Model Selection for Contextual Bandits

Dylan Foster, Akshay Krishnamurthy, and Haipeng Luo

dylanf@mit.edu, akshay@cs.umass.edu, haipengl@usc.edu

USC University of Southern California

Microsoft Research

Overview

Contextual bandits

For $t = 1, \dots, T$:

- Receive context $x_t \in \mathcal{X}$.
- Predict action $a_t \in \mathcal{A} := \{1, \dots, K\}$.
- Receive loss: $\ell_t(a_t)$.

Stochastic variant:

- Assume $(x_t, \ell_t) \sim \mathcal{D}$ i.i.d.
- Learner has access to regression function class $\mathcal{F} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, where

$$\mathbb{E}[\ell(a) \mid x] = f^\star(x, a)$$

  for some $f^\star \in \mathcal{F}$.
- Induced policies: $\pi_f(x) = \operatorname{argmin}_{a \in \mathcal{A}} f(x, a)$.

Goal: Low regret against best policy $\pi^\star := \pi_{f^\star}$:

$$\mathrm{Reg} := \sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(\pi^\star(x_t)). \tag{1}$$

Our question

- Previous results in contextual bandits assume $\mathcal{F}$ is given.
- Model selection [Vapnik'82]: Choose $\mathcal{F}$ in data-dependent fashion.

Can model selection guarantees be achieved in contextual bandit learning?

More broadly, we seek to understand the algorithmic principles and fundamental limits of model selection in interactive settings.

The model selection problem for contextual bandits

- For fixed $\mathcal{F}$, typically expect $\mathrm{Reg} = \sqrt{T\mathrm{comp}(\mathcal{F})}$, e.g. $\mathrm{comp}(\mathcal{F}) = \log|\mathcal{F}|$ for finite classes, $\mathrm{comp}(\mathcal{F}) = d$ for d-dimensional linear classes.
- Assume that $\mathcal{F}$ is structured as a nested sequence of classes

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{F}_M = \mathcal{F},$$

  and define $m^\star = \min\{m : f^\star \in \mathcal{F}_m\}$.
- The model selection problem asks:

  Given that $m^\star$ is not known in advance, can we achieve regret scaling as $O(\sqrt{T\mathrm{comp}(\mathcal{F}_{m^\star})})$, rather than the less favorable $O(\sqrt{T\mathrm{comp}(\mathcal{F})})$?
- Also acceptable: $O(T^{1-\alpha}\mathrm{comp}^\alpha(\mathcal{F}_{m^\star}))$ for any $\alpha \leq 1/2$.

Our result: Model selection for linear contextual bandits

- We assume that each class $\mathcal{F}_m$ consists of linear functions of the form

$$\mathcal{F}_m := \left\{(x, a) \mapsto \langle \beta, \phi^m(x, a)\rangle \mid \beta \in \mathbb{R}^{d_m}\right\},$$

  where $\phi^m : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^{d_m}$ is a fixed feature map.
- If $m^\star$ is known, optimal regret is $\sqrt{d_{m^\star}T}$ [Chu et al.'11].

Main result: With *no prior knowledge* of $m^\star$, we achieve:

$$\mathbf{Reg} = \tilde{O}(\sqrt{d_{m^\star}T} + T^{3/4})$$

Also achieve $\mathbf{Reg} = \tilde{O}(d_{m^\star}^{1/3}T^{2/3})$, which is tighter for $d_{m^\star} \leq T^{1/4}$.

Only positive model selection result we are aware of for *any* contextual bandit setting.

Overview of main result

Statistical assumptions:

- Nested maps: ϕ^m has ϕ^{m-1} as first d_{m-1} features.
- Feature maps ϕ^m and losses $\ell(a)$ are τ-subgaussian.
- For all m, $\Sigma_m := \frac{1}{K}\sum_{a \in \mathcal{A}} \mathbb{E}_{x \sim \mathcal{D}}\left[\phi^m(x, a)\phi^m(x, a)^\top\right] \succeq \gamma I$.

