Contextual Bandits with Continuous Actions:

Smoothing, Zooming, and Adapting

Akshay Krishnamurthy, John Langford, Aleksandrs Slivkins, Chicheng Zhang

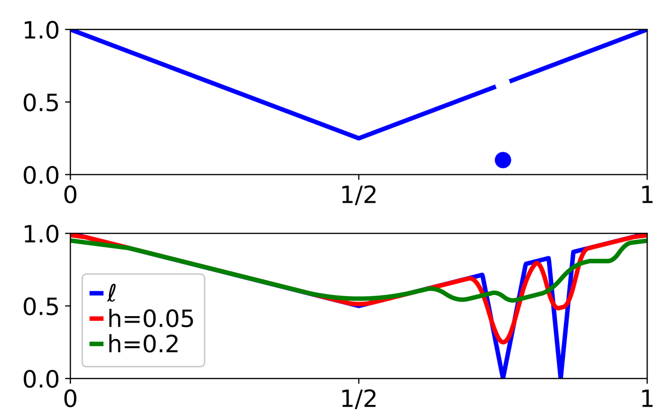
Microsoft Research

The contextual bandit protocol

For round $t = 1, \ldots, T$:

- Observe context $x_t \in \mathcal{X}$ (nature chooses $\ell_t : \mathcal{A} \mapsto [0,1]$)
- Choose action $a_t \in \mathcal{A}$ (possibly randomized)
- Observe loss $\ell_t(a_t) \in [0,1]$

$$\mathrm{Regret}(T, \Pi) \triangleq \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t)\right] - \min_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\pi(x_t))\right]$$



How do we handle continuous action spaces in the contextual bandit protocol?

- Contextual bandits with finite action sets well studied, regret scales with number of actions.
- Lipschitz bandits is a special case, but requires smoothness assumptions.

- **Main idea:** We replace actions with "smoothed" actions and policies with smoothed policies, enabling standard techniques, which we refine.
- **The point:** No smoothness assumptions required. They are baked into the benchmark. (We recover existing results with smoothness.)

Smoothed Regret

Definition 1. *For bandwidth $h \geq 0$, define $\texttt{Smooth}_h : \mathcal{A} \to \Delta(\mathcal{A})$ and policy class*

$$\Pi_h = \{\texttt{Smooth}_h(\pi) : x \mapsto \texttt{Smooth}_h(\pi(x)) : \pi \in \Pi\}$$

New performance measure: $\mathrm{Regret}(T, \Pi_h)$.

- **Example:** $\mathcal{A} = [0,1]$ with metric $\rho(a, a') = |a - a'|$. $\texttt{Smooth}_h(a) = \mathrm{Unif}(\{a' : \rho(a,a') \leq h\})$.
- **Intuition:** Smoothing allows us to focus on "estimation error," yielding assumption-free results.
- **Intuition:** The "smoothed loss" $\ell(a) \triangleq \mathbb{E}_{b \sim \texttt{Smooth}_h(a)} \ell(b)$ is $1/h$-Lipschitz, so prior work yields $O(T^{2/3}(1/h)^{1/3})$ in non-contextual setting (generalizes to $O(T^{2/3}(1/h \log|\Pi|)^{1/3})$ for contextual setting).

Theorem 2. *In the adversarial setting, for $h \geq 0$, $\texttt{EXP4}$ with $\Xi = \Pi_h$ guarantees*

$$\mathrm{Regret}(T, \Pi_h) \leq O\left(\sqrt{T/h \log|\Pi|}\right).$$

Main observation: For policy $\xi : x \mapsto \mathrm{Unif}(\{a' : \rho(\pi(x), a') \leq h\})$, we have

$$\mathop{\mathbb{E}}_{a_t, \xi} \hat{\ell}_t(\xi)^2 \leq \frac{1}{h} \int \mathop{\mathbb{E}}_{\xi \sim P_t} \frac{\xi(a \mid x_t)}{p_t(a \mid x_t)} d\nu(a) \leq 1/h.$$

Otherwise standard proof is unchanged!

- Optimal for adversarial setting with no further assumptions.
- For L-lipschitz losses by tuning h, we get $O(T^{2/3}(L \log|\Pi|)^{1/3})$ regret for Lipschitz (contextual) bandits, recovering the existing optimal rate.
- But doesn't require smoothness assumptions to get meaningful guarantee!

Zooming CB

Question: Better regret for benign instances?
Answer: Yes, by generalizing prior "zooming" algorithms.

Stochastic setting where $(x_t, \ell_t) \sim \mathcal{D}$ for each t.

