- Multi-armed bandits (MAB)

So far in online learning: agent receive instructive feedback

MAB: A basic model of agent making sequential decisions and receiving evaluative feedback.

A slot machines

a slot machine "one armed bandit".

① Setup: action set $A = \{1, \cdots A\}$. (arm set)

For $t = 1, 2, \cdots T$:

agent takes action a_t (arm)

agent receives reward $r_t = f^*(a_t) + \xi_t$ → (zero mean, 1-SG noise)

$f^*: A \to [0,1]$,

agent's expected reward $= \mathbb{E}\left[\sum_{t=1}^{T} r_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} f^*(a_t)\right]$

optimal strategy if f^* is known: take $a^* = \underset{a \in A}{\operatorname{argmax}}\, f^*(a)$

optimal expected reward $= T \cdot f^*(a^*)$.

Performance measure: regret:

$$\mathrm{Reg}(T) = T f^*(a^*) - \mathbb{E}\left[\sum_{t=1}^{T} f^*(a_t)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \left(f^*(a^*) - f^*(a_t)\right)\right]$$

How to design a strategy for the agent?

* need to learn the f^* fn thru noisy observations (exploration) → Take diverse actions

* need to take actions a_t that has large f^* value (exploitation) → Take actions that maximizes f^*.

Conflicting!

② Algorithms for MAB.

First idea: explore-then-exploit (commit)

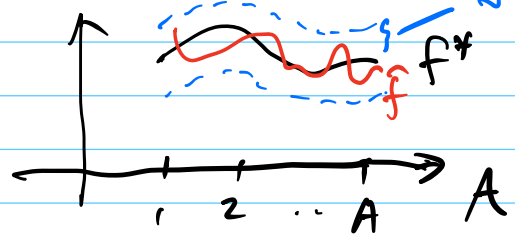
Phase 1. use first T_0 rounds to estimate f^* by taking actions in a round robin fashion $\to \hat{f}$

Phase 2. from T_0+1 round on. take action $\hat{a} = \arg\max_{a \in A} \hat{f}(a)$

Analysis sketch:

Phase 1 regret $\le O(T_0)$

Phase 2:



Hoeffding ↓

$\max_a |\hat{f}(a) - f^*(a)| \le O(\sqrt{\frac{A}{T_0}})$

$\Rightarrow f^*(a^*) - f^*(\hat{a}) \le O(\sqrt{\frac{A}{T_0}})$

$\Rightarrow$ Phase 2 regret $\le O((T-T_0)\sqrt{\frac{A}{T_0}})$

$\Rightarrow Reg(T) \le O(T_0 + T\sqrt{\frac{A}{T_0}}) = O(A^{\frac{1}{3}} T^{\frac{2}{3}})$.

↑ choosing T_0 optimally

Second idea: intersperse explore & exploit

ε-greedy.

For $t = 1, 2, \cdots T$:

Flip a coin Z w/ head prob. ε.

(Explore) If Z = head, take $a_t \sim \text{unif}(\{1 \cdots A\})$

(Exploit) If Z = tail, take $a_t = \left(\hat{a}_t = \arg\max_a \hat{f}_t(a) \right)$

(Notation: $\hat{f}_t(a)$ = avg reward of a at t

$$\frac{\text{total reward of } a}{\#\text{ times } a \text{ taken}} = \frac{\sum_{i=1}^{t-1} I(a_i = a)\, r_i}{\left(N_{t-1}(a) \longrightarrow \sum_{i=1}^{t-1} I(a_i = a)\right)}$$

Analysis sketch

- at step t, each a has $\approx t\frac{\varepsilon}{A}$ reward samples

$$\Rightarrow \max_a \left| \hat{f}_t(a) - f^*(a) \right| \le O\left(\sqrt{\frac{A}{t\varepsilon}}\right).$$

$$\Rightarrow f^*(a^*) - f^*(\hat{a}_t) \le O\left(\sqrt{\frac{A}{t\varepsilon}}\right)$$

$\Rightarrow$ "on average", a_t is $\left(\varepsilon + \sqrt{\frac{A}{t\varepsilon}}\right)$-optimal

- $\text{Reg}(T) \le O\left(\sum_{t=1}^{T} \left(\varepsilon + \sqrt{\frac{A}{t\varepsilon}} \right) \right)$

$$= O\left(\varepsilon T + \sqrt{\frac{AT}{\varepsilon}} \right).$$

setting ε optimally

$$= O\left(A^{\frac{1}{3}} T^{\frac{2}{3}} \right).$$

Can we do better?

