Risk–Aversion in Multi–armed Bandits

Anonymous Author(s)
Affiliation
Address
email

Abstract

In stochastic multi–armed bandits the objective is to solve the exploration–exploitation dilemma and ultimately maximize the expected reward. Nonetheless, in many practical problems, maximizing the expected reward is not the most desirable objective. In this paper, we introduce a novel setting based on the principle of risk–aversion where the objective is to compete against the arm with the best risk–return trade–off. This setting proves to be intrinsically more difficult than the standard multi-arm bandit setting due in part to an exploration risk which introduces a regret associated to the variability of an algorithm. Using variance as a measure of risk, we introduce two new algorithms, we investigate their theoretical guarantees, and we report preliminary empirical results.

1 Introduction

The multi–armed bandit [11] is the most simple yet powerful model for formalizing the problem of on–line learning with partial feedback, which encompasses a large number of real–world applications, such as clinical trials, online advertisements, adaptive routing, and cognitive radio. In the stochastic multi–armed bandit model, a learner chooses among several arms (e.g., different treatments), each of which is characterized by an independent reward distribution (e.g., the effectiveness of the treatment). At each point in time, the learner selects one arm and receives a noisy reward observation from that arm (e.g., the effect of the treatment on one patient). Given a finite number of n rounds (e.g., patients involved in the clinical trial), the learner faces a dilemma between repeatedly exploring all the arms and collecting information about their rewards versus exploiting current reward estimates and selecting the arm with the highest estimated reward. Roughly speaking, the objective of the learner is to solve this exploration–exploitation dilemma and accumulate as much reward as possible over n rounds. In particular, multi–arm bandit literature typically focuses on the problem of finding a learning algorithm capable of maximizing the expected cumulative reward (i.e., the reward collected over n rounds averaged over all the possible realizations from the observations), thus implying that the best arm returns the highest expected reward. Nonetheless, in many practical problems, maximizing the expected reward is not always the most desirable objective. For instance, in clinical trials, the treatment which works best *on average* might also have considerable *variability*; resulting in adverse side effects for some patients. In this case, a treatment which is less effective on average but consistently effective on different patients is preferable w.r.t. an effective but risky treatment. More generally, some application objectives require an effective trade–off between risk and reward.

A large part of decision–making theory focuses on the definition and management of risk (see e.g., [7] for an introduction to risk with an expected utility theory perspective) and has mostly been studied in on–line learning within the so–called expert advice setting (i.e., adversarial full–information on–line learning). In particular, [6] showed that in general, although it is possible to achieve a small regret w.r.t. to the best expert in expectation, it is not possible to compete against the expert which best trades off between average return and risk. On the other hand, it is possible to define no–regret algorithms for simplified measures of risk–return. [13] studied the case of pure risk minimization (notably variance minimization) in an on-line setting where at each step the learner is given a covariance matrix and it has to choose a vector of weights so as to minimize the variance. The regret is then computed over horizon n and compared to the fixed weights minimizing the variance in hindsight. In the multi–arm bandit domain, the most interesting results are by [3] and [12]. [3] introduced an analysis of the expected regret and its distribution, revealing that an anytime version of *UCB* [5] and *UCB-V* might have large regret with some non-negligible probability.[1] This analysis is further extended by [12] who derived negative results showing that no anytime algorithm can achieve a regret with both a small expected regret and exponential tails. Although these results represent an important step towards the analysis of risk within bandit algorithms, they are limited to the case where an algorithm's cumulative reward is compared to the reward obtained by pulling the arm with the highest expectation.

In this paper, we focus on the problem of competing against the arm with the best risk–return trade–off. In particular, we refer to the first and most popular measure of risk–return, the mean–variance model introduce by [8].

The rest of the paper is organized as follows. In Section 2 we introduce notation and define the mean–variance bandit problem. In Section 3 we introduce a confidence–bound algorithm and study its theoretical properties. In Section 5 we report a set of numerical simulations aiming at validating the theoretical results. Finally, in Section 6 we conclude with a discussion on possible extensions. The proofs are reported in the supplementary material.

2 Risk–averse Multi–arm Bandit

In this section we introduce the main notation used throughout the paper and define the mean–variance multi–arm bandit problem.

We consider the standard multi–arm bandit setting with K arms, each characterized by a distribution ν_i bounded in the interval $[0, 1]$. Each distribution has a mean μ_i and a variance σ_i^2. The bandit problem is defined over a finite horizon of n rounds. We denote by $X_{i,s} \sim \nu_i$ the s-th random sample drawn from the distribution of arm i. All arms and samples are independent. In the multi–arm bandit protocol, at each round t, an algorithm selects arm I_t and observes sample $X_{I_t,T_{i,t}}$, where $T_{i,t}$ is the number of samples observed from arm i up to time t (i.e., $T_{i,t} = \sum_{s=1}^{t} \mathbb{I}\{I_t = i\}$).

While in the standard literature on multi–armed bandits the objective is to select the arm leading to the highest reward in *expectation* (the arm with the largest expected value μ_i), here we focus on the problem of finding the arm which effectively trades off between its expected reward (i.e., the *return*) and its variability (i.e., the *risk*). Although a large number of models for risk–return trade–off have been proposed, here we focus on the most popular and simple model: the mean–variance model proposed by [8],[2] where the return of an arm is measured by the expected reward and its risk by its variance.

Definition 1. *The mean–variance of an arm i with mean μ_i, variance σ_i^2 and coefficient of absolute risk tolerance ρ is defined as*[3] $\mathrm{MV}_i = \sigma_i^2 - \rho\mu_i$.

Thus it easily follows that the best arm minimizes the mean–variance, that is $i^* = \arg\min_{i=1,\ldots,K} \mathrm{MV}_i$. We notice that we can obtain two extreme settings depending on the value of risk tolerance ρ. As $\rho \to \infty$, the mean–variance of arm i tends to the opposite of its expected value μ_i and the problem reduces to the standard expected reward maximization traditionally considered in multi–arm bandit problems. With $\rho = 0$, the mean–variance formulation reduces to the variance σ_i^2 and the variance minimization problem.

Given $\{X_{i,s}\}_{s=1}^{t}$ i.i.d. samples from the distribution ν_i, we define the empirical mean–variance of an arm i with t samples as $\widehat{\mathrm{MV}}_{i,t} = \hat{\sigma}_{i,t}^2 - \rho\hat{\mu}_{i,t}$, where

$$\hat{\mu}_{i,t} = \frac{1}{t}\sum_{s=1}^{t} X_{i,s}, \qquad \hat{\sigma}_{i,t}^2 = \frac{1}{t}\sum_{s=1}^{t} \big(X_{i,s} - \hat{\mu}_{i,t}\big)^2. \tag{1}$$

[1]Although the analysis is mostly directed to the pseudo–regret, as commented in Remark 2 at page 23 of [3], it can be extended to the true regret.

[2]We discuss the limitations of this model and possible extensions to other models of risk in Section 6.

[3]The coefficient of risk tolerance is the inverse of the more popular coefficient of risk aversion $A = 1/\rho$.

We now consider a learning algorithm $\mathcal{A}$ and its corresponding performance over n rounds. Similar to a single arm i we define its empirical mean–variance as

$$\widehat{\mathrm{MV}}_n(\mathcal{A}) = \hat{\sigma}_n^2(\mathcal{A}) - \rho\hat{\mu}_n(\mathcal{A}), \tag{2}$$

where

$$\hat{\mu}_n(\mathcal{A}) = \frac{1}{n}\sum_{t=1}^{n} Z_t, \qquad \hat{\sigma}_n^2(\mathcal{A}) = \frac{1}{n}\sum_{t=1}^{n}\big(Z_t - \hat{\mu}_n(\mathcal{A})\big)^2, \tag{3}$$

with $Z_t = X_{I_t,T_{i,t}}$, that is the reward collected by the algorithm at time t. This leads to a natural definition of the (random) regret at each single run of the algorithm as the difference in the mean–variance performance of the algorithm compared to the best arm.

Definition 2. *The regret for a learning algorithm $\mathcal{A}$ over n rounds is defined as*

$$\mathcal{R}_n(\mathcal{A}) = \widehat{\mathrm{MV}}_n(\mathcal{A}) - \widehat{\mathrm{MV}}_{i^*,n}. \tag{4}$$

Given this definition, the objective is to design an algorithm whose regret decreases as the number of rounds increases (in high probability or in expectation).

We notice that the previous definition actually depends on *unobserved* samples. In fact, $\widehat{\mathrm{MV}}_{i^*,n}$ is computed on n samples i^* which are not actually observed when running $\mathcal{A}$. This matches the definition of *true* regret in standard bandits (see e.g., [3]). Thus, in order to clarify the main components characterizing the regret, we introduce additional notation. Let

$$Y_{i,t} = \begin{cases} X_{i^*,t} & \text{if } i = i^* \\ X_{i^*,t'} \text{ with } t' = T_{i^*,n} + \sum\limits_{j<i, j\neq i^*} T_{j,n} + t & \text{otherwise} \end{cases}$$

be a renaming of the samples from the optimal arm, such that while the algorithm was pulling arm i for the t-th time, $Y_{i,t}$ is the unobserved sample from i^*. Then we define the corresponding mean and variance as

$$\tilde{\mu}_{i,T_{i,n}} = \frac{1}{T_{i,n}}\sum_{t=1}^{T_{i,n}} Y_{i,t}, \qquad \tilde{\sigma}^2_{i,T_{i,n}} = \frac{1}{T_{i,n}}\sum_{t=1}^{T_{i,n}}\big(Y_{i,t} - \tilde{\mu}_{i,T_{i,n}}\big)^2. \tag{5}$$

