Thompson Sampling on Symmetric α-Stable Bandits

Abhimanyu Dubey and Alex Pentland

Massachusetts Institute of Technology

dubeya@mit.edu

IJCAI 2019
August 14, 2019

Multi-Armed Bandits

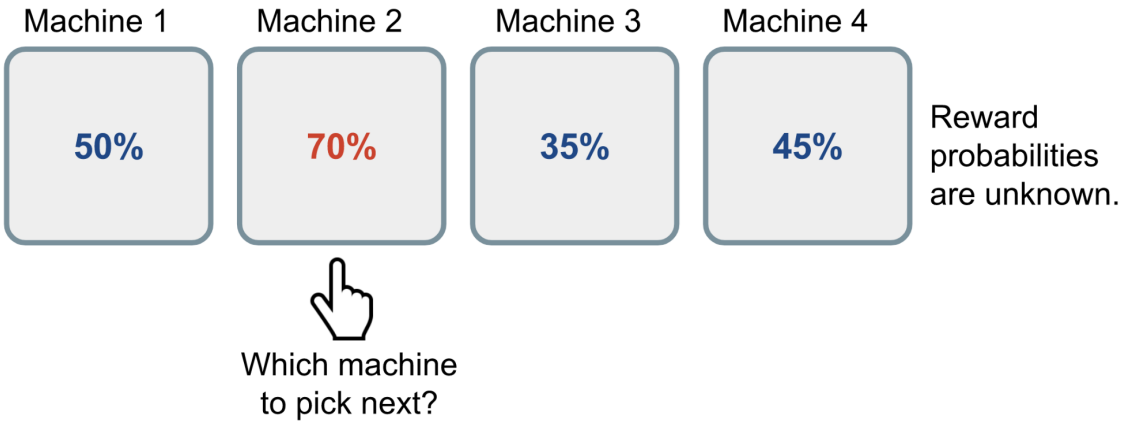

Figure: Bernoulli bandit problem.

Stochastic Multi-Armed Bandits

- K actions ("arms") that return rewards r_k sampled i.i.d. from K different distributions, each with mean μ_k.
- The problem proceeds in rounds; at each round t, the agent chooses action $a(t)$, and obtains a randomly drawn reward $r_{a(t)}(t)$ from the corresponding distribution.
- The goal is to minimize *regret*:

$$R(T) = \underbrace{T \cdot \mu^*}_{\substack{\text{best possible} \\ \text{avg. reward}}} - \underbrace{\sum_{k \in [K]} \mu_k \mathbb{E}[n_k(T)]}_{\substack{\text{obtained} \\ \text{avg. reward}}} = \underbrace{\sum_{k \in [K]} (\mu^* - \mu_k) \mathbb{E}[n_k(T)]}_{\substack{\text{average "loss" from} \\ \text{suboptimal decisions}}}$$

Some Intuition

- **Exploration:** An agent can choose different arms to obtain a better estimate of their average rewards.
- **Exploitation:** An agent can repeatedly choose the arm it believes to be optimal.
- **Priors:** The agent may have prior information about the reward distributions.

The central dilemma is to balance exploration and exploitation, and efficiently use prior information (if available).

Thompson Sampling

Earliest heuristic for the problem [Tho33]; uses a Bayesian approach

- Assume a prior distribution over the reward params, $\theta_k \sim p(\cdot|\eta_k)$
- For $t \in [T]$, sample params for each arm from the posterior:

$$\hat{\theta}_k \sim p(\theta_k|\eta_k, r_k(1), r_k(2), ...)$$

- Choose action that maximizes mean given the posterior samples.If $\mu_k = f_k(\hat{\theta}_k)$ for some function f_k, then:

$$a(t) = \arg\max_{k \in [K]} \mu_k = \arg\max_{k \in [K]} f_k(\hat{\theta}_k)$$

- Update posterior for chosen arm with the reward recieved.

Performs very well in practice [CL11], with theoretical interest in its optimality [AG13, AG12].

Heavy-Tailed Distributions

Most of the existing analysis of the problem is on well-behaved rewards:

- bounded support (e.g. rewards are in [0, 1])
- sub-Gaussian (tails decay faster than Gaussian)

There is evidence, however, that suggests that real-world data exhibit very heavy tails:

- Stock prices [Nol03]
- Presence times in online networks [VOD+06]
- Labels in social media [MGR+18]

We want to design machine learning algorithms that are robust to heavy tails and provide more accurate decision-making in real-world scenarios.

α-Stable Distributions

α-Stable distributions comprise all continuous distributions that are closed under linear transformations.

- That is, if X and Y are stable, then $Z = X + Y$ is also stable.
  - e.g. Gaussian, Lévy, Cauchy
- Do not (generally) admit an analytical density function.
- Do not admit moments higher than order α, where $\alpha \leq 2$.
  - i.e., have infinite variance, and are heavy-tailed (except for $\alpha = 2$).
- The empirical mean exhibits polynomial deviations.
  - i.e. we cannot use typical Chernoff bounds (MGF does not exist).

