Journal of Artificial Intelligence Research 82 (2025) 1773-1806 Submitted 11/2024; published 03/2025

Generalizing the Regret: an Analysis of Lower and Upper Bounds

Marco Mussi MARCO.MUSSI@POLIMI.IT
Alberto Maria Metelli ALBERTOMARIA.METELLI@POLIMI.IT
Politecnico di Milano
Piazza Leonardo da Vinci 32, Milan, 20133, Italy

Abstract

The *(expected cumulative) regret* is the customary index to judge the performance of online sequential decision-making algorithms. In the traditional form, it is defined as the expected sum over a learning horizon T of the sub-optimality gaps Δ_{I_t} (i.e., expected instantaneous regret) the agent suffers when playing arm I_t at round t. In this paper, we propose and investigate a generalization of this notion, named *g-(expected cumulative) regret*, obtained by applying a transformation function g to the sub-optimality gaps, making the agent suffer $g(\Delta_{I_t})$ instead of just Δ_{I_t}. Intuitively, function g embeds the "perception" that the agent manifests when performing a sub-optimal decision. We first show that sublinear g-regret is not achievable for a generic transformation function g. Then, we introduce a mild condition on g and provide instance-dependent and worst-case (i.e., minimax) lower bounds for the g-regret. Finally, we show that state-of-the-art stochastic bandit algorithms with no modification surprisingly display optimal performances for the g-regret. Specifically, we prove that `UCB1` matches (up to constant factors) the instance-dependent lower bound regardless of function g and that `MOSS` matches (up to constant factors) the minimax lower bound at least for a wide class of transformation functions.

1. Introduction

The notion of *regret* (Bubeck and Cesa-Bianchi, 2012; Lattimore and Szepesvári, 2020) is widespread to assess the performance of online learning algorithms. The stochastic *multi-armed bandit* framework (MAB, Lattimore and Szepesvári, 2020) considers the setting in which a learning agent is faced with a finite set of $K \in \mathbb{N}$ options (i.e., arms) and has to identify the best performing one. When an arm $i \in [\![K]\!] \coloneqq \{1, \dots, K\}$ is played, the agent observes a feedback $X \sim \nu_i$ (i.e., reward) sampled from the probability distribution ν_i with expected value μ_i (i.e., expected reward). We assume an additive noise model where $X = \mu_i + \epsilon$ and ϵ is a zero-mean σ^2-subgaussian random variable.[1] In such a case, the agent suffers a loss (i.e., expected instantaneous regret) that equals the performance sub-optimality gap $\Delta_i \coloneqq \mu_1 - \mu_i$, having assumed w.l.o.g. that 1 is the unique optimal arm (i.e., $1 = \arg\max_{i \in [\![K]\!]} \mu_i$). A regret-minimization algorithm $\mathfrak{A}$ seeks to control the *(expected cumulative) regret*, i.e., the sum of the sub-optimality gaps suffered over a learning horizon

1. A zero-mean random variable ϵ is σ^2-subgaussian if it holds that $\mathbb{E}[\exp(\xi\epsilon)] \leqslant \exp(\sigma^2\xi^2/2), \ \forall \xi \in \mathbb{R}$.

$T \in \mathbb{N}$ for a bandit instance $\boldsymbol{\nu} = (\nu_i)_{i \in [\![K]\!]}$:[2]

$$\mathbb{E}_{\boldsymbol{\nu}}[R(\mathfrak{A}, T)] := \mathbb{E}_{\boldsymbol{\nu}}\left[\sum_{t=1}^{T} \Delta_{I_t}\right] = \mathbb{E}_{\boldsymbol{\nu}}\left[\sum_{t=1}^{T} (\mu_1 - \mu_{I_t})\right], \tag{1}$$

where $I_t \in [\![K]\!]$ is the arm played at round $t \in [\![T]\!]$ and the expectation is computed w.r.t. the randomness of the rewards and the possible randomness of algorithm $\mathfrak{A}$.

The regret minimization problem in stochastic MABs has been widely investigated in the literature and it is currently well-understood. Traditionally, the study of the regret is conducted under two analysis approaches: *instance-dependent* and *worst-case*. In the former case, we characterize the complexity of the learning problem by the sub-optimality gaps Δ_i of the specific MAB instance. In such a case, every *consistent*[3] algorithm $\mathfrak{A}$ suffers asymptotically a regret lower bounded by (Lai and Robbins, 1985):

$$\liminf_{T \to +\infty} \frac{\mathbb{E}_{\boldsymbol{\nu}}[R(\mathfrak{A}, T)]}{\log T} \geqslant 2\sigma^2 \sum_{i \in [\![K]\!] \setminus \{1\}} \frac{1}{\Delta_i}. \tag{2}$$

The famous `UCB1` algorithm (Auer et al., 2002) matches the lower bound up to constant factors; while a slight modification of the exploration bonus of `UCB1` allows matching even the constants of the lower bound (Lattimore and Szepesvári, 2020). In the latter case, instead, we focus on characterizing the performance of the algorithm over the whole class of bandit instances with a fixed number of arms K. The worst-case (a.k.a. minimax) regret lower bound suffered by any learning algorithm $\mathfrak{A}$, if $T > K - 1$, is (Auer et al., 1995; Lattimore and Szepesvári, 2020):

$$\mathbb{E}_{\boldsymbol{\nu}}[R(\mathfrak{A}, T)] \geqslant \frac{\sigma}{27}\sqrt{(K-1)T}. \tag{3}$$

The `MOSS` algorithm (Audibert and Bubeck, 2009) matches this lower bound up to constant factors. More modern algorithms managed to achieve optimality in both asymptotic instance-dependent and the worst-case regimes (e.g., `ADA-UCB`, Lattimore, 2018).

The traditional regret definition (Equation 1) assumes that, at each round $t \in [\![T]\!]$, the agent suffers a loss that is equal to the sub-optimality gap Δ_{I_t} of the played arm I_t. This definition implies that the agent "perception" towards playing this sub-optimal option is *linear* in its sub-optimality. However, in several real-world scenarios, an agent may manifest different *non-linear* perceptions over the experienced sub-optimality. For instance, in economic scenarios, it is well-known that humans are more sensitive to large losses of money and they might be more willing to accept negligible money losses. In this sense, the perception of the experienced losses might encode a form of *safety* which amplifies, for instance, the impact of very bad choices or, contrary, a form of *carelessness* that attenuates, for instance, the regret brought by slightly sub-optimal actions. This form of perception that distinguishes between very bad and slightly sub-optimal actions can only be represented by a non-linear transformation of sub-optimality gaps. In more general terms, the perception

2. Whenever clear from the context, we will abbreviate *expected cumulative regret* with just *regret*.
3. An algorithm $\mathfrak{A}$ is consistent over a class of bandits if, for every bandit $\boldsymbol{\nu}$ in the class, it holds that $\lim_{T \to +\infty} \frac{\mathbb{E}_{\boldsymbol{\nu}}[R(\mathfrak{A}, T)]}{T^p} = 0$ for some $p > 0$ (Lattimore and Szepesvári, 2020).

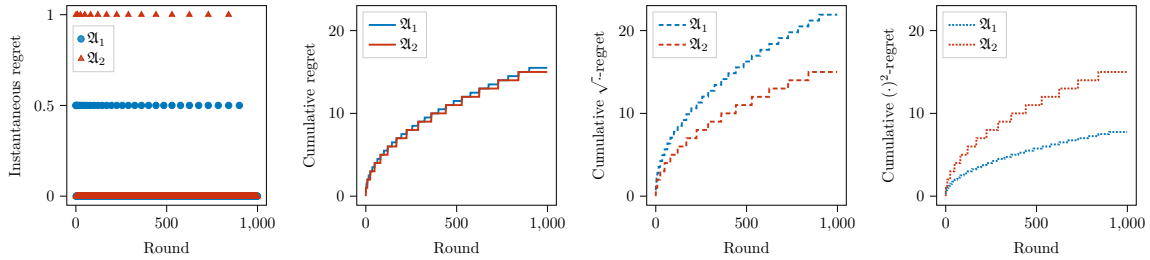

Figure 1: Example of execution of two algorithms $\mathfrak{A}_1$ (blue) and $\mathfrak{A}_2$ (red) with the corresponding expected instantaneous regret (left), g-(expected cumulative) regret with $g \in \{\mathrm{Id}, \sqrt{\cdot}, (\cdot)^2\}$ (others).

the agent expresses towards playing a sub-optimal option I_t becomes a non-linear function of its sub-optimality gap Δ_{I_t}. We propose to model this perception discrepancy through a (non-linear) transformation function g that maps the sub-optimality gap Δ_{I_t} to the loss $g(\Delta_{I_t})$ actually perceived by the agent. g can reduce or magnify the effect of each option; consequently, algorithms suffering the same regret (in the traditional sense) might display different performances in terms of g-regret, as shown in the following example.

Example 1. *Consider the two instantaneous regret plots presented in Figure 1 (left) obtained from the execution of algorithms $\mathfrak{A}_1$ (blue) and $\mathfrak{A}_2$ (red). They suffer the same expected cumulative regret (according to Equation 1, Figure 1 (middle left)), but $\mathfrak{A}_1$ plays a larger number of times an arm with small sub-optimality gap ($\Delta = 1/2$), while the $\mathfrak{A}_2$ plays a largely sub-optimal arm ($\Delta = 1$) a smaller number of times. Depending on g, the corresponding g-regrets appear different (Figure 1 (middle right)-(right)). Choosing $g = \sqrt{\cdot}$, we observe that $\mathfrak{A}_1$ has a worse $\sqrt{\cdot}$-regret (Figure 1 (middle right)), whereas, when $g = (\cdot)^2$, we note that the larger $(\cdot)^2$-regret is suffered by $\mathfrak{A}_2$ (Figure 1 (right)).*

Original Contributions. In this paper, we introduce a generalization of the customary notion of *(expected cumulative) regret* by means of a function g acting on the sub-optimality gaps. This way, through the choice of function g, we are able to encode the "perception" the agent manifests towards playing a sub-optimal option. We study this novel notion of *g-(expected cumulative) regret* from the perspective of the learnability, lower bounds, and upper bounds enjoyed by existing algorithms (i.e., `UCB1` and `MOSS`) for both the instance-dependent and worst-case regimes. We show that, surprisingly, with no knowledge of function g and with no need for modification, `UCB1` and `MOSS` are instance-dependent and minimax optimal (up to constant terms) for the g-regret, respectively. This shows that, remarkably, common algorithms, designed for traditional regret minimization, preserve optimality for this strictly larger class of performance indexes, i.e., the g-regret. The contributions of the paper can be summarized as follows:

- In Section 2, we formally introduce the notion of g-regret. Then, we show that the g-regret minimization problem is non-learnable for a generic function g (Theorem 1). Thus, we propose an assumption, enforcing that the optimal arm does not change when performing the transformation via function g, under which the learning problem becomes feasible (Assumption 1).

- In Section 3, we focus on the instance-dependent analysis. We start by deriving a novel asymptotic instance-dependent lower bound of order $\Omega\left(\sigma^2 \log T \sum_{i\in[\![K]\!]\setminus\{1\}} \frac{g(\Delta_i)}{\Delta_i^2}\right)$ for the g-regret (Theorem 2). Then, we show that for any function g fulfilling Assumption 1, `UCB1` (Auer et al., 2002) matches the lower bound up to constants (Theorem 3).
- In Section 4, we focus on the worst-case analysis. We first provide a novel implicit minimax lower bound for the g-regret (Theorem 4). Then, we show that for the particular class of monomial functions $g = (\cdot)^\alpha$ and for sufficiently large T, the lower bound takes the explicit form: $\Omega\left(\sigma^\alpha K^{\alpha/2} T^{1-\alpha/2}\right)$ for $\alpha \in [0, 2)$ and $\Omega\left(\sigma^2 K \log \frac{T}{K\sigma^2}\right)$ for $\alpha \geqslant 2$ (Corollary 5). Finally, we show that, for this class of g functions, the `MOSS` (Audibert and Bubeck, 2009) matches the lower bound up to constants (Theorem 6).

Omitted proofs can be found in Appendices A and B for Sections 3 and 4, respectively.

2. The g-Expected Cumulative Regret

In this section, we introduce our novel notion of *g-(expected cumulative) regret* (Section 2.1) and discuss the conditions under which the resulting problem is learnable (Section 2.2).

2.1 Definition of g-Expected Cumulative Regret

We consider an instance of the multi-armed bandit problem $\boldsymbol{\nu} = (\nu_i)_{i\in[\![K]\!]}$ with a finite number of arms $K \in \mathbb{N}$ and let $T \in \mathbb{N}$ be the learning horizon. At every round $t \in [\![T]\!]$, the agent selects an arm $I_t \in [\![K]\!]$, plays it, and observes a realization of the reward $X_t = \mu_{I_t} + \epsilon_t$, where μ_{I_t} is the expected reward and ϵ_t is a zero-mean σ^2-subgaussian random noise, independent conditioned to the past.[4] As customary in the bandit literature, we assume that the rewards are bounded in expectation, as $\mu_i \in [0, 1]$ for all $i \in [\![K]\!]$, so that the sub-optimality gaps Δ_i are bounded in $[0, 1]$. Finally, w.l.o.g. we assume that arm 1 is the unique optimal arm, i.e., $1 = \arg\max_{i\in[\![K]\!]} \mu_i$. We are now ready to formally define the notion of *g-(expected cumulative) regret*.

Definition 1 (g-(expected cumulative) regret)**.** *Let $\mathfrak{A}$ be a learning algorithm and let $T \in \mathbb{N}$ be a learning horizon. Let $g : [0, 1] \to [0, 1]$ be a transformation function. The* g-(expected cumulative) regret *induced by g is defined as:*

$$\mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A}, T)] := \mathbb{E}_{\boldsymbol{\nu}}\left[\sum_{t=1}^{T} g(\Delta_{I_t})\right] = \mathbb{E}_{\boldsymbol{\nu}}\left[\sum_{t=1}^{T} g(\mu_1 - \mu_{I_t})\right], \tag{4}$$

where $I_t \in [\![K]\!]$ is the arm played at round $t \in [\![T]\!]$ and the expectation is computed w.r.t. the randomness of the rewards and the possible randomness of algorithm $\mathfrak{A}$.

As already mentioned, the g-regret is obtained by transforming, through function g, the sub-optimality gaps Δ_{I_t} of the played arms I_t into $g(\Delta_{I_t})$. Clearly, if $g = \mathrm{Id}$, where Id is the identity function, Id-regret equals the traditional regret of Equation (1). It is worth noting that in Definition 1, the transformation function g is required to map the sub-optimality gaps (which have domain $[0, 1]$) into the same co-domain $[0, 1]$. This choice is w.l.o.g. and allows to carry out a clean analysis.

4. We consider subgaussian noise as it is more general and includes bounded (and Bernoulli) random variables (we recall that every random variable bounded in $[a, b]$ is subgaussian with $\sigma^2 = (b - a)^2/4$).

