Minimal Exploration in Structured Stochastic Bandits

Richard Combes (Centrale-Supelec, L2S), Stefan Magureanu (KTH) and Alexandre Proutiere (KTH)

1. Our Contribution

We investigate the stochastic Multi-Armed Bandit with finitely many arms and generic structure. We provide a generic regret lower bound and design OSSB, a generic algorithm that is asymptotically optimal for any structured MAB. We further provide a finite time analysis of OSSB.

2. Model

We consider the most general model for a stochastic structured MAB.

- The set of arms $\mathcal{X}$ is finite
- Problem is parameterized by an unknown parameter $\theta \in \Theta$
- When arm $x \in \mathcal{X}$ is selected, one observes $Y(n,x) \sim \nu(\theta(x))$ with expectaton $\theta(x)$
- Successive observations from arm x, $(Y(n,x))_n$ are i.i.d.
- When arm $x \in \mathcal{X}$ is selected, one receives a (not observed) reward $\mu(x,\theta)$
- The mapping $(x,\theta) \mapsto \mu(x,\theta)$ is known, θ is unknown
- The goal is design a policy π minimizing the regret when T is large, with $x^\pi(t)$ is the arm selected at round t by policy π:

$$R^\pi(T,\theta) = T \max_{x\in\mathcal{X}} \mu(x,\theta) - \sum_{t=1}^{T} \mathbb{E}(\mu(x^\pi(t),\theta)).$$

3. Structures Covered by Our Model

This model covers many popular bandit models of the literature.

- Classical Bandits: the parameter set is $\Theta = [0,1]^{|\mathcal{X}|}$, the observation distribution $\nu(\theta(x))$ is any bounded distribution with mean $\theta(x)$, and for all $x \in \mathcal{X}$ the reward equals the mean observation $\mu(x,\theta) = \theta(x)$.
- Linear Bandits: the set of arms $\mathcal{X}$ is a finite subset of $\mathbb{R}^d$; the parameter set Θ is the set of θ such that $\theta(x) = \langle \phi, x \rangle$ for all $x \in \mathcal{X}$ for some $\phi \in \mathbb{R}^d$; the observation distribution $\nu(\theta(x))$ is a Gaussian distribution with unit variance and mean $\theta(x)$, and the reward equals the mean observation $\mu(x,\theta) = \theta(x)$.
- Dueling Bandits: The set of arms $\mathcal{X}$ is $\{1,\dots,d\}^2\}$, and arms are $x = (i,j)$. Parameter θ is a preference matrix such that $\theta(i,j) = 1 - \theta(j,i)$, and $\theta(i,i) = \frac{1}{2}$, and there exists $i^\star$ (Condorcet winner) such that $\min_{i \neq i^\star} \theta(i^\star, i) > \frac{1}{2}$, the observation distribution $\nu(a)$ is the Bernoulli distribution with mean a; the rewards are $\mu((i,j),\theta) = \frac{1}{2}(\theta(i^\star,i) + \theta(i^\star,j) - 1)$. Note: the best arm is $(i^\star, i^\star)$ and has zero reward.
- Lipschitz Bandits: The set of arms $\mathcal{X}$ is a finite metric space endowed with a distance ℓ. The mapping $x \mapsto \theta(x)$ is Lipschitz continuous with respect to ℓ, and the set of parameters is:

$$\Theta = \{\theta : |\theta(x) - \theta(y)| \le \ell(x,y) \quad \forall x,y \in \mathcal{X}\}.$$

  the reward equals the mean observation $\mu(x,\theta) = \theta(x)$.
- Unimodal Bandits: The set of arms $\mathcal{X}$ is $\mathcal{X} = \{1,\dots,|\mathcal{X}|\}$, and finite and the set of parameters Θ is the set of unimodal function i.e. $x \mapsto \theta(x)$ is unimodal: it is strictly increasing on $\{1,\dots,x^\star\}$ and strictly decreasing on $\{x^\star,\dots,|\mathcal{X}|\}$. The reward equals the mean observation $\mu(x,\theta) = \theta(x)$.