Given these additional definitions, we can rewrite the regret as (see Appendix A.1)

$$\begin{aligned}\mathcal{R}_n(\mathcal{A}) = {} & \frac{1}{n}\sum_{i\neq i^*} T_{i,n}\Big[(\hat{\sigma}^2_{i,T_{i,n}} - \rho\hat{\mu}_{i,T_{i,n}}) - (\tilde{\sigma}^2_{i,T_{i,n}} - \rho\tilde{\mu}_{i,T_{i,n}})\Big] \\ & + \frac{1}{n}\sum_{i=1}^{K} T_{i,n}\big(\hat{\mu}_{i,T_{i,n}} - \hat{\mu}_n(\mathcal{A})\big)^2 - \frac{1}{n}\sum_{i=1}^{K} T_{i,n}\big(\tilde{\mu}_{i,T_{i,n}} - \hat{\mu}_{i^*,n}\big)^2. \end{aligned} \tag{6}$$

Since the last term is always negative and small [4], our analysis focuses on the first two terms which reveal two interesting characteristics of $\mathcal{A}$. First, an algorithm $\mathcal{A}$ suffers a regret whenever it chooses a suboptimal arm $i \neq i^*$ and the regret corresponds to the difference in the empirical mean–variance of i w.r.t. the optimal arm i^*. Such a definition has a strong similarity to the standard definition of regret, where i^* is the arm with highest expected value and the regret depends on the number of times suboptimal arms are pulled and their respective gaps w.r.t. the optimal arm i^*. In contrast to the standard formulation of regret, $\mathcal{A}$ also suffers from an additional regret from the variance $\hat{\sigma}_n^2(\mathcal{A})$ which depends on the variability of pulls $T_{i,n}$ over different arms. Recalling the definition of the mean $\hat{\mu}_n(\mathcal{A})$ as the weighted mean of the empirical means $\hat{\mu}_{i,T_{i,n}}$ with weights $T_{i,n}/n$ (see eq. 3), we notice that this second term is a weighted variance of the means and illustrates the exploration risk of the algorithm. In fact, if an algorithm simply selects and pulls a single arm from the beginning, it would not suffer any exploration risk (secondary regret) since $\hat{\mu}_n(\mathcal{A})$ would coincide with $\hat{\mu}_{i,T_{i,n}}$ for the chosen arm and all other components would have zero weight. On the other hand, an algorithm accumulates exploration risk through this second term as the mean $\hat{\mu}_n(\mathcal{A})$ deviates from any specific arm, where the maximum exploration risk resulting from a mean $\hat{\mu}_n(\mathcal{A})$ furthest from all arm means.

[4] More precisely, it can be shown that this term decreases with rate $O(K\log(1/\delta)/n)$ with probability $1-\delta$

The previous definition of regret can be further elaborated and obtain the upper bound (see App. A.1)

$$\mathcal{R}_n(\mathcal{A}) \leq \frac{1}{n}\sum_{i\neq i^*} T_{i,n}\widehat{\Delta}_i + \frac{1}{n^2}\sum_{i=1}^{K}\sum_{j\neq i} T_{i,n}T_{j,n}\widehat{\Gamma}^2_{i,j}, \tag{7}$$

where $\widehat{\Delta}_i = (\hat{\sigma}^2_{i,T_{i,n}} - \tilde{\sigma}^2_{i,T_{i,n}}) - \rho(\hat{\mu}_{i,T_{i,n}} - \tilde{\mu}_{i,T_{i,n}})$ and $\widehat{\Gamma}^2_{i,j} = (\hat{\mu}_{i,T_{i,n}} - \hat{\mu}_{j,T_{j,n}})^2$. Unlike the definition in eq. 6, this upper bound explicitly illustrates the relationship between the regret and the number of pulls $T_{i,n}$; suggesting that a bound on the pulls is sufficient to bound the regret.

Finally, we can also introduce a definition of the pseudo-regret.

Definition 3. *The pseudo regret for a learning algorithm $\mathcal{A}$ over n rounds is defined as*

$$\widetilde{\mathcal{R}}_n(\mathcal{A}) = \frac{1}{n}\sum_{i\neq i^*} T_{i,n}\Delta_i + \frac{2}{n^2}\sum_{i=1}^{K}\sum_{j\neq i} T_{i,n}T_{j,n}\Gamma^2_{i,j}, \tag{8}$$

where $\Delta_i = \mathrm{MV}_i - \mathrm{MV}_{i^*}$ and $\Gamma_{i,j} = \mu_i - \mu_j$.

In the following we will denote the two components of the pseudo–regret as

$$\widetilde{\mathcal{R}}^{\Delta}_n(\mathcal{A}) = \frac{1}{n}\sum_{i\neq i^*} T_{i,n}\Delta_i, \quad \text{and} \quad \widetilde{\mathcal{R}}^{\Gamma}_n(\mathcal{A}) = \frac{2}{n^2}\sum_{i=1}^{K}\sum_{j\neq i} T_{i,n}T_{j,n}\Gamma^2_{i,j}. \tag{9}$$

Where $\widetilde{\mathcal{R}}^{\Delta}_n(\mathcal{A})$ constitutes the standard regret derived from the traditional formulation of the multi-arm bandit problem and $\widetilde{\mathcal{R}}^{\Gamma}_n(\mathcal{A})$ denotes the exploration risk. This regret can be shown to be close to the true regret up to small terms with high probability.

Lemma 1. *Given definitions 2 and 3,*

$$\mathcal{R}_n(\mathcal{A}) \leq \widetilde{\mathcal{R}}_n(\mathcal{A}) + (5+\rho)\sqrt{\frac{2K\log 6nK/\delta}{n}} + 4\sqrt{2}\frac{K\log 6nK/\delta}{n},$$

with probability at least $1-\delta$.

The previous lemma shows that any (high–probability) bound on the pseudo–regret immediately translates into a bound on the true regret. Thus, in the following we will report most of the theoretical analysis according to $\widetilde{\mathcal{R}}_n(\mathcal{A})$. Nonetheless, it is interesting to notice a major difference in the relationship between the true and pseudo–regret here and in the standard bandit problem. In fact, it is possible to show that, in this case, the pseudo–regret is not an unbiased estimator of the true regret, i.e., $\mathbb{E}[\mathcal{R}_n] \neq \mathbb{E}[\widetilde{\mathcal{R}}_n]$. Thus, in order to bound the expectation of $\mathcal{R}_n$ we need to build on the high–probability result from Lemma 1.

3 The Mean–Variance Lower Confidence Bound Algorithm

In this section we introduce a novel risk–averse bandit algorithm whose objective is to identify the arm which best trades off risk and return. The algorithm is a natural extension of *UCB1* [5] and we report a theoretical performance analysis on how well it balances the exploration needed to identify the best arm versus the risk of pulling arms with different means.

3.1 The Algorithm

We propose an index–based bandit algorithm which estimates the mean–variance of each arm and selects the optimal arm according to the optimistic confidence–bounds on the current estimates. A sketch of the algorithm is reported in Figure 1. For each arm, the algorithm keeps track of the empirical mean–variance $\widehat{\mathrm{MV}}_{i,s}$ computed according to s samples. We can build high–probability confidence bounds as an immediate application of the Chernoff–Hoeffding inequality (see e.g., [1] for the bound on the variance) for terms $\hat{\mu}$ and $\hat{\sigma}^2$ in the empirical mean–variance.

Input: Confidence δ
for $t = 1, \dots, n$ **do**
for $i = 1, \dots, K$ **do**
Compute $B_{i,T_{i,t-1}} = \widehat{\text{MV}}_{i,T_{i,t-1}} - (5+\rho)\sqrt{\frac{\log 1/\delta}{2T_{i,t-1}}}$
end for
Return $I_t = \arg\min_{i=1,\dots,K} B_{i,T_{i,t-1}}$
Update $T_{i,t} = T_{i,t-1} + 1$
Observe $X_{I_t,T_{i,t}} \sim \nu_{I_t}$
Update $\widehat{\text{MV}}_{i,T_{i,t}}$
end for

Figure 1: Pseudo-code of the *MV-LCB* algorithm.

Lemma 2. *Let $\{X_{i,s}\}$ be i.i.d. random variables bounded in $[0,1]$ from the distribution ν_i with mean μ_i and variance σ_i^2, and $\hat{\mu}_{i,s}$ and $\hat{\sigma}^2_{i,s}$ be the empirical mean and variance computed as in Equation 1, then*

$$\mathbb{P}\left[\exists i = 1, \dots, K, s = 1, \dots, n, |\widehat{\text{MV}}_{i,s} - \text{MV}_i| \geq (5+\rho)\sqrt{\frac{\log 1/\delta}{2s}}\right] \leq 6nK\delta,$$

The algorithm in Figure 1 implements the principle of optimism in the face of uncertainty used in some multi–arm bandit algorithms. On the basis of the previous confidence bounds, we define a lower–confidence bound on the mean–variance of arm i when it has been pulled s times as

$$B_{i,s} = \widehat{\text{MV}}_{i,s} - (5+\rho)\sqrt{\frac{\log 1/\delta}{2s}}, \tag{10}$$

where δ is an input parameter of the algorithm. Given the index of each arm at each round t, the algorithm simply selects the arm with the smallest mean–variance index, i.e., $I_t = \arg\min_i B_{i,T_{i,t-1}}$. We refer to this algorithm as the mean–variance lower–confidence bound (*MV-LCB*) algorithm.

Remark 1. We notice that the algorithm reduces to *UCB1* whenever $\rho \to \infty$. This is coherent with the fact that for $\rho \to \infty$ the mean–variance problem reduces to the maximization of the cumulative reward, for which *UCB1* is already known to be nearly-optimal. On the other hand, for $\rho = 0$, which leads to the problem of cumulative reward variance minimization, the algorithm plays according to a lower–confidence–bound on the variances.

Remark 2. The *MV-LCB* algorithm is parameterized by a parameter δ which defines the confidence level of the bounds employed in the definition of the index (10). In Theorem 1 we show how to optimize the parameter when the horizon n is known in advance. On the other hand, if n is not known, it is possible to design an anytime version of *MV-LCB* by defining a non-decreasing exploration sequence $(\varepsilon_t)_t$ instead of the term $\log 1/\delta$.

3.2 Theoretical Analysis

In this section we report the analysis of the regret $\mathcal{R}_n(\mathcal{A})$ of *MV-LCB* (Fig. 1). As highlighted in eq. 7, it is enough to analyze the number of pulls for each of the arms to recover a bound on the regret. Although the proofs are reported in the supplementary material, we notice here that they are mostly based on similar arguments to the proof of *UCB*.

