Stochastic Contextual Bandits

Source: Bandit Algorithms, L & S, Chap 19.

Setting:

At time $t = 1, 2, \ldots, n$:

1. Player observes context $C_t \in \mathcal{C}$.
2. Player plays action $A_t \in \mathcal{A}_t$
3. Nature chooses reward $X_t = r(C_t, A_t) + \eta_t$

$r : \mathcal{C} \times \mathcal{A}_t \longmapsto \mathbb{R}$, a real valued reward function.

$\eta_t \sim$ 1-subGaussian noise.

$$I_t = (C_1, A_1, X_1, C_2, A_2, X_2, \ldots, C_t, A_t)$$

$$\mathbb{E}\left[e^{\lambda \eta_t} \mid I_t\right] \leq e^{\lambda^2/2} \quad \text{a.s.} \quad \text{(i.e., conditionally 1-subG noise).}$$

3(a): r: a known function.

4. $$R_n(\pi, \nu) = \mathbb{E}\left[\sum_{t=1}^{n} \max_{a \in \mathcal{A}_t} r(C_t, a) - \sum_{t=1}^{n} X_t\right]$$

(Note:
Back to stochastic setting: At each time, a different action can potentially be chosen).

$r(\cdot,\cdot)$ through a Feature Map:

$$r(c,a) = \langle \theta^*, \psi(c,a) \rangle, \quad \theta^* \in \mathbb{R}^d,$$

$\psi(c,a)$: feature map $\in \mathbb{R}^d$

Example: c = visitor to a website; one-hot encoding.
a = ad to place; one-hot encoding.

Leads to **Linear Bandit**; directly work in feature space.

Model: $X_t = \langle \theta^*, A_t \rangle + \eta_t$

θ^*: unknown vector $\in \mathbb{R}^d$.

A_t: action by player $\in \mathcal{A}_t$

$\mathbb{E}[\eta_t | I_t] \leq e^{\lambda^2/2}$, where

$$I_t = \{ A_1, A_1, X_1, \cdots, A_{t-1}, A_{t-1}, X_{t-1}, A_t, A_t \}$$

$$R_n = E\left[\sum_{t=1}^{n} \max_{a \in A_t} \langle \theta^*, a \rangle - \sum_{t=1}^{n} X_t \right]$$

(a) $A_t = \{ e_1, e_2, \cdots, e_d \}$ unit vectors in $\mathbb{R}^d$.

Then,

$$\langle \theta^*, e_i \rangle = \theta_i^* \quad , \text{ i.e.,}$$

$$X_t = \theta_i^* + \eta_t \longrightarrow 1\text{-subG.}$$

$\Rightarrow$ Standard unstructured bandit environment

(b) $A_t = \{ \psi(C_t, i), \; i = 1, 2, \cdots, k \}$

$\Rightarrow$ contextual linear bandit.

Linear Bandit: (see Sec 19.2 in textbook)

$X_t = \langle \theta^*, A_t \rangle + \eta_t$: Reward at time t

$A_t \in \mathcal{A}_t$, $A_t \in \mathbb{R}^d$

θ^* unknown, $\theta^* \in \mathbb{R}^d$.

For this class: (constraints on action space).

(a) $|\langle \theta^*, a \rangle| \leq 1$, $\forall\ a \in \bigcup_{t=1}^{n} \mathcal{A}_t$

(b) $\|a\|_2 \leq L < \infty$ $\forall\ a \in \bigcup_{t=1}^{n} \mathcal{A}_t$

Information at end of time $(t-1)$:

$$H_{t-1} = \{\mathcal{A}_1, A_1, X_1, \mathcal{A}_2, A_2, X_2, \cdots, \mathcal{A}_{t-1}, A_{t-1}, X_{t-1}\}$$

Algorithm: Lin UCB (Linear UCB).

Note: There are several similar variants, including

* LinRel (Linear Reinforcement Learning)
* OFUL (Optimism in the Face of Uncertainty for Linear bandits)

Idea: Given H_{t-1}, estimate a set $C_t \subseteq \mathbb{R}^d$ such that

(1) (a) $\theta^* \in C_t$ w.h.p.
(b) C_t gets smaller with increasing t

(ideally, $C_t = \{\theta^*\}$).

