} \subset \mathcal{X}(8b_L) \tag{12}$$

with

$$b_L \leq \Big(\frac{2\alpha(2\alpha+d-\alpha\beta)n}{D\log(\frac{2n}{\delta})}\Big)^{-\frac{\alpha}{2\alpha+d-\alpha\beta}}.$$

This shows that by recommending any $x(n) \in \mathcal{A}_{L+1}$, we have: $M(f) - f(x(n)) \leq 8b_L$.

Step 7: Bound on the cumulative regret.

We can now bound with high-probability the *pseudo-regret* up to time $T \leq n$: $R_T = TM(f) -$

$\sum_{t=1}^T f(X_t)$. Define $\Delta_l = 8b_{l-1}$, and recall that $\forall x \in C$ such that $C \in \mathcal{A}_l$, we have $M(f) - f(x) \leq 8b_{l-1}$. We can naively bound the regret by splitting the regret before the reaching depth $L(T)$ and beyond this depth:

$$\begin{aligned}
R_T &= TM(f) - \sum_{t=1}^{T} f(X_t) \\
&\leq 2^d(M(f)-m(f))t_1 + \sum_{l=2}^{L(T)} t_l N_l \Delta_l + T\Delta_{L(T)} \\
&\leq A + 2^d B' 28 \log(1/\delta_{L(T)}) \sum_{l=1}^{L(T)} 2^{l(\alpha+d-\alpha\beta)} + T\Delta_{L(T)} \\
&\leq A + 2^d B' 28 \log(1/\delta_{L(T)}) \frac{2^{L(T)(\alpha+d-\alpha\beta)}}{\alpha+d-\alpha\beta} + 8T\Big(\frac{D\log(\frac{2n}{\delta})}{2\alpha(2\alpha+d-\alpha\beta)T}\Big)^{\frac{\alpha}{2\alpha+d-\alpha\beta}} \\
&\leq A + 2^d B' 28 \frac{(d+1)}{2\alpha(\alpha+d-\alpha\beta)} \log(\frac{2n}{\delta})\Big(\frac{2\alpha(2\alpha+d-\alpha\beta)T}{D\log(\frac{2n}{\delta})}\Big)^{\frac{\alpha+d-\alpha\beta}{2\alpha+d-\alpha\beta}} \\
&\qquad +14\Big(\frac{D\log(\frac{n}{\delta})}{2\alpha(2\alpha+d-\alpha\beta)}\Big)^{\frac{\alpha}{2\alpha+d-\alpha\beta}} T^{\frac{\alpha+d-\alpha\beta}{2\alpha+d-\alpha\beta}} \\
&\leq A + 2^d B' 14(d+1) D \left(\frac{\log(\frac{2n}{\delta})}{2\alpha(\alpha+d-\alpha\beta)}\right)^{\frac{\alpha}{2\alpha+d-\alpha\beta}} T^{\frac{\alpha+d-\alpha\beta}{2\alpha+d-\alpha\beta}},
\end{aligned}$$

with $A \leq (M(f)-m(f))(d+1)2^{2\alpha+d}\log(2/\delta)$ and $m(f) = \inf_x f(x)$. Importantly, this holds on ξ for all $T \leq n$.

Setting $T = n$, we can also get a bound in expectation:

$$\mathbb{E}(R_n) \leq A + 2^d B' 14(d+1) D \left(\frac{\log(\frac{2n}{\delta})}{2\alpha(\alpha+d-\alpha\beta)}\right)^{\frac{\alpha}{2\alpha+d-\alpha\beta}} n^{\frac{\alpha+d-\alpha\beta}{2\alpha+d-\alpha\beta}} + 4(M(f)-m(f))n\delta,$$

and setting $\delta = 1/\sqrt{n}$ yields the result. As we assumed that f takes values in $[0,1]$, we can upper bound $M(f) - m(f) \leq 1$.

Appendix B. Proofs of Section 3

B.1. Proof of Theorem 1

Proof Let $\alpha > 0$, $\beta \geq 0$ such that $\alpha\beta < d$. The case $\alpha\beta = d$ corresponds to the usual $\mathcal{O}\left(n^{-1/2}\right)$ bound, which can easily be obtained using classical techniques with two hypothesis. Define $K = \lceil \Delta^{\frac{\alpha\beta-d}{\alpha}} \rceil$, and Δ such that:

$$\Delta = \sqrt{\frac{K}{n}},$$

with n large enough such that $K \geq \frac{16\exp(2)}{3}$. One can easily check that we have $\Delta = \mathcal{O}\left(n^{-\frac{\alpha}{2\alpha+d-\alpha\beta}}\right)$ and $K = \mathcal{O}\left(n^{\frac{d-\alpha\beta}{2\alpha+d-\alpha\beta}}\right)$ which grows with n.