We derive the following regret bound in high probability and expectation.

Theorem 1. *Let the optimal arm i^* be unique and $b = 2(5+\rho)$, the MV-LCB algorithm achieves a pseudo–regret bounded as*

$$\widetilde{\mathcal{R}}_n(\mathcal{A}) \leq \frac{b^2 \log 1/\delta}{n}\left(\sum_{i \neq i^*} \frac{1}{\Delta_i} + 4\sum_{i \neq i^*} \frac{\Gamma^2_{i^*,i}}{\Delta_i^2} + \frac{2b^2 \log 1/\delta}{n} \sum_{i \neq i^*} \sum_{\substack{j \neq i \\ j \neq i^*}} \frac{\Gamma^2_{i,j}}{\Delta_i^2 \Delta_j^2}\right) + \frac{5K}{n},$$

with probability at least $1 - 6nK\delta$*. Similarly, if MV-LCB is run with* $\delta = 1/n^2$ *then*

$$\mathbb{E}[\widetilde{\mathcal{R}}_n(\mathcal{A})] \leq \frac{2b^2 \log n}{n}\Big(\sum_{i \neq i^*} \frac{1}{\Delta_i} + 4\sum_{i \neq i^*} \frac{\Gamma_{i^*,i}^2}{\Delta_i^2} + \frac{4b^2 \log n}{n} \sum_{i \neq i^*} \sum_{\substack{j \neq i \\ j \neq i^*}} \frac{\Gamma_{i,j}^2}{\Delta_i^2 \Delta_j^2}\Big) + (17 + 6\rho)\frac{K}{n}.$$

Remark 1 (the bound). Let $\Delta_{\min} = \min_{i \neq i^*} \Delta_i$ and $\Gamma_{\max} = \max_i |\Gamma_i|$, then a rough simplification of the previous bound leads to

$$\mathbb{E}[\widetilde{\mathcal{R}}_n(\mathcal{A})] \leq O\Big(\frac{K}{\Delta_{\min}} \frac{\log n}{n} + K^2 \frac{\Gamma_{\max}^2}{\Delta_{\min}^4} \frac{\log^2 n}{n}\Big).$$

First we notice that the regret decreases as $O(\log^2 n/n)$, implying that *MV-LCB* is a consistent algorithm. As already highlighted in Definition 2, the regret is composed by two main terms. The first term is due to the difference in the mean–variance of the best arm and the arms pulled by the algorithm, while the second term denotes the additional variance introduced by the exploration risk of pulling arms with different means. In particular, it is interesting to notice that this additional term depends on the squared difference in the means of the arms $\Gamma_{i,j}^2$. Thus, if all the arms have the same mean, this term would be zero.

Remark 2 (worst–case analysis). We can further study the result of Theorem 1 by considering the worst–case performance of *MV-LCB*, that is the performance when the distributions of the arms are chosen so as to maximize the regret. In order to illustrate our argument we consider the simple case of $K = 2$ arms, $\rho = 0$ (variance minimization), $\mu_1 \neq \mu_2$, and $\sigma_1^2 = \sigma_2^2 = 0$ (deterministic arms). [5] In this case we have a variance gap $\Delta = 0$ and $\Gamma^2 > 0$. According to the definition of *MV-LCB*, the index $B_{i,s}$ would simply reduce to $B_{i,s} = \sqrt{\frac{\log 1/\delta}{s}}$, thus forcing the algorithm to pull both arms uniformly (i.e., $T_{1,n} = T_{2,n} = n/2$ up to rounding effects). Since the arms have the same variance, there is no direct regret in pulling either one or the other. Nonetheless, the algorithm has an additional variance due to the difference in the samples drawn from distributions with different means. In this case, the algorithm suffers a constant (true) regret

$$\mathcal{R}_n(\textit{MV-LCB}) = 0 + \frac{T_{1,n}T_{2,n}}{n^2}\Gamma^2 = \frac{1}{4}\Gamma^2,$$

independent from the number of rounds n. This argument can be generalized to multiple arms and $\rho \neq 0$, since it is always possible to design an environment (i.e., a set of distributions) such that $\Delta_{\min} = 0$ and $\Gamma_{\max} \neq 0$. [6] This result is not surprising. In fact, two arms with the same mean–variance are likely to produce similar observations, thus leading *MV-LCB* to pull the two arms repeatedly over time, since the algorithm is designed to try to discriminate between similar arms. Although this behavior does not suffer from any regret in pulling the "suboptimal" arm (the two arms are equivalent), it does introduce an additional variance, due to the difference in the means of the arms ($\Gamma \neq 0$), which finally leads to a regret the algorithm is not "aware" of. This argument suggests that, for any n, it is always possible to design an environment for which *MV-LCB* has a constant regret. This finding will be further investigated in the numerical simulations in Section 5. This result is particularly interesting since it reveals a huge gap between the mean–variance problem and the standard expected regret minimization problem. In fact, in the latter case, *UCB* is known to have a worst–case regret per round of $\Omega(1/\sqrt{n})$ [4], while in the worst case, *MV-LCB* suffers a constant regret. In the next section we introduce a simple algorithm able to deal with this problem and achieve a vanishing worst–case regret.

4 The Exploration–Exploitation Algorithm

Although for any fixed problem (with $\Delta_{\min} > 0$) the *MV-LCB* algorithm introduced in the previous section has a vanishing regret, for any value of n, it is always possible to find an environment for which its regret is constant. In this section, we analyze a simple algorithm where exploration and exploitation are two distinct phases. The *ExpExp* algorithm divides the time horizon n into two distinct phases of length τ and $n - \tau$ respectively. During the first phase all the arms are explored

[5]Note that in this case (i.e., $\Delta = 0$), Theorem 1 does not hold, since the optimal arm is not unique.

[6]Notice that this is always possible for a large majority of distributions for which the mean and variance are independent or mildly correlated.

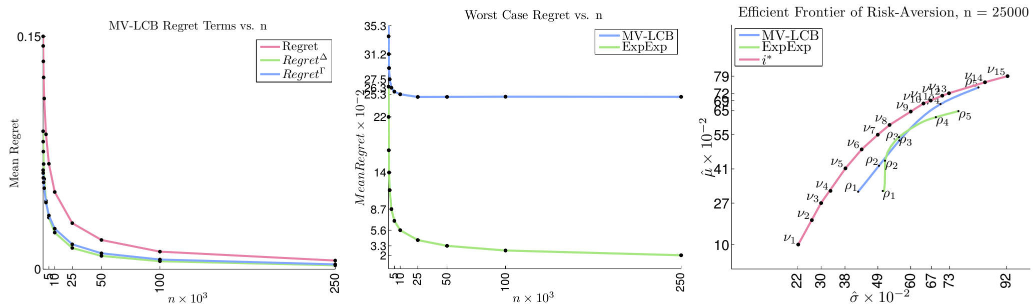

Figure 2: Regret of *MV-LCB* and *ExpExp* in different scenarios.

uniformly, thus collecting τ/K samples each [7]. Once the exploration phase is over, the mean–variance of each arm is computed and the arm with the smallest estimated mean–variance $\mathrm{MV}_{i,\tau/K}$ is repeatedly pulled until the end of the experiment.

The *MV-LCB* is specifically designed to minimize the probability of pulling the wrong arms, so whenever there are two equivalent arms (i.e., arms with the same mean–variance), the algorithm tends to pull them the same number of times, at the cost of potentially introducing an additional variance which might result in a constant regret. On the other hand, *ExpExp* stops exploring the arms after τ rounds and then elicits one arm as the best and keeps pulling it for the remaining $n-\tau$ rounds. Intuitively, the parameter τ should be tuned so as to meet different requirements. The first part of the regret (i.e., the regret coming from pulling the suboptimal arms) suggests that the exploration phase τ should be long enough for the algorithm to select the empirically best arm $\hat{i}^*$ at τ equivalent to the actual optimal arm i^* with high probability; and at the same time, as short as possible to reduce the number of times the suboptimal arms are explored. On the other hand, the second part of the regret (i.e., the variance of pulling arms with different means) is minimized by taking τ as small as possible (e.g., $\tau = 0$ would guarantee a zero regret). The following theorem illustrates the optimal trade-off between these contrasting needs.

Theorem 2. *Let ExpExp be run with $\tau = K(n/14)^{2/3}$, then for any choice of distributions $\{\nu_i\}$ the expected regret is $\mathbb{E}[\widetilde{\mathcal{R}}_n(\mathcal{A})] \le 2\frac{K}{n^{1/3}}$.*

Remark 1 (the bound). We first notice that this bound suggests that *ExpExp* performs worse than *MV-LCB* on easy problems. In fact, Theorem 1 demonstrates that *MV-LCB* has a regret decreasing as $O(K\log(n)/n)$ whenever the gaps Δ are not small compared to n. Nonetheless, the previous bound is distribution independent and indicates the worst performance possible with *ExpExp*. On the other hand, in the remarks of Theorem 1 we highlighted the fact that for any value of n, it is always possible to design an environment which leads *MV-LCB* to suffer a constant regret. This opens the question whether it is possible to design an algorithm which works as well as *MV-LCB* on easy problems and as robustly as *ExpExp* on difficult problems.

Remark 2 (exploration phase). The previous result can be improved by changing the exploration strategy used in the first τ rounds. Instead of a pure uniform exploration of all the arms, we could adopt a best–arm identification algorithms such as *Successive Reject* and *UCB-E* which maximize the probability of returning the best arm given a fixed budget of rounds τ (see e.g., [2]).