Theorem

ModCB (Algorithm 2) with preprocessing guarantees that with probability at least $1 - \delta$,

$$\mathrm{Reg} \leq \begin{cases} \tilde{O}\left(\frac{\tau^4}{\gamma^3} \cdot (Kd_{m^\star})^{1/3}(T \log m^\star)^{2/3}\right), & \kappa = 1/3. \\ \tilde{O}\left(\frac{\tau^3}{\gamma^2} \cdot K^{1/4}(T \log m^\star)^{3/4} + \frac{\tau^5}{\gamma^4} \cdot \sqrt{KTd_{m^\star} \log m^\star}\right), & \kappa = 1/4. \end{cases}$$

Estimating prediction loss with sublinear # samples

Key idea: We can evaluate if a bigger model would improve error without actually learning the model!

Consider following "residual estimation" setup:

- Receive pairs $(x_1, y_1), \dots, (x_n, y_n)$ i.i.d. from a distribution $\mathcal{D} \in \Delta(\mathbb{R}^d \times \mathbb{R})$, where $x \sim \mathsf{subG}_d(\tau^2)$ and $y \sim \mathsf{subG}(\sigma^2)$. Define $\Sigma = \mathbb{E}\left[xx^\top\right]$.
- Suppose x can be partitioned into features $x = (x^{(1)}, x^{(2)})$, where $x^{(1)} \in \mathbb{R}^{d_1}$ and $x^{(2)} \in \mathbb{R}^{d_2}$, and $d_1 + d_2 = d$. Define

$$\beta^\star = \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \mathbb{E}(\langle \beta, x\rangle - y)^2, \quad \text{and} \quad \beta_1^\star = \operatorname*{argmin}_{\beta \in \mathbb{R}^{d_1}} \mathbb{E}\left(\langle \beta, x^{(1)}\rangle - y\right)^2.$$

Goal: estimate the residual error incurred by restricting to features $x^{(1)}$:

$$\mathcal{E} := \mathbb{E}\left(\langle \beta_1^\star, x^{(1)}\rangle - \langle \beta^\star, x\rangle\right)^2.$$

Algorithm 1 EstimateResidual

Inputs: Examples $\{(x_s, y_s)\}_{s=1}^n$, second moment matrix estimates $\widehat{\Sigma} \in \mathbb{R}^{d \times d}$, and $\widehat{\Sigma}_1 \in \mathbb{R}^{d_1 \times d_1}$.

Define $d_2 = d - d_1$ and

$$\widehat{R} = \widehat{D}^\dagger - \widehat{\Sigma}^\dagger \quad \text{where} \quad \widehat{D} = \begin{pmatrix} \widehat{\Sigma}_1 & 0_{d_1 \times d_2} \\ 0_{d_2 \times d_1} & 0_{d_2 \times d_2} \end{pmatrix}.$$

Return estimate

$$\widehat{\mathcal{E}} = \frac{1}{\binom{n}{2}} \sum_{s<t} \left\langle \widehat{\Sigma}^{1/2}\widehat{R}x_s y_s, \widehat{\Sigma}^{1/2}\widehat{R}x_t y_t \right\rangle.$$

Theorem

Suppose we take $\widehat{\Sigma}$ and $\widehat{\Sigma}_1$ to be the empirical second moment matrices formed from m iid unlabeled samples. Then once $m \geq C(d + \log(2/\delta))\tau^4/\lambda_{\min}(\Sigma)$, EstimateResidual guarantees that with probability at least $1 - \delta$,

$$\left|\widehat{\mathcal{E}} - \mathcal{E}\right| \leq \frac{1}{2}\mathcal{E} + \tilde{O}\left(\frac{\sigma^2\tau^4}{\lambda_{\min}^2(\Sigma)} \cdot \frac{d^{1/2}\log^2(2d/\delta)}{n} + \frac{\tau^6}{\lambda_{\min}^4(\Sigma)} \cdot \frac{d\log(2/\delta)}{m}\right).$$

- Generalizes sublinear "variance estimation" results of Dicker'14 and Kong and Valiant'18. Improves sample complexity by estimating difference in loss rather than loss itself.

New model selection algorithm

Algorithm 2 ModCB (Model Selection for Contextual Bandits)

Input:

- Feature maps $\{\phi^m(\cdot, \cdot)\}_{m \in [M]}$, where $\phi^m(x, a) \in \mathbb{R}^{d_m}$, and time $T \in \mathbb{N}$.
- Subgaussian parameter $\tau > 0$, second moment parameter $\gamma > 0$.
- Failure prob. $\delta \in (0, 1)$, exploration param. $\kappa \in (0, 1)$.