(2) Use C_t to determine an optimistic estimate of θ^* using the index for arm $a \in \mathbb{R}^d$ (note: we have an infinite number of arms in this setting) as follows:

$$UCB_t(a) = U_t(a) = \max_{\theta \in C_t} \langle \theta, a \rangle$$

(3) Play:

$$A_t = \max_{a \in A_t} U_t(a).$$

Construction of Confidence Ellipsoid $C_t \subseteq \mathbb{R}^d$:

Recall: At begining of time t, we have observed

$$(A_1, X_1), (A_2, X_2), \dots, (A_{t-1}, X_{t-1}),$$

where $X_s = \langle \theta^*, A_s \rangle + \eta_s$

- $X_s \in \mathbb{R}$
- $A_s \in \mathbb{R}^d$
- η_s → 1-subG noise

Approach: Use a regularized least-square estimator to estimate $\hat{\theta}_{t-1}$, and use an ellipsoid centered at $\hat{\theta}_{t-1}$, and with axes depending on the empirical covariance matrix of the estimator.

$$\hat{\theta}_{t-1} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left(\sum_{s=1}^{t-1} (X_s - \langle \theta, A_s \rangle)^2 + \lambda \|\theta\|_2^2 \right). \quad \text{→ (*)}$$

$\lambda \geq 0$, regularizer. For now, assume $\lambda > 0$.

C_t construction: Solve (*):

$$\hat{\theta}_{t-1} = V_{t-1}^{-1} \sum_{s=1}^{t-1} A_s X_s$$

($A_s \in \mathbb{R}^d$, $X_s \in \mathbb{R}$)

$$V_{t-1} = \left(\lambda I + \sum_{s=1}^{t-1} A_s A_s^\top\right), \quad V_0 = \lambda I .$$

V_{t-1}: $d \times d$, symmetric, p.d. $\Rightarrow$ V_{t-1}^{-1} exists

$$C_t = \left\{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_{t-1}\|^2_{V_{t-1}} \le \beta_t \right\}$$

$$= \left\{ \theta \in \mathbb{R}^d : (\theta - \hat{\theta}_{t-1})^\top V_{t-1} (\theta - \hat{\theta}_{t-1}) \le \beta_t \right\},$$

with

$\{\beta_t, t = 1, 2, \ldots, n\}$ an increasing sequence ($\beta_1 \ge 1$):

$$\sqrt{\beta_n} = \sqrt{\lambda}\, L + \sqrt{2 \ln(1/\delta) + d \ln\left(\frac{d\lambda + nL^2}{d\lambda}\right)}$$

$\sqrt{\beta_n}$: time horizon; L: action norm-bound; δ: prob. $\in (0,1)$, chosen as $\delta = 1/n$; d: dimension.

Roughly, $\beta_n \sim C' \ln(n)$ → (**)

Note: Suppose $A_t = \{e_1, e_2, \cdots, e_d\}$, i.e., the unit vectors in d-dimensions (essentially, a d-armed unstructured bandit, because $\langle \theta^*, e_j \rangle = \theta^*_j$).

Let $T_j(t-1)$ = number of times e_j was played until time $(t-1)$.

Observe $A_s A_s^T = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} (0 \cdots 0 \; 1 \; 0 \cdots 0) = \begin{pmatrix} 0 & & 0 \\ & 1 & \\ 0 & & \ddots \end{pmatrix}$ (j^{th} position)

$$\therefore \quad V_{t-1} = \lambda I + \begin{pmatrix} T_1(t-1) & & 0 \\ & T_2(t-1) & \\ 0 & & \ddots \; T_d(t-1) \end{pmatrix}$$

$$\Rightarrow \hat{\theta}_{t-1} = \begin{pmatrix} \ddots & & 0 \\ & \frac{1}{\lambda + T_j(t-1)} & \\ 0 & & \ddots \end{pmatrix} \begin{pmatrix} \ddots & & 0 \\ & \sum_{s : j \text{ played}} X_s & \\ 0 & & \ddots \end{pmatrix}$$

$$= \begin{pmatrix} \vdots \\ \frac{1}{\lambda + T_j(t-1)} \sum_{s \in j} X_s \\ \vdots \end{pmatrix}$$

i.e., empirical mean, but with λ over-correction.

asymptotically ok as $T_j \to \infty$

$\{s \in j\}$ shorthand for $\{s : A_s = e_j\}$.

$$C_t = \left\{ (\theta - \hat{\theta}_{t-1})^T \begin{pmatrix} \ddots & & 0 \\ & \lambda + T_j(t-1) & \\ 0 & & \ddots \end{pmatrix} (\theta - \hat{\theta}_{t-1}) \leq \beta_t \right\}$$

$$= \left\{ \theta \in \mathbb{R}^d : \sum_{j=1}^{d} (\theta_j - \hat{\theta}_{t-1,j})^2 (\lambda + T_j(t-1)) \leq \beta_t \right\}$$

$$UCB_t(e_j) = U_t(e_j) = \max_{\theta \in C_t} \underbrace{\langle \theta, e_j \rangle}_{\theta_j}$$

choose $\theta_i = \hat{\theta}_{t-1,i} \quad \forall i \neq j$ (so terms are '0' and provides max flexibility along j^{th} coordinate in C_t)

$$= \left(\hat{\theta}_{t-1,j} + \sqrt{\frac{\beta_t}{\lambda + T_j(t-1)}} \right).$$

$$= \frac{\sum_{s \in j} x_s}{\lambda + T_j(t-1)} + \sqrt{\frac{\beta_t}{\lambda + T_j(t-1)}}$$

Scales as $\ln(t)$.

i.e., we get back the usual UCB algorithm.