5 Numerical Simulations

In this section we report numerical simulations aimed at validating the main theoretical findings reported in the previous sections. In all the following graphs we study the true regret $\mathcal{R}_n(\mathcal{A})$ averaged over 500 runs. We first consider the variance minimization problem ($\rho = 0$) for $K = 2$ Gaussian arms with $\mu_1 = 1.0$, $\mu_2 = 0.5$, $\sigma_1^2 = 0.05$, and $\sigma_2^2 = 0.25$ and we run *MV-LCB*. [8] In Figure 2 we report the true regret $\mathcal{R}_n$ (as in the original definition in eq. 4) and its two components $\mathcal{R}_n^{\widehat{\Delta}}$ and $\mathcal{R}_n^{\widehat{\Gamma}}$ (these two values are defined as in eq. 9 with $\widehat{\Delta}$ and $\widehat{\Gamma}$ replacing Δ and Γ). As expected (see e.g., Theorem 1), the regret tends to zero as n increases and it is obtained as the composition of the regret

[7] In the definition and in the following analysis we ignore rounding effects

[8] Notice that although in the paper we assumed the distributions to be bounded in $[0, 1]$ all the results can be extended to sub-Gaussian distributions.

from pulling suboptimal arms and the regret of pulling arms with different means. Indeed, if we considered two distributions with $\mu_1 = \mu_2$, the average regret would coincide with $\mathcal{R}_n^{\widehat{\Delta}}$. Furthermore, as shown in Theorem 1 the two regret terms decrease with the same rate $O(\log n/n)$.

A detailed analysis of the impact of Δ and Γ on the performance of *MV-LCB* is reported in the supplementary material (Appendix D). Here we only report the study of the worst–case performance of *MV-LCB* and and we compare it to *ExpExp* (see Figure 2). In order to have a fair comparison, for any value of n and for each of the two algorithms, we select the pair Δ_w, Γ_w which corresponds to the largest regret (we search in a grid of values with $\mu_1 = 1.5$, $\mu_2 \in [0.4; 1.5]$, $\sigma_1^2 \in [0.0; 0.25]$, and $\sigma_2^2 = 0.25$, so that $\Delta \in [0.0; 0.25]$ and $\Gamma \in [0.0; 1.1]$). As discussed in Section 4, while the worst–case regret of *ExpExp* keeps decreasing over n, it is always possible to find a problem for which regret of *MV-LCB* stabilizes to a constant.

While in the previous experiments we considered the case of variance minimization, in Figure 2 we report results for a wide range of risk tolerance $\rho \in [0.0; 10.0]$ and $K = 15$ arms. We choose the means and variances so that a set of arms is always dominated (i.e., for any ρ, $\mathrm{MV}_i^\rho > \mathrm{MV}_{i_\rho^*}^\rho$), while the optimal arm i_ρ^* changes depending on the value of ρ. In Figure 2 we arranged the arms and the algorithms performance in a standard deviation–mean plot. While the red line connects the arms that are optimal for some value of ρ, the green and blue lines show the standard deviations and means of *ExpExp* and *MV-LCB* for $n = 25,000$. Each point on the two lines corresponds to the performance of different values of ρ. We notice that in this problem, where a lot of arms have big gaps (e.g., all the dominated arms have a large gap for any value of ρ), *MV-LCB* tends to perform better than *ExpExp*. In Appendix D we report additional results.

6 Conclusions

Large part of the literature in multi–armed bandit focuses on the problem of minimizing the regret w.r.t. the arm with the highest return. Nonetheless, this is not always the best option, since the optimal arm in expectation may have a large risk. In this paper, we introduced a novel multi–armed setting where the objective is to perform as well as the arm with the best risk–return trade–off. In particular, we relied on the mean–variance model introduced in [8] to measure the performance of the arms and we defined the regret of a learning algorithm. We proposed two novel algorithms to solve the mean–variance bandit problem and we reported their corresponding theoretical analysis. While *MV-LCB* shows a small regret of order $O(\log n/n)$ on "easy" problems (i.e., where the mean–variance gaps Δ are big w.r.t. n), we showed that it has a constant worst–case regret. On the other hand, we proved that *ExpExp* have a vanishing worst–case regret at the cost of a worse performance on the "easy" problems. To the best of our knowledge this is the first work introducing risk–aversion in the multi–armed setting and it opens a series of interesting questions.

Lower bound. In this paper we introduced two algorithms, *MV-LCB* and *ExpExp*. As discussed in remarks of Theorem 1 and of Theorem 2, *MV-LCB* has a regret of order $O(\sqrt{K/n})$ on easy problems and $O(1)$ on difficult problems, while *ExpExp* achieves the same regret $O(K/n^{1/3})$ over all the problems. The main open question is whether $O(K/n^{1/3})$ is actually the best possible achievable rate (in the worst–case) for this problem or a better rate is possible. This question is of particular interest since the standard reward expectation maximization problem has a known lower–bound of $\Omega(\sqrt{1/n})$ and minimax rate of $\Omega(1/n^{1/3})$ for the mean–variance problem; implying that the risk–averse bandit problem is intrinsically more difficult than standard bandit problems.

Different measures of return–risk. Considering alternative notions of risk is a natural extension to the previous setting. In fact, over the years the mean–variance model has often been criticized. From a point of view of the expected utility theory, the mean–variance model is justified only under a Gaussianity assumption on the arm distributions. Furthermore, the variance is a symmetric measure of risk, while it is often the case that only one–sided deviations from the mean are not desirable (e.g., in finance only losses w.r.t. to the expected return are considered as a risk, while any positive deviation is not considered as a real risk). A popular measure of risk–return is the α value–at–risk (i.e., the quantile). The main challenge in this case is the estimation of the value–at–risk for each arm. In fact, while the cumulative distribution of a random variable can be reliably estimated (see e.g., [9]), the quantile is much more difficult, in particular when the α level corresponds to values where the probability density is close to zero (e.g., a 0.95 quantile for a Gaussian distribution). Thus, unlike the standard case where we consider either a bounded or sub-Gaussian distribution, it would be preferable to deal with distributions with fat tails.

References

[1] András Antos, Varun Grover, and Csaba Szepesvári. Active learning in heteroscedastic noise. *Theoretical Computer Science*, 411:2712–2728, June 2010.

[2] J.-Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *Proceedings of the Twenty-third Conference on Learning Theory (COLT'10)*, 2010.

[3] J-Y. Audibert, R. Munos, and Cs Szepesvari. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410:1876–1902, 2009.

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

[5] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47:235–256, 2002.

[6] Eyal Even-Dar, Michael Kearns, and Jennifer Wortman. Risk-sensitive online learning. In *Proceedings of the 17th international conference on Algorithmic Learning Theory (ALT'06)*, pages 199–213, 2006.

[7] Christian Gollier. *The Economics of Risk and Time*. The MIT Press, 2001.

[8] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[9] P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, 18(3):pp. 1269–1283, 1990.

[10] A. Maurer and M. Pontil. Empirical bernstein bounds and sample-variance penalization. In *Proceedings of the Twenty-Second Annual Conference on Learning Theory*, pages 115–124, 2009.

[11] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the AMS*, 58:527–535, 1952.

[12] Antoine Salomon and Jean-Yves Audibert. Deviations of stochastic bandit regret. In *Proceedings of the 22nd international conference on Algorithmic learning theory (ALT'11)*, pages 159–173, 2011.

[13] Manfred K. Warmuth and Dima Kuzmin. Online variance minimization. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT'06)*, pages 514–528, 2006.

A The Regret

A.1 The True Regret

We recall the definition of the (empirical) regret as

$$\mathcal{R}_n(\mathcal{A}) = \widehat{\mathrm{MV}}_n(\mathcal{A}) - \widehat{\mathrm{MV}}_{i^*,n}.$$

Given the definitions reported in the main paper, we first elaborate on the two mean terms in the regret as

$$\hat{\mu}_{i^*,n} = \frac{1}{n}\sum_{i=1}^{K}\sum_{t=1}^{T_{i,n}} Y_{i,t} = \frac{1}{n}\sum_{i=1}^{K} T_{i,n}\tilde{\mu}_{i,T_{i,n}},$$

and

$$\hat{\mu}_n(\mathcal{A}) = \frac{1}{n}\sum_{i=1}^{K}\sum_{t=1}^{T_{i,n}} X_{i,t} = \frac{1}{n}\sum_{i=1}^{K} T_{i,n}\hat{\mu}_{i,T_{i,n}}.$$

Similarly, the two variance terms can be written as

$$\begin{aligned}
\hat{\sigma}^2_n(\mathcal{A}) &= \frac{1}{n}\sum_{i=1}^{K}\sum_{t=1}^{T_{i,n}} \big(X_{i,t} - \hat{\mu}_n(\mathcal{A})\big)^2 \\
&= \frac{1}{n}\sum_{i=1}^{K}\sum_{t=1}^{T_{i,n}} \big(X_{i,t} - \hat{\mu}_{i,T_{i,n}}\big)^2 + \frac{1}{n}\sum_{i=1}^{K}\sum_{t=1}^{T_{i,n}} \big(\hat{\mu}_{i,T_{i,n}} - \hat{\mu}_n(\mathcal{A})\big)^2 + \frac{2}{n}\sum_{i=1}^{K}\sum_{t=1}^{T_{i,n}} \big(X_{i,t} - \hat{\mu}_{i,T_{i,n}}\big)\big(\hat{\mu}_{i,T_{i,n}} - \hat{\mu}_n(\mathcal{A})\big) \\
&= \frac{1}{n}\sum_{i=1}^{K} T_{i,n}\hat{\sigma}^2_{i,T_{i,n}} + \frac{1}{n}\sum_{i=1}^{K} T_{i,n}\big(\hat{\mu}_{i,T_{i,n}} - \hat{\mu}_n(\mathcal{A})\big)^2 + 0,
\end{aligned}$$

and

$$\begin{aligned}
\sigma^2_{i^*,n} &= \frac{1}{n}\sum_{i=1}^{K}\sum_{t=1}^{T_{i,n}} \big(Y_{i,t} - \hat{\mu}_{i^*,n}\big)^2 \\
&= \frac{1}{n}\sum_{i=1}^{K}\sum_{t=1}^{T_{i,n}} \big(Y_{i,t} - \tilde{\mu}_{i,T_{i,n}}\big)^2 + \frac{1}{n}\sum_{i=1}^{K}\sum_{t=1}^{T_{i,n}} \big(\tilde{\mu}_{i,T_{i,n}} - \hat{\mu}_{i^*,n}\big)^2 + \frac{2}{n}\sum_{i=1}^{K}\sum_{t=1}^{T_{i,n}} \big(Y_{i,t} - \tilde{\mu}_{i,T_{i,n}}\big)\big(\tilde{\mu}_{i,T_{i,n}} - \hat{\mu}_{i^*,n}\big) \\
&= \frac{1}{n}\sum_{i=1}^{K} T_{i,t}\tilde{\sigma}^2_{i,T_{i,n}} + \frac{1}{n}\sum_{i=1}^{K} T_{i,n}\big(\tilde{\mu}_{i,T_{i,n}} - \hat{\mu}_{i^*,n}\big)^2 + 0.
\end{aligned}$$

