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

Peter Bartlett

1. Adversarial bandits
   - Definition: sequential game.
   - Lower bounds on regret from the stochastic case.
   - Exp3: exponential weights strategy.

Adversarial bandits

Repeated game: strategy chooses I_t, then adversary chooses $(x_{1,t}, \ldots, x_{k,t})$.

Aim to minimize regret,

$$R_n = \max_j \sum_{t=1}^{n} x_{j,t} - \sum_{t=1}^{n} x_{I_t,t},$$

or pseudo-regret

$$\overline{R}_n = \max_j \mathbb{E} \sum_{t=1}^{n} x_{j,t} - \mathbb{E} \sum_{t=1}^{n} x_{I_t,t},$$

Adversarial bandits

Some scenarios:

- I_t deterministic. Hopeless.
- Must have I_t randomized: strategy plays a distribution. Regret is random. Could consider expected regret, or high probability regret bounds.
- Adversary chooses $x_{j,t}$ independent of strategy's previous random outcomes I_t. *Oblivious adversary.*
- Adversary chooses $x_{j,t}$ with knowledge of strategy's previous random outcomes I_t. *Adaptive adversary* or *nonoblivious adversary*. But then what does regret mean?

Adversarial bandits: Lower bounds

It's clear that $\overline{R}_n \leq \mathbb{E} R_n$. So a lower bound on pseudo-regret gives lower bounds on expected regret. Lower bounds in the stochastic setting suffice here (the adversary can certainly choose a sequence randomly).

Theorem: For any strategy and any n, there is an oblivious adversary playing $x_{j,t} \in \{0, 1\}$ for which

$$\overline{R}_n \geq \frac{1}{18} \min\{\sqrt{nk}, n\}.$$

In particular, it suffices for the adversary to play i.i.d. Bernoulli rewards to achieve a pseudo-regret that is this large.

Adversarial bandit strategies

How should a strategy choose the distribution over I_t in the adversarial setting?

- Exploration vs exploitation remains an important issue.
- Exploitation appears to be even more dangerous.
- Let's digress and consider the corresponding full information game. It's standard (and, we'll see, convenient) to pose it in terms of losses, rather than rewards. For $x_{i,t}$ in $[0, 1]$, think of $\ell_{i,t} = 1 - x_{i,t}$, so that $\ell_{i,t}$ is in $[0, 1]$.

An aside: A full information prediction game

Consider the following repeated game: at time t,

1. strategy chooses a distribution p_t over k experts ($I_t \sim p_t$),
2. adversary chooses a loss vector $\ell_t = (\ell_{1,t}, \ldots, \ell_{k,t}) \in [0,1]^k$,
3. **strategy sees** ℓ_t.

The aim is to ensure that, for all choices of the adversary,

$$\overline{R}_n = \sum_{t=1}^{n} \mathbb{E}\ell_{I_t,t} - \min_j \sum_{t=1}^{n} \ell_{j,t}$$

is not too large.

An aside: A full information prediction game

Exponential Weights Strategy (with parameter η)

set $p_{1,1} = \cdots = p_{k,1} = 1/k$.

for $t = 1, 2, \ldots, n$, choose $I_t \sim p_t$, observe ℓ_t.

$$L_{i,t} = \sum_{s=1}^{t} \ell_{i,s},$$

$$p_{i,t+1} = \frac{\exp\left(-\eta L_{i,t}\right)}{\sum_{j=1}^{k} \exp\left(-\eta L_{j,t}\right)}.$$

An aside: A full information prediction game

Theorem: The exponential weights strategy with parameter η incurs regret

$$\overline{R}_n \leq \frac{n\eta}{8} + \frac{\log k}{\eta}.$$

Choosing $\eta = \sqrt{8 \log k / n}$ gives $\overline{R}_n \leq \sqrt{n \log k / 2}$.

An aside: A full information prediction game

Proof idea: For this choice of p_t,

$$\Phi_t = -\frac{1}{\eta} \log \left(\sum_{i=1}^{k} \exp(-\eta L_{i,t}) \right)$$

is a *measure of progress*. When $\mathbb{E}\ell_{I_t,t}$ is big, there is a big increase from Φ_{t-1} to Φ_t. But $\Phi_n \leq \min_j L_{j,n}$.