A better idea: optimism in face of uncertainty.
(abbrev. optimism principle)

In words:
Act according to the best plausible world.
"optimistic world model"

Why it works:

If the optimistic world model is

- correct ✓ → no regret ✓
- wrong ✓ → learn new things, avoid making the same mistake in the future ✓

Real world agents are using this (perhaps w/o realizing it)
the optimism bias (Sharot, 2011).

How to define the "best plausible world" in MAB?

- For every action a, what's the highest plausible $f^*(a)$ given data?

upper confidence bound for $f^*(a)$!
(UCB)

The UCB Algorithm:

For $t = 1, \cdots T$:

- Define $UCB_t(a) = \hat{f}_t(a) + b_t(a)$ → define to be 0 if $N_{t-1}(a) = 0$.

confidence width ⇒ "bonus" of action a.

$$b_t(a) = \sqrt{\frac{\ell}{N_{t-1}(a)+1}}, \quad \ell = 8\ln(2AT).$$

- choose $a_t = \underset{a \in A}{\operatorname{argmax}}\ UCB_t(a)$.

(optimism property)

Validity of the UCB's:

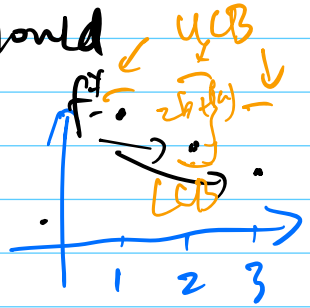
Lemma: there exists an event E. $P(E) \geq 1 - \frac{2}{T}$.

and when E happens:

(*) $\forall a, \forall t, \quad \left| \hat{f}_t(a) - f^*(a) \right| \leq b_t(a) = \sqrt{\frac{\ell}{N_{t-1}(a)+1}}$

($\Rightarrow UCB_t(a) \geq \mu(a)$)

Intuition: arm a reward estimate should have precision $\propto \frac{1}{\sqrt{\#\text{samples in arm } a}}$



pf idea:

First idea: $E_{t,a} = \left\{ \left| \hat{f}_t(a) - f^*(a) \right| \leq b_t(a) \right\}$

Hoeffding $\Rightarrow P(E_{t,a}) \geq 1-\delta$?

Problem: $\hat{f}_t(a)$ is an average over $N_{t-1}(a)$ samples, $N_{t-1}(a)$ itself is random!

A fix: $E_{t,a,n} = \left\{ \left| \hat{f}_t(a) - f^*(a) \right| \leq \sqrt{\frac{1}{n+1}}, \; N_{t-1}(a) = n \right\}$

Hoeffding $\Rightarrow P(E_{t,a,n}) \geq 1 - \frac{2}{AT^3}$

Define $E = \bigcap_{t=1}^{T} \bigcap_{a=1}^{A} \bigcap_{n=1}^{T} E_{t,a,n}$.

$\Rightarrow P(E) \geq 1 - \frac{2}{T}$.

can show that on E, (*) happens.

More details: Slivkins, "Introduction to Multi-armed bandits", 2019. See. 1.3.1.

Regret analysis of UCB:

Thm: Assume that UCB guarantees that $Reg(T) \leq \tilde{O}(\sqrt{AT})$

big-O ignoring log factors

Remark: — regret bound much better than $AT^{2/3}$ in ETC & ε-greedy

— in general, unimprovable (there is a lower)

- can also show UCB has instance dept bound:

$$Reg(T) = \tilde{O}\left(\sum_{a \neq a^*} \frac{\ln T}{\Delta_a}\right), \quad \Delta_a = f^*(a^*) - f^*(a)$$

Δ_a larger, arm is less confusing.

Pf:

Recall: $$Reg(T) = \mathbb{E}\left[\sum_{t=1}^{T} \underbrace{\left(f^*(a^*) - f^*(a_t)\right)}_{r(t)}\right]$$

$$= \mathbb{E}\left[\left(\sum_{t=1}^{T} reg(t)\right)\left(I(E) + I(E^c)\right)\right]$$

Note: $$\mathbb{E}\left[\left(\sum_{t=1}^{T} reg(t)\right) I(E^c)\right] \leq T \cdot \mathbb{E}[I(E^c)]$$

$$= T \cdot P(E^c)$$

$$\leq T \cdot \frac{2}{T} = 2. \quad (1)$$

we just need to control $\sum_{t=1}^{T} reg(t)$ when E happens.