4. Regret Lower Bounds

Assumption 1. The optimal arm $x^\star(\theta)$ is unique.

Theorem 1. Let $\pi \in \Pi$ be a uniformly good algorithm. For any $\theta \in \Theta$, we have:

$$\liminf_{T\to\infty} \frac{R^\pi(T,\theta)}{\ln T} \ge C(\theta), \tag{1}$$

where $C(\theta)$ is the value of the optimization problem:

$$\underset{\eta(x)\ge 0,\, x\in\mathcal{X}}{\text{minimize}} \sum_{x\in\mathcal{X}} \eta(x)(\mu^\star(\theta) - \mu(x,\theta)) \tag{2}$$

$$\text{s.t.} \sum_{x\in\mathcal{X}} \eta(x) D(\theta,\lambda,x) \ge 1\,,\ \forall \lambda \in \Lambda(\theta), \tag{3}$$

where $D(\theta,\lambda,x)$ is the Kullback-Leibler divergence between $\nu(\theta(x))$ and $\nu(\lambda(x))$ and:

$$\Lambda(\theta) = \{\lambda \in \Theta : D(\theta,\lambda,x^\star(\theta)) = 0, x^\star(\theta) \neq x^\star(\lambda)\}. \tag{4}$$

is the set of parameters λ where the optimal arm $x^\star(\lambda)$ is different from $x^\star(\theta)$ and cannot be distinguished from θ by sampling $x^\star(\theta)$.

5. Bounds for Specific Structures

The regret lower bound covers previously known lower bounds for specific structured bandits. Also, the solution of (2)-(3) is often tractable.

- Classical bandits: (Lai, 1985) the solution of (2)-(3) is:

$$c(x,\theta) = \frac{1}{d(\theta(x),\theta(x^\star))}.$$

- Linear bandits: for Gaussian rewards (Lattimore et al. 2016), (2)-(3) is equivalent to the convex program:

$$\underset{\eta(x)\ge 0,\, x\in\mathcal{X}}{\text{minimize}} \sum_{x\in\mathcal{X}} \eta(x)(\theta(x^\star) - \theta(x))$$

$$\text{s.t. } x^\top \text{inv}\left(\sum_{z\in\mathcal{X}} \eta(z) z z^\top\right) x \le \frac{(\theta(x^\star) - \theta(x))^2}{2}, \forall x \neq x^\star.$$

- Lipschitz bandits: for Bernoulli rewards (Magureanu et al. 2014), (2)-(3) is equivalent to a linear program ($|\mathcal{X}|$ variables and $2|\mathcal{X}|$ constraints):

$$\underset{\eta(x)\ge 0,\, x\in\mathcal{X}}{\text{minimize}} \sum_{x\in\mathcal{X}} \eta(x)(\theta(x^\star) - \theta(x))$$

$$\text{s.t.} \sum_{z\in\mathcal{X}} \eta(z) d(\theta(z), \max\{\theta(z), \theta(x^\star) - \ell(x,z)\}) \ge 1\,,\ \forall x \neq x^\star.$$

- Dueling bandits: if there exists a Condorcet winner $i^\star$ (Komiyama, 2016), the solution of (2)-(3) is: (where $x = (i,j)$)

$$c(x,\theta) = \mathbb{1}\left\{ j \in \arg\min_{j'} \frac{\mu((i,j'),\theta)}{d(\theta(i,j'),1/2)} \right\} \frac{1}{d(\theta(i,j),1/2)}.$$

- Unimodal bandits: (Combes et al. 2014) the solution of (2)-(3) is:

$$c(x,\theta) = \frac{\mathbb{1}\{|x - x^\star| = 1\}}{d(\theta(x),\theta(x^\star))}.$$

6. The OSSB Algorithm

OSSB(ε,γ) Pseudocode.