Consider the grid G which partitions $[0,1]^d$ into $N = \lceil \Delta^{-d/\alpha} \rceil$ disjoint hypercubes, and let us index the cells arbitrarily (for example using Cantor's pairing argument in d dimensions). In what follows, we will write

$$\mathcal{S} = \bigcup_{k \leq K} H_k.$$

Fix $M \in [1/2, 1]$. We define the function $\phi_s(x)$ for $0 \leq s \leq K$ and $x \in [0,1]^d$.

$$\phi_s(x) = \begin{cases} \max\{M - \Delta, M - |x - x_i|_\infty^\alpha\}, & \text{if} \quad x \in H_i, i = s, \\ M - \Delta, & \text{if} \quad x \in H_i, i \neq s \\ \max\{0, M - \Delta - \text{dist}_\infty(x, \mathcal{S})^\alpha\}, & \text{if} \quad x \in \mathcal{S}^C, \end{cases}$$

where $\text{dist}_\infty(x, \mathcal{S}) \doteq \inf\{|x - z|_\infty, z \in \mathcal{S}\}$. It is clear that for any $s \in \{0, ..., K\}$, $\phi_s \in \Sigma(1, \alpha)$. We will now show that Assumption 2 for some $B > 0$ is satisfied for ϕ_s, $\forall s \in \{0, ..., K\}$. For any $0 < \epsilon < \Delta < 1$ and any ϕ_s, we have:

$$\mu(\mathcal{X}(\epsilon)) \leq \epsilon^{d/\alpha} \leq \epsilon^\beta,$$

as we have $\alpha\beta \leq d$. Now considering $\epsilon = \Delta$:

$$\mu(\mathcal{X}(\epsilon)) \leq K\Delta^{d/\alpha} \leq 2\epsilon^\beta,$$

as we have set $K = \lceil \Delta^{(\alpha\beta - d)/\alpha} \rceil \leq 2\Delta^{(\alpha\beta - d)/\alpha}$. Finally, we consider $\epsilon \in]\Delta, 1/2]$, and we have:

$$\begin{aligned} \mu(\mathcal{X}(\epsilon)) &\leq \mu(\mathcal{X}(\Delta)) + \mu(\{x : \Delta < M - \phi_s(x) \leq \epsilon\}) \\ &\leq 2\Delta^\beta + \epsilon^{d/\alpha} \\ &\leq 3\epsilon^\beta. \end{aligned}$$

So we have by construction :

- For any $s \leq K$, $\phi_s \in \mathcal{P}(\alpha, \beta)$ with $\lambda = 1$ as the constant in Assumption 1.
- for any $s, t \leq K$, and any $x \in \mathcal{S}^C$, $\phi_s(x) = \phi_t(x)$ (one cannot distinguish problem i from problem j in $\mathcal{S}^C$)
- for any $s \in \{1, ..., K\}$, the maximum of ϕ_s is attained only in x_s with value $\phi_s(x_s) = M$. This shows that the value at the maximum for ϕ_s for $s \in \{1, ..., K\}$ is fixed and known to the learner.
- $\forall x \notin H_s$, $\phi_s(x) = \phi_0(x)$: one cannot distinguish problem s from problem 0 outside of a small neighborhood around x_s.
- For any $1 \leq s \leq K, \forall x \notin H_s, M - \phi_s(x) \geq \Delta$

We now define $\mathcal{H}_K$ the set of recommendation problems such that for any $s \in \{0, .., K\}$, the problem s is characterized by the mean-pay off function ϕ_s, with zero-mean Gaussian noise of variance 1, such that the observations are, conditionally on $X_t = x$, i.i.d. with distribution $Y_t \sim \mathcal{N}(\phi_s(x), 1)$. Let us fix a strategy (algorithm) with two components: a (possibly randomized) *sampling* mechanism, which characterizes the next sampling point X_t based on the previous observations $\{(X_i, Y_i)\}_{i<t}$,

and a (possibly randomized) *recommendation* $x(n)$ based on all the collected samples $\{(X_i, Y_i)\}_{i\leq n}$, which the algorithm outputs at the end of the game incurring the simple regret $M(\phi_s) - \phi_s(x(n))$. We write $\mathbb{P}_s$, $\mathbb{E}_s$, for the probability and expectation under the problem s (uniquely characterized by the function ϕ_s), when the previously mentioned strategy is used.

For a sample $\{(X_i, Y_i)\}_{i\leq n}$ collected under problem 0 by the previously introduced algorithm, we consider the log-likelihood ratio $L_{n,s} \doteq L_{n,s}(\{(X_i, Y_i)\}_{i\leq n})$ for $s \in \{1, ..., K\}$:

$$\begin{aligned}
L_{n,s} &= \sum_{t=1}^{n} \log\left(\frac{\mathbb{P}_0(Y_t|X_t)}{\mathbb{P}_s(Y_t|X_t)}\right) = \sum_{t=1}^{n} \frac{1}{2}\left((Y_t - \phi_s(X_t))^2 - (Y_t - \phi_0(X_t))^2\right) \\
&= \sum_{t=1}^{n} \frac{1}{2}(\phi_0(X_t) - \phi_s(X_t))(2Y_t - \phi_0(X_t) - \phi_s(X_t)) \\
&= \sum_{t=1}^{n} \frac{1}{2}(\phi_s(X_t) - \phi_0(X_t))(\phi_s(X_t) + \phi_0(X_t) - 2Y_t) \\
&\leq \sum_{t=1}^{n} \frac{1}{2}(\phi_s(X_t) - \phi_0(X_t))(2\phi_s(X_t) - 2Y_t) \\
&\leq \sum_{t=1}^{n}(\phi_s(X_t) - \phi_0(X_t))(\phi_s(X_t) - Y_t), \qquad (13)
\end{aligned}$$

