Risk aversion in multi-arm bandits

Oussama Ennafii

Ecole Normale Superieur de Cachan

January 15, 2015

Introduction:

- In the classical multi-arm bandit(MAB) online setting, the objectif is to find the best arm in terms of expectation.
- Problem: if the arm with the best mean value have heavy tails!
- We need to evaluate risk and incorporate risk-aversion into the model.
- Problem: There is no agreed upon definition of risk!
- For each definition there is a possible different soultion!

Risk minimization:

- Here the learning is not done online: we sample at each time rewards from each arm and we learn then what are the arms with less risk. We are more intersted in the sample complexity necessary to get a certain level of precision.
- But first we need a definition of risk!

Risk minimization:

Definition:

For each arm i:

- The value at risk(V@R) is defined as:

$$V@R_\lambda(i) = -q_i(\lambda)$$

- and the average/conditional value at risk(AV@R) as:

$$AV@R_\lambda(i) = \frac{1}{\lambda}\int \mathbb{I}(0 \leq r \leq \lambda)V@R_r(i)dr$$

Risk minimization:

We can estimate easily the quantile using the order statistic and the Reiman sum for the AV@R:

$$\widehat{V@R}_\lambda(i) = -X^{(i)}_{(\lceil \lambda T_t^{(i)} \rceil)}$$

$$\widehat{AV@R}_\lambda(i) = 1/\lambda\Big(\sum_{j=0}^{\lfloor \lambda T_t^{(i)} \rfloor - 1} \frac{1}{\lceil \lambda T_t^{(i)} \rceil} X^{(i)}_{(j+1)} + \Big(\lambda - \frac{\lfloor \lambda T_t^{(i)} \rfloor}{T_t^{(i)}}\Big) X^{(i)}_{(\lceil \lambda T_t^{(i)} \rceil)}\Big)$$

Risk minimization:

Problem: Non linearity: If an arm minimizes risk at time t It does not mean it will minimize the cummulative risk at time $t+1$. There is no optimal arm to pull each time!

Example: Three arms

$$X_t^1 \sim -10 - 10.Bernoulli(0.1)$$

$$X_t^2 \sim -5 - 10.Bernoulli(1/2)$$

$$X_t^3 = -14$$

$$min_{i_1,i_2,i_3} V@R(X_1^{i_1} + X_2^{i_2} + X_3^{i_3}) = V@R(X_1^1 + X_2^2 + X_3^3)$$

Risk minimization:

CuRisk Algorithm

1: **Input:**Arms$\{1,\ldots,K\}$, Number of values possible for each arm:P, scalar $r \in [0,1]$ and rewards: $\mathscr{X}^i = \{X_t^i, t = 1,\ldots,N\}$, $\forall i = 1,\ldots,K$.

2: **Output:** Arm choices $i_1,\ldots,i_T$.

3: for $i = 1,\ldots,K$ do

4: Compute:

$$\hat{d}_i(k) = \frac{1}{|\mathscr{X}^i|} \sum_{X \in mathscrX^i} \mathbb{I}(X = v_k), \forall k = 1,\ldots,P$$

5: end for

Risk minimization:

CuRisk Algorithm(continues:)

6: Solve (ALC-VAR):

$$max_{m_l:l=1,\dots,K} \text{sup} \quad x$$

s.t $\sum m_l = \tau$ and

$$\sum_{k=1,\dots,\lfloor xP \rfloor} \frac{1}{2kr^k} \sum_{1,\dots,2k} (-1)^j R[\prod_{l=1,\dots,K} \hat{D}_l^{m_l}(re^{\sqrt{-1}j\pi/k})]$$

7: for $t = 1, \dots, \tau$ do

8: Output each arm i m_i^* times.