An aside: A full information prediction game

For $\ell_{i,t} \in [0,1]$, Hoeffding's inequality shows that

$$\log \mathbb{E} \exp\left(-\eta\left(\ell_{I_t,t} - \mathbb{E}\ell_{I_t,t}\right)\right) \le \frac{\eta^2}{8},$$

that is,

$$\mathbb{E}\ell_{I_t,t} \le \frac{\eta}{8} - \frac{1}{\eta} \log \mathbb{E} \exp\left(-\eta \ell_{I_t,t}\right).$$

But the choice of p_t means that the sum of these c.g.f.s telescopes:

$$\begin{aligned}\sum_{t=1}^{n} \log \mathbb{E} \exp\left(-\eta \ell_{I_t,t}\right) &= \sum_{t=1}^{n} \log\left(\frac{\sum_{i=1}^{k} \exp\left(-\eta L_{i,t}\right)}{\sum_{i=1}^{k} \exp\left(-\eta L_{i,t-1}\right)}\right) \\ &= \log\left(\sum_{i=1}^{k} \exp\left(-\eta L_{i,n}\right)\right) - \log k.\end{aligned}$$

An aside: A full information prediction game

Thus,

$$\sum_{t=1}^{n} \mathbb{E}\ell_{I_t,t} \leq \frac{\eta n}{8} + \frac{\log k}{\eta} - \frac{1}{\eta} \log \underbrace{\left(\sum_{i=1}^{k} \exp\left(-\eta L_{i,n}\right) \right)}_{\geq \exp(-\eta L_{j,n})}$$

$$\leq \frac{\eta n}{8} + \frac{\log k}{\eta} + \min_j L_{j,n}.$$

$$\overline{R}_n = \sum_{t=1}^{n} \mathbb{E}\ell_{I_t,t} - \min_j \sum_{t=1}^{n} \ell_{j,t} \leq \frac{\eta n}{8} + \frac{\log k}{\eta}.$$

Back to bandits

What can the strategy do when it only sees $\ell_{I_t,t}$?

Importance sampling: in the strategy, replace $\ell_{i,t}$ by

$$\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_{i,t}} 1[I_t = i].$$

This is an unbiased estimate of the unseen losses:

$$\ell_{j,t} = \mathbb{E}\tilde{l}_{j,t}.$$

Exp3: Exponential weights

Strategy Exp3

set p_1 uniform on $\{1, \ldots, k\}$.

for $t = 1, 2, \ldots, n$, choose $I_t \sim p_t$, observe $\ell_{I_t,t}$.

$$\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_{i,t}} 1[I_t = i],$$

$$\tilde{L}_{i,t} = \sum_{s=1}^{t} \tilde{\ell}_{i,s},$$

$$p_{i,t+1} = \frac{\exp\left(-\eta \tilde{L}_{i,t}\right)}{\sum_{j=1}^{k} \exp\left(-\eta \tilde{L}_{j,t}\right)}.$$

Back to bandits

Then the regret involves $\sum_t (\ell_{I_t,t} - \ell_{j,t})$, and

$$\ell_{I_t,t} = \sum_{i=1}^{k} p_{i,t}\tilde{\ell}_{i,t} \qquad\qquad \ell_{j,t} = \mathbb{E}\tilde{l}_{j,t}.$$

But we can no longer appeal to Hoeffding's inequality, because $\tilde{l}_{i,t}$ is unbounded. Happily, we only need an upper bound on the c.g.f. $\Gamma(\lambda)$ for negative values of $\lambda = -\eta$. This corresponds to a lower tail concentration inequality for the non-negative random variable $\tilde{\ell}_{I_t,t}$. But non-negative random variables with finite variance have sub-Gaussian lower tails:

Sub-Gaussian lower tails for non-negative r.v.s

Lemma: For $X \geq 0$ and $\lambda > 0$,

$$\log \mathbb{E} e^{-\lambda X} \leq \frac{\lambda^2}{2} \mathbb{E} X^2 - \lambda \mathbb{E} X,$$

hence

$$\mathbb{E} X \leq \frac{1}{\lambda} \log \mathbb{E} e^{-\lambda X} + \frac{\lambda}{2} \mathbb{E} X^2.$$

Proof.

$$\begin{aligned} \log \mathbb{E} \exp\left(-\lambda(X - \mathbb{E}X)\right) &= \log \mathbb{E} \exp\left(-\lambda X\right) + \lambda \mathbb{E} X \\ &\leq \mathbb{E} \exp\left(-\lambda X\right) - 1 + \lambda \mathbb{E} X \\ &\leq \mathbb{E} \frac{\lambda^2 X^2}{2}, \end{aligned}$$

because $\log y \leq y - 1$ for all y and $e^{-x} \leq 1 - x + x^2/2$ for $x \geq 0$. $\square$

Exp3: Exponential weights

Theorem: Exp3 with parameter η incurs regret

$$\overline{R}_n \leq \frac{n\eta k}{2} + \frac{\log k}{\eta}.$$

Choosing $\eta = \sqrt{2\log k/(nk)}$ gives $\overline{R}_n \leq \sqrt{2nk\log k}$.

