Risk–Aversion in Multi–armed Bandits

Amir Sani AMIR.SANI@INRIA.FR
Alessandro Lazaric ALESSANDRO.LAZARIC@INRIA.FR
Rémi Munos REMI.MUNOS@INRIA.FR
INRIA Lille - Nord Europe, Team SequeL, France

1. Introduction

The multi–armed bandit (Robbins, 1952) 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. 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 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 may be preferable w.r.t. an effective but risky treatment. In general, some application objectives require an effective trade–off between risk and reward.

A large part of decision–making theory focuses on defining and managing risk (see e.g., (Gollier, 2001) for an introduction to risk from 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, (Even-Dar et al., 2006) 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. (Warmuth & Kuzmin, 2006) 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 n and compared to the fixed weights minimizing the variance in hindsight. In the multi–arm bandit domain, the most interesting results are by (Audibert et al., 2009) and (Salomon & Audibert, 2011). (Audibert et al., 2009) introduced an analysis of the expected regret and its distribution, revealing that an anytime version of *UCB* (Auer et al., 2002) and *UCB-V* might have large regret with some non-negligible probability.[1] This analysis is further extended by (Salomon & Audibert, 2011) 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 preliminary paper, we focus on the problem of competing against the arm with the best risk–return trade–off. In particular, we refer to the most popular measure of risk–return, the mean–variance model introduced by Markowitz (1952). We formalize the problem, introduce a confidence–bound based algorithm, discuss its properties, and finally report some preliminary results. We conclude the paper with a list of open problems.

[1] Although the analysis is mostly directed to the pseudo–regret, as commented in Remark 2 at page 23 of (Audibert et al., 2009), it can be extended to the true regret.

2. The Mean–Variance Bandit Problem

We consider the standard multi–arm bandit setting with K arms characterized by a distribution ν_i in $[0,1]$. Each distribution has a mean μ_i and 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 an arm I_t and observes a 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 multi–armed bandits literature the objective is to select the arm which leads to the highest reward in *expectation*, 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 return–risk trade–off have been proposed, here we focus on the most popular and simple model: the single period mean–variance model proposed by Markowitz (1952). In this model, the return is measured by the expected reward and the risk by the 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 it easily follows that the best arm minimizes the mean–variance, that is $i^* = \arg\min_{i=1,\dots,K} \mathrm{MV}_i$. 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 reduces to minimizing the variance σ_i^2 and the objective becomes variance minimization.

We now consider a learning algorithm $\mathcal{A}$ and its corresponding performance over n rounds. We define the *empirical* mean–variance of $\mathcal{A}$ as

$$\widehat{\mathrm{MV}}_n(\mathcal{A}) = \hat{\sigma}_n^2(\mathcal{A}) - \rho\hat{\mu}_n(\mathcal{A}), \tag{1}$$

where

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

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*

$$\mathcal{R}_n(\mathcal{A}) = \widehat{\mathrm{MV}}_n(\mathcal{A}) - \widehat{\mathrm{MV}}_{i^*,n}. \tag{2}$$

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). In order to have a better understanding of the elements composing the regret, we introduce a definition of the pseudo-regret.

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

$$\widetilde{\mathcal{R}}_n(\mathcal{A}) = \frac{1}{n}\sum_{i\neq i^*} T_{i,n}\Delta_i + \frac{2}{n^2}\sum_{i=1}^{K}\sum_{j\neq i} T_{i,n}T_{j,n}\Gamma_{i,j}^2, \tag{3}$$

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

In the following we will denote by $\widetilde{\mathcal{R}}_n^{\Delta}$ and $\widetilde{\mathcal{R}}_n^{\Gamma}$ the first and second term of the pseudo–regret respectively. It can be easily shown that the pseudo–regret is close to the regret $\mathcal{R}_n(\mathcal{A})$.

Lemma 1. *Given definitions 2 and 3,*

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

with probability at least $1 - 6nK\delta$.

On closer inspection, the definition of regret for $\mathcal{A}$ reveals that it can be determined by two different characteristics of the algorithm. Similar to the mean case ($\tau = \infty$), an algorithm $\mathcal{A}$ suffers a regret whenever a suboptimal arm $i \neq i^*$ is pulled and the regret corresponds to the difference in the mean–variance of i w.r.t. the optimal arm i^* (the gap Δ_i). Nonetheless, the variance of an algorithm $\mathcal{A}$ is not only due to the variance of the arms actually pulled by $\mathcal{A}$ but also on how different they are (see the $\Gamma_{i,j}$ term). In particular, we notice that this also has a strong relationship to the number of pulls $T_{i,n}$. In fact, if the algorithm consistently pulls any single arm, then it would only suffer from the regret of the arm but the second term in the regret would be zero. On the other hand, if a learning algorithm repeatedly explores different arms then it may suffer an additional "exploration" regret.

3. The Mean–Variance Confidence–Bound Algorithm

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

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.