Insight: can bound $reg(t)$ by $b_t(a_t)$: bonus on the action taken

$$reg(t) = f^*(a^*) - f^*(a_t)$$

validity of UCB $\leq UCB_t(a^*) - f^*(a_t)$

choice of a_t $\leq UCB_t(a_t) - f^*(a_t)$

$$= \hat{f}_t(a_t) + b_t(a_t) - f^*(a_t)$$

event E $\quad \le 2 b_t(a_t)$.

Insight 2: $\sum_{t=1}^{T} b_t(a_t)$ is controlled. $\quad b_t(a) = \sqrt{\frac{\ell}{N_{t-1}(a)+1}}$

(every time a is chosen as a_t, its bonus $b_t(a)$ will shrink a bit)

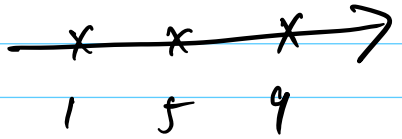
$$\sum_{t=1}^{T} \text{reg}(t)$$

$$\le 2 \sum_{t=1}^{T} b_t(a_t)$$

$$= 2 \sum_{a=1}^{A} \sum_{t: a_t = a} b_t(a) \qquad \text{(group according to } a_t\text{)}$$

a decreasing sequence

say $a = 1$.



$$\sqrt{\frac{\ell}{0+1}} \quad \sqrt{\frac{\ell}{1+1}} \quad \sqrt{\frac{\ell}{2+1}}$$

$$= 2 \sum_{a=1}^{A} \sum_{t: a_t = a} \sqrt{\frac{\ell}{N_{t-1}(a)+1}}$$

$$= 2 \sum_{a=1}^{A} \sum_{n=1}^{N_T(a)} \sqrt{\frac{\ell}{n}}$$

$$\left(\sum_{n=1}^{N} \sqrt{\frac{1}{n}} \le 2\sqrt{N} \right)$$

$$= 4\sqrt{\ell} \sum_{a=1}^{A} \sqrt{N_T(a)}$$

$$\left(\frac{1}{A} \sum_{a=1}^{A} \sqrt{N_T(a)} \right.$$

$$\le 4\sqrt{\ell}\sqrt{AT}$$

$$\leq \sqrt{\frac{1}{A}\sum_{a=1}^{A} N_T(a)}$$

$$= \sqrt{\frac{T}{A}}$$

by Jensen & concavity of $x \to \sqrt{x}$)

Therefore,

$$\mathbb{E}\left[\left(\sum_{t=1}^{T} reg(t)\right) I(E)\right] \leq 4\sqrt{lAT} \quad (2)$$

summing up (1) (2)

$$\Rightarrow \quad Reg(T) \leq 4\sqrt{lAT} + 2 = \tilde{O}(\sqrt{AT}).$$

Remarks on designing optimism-based algs:

(1) Define bonus so UCB's are valid wh.p. (so $reg_t \leq 2b_t(a_t)$)

(2). subject to (1). design bonus to be as tight as possible. (so $\sum_t b_t(a_t)$ is small).

Stochastic linear (contextual) bandits

① setup.

Application: recommendation with personalization

MAB: recommend a single product so that it has maximum overall satisfaction over population

Contextual Bandits: ~~for~~ different users. recommend different products that fit their respective needs.

Protocol:

For $t = 1, 2, \cdots T$:

(generalizes MAB by allowing contexts)

observes context $x_t \in \mathcal{X}$

takes action $a_t \in \mathcal{A}$

receive reward $r_t = \underbrace{f^*(x_t, a_t)}_{\text{unknown}} + \underbrace{\xi_t}_{\text{zero-mean, 1-SG}}$

Goal: maximize $\mathbb{E}\left[\sum_{t=1}^{T} r_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} f^*(x_t, a_t)\right]$

This lecture:- Assumption (Linear Realizability) $f^* \in \mathcal{F} = \{ f(x,a) = \langle \theta, \phi(x,a) \rangle : \|\theta\|_2 \le 1 \}$ with ϕ known

($\theta \in \mathbb{R}^d$, $\phi(x,a) \in \mathbb{R}^d$)

- assumption: $\|\phi(x,a)\|_2 \le 1$ for all x, a.

