Stat 260/CS 294-102. Learning in Sequential Decision Problems.

Peter Bartlett

1. Contextual bandits.
   - Infinite comparison classes.
     - Examples: parameterized policies.
     - Recall: finite ϵ-covers and Exp4.
       * Random ϵ-covers for VC classes.
     - Greedy optimization of regularized regret.

Recall: Contextual bandits

At each round:

- See $X_t \in \mathcal{X}$.
- Choose $I_t \in \mathcal{A}$, $\mathcal{A} = \{1, \ldots, k\}$.
- Receive reward $Y_{I_t,t} \in \mathbb{R}$.

Stochastic/adversarial model for $(X, Y) \in \mathcal{X} \times \mathbb{R}^{\mathcal{A}}$.

Pseudo-regret:

$$\overline{R}_n = \sup_{\pi \in \Pi} \mathbb{E} \sum_{t=1}^{n} Y_{\pi(X_t),t} - \mathbb{E} \sum_{t=1}^{n} Y_{I_t,t}.$$

where Π is *comparison class* of policies $\pi : \mathcal{X} \to \mathcal{A}$ (or the stochastic version, $\pi : \mathcal{X} \to \Delta^{\mathcal{A}}$).

Infinite comparison classes

For instance, linear threshold functions for $\mathcal{X} \subseteq \mathbb{R}^d$:

$$\pi(x) = \arg\max_{j \in \mathcal{A}} \phi(x, j)'\theta.$$

where $\theta \in \mathbb{R}^d$ is a parameter vector and $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ is a *feature map*.

Infinite comparison classes: Finite covers

Theorem: For a class Π with $S_\Pi(n) \leq Bn^d$, there is a strategy such that, under the i.i.d. stochastic model: $(X_t, Y_t) \sim P$, with probability $1-\delta$,

$$\overline{R}_n = O\left(\sqrt{nkd\log\frac{n}{d}}\right).$$

This strategy takes $\tilde{O}(n^{d/2})$ time per round.

Infinite comparison classes: Finite covers

Approach:

1. Construct a finite ϵ-cover $\hat{\Pi}$ of Π, with respect to the pseudo-metric

$$\rho(\hat{\pi}, \pi) = \Pr\left(\pi(X_t) \neq \hat{\pi}(X_t)\right).$$

   - e.g., construct a cover $\hat{\Pi}$ of Π from a cover of Θ.
   - e.g., construct a cover $\hat{\Pi}$ of Π as the (random) set of representatives of each element of

$$\left\{(\pi(X_1), \ldots, \pi(X_m)) : \pi \in \Pi\right\}.$$

2. Use Exp4 on $\hat{\Pi}$.

Combinatorial dimensions

- The size of $\hat{\Pi}$ is no more than

$$S_\Pi(n) := \max_{x_1,\ldots,x_n \in \mathcal{X}} \left|\{(\pi(x_1),\ldots,\pi(x_n)) : \pi \in \Pi\}\right|.$$

- Combinatorial dimensions (like the VC-dimension and its generalizations to k-valued functions) determine the rate of growth of $S_\Pi(n)$: for $d = d_{VC}(\Pi)$,

$$S_\Pi(n) \begin{cases} = 2^n & \text{if } n \leq d, \\ \leq (e/d)^d \, n^d & \text{if } n > d. \end{cases}$$

VC-dimension bounds for parameterized families

Consider a parameterized class of k-valued functions,

$$\Pi = \{x \mapsto f(x, \theta) : \theta \in \mathbb{R}^p\},$$

where $f : \mathbb{R}^m \times \mathbb{R}^p \to \{1, \ldots, k\}$.

Suppose that f can be computed using no more than t operations of the following kinds:

1. arithmetic ($+$, $-$, $\times$, $/$),
2. comparisons ($>$, $=$, $<$),
3. output a constant in $\{1, \ldots, k\}$

Theorem: $d_{VC}(\Pi) = O(pt \log k)$.

(And a similar story applies, with a worse dependence on t, if we include the exponential function in the set of operations.)