9: end for

Risk minimization:

Theorem:

We suppose that the rewards are independent w.r.t time and arms. If each arm distribution takes values in: $\{v_k := 1, \ldots, P\}$ and $\exists \gamma > 0$ s.t $\nu(k) \geq \gamma$ and if:

$$P \geq \frac{(\lambda + \gamma)(1 - r^2)^2 + 2}{\epsilon\gamma(1 - r^2)^2}$$

$$N \geq \frac{32\tau^2}{(K\gamma\epsilon - \lambda - \gamma)^2} log(\frac{4.2^K.\tau n^\tau}{\delta})$$

then the output of the CuRisk Algorithm yields with probability 1_δ:

$$|min_{a_1,\ldots,a_\tau} V@R(\sum_{t\leq\tau} X_t^{a_t}) - V@R(\sum_{t\leq\tau} X_t^{i_t})| \leq 2\epsilon$$

Towards risk-reward trade-off:

- This result is more general. What happens if the best arm in terms of mean value is also the best in terms of risk?
- In that case we can use the MaRaB algorithm. It is a lower confidence bound algorithm: At round t, we choose the arm:

$$I_t := argmax \widehat{AV@R}_i - C\sqrt{\frac{log(\lceil t\lambda \rceil)}{\lceil \lambda T_t^{(i)} \rceil}}$$

The MV-LCB:

Definition:

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:

$$MV_i = \sigma_i^2 - \rho\mu_i$$

- The optimal arm is the one with the best mean-variance value. It is independent from the previous results.

The MV-LCB:

Definition:

The regret, in the mean-variance setting, for a learning algorithm A over T rounds is defined as:

$$\mathscr{R}_T(A) = \widehat{MV}_t(A) - \widehat{MV}_t^{(i^*)}$$

We define also the pseudo-regret:

$$\tilde{\mathscr{R}}_T(A) = \frac{1}{T} \sum_{i \neq i^*} T_T^{(i)} \Delta_i^2 + \frac{1}{T^2} \sum_{i=1,\dots,K} \sum_{i \neq j} T_T^{(i)} T_T^{(j)} \Gamma_{i,j}^2$$

where $\Delta_i = MV_i - MV_{i^*}$ and $\Gamma_{i,j} = \mu_i - \mu_j$.

The MV-LCB:

Lemma:

With probability at least $1-\delta$:

$$\mathscr{R}_T(A) \leq \tilde{\mathscr{R}}_T(A) + (5+\rho)\sqrt{\frac{2K\log(6TK/\delta)}{n}} + 4\sqrt{2}\frac{K\log(6TK/\delta)}{n}$$

The MV-LCB:

The MV-LCB algorithm:

1. **Input:**Confidence δ
2. for $t = 1, \dots, T$ do
3. for $i = 1, \dots, K$ do
4. Compute $B^{(i)}_{T^{(i)}_{t-1}} = \widehat{MV}^{(i)}_{T^{(i)}_{t-1}} - (5+\rho)\sqrt{\frac{log(1/\delta)}{2T^{(i)}_{t-1}}}$
5. end for
6. return $I_t = argmin_{i=1,\dots,K} B^{(i)}_{T^{(i)}_{t-1}}$
7. update $T^{(i)}_t = T^{(i)}_{t-1} + 1$

The MV-LCB:

The MV-LCB algorithm:

8. observe $X^{(I_t)}_{T^{(i)}_t} \sim \nu_{I_t}$
9. update $\widehat{MV}^{(i)}_{T^{(i)}_t}$
10. end for

The MV-LCB:

Roughly,we can bound the pseudo-regret as:

$$\mathbb{E}\tilde{\mathscr{R}}_n(A) \leq O(\frac{K}{\Delta_{min}}\frac{log(n)}{n} + K\frac{\Gamma_{max}^2}{\Delta_{min}^4}\frac{log(n)^2}{n})$$

Example of worst-case scenario:$\rho = 0$, $K = 2$, $\sigma_1 = \sigma_2$ and $\mu_1 \neq \mu_2$:

$$\mathscr{R}_n(MV - LCB) = 1/4\Gamma^2 > 0$$

The ExpExp algorithm:

We run the MV-LCB in the first phase up to time τ. In the second phase, we exploit the rewards of the best arm yielded in the first phase.