- Notation:

$f^* = \langle \theta^*, \phi(x,a) \rangle$; θ^* needs to be learned.

What's the best strategy had f^* been known?

$\forall x$, take action $a = \pi_{f^*}(x) := \arg\max_{a \in A} f^*(x, a)$.

optimal policy

Regret notion:

$$\text{Reg}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \max_a f^*(x_t, a) - \sum_{t=1}^{T} f^*(x_t, a_t)\right].$$

PReg(T): pseudo regret

reward by following optimal policy

reward of the agent

Practical relevance: "A contextual bandit approach to personalized News Article Recommendation". (Li, Chu, Langford, Schapire, 2010, WWW) Test of Time Award.

MAB as linear bandits:

Every K-armed MAB problem can be viewed as a K-dimensional Linear bandit problem by defining:

- $x_t = z_0 \quad \forall t$ (dummy context)
- $\phi(z_0, a) = e_a = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \leftarrow a$-th location $\in \mathbb{R}^k$
- $\theta^* = (f^*(1) \cdots f^*(K))$, thus

$$f^*(a) = \langle \theta^*, \phi(z_0, a) \rangle \quad \forall a.$$

② Designing algorithms for Linear bandits:

Designing algorithm ...

Idea: optimism principle. again.

↓

Act according to the best plausible world.

* "world model" = reward model = Θ.
* "plausible world" = set of Θ's that can plausibly be θ^*. i.e. can explain observations $(x_s, a_s, r_s)_{s=1}^{t-1}$ At t:

Similar to constructing confidence intervals in MAB. here we construct confidence set for θ^*, say.

$$\mathbb{P}(\forall t, \theta^* \in \mathcal{C}_t) \geq 1 - \frac{1}{T}.$$

Algorithm (OFUL: optimism in the face of uncertainty for Linear bandits)

For $t = 1, 2, \cdots T$:

- construct confidence set $\mathcal{C}_t$ for θ^*.
- Observe x_t.
- Take action $a_t = \underset{a \in A}{\operatorname{argmax}} \max_{\theta \in \mathcal{C}_t} \langle \theta, \phi(x_t, a) \rangle$

The largest plausible reward action a can get

Q1: How to construct confidence set $\mathcal{C}_t$?

Q2 How to analyze the OFUL alg?

Q1: Recall: in MAB, confidence interval for $f^*(a)$

$$= \left[\underbrace{\hat{f}_t(a)}_{\text{center: some best guess of } f^*(a)} \pm \underbrace{b_t(a)}_{\text{width deduced by concentration ineq. (eg. Hoeff ding)}} \right]$$

What's our "best guess" of θ^* based on $(x_s, a_s, r_s)_{s=1}^{t-1}$?

Recall: $r_s = \langle \theta^*, \phi(x_s, a_s) \rangle + \varepsilon_s$

Estimating θ^* from data is a linear regression problem

Standard solution: least squares + regularization

$$\hat{\theta}_t(\lambda) = \underset{\theta}{\operatorname{argmin}} \sum_{s=1}^{t-1} \left(\langle \theta, \underbrace{\phi(x_s, a_s)}_{\phi_s} \rangle - r_s \right)^2 + \lambda \|\theta\|_2^2$$

How close is $\hat{\theta}_t(\lambda)$ to θ^*?

Note: in general, cannot claim, say $\|\hat{\theta}_t(\lambda) - \theta^*\|_2 \le \frac{1}{\sqrt{t}}$

Problem: data may be highly "degenerate"

Example:

$d = 2$.

$\phi_1 = \phi_2 = \cdots = \phi_{t-2} = e_1 \quad \phi_{t-1} = e_2$

$r_1 = \theta_1^* + \varepsilon_1$

$r_2 = \theta_1^* + \varepsilon_2$

$\vdots$

$r_{t-2} = \theta_1^* + \varepsilon_{t-2}$

$r_{t-1} = \theta_2^* + \varepsilon_t$ (only one data pt is related).