Remark 1 (Is g-regret minimization just regret minimization in a different bandit?)**.** *The reader might be tempted to conjecture that the g-regret minimization problem in bandit $\boldsymbol{\nu}$ can be equivalently stated as a traditional regret minimization problem in a transformed bandit $\boldsymbol{\nu}_g$. Unfortunately, this is not the case. First of all, we note that it is possible to define the transformed bandit $\boldsymbol{\nu}_g$ in such a way that its sub-optimality gaps Δ_i^g are equivalent to the transformed sub-optimality gaps $g(\Delta_i)$ in the original bandit $\boldsymbol{\nu}$, by enforcing the conditions:*

$$g(\Delta_i) = \Delta_i^g \iff g(\mu_1 - \mu_i) = \mu_1^g - \mu_i^g, \ \forall i \in [\![K]\!] \backslash \{1\},$$

where μ_i^g are the expected rewards of the transformed bandit $\boldsymbol{\nu}_g$. However, in order to carry out this transformation we would need to know the expected rewards μ_i in the original bandit (which are usually unknown) and function g as well.

We now present some examples of intuitive instances of generalized regret and we discuss their related g functions.

Example 2. *This generalized notion of regret is useful to take into account different* perceptions *of making a mistake in choosing the arm. Examples of that can be:*

- *Learn the ε-*optimal arms*: this definition of g is useful when we do not care about small mistakes. We define $g(\Delta) := \max\{0, \Delta - \varepsilon\}$, to assign zero regret to the arms that are ε-close to the optimal one ($\varepsilon > 0$). For instance, in online recommendations (e.g., product or movie recommendations) we might be satisfied with recommending an item that is slightly less preferred than the optimal choice as long as it is close in preference. For example, we might be equally satisfied with two similar movies, even if one is technically a better fit.*
- *Count the number of* mistakes*: this definition is useful in critical tasks. We select $g(\Delta) := \mathbb{1}\{\Delta > 0\} = \Delta^0$, in order to penalize every suboptimal arm in the same way. For instance, in critical applications like medical diagnostics, any wrong diagnosis can have serious consequences, regardless of how "close" it is to the correct diagnosis. If the optimal choice is diagnosing a certain disease, by choosing any other diagnosis, we should incur a penalty to ensure the model avoids any mistakes.*
- *Penalize very suboptimal arms or* "loss-aversion"*. We select $g(\Delta) := \sqrt{\Delta}$, or in general $g(\Delta) = \Delta^\alpha$, with $\alpha \in (0, 1)$, in order to magnify losses (since $\Delta \in [0, 1]$). For instance, in finance, investors often show loss aversion, where large mistakes in asset selection (e.g., choosing poor-performing stocks) are penalized more heavily than smaller mistakes.*

2.2 Non-Learnability for Generic g

In this section, we show that if g is chosen arbitrarily then the resulting g-regret minimization problem might become non-learnable. We start presenting an impossibility result (Theorem 1) and, then, we state an assumption that rules out this possibility, ensuring the learnability of the g-regret minimization problem (Assumption 1).

Theorem 1. *There exists a class of Gaussian MAB problems and a transformation function $g : [0, 1] \to [0, 1]$ such that any learning algorithm $\mathfrak{A}$ suffers g-regret lower bounded by:*

$$\mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A}, T)] \geqslant \frac{T}{8\sqrt{e}}.$$

Proof. We consider two Gaussian bandits $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$ with variance 1 and expected rewards $\boldsymbol{\mu} = (1, 1/2 + 1/\sqrt{T}, 1/2)^\top$ and $\boldsymbol{\mu}' = (1, 1/2, 1/2 + 1/\sqrt{T})^\top$, respectively. We consider the transformation function:

$$g(x) = \begin{cases} 1 & \text{if } x < 1/2 \\ x - 1/2 & \text{if } x \geqslant 1/2 \end{cases}.$$

The transformed sub-optimality gaps are $\boldsymbol{\Delta} = (1, 1, 0)^\top$ and $\boldsymbol{\Delta}' = (1, 0, 1)^\top$, respectively. Thus, in bandit $\boldsymbol{\nu}$ the optimal arm is 3, while in $\boldsymbol{\nu}'$ the optimal arm is 2. Thus:

$$\begin{aligned}
\max\{\mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A},T)], \mathbb{E}_{\boldsymbol{\nu}'}[R_g(\mathfrak{A},T)]\} &\geqslant \frac{1}{2}(\mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A},T)] + \mathbb{E}_{\boldsymbol{\nu}'}[R_g(\mathfrak{A},T)]) && (5)\\
&\geqslant \frac{T}{4}(\mathbb{P}_{\boldsymbol{\nu}}(N_2(T) \geqslant T/2) + \mathbb{P}_{\boldsymbol{\nu}'}(N_2(T) \leqslant T/2)) && (6)\\
&\geqslant \frac{T}{8}\exp\left(-\mathbb{E}_{\boldsymbol{\nu}}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}(\nu_{I_t} \| \nu'_{I_t})\right]\right) && (7)\\
&\geqslant \frac{T}{8\sqrt{e}}, && (8)
\end{aligned}$$

where line (5) follows from $\max\{a, b\} \geqslant \frac{1}{2}(a + b)$ for $a, b \geqslant 0$; line (6) is obtained by observing that every time arm 2 is played (resp. not played) in bandit $\boldsymbol{\nu}$ (resp. $\boldsymbol{\nu}'$) we suffer a g-instantaneous regret of 1, having denoted with $N_2(T)$ the number of times arm 2 is played over the horizon T; line (7) follows from the Bretagnolle-Huber's inequality (Bretagnolle and Huber 1978; see Theorem 14.2, Lattimore and Szepesvári, 2020), line (8) comes from observing that:

$$\begin{aligned}
D_{\mathrm{KL}}(\nu_1 \| \nu'_1) &= 0, \\
D_{\mathrm{KL}}(\nu_2 \| \nu'_2) &= D_{\mathrm{KL}}(\nu_3 \| \nu'_3) = \frac{1}{2T},
\end{aligned}$$

and bounding $D_{\mathrm{KL}}(\nu_{I_t} \| \nu'_{I_t}) \leqslant 1/(2T)$. □

The result essentially shows that if the transformation function g can freely change the optimal arm, we have no hope of conceiving a g-regret minimization algorithm that suffers sub-linear g-regret. This is supported by intuition since we lose the usual relation between the amount of samples needed to distinguish arms and their contribution to the regret. Thus, we introduce the following *optimality preserving* assumption on function g which guarantees that the optimal arm is not altered by the transformation, and, therefore, an agent remains able to recognize the optimal arm.

Assumption 1 (Optimality Preserving Function)**.** *Let $g : [0, 1] \to [0, 1]$. g is* optimality preserving*, i.e., $g(0) = 0$.*

Thus, any transformation function g fulfilling Assumption 1, preserves 1 as the optimal arm (with the possibility of generating additional optimal arms). For obtaining more explicit results in some of the subsequent sections, we introduce a further particularization of the transformation function g, i.e., *monomial* functions, as formalized in the following assumption that, clearly, implies Assumption 1.

Assumption 2 (Monomial Function)**.** *g is a monomial function, i.e., $g(x) = x^\alpha$ for $\alpha \in [0, +\infty)$.*

This class of functions allows managing cases in which we want to amplify ($\alpha \in (0, 1)$) or attenuate ($\alpha \in (1, +\infty)$) the regret, and manage extreme cases such as the one in which we want to count the number of mistakes ($\alpha = 0$, see Example 2).

3. Instance-Dependent Analysis

In this section, we provide the instance-dependent analysis for the g-regret minimization problem. First, we present an asymptotic lower bound to the g-regret depending on the transformation g (Section 3.1). Then, we show that the classical `UCB1` with no modifications matches (up to constant terms) the lower bound even without the knowledge of the transformation function g (Section 3.2).

3.1 Asymptotic Instance-Dependent Lower Bound

The following result provides the asymptotic instance-dependent lower bound for the g-regret.

Theorem 2 (Asymptotic Instance-Dependent Lower Bound)**.** *Let g fulfilling Assumption 1. For any consistent algorithm $\mathfrak{A}$, there exists a σ^2-subgaussian MAB $\boldsymbol{\nu}$ such that the asymptotic g-regret is lower bounded by:*

$$\liminf_{T \to +\infty} \frac{\mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A}, T)]}{\log T} \geqslant 2\sigma^2 \sum_{i \in [\![K]\!] \setminus \{1\}} \frac{g(\Delta_i)}{\Delta_i^2}.$$

Proof Sketch. The full proof is provided in Appendix A.1. This result can be obtained by rewriting the g-(expected cumulative) regret using the Wald's Identity (Wald, 1944):

$$\mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A}, T)] = \sum_{i \in [\![K]\!] \setminus \{1\}} g(\Delta_i) \mathbb{E}_{\boldsymbol{\nu}}[N_i(T)], \tag{9}$$

where $N_i(T)$ is the number of pulls of arm $i \in [\![K]\!] \setminus \{1\}$ up to round T and we excluded arm 1 from the summation since $g(0) = 0$ thanks to Assumption 1. We can now lower bound the expected number of pulls $\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)]$ as in the original instance-dependent lower bound proof, with no additional technical challenges. □

It is worth noting that, as apparent in the proof, under Assumption 1, the transformation function g does not affect the lower bound on the expected number of pulls $\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)]$, but just the instantaneous regret $g(\Delta_i)$. This particularly convenient decomposition allows us to show, as done in the next section, that `UCB1` achieves instance-dependent optimality.

3.2 Instance-Dependent Upper Bound for `UCB1`

We now illustrate that the asymptotic lower bound of Theorem 2 is matched (up to constant factors) by `UCB1` (Auer et al., 2002; Bubeck, 2010).[5]

5. The pseudo-code of `UCB1` is reported in Algorithm 1.

Algorithm 1 `UCB1` (Auer et al., 2002; Bubeck, 2010).

Require: number of arms K, exploration parameter $a > 2$, subgaussianity parameter σ
  $N_i \leftarrow 0,\ \widehat{\mu}_i \leftarrow 0,\ \mathrm{UCB}_i \leftarrow +\infty, \quad \forall i \in [\![K]\!]$
  for $t \in [\![T]\!]$ **do**
    Select $I_t \in \arg\max_{i \in [\![K]\!]} \mathrm{UCB}_i$
    Play I_t and observe reward X_t
    Update $\widehat{\mu}_{I_t} \leftarrow \dfrac{\widehat{\mu}_{I_t} N_{I_t} + X_t}{N_{I_t} + 1},\ N_{I_t} \leftarrow N_{I_t} + 1$
    Compute $\mathrm{UCB}_i \leftarrow \widehat{\mu}_i + \sigma\sqrt{\dfrac{a \log t}{N_i}}, \quad \forall i \in [\![K]\!]$
  end for

Theorem 3 (Instance-Dependent Upper Bound for `UCB1`)**.** *Let g fulfilling Assumption 1 and $\boldsymbol{\nu}$ be a σ^2-subgaussian MAB. The g-regret of* `UCB1` *with $a > 2$ is bounded by:*

$$\mathbb{E}_{\boldsymbol{\nu}}[R_g(\texttt{UCB1}, T)] \leqslant 4a\sigma^2 \log T \sum_{i \in [\![K]\!] \setminus \{1\}} \frac{g(\Delta_i)}{\Delta_i^2} + \left(\frac{2(a+2)}{a-2} + \frac{2}{\log\left(\frac{a+2}{4}\right)} \left(\frac{a+2}{a-2}\right)^2 \right) \sum_{i \in [\![K]\!] \setminus \{1\}} g(\Delta_i).$$

Proof Sketch. The full proof is provided in Appendix A.2. This result is obtained by rewriting the g-regret as in Equation (9), and, then, by upper bounding $\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)]$. □

The results of Theorem 2 and Theorem 3 prove that `UCB1` is instance-dependent optimal (up to constant terms), no matter the shape of function g, provided that it fulfills Assumption 1. From a technical perspective, this is somehow expected, since both the lower and the upper bounds are obtained by controlling the number of expected pulls $\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)]$, that does not depend on function g, and, then, rewriting the g-regret as in Equation (9). In a more general sense, we have demonstrated that, from the instance-dependent perspective, the class of performance indexes represented by the g-regrets (with g fulfilling Assumption 1) does not require exploration strategies different from that employed for the traditional regret.

Analysis under Assumption 2. Let us now particularize these results under Assumption 2, enforcing $g = (\cdot)^\alpha$, to obtain a more interpretable result. The following discussion holds for both the lower bounds and the `UCB1` upper bounds, as they present the same rate (neglecting constant terms):[6]

$$\mathbb{E}_{\boldsymbol{\nu}}[R_{(\cdot)^\alpha}(\mathfrak{A}, T)] = \Theta\left(\textstyle\sum_{i \in [\![K]\!] \setminus \{1\}} \Delta_i^{\alpha-2} \log T\right).$$

Figure 2 depict the behavior of $\Delta_i^{\alpha-2}$ for some interesting values of α. First of all, we note that when $\alpha = 1$ (i.e., $g = \mathrm{Id}$), we recover the usual bound on the regret. When $\alpha \in [0, 2)$ (i.e., $\alpha - 2 < 0$), the g-regret displays the characteristic behavior where small values of Δ_i lead to a large impact on the regret. Importantly, when $\alpha = 0$, the g-regret corresponds to the expected number of pulls of sub-optimal arms over the horizon T:

$$\mathbb{E}_{\boldsymbol{\nu}}[R_{(\cdot)^0}(\mathfrak{A}, T)] = \mathbb{E}_{\boldsymbol{\nu}}\left[\textstyle\sum_{t=1}^T \mathbb{1}\{I_t \neq 1\}\right] = \Theta\left(\textstyle\sum_{i \in [\![K]\!] \setminus \{1\}} \frac{\log T}{\Delta_i^2}\right).$$

6. We use the notation $f(x) = \Theta(g(x))$ to denote when, at the same time, it holds $f(x) = \Omega(g(x))$ and $f(x) = \mathcal{O}(g(x))$.

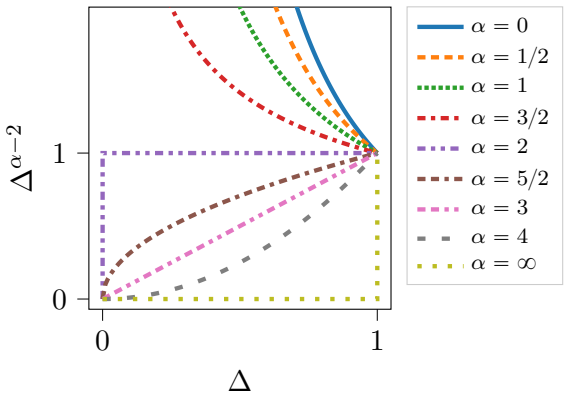

Figure 2: Behavior of the instance-dependent regret considering $g(\Delta) = \Delta^\alpha$ for representative values of α.

On the other hand, when $\alpha = 2$, the dependence on the sub-optimality gaps Δ_i completely disappears. Finally, when $\alpha > 2$ (i.e., $\alpha - 2 > 0$), the g-regret grows as the sub-optimality gaps increase, leading to the behavior in which small values of Δ_i lead to a small impact to the regret. This is supported by the intuition that since $\Delta_i \in [0, 1]$, large exponents α reduce the effect of playing sub-optimal arms on the g-regret.