Summary: Infinite $\Pi \subseteq \{1, \ldots, k\}^{\mathcal{X}}$

- If any strategy can compete with an infinite Π for all distributions on $\mathcal{X} \times [0,1]^k$, then $S_\Pi(n)$ must have polynomial growth, say $O(n^d)$.
- In that case, we can use i.i.d. data to build an ϵ-cover of Π of size $O(S_\Pi(n)) = O(n^d)$.
- Running Exp4 with this class of experts gives regret

$$\overline{R}_n = O\left(\sqrt{nkd \log n}\right).$$

- The drawback is *computational*: $S_\Pi(n)$ is polynomial in n, but exponential in the dimension d. For example, for

$$\pi(x) = \arg\max_{j \in \mathcal{A}} \phi(x, j)'\theta,$$

the computation grows exponentially with the number of features.

An alternative approach: Reduction to classification

The high-level idea:

- Gather relevant data $(x_t, a_t, r_t(a_t), p_t(a_t))$. (Here, (x_t, r_t) are i.i.d.)
- Transform data to $(x, \ell) \in \mathcal{X} \times \mathbb{R}^{\mathcal{A}}$ pairs.
- Find a $\pi^t \in \Pi$ to minimize empirical risk,

$$\frac{1}{t}\sum_{s=1}^{t} \ell_s(\pi^t(x_s)).$$

- Use π^t to update how strategy makes subsequent choices.

Assumes we have access to an efficient empirical risk minimization oracle.

An alternative approach: Reduction to classification

Example: ϵ-greedy.

- With probability ϵ, explore: choose a_t uniformly.
- Otherwise, choose $a_t \sim \pi^t$.
- Use exploration data for losses,

$$\ell_t(a) = \frac{(1 - r_t(a_t))1[a = a_t]}{p_t(a_t)} = k(1 - r_t(a_t))1[a = a_t].$$

- Uniform convergence ensures π^t has per-trial regret $O(1/\sqrt{\epsilon t})$. Regret from exploration trials is $O(\epsilon n)$.
- Optimizing gives $\epsilon \sim n^{-1/3}$, with $\overline{R}_n = O\left(n^{2/3}\right)$.

(Or run ϵ-greedy with the doubling trick—also called *epoch-greedy*.)

Separating exploration and exploitation gives sub-optimal $\Omega(n^{2/3})$ regret.

Combining exploration and exploitation

- Maintain distribution q_t over Π.
- Observe x_t, choose $a_t \sim p_t$ where

$$p_t(a) = \mathbb{E}_{\pi \sim q_t} \pi(a|x_t).$$

- Gather relevant data $(x_t, a_t, r_t(a_t), p_t(a_t))$.
- Transform data to $(x, \ell) \in \mathcal{X} \times \mathbb{R}^{\mathcal{A}}$ pairs.
- Find a $\pi^t \in \Pi$ to minimize empirical risk,

$$\frac{1}{t} \sum_{s=1}^{t} \ell_s(\pi^t(x_s)).$$

- Use π^t to update q_t.

An alternative approach: Reduction to classification

Exp4 used

$$\tilde{\ell}_t(a) = \frac{(1 - r_t(a_t))1[a = a_t]}{p_t(a_t)},$$

and maintained exponential weights over Π based on cumulative sums of

$$\tilde{y}_t(\pi) = \mathbb{E}_{a \sim \pi} \tilde{\ell}_t(a).$$

But this required enumeration over Π. Instead, we will

1. Give the strategy access to Π only via empirical risk minimization.
2. Determine the distribution q_t by the set of π^ts.

Reduction to classification

Assume we have access to an efficient empirical risk minimization oracle.

- (Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang, 2011):
  Ellipsoid method to choose q_t.
  Polynomial (in n and k) number of calls to oracle.

- (Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, Robert E. Schapire, 2014):
  Gradient descent approach.
  $O(\sqrt{kn})$ calls to oracle.

We'll look at (Agarwal et al, 2014).