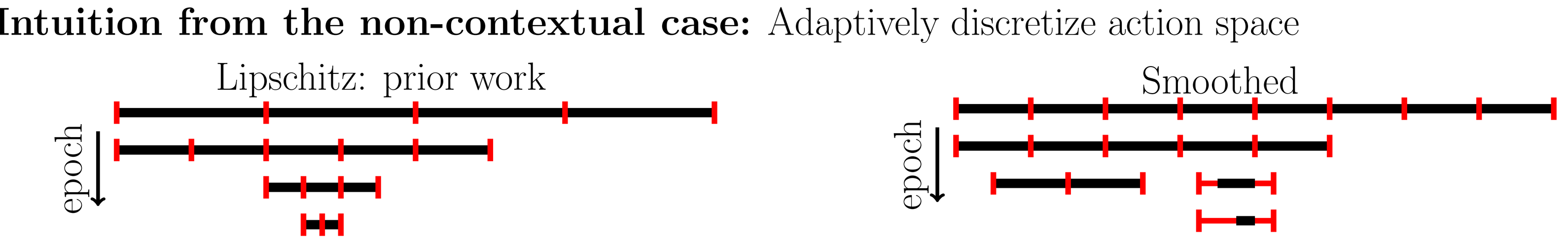
Algorithm 1 `SmoothPolicyElimination`

Set $\Pi^{(1)} = \Pi$.
for each epoch $m = 1, 2, \ldots,$ **do**
  Set $V_m = \mathbb{E}_{x \sim \mathcal{D}}\, \nu(\bigcup_{\pi \in \Pi^{(m)}} B_h(\pi(x)))$.
  Set radius $r_m = 2^{-m}$, epoch length $n_m \approx \frac{V_m \log|\Pi|}{r_m^2 h}$, exploration probability $\mu_m = r_m$.
  Find distribution Q_m over Π_m minimizing

$$\max_{\pi \in \Pi^{(m)}} \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \mathop{\mathbb{E}}_{a \sim \texttt{Smooth}_h(\pi(x))} \left[\frac{1}{q_m(a \mid x)}\right],$$
$$q_m(a \mid x) = \mu_m + (1 - \mu_m) \mathop{\mathbb{E}}_{\pi \sim Q_m} \texttt{Smooth}_h(\pi(x_t))$$

  For each of n_m rounds, observe x_t play $a_t \sim q_m(\cdot \mid x_t)$ observe loss $\ell_t(a_t)$.
  For each $\pi \in \Pi^{(m)}$, let $\hat{L}_m(\pi)$ be the median-of-means importance weighted estimate of $\mathbb{E}\,\ell(\pi(x))$.
  Set $\Pi^{(m+1)} = \left\{\pi \in \Pi^{(m)} : \hat{L}_m(\pi) \leq \min_{\pi' \in \Pi^{(m)}} \hat{L}_m(\pi') + 3r_m\right\}$.

Intuition from the non-contextual case: Adaptively discretize action space



Theorem 3. *For $h \geq 0$, `SmoothPolicyElimination` guarantees*

$$\mathrm{Regret}(T, \Pi_h) \leq \tilde{O}\left(\inf_{\epsilon_0 \geq 1/T} T\epsilon_0 + \theta_h(\epsilon_0) \cdot \log(|\Pi|)\right).$$

For L Lipschitz losses, a similar algorithm guarantees (set $h_m = 2^{-m}, r_m = L2^{-m}$)

$$\mathrm{Regret}(T, \Pi) \leq \tilde{O}\left(\inf_{\epsilon_0 \geq 1/T} TL\epsilon_0 + \psi_L(\epsilon_0)/L \cdot \log(|\Pi|)\right).$$

Smoothing and zooming coefficients θ_h, ψ_L are small in favorable instances.

Remarks

- Coefficients, θ_h and ψ_L measure size of action space for good policies. Small in favorable instances.
- Best case: smoothed $\sqrt{T \log|\Pi|} + 1/h \log|\Pi|$, Lipschitz $\sqrt{T \log|\Pi|}$.
- Lipschitz result generalizes "zooming dimension" results to contextual case.
- Akin to gap-dependent bound.

Key ideas

- Refined analysis so variance scales with "characteristic volume" V_m.
- Duality certifies small objective value, which controlling variance of loss estimates.
- Median-of-means avoids range dependence (uniform probability insufficient!).

Zooming coefficients: Let $M_h(\epsilon, \delta) \triangleq \mathbb{E}_{x \sim \mathcal{D}}[\mathcal{N}_\delta(\Pi_{h,\epsilon}(x))]$, where $\mathcal{N}$ is the covering number at scale δ and $\Pi_{h,\epsilon}(x) = \{\pi(x) : \mathbb{E}_{\mathcal{D}}\,\ell(\pi_h(x')) \leq \min_{\pi \in \Pi} \mathbb{E}_{\mathcal{D}}\,\ell(\pi_h(x')) + \epsilon\}$.

$$\theta_h(\epsilon_0) \triangleq \sup_{\epsilon \geq \epsilon_0} M_h(12\epsilon, h)/\epsilon, \qquad \psi_L(\epsilon_0) \triangleq \sup_{\epsilon \geq \epsilon_0} M_0(12L\epsilon, \epsilon)/\epsilon.$$

We have $\max\{1/\epsilon, 1/h\} \leq \theta_h(\epsilon) \leq 1/(h\epsilon)$ and $1/\epsilon \leq \psi_L(\epsilon) \leq 1/\epsilon^2$.

Adaptive CB

Question: Can we compete with Π_h for all h? Can we adapt to Lipschitz constant?

