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

Peter Bartlett

1. Lower bounds on regret for multi-armed bandits.

Stochastic bandit problem: notation.

- k arms.
- Arm j has unknown reward distribution P_{θ_j}, for $\theta_j \in \Theta$.
- Reward: $X_{j,t} \sim P_{\theta_j}$.
- Mean reward: $\mu_j = \mathbb{E} X_{j,1}$.
- Best: $\mu^* = \max_{j^*=1,\ldots,k} \mu_{j^*}$.
- Gap: $\Delta_j = \mu^* - \mu_j$.
- Number of plays: $T_j(s) = \sum_{t=1}^{s} 1[I_t = j]$.

Lower bounds on regret.

Because

$$\overline{R}_n = n \max_{j^*=1,\ldots,k} \mathbb{E}\mu_{j^*} - \mathbb{E}\sum_{t=1}^{n} X_{I_t,t} = \sum_{j=1}^{k} \mathbb{E}T_j(n)\Delta_j,$$

we need to understand how $\mathbb{E}T_j(n)$ behaves for $j \neq j^*$.

We'll see that (asymptotically)

$$\mathbb{E}T_j(n) \gtrsim \frac{\log n}{D_{KL}(P_{\theta_j}, P_{\theta^*})}.$$

Here, when $P \ll Q$,

$$D_{KL}(P, Q) = \int \log \frac{dP}{dQ} dP.$$

Lower bounds on regret.

Key insight: Consider two bandit problems:

$$\theta = (\theta_1, \theta_2, \ldots, \theta_k),$$

$$\theta = (\theta_1, \theta_2', \ldots, \theta_k),$$

with

$$\mu_1 > \mu_2 \geq \mu_3 \geq \cdots \geq \mu_k,$$

$$\mu_2' \gtrsim \mu_1 > \mu_3 \geq \cdots \geq \mu_k.$$

If a strategy performs well for θ, and P_{θ_2} and $P_{\theta_{2'}}$ are close, then the same data is likely under both, so it must perform poorly for θ'.

The lower bound will require the strategy to perform well for all θ (c.f. a stopped clock).

(And the right way of measuring "close" is via a change of measure between P_{θ_2} and $P_{\theta_2'} \approx P_{\theta_1}$, which leads to the KL-divergence.)

Lower bounds on regret.

[Radon-Nikodym derivative] For any event A,

$$P_{\theta'}(A) = \int_A \frac{dP_{\theta'}}{dP_\theta}\, dP_\theta.$$

Need to have $P_{\theta'} \ll P_\theta$.
(i.e., $P_{\theta'}$ is absolutely continuous wrt P_θ,
i.e., if $P_\theta(E) = 0$ then $P_{\theta'}(E) = 0$.)

Lower bounds on regret.

Fix a strategy, and write:

$X_{j,s}$ = outcome from pull s of arm j,

$\mathbb{P}$ = joint distribution over $\{I_t, X_{j,s}\}$ under distribution P_θ,

$\mathbb{P}'$ = joint distribution under distribution $P_{\theta'}$.

Lower bounds on regret.

For an event $A \subseteq \{T_2(n) = n_2\}$, we can write

$$\begin{aligned}
\mathbb{P}'(A) &= \int_A \prod_{s=1}^{n_2} \frac{dP_{\theta_2'}}{dP_{\theta_2}}(X_{2,s})\, d\mathbb{P} \\
&= \int_A \exp\left(\sum_{s=1}^{n_2} \log \frac{dP_{\theta_2'}}{dP_{\theta_2}}(X_{2,s}) \right)\, d\mathbb{P} \\
&= \int_A e^{-L_{n_2}}\, d\mathbb{P},
\end{aligned}$$

where

$$L_{n_2} = \sum_{s=1}^{n_2} \log \frac{dP_{\theta_2}}{dP_{\theta_2'}}(X_{2,s}).$$

So if $A \subseteq \{T_2(n) = n_2 \text{ and } L_{n_2} \leq c_n\}$, (data from θ could plausibly have come from θ')
then $\mathbb{P}'(A) \geq e^{-c_n}\mathbb{P}(A)$, that is, $\mathbb{P}(A) \leq e^{c_n}\mathbb{P}'(A)$.

Lower bounds on regret.

Fix sequences f_n and c_n (we'll pick them later).

$\mathbb{P}\left(T_2(n) < f_n\right)$ (suboptimal arm not chosen too often)

$$\leq \mathbb{P}\left(T_2(n) < f_n \;\&\; L_{T_2(n)} \leq c_n\right) + \mathbb{P}\left(T_2(n) < f_n \;\&\; L_{T_2(n)} > c_n\right)$$

$$\leq e^{c_n} \mathbb{P}'\left(T_2(n) < f_n \;\&\; L_{T_2(n)} \leq c_n\right) + \mathbb{P}\left(T_2(n) < f_n \;\&\; L_{T_2(n)} > c_n\right)$$

$$\leq e^{c_n} \underbrace{\mathbb{P}'\left(T_2(n) < f_n\right)}_{\text{(optimal arm not chosen often)}} + \underbrace{\mathbb{P}\left(T_2(n) < f_n \;\&\; L_{T_2(n)} > c_n\right)}_{\text{(and data from } \theta \text{ unlikely to have come from } \theta'\text{)}}.$$

Lower bounds on regret.