Reduction to classification

Gradient Descent Strategy

for epoch i

$q_i =$ distribution over Π (approximately) minimizing

$$\mathbb{E}_{\pi \sim q} \hat{R}_t(\pi) + k\mu \hat{\mathbb{E}}_x d_{KL}\left(\mathsf{U}, q^{\mu}(\cdot|x)\right)).$$

for t in epoch i

observe x_t, play $a_t \sim p_t$, where $p_t(a) = \mathbb{E}_{\pi \sim q_i} \pi(a|x_t)$,

observe $r_t(a_t)$

Reduction to classification

Here, U is uniform on $\mathcal{A} = \{1, \ldots, k\}$,

$$\mu = \sqrt{\frac{\log\left(|\Pi|/\delta\right)}{kt}},$$

(similar result with VC-dimension)

$$q^{\mu}(a|x) = (1-\mu)\sum_{\pi\in\Pi} q(\pi)1[\pi(x) = a] + \mu U(a)$$

and the *empirical per-trial regret* is defined by

$$\hat{R}_t(\pi) = \hat{L}_t(\pi) - \min_{\pi\in\Pi} \hat{L}_t(\pi),$$

$$\hat{L}_t(\pi) = \hat{\mathbb{E}}_{(x,\tilde{\ell})}\tilde{\ell}(\pi(x)).$$

Reduction to classification

- The criterion for q combines empirical regret (exploitation) with distance to uniform (exploration).
- It can be approximately minimized by a coordinate descent approach: choose the $\pi \in \Pi$ that is best aligned with the negative gradient.
- Finding the descent direction is (roughly) equivalent to choosing

$$\arg\min_{\pi \in \Pi} \left(\hat{R}_t(\pi) - \hat{\mathbb{E}}_x \frac{\mu}{(1-\mu)q(\pi(x)|x) + \mu/k} \right)$$

which is $$\arg\min_{\pi \in \Pi} \hat{\mathbb{E}}_{x,\ell} \ell(\pi(x))$$

where $$\ell_s(a) = \tilde{\ell}_s(a) - \frac{\mu}{(1-\mu)q(\pi(x_s)|x_s) + \mu/k}.$$

Reduction to classification

Theorem: Under the i.i.d. stochastic model, $(x_t, r_t) \sim P$, with probability at least $1-\delta$,

$$R_n = O\left(\sqrt{kn\log\left(\frac{n|\Pi|}{\delta}\right)} + k\log\left(\frac{n|\Pi|}{\delta}\right)\right).$$

(similar result with VC-dimension)

Idea of proof:

Solution q_i to optimization problem has

1. small empirical regret:

$$\mathbb{E}_{\pi\sim q}\hat{R}_t(\pi) \le ck\mu_i,$$

2. low variance: for all $\pi \in \Pi$,

$$\hat{\mathbb{E}}_x \frac{\mu_i}{(1-\mu_i)q(\pi(x)|x)+\mu_i} \le c\left(k\mu_i + \hat{R}_t(\pi)\right).$$

Reduction to classification

Hence, once $t = \Omega(k \log |\Pi|)$, with high probability, for all $\pi \in \Pi$,

$$R(\pi) := \min_{\pi^* \in \Pi} \mathbb{E}_{r,x} \left(r(\pi^*(x) - r(\pi(x)) \right) \leq 2\hat{R}_t(\pi) + O(k\mu_i).$$

So

$$\mathbb{E}_{\pi \sim q_i} R(\pi) = O(k\mu_i).$$

Summing across time gives the regret bound.

Summary: Reduction to classification

- Maintain distribution q_t over Π.
- Observe x_t, choose $a_t \sim p_t = \mathbb{E}_{\pi \sim q_t} \pi(\cdot|x_t)$.
- Transform relevant data $(x_t, a_t, r_t(a_t), p_t(a_t))$ to $(x, \ell) \in \mathcal{X} \times \mathbb{R}^{\mathcal{A}}$ pairs.
- Find a $\pi^t \in \Pi$ to minimize empirical risk,

$$\frac{1}{t} \sum_{s=1}^{t} \ell_s(\pi^t(x_s)).$$

- Use π^t to update q_t:
  coordinate descent of regularized empirical regret.
- Regularization ensures empirical regret bounds regret.