In general, with more actions, arms provide richer information about all coordinates, thus providing a ellipsoid confidence region.

Regret Analysis: Take 1

Assumptions: (19.1 in textbook)

1. $|\langle \theta^*, a \rangle| \leq 1 \quad \forall\ a \in \bigcup_{t=1}^{n} A_t$
2. $\|a\|_2 \leq L \quad \forall\ a \in \bigcup_{t=1}^{n} A_t$

} both are reasonable

3. $\exists\ \delta \in (0,1)$ s.t. with prob $\geq (1-\delta)$,

$\forall\ t \in \{1, 2, \ldots, n\}, \quad \theta^* \in C_t$ → confidence set defined above

→ This is unjustified at present. C_t is centered around $\hat{\theta}_t$. To justify this assumption, we need to show $\|\theta^* - \hat{\theta}_t\|$ is "small", and that C_t is large enough to contain θ^*.

In future takes on this, this will be justified more carefully.

Thm (19.2 in textbook). Suppose assumption above holds. Then w.p. $\geq (1-\delta)$,

$$\hat{R}_n \doteq \left(\sum_{t=1}^{n} \max_{a_t \in A_t} \langle \theta^*, a_t \rangle - \sum_{t=1}^{n} X_t \right)$$

$$\leq \sqrt{8 n \beta_n \ln\left(\frac{\det(V_n)}{\det(V_0)} \right)}$$

$$\leq \sqrt{8 n \beta_n \ln\left(\frac{d\lambda + nL^2}{d\lambda} \right)}$$

Corollary (19.3 in textbook): Suppose β_n chosen as (*) and $\delta = 1/n$. Then,

$$R_n \leq C d \sqrt{n} \ln(nL).$$

$C > 0$ a fixed constant.

(19.4 in text)

Lemma: V_0 a p.d. matrix, $\{x_1, x_2, \dots x_n\} \in \mathbb{R}^d$, with $\max_{1 \leq k \leq n} \|x_k\|_2 \leq L < \infty$. Then, with $V_t = V_0 + \sum_{s=1}^{t} x_s x_s^T$,

$$\sum_{t=1}^{n} \left(1 \wedge \underbrace{x_t^\top V_{t-1}^{-1} x_n}_{\|x_t\|^2_{V_{t-1}^{-1}}}\right) \le 2 \ln\left(\frac{\det(V_n)}{\det(V_0)}\right)$$

$$\le 2d \ln\left(\frac{\text{trace}(V_0) + nL^2}{d \cdot (\det(V_0))^{1/d}}\right).$$

Pf: For any $u \ge 0$, we have $(u \wedge 1) \le 2\ln(1+u)$. ($\wedge$ = min)

Thus,

$$\sum_{t=1}^{n} \left(1 \wedge x_t^\top V_{t-1}^{-1} x_t\right) \le \sum_{t=1}^{n} \ln\left(1 + x_t^\top V_{t-1}^{-1} x_t\right)$$

Now, recall $V_t = V_{t-1} + x_t x_t^\top$

$$= V_{t-1}^{1/2}\left(I + V_{t-1}^{-1/2} x_t x_t^\top V_{t-1}^{-1/2}\right) V_{t-1}^{1/2}$$

$$\therefore \det(V_t) = \det(V_{t-1}) \det\left(I + V_{t-1}^{-1/2} x_t x_t^\top V_{t-1}^{-1/2}\right)$$

$$= \det(V_{t-1}) \cdot \det\left(I + \left(V_{t-1}^{-1/2} x_t\right)\left(V_{t-1}^{-1/2} x_t\right)^\top\right)$$