where we use: $0 \leq \phi_s(x) - \phi_0(x) \leq \Delta$ for all $x \in H_s$ in the fourth line.
We now consider $\mathbb{E}_0(L_{n,s})$:

$$\begin{aligned}
\mathbb{E}_0(L_{n,s}) &\leq \sum_{t=1}^{n} \mathbb{E}_0\left((\phi_s(X_t) - \phi_0(X_t))(\phi_s(X_t) - Y_t)\right) \\
&\leq \sum_{t=1}^{n} \mathbb{E}_0\Big(\mathbb{E}_0\big((\phi_s(X_t) - \phi_0(X_t))(\phi_s(X_t) - Y_t)\big|X_t\big)\Big) \\
&\leq \sum_{t=1}^{n} \mathbb{E}_0\Big((\phi_s(X_t) - \phi_0(X_t))(\phi_s(X_t) - \mathbb{E}_0\big(Y_t\big|X_t\big)\Big) \\
&\leq \sum_{t=1}^{n} \mathbb{E}_0\left((\phi_s(X_t) - \phi_0(X_t))^2\right) \\
&\leq \sum_{t=1}^{n} \mathbb{E}_0\left((\phi_s(X_t) - \phi_0(X_t))^2\big|X_t \in H_s\right)\mathbb{P}_0(X_t \in H_s) \\
&\leq \max_{x\in H_s}(\phi_s(x) - \phi_0(x))^2 \sum_{t=1}^{n} \mathbb{P}_0(X_t \in H_s) \\
&\leq \Delta^2 \sum_{t=1}^{n} \mathbb{P}_0(X_t \in H_s) \\
&\leq \Delta^2 \mathbb{E}_0(T_s(n))
\end{aligned}$$

where we use the fact that the function evaluations Y_t are independent and identically distributed as $\mathcal{N}(\phi_0(X_t), 1)$ conditionally on X_t, and we denote $\mathbb{E}_0(T_s(n)) = \sum_{t=1}^n \mathbb{P}_0(X_t \in H_s)$ the expected number of samples collected in H_s by the strategy under problem 0.

We now state the two main technical lemmas we will use.

Lemma 1 *For any event $\mathcal{E} \in \mathcal{F}_n = \sigma(X_1, Y_1, ..., X_n, Y_n)$ we have:*

$$\mathbb{E}_0(L_{n,s}|\, \mathcal{E}) \geq \log\left(\frac{\mathbb{P}_0(\mathcal{E})}{\mathbb{P}_s(\mathcal{E})}\right).$$

Proof Use the change of measure identity and conditional Jensen's inequality (see Kaufmann et al. (2016), proof of Lemma 19). ■

Lemma 2 *Let ρ_0, ρ_1 be two probability distributions supported on some set $\mathcal{X}$, with ρ_1 absolutely continuous with respect to ρ_0. Then for any measurable function $\tau : \mathcal{X} \to \{0, 1\}$, one has:*

$$\mathbb{P}_{X\sim\rho_0}(\tau(X) = 1) + \mathbb{P}_{X\sim\rho_1}(\tau(X) = 0) \geq \frac{1}{2}\exp\big(-\mathrm{KL}(\rho_0, \rho_1)\big).$$

The proof can be found in Tsybakov (2009) (Chapter 2, Theorem 2.2, Conclusion (iii)).

We now consider a realization of both the samples $\{(X_i, Y_i)\}_{i\leq n}$ and the recommendation $x(n)$ output by the strategy. We write $g(x(n)) = \arg\min_{k\leq K} |x(n) - x_k|_\infty$, which simply maps the recommendation $x(n)$ to the closest x_k (which correspond to the K possible maxima for our set of problems) in infinity norm. We define ρ_0, ρ_s as the distribution of $g(x(n))$x (here $\mathcal{X}$ in Lemma 2 corresponds to $\{1, ..., K\}$) under problems 0 and s respectively. By definition of the fixed budget setting, we have $\sum_{k=1}^K \mathbb{E}_0(T_s(n)) \leq n$, so for $K \geq 2$, there exists at least $K/2$ indices $s \in \{1, ..., K\}$ such that $\mathbb{E}_0(T_s(n)) \leq \frac{2n}{K}$. Moreover, there also exists $0.75K$ indices $s \in \{1, ..., K\}$ such that $\mathbb{P}_0(g(x(n)) = s) \leq \frac{4}{3K}$. The intersection of these two sets of indices cannot be empty, and we fix i as one element of this intersection. Finally, we define the test function $\tau : k \to \mathbf{1}\{k = i\}$. Under this choice of ρ_0, ρ_1 and τ, the previous lemma rewrites to:

$$\mathbb{P}_0(g(x(n)) = i) + \mathbb{P}_i(g(x(n)) \neq i) \geq \frac{1}{2}\exp\big(-\mathrm{KL}(\rho_0, \rho_i)\big).$$