(Recall the lower bound $\overline{R}_n = \Omega(\sqrt{nk})$. This strategy matches it to within a $\log k$ factor.)

Exp3: Exponential weights

$$
\begin{aligned}
\sum_{t=1}^{n} \ell_{I_t,t} &= \sum_{t=1}^{n} \sum_{i=1}^{k} p_{i,t} \tilde{\ell}_{i,t} \\
&\leq \sum_{t=1}^{n} \left(\frac{\lambda}{2} \sum_{i=1}^{k} p_{i,t} \tilde{\ell}_{i,t}^2 - \frac{1}{\lambda} \log \left(\sum_{i=1}^{k} p_{i,t} \exp(-\lambda \tilde{\ell}_{i,t}) \right) \right).
\end{aligned}
$$

For the first term, we can bound the variance by k:

$$
\mathbb{E} \sum_{i=1}^{k} p_{i,t} \tilde{\ell}_{i,t}^2 = \mathbb{E} \frac{\ell_{I_t,t}^2}{p_{I_t,t}} \leq \mathbb{E} \frac{1}{p_{I_t,t}} = k.
$$

For the second, as in the full information case, for $\lambda = \eta$ and any j,

$$
-\frac{1}{\eta} \sum_{t=1}^{n} \log \left(\sum_{i=1}^{k} p_{i,t} \exp(-\lambda \tilde{\ell}_{i,t}) \right) \leq \frac{\log k}{\eta} + \mathbb{E} \tilde{L}_{j,n}.
$$

Exp3: Exponential weights

Hence,

$$\begin{aligned}\overline{R}_n &= \mathbb{E}\sum_{t=1}^{n}\ell_{I_t,t} - \min_j \mathbb{E}\sum_{t=1}^{n}\ell_{j,t} \\ &= \mathbb{E}\sum_{t=1}^{n}\ell_{I_t,t} - \min_j \mathbb{E}\tilde{L}_{j,n} \\ &\leq \frac{\eta n k}{2} + \frac{\log k}{\eta}.\end{aligned}$$

Exp3: Exponential weights

- Auer, Cesa-Bianchi, Freund and Schapire introduced Exp3 (but with a uniform distribution mixed in, to keep $1/p_{I_t,t}$ bounded).
- Stoltz showed that the uniform distribution is not necessary, through the sub-Gaussian lower tails idea. (See the 'Readings' page on the website.)
- Notice that the η parameter is set using the time horizon n. There are two approaches to avoiding this:
  - Run Exp3 in multiple epochs, doubling n in each epoch, and then the regret is no more than the sum of the regrets, which grows like the square root of the total time horizon.
  - Set η_t on a decreasing schedule: $\eta_t = \sqrt{\log k/(tk)}$. Since η_t is decreasing, the telescoping sum becomes a sum that is no more than 0, and the $kn\eta/2$ becomes $k\sum_t \eta_t/2 = O(\sqrt{nk\log k})$.

Exp3: Exponential weights

- High probability? Variances of the losses are big ($\mathbb{E}\tilde{\ell}_{j,t}^2 = \tilde{\ell}_{j,t}^2/p_{j,t}$). Mixing the uniform distribution does not help (for a small regret, the uniform component must be as small as $n^{-1/2}$). Need a new strategy. Instead, use a biased estimate in place of $\tilde{\ell}_{j,t}$: subtract a small offset so $\tilde{L}_{j,t}$ becomes a lower confidence bound on the cumulative expected loss $L_{j,t}$ (or, in the case of gains, an upper confidence bound on the cumlative expected reward).
- Log factor? Can eliminate the $\log k$ factor. One approach is to replace exponential weights with a generalization, *mirror descent*, with an appropriate potential function.
- Other comparisons? Can compete with sequences of actions, with a bounded number of switches.