Efficient Optimal Learning for Contextual Bandits

Miroslav Dudik†, Daniel Hsu‡, Satyen Kale†, Nikos Karampatziakis⋆, John Langford†, Lev Reyzin♯, Tong Zhang♭
Yahoo! Research†, Microsoft Research‡, Cornell University⋆, Georgia Institute of Technology♯, Rutgers University♭

Contextual bandit setting

For round $t = 1, 2, \ldots$

1. World presents context information as features $x_t \in \mathcal{X}$ from a feature space $\mathcal{X}$.
2. Learner chooses action $a_t \in A \doteq \{1, \ldots, K\}$ from among K possible actions.
3. World presents reward $r_t \in [0,1]$ for the chosen action a_t.

Note: learner does *not* see rewards for other actions $a \neq a_t$ in round t. The goal of the learner is to maximize its cumulative reward $\sum_{t=1}^T r_t$ over T rounds.

I.I.D. setting. Assume context and actions' rewards $(x, \vec{r})$ for each round are drawn independently from a fixed distribution D over $\mathcal{X} \times [0,1]^K$.

Regret to a policy class. Fix a set of N policies Π mapping contexts $x \in \mathcal{X}$ to actions $a \in A$. The *best policy* $\pi_{\max}$ maximizes the expected instantaneous reward

$$\eta_D(\pi) = \mathbb{E}_{(x,\vec{r})\sim D}[r(\pi(x))]$$

over all $\pi \in \Pi$. The *regret* to the expected performance of the best policy over T rounds is $\sum_{t=1}^T (\eta_D(\pi_{\max}) - r_t)$. The regret of the learner over T rounds is bounded by ϵ with probability at least $1-\delta$ if

$$\Pr\left[\sum_{t=1}^T (\eta_D(\pi_{\max}) - r_t) \leq \epsilon\right] \geq 1 - \delta$$

where the probability is taken over the random pairs $(x_t, \vec{r}_t) \sim D$ for $t = 1, \ldots, T$, and any internal randomness used by the learner.

Drawbacks of previous approaches

Previous algorithms for this setting are either *measure based* (hence, computationally inefficient in general), or *regret-suboptimal* (*i.e.*, regret bound after T time steps is $\omega(\sqrt{T})$).

Exp4/Exp4.P [1, 2] maintain weights for each policy based on an importance weighted estimate of its cumulative reward. Regret bound is $O(\sqrt{TK \log(N/\delta)})$ w.p. $\geq 1-\delta$, but the computation is $\Omega(N)$ in general.

ϵ-greedy/epoch-greedy [4] are efficient given a cost-sensitive learning algorithm, but have sub-optimal regret bound of $O(T^{2/3})$.

New algorithmic contributions

1. Algorithm PolicyElimination for contextual bandits achieving optimal regret bound $\tilde{O}(\sqrt{TK \log(N/\delta)})$; intuitively based on a *non-constructive minimax argument* for choosing a distribution over policies such that the reward estimates for each policy have low variance.
2. Algorithm RandomizedUCB, also achieving optimal regret bound $\tilde{O}(\sqrt{TK \log(N/\delta)})$; selection of distribution over policies in each round t can be computed in $\mathrm{poly}(t, \log(N))$ time, given a *cost-sensitive classification learning algorithm* for policy class Π.

Algorithm 1: PolicyElimination

PolicyElimination maintains a candidate set of policies, throwing out policies that are proved, using confidence intervals, to be suboptimal. Confidence interval for a policy π's reward is centered around an importance-weighted estimator

$$\eta_t(\pi) \doteq \frac{1}{t} \sum_{(x_\tau, a_\tau, r_\tau, p_\tau)\in h_t} \frac{r_\tau}{p_\tau} \cdot \mathbb{I}\{\pi(x_\tau) = a_\tau\}$$

based on the history after t rounds (*i.e.*, scale reward of selected action a_τ in round τ by $1/p_\tau$, set reward for other actions in round τ to zero).

For a distribution P over policies Π, let

$$W_P(x, a) \doteq \sum_{\pi\in\Pi} P(\pi) \cdot \mathbb{I}\{\pi(x) = a\}$$

denote the induced distribution over actions $a \in A$ given the context x.

Inputs: Π, δ, K, $D_{\mathcal{X}}$ (marginal of D over $\mathcal{X}$).
Initialize: $\Pi_0 \doteq \Pi$ and history $h_0 \doteq \emptyset$.
Define: $\delta_t \doteq \delta \,/\, 4Nt^2$, $b_t \doteq 2\sqrt{\frac{2K \ln(1/\delta_t)}{t}}$,

$$\mu_t \doteq \min\{\tfrac{1}{2K}, \sqrt{\tfrac{\ln(1/\delta_t)}{2Kt}}\}.$$

For each round $t = 1, \ldots, T$, observe x_t and do:

1. Choose distribution P_t over Π_{t-1} s.t. $\forall\, \pi \in \Pi_{t-1}$:
$$\mathbb{E}_{x\sim D_{\mathcal{X}}}\left[\frac{1}{(1-K\mu_t)W_{P_t}(x,\pi(x)) + \mu_t}\right] \leq 2K.$$
2. Let $W_t'(a) \doteq (1 - K\mu_t)W_{P_t}(x_t, a) + \mu_t$ for all $a \in A$.
3. Randomly choose action $a_t \sim W_t'$.
4. Observe reward r_t.
5. Let $\Pi_t \doteq \Big\{\pi \in \Pi_{t-1}:$
$$\eta_t(\pi) \geq \Big(\max_{\pi'\in\Pi_{t-1}} \eta_t(\pi')\Big) - 2b_t\Big\}.$$
6. Let $h_t \doteq h_{t-1} \cup (x_t, a_t, r_t, W_t'(a_t))$.