The data reveals little information about θ_2^*. so cannot hope to show $|\hat{\theta}_2^t - \theta_2^*| = O(\frac{1}{\sqrt{t}})$.

In general. we will show the following estimation guarantee:

Lemma: $\exists$ event E. $P(E) \geq 1 - \frac{1}{T}$, on E for all t,

$\theta^* \in C_t(\lambda) = \cdots$ $\quad \|\hat{\theta}^t(\lambda) - \theta^*\|_{V_{t-1}(\lambda)} \leq \beta_t(\lambda) = \tilde{O}(\sqrt{\lambda} + d)$

Specifically. $\lambda = 1 \Rightarrow$

$\theta^* \in C_t(1) = \cdots$ $\quad \|\hat{\theta}^t(1) - \theta^*\|_{V_{t-1}(1)} \leq \beta_t(1) = \tilde{O}(d)$

This, although not a pointwise guarantee, turns out to be still useful.

Here: $\|x\|_M := \sqrt{x^T M x}$ For $M > 0$ This is a norm: $\|x\|_M \geq 0$.

$= \|M^{\frac{1}{2}} x\|_2$ psd $M = U(\lambda_1 \cdots \lambda_d) U^T$ $M^{\frac{1}{2}} = U(\sqrt{\lambda_1} \cdots \sqrt{\lambda_d}) U^T$

$\|\alpha x\|_M = |\alpha| \, \|x\|_M$

$\|x+y\|_M \leq \|x\|_M + \|y\|_M$ generalized Cauchy Schwarz.

$\langle x, y \rangle \leq \|x\|_M \|y\|_{M^{-1}}$

$V_{t-1}(\lambda) = \sum_{s=1}^{t-1} \phi_s \phi_s^T + \lambda I$

Intuition check on previous example w/ $\lambda = 0$:

$\hat{\theta}^t = \underset{\theta}{\operatorname{argmin}} \sum_{s=1}^{t-2} (\theta_1 - (\theta_1^* + \varepsilon_s))^2 + (\theta_2 - (\theta_2^* + \varepsilon_{t-1}))^2$

$\theta_1 = \frac{1}{t-2} \sum_{s=1}^{t-2} (\theta_1^* + \varepsilon_s)$ $\qquad \theta_2 = \theta_2^* + \varepsilon_{t-1}$

Let $z = \hat{\theta}^t - \theta^*$

$\Rightarrow |z_1| \leq O(\sqrt{\frac{1}{t}})$, $|z_2| \leq O(1)$.

* Note: $V_{t-1} = (t-1) e_1 e_1^T + 1 \cdot e_2 e_2^T = \begin{pmatrix} t-1 & 0 \\ 0 & 1 \end{pmatrix}$

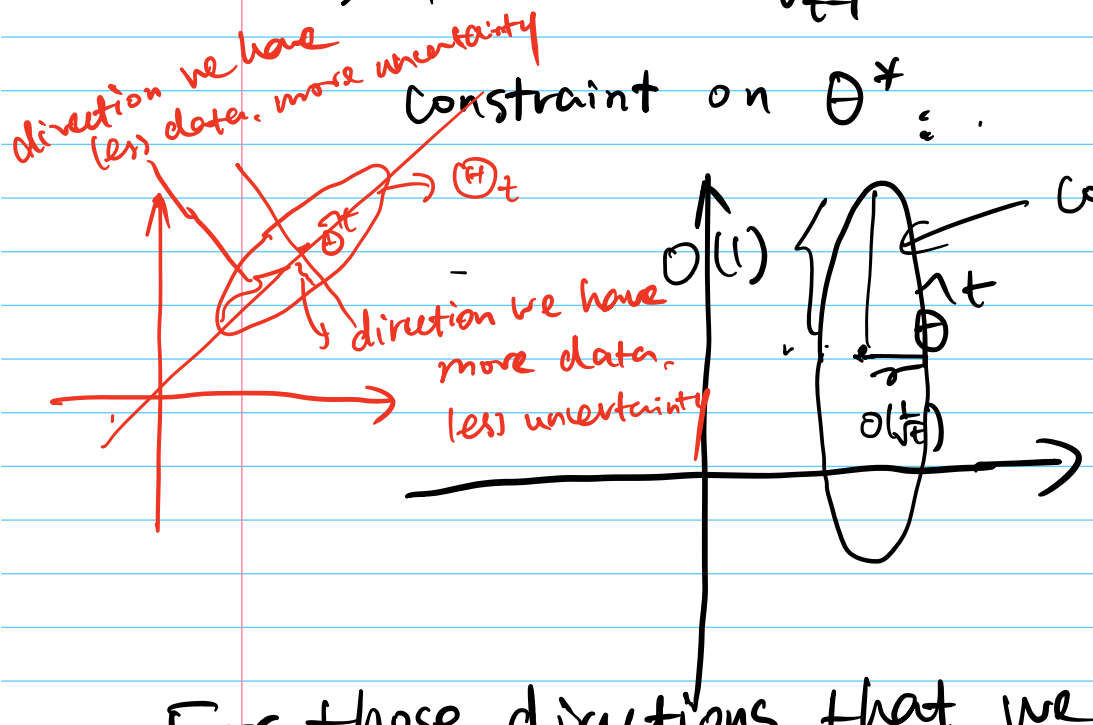
$\Rightarrow \|z\|_{V_{t-1}}^2 = (z_1, z_2) \begin{pmatrix} t-1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$