Definitions:

- Define $\delta_0 = \delta/10M^2T^2$ and $\mu_t = (K/t)^\kappa \wedge 1$.
- Define $\alpha_{m,t} = C_1 \cdot \left(\frac{\tau^6}{\gamma^4} \cdot \frac{d_m^{1/2}\log^2(2d_m/\delta_0)}{K^\kappa t^{1-\kappa}} + \frac{\tau^{10}}{\gamma^8} \cdot \frac{d_m \log(2/\delta_0)}{t}\right)$.
- Define $T_m^{\min} = C_2 \cdot \left(\frac{\tau^4}{\gamma^2} \cdot d_m \log(2/\delta_0) + \log^{\frac{1}{1-\kappa}}(2/\delta_0) + K\right) + 1$.

Initialization:

- $\widehat{m} \leftarrow 1$. `// Index of candidate policy class.`
- Exp4-IX$_1 \leftarrow$ Exp4-IX(Π_1, T, δ_0).
- $S \leftarrow \{\emptyset\}$. `// Times at which uniform exploration takes place.`

for $t = 1, \dots, T$ **do**

Receive x_t.

with probability $1 - \mu_t$

Feed x_t into Exp4-IX$_{\widehat{m}}$ and take a_t to be the predicted action.

Update Exp4-IX$_{\widehat{m}}$ with $(x_t, a_t, \ell_t(a_t))$.

otherwise

Sample a_t uniformly from $\mathcal{A}$ and let $S \leftarrow S \cup \{t\}$.

`/* Test whether we should move on to a larger policy class. */`

$\widehat{\Sigma}_i \leftarrow \frac{1}{K}\sum_{a \in \mathcal{A}}\sum_{s=1}^t \phi^i(x_s, a)\phi^i(x_s, a)^\top$ for each $i \geq \widehat{m}$.

$H_i \leftarrow \{(\phi^i(x_s, a_s), \ell(a_s))\}_{s \in S}$ for each $i > \widehat{m}$.

$\widehat{\mathcal{E}}_{\widehat{m},i} \leftarrow$ EstimateResidual$\left(H_i, \widehat{\Sigma}_{\widehat{m}}, \widehat{\Sigma}_i\right)$ for each $i > \widehat{m}$. `// Gap estimate.`

if there exists $i > \widehat{m}$ such that $\widehat{\mathcal{E}}_{\widehat{m},i} \geq 2\alpha_{i,t}$ and $t \geq T_i^{\min}$ **then**

Let $\widehat{m}$ be the smallest such i.

Re-initialize Exp4-IX$_{\widehat{m}} \leftarrow$ Exp4-IX$(\Pi_{\widehat{m}}, T, \delta_0)$.

end if

end for

Proof sketch for main theorem.

- Assume $m = 2$ and $\kappa = 1/4$ for simplicity.
- Two cases based on whether $f^\star \in \mathcal{F}_1$ or $f^\star \in \mathcal{F}_2$.

Case 1:

- With high probability, algorithm never switches from class Π_1.
- Total contribution to regret is $\tilde{O}(\sqrt{d_1 T})$ from Exp4-IX and $\tilde{O}(T^{3/4})$ from uniform exploration.

Case 2:

- Let $\hat{T}$ denote the first round where $\widehat{m} = 2$, or T if the algorithm never advances. Then regret is bounded as

$$\mathrm{Reg} \leq O\left(T^{3/4}\right) + \tilde{O}\left(\sqrt{\hat{T}d_1}\right) + \hat{T}\Delta_{1,2} + \tilde{O}\left(\sqrt{(T - \hat{T})d_2}\right),$$

  where $\Delta_{i,j} = L_i^\star - L_j^\star$, and $L_i^\star = \min_{\pi \in \Pi_i} L(\pi)$.
- **All that remains is to bound the gap $\Delta_{1,2}$.**
- Since we didn't switch until $\hat{T}$, EstimateResidual guarantees that

$$\Delta_{1,2}\hat{T} \leq \tilde{O}\left(\sqrt{\mathcal{E}_{1,2}\hat{T}}\right) \leq \tilde{O}\left(\hat{T}^{5/8}d_2^{1/4}\right) \leq \tilde{O}\left(T^{3/4} \vee \sqrt{d_2\hat{T}}\right).$$

Only works due to sublinear loss estimation guarantee! With naive estimator, any choice of μ_t and κ leads to vacuous guarantees!