Putting together these terms, we obtain the regret (see eq. 4)

$$\begin{aligned}
\mathcal{R}_n(\mathcal{A}) = \frac{1}{n}\sum_{i\neq i^*} T_{i,n}\Big[(\hat{\sigma}^2_{i,T_{i,n}} - \tilde{\sigma}^2_{i,T_{i,n}}) - \rho(\hat{\mu}_{i,T_{i,n}} - \tilde{\mu}_{i,T_{i,n}})\Big] \\
+ \frac{1}{n}\sum_{i=1}^{K} T_{i,n}\big(\hat{\mu}_{i,T_{i,n}} - \hat{\mu}_n(\mathcal{A})\big)^2 - \frac{1}{n}\sum_{i=1}^{K} T_{i,n}\big(\tilde{\mu}_{i,T_{i,n}} - \hat{\mu}_{i^*,n}\big)^2
\end{aligned}$$

If we further elaborate the second term, we obtain

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{K} T_{i,n}\big(\hat{\mu}_{i,T_{i,n}} - \hat{\mu}_n(\mathcal{A})\big)^2 &= \frac{1}{n}\sum_{i=1}^{K} T_{i,n}\Big(\hat{\mu}_{i,T_{i,n}} - \frac{1}{n}\sum_{j=1}^{K} T_{j,n}\hat{\mu}_{j,T_{j,n}}\Big)^2 \\ &= \frac{1}{n}\sum_{i=1}^{K} T_{i,n}\Big(\sum_{j=1}^{K}\frac{T_{j,n}}{n}(\hat{\mu}_{i,T_{i,n}} - \hat{\mu}_{j,T_{j,n}})\Big)^2 \\ &\le \frac{1}{n}\sum_{i=1}^{K} T_{i,n}\sum_{j=1}^{K}\frac{T_{j,n}}{n}(\hat{\mu}_{i,T_{i,n}} - \hat{\mu}_{j,T_{j,n}})^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{K}\sum_{j\neq i} T_{i,n}T_{j,n}(\hat{\mu}_{i,T_{i,n}} - \hat{\mu}_{j,T_{j,n}})^2.\end{aligned}$$

Using the definitions $\widehat{\Delta}_i = (\hat{\sigma}^2_{i,T_{i,n}} - \tilde{\sigma}^2_{i,T_{i,n}}) - \rho(\hat{\mu}_{i,T_{i,n}} - \tilde{\mu}_{i,T_{i,n}})$ and $\widehat{\Gamma}^2_{i,j} = (\hat{\mu}_{i,T_{i,n}} - \hat{\mu}_{j,T_{j,n}})^2$ we finally obtain an upper–bound on the regret of the form

$$\mathcal{R}_n(\mathcal{A}) \le \frac{1}{n}\sum_{i\neq i^*} T_{i,n}\widehat{\Delta}_i + \frac{1}{n^2}\sum_{i=1}^{K}\sum_{j\neq i} T_{i,n}T_{j,n}\widehat{\Gamma}^2_{i,j}.$$

In the following we refer to the two terms as $\mathcal{R}_n^{\widehat{\Delta}}$ and $\mathcal{R}_n^{\widehat{\Gamma}}$.

A.2 The Pseudo–Regret

Similar to what is done in the standard bandit problem, we can introduce a different notion of regret. Starting from the last equation in the previous section, we define the pseudo–regret

$$\widetilde{\mathcal{R}}_n(\mathcal{A}) = \frac{1}{n}\sum_{i\neq i^*} T_{i,n}\Delta_i + \frac{2}{n^2}\sum_{i=1}^{K}\sum_{j\neq i} T_{i,n}T_{j,n}\Gamma^2_{i,j},$$

where the empirical values $\widehat{\Delta}_i$ and $\widehat{\Gamma}_{i,j}$ are substituted by their corresponding exact values [9]. In the following we show that the true and pseudo regrets different for values that tend to zero with high probability.

Proof. (Lemma 1)

We define a high–probability event in which the empirical values and the true values only differ for small quantities

$$\mathcal{E} = \Big\{\forall i = 1,\dots,K,\ \forall s = 1,\dots,n,\ \ \big|\hat{\mu}_{i,s} - \mu_i\big| \le \sqrt{\frac{\log 1/\delta}{2s}}\ \text{ and }\ \big|\hat{\sigma}^2_{i,s} - \sigma^2_i\big| \le 5\sqrt{\frac{\log 1/\delta}{2s}}\Big\}.$$

Using Chernoff–Hoeffding inequality and a union bound over arms and rounds, we have that $\mathbb{P}[\mathcal{E}^C] \le 6nK\delta$. Under this event we rewrite the empirical $\widehat{\Delta}_i$ as

$$\begin{aligned}\widehat{\Delta}_i &= \Delta_i - (\sigma^2_i - \sigma^2_{i^*}) + \rho(\mu_i - \mu_{i^*}) + (\hat{\sigma}^2_{i,T_{i,n}} - \tilde{\sigma}^2_{i,T_{i,n}}) - \rho(\hat{\mu}_{i,T_{i,n}} - \tilde{\mu}_{i,T_{i,n}}) \\ &\le \Delta_i + 2(5+\rho)\sqrt{\frac{\log 1/\delta}{2T_{i,n}}}.\end{aligned}$$

Similarly, $\widehat{\Gamma}_{i,j}$ is upper–bounded as

$$\begin{aligned}|\widehat{\Gamma}_{i,j}| &= |\Gamma_{i,j} - \mu_i + \mu_j + \hat{\mu}_{i,T_{i,n}} - \hat{\mu}_{j,T_{j,n}}| \\ &\le |\Gamma_{i,j}| + \sqrt{\frac{\log 1/\delta}{2T_{i,n}}} + \sqrt{\frac{\log 1/\delta}{2T_{j,n}}}.\end{aligned}$$

[9] Notice that the factor 2 in front of the second term is due to a rough upper bounding used in the proof of Lemma 1.

Thus the regret can be written as

$$
\begin{aligned}
\mathcal{R}_n(\mathcal{A}) &\le \frac{1}{n}\sum_{i\neq i^*} T_{i,n}\Big(\Delta_i + 2(5+\rho)\sqrt{\frac{\log 1/\delta}{2T_{i,n}}}\Big) + \frac{1}{n^2}\sum_{i=1}^{K}\sum_{j\neq i} T_{i,n}T_{j,n}\Big(|\Gamma_{i,j}| + \sqrt{\frac{\log 1/\delta}{2T_{i,n}}} + \sqrt{\frac{\log 1/\delta}{2T_{j,n}}}\Big)^2 \\
&\le \frac{1}{n}\sum_{i\neq i^*} T_{i,n}\Delta_i + \frac{5+\rho}{n}\sum_{i\neq i^*}\sqrt{2T_{i,n}\log 1/\delta} + \frac{2}{n^2}\sum_{i=1}^{K}\sum_{j\neq i} T_{i,n}T_{j,n}\Gamma_{i,j}^2 \\
&\quad + \frac{2\sqrt{2}}{n^2}\sum_{i=1}^{K}\sum_{j\neq i} T_{j,n}\log 1/\delta + \frac{2\sqrt{2}}{n^2}\sum_{i=1}^{K}\sum_{j\neq i} T_{i,n}\log 1/\delta \\
&\le \frac{1}{n}\sum_{i\neq i^*} T_{i,n}\Delta_i + \frac{2}{n^2}\sum_{i=1}^{K}\sum_{j\neq i} T_{i,n}T_{j,n}\Gamma_{i,j}^2 + (5+\rho)\sqrt{\frac{2K\log 1/\delta}{n}} + 4\sqrt{2}\frac{K\log 1/\delta}{n}.
\end{aligned}
$$

where in the next to last passage we used Jensen's inequality for concave functions and rough upper bounds on other terms ($K-1<K$, $\sum_{i\neq i^*} T_{i,n}\le n$). By recalling the definition of $\widetilde{\mathcal{R}}_n(\mathcal{A})$ we finally obtain

$$
\mathcal{R}_n(\mathcal{A}) \le \widetilde{\mathcal{R}}_n(\mathcal{A}) + (5+\rho)\sqrt{\frac{2K\log 1/\delta}{n}} + 4\sqrt{2}\frac{K\log 1/\delta}{n},
$$

with probability $1-6nK\delta$. Thus we can conclude that any upper bound on the pseudo–regret $\widetilde{\mathcal{R}}_n(\mathcal{A})$ is a valid upper bound for the true regret $\mathcal{R}_n(\mathcal{A})$ as well, up to a decreasing term of order $O(\sqrt{K/n})$.

□

B *MV-LCB* Theoretical Analysis

In order to simplify the notation in the following we use $b = 2(5+\rho)$.

Proof. (Theorem 1)

We begin by defining a high–probability event $\mathcal{E}$ as

$$
\mathcal{E} = \Big\{\forall i=1,\dots,K,\ \forall s=1,\dots,n,\ \ \big|\hat{\mu}_{i,s}-\mu_i\big| \le \sqrt{\frac{\log 1/\delta}{2s}}\ \text{ and }\ \big|\hat{\sigma}^2_{i,s}-\sigma^2_i\big| \le 5\sqrt{\frac{\log 1/\delta}{2s}}\Big\}.
$$

Using Chernoff–Hoeffding inequality and a union bound over arms and rounds, we have that $\mathbb{P}[\mathcal{E}^C] \le 6nK\delta$.