4. Worst-Case Analysis

In this section, we focus on the worst-case analysis of the g-regret minimization problem. First, we present the worst-case (or minimax) lower bound for g-regret minimization in an implicit form, and, then, we devote particular attention to the case $g = (\cdot)^\alpha$ (Assumption 2) to obtain a more explicit result (Section 4.1). Finally, we present an upper bound of the g-expected cumulative regret suffered by `MOSS` (Audibert and Bubeck, 2009, 2010) under Assumption 2, showing its minimax optimality up to constants (Section 4.2).

4.1 Minimax Lower Bound

The following result provides the minimax lower bound for the g-regret.

Theorem 4 (Minimax Lower Bound)**.** *Let g fulfilling Assumption 1 and $T \in \mathbb{N}$ be the learning horizon. For any algorithm $\mathfrak{A}$, there exists a σ^2-subgaussian MAB $\boldsymbol{\nu}$ such that the g-regret is lower bounded by:*

$$\mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A}, T)] \geqslant \sup_{\Delta \in [0,1]} \left\{ g\left(\frac{\Delta}{2}\right) \frac{(K-1)\sigma^2}{\Delta^2} \log\left(\frac{T\Delta^2 e}{16(K-1)\sigma^2}\right) \right\}. \tag{10}$$

Proof Sketch. The full proof is provided in Appendix B.1. The instances are Gaussian bandits with σ^2 variance where the base instance has expected rewards $(\Delta/2, 0, \ldots, 0)^\top$, where 1 is the optimal arm. The alternative instance is constructed identically to the base instance with the only modification to the arm i that has been pulled the smallest number of times in the base instance whose expected reward is set to Δ as in (Theorem 15.2, Lattimore and Szepesvári, 2020), becoming the optimal arm. The technical novelty lies in

the construction of *two different lower bounds* on the g-regret that are averaged for obtaining the presented result, similarly as done in (Bubeck et al., 2013). The first one simply lower bounds the g-regret as a function of the number of times arm 1 is not pulled in the base instance $\sup_{\boldsymbol{\nu}} \mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A}, T)] \geqslant g(\Delta/2)(T - \mathbb{E}_{\boldsymbol{\nu}}[N_1(T)]) \geqslant g(\Delta/2)(K-1)\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)]$. The second one follows the usual change of measure arguments based on the Bretagnolle-Huber's inequality, leading to the result $\sup_{\boldsymbol{\nu}} \mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A}, T)] \geqslant T/2 g(\Delta/2) \exp\left(-\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)]\Delta^2/(2\sigma^2)\right)$. By taking the value of $\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)]$ that minimizes the average of these lower bounds (which can be computed in closed form by vanishing the derivative) we obtain the result. □

The result of Theorem 4 provides an implicit form for the regret bound and the solution of the optimization problem in Δ heavily depends on the form of the transformation function g. This bound can be made more explicit by considering the class of monomial functions, i.e., under Assumption 2, as illustrated in the following result.

Corollary 5 (Minimax Lower Bound for $g = (\cdot)^\alpha$)**.** *Let g fulfilling Assumption 2 and $T \in \mathbb{N}$ be the learning horizon. For any algorithm $\mathfrak{A}$, there exists a σ^2-subgaussian MAB $\boldsymbol{\nu}$ such that the g-regret is lower bounded by:*

$$\mathbb{E}_{\boldsymbol{\nu}}[R_{(\cdot)^\alpha}(\mathfrak{A}, T)] \geqslant \begin{cases} \dfrac{\sigma^\alpha (K-1)^{\alpha/2} T^{1-\alpha/2}}{2^{3-\alpha} e^{\alpha/2}(2-\alpha)} & \text{if } \alpha \in [0,2) \text{ and } T > \overline{T} \coloneqq 4\sigma^2(K-1)e^{\alpha/(2-\alpha)} \\ \frac{(K-1)\sigma^2}{2^\alpha} \log\left(\frac{Te}{16(K-1)\sigma^2}\right) & \text{if } \alpha \in [2, +\infty) \end{cases}.$$

Proof Sketch. The full proof is provided in Appendix B.1. The proof starts from Theorem 4 and, then, is based on solving the maximization problem over $\Delta \in [0,1]$. Specifically, when $\alpha \geqslant 2$, the expression of Equation (10) is increasing in Δ; thus, we choose $\Delta = 1$. Instead, when $\alpha \in [0,2)$, the resulting is a concave function in Δ and we find the optimal value of Δ by vanishing the derivative. The lower bound $\overline{T}$ on the learning horizon T is obtained given that there exist cases in which the optimal Δ for small time horizons is greater than 1, so we must select $\Delta = 1$. □

Corollary 5 sheds light on the behavior of the lower bound as a function of the exponent α. First of all, we note that when $\alpha = 1$, we recover the minimax lower bound $\Omega(\sigma\sqrt{KT})$ for the traditional regret. When $\alpha \in [0,2)$ (and the learning horizon T is sufficiently large) the lower bound displays an order of $\Omega(\sigma^\alpha K^{\alpha/2} T^{1-\alpha/2})$. Focusing on the dependence on the learning horizon T, we observe that the exponent decreases with α, suggesting that smaller α leads to a more challenging g-regret minimization problem. In particular, when $\alpha = 0$, i.e., as already observed, when we count the expected number of times a sub-optimal arm is pulled, the minimax lower bound degenerates to linear $\Omega(T)$. This is expected since, in such a case, we are losing the usual trade-off between exploration and exploitation (i.e., radically unbalanced towards exploration) as all sub-optimal arms equally impact the g-regret, regardless of their sub-optimality gap. Instead, when $\alpha \geqslant 2$, we observe that the minimax lower bound no longer depends on α (apart from some constants), recovering a logarithmic order $\Omega\left(\sigma^2 K \log \frac{T}{\sigma^2 K}\right)$. This suggests that, as α goes beyond the threshold value of 2, the complexity of the g-regret minimization problem is not affected by α. Contrary to the case of $\alpha = 0$, here the exploration/exploitation trade-off is radically unbalanced towards exploitation since the impact of the sub-optimality gap on the g-regret is attenuated by the large exponent α. A graphical illustration of this landscape is presented in Figure 3.

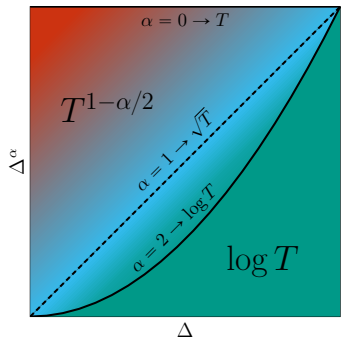

Figure 3: Graphical representation of the g-regret rate with $g = (\cdot)^\alpha$, highlighting the dependence on T only, as a function of the exponent $\alpha \geqslant 0$.

Algorithm 2 `MOSS` (Audibert and Bubeck, 2009, 2010).

Require: number of arms K, learning horizon T, subgaussianity parameter σ

$N_i \leftarrow 0,\ \widehat{\mu}_i \leftarrow 0,\ \mathrm{UCB}_i \leftarrow +\infty, \quad \forall i \in [\![K]\!]$

for $t \in [\![T]\!]$ **do**

Select $I_t \in \arg\max_{i \in [\![K]\!]} \mathrm{UCB}_i$

Play I_t and observe reward X_t

Update $\widehat{\mu}_{I_t} \leftarrow \dfrac{\widehat{\mu}_{I_t} N_{I_t} + X_t}{N_{I_t} + 1},\ \ N_{I_t} \leftarrow N_{I_t} + 1$

Compute $\mathrm{UCB}_{I_t} \leftarrow \widehat{\mu}_{I_t} + \sigma \sqrt{\dfrac{4}{N_{I_t}} \log^+ \left(\dfrac{T}{K\, N_{I_t}} \right)}$

where $\log^+(x) \coloneqq \log(\max\{1, x\})$

end for

4.2 Worst-Case Upper Bound for `MOSS`

In this section, we derive the upper bound on the g-expected cumulative regret of `MOSS` (Audibert and Bubeck, 2009, 2010), whose pseudo-code is provided in Algorithm 2. `MOSS` is an algorithm known to be minimax optimal up to constant factors for traditional regret. Our analysis limits to the case in which function g fulfills Assumption 2, i.e., monomial function $g = (\cdot)^\alpha$.[7]

Theorem 6 (`MOSS` Minimax Upper Bound for $g(x) = x^\alpha$)**.** *Let g fulfilling Assumption 2 and $\boldsymbol{\nu}$ be a σ^2-subgaussian MAB. The g-regret of* `MOSS` *is bounded by:*

$$\sup_{\boldsymbol{\nu}} \mathbb{E}_{\boldsymbol{\nu}}[R_{(\cdot)^\alpha}(\texttt{MOSS}, T)] \leqslant \begin{cases} \left(\frac{8 \cdot 2^{3\alpha}}{2-\alpha}\right) \sigma^\alpha K^{\alpha/2}\ T^{1-(\alpha/2)} + \sigma^\alpha K & \text{if } \ \alpha \in [0, 2) \\ 37 K \sigma^2 \log\left(\frac{T}{K\sigma^2}\right) + \sigma^2 \log\left(\frac{1}{4\sigma^2}\right) + 73\sigma^2 K + K + \sigma^2 & \text{if } \ \alpha = 2 \text{ and } T \geqslant \widetilde{T} \\ K\sigma^2 \left(\frac{69}{1-(2/\alpha)} + 37 \log\left(\frac{T}{\sigma^2 K}\right)\right) + K(1 + \sigma^\alpha) & \text{if } \ \alpha \in (2, \infty) \text{ and } T \geqslant \overline{T} \end{cases},$$

7. We discuss the reasons for this choice in Remark 2.

where $\widetilde{T} = \max\{e\sigma^2 K, 15K\}$ *and* $\overline{T} = \max\{e\sigma^2 K, 8^{-2\alpha/(2-\alpha)}/K\}$. *For the case* $\alpha \in (2, \infty)$ *and* $T < \overline{T}$ *and the case* $\alpha = 2$ *and* $T < \widetilde{T}$, *the g-regret is still logarithmic and the exact expression is reported in the proof.*

Proof Sketch. The full proof is provided in Appendix B.2. First of all, the proof is carried out for 1-subgaussian rewards and, subsequently, translating the obtained result for generic σ^2-subgaussian rewards (Lemma 3). The proof for the 1-subgaussian case follows the decomposition of the regret partitioning the arms based on the values of the sub-optimality gaps Δ_i, similarly to the original proof (Audibert and Bubeck, 2009), using the threshold $\max\{2\Delta, 8\sqrt{K/T}\}$, where Δ is a random variable suitably defined in the proof. However, the derivation takes different paths depending on whether $\alpha \in [0, 2)$ or $\alpha \geqslant 2$. Specifically, for the case $\alpha \geqslant 2$, tighter bounds on some relevant quantities are needed. □

By comparing this result with the lower bound of Corollary 5, we observe that, for sufficiently large T, all the relevant quantities, i.e., T, K and σ, are tight for every value of α, up to constant factors. This implies that `MOSS` preserves the minimax optimality for this large class of generalized regret functions. This further confirms that the g-regret minimization problem, although representing a strict generalization of the regret minimization problem, does not require different exploration strategies beyond the ones that can be employed for the traditional regret minimization problem.

Remark 2 (On the worst-case upper bound for `MOSS` without Assumption 2)**.** *The worst-case upper bound for* `MOSS` *presents analytical challenges when attempting to study it for a generic g fulfilling Assumption 1 only. Indeed, following the standard decomposition of the g-regret of (Audibert and Bubeck, 2009, 2010) as in Equation* (37) *for a generic g, we obtain:*

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\nu}}[g(\Delta)] &= \int_0^{+\infty} \mathbb{P}\left(g(2\Delta) > x\right) \mathrm{d}x \\
&= \int_0^{+\infty} \mathbb{P}\left(\Delta > \frac{g^{-1}(x)}{2}\right) \mathrm{d}x \\
&\leqslant \int_0^{+\infty} \min\left\{1, \frac{60K}{T(g^{-1}(x))^2}\right\} \mathrm{d}x \quad \text{(Lemma 9.3, Lattimore and Szepesvári, 2020)} \\
&\leqslant g\left(\frac{60K}{T}\right) + \int_{g\left(\frac{60K}{T}\right)}^{+\infty} \frac{60K}{T(g^{-1}(x))^2} \, \mathrm{d}x,
\end{aligned}$$

where g^{-1} *is the inverse function of g (that may be not well-defined). Even when function g is invertible, the integral for a generic g cannot be computed in closed form (even proceeding with the intuitive variable substitution* $y = g^{-1}(x)$*), preventing from obtaining an explicit rate for the regret.*

5. Related Works

In this section, we revise the literature that shares connections with our formulation, focusing on the approaches that somehow alter the objective of the learning process. Specifically, we survey: *safe exploration*, *risk-averse* learning, and *lenient regret* in MABs.

Safe Exploration in MABs. Safe exploration focuses on ensuring that during the learning phase, the performance does not fall below a certain threshold with high probability (Berkenkamp et al., 2017). In these works, the optimal arm does not change, and prescribing to learn *safely*, in practice, slows down the learning process only, generating additional terms in the regret. This type of algorithms is useful in safety-critical tasks (e.g., Cheng et al., 2019) and applies in many bandit formulations (e.g., Amani et al., 2019; Garcelon et al., 2020). The idea behind the algorithms may vary based on the specific setting. On the one hand, in a continuous-arms setting, we may exploit the regularity of the reward by means of Gaussian Processes (Sui et al., 2015; Schreiter et al., 2015; Amani et al., 2020a), or the particular structure of the rewards, as in the case of stochastic linear bandits (Amani et al., 2020b; Khezeli and Bitar, 2020) and contextual linear bandits (Kazerouni et al., 2017). On the other hand, other works propose different solutions to deal with safe exploration such as (Jagerman et al., 2020) that consider an initial policy known to the learner and propose an algorithm (using off-policy evaluation) that improves it only when there is high confidence that the performance is not worst than the previous one. Nevertheless, while supervising the exploration in an explicit way, we remark that these works, differently from ours, do not alter the definition of regret.

Risk-Averse Learning in MABs. Risk-averse MABs is a widely studied setting for critical task (e.g., Huo and Fu, 2017) or scenarios in which the stochasticity cannot be neglected (e.g., heavy-tail MABs, Kagrecha et al., 2019).[8] Thus, to evaluate the goodness of an arm, we can no longer rely on its expected value, as it does not capture the variability or uncertainty of outcomes. Usually, in risk-averse learning, we look at maximizing some quantity characterizing the distribution of the rewards (Cassel et al., 2018). The widely used indexes in this field are the Conditional Value at Risk (CVaR, Galichet et al., 2013; Curi et al., 2020; Chang et al., 2020; Khajonchotpanya et al., 2021) and the Mean-Variance (Sani et al., 2012; Vakili and Zhao, 2015, 2016). Other works consider wider classes of risk measures (e.g., Lipschitz risk functionals, Huang et al., 2021) or combinations of those (Yu and Nikolova, 2013; Kagrecha et al., 2019). Differently from our formulation in which the transformation is applied to sub-optimality gaps (which are deterministic quantities), risk-averse methods are based on indexes that take into account stochasticity and, as a possible effect, might change the notion of optimal arm.

