Risk–Aversion in Multi–armed Bandits

Amir Sani **Alessandro Lazaric** **Rémi Munos**
INRIA Lille - Nord Europe, Team SequeL
{amir.sani,alessandro.lazaric,remi.munos}@inria.fr

Abstract

Stochastic multi–armed bandits 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 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 define two algorithms, investigate their theoretical guarantees, and report preliminary empirical results.

1 Introduction

The multi–armed bandit [13] elegantly formalizes 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 characterized by an independent reward distribution (e.g., the treatment effectiveness). 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 arms and collecting reward information versus exploiting current reward estimates by selecting the arm with the highest estimated reward. Roughly speaking, the learning objective is to solve this exploration–exploitation dilemma and accumulate as much reward as possible over n rounds. 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 possible observation realizations), 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 may be preferable to an effective but risky treatment. More generally, some applications require an effective trade–off between risk and reward.

There is no agreed upon definition for risk. A variety of behaviours result in an uncertainty which might be deemed unfavourable for a specific application and referred to as a risk. For example, an algorithm which is consistent over multiple runs may not satisfy the desire for a solution with low variability in every single realization of the algorithm. Two foundational risk modeling paradigms are Expected Utility theory [12] and the historically popular and accessible Mean-Variance paradigm [10]. A large part of decision–making theory focuses on defining and managing risk (see e.g., [9] for an introduction to risk from an expected utility theory perspective).

Risk 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, [8] showed that in general, although it is possible to achieve a small regret w.r.t. to the expert with the best average performance, 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. [16] 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 must choose a weight vector that minimizes the variance. The regret is then computed over horizon and compared to the fixed weights minimizing the variance in hindsight. In the multi–arm bandit domain, the most interesting results are by [5] and [14]. [5] introduced an analysis of the expected regret and its distribution, revealing that an anytime version of *UCB* [6] and *UCB-V* might have large regret with some non-negligible probability.[1] This analysis is further extended by [14] who derived negative results which show 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 popular mean–variance model introduced by [10]. In Sect. 2 we introduce notation and define the mean–variance bandit problem. In Sect. 3 and 4 we introduce two algorithms and study their theoretical properties. In Sect. 5 we report a set of numerical simulations aiming at validating the theoretical results. Finally, in Sect. 7 we conclude with a discussion on possible extensions. The proofs and additional experiments are reported in the extended version [15].

2 Mean–Variance Multi–arm Bandit

In this section we introduce the notation 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 bandit literature 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 historically popular and simple model: the mean–variance model proposed by [10],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*[2] $\mathrm{MV}_i = \sigma_i^2 - \rho\mu_i$.

Thus the optimal arm is the arm with the smallest mean-variance, that is $i^* = \arg\min_i \mathrm{MV}_i$. 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 reduces to σ_i^2 and the objective becomes variance minimization.

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

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

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

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

[1]The analysis is for the pseudo–regret but it can be extended to the true regret (see Remark 2 at p.23 of [5]).
[2]The coefficient of risk tolerance is the inverse of the more popular coefficient of risk aversion $A = 1/\rho$.

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*

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

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, $\widehat{\text{MV}}_{i^*,n}$ 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., [5]). 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^*. The corresponding mean and variance is

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

Given these additional definitions, we can rewrite the regret as (see App. A.1 in [15])

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

Since the last term is always negative and small [3], 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 an additional regret from the variance $\hat{\sigma}^2_n(\mathcal{A})$, 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 peaks at the mean $\hat{\mu}_n(\mathcal{A})$ furthest from all arm means.

The previous definition of regret can be further elaborated to 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.

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

Input: Confidence δ
for $t = 1, \dots, n$ **do**
  for $i = 1, \dots, K$ **do**
    Compute $B_{i,T_{i,t-1}} = \widehat{\mathrm{MV}}_{i,T_{i,t-1}} - (5+\rho)\sqrt{\frac{\log 1/\delta}{2T_{i,t-1}}}$
  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 $\widehat{\mathrm{MV}}_{i,T_{i,t}}$
end for

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

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

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

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

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

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

Where $\widetilde{\mathcal{R}}_n^{\Delta}(\mathcal{A})$ constitutes the standard regret derived from the traditional formulation of the multi-arm bandit problem and $\widetilde{\mathcal{R}}_n^{\Gamma}(\mathcal{A})$ 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, we report most of the theoretical analysis according to $\widetilde{\mathcal{R}}_n(\mathcal{A})$. Nonetheless, it is interesting to notice the major difference between the true and pseudo–regret when compared to the standard bandit problem. In fact, it is possible to show in the risk–averse case that 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, to bound the expectation of $\mathcal{R}_n$ we build on the high–probability result from Lemma 1.

3 The Mean–Variance Lower Confidence Bound Algorithm

In this section we introduce a 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* [6] and we report a theoretical performance analysis on how its mean–variance.