We now introduce the definition of the algorithm. Let consider any time t when arm $i\neq i^*$ is pulled (i.e., $I_t = i$). By definition of the algorithm in Figure 1, i is selected if its corresponding index $B_{i,T_{i,t-1}}$ is bigger than for any other arm, notably the best arm i^*. By recalling the definition of the index and the empirical mean–variance at time t, we have

$$
\begin{aligned}
&\hat{\sigma}^2_{i,T_{i,t-1}} - \rho\hat{\mu}_{i,T_{i,t-1}} - (5+\rho)\sqrt{\frac{\log 1/\delta}{2T_{i,t-1}}} = B_{i,T_{i,t-1}} \le \\
&\qquad \le B_{i^*,T_{i^*,t-1}} = \hat{\sigma}^2_{i^*,T_{i^*,t-1}} - \rho\hat{\mu}_{i^*,T_{i^*,t-1}} - (5+\rho)\sqrt{\frac{\log 1/\delta}{2T_{i^*,t-1}}}.
\end{aligned}
$$

Over all the possible realizations, we now focus on the realizations in $\mathcal{E}$. In this case, we can rewrite the previous condition as

$$
\sigma^2_i - \rho\mu_i - 2(5+\rho)\sqrt{\frac{\log 1/\delta}{2T_{i,t-1}}} \le B_{i,T_{i,t-1}} \le B_{i^*,T_{i^*,t-1}} \le \sigma^2_{i^*} - \rho\mu_{i^*}.
$$

Let time t be the last time when arm i is pulled until the final round n, then $T_{i,t-1} = T_{i,n} - 1$ and

$$T_{i,n} \leq \frac{2(5+\rho)^2}{\Delta_i^2} \log \frac{1}{\delta} + 1,$$

which suggests that the suboptimal arms are pulled only few times with high probability. Plugging the bound in the regret in eq. 8 leads to the final statement

$$\widetilde{\mathcal{R}}_n(\mathcal{A}) \leq \frac{1}{n} \sum_{i \neq i^*} \frac{b^2 \log 1/\delta}{\Delta_i} + \frac{1}{n} \sum_{i \neq i^*} \frac{4b^2 \log 1/\delta}{\Delta_i^2} \Gamma_{i^*,i}^2 + \frac{1}{n^2} \sum_{i \neq i^*} \sum_{\substack{j \neq i \\ j \neq i^*}} \frac{2b^4 (\log 1/\delta)^2}{\Delta_i^2 \Delta_j^2} \Gamma_{i,j}^2 + \frac{5K}{n},$$

with probability $1 - 6nK\delta$.

We now move from the previous high–probability bound to a bound in expectation. The pseudo–regret is (roughly) bounded as $\widetilde{\mathcal{R}}_n(\mathcal{A}) \leq 2 + \rho$ (by bounding $\Delta_i \leq 1 + \rho$ and $\Gamma \leq 1$), thus

$$\mathbb{E}[\widetilde{\mathcal{R}}_n(\mathcal{A})] = \mathbb{E}\big[\widetilde{\mathcal{R}}_n(\mathcal{A}) \mathbb{I}\{\mathcal{E}\}\big] + \mathbb{E}\big[\widetilde{\mathcal{R}}_n(\mathcal{A}) \mathbb{I}\{\mathcal{E}^C\}\big] \leq (2+\rho)\mathbb{P}\big[\mathcal{E}^C\big].$$

By taking u equal to the previous high–probability bound and recalling that $\mathbb{P}[\mathcal{E}^C] \leq 6nK\delta$, we have

$$\mathbb{E}[\widetilde{\mathcal{R}}_n(\mathcal{A})] \leq \frac{1}{n} \sum_{i \neq i^*} \frac{b^2 \log 1/\delta}{\Delta_i} + \frac{1}{n} \sum_{i \neq i^*} \frac{4b^2 \log 1/\delta}{\Delta_i^2} \Gamma_{i^*,i}^2 + \frac{1}{n^2} \sum_{i \neq i^*} \sum_{\substack{j \neq i \\ j \neq i^*}} \frac{2b^4 (\log 1/\delta)^2}{\Delta_i^2 \Delta_j^2} \Gamma_{i,j}^2$$
$$+ \frac{5K}{n} + (2+\rho) 6nK\delta.$$

The final statement of the lemma follows by tuning the parameter $\delta = 1/n^2$ so as to have a regret bound decreasing with n. □

While a high–probability bound for $\mathcal{R}_n$ can be immediately obtained form Lemma 1, the expectation of $\mathcal{R}_n$ is reported in the next corollary.

Proof. Since the mean–variance $-\rho \leq \widehat{\text{MV}} \leq 1/4$, the regret is bounded by $-1/4 - \rho \leq \mathcal{R}_n(\mathcal{A}) \leq 1/4 + \rho$. Thus we have

$$\mathbb{E}[\mathcal{R}_n(\mathcal{A})] = \int_{-1/4-\rho}^{u} t f_t(t) dt + \int_{u}^{1/4+\rho} t f_t(t) dt \leq u \mathbb{P}[\mathcal{R}_n(\mathcal{A}) \leq u] + \Big(\frac{1}{4} + \rho\Big) \mathbb{P}[\mathcal{R}_n(\mathcal{A}) > u].$$

By taking u equal to the previous high–probability bound and recalling that $\mathbb{P}[\mathcal{E}^C] \leq 6nK\delta$, we have

$$\mathbb{E}[\mathcal{R}_n(\mathcal{A})] \leq \frac{1}{n} \sum_{i \neq i^*} \frac{b^2 \log 1/\delta}{\Delta_i} + \frac{1}{n} \sum_{i \neq i^*} \frac{4b^2 \log 1/\delta}{\Delta_i^2} \Gamma_{i^*,i}^2 + \frac{1}{n^2} \sum_{i \neq i^*} \sum_{\substack{j \neq i \\ j \neq i^*}} \frac{2b^4 (\log 1/\delta)^2}{\Delta_i^2 \Delta_j^2} \Gamma_{i,j}^2$$
$$+ \frac{5K}{n} + b\sqrt{\frac{K \log 1/\delta}{2n}} + 4\sqrt{2} \frac{K \log 1/\delta}{n} + \Big(\frac{1}{4} + \rho\Big) 6nK\delta.$$

The final statement of the lemma follows by tuning the parameter $\delta = 1/n^2$ so as to have a regret bound decreasing with n. □

C Exp–Exp Theoretical Analysis

The length of the exploration phase is τ and during the exploitation phase the algorithm keeps pulling the arm $\hat{i}^*$ with the smallest empirical variance estimated during the exploration phase. As a result, the number of pulls of each arm is

$$T_{i,n} = \frac{\tau}{K} + (n - \tau) \mathbb{I}\{i = \hat{i}^*\} \tag{11}$$

We analyze the two terms of the regret separately.

$$\widetilde{\mathcal{R}}_n^{\Delta} = \frac{1}{n}\sum_{i\neq i^*}\Big(\frac{\tau}{K} + (n-\tau)\mathbb{I}\{i=\hat{i}^*\}\Big)\Delta_i = \frac{\tau}{nK}\sum_{i\neq i^*}\Delta_i + \frac{n-\tau}{n}\sum_{i\neq i^*}\underbrace{\Delta_i\mathbb{I}\{i=\hat{i}^*\}}_{(c)}.$$

We notice that the only random variable in this formulation is the best arm $\hat{i}^*$ at the end of the exploration phase. We thus compute the expected value of $\widetilde{\mathcal{R}}_n^{\Delta}$.

$$\begin{aligned}
\mathbb{E}[(c)] &= \mathbb{P}[i=\hat{i}^*]\Delta_i = \mathbb{P}[\forall j\neq i,\ \hat{\sigma}_{i,1}^2 \leq \hat{\sigma}_{j,1}^2]\Delta_i \\
&\leq \mathbb{P}[\hat{\sigma}_{i,1}^2 \leq \hat{\sigma}_{i^*,1}^2]\Delta_i = \mathbb{P}\big[(\hat{\sigma}_{i,1}^2 - \sigma_i^2) + (\sigma_{i^*}^2 - \hat{\sigma}_{i^*,1}^2) \leq \Delta_i\big]\Delta_i \\
&\leq 2\Delta_i \exp\Big(-\frac{\tau}{K}\Delta_i^2\Big)
\end{aligned}$$

The second term in the regret can be bounded as follows.

$$\begin{aligned}
\widetilde{\mathcal{R}}_n^{\Gamma} &= \frac{1}{n^2}\sum_{i=1}^K\sum_{j\neq i}\Big(\frac{\tau}{K} + (n-\tau)\mathbb{I}\{i=\hat{i}^*\}\Big)\Big(\frac{\tau}{K} + (n-\tau)\mathbb{I}\{j=\hat{i}^*\}\Big)\Gamma_{i,j}^2 \\
&= \frac{1}{n^2}\sum_{i=1}^K\sum_{j\neq i}\Big(\frac{\tau^2}{K^2} + (n-\tau)^2\mathbb{I}\{i=\hat{i}^*\}\mathbb{I}\{j=\hat{i}^*\} + \frac{\tau}{K}(n-\tau)\mathbb{I}\{j=\hat{i}^*\} + \frac{\tau}{K}(n-\tau)\mathbb{I}\{i=\hat{i}^*\}\Big)\Gamma_{i,j}^2 \\
&= \frac{\tau^2}{n^2K^2}\sum_{i=1}^K\sum_{j\neq i}\Gamma_{i,j}^2 + 2\frac{(n-\tau)\tau}{Kn^2}\sum_{i=1}^K\sum_{j\neq i}\Gamma_{i,j}^2\mathbb{I}\{i=\hat{i}^*\} \\
&\leq \frac{\tau}{n^2} + 2\frac{(n-\tau)\tau}{n^2} \leq 2\frac{\tau}{n}
\end{aligned}$$

Grouping all the terms, *ExpExp* has an expected regret bounded as

$$\mathbb{E}[\widetilde{\mathcal{R}}_n(\mathcal{A})] \leq 2\frac{\tau}{n} + 2\sum_{i\neq i^*}\Delta_i\exp\Big(-\frac{\tau}{K}\Delta_i^2\Big)$$