Now $\det(I + yy^\top) = (1 + y^\top y)$ ($\because$ eigenvalues are $1+y^\top y, 1, 1, \ldots, 1$)

$$= \det(V_{t-1})\left(1 + x_t^\top V_{t-1}^{-1} x_t\right).$$

$$\therefore \quad \frac{\det(V_t)}{\det(V_{t-1})} = \left(1 + x_t^\top V_{t-1}^{-1} x_t\right)$$

$$\Rightarrow 2\sum_{t=1}^{n} \ln\left(1 + x_t^\top V_{t-1}^{-1} x_t\right) = 2\sum_{t=1}^{n} \ln\left(\frac{\det(V_t)}{\det(V_{t-1})}\right)$$

$$= 2\,\frac{\det(V_n)}{\det(V_0)}.$$

$$\Rightarrow \sum_{t=1}^{n} \left(1 + x_t^\top V_{t-1}^{-1} x_t\right) \leq 2\,\frac{\det(V_n)}{\det(V_0)}.$$

Now recall V_n is p.d. with rank $= d$.

$$\left(\det(V_n)\right)^d = \left(\prod_{i=1}^{d} \lambda_i\right)^d \leq \left(\frac{1}{d}\sum_{i=1}^{d} \lambda_i\right)$$

λ_i: eigenvalues of V_n

AM-GM inequality

$\sum_{i=1}^{d} \lambda_i$ = trace (V_n)

$$\leq \left(\frac{\text{trace}(V_0) + nL^2}{d}\right). \qquad \blacksquare$$

Pf of Thm 19.2: Recall from Assumption that $\theta^* \in C_t$ $\forall t = 1, 2, \cdots, n$. Let $A_t^* = \underset{a_t \in A_t}{\text{argmax}} \langle \theta^*, a_t \rangle$

and $r_t = \langle \theta^*, A_t^* - A_t \rangle$

Recall the LinUCB algorithm: $U_t(a) = \max_{\theta \in C_t} \langle \theta, a \rangle$.

$A_t = \underset{a \in A_t}{\text{argmax}}\ U_t(a)$

Let $\tilde{\theta}_t \in C_t$ be s.t. $\langle \tilde{\theta}_t, A_t \rangle = U_t(A_t)$

(i.e., the upper confidence vector corresponding to the action taken).

$\therefore \langle \theta^*, A_t^* \rangle \leq U_t(A_t^*) \leq U_t(A_t) = \langle \tilde{\theta}_t, A_t \rangle$.

$\theta^* \in C_t$; A_t is the optimizing action in A_t

$\therefore$

$r_t = \langle \theta^*, A_t^* - A_t \rangle = \langle \theta^*, A_t^* \rangle - \langle \theta^*, A_t \rangle$

$\leq \langle \tilde{\theta}_t, A_t \rangle - \langle \theta^*, A_t \rangle = \langle \tilde{\theta}_t - \theta^*, A_t \rangle$

Cauchy Schwartz inequality $\leq \|A_t\|_{V_{t-1}^{-1}} \|\tilde{\theta}_t - \theta^*\|_{V_{t-1}}$

V_{t-1} → symmetric, p.d. matrix

defn. of C_t $\leq 2 \|A_t\|_{V_{t-1}^{-1}} \sqrt{\beta_t}$

Further, by Assumption (a), $|\langle \theta^*, a \rangle| \le 1 \quad \forall a \in \mathcal{A}_t$

$\Rightarrow \quad r_t \le 2.$

$\{\beta_t, t \ge 1\}$ increasing $\Rightarrow \beta_n \ge \max\{1, \beta_t\}.$

$$\therefore \quad r_t \le 2 \wedge 2\sqrt{\beta_t}\, \|A_t\|_{V_{t-1}^{-1}}$$

$$\le 2\sqrt{\beta_t}\left(1 \wedge \|A_t\|_{V_{t-1}^{-1}}\right)$$

$$\therefore \quad \hat{R}_n = \sum_{t=1}^{n} r_t \le \sqrt{n \sum_{t=1}^{n} r_t^2}$$

Cauchy-Schwartz ineq.

$$\le 2\sqrt{n\beta_n \sum_{t=1}^{n} \left(1 \wedge \|A_t\|_{V_{t-1}^{-1}}\right)}$$

Lemma

$$\le \sqrt{8 n \beta_n \ln\left(\frac{d\lambda + nL^2}{d\lambda}\right)}$$

$\blacksquare$

$$\sum_{t=1}^{n} r_t = n \cdot \frac{1}{n} \sum_{t=1}^{n} r_t$$

$$\le n \sqrt{\frac{1}{n} \sum_{t=1}^{n} r_t^2}$$

$$= \sqrt{n \sum_{t=1}^{n} r_t^2}$$

Cauchy Schwartz Inequality.

$$\langle a, b \rangle = a^T V^{-1/2} V^{1/2} b = \left(V^{-1/2} a\right)^T \left(V^{1/2} b\right)$$

$$\leq \left(a^T V^{-1} a\right)^{1/2} \left(b^T V b\right)^{1/2}.$$

$$= \|a\|_{V^{-1}} \|b\|_V$$