$s(0) \leftarrow 0, N(x,1), m(x,1) \leftarrow 0\,, \forall x \in \mathcal{X}$ {Initialization}
for $t = 1,\dots,T$ do
  Compute the optimization problem (2)-(3) solution $(c(x,m(t)))_{x\in\mathcal{X}}$ where $m(t) = (m(x,t))_{x\in\mathcal{X}}$
  if $N(x,t) \ge c(x,m(t))(1+\gamma)\ln\, t, \forall x$ then
    $s(t) \leftarrow s(t-1)$
    $x(t) \leftarrow x^\star(m(t))$ {Exploitation}
  else
    $s(t) \leftarrow s(t-1) + 1$
    $\overline{X}(t) \leftarrow \arg\min_{x\in\mathcal{X}} \frac{N(x,t)}{c(x,m(t))}$
    $\underline{X}(t) \leftarrow \arg\min_{x\in\mathcal{X}} N(x,t)$
    if $N(\underline{X}(t),t) \le \varepsilon s(t)$ then
      $x(t) \leftarrow \underline{X}(t)$ {Estimation}
    else
      $x(t) \leftarrow \overline{X}(t)$ {Exploration}
    end if
  end if
  Play $x(t)$ and update statistics.
end for

OSSB(ε,γ) is provably asymptotically optimal (we give a finite time analysis).

Assumption 2. For all x, the mapping $(\theta,\lambda) \mapsto D(x,\theta,\lambda)$ is continuous at all points where it is not infinite.

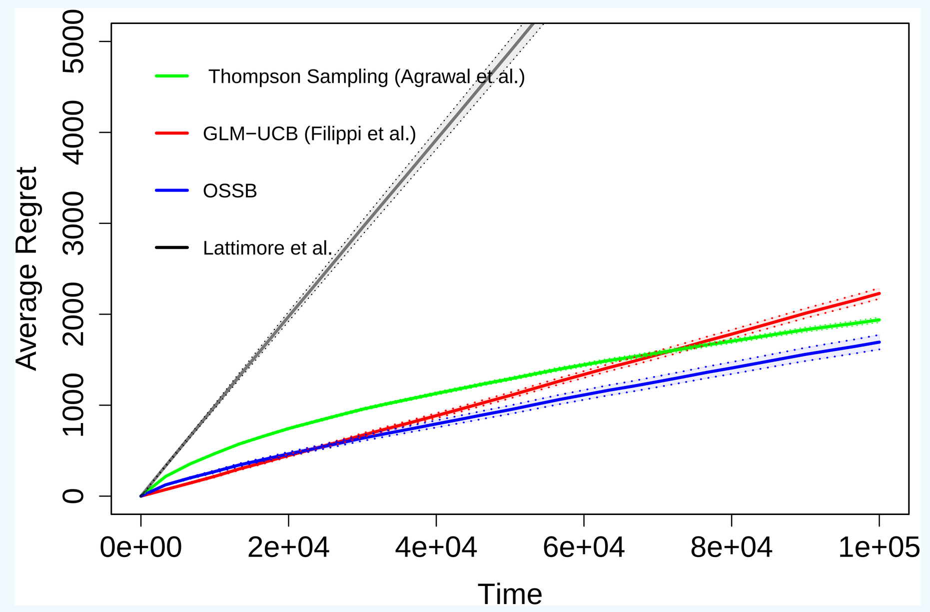
Assumption 3. For all x, the mapping $\theta \to \mu(x,\theta)$ is continuous.

Theorem 2. Under Assumptions 1, 2 and 3, for Bernoulli and Gaussian observation, then under the algorithm $\pi =$OSSB(ε,γ) with $\varepsilon < \frac{1}{|\mathcal{X}|}$ we have:

$$\limsup_{T\to\infty} \frac{R^\pi(T)}{\ln T} \le C(\theta) F(\varepsilon,\gamma,\theta),$$

with F a function such that for all θ, we have $F(\varepsilon,\gamma,\theta) \to 1$ as $\varepsilon,\gamma \to 0$.

7. Numerical Results



Performance of OSSB(0,0) vs. state-of-the-art algorithms.

- We consider linear bandits with 81 arms and 10 random instances.
- OSSB works well in finite time (competitive with the state of the art).
- Since OSSB is more complex to implement than other algorithms, reducing its complexity is an interesting topic of future research.