We can now move to the worst–case analysis of the regret. Let $f(\Delta_i) = \Delta_i\exp\Big(-\frac{\tau}{K}\Delta_i^2\Big)$, the "adversarial" choice of the gap is determined by maximizing the regret and it corresponds to

$$\begin{aligned}
f'(\Delta_i) &= \exp\Big(-\frac{\tau}{K}\Delta_i^2\Big) + \Delta_i\bigg(-2\frac{\tau}{K}\Delta_i\exp\Big(-\frac{\tau}{K}\Delta_i^2\Big)\bigg) \\
&= \Big(1 - 2\frac{\tau}{K}\Delta_i^2\Big)\exp\Big(-\frac{\tau}{K}\Delta_i^2\Big) = 0,
\end{aligned}$$

which leads a worst–case choice of the gap as

$$\Delta_i = \sqrt{\frac{K}{2\tau}}.$$

The worst–case regret is then

$$\mathbb{E}[\widetilde{\mathcal{R}}_n(\mathcal{A})] \leq 2\frac{\tau}{n} + (K-1)\sqrt{2K}\frac{1}{\sqrt{\tau}}\exp(-0.5) \leq 2\frac{\tau}{n} + K^{3/2}\frac{1}{\sqrt{\tau}}$$

We can now choose the parameter τ minimizing the worst–case regret. Taking the derivative of the regret w.r.t. τ we obtain

$$\frac{d\mathbb{E}[\widetilde{\mathcal{R}}_n(\mathcal{A})]}{d\tau} = \frac{2}{n} - \frac{1}{2}\Big(\frac{K}{\tau}\Big)^{3/2} = 0,$$

thus leading to the optimal parameter $\tau = (n/4)^{2/3}K$. The final regret is thus bounded as

$$\mathbb{E}[\widetilde{\mathcal{R}}_n(\mathcal{A})] \leq 3\frac{K}{n^{1/3}}.$$

D Additional Simulations

D.1 Comparison between *MV-LCB* and *ExpExp* with $K = 2$

We consider the variance minimization problem ($\rho = 0$) with $K = 2$ Gaussian arms with different means and variances. In particular, we consider a grid of values with $\mu_1 = 1.5$, $\mu_2 \in [0.4; 1.5]$, $\sigma_1^2 \in [0.0; 0.25]$, and $\sigma_2^2 = 0.25$, so that $\Delta \in [0.0; 0.25]$ and $\Gamma \in [0.0; 1.1]$ and number of rounds $n \in [50; 2.5 \times 10^5]$. Figures 3 and 4 report the mean regret for different values of n. The colors are renormalized in each plot so that dark blue corresponds to the smallest regret and red to the largest regret. The results confirm the theoretical findings of Theorem 1 and 2. In fact, for simple problems (large gaps Δ) *MV-LCB* converges to a zero–regret faster than *ExpExp*, while for Δ close to zero (i.e., equivalent arms), *MV-LCB* has a constant regret which does not decrease with n and the regret of *ExpExp* slowly decreases to zero.

D.2 Risk tolerance sensitivity

In section we report numerical results for different values of the risk tolerance parameter ρ and $K = 15$ arms. We consider the two settings reported in Figure 7.

As we notice, in both configurations the performance of *MV-LCB* and *ExpExp* approaches the one of the optimal arm i^*_ρ for each specific ρ as n increases. Nonetheless, in configuration 1 the large number of suboptimal arms (e.g., arms with large gaps) allows *MV-LCB* to outperform *ExpExp* and converge faster to the optimal arm (and thus zero regret). On the other hand, in configuration 2 there are more arms with similar performance and for some values of ρ *ExpExp* eventually achieves a better performance than *MV-LCB*.

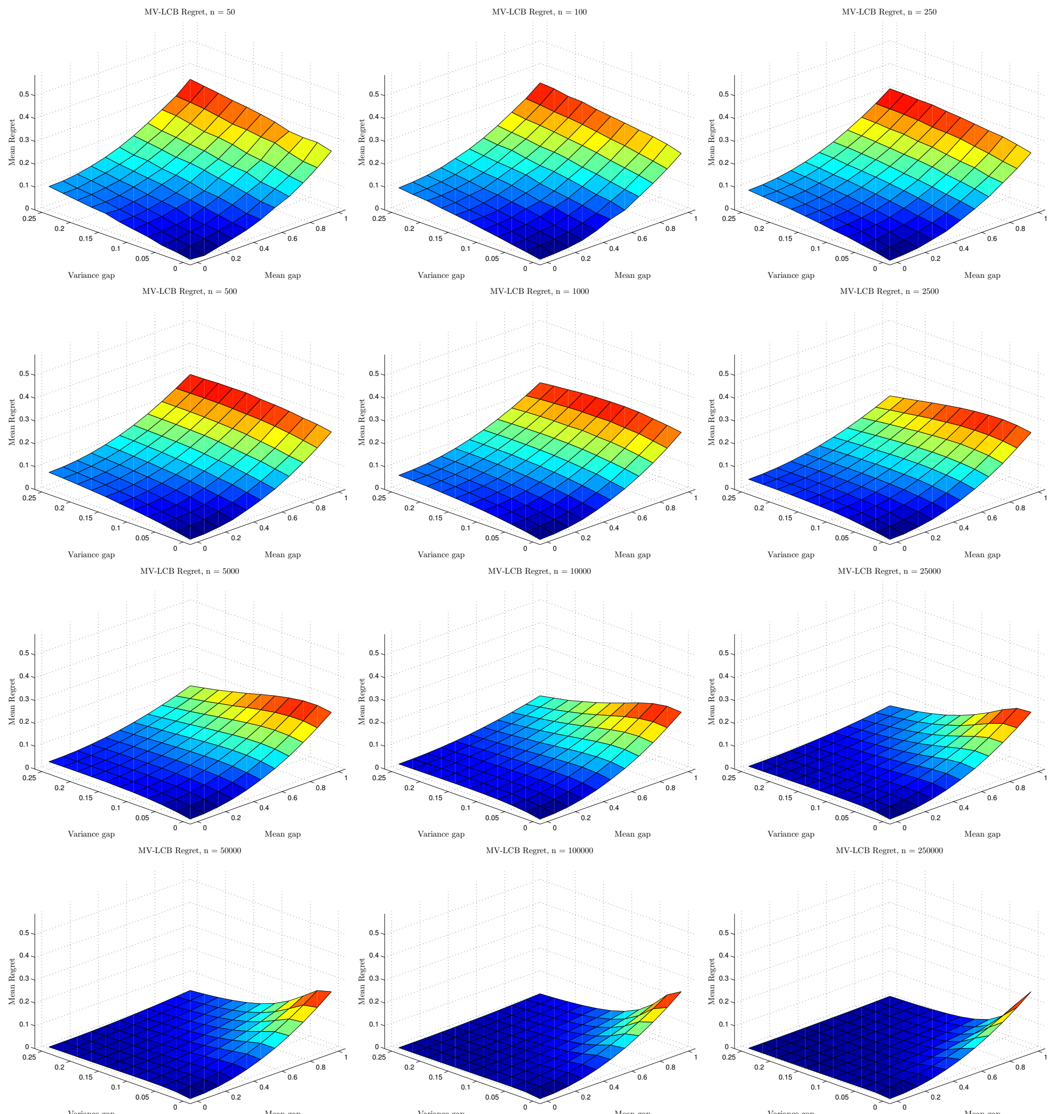

Figure 3: Regret $\mathcal{R}_n$ of *MV-LCB*.

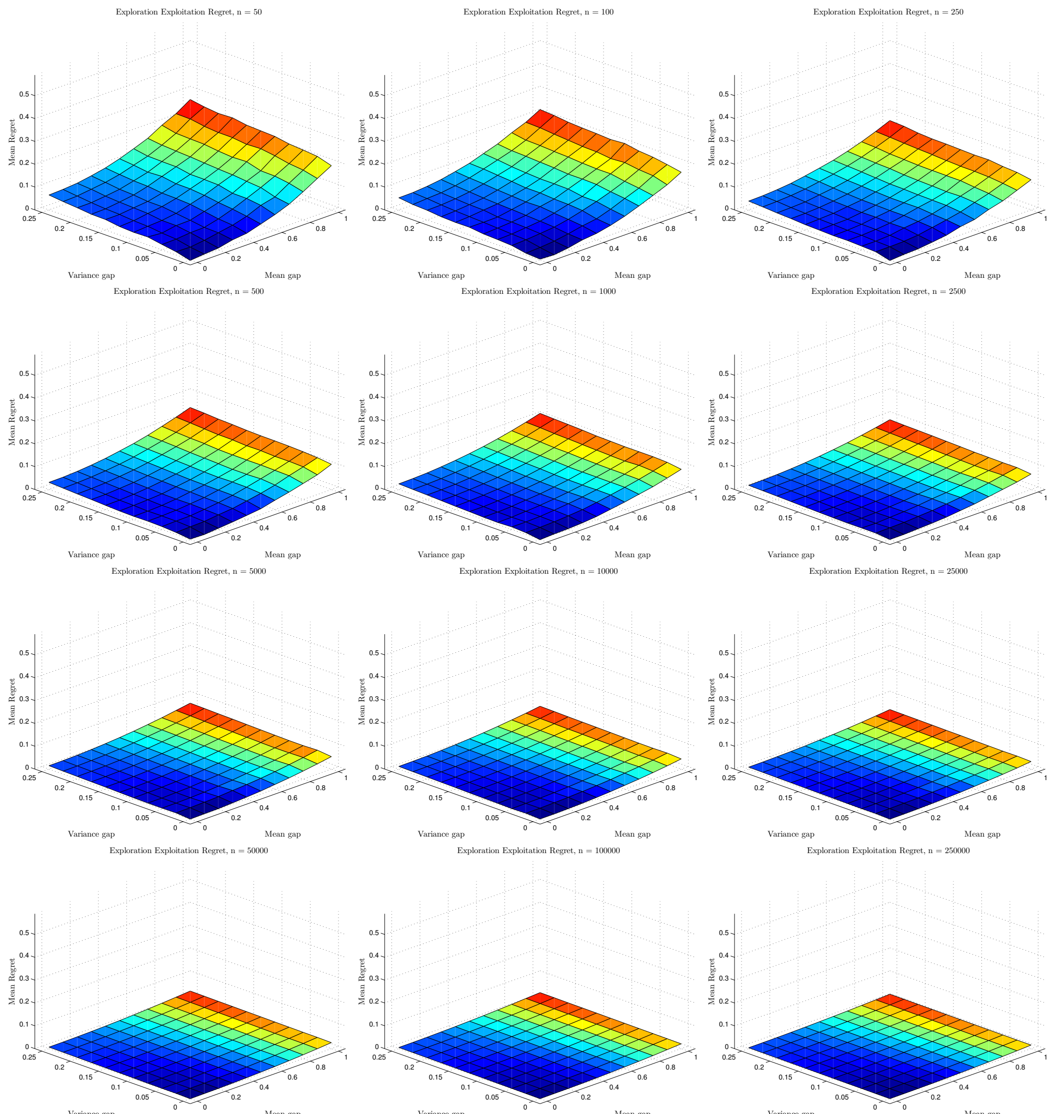

Figure 4: Regret $\mathcal{R}_n$ of *ExpExp*.

