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

Peter Bartlett

1. Linear bandits.
   - Exponential weights with unbiased loss estimates.
   - Controlling loss estimates and their variance.

Linear bandits

At round t,

- Strategy chooses $a_t \in \mathcal{A} \subset \mathbb{R}^d$.
- Adversary chooses loss $\ell_t \in \mathcal{A}^* \subset [-1, 1]^d$.
- Strategy sees loss $\ell_t(a_t)$.

Loss is *linear* in action.

Aim to minimize pseudo-regret:

$$\overline{R}_n = \mathbb{E} \sum_{t=1}^{n} \ell_t(a_t) - \inf_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{n} \ell_t(a).$$

Example: Packet routing

Consider the problem of packet-routing in a network (V, E). At round t,

- Strategy chooses a path $a_t \in \mathcal{A} \subset \{0,1\}^E$ from origin node to destination node.
- Adversary chooses delays $\ell_t \in \mathcal{L} = [0,1]^E$.
- See loss $\ell_t \cdot a_t$ (total delay).

Aim to minimize pseudo-regret:

$$\overline{R}_n = \mathbb{E} \sum_{t=1}^{n} \ell_t \cdot a_t - \inf_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{n} \ell_t \cdot a.$$

Loss is *linear* in action.

Linear bandits vs k-armed bandits

This problem is closely related to the classical k-armed bandit problem: At round t:

- Strategy chooses $a_t \in \mathcal{A} = \{1, \ldots, k\}$.
- Adversary chooses $\ell_t \in \mathcal{L} = [0,1]^{\mathcal{A}}$.
- See loss $\ell_t(a_t)$.

Aim to minimize pseudo-regret:

$$\overline{R}_n = \mathbb{E} \sum_{t=1}^{n} \ell_t(a_t) - \inf_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{n} \ell_t(a).$$

Linear bandits vs k-armed bandits

This is unchanged (up to a constant factor) if we instead define

$$\mathcal{A} = \{e_1, \ldots, e_k\} \subset \mathbb{R}^k,$$
$$\mathcal{L} = \mathcal{A}^* \cap [-1, 1]^{\mathcal{A}},$$

(bounded linear functions on $\mathcal{A}$).

And allowing the strategy to choose a in the convex hull of $\mathcal{A}$ does not change the pseudo-regret

$$\overline{R}_n = \mathbb{E} \sum_{t=1}^{n} \ell_t(a_t) - \inf_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{n} \ell_t(a).$$

(But it might make the game easier for the strategy since it changes the information that the strategy sees.)

Finite covers

For a compact $\mathcal{A} \subseteq \mathbb{R}^d$, we can construct an ϵ-cover of size $O(1/\epsilon^d)$. Since we're aiming for $O(\sqrt{n})$ regret, we can think of $\mathcal{A}$ as having cardinality $|\mathcal{A}| = O(n^{d/2})$, so $\log |\mathcal{A}| = O(d \log n)$.

Exponential weights for linear bandits

Given $\mathcal{A}$, distribution μ on $\mathcal{A}$, mixing coefficient $\gamma > 0$, learning rate $\eta > 0$,

set q_1 uniform on $\mathcal{A}$.

for $t = 1, 2, \ldots, n$,

1. $p_t = (1-\gamma)q_t + \gamma\mu$
2. choose $a_t \sim p_t$
3. observe $\ell_t^T a_t$
4. update $q_{t+1}(a) \propto q_t(a)\exp(-\eta\tilde{\ell}_t^T a)$,

where

$$\tilde{\ell}_t = \left(\mathbb{E}_{a\sim p_t} a a^T\right)^{\dagger} a_t a_t^T \ell_t.$$

Unbiased loss estimates

Strategy observes $a_t^T \ell_t$ and a_t, so it can compute

$$\tilde{\ell}_t = \left(\mathbb{E}_{a \sim p_t} a a^T\right)^\dagger a_t \left(a_t^T \ell_t\right).$$

Also,

$$\mathbb{E}_{a_t \sim p_t} \tilde{\ell}_t = \left(\mathbb{E}_{a \sim p_t} a a^T\right)^\dagger \left(\mathbb{E}_{a_t \sim p_t} a_t a_t^T\right) \ell_t = \ell_t.$$

Regret bound

Theorem: For $\sup_{a\in\mathcal{A}} \left|\tilde{\ell}_t^T a\right| \leq 1$ and $\eta < 1/2$,

$$\overline{R}_n \leq \gamma n + \frac{\log|\mathcal{A}|}{\eta} + (e-2)\eta \sum_{t=1}^{n} \mathbb{E}_{a\sim p_t} \left(\tilde{\ell}_t^T a\right)^2 .$$

So we need to control the magnitude of the loss estimates,

$$\sup_{a\in\mathcal{A}} \left|\tilde{\ell}_t^T a\right|$$

and the variance term,

$$\mathbb{E}_{a\sim p_t} \left(\tilde{\ell}_t^T a\right)^2 .$$

Exponential weights for linear bandits

- (Dani, Hayes, Kakade, 2008):
  For μ uniform over *barycentric spanner*,

$$\overline{R}_n = \tilde{O}\left(\log|\mathcal{A}|\sqrt{dn} + d^{3/2}\sqrt{n}\right) = \tilde{O}\left(d^{3/2}\sqrt{n}\right).$$

- (Cesa-Bianchi and Lugosi, 2009):
  If smallest non-zero eigenvalue of $\mathbb{E}_{a\sim\mu}[aa^T]$ is $\Omega(1/d)$,

$$\overline{R}_n = \tilde{O}\left(\sqrt{dn\log|\mathcal{A}|}\right) = \tilde{O}\left(d\sqrt{n}\right).$$

  And for several interesting $\mathcal{A}$, μ uniform over $\mathcal{A}$ suffices.

- (Bubeck, Cesa-Bianchi and Kakade, 2009):
  Johns Theorem gives a suitable μ.

$$\overline{R}_n = \tilde{O}\left(\sqrt{dn\log|\mathcal{A}|}\right) = \tilde{O}\left(d\sqrt{n}\right).$$