Stochastic Multi-armed Bandits: Optimal Trade-off among Optimality, Consistency, and Tail Risk

David Simchi-Levi [1] Zeyu Zheng [2] Feng Zhu [1]

[1] Institute for Data, Systems, and Society, Massachusetts Institute of Technology

[2] Industrial Engineering & Operations Research, University of California, Berkeley

December 2023

NeurIPS 2023 Spotlight

Contents

Contents

Motivation

- The multi-armed bandit (MAB) problem and its various extensions have been extensively investigated.
  - [Bubeck and Cesa-Bianchi, 2012], [Russo et al., 2018], [Slivkins et al., 2019], [Lattimore and Szepesvári, 2020], ...

Motivation

- Most work typically focus on achieving optimal expected regret.
- Well-known common goals:
  - worst-case expected regret $\tilde{O}(\sqrt{T})$
  - instance-dependent expected regret $\tilde{O}(1)$

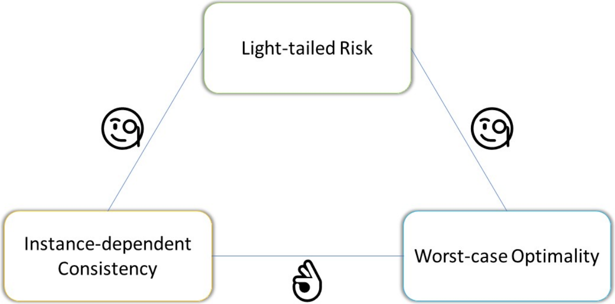
Motivation

- Most work typically focus on achieving optimal expected regret.
- Well-known common goals:
  - worst-case expected regret $\tilde{O}(\sqrt{T})$
  - instance-dependent expected regret $\tilde{O}(1)$
- What if we are willing to give up a little bit on regret *expectation* ...

Motivation

- Most work typically focus on achieving optimal expected regret.
- Well-known common goals:
  - worst-case expected regret $\tilde{O}(\sqrt{T})$
  - instance-dependent expected regret $\tilde{O}(1)$
- What if we are willing to give up a little bit on regret *expectation* ...
  - regret *tail* $\mathbb{P}(\text{regret} > x)$ decays faster for large x?

Motivation: Main Question



What is the (optimal) trade-off between expectation $\mathbb{E}[\text{regret}]$ and tail risk $\mathbb{P}(\text{regret} > x)$?

Motivation: Literature

- Not much work on understanding the regret *distribution* of stochastic bandit policies.
    - [Audibert et al., 2009], [Salomon and Audibert, 2011]: standard bandit algorithms generally have undesirable concentration properties around the instance-dependent mean $O(\ln T)$.
    - [Ashutosh et al., 2021]: a policy with an $O(\ln T)$ regret can be fragile to mis-specified risk parameter (e.g., the parameter for subgaussian noises).
    - [Fan and Glynn, 2022]: for optimized UCB policies, *the probability of incurring a linear regret* is very heavy-tailed: at least $\Omega(1/T)$. Meanwhile, heavy-tailed risk exists for more general UCB policies.

Contents

Model

- Time horizon T; Number of arms K; Mean reward $\theta \in [0,1]^K$.
- In each time $t \in [T]$, a policy π pulls an arm $a_t \in [K]$ and collects a reward $r_{t,a_t} = \theta_{a_t} + \epsilon_{t,a_t}$.
  - ϵ_{t,a_t} is an independent zero-mean σ-subgaussian noise term.
- Fixed-time, known T:

$$\pi_t(T) : \{a_1, r_{1,a_1}, \cdots, a_{t-1}, r_{t-1,a_{t-1}}\} \cup \{T\} \longmapsto a_t.$$

- Pseudo Regret

$$R^\pi_\theta(T) = \theta_* \cdot T - \sum_{t=1}^{T} \theta_{a_t}. \quad (\theta_* = \max_k \theta_k)$$

Model: Core Concepts

1. Regret Expectation. Fix $\alpha \in [1/2, 1)$ and $\beta \in [0, 1)$. We differentiate between two scenarios: worst-case and instance-dependent.

Model: Core Concepts

1. Regret Expectation. Fix $\alpha \in [1/2, 1)$ and $\beta \in [0, 1)$. We differentiate between two scenarios: worst-case and instance-dependent.