We now use the tower rule (its countable - finite - version) and Lemma 1:

$$\begin{aligned}\mathbb{E}_0(L_{n,i}) &= \sum_{k=1}^K \mathbb{E}_0(L_{n,i}|g(x(n)) = k)\mathbb{P}_0(g(x(n) = k) \\ &\geq \sum_{k=1}^K \log\left(\frac{\mathbb{P}_0(g(x(n) = k)}{\mathbb{P}_i(g(x(n) = k)}\right)\mathbb{P}_0(g(x(n) = k),\end{aligned}$$

and we remark that the quantity on right hand side of the last inequality is precisely $\mathrm{KL}(\rho_0, \rho_i)$ for our choice of ρ_0, ρ_i. Combining this with our previous bound in Equation (13): $\mathbb{E}_0(L_{n,i}) \leq \mathbb{E}(T_i(n))\Delta^2 \leq \frac{2n}{K}\Delta^2$, with $\Delta = \sqrt{\frac{K}{n}}$, we get:

$$\mathbb{P}_i(g(x(n)) \neq i) \geq \frac{1}{2}\exp(-2) - \frac{4}{3K}.$$

with $K \geq \frac{16\exp(2)}{3}$, this yields:

$$\max_{s\in\{1,\ldots,K\}} \mathbb{P}_s(g(x(n)) \neq i) \geq \frac{1}{4}\exp(-2).$$

Thus, with constant probability, it holds that $g(x(n)) \neq i$, and by definition of $g(x(n))$ we have $x(n) \notin H_i$. The simple regret associated with recommending $x(n)$ can then be bounded by using the definition of ϕ_i:

$$M - \phi_i(x(n)) \geq \Delta.$$

In the corresponding passive setting where the sampled locations X_t are independent, identically distributed uniformly at random over $[0,1]^d$, we have instead for all s: $\mathbb{E}(T_s(n)) \leq \mathcal{O}\left(n\Delta^{d/\alpha}\right)$ and setting instead $\Delta = \mathcal{O}\left(n^{-\alpha/(2\alpha+d)}\right)$ we get the rate $\mathcal{O}\left(n^{-\alpha/(2\alpha+d)}\right)$. Here, β plays no role in the rate, which shows that sampling actively is very beneficial as soon as $\beta > 0$.

■

B.2. Proof of Theorem 3

Proof Let $\gamma > \alpha > 0$ the two smoothness parameters and $\beta \geq 0$ such that $\gamma\beta \leq d$. Define $K = \lceil \Delta^{\frac{\alpha\beta-d}{\alpha}} \rceil \geq 2$, and Δ such that:

$$\Delta = \frac{K}{R_{\gamma,\beta}(n)},$$

with $R_{\gamma,\beta}(n)$ such that $R_{\gamma,\beta}(n)^{(2\alpha+d-\alpha\beta)/(\alpha+d-\alpha\beta)} \leq \frac{n}{16}\exp(-2)$. Importantly, we will consider strategies such that for any problem in $\mathcal{P}(\gamma,\beta)$, their expected regret is smaller than $R_{\gamma,\beta}(n)$. Consider the grid G which partitions $[0,1/2]^d$ into $N = \lceil \Delta^{-d/\alpha} \rceil$ disjoint hypercubes. We index the cells of G as $(H_k)_{k\leq N}$ as in the proof of Theorem 1. We also define H_0 the hypercube $[1-\Delta^{1/\gamma}, 1] \times \ldots \times [1-\Delta^{1/\gamma}, 1]$.

In what follows, we will write

$$\mathcal{S} = \bigcup_{0\leq k\leq K} H_k.$$

Fix $M \in [1/2, 1]$. We define the function $\phi_s(x)$ for $0 \leq s \leq K$ and $x \in [0,1]^d$.

$$\phi_s(x) = \begin{cases} \max\{M-\Delta, M-\Delta/2-|x-x_i|_\infty^\gamma\}, & \text{if} \quad x \in H_0 \\ \max\{M-\Delta, M-|x-x_i|_\infty^\alpha\}, & \text{if} \quad x \in H_i, i = s \\ M-\Delta, & \text{if} \quad x \in H_i, i \neq s \\ \max\{0, M-\Delta-\text{dist}_\infty(x,\mathcal{S})^\gamma\}, & \text{if} \quad x \in \mathcal{S}^C, \end{cases}$$

where $\text{dist}_\infty(x,\mathcal{S}) \doteq \inf\{|x-z|_\infty, z \in \mathcal{S}\}$. It is clear that for $s = 0$, we have $\phi_0 \in \Sigma(1,\gamma)$. By the nestedness of the smoothness classes for any $1 \leq s \leq K$ we have $\phi_s \in \Sigma(1,\alpha)$ as $\alpha \leq \gamma$.
We will now show that Assumption 2 for some $B > 0$ is satisfied for ϕ_s, $\forall s \leq K$. For any $0 < \epsilon < \Delta < 1$, we have:

$$\mu(\mathcal{X}(\epsilon)) \leq \epsilon^{d/\gamma} \leq \epsilon^\beta,$$

as we have $\gamma\beta \leq d$. Now considering $\epsilon = \Delta$:

$$\mu(\mathcal{X}(\epsilon)) \leq K\Delta^{d/\alpha} + \Delta^{d/\gamma} \leq 2\Delta^{\beta},$$

as we have set $K = \lceil \Delta^{(\alpha\beta - d)/\alpha} \rceil \leq 2\Delta^{(\alpha\beta-d)/\alpha}$. Finally, we consider $\epsilon \in]\Delta, 1/2]$, and we have:

$$\begin{aligned} \mu(\Omega(\epsilon)) &\leq \mu(\mathcal{X}(\Delta)) + \mu(\{x : \Delta < M - \phi_s(x) \leq \epsilon\}) \\ &\leq 2\Delta^{\beta} + \epsilon^{d/\gamma} \\ &\leq 3\epsilon^{\beta}. \end{aligned}$$