Lenient Regret. The notion of *lenient regret* (Merlis and Mannor, 2021) has been introduced to account for the scenarios in which suboptimality gaps below a certain threshold ε are ignored in the regret. Formally, a function $f : [0,1] \to \mathbb{R}_{\geqslant 0}$ such that $f(\Delta) = 0$ if $\Delta \in [0, \varepsilon]$ and $f(\Delta) > 0$ if $\Delta > \varepsilon$ is considered in the computation of the lenient regret $\mathbb{E}_{\boldsymbol{\nu}}[\sum_{t=1}^{T} f(\Delta_{I_t})]$. As the authors state, while it is natural to choose function f as monotonically increasing, this condition is not requested for their analysis. Thus, the lenient regret formulation is not fully comparable with ours since it enforces the further constrained of having $f(\Delta) = 0$ if $\Delta \in [0, \varepsilon]$, but, at the same time, it does not require the monotonicity. Merlis and Mannor (2021) show that, for the class of consistent algorithms for the lenient regret, it is possible to achieve instance-dependent sub-logarithmic regret. We remark that, differently from what is done in (Merlis and Mannor, 2021) in Theorem 2, we are assuming the consistency of algorithm $\mathfrak{A}$ for the standard regret $R(\mathfrak{A}, T)$. This makes our result not

8. For a complete review on risk-aversion in MABs, please refer to (Tan et al., 2022).

		Lower Bound	Upper Bound	Match?
Instance Dependent	General g	$\sigma^2 \log T \sum_{i \in [\![K]\!] \setminus \{1\}} \frac{g(\Delta_i)}{\Delta_i^2}$	$\sigma^2 \log T \sum_{i \in [\![K]\!] \setminus \{1\}} \frac{g(\Delta_i)}{\Delta_i^2}$ (UCB1)	✓
	$g = (\cdot)^\alpha$	$\sigma^2 \log T \sum_{i \in [\![K]\!] \setminus \{1\}} \Delta^{\alpha-2}$	$\sigma^2 \log T \sum_{i \in [\![K]\!] \setminus \{1\}} \Delta_i^{\alpha-2}$ (UCB1)	✓
Worst-Case	General g	$\sup_{\Delta \in [0,1]} \left\{ g\left(\frac{\Delta}{2}\right) \frac{K\sigma^2}{\Delta^2} \log\left(\frac{T\Delta^2}{K\sigma^2}\right) \right\}$	? (see Remark 2)	?
	$g = (\cdot)^\alpha$	$\begin{cases} \sigma^\alpha K^{\alpha/2} T^{1-\alpha/2} & \text{if } \alpha \in [0,2) \\ \sigma^2 K \log\left(\frac{T}{\sigma^2 K}\right) & \text{if } \alpha \in [2,\infty) \end{cases}$	$\begin{cases} \sigma^\alpha K^{\alpha/2}\, T^{1-\alpha/2} & \text{if } \alpha \in [0,2) \\ \sigma^2 K \log\left(\frac{T}{\sigma^2 K}\right) & \text{if } \alpha \in [2,\infty) \end{cases}$ (MOSS)	✓†

Table 1: Comparison of instance-dependent and worst-case lower and upper bounds. All the results are reported considering the dominating terms in the bounds and neglecting constants. † for sufficiently large T (see Corollary 5 and Theorem 6).

comparable with (Theorem 1, Merlis and Mannor, 2021). Moreover, the authors provide a Thompson-sampling-like algorithm that matches the instance-dependent lower bound. Unfortunately, no worst-case bound is provided.

6. Discussion and Conclusions

In this paper, we presented a generalization of the customary concept of regret through a function g acting on the sub-optimality gaps. After having formally presented the formulation of the g-regret, we proved that the transformation function must at least preserve the optimal arm for the existence of no-regret algorithms. Then, we showed that UCB1 is instance-dependent optimal (up to constant terms) for every g preserving the optimality of the optimal arm. Then, we provided a minimax lower bound for generic functions g and we derived an explicit form for the class of the monomial functions. Then, we verified that MOSS preserves its minimax optimality for the g-regret (up to constant factors), at least for the monomial function. A summary of the results obtained in this paper is reported in Table 1. This work has illustrated how this strict superclass of performance indexes for evaluating online learning algorithms does not require conceiving different exploration strategies and can be effectively tackled with algorithms designed for the traditional regret. Future works should include the analysis of the MOSS algorithm for a generic transformation function g as well as the extension of the definition of g-regret to functions g_t indexed by the round t to account for the possible non-stationary perception of the agent on the regret.

Acknowledgements

Funded by the European Union – Next Generation EU within the project NRPP M4C2, Investment 1.3 DD. 341 – 15 March 2022 – FAIR – Future Artificial Intelligence Research – Spoke 4 – PE00000013 – D53C22002380006.

Appendix A. Omitted Proofs of the Instance-Dependent Analysis (Section 3)

In this appendix, we provide the formal proofs related to the instance-dependent bounds. More in detail, in Appendix A.1, we provide proofs for the lower bound, while in Appendix A.2, we provide proofs for the upper bound.

A.1 Lower Bounds

Theorem 2 (Asymptotic Instance-Dependent Lower Bound)**.** *Let g fulfilling Assumption 1. For any consistent algorithm $\mathfrak{A}$, there exists a σ^2-subgaussian MAB $\boldsymbol{\nu}$ such that the asymptotic g-regret is lower bounded by:*

$$\liminf_{T\to+\infty} \frac{\mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A},T)]}{\log T} \geqslant 2\sigma^2 \sum_{i\in[\![K]\!]\setminus\{1\}} \frac{g(\Delta_i)}{\Delta_i^2}.$$

Proof. The proof of this theorem starts by rewriting using Wald's Identity (Wald, 1944) the expected regret as the summation over all the arms of the expected number of pulls, multiplied by what we lose at each pull, i.e., $g(\Delta_i)$:

$$\mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A},T)] = \sum_{i\in[\![K]\!]} g(\Delta_i)\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)]. \tag{11}$$

Once we have this result, we can proceed to lower bound the expected number of pulls as in (Theorem 16.2, Lattimore and Szepesvári, 2020). What we get is:

$$\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)] \geqslant \frac{\log T}{D_{\mathrm{KL}}(\nu_1,\nu_i)},$$

where $D_{\mathrm{KL}}(\cdot,\cdot)$ is the Kullback-Leibler divergence, that, for σ^2-subgaussian random variables is simmetric, equal to:

$$D_{\mathrm{KL}}(\nu_1,\nu_i) = \frac{(\mu_1-\mu_i)^2}{2\sigma^2}. \tag{12}$$

We can now join all this information to get a lower bound on the instance-dependent expected regret, starting from Equation (11):

$$\liminf_{T\to\infty} \mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A},T)] = \sum_{i\in[\![K]\!]\setminus\{1\}} g(\Delta_i)\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)] \tag{13}$$

$$\geqslant \sum_{i\in[\![K]\!]\setminus\{1\}} g(\Delta_i)\left(\frac{2\sigma^2\log T}{\Delta_i^2}\right), \tag{14}$$

where in Equation (13) we removed the regret of the optimal arm, which is equal to zero (thanks to Assumption 1), and Equation (14) is obtained by replacing the value of the Kullback-Leiber Divergence (Equation 12). □

A.2 Upper Bounds

Theorem 3 (Instance-Dependent Upper Bound for `UCB1`)**.** *Let g fulfilling Assumption 1 and $\boldsymbol{\nu}$ be a σ^2-subgaussian MAB. The g-regret of* `UCB1` *with $a > 2$ is bounded by:*

$$\mathbb{E}_{\boldsymbol{\nu}}[R_g(\texttt{UCB1}, T)] \leqslant 4a\sigma^2 \log T \sum_{i \in [\![K]\!] \setminus \{1\}} \frac{g(\Delta_i)}{\Delta_i^2} + \left(\frac{2(a+2)}{a-2} + \frac{2}{\log\left(\frac{a+2}{4}\right)} \left(\frac{a+2}{a-2}\right)^2 \right) \sum_{i \in [\![K]\!] \setminus \{1\}} g(\Delta_i).$$

Proof. The proof of this statement can be obtained by rewriting the regret w.r.t. the expected number of pulls or each arm:

$$\mathbb{E}_{\boldsymbol{\nu}}[R_g(\texttt{UCB1}, T)] = \sum_{i \in [\![K]\!]} g(\Delta_i) \mathbb{E}_{\boldsymbol{\nu}}[N_i(T)]. \tag{15}$$

Then, the expected number of pulls can be bounded using Lemma 1 as:

$$\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)] \leqslant \frac{4a\sigma^2 \log T}{\Delta_i^2} + \frac{2(a+2)}{a-2} + \frac{2}{\log\left(\frac{a+2}{4}\right)} \left(\frac{a+2}{a-2}\right)^2, \tag{16}$$

for all the suboptimal arms $i \neq 1$.
Given Assumption 1, ensuring $g(0) = 0$, we know that the pulls of the optimal arm (i.e., 1, in our case) does not increase the regret, so we can rewrite Equation (15) as:

$$\mathbb{E}_{\boldsymbol{\nu}}[R_g(\texttt{UCB1}, T)] = \sum_{i \in [\![K]\!] \setminus \{1\}} g(\Delta_i) \mathbb{E}_{\boldsymbol{\nu}}[N_i(T)] \tag{17}$$

$$\leqslant \sum_{i \in [\![K]\!] \setminus \{1\}} g(\Delta_i) \left(\frac{4a\sigma^2 \log T}{\Delta_i^2} + \frac{2(a+2)}{a-2} + \frac{2}{\log\left(\frac{a+2}{4}\right)} \left(\frac{a+2}{a-2}\right)^2 \right), \tag{18}$$

where in Equation (17) we removed the regret of the optimal arm, which is equal to zero (thanks to Assumption 1), and Equation (18) is obtained by replacing the value of the upper bound on the expected number of pulls (Equation 16). □

Lemma 1. *Given an instance $\boldsymbol{\nu}$, the expected number of pulls of each suboptimal arm for* `UCB1` *on a σ^2-subgaussian MAB is:*

$$\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)] \leqslant \frac{4a\sigma^2 \log T}{\Delta_i^2} + \frac{2(a+2)}{a-2} + \frac{2}{\log\left(\frac{a+2}{4}\right)} \left(\frac{a+2}{a-2}\right)^2.$$

Proof. This proof of the expected number of pulls for σ^2-subgaussian variables extends the one of (Bubeck, 2010, Theorem 2.2) following the derivation of (Mussi et al., 2024, Theorem 4.2).

Given an instance $\boldsymbol{\nu}$, and considering a suboptimal action $i \in [\![K]\!]$, which suffers a suboptimality gap of Δ_i, we want to show that if $I_t = i$, then one of the three following equations is true:

$$\mathrm{UCB}_1(t) \leqslant \mu_1, \tag{19}$$

or:

$$\hat{\mu}_i(t-1) > \mu_i + \sigma \sqrt{\frac{a \log t}{N_i(t-1)}}, \tag{20}$$

or:

$$N_i(t-1) < \frac{4\sigma^2 a \log T}{\Delta_i^2}, \tag{21}$$

where $\mathrm{UCB}_i(t)$ is the upper confidence bound of the optimal arm for component i at time t, and $\hat{\mu}_i(t)$ is the estimated value of the mean of arm i after $N_i(t-1)$ pulls. For absurd, if we assume that the three equations are false, then we have:

$$\begin{aligned}
\mathrm{UCB}_1(t) &> \mu_1 \\
&= \mu_i + \Delta_i \\
&\geqslant \mu_i + 2\sqrt{\frac{\sigma^2 a \log t}{N_i(t-1)}} \\
&\geqslant \hat{\mu}_i(t) + \sqrt{\frac{\sigma^2 a \log t}{N_i(t-1)}} \\
&= \mathrm{UCB}_i(t),
\end{aligned}$$

which implies that $I_t \neq i$.

Now, we bound the probability that Equation (19) or Equation (20) hold true. Similar to the original proof, we use a peeling argument together with Hoeffding's maximal inequality, which is a consequence of Azuma-Hoeffding inequality. Note that:

$$\begin{aligned}
\mathbb{P}(\text{Eq. (19) is true}) &\leqslant \mathbb{P}\left(\exists s \in \{1,\dots,t\} : \hat{\mu}_1[s] + \sqrt{\frac{\sigma^2 a \log t}{s}} \leqslant \mu_1\right) \\
&= \mathbb{P}\left(\exists s \in \{1,\dots,t\} : \sum_{l=1}^{s}(X_1[l] - \mu_1) \leqslant -\sqrt{\sigma^2 a s \log t}\right),
\end{aligned}$$

where $\hat{\mu}_1[s]$ denotes the estimator computed with s samples and $X_1[l]$ is the l-th reward we observe related to arm 1. This result will provide an upper bound on the probability that the sum of independent bounded random variables deviates from its expected value. In particular, we rewrite the probability as the event that there exists some time s (from 1 up to t) for which the sum of deviations $\sum_{l=1}^{s}(X_1[l] - \mu_1)$ exceeds a certain threshold. The reason for which we consider all the s is that this is the worst-case scenario.