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 on empirical mean–variance through an application of the Chernoff–Hoeffding inequality (see e.g., [1] for the bound on the variance) on terms $\hat{\mu}$ and $\hat{\sigma}^2$.

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 the empirical mean $\hat{\mu}_{i,s}$ and variance $\hat{\sigma}_{i,s}^2$ computed as in Equation 1, then*

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

The algorithm in Figure 1 implements the principle of optimism in the face of uncertainty used in many 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

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

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 *MV-LCB* reduces to *UCB1* for $\rho \to \infty$. This is coherent with the fact that for $\rho \to \infty$ the mean–variance problem reduces to expected reward maximization, for which *UCB1* is known to be nearly-optimal. On the other hand, for $\rho = 0$ (variance minimization), the algorithm plays according to a lower–confidence–bound on the variances.

Remark 2. The *MV-LCB* algorithm has a parameter δ defining the confidence level of the bounds employed in (10). In Thm. 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. The proofs (reported in [15]) 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*

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

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 Def. 2, the regret is mainly composed by two 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, this additional term depends on the squared difference of the arm means $\Gamma_{i,j}^2$. 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 Thm. 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). [4] 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 $B_{i,s} = \sqrt{\log(1/\delta)/s}$, 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 $\Gamma_{\max} \neq 0$. [5] 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 is particularly interesting since it reveals a huge gap between the mean–variance and the standard expected regret minimization problem and will be further investigated in the numerical simulations in Sect. 5. In fact, *UCB* is known to have a worst–case regret of $\Omega(1/\sqrt{n})$ [3], 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

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 uniformly, thus collecting τ/K samples each [6]. 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.

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 $\mathbb{E}[\widetilde{\mathcal{R}}_n(\mathcal{A})] \leq 2\frac{K}{n^{1/3}}$.*

Remark 1 (the bound). We first notice that this bound suggests that *ExpExp* performs worse than *MV-LCB* on easy problems. In fact, Thm. 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, while in the remarks of Thm. 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. On the other hand, the previous bound for *ExpExp* is distribution independent and indicates the regret is still a decreasing function of n even in the worst

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

[5] Notice that this is always possible for a large majority of distributions with independent mean and variance.

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

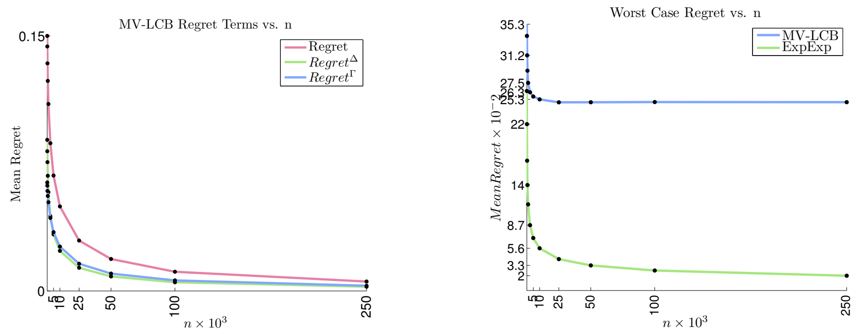

Figure 2: Regret of *MV-LCB* and *ExpExp* in different scenarios.

case. 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* or *UCB-E*, which maximize the probability of returning the best arm given a fixed budget of rounds τ (see e.g., [4]).

5 Numerical Simulations

In this section we report numerical simulations aimed at validating the main theoretical findings reported in the previous sections. In the following graphs we study the true regret $\mathcal{R}_n(\mathcal{A})$ averaged over 500 runs. We first consider the variance minimization problem ($\rho = 0$) with $K = 2$ Gaussian arms set to $\mu_1 = 1.0$, $\mu_2 = 0.5$, $\sigma_1^2 = 0.05$, and $\sigma_2^2 = 0.25$ and run *MV-LCB* [7]. 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., Thm. 1), the regret is characterized by the regret realized from pulling suboptimal arms and arms with different means (Exploration Risk) and tends to zero as n increases. Indeed, if we considered two distributions with equal means ($\mu_1 = \mu_2$), the average regret coincides with $\mathcal{R}_n^{\widehat{\Delta}}$. Furthermore, as shown in Thm. 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 App. D in [15]. Here we only compare the worst–case performance of *MV-LCB* 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 Sect. 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. For numerical results with multiple values of ρ and 15 arms, see App. D in [15].