(a) A fixed-time policy π is said to be worst-case α-optimal or simply, α-**optimal**, if for any $\varepsilon > 0$, we have

$$\sup_{\theta} \mathbb{E}\left[R_{\theta}^{\pi}(T)\right] = o(T^{\alpha+\varepsilon}).$$

$\alpha = 1/2$: worst-case optimality

Model: Core Concepts

1. Regret Expectation. Fix $\alpha \in [1/2, 1)$ and $\beta \in [0, 1)$. We differentiate between two scenarios: worst-case and instance-dependent.

(a) A fixed-time policy π is said to be worst-case α-optimal or simply, α-**optimal**, if for any $\varepsilon > 0$, we have

$$\sup_{\theta} \mathbb{E}\left[R_\theta^\pi(T)\right] = o(T^{\alpha+\varepsilon}).$$

$\alpha = 1/2$: worst-case optimality

(b) A fixed-time policy π is said to be instance-dependent β-consistent or simply, β-**consistent**, if for any underlying true mean vector θ and any $\varepsilon > 0$, we have

$$\mathbb{E}\left[R_\theta^\pi(T)\right] = o(T^{\beta+\varepsilon}).$$

$\beta = 0$: instance-dependent consistency

Model: Core Concepts

2. Regret Tail Risk. We differentiate between two scenarios: worst-case and instance-dependent.

Model: Core Concepts

2. Regret Tail Risk. We differentiate between two scenarios: worst-case and instance-dependent.

(a) A fixed-time policy π enjoys worst-case **light-tailed risk**, if there exists a constant $c \in (0, 1/2)$ such that

$$\sup_{\theta} \mathbb{P}\left(R_\theta^\pi(T) > cT\right) = O(\exp(-\mathrm{poly}(T))).$$

Model: Core Concepts

2. Regret Tail Risk. We differentiate between two scenarios: worst-case and instance-dependent.

(a) A fixed-time policy π enjoys worst-case **light-tailed risk**, if there exists a constant $c \in (0, 1/2)$ such that

$$\sup_{\theta} \mathbb{P}\left(R_{\theta}^{\pi}(T) > cT\right) = O(\exp(-\mathrm{poly}(T))).$$

(b) A fixed-time policy π enjoys instance-dependent **light-tailed risk**, if for any underlying true mean vector θ, there exists a constant $c \in (0, 1/2)$ such that

$$\mathbb{P}\left(R_{\theta}^{\pi}(T) > cT\right) = O(\exp(-\mathrm{poly}(T))).$$

Contents

Main Results: Optimal Trade-off

- Optimal regret tail risk for the family of policies that obtain both $\tilde{O}(T^{\alpha})$ worst-case and $\tilde{O}(T^{\beta})$ instance-dependent expected regret (explicit tail bounds are given in the paper):

$-\ln \sup_{\theta} \mathbb{P}(R_{\theta}^{\pi}(T) > x)$ (worst-case scenario)	$\widetilde{\Theta}((x/T^{1-\alpha}) \wedge T^{\beta})$ for $x = \widetilde{\Omega}(T^{\alpha})$
$-\ln \mathbb{P}(R_{\theta}^{\pi}(T) > x)$ (instance-dependent scenario)	$\widetilde{\Theta}(T^{\beta})$ for $x = \widetilde{\Omega}(T^{\beta})$

Main Results: Optimal Trade-off

- Optimal regret tail risk for the family of policies that obtain both $\tilde{O}(T^{\alpha})$ worst-case and $\tilde{O}(T^{\beta})$ instance-dependent expected regret (explicit tail bounds are given in the paper):

$-\ln\sup_{\theta}\mathbb{P}(R_{\theta}^{\pi}(T) > x)$ (worst-case scenario)	$\widetilde{\Theta}((x/T^{1-\alpha}) \wedge T^{\beta})$ for $x = \widetilde{\Omega}(T^{\alpha})$
$-\ln\mathbb{P}(R_{\theta}^{\pi}(T) > x)$ (instance-dependent scenario)	$\widetilde{\Theta}(T^{\beta})$ for $x = \widetilde{\Omega}(T^{\beta})$

More sub-optimality and inconsistency leaves space for more light-tailed regret distribution.