We now apply the peeling argument with a geometric grid over the time interval $[1, t]$. This step, which is an analytical improvement w.r.t. perform a union bound over t, has the advantage of reducing the minimum admissible value of a, i.e., the parameter regulating exploration by a factor 2. Given $\beta \in (0,1)$, we note that if $s \in \{1,\dots,t\}$, then $\exists j \in \left\{0,\dots,\frac{\log t}{\log 1/\beta}\right\} : \beta^{j+1}t < s \leqslant \beta^j t$. As such, we get:

$$\begin{aligned}
\mathbb{P}(\text{Eq. (19) is true}) &\leqslant \sum_{j=0}^{\frac{\log t}{\log 1/\beta}} \mathbb{P}\left(\exists s : \beta^{j+1}t < s \leqslant \beta^j t, \sum_{l=1}^{s}(X_1[l] - \mu_1) \leqslant -\sqrt{\sigma^2 a s \log t}\right) \\
&\leqslant \sum_{j=0}^{\frac{\log t}{\log 1/\beta}} \mathbb{P}\left(\exists s : \beta^{j+1}t < s \leqslant \beta^j t, \sum_{l=1}^{s}(X_1[l] - \mu_1) \leqslant -\sqrt{\sigma^2 a \beta^{j+1} t \log t}\right).
\end{aligned}$$

We now bound this last term using Hoeffding's maximal inequality, which gives:

$$\begin{aligned}\mathbb{P}(\text{Eq. (19) is true}) &\leqslant \sum_{j=0}^{\frac{\log t}{\log 1/\beta}} \exp\left(-\frac{\left(\sqrt{\sigma^2 a\beta^{j+1} t\log t}\right)^2}{2\sigma^2\beta^j t}\right)\\ &\leqslant \sum_{j=0}^{\frac{\log t}{\log 1/\beta}} \exp\left(-\frac{a\beta\log t}{2}\right)\\ &\leqslant \left(\frac{\log t}{\log 1/\beta}+1\right)\frac{1}{t^{\frac{\beta a}{2}}}.\end{aligned}$$

Using the same argument, it can be proven that:

$$\mathbb{P}(\text{Eq. (20) is true}) \leqslant \left(\frac{\log t}{\log 1/\beta}+1\right)\frac{1}{t^{\frac{\beta a}{2}}}.$$

We can now write:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\nu}}\left[N_i(T)\right] = \mathbb{E}_{\boldsymbol{\nu}}\left[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}\right] &\leqslant u + \mathbb{E}\left[\sum_{t=u+1}^{T}\mathbb{1}_{\{I_t=i \text{ and Eq. (21) is false}\}}\right]\\ &\leqslant u + \mathbb{E}\left[\sum_{t=u+1}^{T}\mathbb{1}_{\{\text{Eq. (19) or Eq. (20) is true}\}}\right]\\ &= u + \sum_{t=u+1}^{T}\left(\mathbb{P}(\text{Eq. (19) is true}) + \mathbb{P}(\text{Eq. (20) is true})\right),\end{aligned}$$

where $u = \left\lceil \frac{4\sigma^2 a\log T}{\Delta_i^2}\right\rceil$.
We can now upper bound the probability of Equations (19) and (20) holds:

$$\begin{aligned}&\sum_{t=u+1}^{T}\left(\mathbb{P}(\text{Eq. (19) is true}) + \mathbb{P}(\text{Eq. (20) is true})\right)\\ &\quad\leqslant 2\sum_{t=u+1}^{T}\left(\frac{\log t}{\log 1/\beta}+1\right)\frac{1}{t^{\frac{\beta a}{2}}}\\ &\quad\leqslant 2\int_1^{+\infty}\left(\frac{\log t}{\log 1/\beta}+1\right)\frac{1}{t^{\frac{\beta a}{2}}}dt\\ &\quad= 2\left[\left(\frac{\log t}{\log 1/\beta}+1\right)\left(\frac{2}{2-a\beta}t^{1-\frac{a\beta}{2}}\right)\right]_1^{+\infty} - \frac{4}{(2-a\beta)\log 1/\beta}\int_1^{+\infty}t^{-\frac{a\beta}{2}}dt \qquad (22)\\ &\quad= -\frac{4}{2-a\beta} - \frac{8}{(2-a\beta)^2\log 1/\beta}\left[t^{1-\frac{a\beta}{2}}\right]_1^{+\infty} \qquad (23)\\ &\quad= -\frac{4}{2-a\beta} + \frac{8}{(2-a\beta)^2\log 1/\beta},\end{aligned}$$

where Equation (22) is obtained via integration by parts and the first term of Equation (23) is obtained imposing $a\beta > 2$. Substituting now $\beta = \frac{4}{a+2}$, which verifies $\beta \in (0,1)$ for $a > 2$, we obtain:

$$\begin{aligned}\sum_{t=u+1}^{T} (\mathbb{P}(\text{Eq. (19) is true}) + \mathbb{P}(\text{Eq. (20) is true})) &\leqslant -\frac{4}{2-\frac{4a}{a+2}} + \frac{8}{\left(2-\frac{4a}{a+2}\right)^2}\frac{1}{\log\left(\frac{a+2}{4}\right)} \\ &= -\frac{2(a+2)}{2-a} + \frac{2(a+2)^2}{(2-a)^2}\frac{1}{\log\left(\frac{a+2}{4}\right)} \\ &= \frac{2(a+2)}{a-2} + \frac{2}{\log\left(\frac{a+2}{4}\right)}\left(\frac{a+2}{a-2}\right)^2.\end{aligned}$$

Rearranging the upper bound on the expected number of pulls given the three cases presented above, we get:

$$\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)] \leqslant \frac{4a\sigma^2 \log T}{\Delta_i^2} + \frac{2(a+2)}{a-2} + \frac{2}{\log\left(\frac{a+2}{4}\right)}\left(\frac{a+2}{a-2}\right)^2.$$

□

Appendix B. Omitted Proofs of the Worst-Case Analysis (Section 4)

In this appendix, we provide the formal proofs related to the worst-case bounds. More in detail, in Appendix B.1, we provide proofs for the lower bounds, while in Appendix B.2, we provide proofs for the upper bounds.

B.1 Lower Bounds

Theorem 4 (Minimax Lower Bound)**.** *Let g fulfilling Assumption 1 and $T \in \mathbb{N}$ be the learning horizon. For any algorithm $\mathfrak{A}$, there exists a σ^2-subgaussian MAB $\boldsymbol{\nu}$ such that the g-regret is lower bounded by:*

$$\mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A}, T)] \geqslant \sup_{\Delta\in[0,1]} \left\{ g\left(\frac{\Delta}{2}\right) \frac{(K-1)\sigma^2}{\Delta^2} \log\left(\frac{T\Delta^2 e}{16(K-1)\sigma^2}\right)\right\}. \tag{10}$$

Proof. We proceed by constructing a pair of MAB instances that are difficult to be distinguished. Let $\mathfrak{A}$ be an algorithm. Consider the base instance $\boldsymbol{\nu}$ of a Gaussian bandit with σ^2 variance (that is σ^2-subgaussian), whose expected rewards are defined as follows:

$$\boldsymbol{\mu} = \left(\frac{\Delta}{2}, 0, \dots, 0, \dots, 0\right),$$

with $\Delta \in [0,1)$. This for what concerns the first instance. Now we have to build a second instance $\boldsymbol{\nu}'$. Before doing that, we need to define index $i \in [\![K]\!]$ as the index of the arm pulled fewer times by an algorithm $\mathfrak{A}$ on bandit $\boldsymbol{\nu}$:

$$i \coloneqq \operatorname*{arg\,min}_{j\in[\![K]\!]\setminus\{1\}} \mathbb{E}_{\boldsymbol{\nu}}[N_j(T)]. \tag{24}$$

Let us now construct an alternative instance $\boldsymbol{\nu}'$ built from the base instance $\boldsymbol{\nu}$ in which we choose that the expected reward of arm i is increased from 0 (as in the base instance) to Δ:

$$\boldsymbol{\mu}' = \left(\frac{\Delta}{2}, 0, \dots, \underset{\text{position } i}{\Delta}, \dots, 0\right).$$

In this way, the less pulled suboptimal arm on instance $\boldsymbol{\nu}$, will become the best arm on instance $\boldsymbol{\nu}'$. We now lower bound the minimax g-regret as:

$$\sup_{\underline{\boldsymbol{\nu}}} \mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_g(\mathfrak{A}, T)] \geqslant \max\left\{\mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A}, T)], \mathbb{E}_{\boldsymbol{\nu}'}[R_g(\mathfrak{A}, T)]\right\}. \tag{25}$$

In order to get the result, we proceed lower bounding the last expression in two different ways (**LB1** and **LB2**).

Lower Bound 1 (LB1). The first derivation proceeds as follows:

$$\begin{aligned}
\sup_{\underline{\boldsymbol{\nu}}} \mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_g(\mathfrak{A}, T)] &\geqslant \max\left\{\mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A}, T)], \mathbb{E}_{\boldsymbol{\nu}'}[R_g(\mathfrak{A}, T)]\right\} \\
&\geqslant \mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A}, T)] \\
&\geqslant g\left(\frac{\Delta}{2}\right)(T - \mathbb{E}_{\boldsymbol{\nu}}[N_1(T)]),
\end{aligned}$$

where we noted that in bandit $\boldsymbol{\nu}$ the optimal arm is 1 and all other arms suffer $g(\Delta/2)$ as instantaneous g-regret. We can also note that, given that i is the index of the fewer pulled arm in instance $\boldsymbol{\nu}$, we have that:

$$\sum_{j \in [\![K]\!] \setminus \{1\}} \mathbb{E}_{\boldsymbol{\nu}}[N_j(T)] \geqslant (K-1)\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)].$$

Thus, we obtain the first lower bound:

$$\sup_{\underline{\boldsymbol{\nu}}} \mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_g(\mathfrak{A}, T)] \geqslant g\left(\frac{\Delta}{2}\right)(K-1)\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)] =: \mathbf{LB1}. \tag{26}$$

Lower Bound 2 (LB2). To obtain the second lower bound, we proceed with a standard change of measure argument based on Bretagnolle-Huber's inequality:

$$\begin{aligned}
\sup_{\underline{\boldsymbol{\nu}}} \mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_g(\mathfrak{A}, T)] &\geqslant \max\left\{\mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A}, T)], \mathbb{E}_{\boldsymbol{\nu}'}[R_g(\mathfrak{A}, T)]\right\} \\
&\geqslant \frac{1}{2}\left(\mathbb{E}_{\boldsymbol{\nu}}[R_g(\mathfrak{A}, T)] + \mathbb{E}_{\boldsymbol{\nu}'}[R_g(\mathfrak{A}, T)]\right) && (27) \\
&\geqslant \frac{T}{4}\, g\left(\frac{\Delta}{2}\right)\left(\mathbb{P}_{\boldsymbol{\nu}}(N_1(T) < T/2) + \mathbb{P}_{\boldsymbol{\nu}'}(N_1(T) \geqslant T/2)\right) && (28) \\
&\geqslant \frac{T}{8}\, g\left(\frac{\Delta}{2}\right)\exp\left(-D_{\mathrm{KL}}(\mathbb{P}_{\boldsymbol{\nu}} || \mathbb{P}_{\boldsymbol{\nu}'})\right) && (29) \\
&= \frac{T}{8}\, g\left(\frac{\Delta}{2}\right)\exp\left(-\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)]\frac{\Delta^2}{2\sigma^2}\right) =: \mathbf{LB2}, && (30)
\end{aligned}$$

where Equation (27) holds from $\max\{a,b\} \geqslant \frac{1}{2}(a+b)$ when both a and b are non-negative, Equation (28) comes from the observation that in instance $\boldsymbol{\nu}$ the optimal arm is 1 and in instance $\boldsymbol{\nu}'$ the optimal arm is $i \neq 1$ and the g-instantaneous regret when playing a sub-optimal arm in both instances is at least $g(\Delta/2)$, Equation (29) is derived from the Bretagnolle-Huber Inequality (Theorem 14.2, Lattimore and Szepesvári, 2020), and Equation (30) is the Kullback-Leibler divergence between the canonical distributions when arm i changes between the instances (Lemma 15.1, Lattimore and Szepesvári, 2020) for Gaussian reward distributions with variance σ^2.

Combining the Lower Bounds. The two bounds presented above must both holds at the same time, so we can take the maximum of them.

$$\begin{aligned}
\sup_{\underline{\boldsymbol{\nu}}} \mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_g(\mathfrak{A},T)] &\geqslant \max\{\textbf{LB1 (Eq. 26)}, \textbf{LB2 (Eq. 30)}\} \\
&\geqslant g\left(\frac{\Delta}{2}\right) \max\left\{(K-1)\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)], \frac{T}{8}\exp\left(-\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)]\frac{\Delta^2}{2\sigma^2}\right)\right\} \\
&\geqslant \frac{1}{2}\, g\left(\frac{\Delta}{2}\right)\left[(K-1)\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)] + \frac{T}{8}\exp\left(-\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)]\frac{\Delta^2}{2\sigma^2}\right)\right], \qquad (31)
\end{aligned}$$

where Equation (31) holds from $\max\{a,b\} \geqslant \frac{1}{2}(a+b)$ when both a and b are non-negative.

In order to get rid of the dependence on the expected number of pulls of arm i, we minimize $\mathbb{E}_{\boldsymbol{\nu}}[N_i(T)]$ (since we are looking at the performance of the best possible policy), constrained to the fact that this must be in $[0, \frac{T}{K-1}]$. Renaming $x \coloneqq \mathbb{E}_{\boldsymbol{\nu}}[N_i(T)]$ for simplicity, we have:

$$\sup_{\underline{\boldsymbol{\nu}}} \mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_g(\mathfrak{A},T)] \geqslant \frac{1}{2}\, g\left(\frac{\Delta}{2}\right) \min_{x\in[0,\frac{T}{K-1}]} \left\{(K-1)x + \frac{T}{8}\exp\left(-x\frac{\Delta^2}{2\sigma^2}\right)\right\} =: h(x).$$

Since function $h(x)$ is convex in x and, consequently, the minimum point can be found by vanishing the derivative:

$$\frac{\partial}{\partial x}h(x) = K - 1 - \frac{\Delta^2}{16\sigma^2}\frac{T}{2}\exp\left(-x\frac{\Delta^2}{2\sigma^2}\right) = 0 \implies x^* = \frac{2\sigma^2}{\Delta^2}\log\left(\frac{T\Delta^2}{16(K-1)\sigma^2}\right).$$

Substituting the value of x^* into $h(x)$, we get:

$$\begin{aligned}
&\sup_{\underline{\boldsymbol{\nu}}} \mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_g(\mathfrak{A},T)] \\
&\geqslant \frac{1}{2}\, g\left(\frac{\Delta}{2}\right)\left[\frac{2(K-1)\sigma^2}{\Delta^2}\log\left(\frac{T\Delta^2}{16(K-1)\sigma^2}\right) + \frac{T}{8}\exp\left(-\frac{2\sigma^2}{\Delta^2}\log\left(\frac{T\Delta^2}{16(K-1)\sigma^2}\right)\frac{\Delta^2}{2\sigma^2}\right)\right] \\
&= \frac{1}{2}\, g\left(\frac{\Delta}{2}\right)\left[\frac{2(K-1)\sigma^2}{\Delta^2}\log\left(\frac{T\Delta^2}{16K-1)\sigma^2}\right) + \frac{T}{8}\exp\left(-\log\left(\frac{T\Delta^2}{16(K-1)\sigma^2}\right)\right)\right] \\
&= g\left(\frac{\Delta}{2}\right)\frac{(K-1)\sigma^2}{\Delta^2}\log\left(\frac{T\Delta^2 e}{16(K-1)\sigma^2}\right).
\end{aligned}$$

In order to get the lower bound, we can choose the worst-case Δ, i.e., the one that maximizes the lower bound:

$$\sup_{\underline{\boldsymbol{\nu}}} \mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_g(\mathfrak{A}, T)] \geqslant \sup_{\Delta \in [0,1]} \left(g\left(\frac{\Delta}{2}\right) \frac{(K-1)\sigma^2}{\Delta^2} \log\left(\frac{T\Delta^2 e}{16(K-1)\sigma^2}\right)\right).$$