Inspired by `UCB`, we introduce the index–based bandit algorithm reported in Figure 1. For each arm, the algorithm keeps track of the empirical mean–variance computed according to the samples observed so far. In particular, we define

$$\hat{\mu}_{i,s} = \frac{1}{s}\sum_{s'=1}^{s} X_{i,s'} \text{ and } \hat{\sigma}^2_{i,s} = \frac{1}{s}\sum_{s'=1}^{s} X^2_{i,s'} - \hat{\mu}^2_{i,s}, \quad (4)$$

as the empirical mean and variance computed on s observations. Thus, at the beginning of each round t, we define the empirical mean–variance of arm i as $\widehat{\mathrm{MV}}_{i,T_{i,t-1}} = \hat{\sigma}^2_{i,T_{i,t-1}} - \tau\hat{\mu}_{i,T_{i,t-1}}$. For both the terms in the empirical mean–variance we can build high–probability confidence bounds as an immediate application of Chernoff–Hoeffding inequality (see e.g., Antos et al. (2010) for the bound on the variance). The algorithm in Figure 1 implements the popular principle of optimism in face of uncertainty used in most of the multi–arm bandit algorithms. Thus, we define a lower–confidence bound on the mean–variance of arm i at time t when it has been pulled s times so far as

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

Given the index of each arm, at each round t the algorithm simply selects the arm with the smallest mean–variance index, i.e., $I_t = \arg\min_i B_{i,s}$. 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` (Auer et al., 2002) whenever $\tau = \infty$. This is coherent with the fact that when $\tau = \infty$ 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 $\tau = 0$, which leads to the problem of cumulative reward variance minimization, the algorithm is a lower–confidence–bound algorithm on the variance.

In the following we report a theoretical analysis for the expected pseudo–regret. Similar results hold in high–probability and for the true regret $\mathcal{R}_n(\mathcal{A})$ as well.

Theorem 1. *Let the optimal arm i^* be unique and $b = 2(5+\rho)$, if MV-LCB is run with $\delta = 1/n^2$ then*

$$\mathbb{E}[\widetilde{\mathcal{R}}_n(\mathcal{A})] \le \frac{2b^2\log n}{n}\Bigg(\sum_{i\neq i^*}\frac{1}{\Delta_i} + 4\sum_{i\neq i^*}\frac{\Gamma^2_{i^*,i}}{\Delta^2_i} + \frac{4b^2\log n}{n}\sum_{i\neq i^*}\sum_{\substack{j\neq i\\ j\neq i^*}}\frac{\Gamma^2_{i,j}}{\Delta^2_i\Delta^2_j}\Bigg) + (17+6\rho)\frac{K}{n}.$$

Remark 2 (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

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

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 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, it is interesting to note that this additional term depends on the squared difference in the means of the arms $\Gamma^2_{i,j}$. Thus, if all the arms have the same mean, this term would be zero.

Remark 3 (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 $\sigma^2_1 = \sigma^2_2 = 0$ (deterministic arms). [3] 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 $\rho \neq 0$,

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

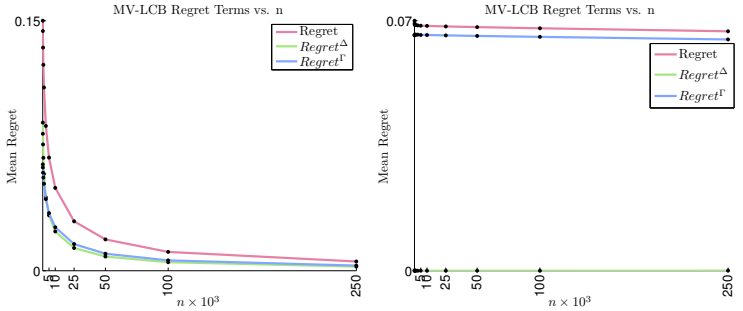

Figure 2. Performance of `MV-LCB` in two different settings.

since it is always possible to design an environment (i.e., a set of distributions) such that $\Delta_{\min} = 0$ and $\Gamma_{\max} \neq 0$. 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 gap between the mean–variance problem and the standard expected regret minimization problem and will be further investigated in the numerical simulations presented in Section 4. In fact, in the latter case, *UCB* is known to have a worst–case regret per round of $\Omega(1/\sqrt{n})$ (Audibert & Bubeck, 2010), while in the worst case, *MV-LCB* suffers a constant regret.

4. 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 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*. In the first plot of Figure 2 we report the true regret $\mathcal{R}_n$ and the two components due to the pull of suboptimal arm and the "exploration risk" ($\mathcal{R}_n^{\Delta}$ and $\mathcal{R}_n^{\Gamma}$ respectively). 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 from pulling suboptimal arms and the regret of pulling arms with different means (Exploration Risk). 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 $O(\log n/n)$. In the second plot of Figure 2 we report the study of the worst–case performance of *MV-LCB* for the configuration suggested in Remark 3. Indeed, as discussed in the remark, in this case the regret of *MV-LCB* does not decrease over time but stabilizes to a constant, confirming that in the worst–case *MV-LCB* can have a very poor performance.

5. Open Problems and Extensions

Lower bound. The previous results suggest that a confidence–bound algorithm trying to minimize the number of suboptimal pulls might have a large "risk" and suffer a constant regret in the worst–case. It is an open problem to understand to which extent it is actually possible to achieve an effective trade–off between the number of pulls on the suboptimal arm and the variability of the algorithm. For this reason, it would be important to derive a distribution–free lower bound for the general mean–variance problem.

Different measures of return–risk. In economics, the mean–variance model has often been criticized. In fact, in expected utility theory, the mean–variance model is justified only under a Gaussian assumption on the distribution of the arms, and the use of one–sided deviations from the expected return are preferable to symmetric measures of risk like the variance (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). Technically speaking, the main challenge in this case is the estimation of the value–at–risk of each arm. In fact, while the cumulative distribution of random variable can be reliably estimated (see e.g., (Massart, 1990)), the quantile is much more difficult, in particular when the level α 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 bounded or sub-gaussian distribution, in this case it would be preferable to deal with distributions with fat tails.

References

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

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

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

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

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

Gollier, Christian. *The Economics of Risk and Time*. The MIT Press, 2001. ISBN 0262072157.

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

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

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

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

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