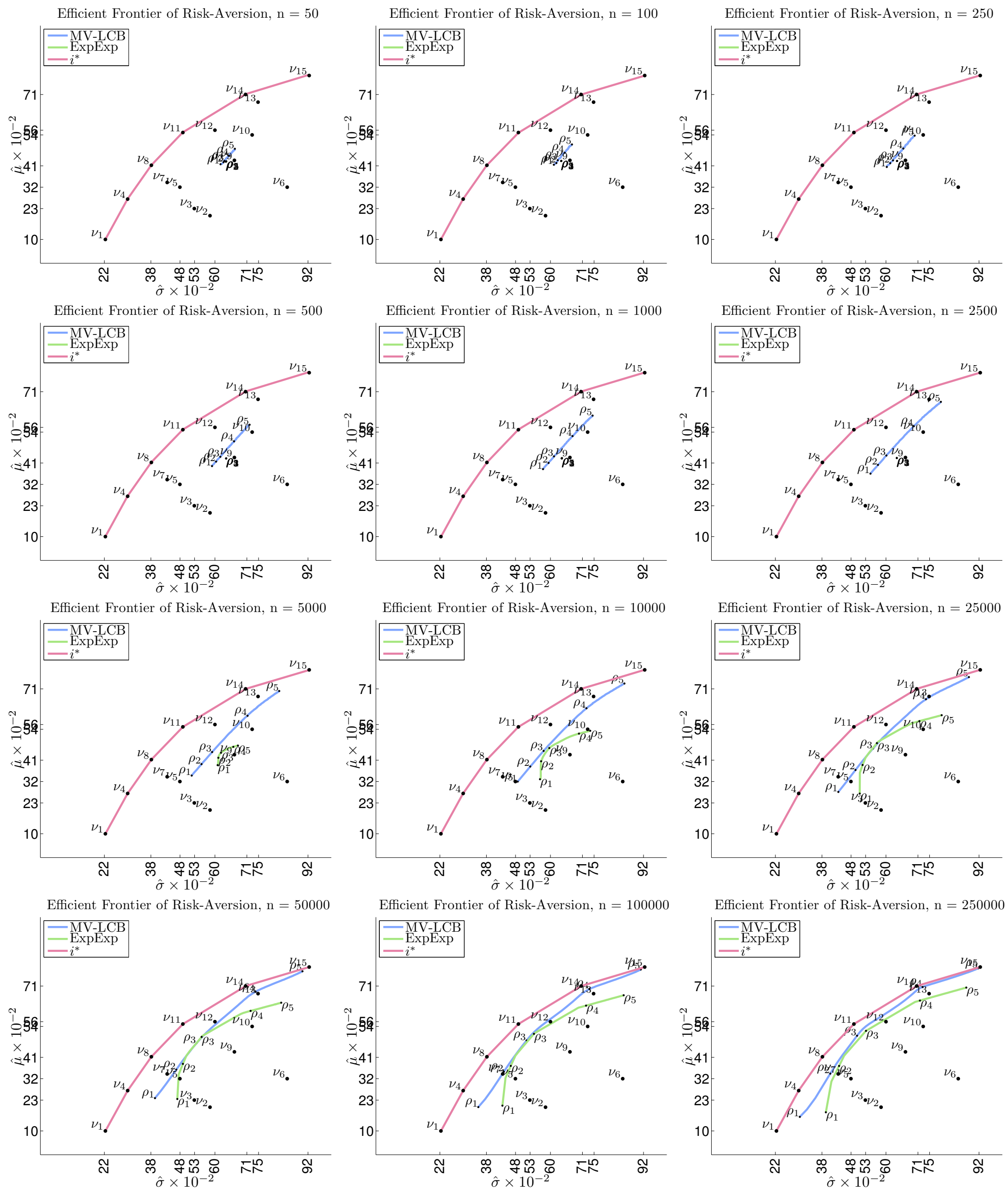

Figure 5: Risk tolerance sensitivity of *MV-LCB* and *ExpExp* for *configuration 1*.

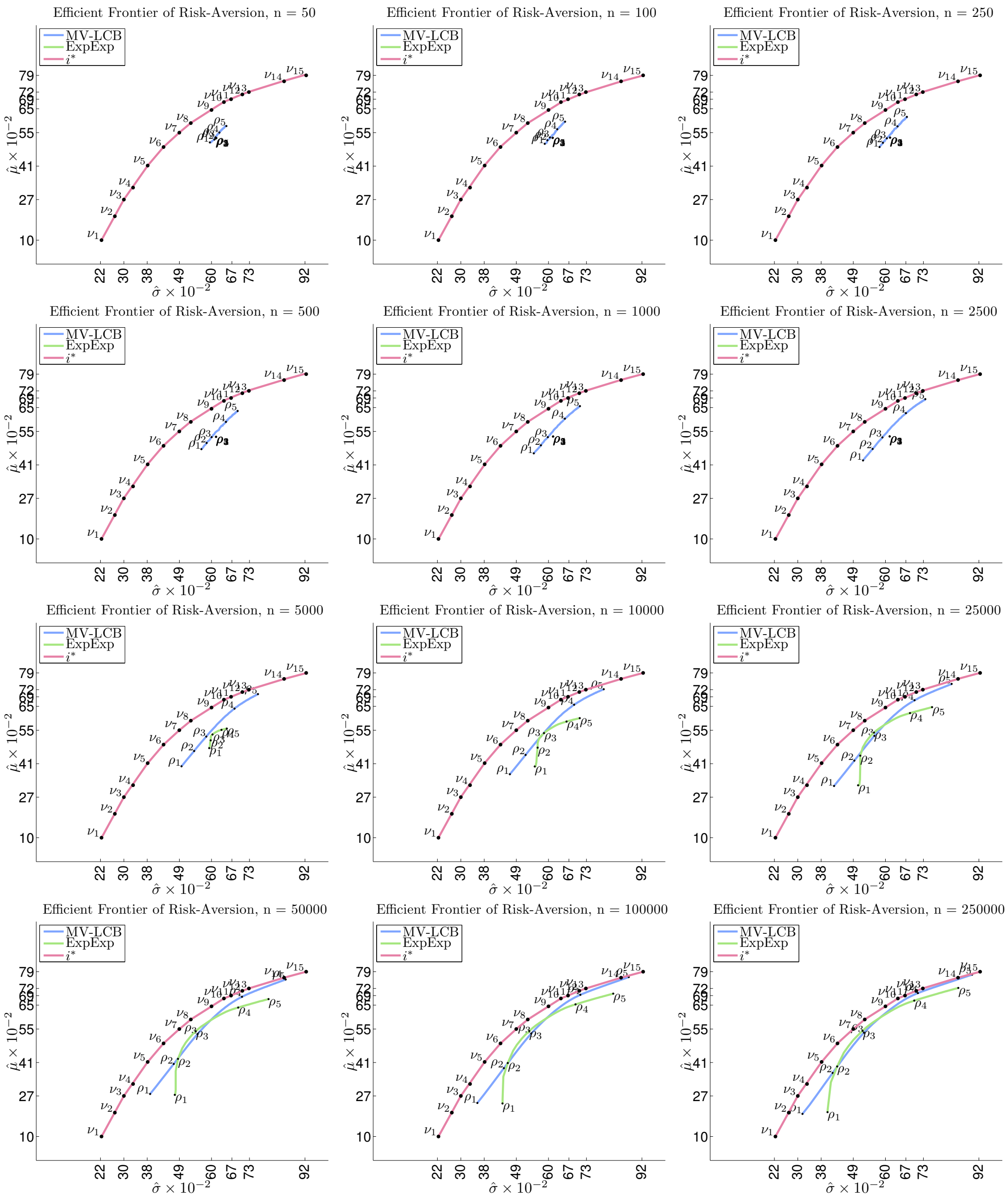

Figure 6: Risk tolerance sensitivity of *MV-LCB* and *ExpExp* for *configuration 2*.

	μ	σ^2
Arm 1	0.10	0.05
Arm 2	0.20	0.34
Arm 3	0.23	0.28
Arm 4	0.27	0.09
Arm 5	0.32	0.23
Arm 6	0.32	0.72
Arm 7	0.34	0.19
Arm 8	0.41	0.14
Arm 9	0.43	0.44
Arm 10	0.54	0.53
Arm 11	0.55	0.24
Arm 12	0.56	0.36
Arm 13	0.67	0.56
Arm 14	0.71	0.49
Arm 15	0.79	0.85

	μ	σ^2
Arm 1	0.1	0.05
Arm 2	0.2	0.0725
Arm 3	0.27	0.09
Arm 4	0.32	0.11
Arm 5	0.41	0.145
Arm 6	0.49	0.19
Arm 7	0.55	0.24
Arm 8	0.59	0.28
Arm 9	0.645	0.36
Arm 10	0.678	0.413
Arm 11	0.69	0.445
Arm 12	0.71	0.498
Arm 13	0.72	0.53
Arm 14	0.765	0.72
Arm 15	0.79	0.854

Figure 7: Configuration 1 and configuration 2.