Theorem 4. *Fix $\alpha \in [0,1]$. `Corral` with `EXP4` (with parameter α) guarantees*

$$\forall h \in (0,1] : \mathrm{Regret}(T, \Pi_h) \leq \tilde{O}\left(T^{\frac{1}{1+\alpha}} h^{-\alpha}\right) \cdot (\log|\Pi|)^{\frac{\alpha}{1+\alpha}}.$$

The same algorithm is Lipschitz-adaptive with rate $\tilde{O}(T^{\frac{1+\alpha}{1+2\alpha}} L^{\frac{\alpha}{1+\alpha}}) \cdot (\log|\Pi|)^{\frac{\alpha}{1+2\alpha}}$. These are the optimal adaptive rates for their respective settings for the non-contextual case.

Remarks

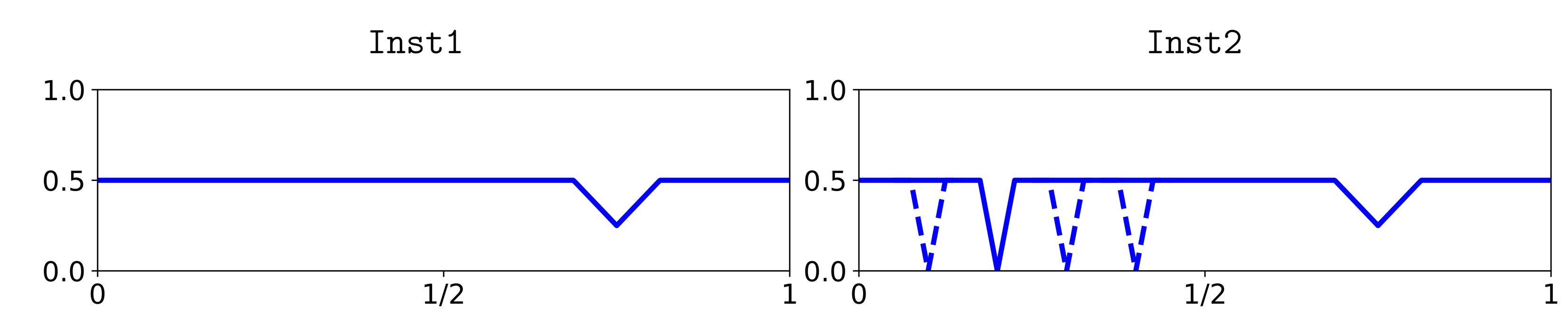
- With $\alpha = 1$ we get $\sqrt{T}/h$ and $T^{2/3}\sqrt{L}$. With $\alpha = 1/2$ we get $T^{2/3}/\sqrt{h}$ or $T^{3/4}L^{1/3}$.
- Lower bounds demonstrate a price of adaptivity, related to Locatelli and Carpentier (2018).
- α traces out Pareto frontier of optimal bounds, which are all incomparable.
- Lipschitz-adaptive algorithms know T, Π and *nothing else*, unlike much prior work. With extra information, it is possible to obtain the optimal non-adaptive rates.

Upper bounds: The algorithm is `Corral` (Agarwal et al., 2016) using copies of `EXP4` as base learners. `Corral` is just online mirror descent with the log-barrier regularizer, where each "arm" is a bandit algorithm. For $\alpha = 1$, when base learners use bandwidths $\mathcal{H}$, `Corral` ensures

$$\forall h \in \mathcal{H} : \mathrm{Regret}(T, \Pi_h) \leq \tilde{O}\left(\frac{|\mathcal{H}|}{\eta} + T\eta + \frac{T\eta}{h} \cdot \log(|\Pi|)\right)$$

- Tuning η, ignoring h, and chosing a logarithmically spaced grid $\mathcal{H}$ gives the Lipschitz result.
- For smoothed regret, since we want all $h \in (0,1]$, some more tricks are needed!

Lower bounds: Inspired by construction of Locatelli and Carpentier (2018).



If learner does well in `Inst1`, it cannot explore enough to find a needle hidden in $[0, 1/2]$ in `Inst2`.

Generalizations

- Results extend to arbitrary metric spaces.
- Results extends to `Smooth` given by a *kernel* $K : a \mapsto \Delta(\mathcal{A})$.
  - Analog of $1/h$ is $\kappa \triangleq \sup_{a,a'} |(Ka)(a')|$, the density value w.r.t., base measure ν.
- **Example:** Finite $\mathcal{A} = \{i/M : i \in [M]\}$ and identity metric recovers standard (contextual) bandits. But can also use non-degenerate metric and kernel to share informationa across actions.
- Can also obtain results for non-contextual case.

Open problem: Computationally (oracle) efficient algorithms for continuous action spaces?

References
1. Agarwal, Luo, Neyshabur, and Schapire. Corralling a band of bandit algorithms. In COLT, 2016.
2. Locatelli and Carpentier. Adaptivity to smoothness in $\mathcal{X}$-armed bandits. In COLT, 2018.

Learn more at: `https://arxiv.org/abs/1902.01520`