So we have by construction :

- For $s = 0$, $\phi_0 \in \mathcal{P}(\gamma, \beta)$ and $M(\phi_0) = M - \Delta/2$
- For any $1 \leq s \leq K$, $\phi_s \in \mathcal{P}(\alpha, \beta)$.
- for any $s, t \leq K$, and any $x \in \mathcal{A}^C$, $\phi_s(x) = \phi_t(x)$ (one cannot distinguish problem i from problem j in $\mathcal{S}^C$)
- for any $1 \leq s \leq K$, the maximum of ϕ_s is attained only in x_s and we have $\phi_s(x_s) = M$. In particular, for any $s \neq 1$, we have $M(\phi_s) = M$.
- $\forall x \notin H_s$, $\phi_s(x) = \phi_0(x)$: one cannot distinguish problem s from problem 0 outside of a small neighborhood around x_s.
- For any $s \leq K, \forall x \notin H_s, M_s - \phi_s(x) \geq \Delta/2$

We now define $\mathcal{H}_K$ the set of recommendation problems such that for any $0 \leq s \leq K$, the problem s is characterized by the mean-pay off function ϕ_s, with zero-mean Gaussian noise of variance 1, such that the observations are, conditionally on $X_t = x$, i.i.d. with distribution $Y_t \sim \mathcal{N}(\phi_s(x), 1)$. Let us fix a strategy (algorithm): it defines a (possibly randomized) *sampling* mechanism, which characterizes the next sampling point X_t based on the previous observations $\{(X_i, Y_i)\}_{i<t}$, for all $t \leq n$. We write $\mathbb{P}_s$, $\mathbb{E}_s$, for the probability and expectation under the problem s (uniquely characterized by the function ϕ_s), when the previously mentioned strategy is used. This strategy is such that for any problem in $\mathcal{P}(\gamma, \beta)$, we have $\mathbb{E}[R_n] \leq R_{\gamma,\beta}(n)$. This assumption will be used to encode the fact the strategy is nearly minimax optimal over the class $\mathcal{P}(\gamma, \beta)$, and that any such strategy is strictly suboptimal over the larger class $\mathcal{P}(\alpha, \beta)$.

As in the proof of Theorem 1, for a sample $\{(X_i, Y_i)\}_{i\leq n}$ collected by the previously introduced algorithm under problem 1, we consider the log-likelihood ratio $L_{n,s} \doteq L_{n,s}(\{(X_i, Y_i)\}_{i\leq n})$ for $1 < s \leq K$:

$$\begin{aligned} L_{n,s} &= \sum_{t=1}^{n} \log\left(\frac{\mathbb{P}_0(Y_t|X_t)}{\mathbb{P}_s(Y_t|X_t)}\right) = \sum_{t=1}^{n} \frac{1}{2}\left((Y_t - \phi_s(X_t))^2 - (Y_t - \phi_0(X_t))^2\right) \\ &= \sum_{t=1}^{n} (\phi_s(X_t) - \phi_0(X_t))(\phi_s(X_t) - Y_t); \end{aligned}$$

which yields as in the proof of Theorem 1:

$$\mathbb{E}_0(L_{n,s}) \leq \mathbb{E}_0(T_s(n))\Delta^2, \tag{14}$$

where $\mathbb{E}_0(T_s(n))$ is the expected number of samples in cell H_s collected by the sampling strategy under problem 0 at the end of the game.
By definition of $R_{\gamma,\beta}(n)$ which bounds the expected regret of the strategy, there exists a cell H_m and an index m such that:

$$\mathbb{E}_0(T_m(n)) \leq \frac{2R_{\gamma,\beta}(n)}{\Delta K},$$

otherwise the strategy has an expected regret strictly greater than $R_{\gamma,\beta}(n)$. Combined with Equation (14), this yields:

$$\mathbb{E}_0(L_{n,m}) \leq \frac{2R_{\gamma,\beta}(n)\Delta}{K} = 2,$$

by definition of $\Delta = \frac{K}{R_{\gamma,\beta}(n)}$.

Consider a realization of the samples $\{(X_i, Y_i)\}_{i\leq n}$. We define ρ_0, ρ_m as the distribution of $T_m(n)$ (here $\mathcal{X}$ in Lemma 2 corresponds to $\{0, ..., n\}$) under problems 0 and m respectively. Finally, we define the test function $\tau : T \rightarrow \mathbf{1}\{T \geq n/2\}$. Under this choice of ρ_0, ρ_m and τ, Lemma 2 yields:

$$\mathbb{P}_0(T_m(n) \geq n/2) + \mathbb{P}_m(T_m(n) < n/2) \geq \frac{1}{2}\exp\big(-\text{KL}(\rho_0, \rho_m)\big).$$