□

Corollary 5 (Minimax Lower Bound for $g = (\cdot)^\alpha$)**.** *Let g fulfilling Assumption 2 and $T \in \mathbb{N}$ be the learning horizon. For any algorithm $\mathfrak{A}$, there exists a σ^2-subgaussian MAB $\boldsymbol{\nu}$ such that the g-regret is lower bounded by:*

$$\mathbb{E}_{\boldsymbol{\nu}}[R_{(\cdot)^\alpha}(\mathfrak{A}, T)] \geqslant \begin{cases} \dfrac{\sigma^\alpha (K-1)^{\alpha/2} T^{1-\alpha/2}}{2^{3-\alpha} e^{\alpha/2}(2-\alpha)} & \text{if } \alpha \in [0,2) \text{ and } T > \overline{T} \coloneqq 4\sigma^2(K-1)e^{\alpha/(2-\alpha)} \\ \frac{(K-1)\sigma^2}{2^\alpha} \log\left(\frac{Te}{16(K-1)\sigma^2}\right) & \text{if } \alpha \in [2, +\infty) \end{cases}.$$

Proof. We start this proof from the result of Theorem 4:

$$\sup_{\underline{\boldsymbol{\nu}}} \mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_g(\mathfrak{A}, T)] \geqslant \sup_{\Delta \in [0,1]} \left(g\left(\frac{\Delta}{2}\right) \frac{(K-1)\sigma^2}{\Delta^2} \log\left(\frac{T\Delta^2 e}{16(K-1)\sigma^2}\right)\right) =: f(\Delta).$$

Considering $g = (\cdot)^\alpha$ we get:

$$\sup_{\underline{\boldsymbol{\nu}}} \mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_{(\cdot)^\alpha}(\mathfrak{A}, T)] \geqslant \frac{\Delta^{\alpha-2}(K-1)\sigma^2}{2^\alpha} \log\left(\frac{T\Delta^2 e}{16(K-1)\sigma^2}\right). \tag{32}$$

Case $\alpha \in (2, \infty)$**.** We immediately observe that for $\alpha \geqslant 2$ the function $f(\Delta)$ is increasing in Δ, so we select $\Delta = 1$ and we get that:

$$\sup_{\underline{\boldsymbol{\nu}}} \mathbb{E}_{\underline{\boldsymbol{\nu}}}[R_{(\cdot)^\alpha}(\mathfrak{A}, T)] \geqslant \frac{(K-1)\sigma^2}{2^\alpha} \log\left(\frac{Te}{16(K-1)\sigma^2}\right).$$

Case $\alpha \in [0, 2)$**.** In this case, we can observe that Equation (32) is a concave function in Δ, so we can search for the value of $\Delta \in [0,1]$ maximizing the function by vanishing the derivative. The first-order derivative of our function is:

$$\begin{aligned} \frac{\partial}{\partial \Delta} f(\Delta) =& \frac{(\alpha-2)(K-1)\sigma^2}{2^\alpha} \Delta^{\alpha-3} \log\left(\frac{T\Delta^2 e}{16(K-1)\sigma^2}\right) + \\ &+ \frac{(K-1)\sigma^2}{2^\alpha} \Delta^{\alpha-2} \left(\frac{1}{\Delta^2} \frac{16(K-1)\sigma^2}{Te} \frac{2Te}{16(K-1)\sigma^2} \Delta\right) \\ =& \frac{(K-1)\sigma^2 \Delta^{\alpha-3}}{2^\alpha} \left[(\alpha-2)\log\left(\frac{T\Delta^2 e}{16(K-1)\sigma^2}\right) + 2\right]. \end{aligned}$$

By enforcing this derivative equal to zero we get:

$$\Delta^* = 4\sigma \sqrt{\frac{(K-1)}{T}} \exp\left(\frac{\alpha/2}{2-\alpha}\right). \tag{33}$$

However, we need to check whether this value Δ^* lies in the interval $[0,1]$, otherwise the optimal value of Δ must be 1 (obtaining the result for the case $\alpha \geqslant 2$) since the function $f(\Delta)$ is increasing:

$$4\sigma\sqrt{\frac{(K-1)}{T}}\exp\left(\frac{\alpha/2}{2-\alpha}\right) \leqslant 1 \implies T \geqslant 16(K-1)\sigma^2\exp\left(\frac{\alpha}{2-\alpha}\right) =: \overline{T}. \tag{34}$$

Given that, we will select the value of Δ^* in Equation (33) for T satisfying Equation (34), and $\Delta = 1$ (as before) otherwise.
For T satisfying Equation (34) the regret is:

$$\begin{aligned}
\sup_{\underline{\nu}} \mathbb{E}_{\underline{\nu}}[R_{(\cdot)^\alpha}(\mathfrak{A},T)] &\geqslant \frac{(K-1)\sigma^2}{2^\alpha}\left(4\sigma\sqrt{\frac{K-1}{T}}\exp\left(\frac{\alpha/2}{2-\alpha}\right)\right)^{\alpha-2}\log\left(\exp\left(\frac{\alpha}{2-\alpha}\right)e\right) \\
&= \frac{(K-1)\sigma^2}{2^\alpha}(4\sigma)^{\alpha-2}\left(\frac{K-1}{T}\right)^{(\alpha-2)/2}\exp\left(\frac{\alpha(\alpha-2)}{2(2-\alpha)}\right)\frac{2}{2-\alpha} \\
&= \frac{(K-1)\sigma^2}{2^\alpha}4^{\alpha-2}\sigma^{\alpha-2}(K-1)^{(\alpha/2)-1}T^{1-(\alpha/2)}\exp\left(\frac{\alpha(\alpha-2)}{-2(\alpha-2)}\right)\frac{2}{2-\alpha} \\
&= \frac{\sigma^\alpha(K-1)^{\alpha/2}T^{1-(\alpha/2)}}{2^{3-\alpha}e^{\alpha/2}(2-\alpha)}.
\end{aligned}$$

□

B.2 Upper Bounds

Lemma 2. *Let g fulfilling Assumption 2 and $\boldsymbol{\nu}$ be a 1-subgaussian MAB with expected payoffs in $[0,1/\sigma]$. The g-regret of* MOSS *is bounded by:*

$$\sup_{\boldsymbol{\nu}} \mathbb{E}_{\boldsymbol{\nu}}[R_{(\cdot)^\alpha}(\texttt{MOSS},T)] \leqslant$$
$$\begin{cases}
\left(\frac{8\cdot 2^{3\alpha}}{2-\alpha}\right)K^{\alpha/2}\,T^{1-(\alpha/2)} + K & \text{if } \alpha \in [0,2) \\
37K\log\left(\frac{T}{K\sigma^2}\right) + \log\left(\frac{1}{4\sigma^2}\right) + (\sigma^{-2}+73)K + 2 & \text{if } \alpha = 2 \text{ and } T \geqslant \widetilde{T} \\
K\sigma^{2-\alpha}\left(\frac{69}{1-(2/\alpha)} + 37\log\left(\frac{T}{\sigma^2 K}\right)\right) + K(1+\sigma^{-\alpha}) & \text{if } \alpha \in (2,\infty) \text{ and } T \geqslant \overline{T}
\end{cases},$$

where $\widetilde{T} = 15K$ and $\overline{T} = \max\{e\sigma^2 K, 8^{-2\alpha/(2-\alpha)}/K\}$. For the case $\alpha \in (2,\infty)$ and $T < \overline{T}$ and the case $\alpha = 2$ and $T < \widetilde{T}$, the g-regret is still logarithmic and the exact expression is reported in the proof.

Proof. The proof of this lemma takes inspiration from one of (Theorem 9.1, Lattimore and Szepesvári, 2020) that demonstrates the regret bound for the standard regret of 1-subgaussian stochastic MAB with expected payoffs in $[0,1]$. We generalize this proof in order to account for the g-expected cumulative regret of 1-subgaussian stochastic MAB with expected rewards in $[0,1/\sigma]$. As usual, we assume w.l.o.g. that the first arm is optimal, so $\mu_1 = \mu^*$. In this refined analysis, the probability that an arm is played linearly often depends on its suboptimality gap. We follow the proof path of (Theorem 9.1, Lattimore and Szepesvári, 2020) and we start by making an argument in terms of the expected amount

of optimism. Define a random variable Δ that measures how far below the index of the optimal arm drops below its true mean:

$$\Delta \coloneqq \left(\mu_1 - \min_{s \leqslant T}\left(\hat{\mu}_1(s) + \sqrt{\frac{4}{s}\log^+\left(\frac{T}{Ks}\right)}\right)\right)^\dagger,$$

where $x^\dagger \coloneqq \max\{0, x\}$. Arms with suboptimality gaps much larger than Δ will not be played too often, while arms with suboptimality gaps smaller than Δ may be played linearly often, but Δ is sufficiently small in expectation that this price is small. Using the basic regret decomposition which involves Wald's Identity (Wald, 1944), for the standard regret we have:

$$\mathbb{E}_{\boldsymbol{\nu}}[R(\mathfrak{A}, T)] = \sum_{i \in [\![K]\!]} \Delta_i \mathbb{E}_{\boldsymbol{\nu}}[N_i(T)], \tag{35}$$

and splitting the actions based on whether or not their suboptimality gaps are smaller or larger than 2Δ leads to (for standard regret):

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\nu}}[R(\texttt{MOSS}, T)] &= \sum_{i \in [\![K]\!] \setminus \{1\}} \Delta_i \mathbb{E}_{\boldsymbol{\nu}}[N_i(T)] \\
&\leqslant \mathbb{E}_{\boldsymbol{\nu}}\left[2T\Delta + \sum_{i:\Delta_i > 2\Delta} \Delta_i N_i(T)\right] \\
&\leqslant \mathbb{E}_{\boldsymbol{\nu}}\left[\underbrace{2T\Delta}_{\texttt{A}} + \underbrace{8\sqrt{KT}}_{\texttt{B}} + \underbrace{\sum_{i:\Delta_i > \max\{2\Delta, 8\sqrt{K/T}\}} \Delta_i N_i(T)}_{\texttt{C}}\right].
\end{aligned} \tag{36}$$

We can generalize this bound to take into account the g-regret with transformation function $g = (\cdot)^\alpha$ and explicitly exploit that the sub-optimality gaps are in $[0, 1/\sigma]$. What we have is:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\nu}}[R_{(\cdot)^\alpha}(\texttt{MOSS}, T)] \\
&\quad= \sum_{i \in [\![K]\!] \setminus \{1\}} \Delta_i^\alpha \mathbb{E}_{\boldsymbol{\nu}}[N_i(T)] \\
&\quad\leqslant \mathbb{E}_{\boldsymbol{\nu}}\left[\underbrace{\min\left\{\frac{1}{\sigma}, 2\Delta\right\}^\alpha T}_{\texttt{A}} + \underbrace{8^\alpha K^{\alpha/2} T^{1-(\alpha/2)}}_{\texttt{B}} + \underbrace{\sum_{i:\Delta_i > \max\{2\Delta, 8\sqrt{K/T}\}} \Delta_i^\alpha N_i(T)}_{\texttt{C}}\right] \\
&\quad= \underbrace{T\ \mathbb{E}_{\boldsymbol{\nu}}\left[\min\left\{\frac{1}{\sigma}, 2\Delta\right\}^\alpha\right]}_{\texttt{A}} + \underbrace{8^\alpha K^{\alpha/2} T^{1-(\alpha/2)}}_{\texttt{B}} + \underbrace{\sum_{i:\Delta_i > \max\{2\Delta, 8\sqrt{K/T}\}} \Delta_i^\alpha \mathbb{E}_{\boldsymbol{\nu}}\left[N_i(T)\right]}_{\texttt{C}}.
\end{aligned} \tag{37}$$

We now discuss how Equation (36), which provides a bound for the standard regret, is related to the corresponding result for the g-regret (with bounded rewards in $[0, 1/\sigma]$ and

$g(x) = x^\alpha$) presented in Equation (37). First, the term $\mathtt{A}$ is modified to explicitly take into account the fact that the 2Δ cannot be greater than $1/\sigma$. Term $\mathtt{B}$ is no more than algebraic calculations observing that term Δ for this term is for sure lower than $8\sqrt{K/T}$ and can be pulled up to T times, so we can bound term $\mathtt{B} \leqslant (8\sqrt{K/T})^\alpha \cdot T = 8^\alpha K^{\alpha/2} T^{1-(\alpha/2)}$. Term $\mathtt{C}$, instead, is no more than the application of the notion of g-regret.
The three terms $\mathtt{A}$, $\mathtt{B}$ and $\mathtt{C}$, show different behavior in the case of $\alpha \in [0, 2)$, $\alpha = 2$ and $\alpha \in (2, \infty)$. The term $\mathtt{B}$ does not require further elaboration right now.

Case $\alpha \in [0, 2)$. We start to bound $\mathtt{A}$ by observing by making use of Lemma 5:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\nu}}[\Delta^\alpha] &= \int_0^\infty \mathbb{P}\left(\Delta^\alpha > x\right) \mathrm{d}x \\ &= \int_0^\infty \mathbb{P}\left(\Delta > x^{1/\alpha}\right) \mathrm{d}x \\ &\leqslant \int_0^\infty \min\left\{1, \frac{15K}{Tx^{2/\alpha}}\right\} \mathrm{d}x,\end{aligned}$$

where the last inequality follows from Lemma 4. Observing that the two components of the minimum are equal for $\bar{x} = \left(\frac{15K}{T}\right)^{\alpha/2}$, we have:

$$\mathbb{E}_{\boldsymbol{\nu}}[\Delta^\alpha] \leqslant \left(\frac{15K}{T}\right)^{\alpha/2} + \int_{\left(\frac{15K}{T}\right)^{\alpha/2}}^\infty \frac{15K}{Tx^{2/\alpha}}\, \mathrm{d}x \leqslant \frac{2}{2-\alpha}\left(\frac{15K}{T}\right)^{\alpha/2}.$$

Thus, the term $\mathtt{A}$ for $\alpha \in [0, 2)$ can be bounded as:

$$\mathtt{A} = T\ \mathbb{E}_{\boldsymbol{\nu}}\left[\min\left\{\frac{1}{\sigma}, 2\Delta\right\}^\alpha\right] \leqslant 2^\alpha T\ \mathbb{E}_{\boldsymbol{\nu}}\left[\Delta^\alpha\right] \leqslant \frac{2^{\alpha+1}}{2-\alpha} 15^{\alpha/2} K^{\alpha/2} T^{1-(\alpha/2)}.$$

The term $\mathtt{C}$ can be bounded as in (Theorem 9.1, Lattimore and Szepesvári, 2020) by resorting the demonstration to the worst-case scenario in the case of the g-expected cumulative regret. For suboptimal arm i, we define:

$$\kappa_i = \sum_{s=1}^T \mathbb{I}\left\{\hat{\mu}_i(s) + \sqrt{\frac{4}{s}\log^+\left(\frac{T}{Ks}\right)} \geqslant \mu_i + \Delta_i/2\right\}.$$