Algorithm 2: RandomizedUCB

RandomizedUCB selects a distribution P_t over policies Π by minimizing an estimate of the instantaneous regret, subject to a constraint that bounds the variance of future reward estimates. Differences from PolicyElimination: (i) chooses distribution P_t over all of Π via convex optimization, (ii) variance constraints use empirical estimate of marginal distribution $D_{\mathcal{X}}$ over $\mathcal{X}$, and are more slack for policies with larger regret.

For any policy $\pi \in \Pi$ and round t, let

$$\Delta_t(\pi) \doteq \max_{\pi'\in\Pi} \eta_t(\pi') - \eta_t(\pi)$$

denote the importance-weighted empirical instantaneous regret to (empirically) best policy through round t, and let $\Delta_t(W_Q) \doteq \mathbb{E}_{\pi\sim Q}[\Delta_t(\pi)]$ for any distribution Q over Π.

Inputs: Π, δ, K.
Initialize: history $h_0 \doteq \emptyset$.
Define: $C_t \doteq 2\log(Nt/\delta)$, $\mu_t \doteq \min\{\frac{1}{2K}, \sqrt{\frac{C_t}{2Kt}}\}$.
For each round $t = 1, \ldots, T$, observe x_t and do:

1. Let P_t be a distribution over Π that approximately solves the optimization problem
$$\min_P \sum_{\pi\in\Pi} P(\pi)\Delta_{t-1}(\pi)$$
s.t. for all distributions Q over Π:
$$\mathbb{E}_{\pi\sim Q}\widehat{\mathbb{E}}_{x_i\sim h_{t-1}}\left[\frac{1}{(1-K\mu_t)W_P(x_i, \pi(x_i)) + \mu_t}\right] \leq \max\left\{4K, \frac{(t-1)\Delta_{t-1}(W_Q)^2}{180C_{t-1}}\right\}.$$
2. Let $W_t'(a) \doteq (1 - K\mu_t)W_{P_t}(x_t, a) + \mu_t$ for all $a \in A$.
3. Randomly choose action $a_t \sim W_t'$.
4. Observe reward r_t.
5. Let $h_t \doteq h_{t-1} \cup (x_t, a_t, r_t, W_t'(a_t))$.

Minimax argument

View policies $\pi \in \Pi$ as functions in $\mathcal{X} \times A \to [0,1]$, with $\pi(x, a) \doteq \mathbb{I}\{\pi(x) = a\}$. For any distribution P over policies Π, W_P (a randomized policy) is a point in the convex hull of Π.

Lemma 1 *Let $\mathcal{C}$ be a compact and convex set of randomized policies (functions $W\colon \mathcal{X} \times A \to [0,1]$ s.t. $\sum_{a\in A} W(x, a) = 1$ for all $x \in \mathcal{X}$). Let $\mu \in (0, 1/K]$, and for any $W \in \mathcal{C}$, let $W'(x, a) \doteq (1 - K\mu)W(x, a) + \mu$.*
For all distributions $D_{\mathcal{X}}$ over $\mathcal{X}$,

$$\min_{W\in\mathcal{C}} \max_{Z\in\mathcal{C}} \mathbb{E}_{x\sim D_{\mathcal{X}}} \mathbb{E}_{a\sim Z(x,\cdot)}\left[\frac{1}{W'(x,a)}\right] \leq \frac{K}{1-K\mu}.$$

This lemma guarantees that the set of distributions over Π_{t-1} satisfying the constraints in Step 1 of PolicyElimination is non-empty, and hence P_t is well-defined. The induced distribution W_t' over actions is a mixture of W_{P_t} with the uniform distribution over actions that guarantees bounded variance for the importance-weighted reward estimates of policies in Π_t.

Using an $\arg\max$ oracle

RandomizedUCB can be implemented using an $\arg\max$ oracle ($\mathcal{AMO}$) for the policy class Π (*i.e.*, a cost-sensitive learner for Π [3]): given a sequence $((x_\tau, \vec{r}_\tau))_{\tau=1}^t$ in $\mathcal{X} \times \mathbb{R}_+^K$, the $\mathcal{AMO}$ returns

$$\arg\max_{\pi\in\Pi} \sum_{\tau=1}^t r_\tau(\pi(x_\tau)).$$

Specifically, the optimization problem in Step 1 of RandomizedUCB can be solved by the ellipsoid method using a separation oracle implemented with the $\mathcal{AMO}$.

References

[1] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. of Comp.*, 2002

[2] A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *AISTATS*, 2011.

[3] A. Beygelzimer, J. Langford, and P. Ravikumar. Error correcting tournaments. In *ALT*, 2009.

[4] J. Langford and T. Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *NIPS*, 2007.