Under $\mathbb{P}'$, arm 2 is optimal, so the first probability,

$$\mathbb{P}'\left(T_2(n) < f_n\right),$$

is the probability that the optimal arm is not chosen too often. This should be small whenever the strategy does a good job (and f_n quantifies what a good job means). We'll ensure $f_n = o(n)$. Then if we assume that, for any $\alpha > 0$, the expected number of pulls that the strategy wastes on sub-optimal arms is $o(n^\alpha)$, that is,

$$\mathbb{E}'\left(n - T_2(n)\right) = o(n^\alpha),$$

Markov's inequality shows that

$$\mathbb{P}'\left(T_2(n) < f_n\right) \leq \frac{\mathbb{E}'(n - T_2(n))}{n - f_n} = o(n^{\alpha-1}).$$

Lower bounds on regret.

The second term is

$$\mathbb{P}\left(T_2(n) < f_n \;\&\; \sum_{s=1}^{T_2(n)} \log \frac{dP_{\theta_2}}{dP_{\theta_2'}}(X_{2,s}) > c_n\right).$$

But notice that the expectation (under $\mathbb{P}$) of each $\log \frac{dP_{\theta_2}}{dP_{\theta_2'}}(X_{2,s})$ term is $D_{KL}(P_{\theta_2}, P_{\theta_2'})$, the KL-divergence of $P_{\theta_2'}$ from P_{θ_2}.

If c_n is a little bigger than $f_n D_{KL}(P_{\theta_2}, P_{\theta_2'})$, the law of large numbers will ensure that this term will go to zero.

Lower bounds on regret.

Choosing (for a suitable $\delta > 0$)

$$f_n = (1-\delta)\frac{\log n}{D_{KL}(P_{\theta_2}, P_{\theta_2'})}$$

ensures $\mathbb{P}\left(T_2(n) < f_n\right) = o(1)$. Hence choosing $P_{\theta_2'}$ suitably close to P_{θ_1} gives

$$\lim_{n\to\infty} \inf \frac{\mathbb{E}T_2(n)}{\log n} \geq \frac{1}{D_{KL}(P_{\theta_2}, P_{\theta^*})}.$$

Lower bounds on regret.

Theorem: [Lai-Robbins, 1985] Suppose P_θ and Θ are such that:

1. Whenever $\mu(\theta_1) > \mu(\theta_2)$, $0 < D_{KL}(P_{\theta_2}, P_{\theta_1}) < \infty$, and
2. (denseness condition on $\mu(\Theta)$)
3. (continuity condition on $\theta_1 \mapsto D_{KL}(\theta_2, \theta_1)$)

If a strategy has, for all $\theta = (\theta_1, \ldots, \theta_k)$ and all $\alpha > 0$, $\overline{R}_n(\theta) = o(n^\alpha)$, then

$$\lim_{n\to\infty} \inf \frac{\overline{R}_n(\theta)}{\log n} \geq \sum_{j:\mu_j<\mu^*} \frac{\mu^* - \mu_j}{D_{KL}(P_{\theta_j}, P_{\theta^*})}.$$

Lower bounds on regret.

Example: Bernoulli. Parameter is μ.

$$D_{KL}(p, q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

The lower bound implies

$$\lim_{n\to\infty} \inf \frac{\overline{R}_n(\theta)}{\log n} \geq \mu^*(1-\mu^*) \sum_{j:\mu_j<\mu^*} \frac{1}{\mu^* - \mu_j}.$$

Lower bounds on regret.

To see this, use the upper bound $\log(x) \leq x - 1$ to give

$$\begin{aligned} D_{KL}(p,q) &= p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \\ &\leq p \frac{p-q}{q} + (1-p) \frac{q-p}{1-q} \\ &= \frac{(p(1-q) - (1-p)q)(p-q)}{q(1-q)} \\ &= \frac{(p-q)^2}{q(1-q)}. \end{aligned}$$

Then the lower bound becomes

$$\sum_{j:\mu_j<\mu^*} \frac{\mu^* - \mu_j}{D_{KL}(P_{\theta_j}, P_{\theta^*})} \geq \mu^*(1-\mu^*) \sum_{j:\mu_j<\mu^*} \frac{1}{\mu^* - \mu_j}.$$

Lower bounds on regret.

Also, this form of the inequality for Bernoulli distributions does not lose much:

Theorem: [Pinsker's inequality]

$$D_{KL}(P,Q) \geq 2d_{TV}(P,Q)^2,$$

where the total variation distance is defined as

$$d_{TV}(P,Q) = \sup\{|P(A) - Q(A)| : \; A \text{ measurable}\}.$$

For Bernoulli distributions, $d_{TV}(p,q) = |p-q|$, so

$$D_{KL}(P_{\theta_j}, P_{\theta^*}) \geq 2(\mu^* - \mu_j)^2.$$

Lower bounds on regret.

An aside:

To prove Pinsker's inequality for Bernoulli, it suffices to calculate the partial derivative of $D_{KL}(p,q) - 2(p-q)^2$ wrt q. Actually, this leads to the proof of Pinsker's inequality for any distribution: $d_{TV}(P,Q) = P(A) - Q(A) = d_{TV}(P_{\mathcal{A}}, Q_{\mathcal{A}})$ for

$$A = \left\{ \frac{dP}{d(P+Q)} > \frac{dQ}{d(P+Q)} \right\}$$

and $P_{\mathcal{A}}$ and $Q_{\mathcal{A}}$ are the induced (Bernoulli) distributions on the elements of the partition $\mathcal{A} = \{A, \bar{A}\}$. But the *partition inequality for KL-divergence* shows that, for any partition, $D_{KL}(P,Q) \geq D_{KL}(P_{\mathcal{A}}, Q_{\mathcal{A}})$.