Main Results: Algorithm Design

- Successive Elimination with the bonus term

$$\mathrm{rad}(n) = \underbrace{\eta_1 \frac{(T/K)^{\alpha}\sqrt{\ln T}}{n}}_{\text{control the worst-case tail}} \wedge \underbrace{\eta_2 \sqrt{\frac{T^{\beta} \ln T}{n}}}_{\text{control the instance-dependent tail}} .$$

Main Results: Algorithm Design

- Successive Elimination with the bonus term

$$\mathrm{rad}(n) = \underbrace{\eta_1 \frac{(T/K)^{\alpha}\sqrt{\ln T}}{n}}_{\text{control the worst-case tail}} \wedge \underbrace{\eta_2 \sqrt{\frac{T^{\beta}\ln T}{n}}}_{\text{control the instance-dependent tail}} .$$

- Special hyperparameters
  - $\eta_1 = \beta = 0$

$$\mathrm{rad}(n) = \eta_2 \sqrt{\frac{\ln T}{n}}$$

  - $\eta_2 = +\infty, \alpha = 1/2$

$$\mathrm{rad}(n) = \eta_1 \frac{\sqrt{T \ln T}}{n\sqrt{K}}$$

Contents

Extensions: Structured Non-stationarity

- We also extend our results to models that allow **structured non-stationarity** beyond standard stochastic MAB problems:
  - a common reward baseline among all arms for each time period [Greenewald et al., 2017, Krishnamurthy et al., 2018, Kim and Paik, 2019, Simchi-Levi and Wang, 2022]

$$r_{t,a_t} = b_t + \theta_{a_t} + \epsilon_{t,a_t}.$$

  We show that a simple modification to our policy design leads to optimal trade-off similar to those for the stochastic MAB model.

Concluding Remarks

- Optimal trade-off and explicit regret tail bounds for K-armed bandit
- Extensions on structured non-stationarity

Concluding Remarks

- Optimal trade-off and explicit regret tail bounds for K-armed bandit
- Extensions on structured non-stationarity
- Unknown T? Heavy-tailed rewards? Thompson sampling?

Thank You!

A follow-up extended version
https://arxiv.org/abs/2304.04341

Contact: fengzhu@mit.edu

Ashutosh, K., Nair, J., Kagrecha, A., and Jagannathan, K. (2021).
Bandit algorithms: Letting go of logarithmic regret for statistical robustness.
In *International Conference on Artificial Intelligence and Statistics*, pages 622–630. PMLR.

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009).
Exploration–exploitation tradeoff using variance estimates in multi-armed bandits.
Theoretical Computer Science, 410(19):1876–1902.

Bubeck, S. and Cesa-Bianchi, N. (2012).
Regret analysis of stochastic and nonstochastic multi-armed bandit problems.
arXiv preprint arXiv:1204.5721.

Fan, L. and Glynn, P. W. (2022).
The fragility of optimized bandit algorithms.
arXiv preprint arXiv:2109.13595.

Greenewald, K., Tewari, A., Murphy, S., and Klasnja, P. (2017).

Action centered contextual bandits.
Advances in neural information processing systems, 30.

Kim, G.-S. and Paik, M. C. (2019).
Contextual multi-armed bandit algorithm for semiparametric reward model.
In *International Conference on Machine Learning*, pages 3389–3397. PMLR.

Krishnamurthy, A., Wu, Z. S., and Syrgkanis, V. (2018).
Semiparametric contextual bandits.
In *International Conference on Machine Learning*, pages 2776–2785. PMLR.

Lattimore, T. and Szepesvári, C. (2020).
Bandit algorithms.
Cambridge University Press.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., et al. (2018).
A tutorial on thompson sampling.

Foundations and Trends® in Machine Learning, 11(1):1–96.

Salomon, A. and Audibert, J.-Y. (2011).
Deviations of stochastic bandit regret.
In *International Conference on Algorithmic Learning Theory*, pages 159–173. Springer.

Simchi-Levi, D. and Wang, C. (2022).
Multi-armed bandit experimental design: Online decision-making and adaptive inference.
Available at SSRN 4224969.

Slivkins, A. et al. (2019).
Introduction to multi-armed bandits.
Foundations and Trends® in Machine Learning, 12(1-2):1–286.