The reason for choosing κ_i in this way is that for arms i with $\Delta_i > 2\Delta$, it holds that the index of the optimal arm is always larger than $\mu_i + \Delta_i/2$, so κ_i is an upper bound on the number of times arm i is played, $N_i(T)$. If $\Delta_i \geqslant 8(K/T)^{1/2}$, then the expectation of $\Delta_i^\alpha \kappa_i$ is bounded using Lemma 6 by:

$$\begin{aligned}&\Delta_i^\alpha \mathbb{E}_{\boldsymbol{\nu}}\left[\kappa_i\right] \\ &\quad\leqslant \frac{1}{\Delta_i^{2-\alpha}} + \Delta_i^\alpha + \frac{8}{\Delta_i^{2-\alpha}}\left(2\log^+\left(\frac{T\Delta_i^2}{K}\right) + \sqrt{2\pi\log^+\left(\frac{T\Delta_i^2}{K}\right)} + 1\right) \\ &\quad\leqslant \frac{1}{\left(8\sqrt{K/T}\right)^{2-\alpha}} + \Delta_i^\alpha +\end{aligned}$$

$$
\begin{aligned}
&+ \frac{8}{\left(8\sqrt{K/T}\right)^{2-\alpha}} \left(2\log^+ \left(\frac{T}{K} \left(8\sqrt{\frac{K}{T}} \right)^2 \right) + \sqrt{2\pi \log^+ \left(\frac{T}{K} \left(8\sqrt{\frac{K}{T}} \right)^2 \right)} + 1 \right) \\
&\leqslant \Delta_i^\alpha + \frac{1}{\left(8\sqrt{K/T}\right)^{2-\alpha}} + \frac{8}{\left(8\sqrt{K/T}\right)^{2-\alpha}} \left(2\log^+ (64) + \sqrt{2\pi \log^+ (64)} + 1 \right) \\
&\leqslant \Delta_i^\alpha + \frac{1}{\left(8\sqrt{K/T}\right)^{2-\alpha}} \left(1 + 8 \left(2\log^+ (64) + \sqrt{2\pi \log^+ (64)} + 1 \right) \right) \\
&\leqslant \Delta_i^\alpha + \frac{117}{\left(8\sqrt{K/T}\right)^{2-\alpha}} \\
&\leqslant 1 + 117 \cdot 8^{\alpha-2} K^{(\alpha/2)-1} T^{1-(\alpha/2)},
\end{aligned}
$$

where the first inequality is obtained by observing that the function is decreasing in Δ_i. Finally, term $\mathtt{C}$ can be therefore bounded by considering that we have at most K arms satisfying this constraint as:

$$
\begin{aligned}
\mathtt{C} &= \sum_{i:\Delta_i > \max\{2\Delta, 8\sqrt{K/T}\}} \Delta_i^\alpha \mathbb{E}_{\boldsymbol{\nu}} \left[N_i(T) \right] \\
&\leqslant \sum_{i:\Delta_i > 8\sqrt{K/T}} \Delta_i^\alpha \mathbb{E}_{\boldsymbol{\nu}} \left[\kappa_i \right] \\
&\leqslant K \left(1 + 117 \cdot 8^{\alpha-2} K^{(\alpha/2)-1} T^{1-(\alpha/2)} \right) \\
&\leqslant K + 117 \cdot 8^{\alpha-2} K^{\alpha/2} T^{1-(\alpha/2)}.
\end{aligned}
$$

To summarize, we obtained that for $\alpha \in [0, 2)$ the minimax g-expected cumulative regret is bounded by:

$$
\begin{aligned}
\sup_{\boldsymbol{\nu}} \mathbb{E}_{\boldsymbol{\nu}}[R_{(\cdot)^\alpha}(\mathtt{MOSS}, T)] &\leqslant \underbrace{\frac{2^{\alpha+1}}{2-\alpha} 15^{\alpha/2} K^{\alpha/2} T^{1-(\alpha/2)}}_{\mathtt{A}} + \underbrace{8^\alpha K^{\alpha/2} T^{1-(\alpha/2)}}_{\mathtt{B}} + \\
&\quad + \underbrace{K + 117 \cdot 8^{\alpha-2} K^{\alpha/2} T^{1-(\alpha/2)}}_{\mathtt{C}} \\
&\leqslant \left(\frac{2^{\alpha+1}}{2-\alpha} 15^{\alpha/2} + 8^\alpha + 117 \cdot 8^{\alpha-2} \right) K^{\alpha/2}\, T^{1-(\alpha/2)} + K \\
&\leqslant \left(\frac{2^{\alpha+1}}{2-\alpha} 2^{2\alpha} + 2^{3\alpha} + 2^7 \cdot 2^{3\alpha-6} \right) K^{\alpha/2}\, T^{1-(\alpha/2)} + K \\
&\leqslant \left(\frac{2 \cdot 2^{3\alpha} + 6 \cdot 2^{3\alpha}}{2-\alpha} \right) K^{\alpha/2}\, T^{1-(\alpha/2)} + K \\
&= \left(\frac{8 \cdot 2^{3\alpha}}{2-\alpha} \right) K^{\alpha/2}\, T^{1-(\alpha/2)} + K,
\end{aligned}
$$

where the last inequality follows from the fact that $\alpha \in [0, 2)$.

Case $\alpha \in (2, \infty)$. We start by working in order to bound $\mathtt{A}$ as before with a refinement in the range of integration by preserving the minimum:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\nu}}\left[\min\left\{\frac{1}{2\sigma}, \Delta\right\}^{\alpha}\right] &= \int_0^{\infty} \mathbb{P}\left(\min\left\{\frac{1}{2\sigma}, \Delta\right\}^{\alpha} > x\right) \mathrm{d}x \\
&= \int_0^{1/(2\sigma)^{\alpha}} \mathbb{P}\left(\min\left\{\frac{1}{2\sigma}, \Delta\right\} > x^{1/\alpha}\right) \mathrm{d}x \\
&\leqslant \int_0^{1/(2\sigma)^{\alpha}} \mathbb{P}\left(\Delta > x^{1/\alpha}\right) \mathrm{d}x \\
&\leqslant \int_0^{1/(2\sigma)^{\alpha}} \min\left\{1, \frac{15K}{Tx^{2/\alpha}}\right\} \mathrm{d}x \qquad (38) \\
&\leqslant \left(\frac{15K}{T}\right)^{\alpha/2} + \int_{\left(\frac{15K}{T}\right)^{\alpha/2}}^{1/(2\sigma)^{\alpha}} \frac{15K}{Tx^{2/\alpha}}\, \mathrm{d}x \\
&\leqslant \left(\frac{15K}{T}\right)^{\alpha/2} + \frac{15K}{T}\frac{x^{1-(2/\alpha)}}{1-(2/\alpha)}\Bigg|_{\left(\frac{15K}{T}\right)^{\alpha/2}}^{1/(2\sigma)^{\alpha}} \\
&\leqslant \left(\frac{15K}{T}\right)^{\alpha/2} + \frac{1}{1-(2/\alpha)}\frac{15K}{T}\left(\frac{1}{(2\sigma)^{\alpha}}\right)^{1-(2/\alpha)} + \\
&\quad - \frac{1}{1-(2/\alpha)}\frac{15K}{T}\left(\frac{15K}{T}\right)^{(\alpha/2)-1} \\
&\leqslant -\frac{2/\alpha}{1-(2/\alpha)}\left(\frac{15K}{T}\right)^{\alpha/2} + \frac{1}{1-(2/\alpha)}\frac{15K}{T}\left(\frac{1}{(2\sigma)^{\alpha}}\right)^{1-(2/\alpha)} \\
&\leqslant \frac{1}{1-(2/\alpha)}\frac{15K}{T}(2\sigma)^{2-\alpha}.
\end{aligned}$$

The term $\mathtt{A}$ for $\alpha \in (2, \infty)$ can be bounded as:

$$\begin{aligned}
\mathtt{A} = T\ \mathbb{E}_{\boldsymbol{\nu}}\left[\min\left\{\frac{1}{\sigma}, 2\Delta\right\}^{\alpha}\right] &= 2^{\alpha}\ T\ \mathbb{E}_{\boldsymbol{\nu}}\left[\min\left\{\frac{1}{2\sigma}, \Delta\right\}^{\alpha}\right] \\
&= 2^{\alpha}\ T\ \frac{1}{1-(2/\alpha)}\frac{15K}{T}(2\sigma)^{2-\alpha} \\
&\leqslant \frac{60}{1-(2/\alpha)} K\sigma^{2-\alpha}.
\end{aligned}$$

The term $\mathtt{C}$ also for the case of $\alpha \in (2, \infty)$ following the same procedure we present for the case $\alpha \in [0, 2)$ by resorting to the proof of the worst-case scenario in the case of the g-expected cumulative regret, observing that the following function is increasing in Δ_i, so we can select $\Delta_i = 1/\sigma$:

$$\begin{aligned}
\Delta_i^{\alpha}\mathbb{E}_{\boldsymbol{\nu}}\left[N_i(T)\right] &\leqslant \frac{1}{\Delta_i^{2-\alpha}} + \Delta_i^{\alpha} + \frac{8}{\Delta_i^{2-\alpha}}\left(2\log^{+}\left(\frac{T\Delta_i^2}{K}\right) + \sqrt{2\pi\log^{+}\left(\frac{T\Delta_i^2}{K}\right)} + 1\right) \\
&\leqslant \sigma^{-\alpha} + \sigma^{2-\alpha}\left(1 + 8\left(2\log^{+}\left(\frac{T}{K\sigma^2}\right) + \sqrt{2\pi\log^{+}\left(\frac{T}{K\sigma^2}\right)} + 1\right)\right),
\end{aligned}$$

and the term $\mathtt{C}$ can be therefore bounded by considering that we have at most K arms satisfying this constraint as:

$$\begin{aligned}\mathtt{C} &\leqslant \sum_{i:\Delta_i > \max\{2\Delta, 8\sqrt{K/T}\}} \Delta_i^{\alpha} \mathbb{E}_{\boldsymbol{\nu}}\left[N_i(T)\right] \\ &\leqslant K\sigma^{-\alpha} + K\sigma^{2-\alpha}\left(1 + 8\left(2\log^+\left(\frac{T}{K\sigma^2}\right) + \sqrt{2\pi\log^+\left(\frac{T}{K\sigma^2}\right)} + 1\right)\right).\end{aligned}$$

To summarize, we obtained that for $\alpha \in (2, \infty)$ the minimax g-expected cumulative regret is bounded by:

$$\begin{aligned}\sup_{\boldsymbol{\nu}} \mathbb{E}_{\boldsymbol{\nu}}[R_{(\cdot)^\alpha}(\mathtt{MOSS}, T)] \leqslant \underbrace{\frac{60}{1-(2/\alpha)} K\sigma^{2-\alpha}}_{\mathtt{A}} + \underbrace{8^{\alpha} K^{\alpha/2} T^{1-(\alpha/2)}}_{\mathtt{B}} + \\ + \underbrace{K\sigma^{-\alpha}\left(1 + \sigma^2\left(9 + 16\log^+\left(\frac{T}{K\sigma^2}\right) + 8\sqrt{2\pi\log^+\left(\frac{T}{K\sigma^2}\right)}\right)\right)}_{\mathtt{C}}.\end{aligned} \tag{39}$$

To simplify this result more readable, we introduce the mild assumption that $T \geqslant eK\sigma^2$, so we have that $\log^+(T/(K\sigma^2)) = \log(T/(K\sigma^2))$ and $\log(T/(K\sigma^2)) \geqslant \sqrt{\log(T/(K\sigma^2))}$. This allows us to further simply the expression:

$$\begin{aligned}\sup_{\boldsymbol{\nu}} \mathbb{E}_{\boldsymbol{\nu}}[R_{(\cdot)^\alpha}(\mathtt{MOSS}, T)] &\leqslant 8^{\alpha} K^{\alpha/2} T^{1-(\alpha/2)} + K\sigma^{-\alpha} + \\ &\quad + K\sigma^{2-\alpha}\left[\frac{60}{1-(2/\alpha)} + 9 + 37\log\left(\frac{T}{\sigma^2 K}\right)\right] \\ &\leqslant 8^{\alpha} K^{\alpha/2} T^{1-(\alpha/2)} + K\sigma^{-\alpha} + K\sigma^{2-\alpha}\left[\frac{69}{1-(2/\alpha)} + 37\log\left(\frac{T}{\sigma^2 K}\right)\right].\end{aligned}$$

We now decide to further simplify the term $8^{\alpha} K^{\alpha/2} T^{1-(\alpha/2)}$ in order to bound it by a constant K for the sake of simplicity in calculations. To make this bound hold we have to impose a constraint on the minimum T:

$$8^{\alpha} K^{\alpha/2} T^{1-(\alpha/2)} \leqslant K \quad \Longrightarrow \quad T \geqslant 8^{-2\alpha/(2-\alpha)}/K.$$

This lead us to the final result for $\alpha \in (2, \infty)$:

$$\sup_{\boldsymbol{\nu}} \mathbb{E}_{\boldsymbol{\nu}}[R_{(\cdot)^\alpha}(\mathtt{MOSS}, T)] \leqslant 8^{\alpha} K^{\alpha/2} T^{1-(\alpha/2)} + K\sigma^{-\alpha} + K\sigma^{2-\alpha}\left[\frac{69}{1-(2/\alpha)} + 37\log\left(\frac{T}{\sigma^2 K}\right)\right],$$

which holds for $T \geqslant \max\{e\sigma^2 K, 8^{-2\alpha/(2-\alpha)}/K\}$.

