Minimax Lower Bound – Sketch.

Source: Chap. 13, Bandit Algorithms, L & S.

Running example: 2 armed bandit, unstructured environment, Gaussian (1) reward,

i.e.

arm 1 • $\{X_{t1}, t=1,2,\dots,n\}$ $X_{t1} \sim N(\Delta, 1)$, $\mu_1 = \Delta$

arm 2 • $\{X_{t2}, t=1,2,\dots,n\}$ $X_{t2} \sim N(0, 1)$, $\mu_2 = 0$

Fact: with m samples, $X_{t2} \sim N(0,1)$

$$\frac{\sqrt{8/\pi}}{\sqrt{m\Delta^2} + \sqrt{m\Delta^2 + 16}} e^{-m\Delta^2/8} \leq P\left(\frac{1}{m}\sum_{t=1}^{m} X_{t2} > \frac{\Delta}{2}\right)$$

$$\leq \frac{\sqrt{8/\pi}}{\sqrt{m\Delta^2} + \sqrt{m\Delta^2 + 32/\pi}} e^{-m\Delta^2/8}$$

Suppose $\Delta = \frac{1}{\sqrt{m}}$.

Then, $P\left(\frac{1}{m}\sum_{r=1}^{m} x_{r2} > \frac{\Delta}{2}\right) \approx e^{-1/8} > 0$

i.e., with a non-vanishing probability, we cannot resolve if samples from arm 2 might have been generated by arm 1.

Similarly,

$$P\left(\frac{1}{m}\sum_{r=1}^{m} x_{r1} < \frac{\Delta}{2}\right) \approx e^{-1/8} > 0$$

and similar reasoning as above.

This means that with constant prob (of $e^{-1/8}$), it is plausible that we play the wrong arm for the entire duration of m time steps

i.e., Regret $\approx e^{-1/8} \cdot \Delta \cdot m = e^{-1/8} \cdot \frac{1}{\sqrt{m}} \cdot \sqrt{m}$

$\approx O(\sqrt{m})$.

The idea for constructing bandit lower bounds:

Construct two alternative models of the bandit system that are "difficult" to distinguish between, while satisfying two criteria:

(a) The best arm in one system is a sub-optimal arm in the other system.

(b) The systems are statistically as "close" as possible (while satisfying (a) above), so that given the samples from arm plays, it is hard to distinguish between them.

---

time horizon

Back to example: choose small $\Delta = \frac{1}{\sqrt{n}}$

System 1 = ν		ν' = System 2	
arm 1	arm 2	arm 1	arm 2
$N(\Delta, 1)$	$N(0, 1)$	$N(\Delta, 1)$	$N(2\Delta, 1)$
best arm			best arm

Fix any policy π, i.e., π is a mapping from past samples and actions to current play decision.

* For either system, π is the same.

$\Rightarrow$ If the player sees the same (action, reward) sequence in either system, the player makes the same action decision at the next play.

Sketch below to show that:

$$\max\left\{ R_n(\pi, \nu) \,,\, R_n(\pi, \nu') \right\} \geq \frac{1}{2}\sqrt{n}$$

i.e., with <u>any</u> fixed policy π, the player suffers a regret of $O(\sqrt{n})$ in one of these two environments.

$\Rightarrow$ **minimax lower bound**

Fix policy π, and let $T_1(n)$, $T_2(n)$ be the number of arm 1 and arm 2 plays.

Consider the environment ν. (system 1)

$E[\cdot] = E_{P_\nu}[\cdot]$

$(\Delta, 0)$

Then, $R_n(\pi, \nu) = (n - E[T_1(n)]) \cdot \Delta$ $\rightarrow \left(\frac{1}{\sqrt{n}}\right)$

$E[T_2(n)]$

Now, in environment ν' (system 2)

$(\Delta, 2\Delta)$

$E'[\cdot] = E_{P_{\nu'}}[\cdot]$

$$R_n(\pi, \nu') = \Delta E'[T_1(n)]$$

Suppose: $E[T_1(n)] \simeq E'[T_1(n)]$, i.e.,

This is key, and so far unjustified in any formal sense .... but seems reasonable, based on Gaussian intuition earlier.

the policy π is unable to distinguish between ν and ν', and thus plays arm 1 the same number of times under either scenario.

Then $R_n(\pi, \nu) \approx \left(\sqrt{n} - \frac{a}{\sqrt{n}}\right)$ $\Delta = \frac{1}{\sqrt{n}}$

$$R_n(\pi, \nu') = \frac{a}{\sqrt{n}}$$

$a = E'[T_1(n)] \approx E[T_1(n)]$,

$1 \leq a \leq n$

$$\Rightarrow \max\left\{R_n(\pi, \nu), R_n(\pi, \nu')\right\} \geq \frac{\sqrt{n}}{2}$$

$a = n/2$

Note: This is not a sketch of proof because we have assumed $E[T_1(n)] \approx E'[T_1(n)]$ under any policy π, and scenarios:
$\nu = (\Delta, 0)$, $\nu' = (\Delta, 2\Delta)$, $\Delta = \frac{1}{\sqrt{n}}$.

In subsequent lectures:

k-armed, unstructured environment $\nu \in \mathcal{E}$, stochastic setting.

(a) Minimax bound — Showing existence of environment $\nu' \in \mathcal{E}$ s.t. $R_n(\pi, \nu') \geq \Theta(\sqrt{nk})$.

(b) Instance dependent bound – For any environment v, showing $R_n(\pi, v) \geq c^* \cdot \ln(n)$, for reasonable π.

Note that c^* depends on $\{\Delta_i\}$, the arm gaps. Specifically, we will see that roughly,

$$c^* \approx \sum_{\{i : \Delta_i > 0\}} \frac{\Delta_i}{d(\mu_i, \mu_1)}$$

(c) High probability bound – Bounds for both stochastic and adversarial environments, that hold with high prob., i.e., results of the form

stochastic: $P\left(R_n(\pi, v) > c \sqrt{Kn \ln(1/\delta)} \right) \geq \delta$

adversarial: $\exists$ rewards $\{x_1, x_2, \cdots, x_n\}$ s.t.

$$P\left(R_n(\pi, x) > c \sqrt{Kn \ln(1/\delta)} \right) \geq \delta$$

for $\delta \in (0, 1)$.