Instance Dependent Lower Bound

Text (Bandit Algorithms, Chap 16) approach:

- Asymptotic (i.e., time horizon $\to \infty$) lower bound for any "reasonable" policy (formally consistent policy, see Defn 16.1) for unstructured stochastic bandits.

$$d_{inf}(P, \mu, \mathcal{M}) = \inf_{P' \in \mathcal{M}} \{ D(P, P') : \mu(P') > \mu \}.$$

($\mu(P')$: mean when arm dist is P')

Then, $$\lim_{n \to \infty} \frac{R_n}{\ln(n)} \geq \sum_{j : \Delta_j > 0} \frac{\Delta_j}{d_{inf}(P_j, \mu_1, \mathcal{M}_j)}$$

(P_j: j^{th} arm dist.; μ_1: best arm mean; $\mathcal{M}_j$: possible set of j^{th} arm dist.)

Roughly, $$\frac{R_n}{\ln(n)} \gtrsim \sum_{j : \Delta_j > 0} \left(\frac{\Delta_j}{d(P_j, P_1)} \right)$$

(asymp.)

- For finite horizon, regret bounds for Gaussian environment with additional conditions on reasonableness of policy.

In this class, I will focus directly on finite horizon bounds. Setting of 2-armed Bernoulli bandit with means μ_1, μ_2, with $\mu_1 - \mu_2 = \Delta > 0$.

Sources: Thanks to Rajat Sen and Subhashini Krishnasamy (Ph.D. grads from UT) for discussions and notes based on:

(1) S. Bubeck, V. Perchet and P. Rigollet, "Bounded regret in stochastic multi-armed bandits", arXiv: 1302.1611

(2) R. Combes, C. Jiang and R. Srikant, "Bandits with budgets: Regret lower bounds and optimal algorithms", ACM Sigmetrics 2015.

Setting: k armed stochastic Bernoulli bandit with unstructured environment $\nu \in \mathcal{E}$,

$$\mathcal{E} = \{ \nu : \nu = (P_1, P_2 \dots, P_k),\ P_j \sim \text{Bernoulli}(\mu_j) \}$$

($\mu_j \in [0,1]$)

First, we restrict to "reasonable" policies:

α-consistent Policy: A policy is α-consistent, with $\alpha \in (0,1)$, if for any $\nu \in \mathcal{E}$, $k \neq k^*$, where $k^* = \text{argmax} \{\mu_1, \mu_2, \dots \mu_k\}$, we have

$$E\left[\sum_{s=1}^{t} \mathbb{1}_{\{A_s = k\}}\right] \leq C t^{\alpha}, \text{ for some } C > 0,$$

($T_k(t)$ over the sum; E w.r.t. (ν, π))

and $t \geq \tau_2$.

Discussion: This condition states that the policies we consider are "reasonable" in the sense that the policy discovers sub-optimal arms quickly enough such that it plays these bad arms only infrequently (sub linear with respect to time).

e.g.: Policy: Play arm 1 always. This is a bad policy because for $\nu = (\text{Bernoulli}(0.1), \text{Bern}(0.3))$, this will incur linear regret. The α-consistency condition disallows such a policy.

* All policies we have considered so far for stochastic bandits are α-consistent.

* Suppose a policy π is NOT α-consistent. Then $\exists\ \nu \in \mathcal{E}$ s.t. $R_n(\pi, \nu)$ incurs $\geq \Omega(t^{\alpha})$ regret. Henceforth, we restrict to α-consistent policies.

Proof below for a two armed bandit, with arms μ_1, μ_2 with $1 > \mu_1 > \mu_2 > 0$.

Theorem: Suppose $\nu \doteq (P_1, P_2)$, with $P_i \sim \text{Bernoulli}(\mu_i)$, with $1 > \mu_1 > \mu_2 > 0$, and π is any α-consistent policy, then for any $\delta \in (0, 1-\mu_1)$, $n \geq \tau_2$,

$$R_n(\pi, \nu) = \Delta_2 \, E[T_2(n)]$$

$$\geq \frac{\Delta}{d(\mu_2, \mu_1+\delta)} \left((1-\alpha) \ln(n) - \ln(8C) \right)$$

(n: time horizon.)

KL divergence for Bernoulli r.vs

from α-consistent policy defn.

Proof:

System 1: ν | System 2: ν'

	arm 1	arm 2
System 1: ν	μ_1	μ_2
System 2: ν'	μ_1	$\mu_1+\delta$

$$E_{\nu,\pi}[T_2(n)] \leq C n^\alpha \quad \forall\, n \geq \tau_2$$

$$E_{\nu',\pi}[T_1(n)] \leq C n^\alpha \quad \forall\, n \geq \tau_2.$$

Above bounds follows from α-consistency of policy π.

Let P = prob. measure induced by (ν, π) over time horizon n. (i.e; dist. over $A_1, X_1, \cdots, A_n, X_n$)

$$P \equiv P_{\nu,\pi}\left(A_1 = a_1, X_1 = x_1, A_2 = a_2, \cdots, A_n = a_n, X_n = x_n\right)$$

Q = prob. measure induced by (ν', π) over time horizon n.

$$A = \left\{ T_2(n) > \frac{n}{2} \right\}.$$

(A is a bad event under P; A^c is bad under Q)

B-H inequality: $P(A) + Q(A^c) \geq \frac{e^{-D(P,Q)}}{2}$

Now,

$$P(A) = P\left(T_2(n) > \frac{n}{2}\right) \overset{\text{Markov's inequality}}{\leq} \frac{E_{\nu,\pi}[T_2(n)]}{\left(\frac{n}{2}\right)}$$

$$\leq \frac{2Cn^{\alpha}}{n} = \frac{2C}{n^{1-\alpha}}$$

Similarly,

$$Q(A^c) = Q\left(T_2(n) \leq \frac{n}{2}\right) = Q\left(T_1(n) > \frac{n}{2}\right)$$

$$\leq 2C / n^{1-\alpha}.$$

$$\therefore \quad \frac{4C}{n^{1-\alpha}} \geq \frac{e^{-D(P,Q)}}{2}$$

$$\Rightarrow \boxed{D(P,Q) \geq (1-\alpha)\ln(n) - \ln(8C) \quad \forall \; n \geq \tau_2.}$$

$\hookrightarrow$ (*)

Claim 1: $D(P,Q) = d(\mu_2, \mu_1+\delta)\, E_{\nu,\pi}[T_2(n)]$.

$\hookrightarrow$ Divergence Decomposition (See Set 10).

Then, (*) + Claim 1 $\Rightarrow$ $\forall \; n \geq \tau_2$,

$$E_{\nu,\pi}[T_2(t)] \geq \frac{1}{d(\mu_2, \mu_1+\delta)} \left((1-\alpha)\ln(n) - \ln(8C) \right)$$

$$\therefore \; R_n(\pi, \nu) = \Delta_2 \, E_{\nu,\pi}[T_2(t)]$$

$$\geq \frac{\Delta_2}{d(\mu_2, \mu_1+\delta)} \left((1-\alpha)\ln(n) - \ln(8C) \right).$$