Case $\alpha = 2$. The term $\mathtt{A}$ can be bounded by modifying the proof of case $\alpha \in [2, \infty)$ starting from Equation (38):

$$\mathtt{A} = T\ \mathbb{E}_{\boldsymbol{\nu}}\left[\min\left\{\frac{1}{\sigma}, 2\Delta\right\}^2\right] \leqslant \int_0^{1/(4\sigma^2)} \min\left\{1, \frac{15K}{Tx}\right\} \mathrm{d}x$$

$$
\begin{aligned}
&\leqslant \frac{15K}{T} + \int_{\frac{15K}{T}}^{1/(4\sigma^2)} \frac{15K}{Tx} \mathrm{d}x \\
&= \frac{15K}{T} + \frac{15K}{T} \log(x) \Big|_{\frac{15K}{T}}^{1/(4\sigma^2)} \\
&= \frac{15K}{T} \left(1 + \log\left(\frac{T}{15K}\right) + \log\left(\frac{1}{4\sigma^2}\right)\right) \\
&\leqslant 1 + \log\left(\frac{1}{4\sigma^2}\right),
\end{aligned}
$$

where the last inequality is derived assuming $T \geqslant 15K$ and observing that $y(1+\log(1/y)) \leqslant 1$ for $y \in [0,1]$.
The term $\mathtt{B}$ is equal to:

$$\mathtt{B} = 8^{\alpha} K^{\alpha/2} T^{1-(\alpha/2)} = 64K.$$

In order to compute term $\mathtt{C}$, we can start from the result of Equation (39), and observe that for $\alpha = 2$ we get (for $T \geqslant eK\sigma^2$):

$$
\begin{aligned}
\mathtt{C} &\leqslant K\sigma^{-\alpha} \left(1 + \sigma^2 \left(9 + 16 \log^+\left(\frac{T}{K\sigma^2}\right) + 8\sqrt{2\pi \log^+\left(\frac{T}{K\sigma^2}\right)}\right)\right) \\
&\leqslant K(\sigma^{-2} + 9) + 37K \log\left(\frac{T}{K\sigma^2}\right).
\end{aligned}
$$

Merging all these three terms together, under the condition that $T > \max\{e\sigma^2 K, 15K\}$, we get:

$$
\begin{aligned}
\sup_{\boldsymbol{\nu}} \mathbb{E}_{\boldsymbol{\nu}}[R_{(\cdot)^\alpha}(\mathtt{MOSS}, T)] &\leqslant 37K \log\left(\frac{T}{K\sigma^2}\right) + \log\left(\frac{1}{4\sigma^2}\right) + 64K + (\sigma^{-2} + 9)K + 1 \\
&\leqslant 37K \log\left(\frac{T}{K\sigma^2}\right) + \log\left(\frac{1}{4\sigma^2}\right) + (\sigma^{-2} + 73)K + 1.
\end{aligned}
$$

□

Theorem 6 ($\mathtt{MOSS}$ Minimax Upper Bound for $g(x) = x^{\alpha}$)**.** *Let g fulfilling Assumption 2 and $\boldsymbol{\nu}$ be a σ^2-subgaussian MAB. The g-regret of $\mathtt{MOSS}$ is bounded by:*

$$
\sup_{\boldsymbol{\nu}} \mathbb{E}_{\boldsymbol{\nu}}[R_{(\cdot)^\alpha}(\mathtt{MOSS}, T)] \leqslant
\begin{cases}
\left(\frac{8 \cdot 2^{3\alpha}}{2-\alpha}\right) \sigma^{\alpha} K^{\alpha/2}\, T^{1-(\alpha/2)} + \sigma^{\alpha} K & \text{if } \alpha \in [0,2) \\
37K\sigma^2 \log\left(\frac{T}{K\sigma^2}\right) + \sigma^2 \log\left(\frac{1}{4\sigma^2}\right) + 73\sigma^2 K + K + \sigma^2 & \text{if } \alpha = 2 \text{ and } T \geqslant \widetilde{T} \\
K\sigma^2 \left(\frac{69}{1-(2/\alpha)} + 37 \log\left(\frac{T}{\sigma^2 K}\right)\right) + K(1 + \sigma^{\alpha}) & \text{if } \alpha \in (2,\infty) \text{ and } T \geqslant \overline{T}
\end{cases},
$$

where $\widetilde{T} = \max\{e\sigma^2 K, 15K\}$ and $\overline{T} = \max\{e\sigma^2 K, 8^{-2\alpha/(2-\alpha)}/K\}$. For the case $\alpha \in (2,\infty)$ and $T < \overline{T}$ and the case $\alpha = 2$ and $T < \widetilde{T}$, the g-regret is still logarithmic and the exact expression is reported in the proof.

Proof. We can see the problem of finding a regret bound for a σ^2-subgaussian stochastic MAB with expected payoffs in $[0,1]$ as a 1-subgaussian stochastic MAB with expected payoffs in $[0,1/\sigma]$ by dividing all by σ. Given that, the proof follows from Lemma 3 and Lemma 2. □

Appendix C. Technical Lemmas

Lemma 3. *Let $\boldsymbol{\nu}$ be a σ^2-subgaussian MAB with sub-optimality gaps in $[0,1]$. Then, running algorithm $\mathfrak{A}$ dividing the observed rewards by σ leads to the g-regret:*

$$\mathbb{E}_{\boldsymbol{\nu}}[R_{(\cdot)^\alpha}(\mathfrak{A},T)] = \sigma^\alpha \mathbb{E}_{\boldsymbol{\nu}'}[R_{(\cdot)^\alpha}(\mathfrak{A},T)],$$

where $\mathbb{E}_{\boldsymbol{\nu}'}[R_{(\cdot)^\alpha}(\mathfrak{A},T)]$ is the g-regret (considering $g=(\cdot)^\alpha$) of $\mathfrak{A}$ for a 1-subgaussian bandit $\boldsymbol{\nu}'$ with expected rewards in $[0,1/\sigma]$.

Proof. Consider a K-armed σ^2-subgaussian bandit with expected rewards μ_i bounded in $[0,1]$, for every $i \in [\![K]\!]$. We can convert this problem in a 1-subgaussian bandit one with expected rewards (and, by consequence, sub-optimality gaps) in $[0,1/\sigma]$ by scaling all the expected rewards by σ.[9] The new expected rewards are given by:

$$\mu_i' = \frac{\mu_i}{\sigma},$$

and the related suboptimality gaps become:

$$\Delta_i' = \mu_1' - \mu_i = \frac{\mu_1 - \mu_i}{\sigma} = \frac{\Delta_i}{\sigma}.$$

The g-expected cumulative regret $\mathbb{E}_{\boldsymbol{\nu}}[R_{(\cdot)^\alpha}(\mathfrak{A},T)]$ of the new problem will become:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\nu}}[R_{(\cdot)^\alpha}(\mathfrak{A},T)] &= \sum_{i \in [\![K]\!]\setminus\{1\}} (\Delta_i')^\alpha \mathbb{E}_{\boldsymbol{\nu}}\left[N_i(T)\right] \\
&= \sum_{i \in [\![K]\!]\setminus\{1\}} \left(\frac{\Delta_i^\alpha}{\sigma^\alpha}\right) \mathbb{E}_{\boldsymbol{\nu}}\left[N_i(T)\right] \\
&= \sigma^{-\alpha} \sum_{i \in [\![K]\!]\setminus\{1\}} \Delta_i^\alpha \mathbb{E}_{\boldsymbol{\nu}}\left[N_i(T)\right] \\
&= \sigma^{-\alpha} \mathbb{E}_{\boldsymbol{\nu}'}[R_{(\cdot)^\alpha}(\mathfrak{A},T)].
\end{aligned}$$

This implies that the regret of the initial problem can be calculated as:

$$\mathbb{E}_{\boldsymbol{\nu}}[R_{(\cdot)^\alpha}(\mathfrak{A},T)] = \sigma^\alpha \mathbb{E}_{\boldsymbol{\nu}'}[R_{(\cdot)^\alpha}(\mathfrak{A},T)],$$

where $\mathbb{E}_{\boldsymbol{\nu}'}[R_{(\cdot)^\alpha}(\mathfrak{A},T)]$ is the regret of the transformed problem, which is a K-armed 1-subgaussian with expected payoffs bounded in $[0,1/\sigma]$. □

9. This can be seen by taking a generic random variable X and with finite first and second-order moments (i.e., with finite mean and variance). This is due to the fact that multiplying by σ a 1-subgaussian random variable, we get a σ^2 subgaussian random variable. This can be proven by taking a generic 1-subgaussian random variable X and observe that for every $k \in \mathbb{R}$, we have $\mathbb{E}[k \cdot X] = k \cdot \mathbb{E}[X]$ and:

$$\forall \lambda \in \mathbb{R}: \qquad \mathbb{E}[\exp(\underbrace{\lambda k}_{=\lambda'} X)] = \mathbb{E}[\exp(\lambda' X)] \leqslant \exp((\lambda')^2/2) = \exp(\lambda^2 k^2/2).$$

Lemma 4 (Lemma 9.3, Lattimore and Szepesvári 2020). *Let $\delta \in (0,1)$ and $X_1, X_2, \dots$ be independent and 1-subgaussian random variables. Let $\hat{\mu}(t) = \frac{1}{t}\sum_{s=1}^{t} X_s$. Then, for any $\Delta > 0$:*

$$\mathbb{P}\left(\exists s \geqslant 1 : \hat{\mu}(s) + \sqrt{\frac{4}{s}\log^+\left(\frac{1}{s\delta}\right)} + \Delta \leqslant 0\right) \leqslant \frac{15\delta}{\Delta^2}.$$

Lemma 5 (Proposition 2.8, Lattimore and Szepesvári 2020). *If $X \geqslant 0$ is a non-negative random variable, then:*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x)\,\mathrm{d}x.$$

Lemma 6 (Lemma 8.2, Lattimore and Szepesvári 2020). *Let $X_1, \dots, X_T$ be a sequence of independent 1-subgaussian random variables, $\hat{\mu}(t) = \frac{1}{t}\sum_{s=1}^{t} X_s$, $\varepsilon > 0$, $a > 0$ and:*

$$\kappa = \sum_{t=1}^{T} \mathbb{I}\left\{\hat{\mu}(t) + \sqrt{\frac{2a}{t}} \geqslant \varepsilon\right\},$$
$$\kappa' = u + \sum_{t=\lceil u \rceil}^{T} \mathbb{I}\left\{\hat{\mu}(t) + \sqrt{\frac{2a}{t}} \geqslant \varepsilon\right\},$$

where $u = 2a\varepsilon^{-2}$. Then, it holds:

$$\mathbb{E}[\kappa] \leqslant \mathbb{E}[\kappa'] \leqslant 1 + \frac{2}{\varepsilon^2}(a + \sqrt{\pi a + 1}).$$

Appendix D. Numerical Examples

In this appendix, we provide numerical examples to empirically validate our findings. We consider the performances of `UCB1` with a bandit made of $K = 10$ arms over 10 runs and comparing the empirical regret (`EXP`, mean $\pm$ std) with the instance-dependent lower (`LB`) and upper (`UB`) bounds, for different choices of function g and for different time horizons $T \in \{1 \cdot 10^5, 5 \cdot 10^5, 1 \cdot 10^6\}$. The results are presented in Table 2. We can observe how the empirical results are consistent with our theoretical findings for all the g and all the time horizons T considered.

T	$g(\Delta) = \max\{0, \Delta - \varepsilon\}$			$g(\Delta) = \sqrt{\Delta}$			$g(\Delta) = \Delta^2$			$g(\Delta) = \Delta$		
	`LB`	`EXP`	`UB`	`LB`	`EXP`	`UB`	`LB`	`EXP`	`UB`	`LB`	`EXP`	`UB`
$1 \cdot 10^5$	3.36	7.61 ± 0.23	26.84	27.01	50.63 ± 3.67	216.09	2.07	4.73 ± 0.17	16.58	9.96	19.74 ± 1.09	79.65
$5 \cdot 10^5$	3.82	8.71 ± 0.17	30.59	30.79	61.61 ± 4.01	246.3	2.36	5.45 ± 0.11	18.9	11.35	23.49 ± 1.04	90.78
$1 \cdot 10^6$	4.03	8.93 ± 0.18	32.21	32.41	66.36 ± 2.25	259.31	2.49	5.63 ± 0.11	19.89	11.95	25.06 ± 0.69	95.58

Table 2: Examples of g-expected cumulative regret lower and upper bounds in comparison with the empirical performance (10 runs, mean $\pm$ std).

References

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Tze L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE Annual Foundations of Computer Science*, pages 322–331. IEEE, 1995.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Annual Conference on Learning Theory (COLT)*, 2009.

Tor Lattimore. Refining the confidence level for optimistic bandit strategies. *Journal of Machine Learning Research*, 19(20):1–32, 2018.

Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Séminaire de probabilités de Strasbourg*, 12:342–363, 1978.

Abraham Wald. On cumulative sums of random variables. *The Annals of Mathematical Statistics*, 15(3):283–296, 1944.

Sébastien Bubeck. *Bandits games and clustering foundations*. PhD thesis, Université des Sciences et Technologie de Lille, 2010.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2785–2836, 2010.

Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *Annual Conference on Learning Theory (COLT)*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 122–134. JMLR, 2013.

Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *Advances in Neural Information Processing Systems (NIPS)*, pages 908–918, 2017.

Richard Cheng, Gábor Orosz, Richard M. Murray, and Joel W. Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 3387–3395, 2019.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9252–9262, 2019.

Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. Improved algorithms for conservative exploration in bandits. In *AAAI Conference on Artificial Intelligence*, pages 3962–3969. AAAI Press, 2020.

Yanan Sui, Alkis Gotovos, Joel W. Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 997–1005. JMLR, 2015.

Jens Schreiter, Duy Nguyen-Tuong, Mona Eberts, Bastian Bischoff, Heiner Markert, and Marc Toussaint. Safe exploration for active learning with gaussian processes. In *Machine Learning and Knowledge Discovery in Databases - European Conference (ECML PKDD)*, volume 9286 of *Lecture Notes in Computer Science*, pages 133–149. Springer, 2015.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Regret bound for safe gaussian process bandit optimization. In *Proceedings of the Annual Conference on Learning for Dynamics and Control (L4DC)*, volume 120 of *Proceedings of Machine Learning Research*, pages 158–159. PMLR, 2020a.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Generalized linear bandits with safety constraints. In *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, pages 3562–3566. IEEE, 2020b.

Kia Khezeli and Eilyan Bitar. Safe linear stochastic bandits. In *AAAI Conference on Artificial Intelligence*, pages 10202–10209. AAAI Press, 2020.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi, and Benjamin Van Roy. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3910–3919, 2017.

Rolf Jagerman, Ilya Markov, and Maarten de Rijke. Safe exploration for optimizing contextual bandits. *ACM Transactions on Information Systems*, 38(3):1–23, 2020.

Xiaoguang Huo and Feng Fu. Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society open science*, 4(11):171377, 2017.

Anmol Kagrecha, Jayakrishnan Nair, and Krishna P. Jagannathan. Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11269–11278, 2019.

Vincent Y. F. Tan, Prashanth L. A., and Krishna P. Jagannathan. A survey of risk-aware multi-armed bandits. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5623–5629, 2022.

Asaf B. Cassel, Shie Mannor, and Assaf Zeevi. A general approach to multi-armed bandits under risk criteria. In *Annual Conference on Learning Theory (COLT)*, volume 75 of *Proceedings of Machine Learning Research*, pages 1295–1306. PMLR, 2018.

Nicolas Galichet, Michèle Sebag, and Olivier Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning (ACML)*, volume 29 of *JMLR Workshop and Conference Proceedings*, pages 245–260. JMLR, 2013.

Sebastian Curi, Kfir Y. Levy, Stefanie Jegelka, and Andreas Krause. Adaptive sampling for stochastic risk-averse learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1036–1047, 2020.

Joel Q. L. Chang, Qiuyu Zhu, and Vincent Y. F. Tan. Risk-constrained thompson sampling for cvar bandits. *CoRR*, abs/2011.08046, 2020.

Najakorn Khajonchotpanya, Yilin Xue, and Napat Rujeerapaiboon. A revised approach for risk-averse multi-armed bandits under cvar criterion. *Operations Research Letters*, 49(4): 465–472, 2021.

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3284–3292, 2012.

Sattar Vakili and Qing Zhao. Mean-variance and value at risk in multi-armed bandit problems. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 1330–1335. IEEE, 2015.

Sattar Vakili and Qing Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):1093–1111, 2016.

Audrey Huang, Liu Leqi, Zachary C. Lipton, and Kamyar Azizzadenesheli. Off-policy risk assessment in contextual bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 23714–23726, 2021.

Jia Y. Yu and Evdokia Nikolova. Sample complexity of risk-averse bandit-arm selection. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2576–2582. IJCAI/AAAI, 2013.

Nadav Merlis and Shie Mannor. Lenient regret for multi-armed bandits. In *AAAI Conference on Artificial Intelligence*, pages 8950–8957. AAAI Press, 2021.

Marco Mussi, Simone Drago, Marcello Restelli, and Alberto M. Metelli. Factored-reward bandits with intermediate observations. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235, pages 36911–36952. PMLR, 2024.