By the tower rule and Lemma 1:

$$\begin{aligned}\mathbb{E}_0(L_{n,s}) &= \sum_{k=0}^{n} \mathbb{E}_0(L_{n,s}|T_m(n) = k)\mathbb{P}_0(T_m(n) = k)\\ &\geq \sum_{k=0}^{n} \log\left(\frac{\mathbb{P}_0(T_m(n) = k)}{\mathbb{P}_s(T_m(n) = k)}\right)\mathbb{P}_0(T_m(n) = k),\end{aligned}$$

which is precisely $\text{KL}(\rho_0, \rho_m)$ for our choice of ρ_0, ρ_m. As $\mathbb{E}_0(L_{n,s}) \leq 2$, we get:

$$\mathbb{P}_0(T_m(n) \geq n/2) + \mathbb{P}_m(T_m(n) < n/2) \geq \frac{1}{2}\exp(-2). \tag{15}$$

We now remark that $\mathbb{P}_0(T_m(n) \geq n/2) \leq \mathbb{P}_0(R_n \geq \frac{n\Delta}{4})$, which can be bounded by Markov's inequality:

$$\begin{aligned}\mathbb{P}_0(R_n \geq \frac{n\Delta}{4}) &\leq \frac{4R_{\gamma,\beta}(n)}{n\Delta} \\ &\leq \frac{4R_{\gamma,\beta}(n)^{(2\alpha+d-\alpha\beta)/(\alpha+d-\alpha\beta)}}{n} \\ &\leq \frac{1}{4}\exp(-2),\end{aligned} \tag{16}$$

as we have set $R_{\gamma,\beta}^{(2\alpha+d-\alpha\beta)/(\alpha+d-\alpha\beta)} \leq \frac{\exp(-2)n}{16}$. Intuitively, Equation (15) tells us that the strategy suffers a regret of order $\mathcal{O}\left(n\Delta\right)$ with constant probability either under problem 0 or problem m.

In order to satisfy the bound $R_{\gamma,\beta}(n)$ on the regret of the strategy when it is facing problem 0, the probability of suffering regret of order $\mathcal{O}(n\Delta)$ under problem 0 cannot be too big (and in fact, for $\gamma > \alpha$, it vanishes), and thus, the strategy errs with constant probability under problem m. In other words, combining Equation Equation (15) and (16), we just showed that:

$$\mathbb{P}_m(R_n > \frac{n\Delta}{4}) \geq \mathbb{P}_m(T_m(n) < n/2) \geq \frac{1}{4}\exp(-2),$$

which implies directly, as R_n is a non-negative random variable:

$$\sup_{f \in \mathcal{P}(\alpha,\beta)} \mathbb{E}[R_n] \geq \mathbb{E}_m[R_n] \geq \frac{n\Delta}{16}\exp(-2) = \frac{n}{16}\exp(-2)R_{\gamma,\beta}(n)^{-\alpha/(\alpha+d-\alpha\beta)}$$

■

Appendix C. Proofs of Section 4

C.1. Proof of Theorem 4

Proof Let $\alpha_i = i/\lfloor\log(n)\rfloor^2$ for $i \in \{1, ..., \lfloor\log(n)\rfloor^3\}$. We write $\texttt{SR}(i)$ for the Subroutine i run with parameter α_i. We define $T_i(T)$ the number of samples allocated to the $\texttt{SR}(i)$ up to time T, and $\widehat{R}_T(i) = T_i(T)M(f) - \sum_{t=1}^{T_i(T)} Y_i(t)$ the regret incurred by $\texttt{SR}(i)$ after it has performed $T_i(T)$ function evaluations. We write the corresponding pseudo-regret $R_T(i) = T_i(T)M(f) - \sum_{t=1}^{T_i(T)} f(X_i(t))$, where $X_i(t)$ is the t-th sampling location chosen by $\texttt{SR}(i)$.

We have $\mathbb{E}(Y_i(t)) = f(X_i(t))$, and claim that $\widehat{R}_T(i) - R_T(i) = \sum_{t=1}^{T_i(T)} (f(X_i(t)) - Y_i(t))$ is a martingale with respect to the filtration $\mathcal{F}_T = \sigma(X_1, Y_1, ..., X_{T-1}, Y_{T-1}, X_T)$.
By standard concentration arguments and a union bound, we have for all i and all $T \leq n$ with probability at least $1-\delta$:

$$|\widehat{R}_T(i) - R_T(i)| \leq 2\sqrt{T_i(t)}\log(n\lfloor\log(n)\rfloor^3/\delta).$$

Fix k arbitrarily and consider the regret $\widehat{R}_n(k)$ that $\texttt{SR}(k)$ has incurred up to time n. Now consider $j \neq k$. The last time T that $\texttt{SR}(j)$ was chosen by the Meta-Strategy, we know that:

$$\begin{aligned} \widehat{R}_T(j) &\leq \widehat{R}_T(k) \\ &\leq R_T(k) + 2\sqrt{T_k(T)}\log(n\lfloor\log(n)\rfloor^3/\delta) \\ &\leq R_n(k) + 2\sqrt{n}\log(n\lfloor\log(n)\rfloor^3/\delta), \end{aligned}$$

where we used the fact that the pseudo-regret is non-decreasing with T. Furthermore, we know that once $\texttt{SR}(j)$ is chosen for the last time, it performs $\sqrt{n}$ function evaluations. This brings $\widehat{R}_j(n) = \widehat{R}_{T+\sqrt{n}}(j) \leq \widehat{R}_T(j) + \sqrt{n}$, as $f(X)$ is in $[0,1]$ for all X, so the regret incurred between time T and $T + \sqrt{n}$ is at most $\sqrt{n}$. If j is never chosen by the Meta-Strategy after the initial exploration phase that allocates $\sqrt{n}$ samples, the same bound trivially holds.

