STATS 710 – Seq. Dec. Making with mHealth Applications

Note: Risk-Averse Bandits

Instructors: Susan Murphy and Ambuj Tewari *Scribe: Brandon Oselio*

1 Introduction

This brief summary will cover the recent work regarding risk-aware bandits. Most bandit algorithm literature is focused on mean regret $\mathbb{E}[\mathcal{R}_T]$, and its corresponding finite-sample and asymptotic behavior. In many applications, however, this may not be a rich enough measure of performance. For example, consider the problem of treatment allocation in clinical trials; while some treatment policy may work well on average, it might also have a large variability, which could make certain patients worse off. This type of outcome is obviously unacceptable. Therefore, we must search for ways to control this type of error.

We will explore two main concepts in this paper. First, we see that [AMS09] provides concentration bounds of regret for both the traditional UCB1 and a variant UCBV. We discuss the implication of these results and some strategies for risk-averse users of UCB algorithms in general. The second topic takes a different route to control risk by changing the definition of traditional regret. A popular replacement for regret is the "mean-variance" model proposed by [Mar52].

2 Concentration Analysis of UCB1

We recall the UCB1 algorithm originally studied by [ACBF02]:

1. $\forall a \in \mathcal{A}$, compute $\bar{R}_t^a$ and $N_t(a)$
2. Pick action a to maximize $\bar{R}_t^a + \sqrt{\frac{\rho \log t}{N_t(a)}}$

Note that the conventional choice of ρ is 2. The choice of ρ controls the amount of exploration that occurs. This is important for the concentration bound that follows. Qualitatively, we can think of a larger ρ as shielding the user from unlucky initializations by imposing more exploration, at the cost of a greater rate for expected regret.

Before presenting the results, we note that these are with respect to pseudo-regret:

$$\tilde{\mathcal{R}}_T(\mathcal{L}, (D_{a\in\mathcal{A}}) = \sum_{a\in\mathcal{A}} N_T(a)\Delta_a, \tag{1}$$

where $\Delta_a = \mu_* - \mu_a$. Note that $\mathbb{E}[\tilde{\mathcal{R}}_T] = \mathbb{E}[\mathcal{R}_T]$. Remark 2 of [AMS09] discusses the similarity of the expectation and concentration bounds for pseudo-regret and regret. For simplicity, we will discuss pseudo-regret only for the rest of this section, and assume that a similar type of bound holds for regret.

Theorem 1 ([AMS09], Theorem 7)**.** *Assume that the distributions of rewards for all arms have support on* $[0, b]$*, and let* $\rho > 1$*. Then we have:*

$$\mathbb{E}[\tilde{\mathcal{R}}_T(\mathcal{L}_{UCB1}, (D_a)_{a\in\mathcal{A}})] \leq \sum_{a:\Delta_a>0} \left[\frac{4b^2}{\Delta_a}\rho \log(n) + \Delta_a\left(\frac{3}{2} + \frac{1}{2(\rho-1)}\right)\right]. \tag{2}$$

Theorem 2 ([AMS09], Theorem 8)**.** *Assume that* $\rho > 1/2$. *Let* $v_a = (2b/\Delta_a)^2$, $r_0 = \sum_a \Delta_a(1 + \rho v_a \log T)$. *Then, for any* $x \geq 1$, *we have:*

$$\mathbb{P}(\tilde{\mathcal{R}}_T > r_0 x) \leq \sum_{a:\Delta_a>0} \left\{ T^{-2\rho x+1} + \frac{((1+\rho v_a \log T)x)^{-2\rho+1}}{2\rho - 1} \right\} \tag{3}$$

These two theorems characterize the trade-off between expected regret and concentration of regret. As ρ increases, we see that the tail bound decreases, but our proven expected regret bound increases at a rate of $\log(n)$. A variant of UCB which is slightly more aggressive (less exploration) and utilizes the empirical variance, is also analyzed in [AMS09], but we omit this due to space.

3 Risk-Averse Measures of Regret

This section is focused on defining alternative notions of regret that incorporate a quantitative measure of risk so that we may directly minimize it. First, we discuss an approach that follows the mean-variance model, originally proposed by [Mar52]. This is the procedure followed by [SLM12], and later developed further by [VZ16]. We first define the mean-variance of an arm as:

$$\mathrm{MV}_a = \sigma_a^2 - \rho\mu_a, \ \rho > 0 \tag{4}$$

where μ_a, σ_a^2 is the mean and variance of the distribution D_a, respectively. We also define the empirical mean and variance in the usual way, and further define the following:

$$\hat{\mu}_t(\mathcal{A}) = \frac{1}{t}\sum_{i=1}^{t} R_t, \quad \hat{\sigma}_t^2(\mathcal{A}) = \frac{1}{t}\sum_{i=1}^{t}(R_i - \hat{\mu}_t(\mathcal{A}))^2, \tag{5}$$

where R_t is the observed reward at timestep t. We can then define the following notion of regret:

$$\mathcal{R}_T(\mathcal{L}, (D_{a\in\mathcal{A}}) = \widehat{\mathrm{MV}}_T(\mathcal{A}) - \widehat{\mathrm{MV}}_{i^*,n} \tag{6}$$

We note that as $\rho \leftarrow \infty$, we restore our standard setting of bandit regret. As $\rho \leftarrow 0$, we attempt to pick the arm with the least variance, i.e., variance minimization. using ρ, we are once again trading off between minimizing variability, and maximizing the mean utility using this reward function.

[SLM12] Then gives two algorithms. One is a UCB based algorithm, which acheives an expected regret bound on the order of $O(\log^2 T/T)$. However, the worst case analysis of regret in this case is constant (see Remark 2), which implies that on certain "hard problems", the algorithm does not do very well. They then introduce an algorithm whose worst case algorithm is better, but explicitly splits apart the exploration and exploitation phases. Therefore, the bound on regret is $O(T^{-1/3})$. Note that this measure of regret implicitly divides by the epoch T - therefore, to compare with bounds that we consider in the class, we should multiply by T.

We conclude this section by mentioning the work of [ZIJC14], which characterizes the types of regret functions that allow for sublinear regret. This includes both the standard bandit regret, as well as the MV regret described above. This is relevant as one might wish to consider a more complicated regret measure that depends on the mean and variance of the arms.

4 Conclusion

In this short exposition, we covered a few ideas for risk-averse bandits. The first approach is to calculate concentration bounds on the standard regret, with which we can quantitatively give a bound on the risk of unlucky runs. We also consider a non-standard regret function that takes into account the variance of the arms. For future work, it would be interesting to calculate concentration bounds on these non-standard regret functions, and compare them with the traditional regret concentration bounds.

References

[ACBF02] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[AMS09] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

[Mar52] Harry Markowitz. Portfolio selection. *The journal of finance*, 7(1):77–91, 1952.

[SLM12] Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012.

[VZ16] Sattar Vakili and Qing Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):1093–1111, Sept 2016.

[ZIJC14] Alexander Zimin, Rasmus Ibsen-Jensen, and Krishnendu Chatterjee. Generalized risk-aversion in stochastic multi-armed bandits. *arXiv preprint arXiv:1405.0833*, 2014.