6 Discussion

In this paper we evaluate the *risk* of an algorithm in terms of the variability of the sequences of samples that it actually generates. Although this notion might resemble other analyses of bandit algorithms (see e.g., the high-probability analysis in [5]), it captures different features of the learning algorithm. Whenever a bandit algorithm is run over n rounds, its behavior, combined with the arms' distributions, generates a probability distribution over sequences of n rewards. While the *quality* of this sequence is usually defined by its cumulative sum (or average), here we say that a sequence of rewards is *good* if it displays a good trade-off between its (empirical) mean and variance. The variance of the sequence does not coincide with the variance of the algorithm over multiple runs. Let us consider a simple case with two arms that deterministically generate 0s and 1s respectively, and two different algorithms. Algorithm $\mathcal{A}_1$ pulls the arms in a fixed sequence at each run (e.g., arm 1, arm 2, arm 1, arm 2, and so on), so that each arm is always pulled $n/2$ times. Algorithm $\mathcal{A}_2$ chooses one arm uniformly at random at the beginning of the run and repeatedly pulls this arm for n rounds. Algorithm $\mathcal{A}_1$ generates sequences such as 010101... which have high variability within

[7] 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.

each run, incurs a high regret (e.g., if $\rho = 0$), but has no variance over multiple runs because it always generates the same sequence. On the other hand, $\mathcal{A}_2$ has no variability in each run, since it generates sequences with only 0s or only 1s, suffers no regret in the case of variance minimization, but has high variance over multiple runs since the two completely different sequences are generated with equal probability. This simple example shows that an algorithm with small standard regret (e.g., $\mathcal{A}_1$), might generate at each run sequences with high variability, while an algorithm with small mean-variance regret (e.g., $\mathcal{A}_2$) might have a high variance over multiple runs.

7 Conclusions

The majority of multi–armed bandit literature focuses on the problem of minimizing the regret w.r.t. the arm with the highest return in expectation. In this paper, we introduced a novel multi–armed bandit 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 [10] to measure the performance of the arms and define the regret of a learning algorithm. We show that defining the risk of a learning algorithm as the variability (i.e., empirical variance) of the sequence of rewards generated at each run, leads to an interesting effect on the regret where an additional *algorithm* variance appears. We proposed two novel algorithms to solve the mean–variance bandit problem and we reported their corresponding theoretical analysis. To the best of our knowledge this is the first work introducing risk–aversion in the multi–armed bandit setting and it opens a series of interesting questions.

Lower bound. As discussed in the remarks of Thm. 1 and Thm. 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 problems. The primary open question is whether $O(K/n^{1/3})$ is actually the best possible achievable rate (in the worst–case) for this problem. This question is of particular interest since the standard reward expectation maximization problem has a known lower–bound of $\Omega(\sqrt{1/n})$, and a minimax rate of $\Omega(1/n^{1/3})$ for the mean–variance problem would imply 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 only justified under a Gaussianity assumption on the arm distributions. It also violates the monotonocity condition due to the different orders of the mean and variance and is not a coherent measure of risk [2]. Furthermore, the variance is a symmetric measure of risk, while it is often the case that only one–sided deviations from the mean are undesirable (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). Popular replacements for the mean–variance are the α value–at–risk (i.e., the quantile) or Conditional Value at Risk (otherwise known as average value at risk, tail value at risk, expected shortfall and lower tail risk) or other coherent measures of risk [2]. While the estimation of the α value–at–risk might be challenging [8], concentration inequalities exist for the CVaR [7]. Another issue in moving from variance to other measures of risk is whether single-period or multi-period risk evaluation should be used. While the single-period risk of an arm is simply the risk of its distribution, in a multi-period evaluation we consider the risk of the sum of rewards obtained by repeatedly pulling the same arm over n rounds. Unlike the variance, for which the variance of a sum of n i.i.d. samples is simply n times their variance, for other measures of risk (e.g., α value–at–risk) this is not necessarily the case. As a result, an arm with the smallest single-period risk might not be the optimal choice over an horizon of n rounds. Therefore, the performance of an algorithm should be compared to the smallest risk that can be achieved by any sequence of arms over n rounds, thus requiring a new definition of regret.

Simple regret. Finally, an interesting related problem is the simple regret setting where the learner is allowed to explore over n rounds and it only suffers a regret defined on the solution returned at the end. It is known that it is possible to design algorithm able to effectively estimate the mean of the arms and finally return the best arm with high probability. In the risk-return setting, the objective would be to return the arm with the best risk-return tradeoff.

Acknowledgments This work was supported by Ministry of Higher Education and Research, Nord-Pas de Calais Regional Council and FEDER through the “contrat de projets état region 2007–2013", European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270327, and PASCAL2 European Network of Excellence.

[8]While the cumulative distribution of a random variable can be reliably estimated (see e.g., [11]), estimating the quantile might be more difficult

References

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

[2] P Artzner, F Delbaen, JM Eber, and D Heath. Coherent measures of risk. *Mathematical finance*, (June 1996):1–24, 1999.

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

[4] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Proceedings of the Twenty-third Conference on Learning Theory (COLT'10)*, 2010.

[5] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410:1876–1902, 2009.

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

[7] David B. Brown. Large deviations bounds for estimating conditional value-at-risk. *Operations Research Letters*, 35:722–730, 2007.

[8] 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.

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

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

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

[12] J Neumann and O Morgenstern. Theory of games and economic behavior. *Princeton University, Princeton*, 1947.

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

[14] 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.

[15] Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-arm bandit. Technical Report hal-00750298, INRIA, 2012.

[16] 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.