This allows us to bound for all $j \neq k$:

$$\widehat{R}_n(j) \leq R_n(k) + 3\sqrt{n}\log(n\lfloor\log(n)\rfloor^3/\delta)$$

By definition of the regret, the regret of the Meta-Strategy can be decomposed as the regret incurred by each $\texttt{SR}(i)$ up to time n:

$$\begin{aligned}\widehat{R}_n &= \sum_i \widehat{R}_n(i)\\ &\leq \lfloor \log(n) \rfloor^3 \left(R_n(k) + 3\sqrt{n}\log(n\lfloor\log(n)\rfloor^3/\delta) \right).\end{aligned}$$

We now consider i^* such that: $\alpha - \frac{1}{\lfloor \log^2(n) \rfloor} \leq \alpha_i^* \leq \alpha$. With probability at least $1-\delta$, we have by Proposition 2:

$$R_n(i^*) \leq D\log(n/\delta) n^{1-\alpha_{i^*}/(2\alpha_{i^*}+d-\alpha_{i^*}\beta)},$$

where we use the fact that $T_{i^*}(n) \leq n$ in the fixed budget setting. We conclude by using Lemma 3, which shows that our discretization over the smoothness parameters does not worsen the rate. ■

Lemma 3 *Let $\alpha > 0.5\sqrt{d/\log(n)}$ and consider $f \in \mathcal{P}(\alpha,\beta)$ and α_i such that: $\alpha - \lfloor\log(n)\rfloor^{-2} \leq \alpha_i \leq \alpha$. Then Subroutine 2 run with parameters α_i, n, δ is such that with probability at least $1-\delta$, we have:*

$$R_n \leq C\log\left(\frac{n}{\delta}\right)^p n^{1-\alpha/(2\alpha+d-\alpha\beta)},$$

where $p<1$ and $C>0$ is a constant that does not depend on n,δ.

Proof By Proposition 2 we have with probability at least $1-\delta$:

$$R_n \leq D\log\left(\frac{n}{\delta}\right)^p n^{1-\alpha_i/(2\alpha_i+d-\alpha_i\beta)}.$$

By considering the exponent $\frac{\alpha_i}{2\alpha_i+d-\alpha_i\beta}$, we have:

$$\begin{aligned}-\frac{\alpha_i}{2\alpha_i+d-\alpha_i\beta} &\leq -\frac{\alpha-\lfloor\log(n)\rfloor^{-2}}{2\alpha+d-\alpha\beta+\beta\lfloor\log(n)\rfloor^{-2}}\\ &\leq -\frac{\alpha}{2\alpha+d-\alpha\beta} + \frac{2\alpha+d}{\lfloor\log(n)\rfloor^2(2\alpha+d-\alpha\beta)^2},\end{aligned}$$

for $\alpha \geq \frac{1}{\lfloor\log(n)\rfloor}\sqrt{\frac{d}{2}}$ and we conclude by remarking that:

$$n^{\frac{2\alpha+d}{\lfloor\log(n)\rfloor^2(2\alpha+d-\alpha\beta)^2}} \leq \exp\left(\frac{\log(n)(2\alpha+d)}{\lfloor\log(n)\rfloor^2(2\alpha+d-\alpha\beta)^2}\right),$$

and thus for $\alpha \geq \frac{1}{2}\sqrt{\frac{d}{\log(n)}}$, this extra factor only worsens the rate by a constant. ■

C.2. Proof of Theorem 5

Proof The proof relies on the same notations and technical tools as in the proof of Theorem 4. We assume that on the event ξ, we have for all $i, T \leq n$:

$$|\widehat{R}_T(i) - R_T(i)| \leq 2\sqrt{T_i(t)} \log(n\lfloor \log(n) \rfloor^3/\delta).$$

with $\mathbb{P}(\xi) \geq 1 - \delta$.
We denote i^* the index of the Subroutine such that with probability at least $1-\delta$, we have for all $T \leq n$:

$$TM(f) - \sum_{t=1}^{T} f(X_{i^*}(t)) \leq R^*(n, \delta).$$

$R^*(n,\delta)$ is the maximum pseudo-regret for $\texttt{SR}(i^*)$ if it had been allocated the entire budget of n of function evaluations. We denote the event where this holds ξ'. We first show that with probability $1-2\delta$, $\texttt{SR}(i^*)$ is never eliminated by the Meta-Strategy. Let $\mathcal{A}_N$ be the set of active Subroutines at the beginning of round N. Assume that $i^* \in \mathcal{A}_N$ at the beginning of round N. We consider $k = \arg\max_{i\in\mathcal{A}_N} \widehat{S}_T(i)$ where $\widehat{S}_T(i) = \sum_{t=1}^{T_i(T)} Y_i(t)$ and $S_T(i) = \sum_{t=1}^{T_i(T)} f(X_i(t))$. We know that on ξ, we have:

$$\begin{aligned}\sum_{t=1}^{T_k(T)} Y_k(t) &\leq \sum_{t=1}^{T_k(T)} f(X_k(t)) + 2\sqrt{T_k(t)} \log(n\lfloor \log(n) \rfloor^3/\delta) \\ &\leq T_k(T)M(f) + 2\sqrt{T_k(t)} \log(n\lfloor \log(n) \rfloor^3/\delta),\end{aligned}$$

where we use $f(X_k(t)) \leq M(f)$ for any $X_k(t)$.
We also have on $\xi \cap \xi'$:

$$\begin{aligned}\sum_{t=1}^{T_{i^*}(T)} Y_{i^*}(t) &\geq \sum_{t=1}^{T} f(X_{i^*}(t)) - 2\sqrt{T_{i^*}(t)} \log(n\lfloor \log(n) \rfloor^3/\delta) \\ &\geq T_{i^*}(T)M(f) - R^*(n,\delta) - 2\sqrt{T_{i^*}(t)} \log(n\lfloor \log(n) \rfloor^3/\delta).\end{aligned}$$