This Work

- **Problem Setting:** We are given a K-armed stochastic bandit problem, where rewards are drawn i.i.d. from symmetric α-stable distributions where $\alpha \in (1, 2)$.
- **Contributions:**
  - Efficient algorithm for approximate Bayesian inference under α-stable rewards.
  - Finite-time regret bounds for naive posterior sampling using the empirical mean.
  - Near-optimal regret bounds for posterior sampling using a robust mean.

Our Approach: Algorithm

- Symmetric α-stable distributions can be considered as a case of *scale mixtures of normals*:
  - i.e. they can be considered as a weighted mixture of a Gaussian density with another α-stable distribution.

$$\underbrace{p_X(x|\mu)}_{\substack{\alpha\text{-stable} \\ \text{density}}} = \int_0^\infty \underbrace{\mathcal{N}(x|\mu, \lambda\sigma^2)}_{\substack{\text{normal} \\ \text{density}}} \cdot \underbrace{p_\Lambda(\lambda)}_{\substack{\text{known } \alpha\text{-stable} \\ \text{density}}} \cdot d\lambda$$

  - This implies that given samples of an auxiliary variable λ, the conditional density of rewards is Gaussian.

$$p(x|\lambda, \mu) \sim \mathcal{N}(\mu, \lambda\sigma^2)$$

  - Thus, we can use a (conjugate) Gaussian prior on the mean rewards:

$$p(x|\lambda) \sim \mathcal{N}(\mu, \lambda\sigma^2) \cdot \mathcal{N}(\mu_0, \eta^2)$$

α-Thompson Sampling

- **Input**: Arms $k \in [K]$, priors $\mathcal{N}(\mu_k^0, \sigma^2)$ for each arm.
- Set $D_k = 1, N_k = 0$ for each arm k.
- For each iteration $t \in [1, T]$:
  - Draw $\bar{\mu}_k(t) \sim \mathcal{N}\left(\frac{\mu_k^0 + N_k}{D_k}, \frac{\sigma^2}{D_k}\right)$ for each arm k.
  - Choose arm $A_t = \arg\max_{k \in [K]} \bar{\mu}_k(t)$, and get reward r_t.
  - Draw $\lambda_k^{(t)}$ using a rejection sampler.
- Set $D_k = D_k + 1/\lambda_k^{(t)}, N_k = N_k + r_t/\lambda_k^{(t)}$.

Our Approach: Analysis

- α-stable distributions do not allow an analytic probability density.
  - We therefore work with the characteristic function (Fourier Transform) of the probability density:

$$\phi(z) = \int_{-\infty}^{\infty} p(x) e^{-2\pi i z x} dx$$

  - We derive concentration results that bound how fast the empirical mean converges to the true mean under α-stable distributions.These results are of independent interest in robust machine learning theory, and are critical for our regret analysis.
  - Using the concentration results, we bound the Bayes regret of our algorithm via an upper confidence-bound decomposition.

Our Contributions

α-Thompson Sampling

We use an auxiliary variable to obtain an efficient algorithm for posterior updation, and subsequently perform Thompson Sampling when rewards are from α-stable distributions.We obtain the first *problem-independent* bound of order $O(K^{\frac{1}{1+\epsilon}} T^{\frac{1+\epsilon}{1+2\epsilon}})$ on the finite-time Bayes Regret of this algorithm.

Robust α-Thompson Sampling

Using a robust mean estimator we propose a version of Thompson Sampling for α-stable bandits that incurs a tight Bayesian Regret of $\tilde{O}\left((KT)^{\frac{1}{1+\epsilon}}\right)$, matching the lower bound (up to logarithmic factors).

References

Shipra Agrawal and Navin Goyal, *Analysis of thompson sampling for the multi-armed bandit problem*, Conference on Learning Theory, 2012, pp. 39–1.

_____, *Further optimal regret bounds for thompson sampling*, Artificial Intelligence and Statistics, 2013, pp. 99–107.

Olivier Chapelle and Lihong Li, *An empirical evaluation of thompson sampling*, Advances in neural information processing systems, 2011, pp. 2249–2257.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten, *Exploring the limits of weakly supervised pretraining*, The European Conference on Computer Vision (ECCV), September 2018.

John P Nolan, *Modeling financial data with stable distributions*, Handbook of heavy tailed distributions in finance, Elsevier, 2003, pp. 105–130.

William R Thompson, *On the likelihood that one unknown probability exceeds another in view of the evidence of two samples*, Biometrika **25** (1933), no. 3/4, 285–294.

Alexei Vázquez, Joao Gama Oliveira, Zoltán Dezsö, Kwang-Il Goh, Imre Kondor, and Albert-László

The End