Theorem

If we run the ExpExp algorithm with $\tau = K(\frac{T}{14})^{\frac{2}{3}}$ then

$$\mathbb{E}\tilde{\mathscr{R}}_n(A) \leq 2\frac{K}{T^{\frac{1}{3}}}$$

Towards a more general risk measure:

Definition:

We denote the risk-aversion level by:

$$\kappa_{\lambda,\nu} = \frac{1}{\lambda} log\ \mathbb{E}exp(\lambda X)$$

We justify this definition by the inequalities:

$$\mathbb{P}(X \geq inf\{\frac{1}{\lambda} log\ \mathbb{E}exp(\lambda X) + \frac{log(1/\delta)}{\lambda} : \lambda > 0\}) \leq \delta$$

and

$$\mathbb{P}(X \leq sup\{\frac{1}{\lambda} log\ \mathbb{E}exp(\lambda X) - \frac{log(1/\delta)}{\lambda} : \lambda > 0\}) \leq \delta$$

Towards a more general risk measure:

- we can characterize it as:

$$\kappa_{-\lambda,\nu} = inf\{\mathbb{E}_{\nu'}(X) + \frac{1}{\lambda}KL(\nu'\|\nu) : \nu' \text{a distribution over } \mathbb{R}\} \leq \mathbb{E}_{\nu}(X)$$

- Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then:

$$\kappa_{-\lambda,\nu} = \mu + \lambda\sigma^2/2$$

- Now the optimal arm is characterised as:

$$i^* = argmin_{i=1,\ldots,K}\kappa_{-\lambda,\nu}$$

Towards a more general risk measure:

Definition:

In this setting, the empirical regret is defined as:

$$\mathscr{R}(\lambda) := \sum_{t} X_t^{(i^*)} - \sum_{i \leq K} \sum_{s \leq T_T^{(i)}} X_t^{(i)}$$

The risk-averse regret as:

$$\bar{\mathscr{R}}(\lambda) := \sum_{i \leq K} (\kappa_{-\lambda,\nu_{i^*}} - \kappa_{-\lambda,\nu_i}) \mathbb{E} T_T^{(i)}$$

Towards a more general risk measure:

Proposition:

For some non negative constants u_i, we define the event where at least one arm is pulled too much:

$$\Omega = \{\exists i \neq i^* : T_T^{(i)} > u_i\}$$

We fix λ such that $\kappa_{-\lambda,\nu}$ exists for all arms. Then with probability at least $1-\delta-\mathbb{P}(\Omega)$, the regret verifies always:

$$\mathscr{R}_T(\lambda) \leq \sum_{i \neq i^*} u_i(\kappa_{-\lambda,\nu_{i^*}} - \kappa_{-\lambda,\nu_i}) + (m^-_{\lambda,\nu_i^*} \sum_{i \neq i^*} u_i + (K-1)\frac{log(2K/\delta)}{\lambda})$$

$$+ inf_{\lambda'>0}\{m^+_{\lambda',\nu_i^*} \sum_{i \neq i^*} u_i + \frac{log(2K/\delta)}{\lambda}\}$$

The RA-UCB:

We work with upper bounded distributions. We define:

$$U_t(i) = sup\{\kappa_{-\lambda,\nu} : \mathbb{K}(\hat{\nu}_t(i), \kappa_{-\lambda,\nu}) \leq C\frac{log(t)}{T_t^{(i)}}\}$$

where:

$$\mathbb{K}(\hat{\nu}_t(i), r) = inf\{KL(\hat{\nu}_t(i)||\nu) : \nu \text{ distribution bounded by } B, \kappa_{-\lambda,\nu} \geq r\}$$

The RA-UCB:

The RA-UCB algorithm:

1. **Input:**Confidence δ
2. for $t = 1, \ldots, T$ do
3. for $i = 1, \ldots, K$ do
4. Compute $U_t(i)$.
5. end for
6. return $I_t = argmin_{i=1,\ldots,K} U_t(i)$
7. update $T_t^{(i)} = T_{t-1}^{(i)} + 1$

The RA-UCB:

The RA-UCB algorithm:

8. observe $X^{(I_t)}_{T^{(i)}_t} \sim \nu_{I_t}$
9. update $\hat{\nu}_t(i)$
10. end for

Conclusion:

- It is more difficult to emcompass risk aversion into the MAB setting.
- The MV-LCB, the ExpExp and the RA-UCB are powerful algorithms that takes into account the risk reward trade-off.
- The MaRaB algorithm is very restrictive.
- The MV-LCB has two drawbacks: it penalizes the algorithm for switching arms and the risk measure used is adequate only for sub-guassian distributions with some symmetry.
- The RA-UCB don't take advantage of possible heavy upper tails.