For any $i \in \mathcal{A}_N$, $\texttt{SR}(i)$ has performed the same number of function evaluations $T_N \doteq N\sqrt{n}$ up to time T at the end of round N. Therefore on $\xi \cap \xi'$ the following holds:

$$\widehat{S}_T(k) - \widehat{S}_{i^*}(k) \leq R^*(n,\delta) + 4\sqrt{T_N} \log\left(\frac{n\lfloor \log(n) \rfloor^3}{\delta}\right),$$

and $i^* \in \mathcal{A}_{N+1}$. As $i^* \in \mathcal{A}_1$, by induction i^* is never eliminated on $\xi \cap \xi'$.
We now consider i such that $\texttt{SR}(i)$ is eliminated at round $N+1$, that is:

$$\widehat{S}_T(k) - \widehat{S}_i(k) \geq R^*(n,\delta) + 4\sqrt{T_{N+1}} \log\left(\frac{n\lfloor \log(n) \rfloor^3}{\delta}\right).$$

On $\xi \cap \xi'$, we know that at round N we had for $k = \arg\max_{i\in\mathcal{A}_N} \widehat{S}_T(i)$:

$$\begin{aligned}\widehat{S}_T(k) &\geq \widehat{S}_T(i^*) \\ &\geq T_N M(f) - R^*(n,\delta) - 2\sqrt{T_N} \log\left(\frac{n\lfloor \log(n) \rfloor^3}{\delta}\right),\end{aligned}$$

where we used the fact that i^* is never eliminated on $\xi \cap \xi$. Since $\mathtt{SR}(i)$ was eliminated at round $N+1$, it implies that at round N we had:

$$\begin{aligned}\widehat{S}_i(k) &\geq \widehat{S}_T(k) - R^*(n,\delta) - 4\sqrt{T_N}\log\left(\frac{n\lfloor\log(n)\rfloor^3}{\delta}\right)\\ &\geq T_N M(f) - 2R^*(n,\delta) - 6\sqrt{T_N}\log\left(\frac{n\lfloor\log(n)\rfloor^3}{\delta}\right),\end{aligned}$$

and on ξ this yields immediately:

$$T_N M(f) - \sum_{t=1}^{T_N} f(X_i(t)) \leq 2R^*(n,\delta) + 8\sqrt{T_N}\log\left(\frac{n\lfloor\log(n)\rfloor^3}{\delta}\right).$$

As $\mathtt{SR}(i)$ is allocated another $\sqrt{n}$ samples before being eliminated at round $N+1$, we can therefore bound its regret on $\xi \cap \xi'$ before being eliminated:

$$\begin{aligned}T_{N+1}M(f) - \sum_{t=1}^{T_{N+1}} f(X_i(t)) &= T_N M(f) - \sum_{t=1}^{T_N} f(X_i(t)) + \sqrt{n}M(f) - \sum_{T_N}^{T_N+\sqrt{n}} f(X_i(t))\\ &\leq 2R^*(n,\delta) + 8\sqrt{T_N}\log\left(\frac{n\lfloor\log(n)\rfloor^3}{\delta}\right) + \sqrt{n}\\ &\leq 2R^*(n,\delta) + 8\sqrt{n}\log\left(\frac{n\lfloor\log(n)\rfloor^3}{\delta}\right) + \sqrt{n}.\end{aligned}$$

Similarly, for i such that $\mathtt{SR}(i)$ is never eliminated, we have:

$$\begin{aligned}T_i(n)M(f) - \sum_{t=1}^{T_i(n)} f(X_i(t)) &\leq 2R^*(n,\delta) + 8\sqrt{T_i(n)}\log\left(\frac{n\lfloor\log(n)\rfloor^3}{\delta}\right)\\ &\leq 2R^*(n,\delta) + 8\sqrt{n}\log\left(\frac{n\lfloor\log(n)\rfloor^3}{\delta}\right).\end{aligned}$$

Finally, we can decompose the pseudo-regret of the Meta-Strategy as the sum of the pseudo-regret of each $\mathtt{SR}(i)$, which yields on $\xi \cap \xi'$:

$$\begin{aligned}R_n &= \sum_i R_i(n)\\ &\leq |\mathcal{A}_1|\left(2R^*(n,\delta) + 8\sqrt{n}\log\left(\frac{n\lfloor\log(n)\rfloor^3}{\delta}\right) + \sqrt{n}\right).\end{aligned}$$

By a union bound we have $\mathbb{P}(\xi \cap \xi') \geq 1 - 2\delta$, which concludes the proof. ■