$$= (t-1)\, z_1^2 + z_2^2$$

$$\le O\left(t \cdot \frac{1}{t} + 1 - \frac{1}{1}\right) = O(1)$$

* Why is $\|z\|^2_{V_{t-1}} \le O(1)$ guarantee useful?

$\Rightarrow \|\theta^* - \hat{\theta}^t\|^2_{V_{t-1}} \le O(1)$ gives a constraint on θ^*.



For those directions that we have more data, our estimate on θ^* is more accurate.

Proof of the Lemma: ($V_{t-1}(\lambda)$ abbrev. V_{t-1} below)

First, $\hat{\theta}_t(\lambda) = V_{t-1}^{-1} \left(\sum_{s=1}^{t-1} \phi_s r_s \right)$

$$= V_{t-1}^{-1} \left(\underbrace{\sum_{s=1}^{t-1} \phi_s \phi_s^T \theta^*}_{V_{t-1} - \lambda I} + \underbrace{\sum_{s=1}^{t-1} \phi_s \eta_s}_{m_{t-1}} \right)$$

$$= \theta^* - V_{t-1}^{-1} \lambda \cdot \theta^* + V_{t-1}^{-1} m_{t-1}$$

$$\Rightarrow \hat{\theta}_t(\lambda) - \theta^* = -V_{t-1}^{-1} \lambda \theta^* + V_{t-1}^{-1} m_{t-1}$$

$$\Rightarrow \| \hat{\theta}_t(\lambda) - \theta^* \|_{V_{t-1}} \leq \| V_{t-1}^{-1} \lambda \theta^* \|_{V_{t-1}} + \| V_{t-1}^{-1} m_{t-1} \|_{V_{t-1}}$$

$$\overset{\text{def of Mahalanobis norm}}{=} \| \lambda \theta^* \|_{V_{t-1}^{-1}} + \| m_{t-1} \|_{V_{t-1}^{-1}}$$

$$\overset{\text{w.p. } 1-\frac{1}{T^2}}{\leq} \sqrt{\lambda} + \sqrt{2 \ln(T^2) + d \ln\left(1 + \frac{t}{d\lambda}\right)}$$

(exercise)

union bound over all $t \Rightarrow$ Lemma.

In the last step, to deal with $\| m_{t-1} \|_{V_{t-1}^{-1}}$, we use self-normalized Tail Inequality for vector-valued martingales:

martingale

Lemma (Abbasi-Yadkori, Pál, Szepesvári, 2011)

$\forall t$,

$$P\left(\left\| \sum_{s=1}^{t} \phi_s \eta_s \right\|_{V_t^{-1}(\lambda)} \leq \sqrt{2 \ln \frac{1}{\delta} + \ln \underbrace{\frac{\det(V_t(\lambda))}{\det(V_0(\lambda))}}_{\leq d \ln\left(1 + \frac{t}{d\lambda}\right)}} \right) \geq 1 - \delta$$

Intuition: when all $\eta_s \sim N(0,1)$ and ϕ_s's are chosen ahead of time

$$m_t = \sum_{s=1}^{t} \phi_s \eta_s \overset{d}{\approx} N(0, V_t(\lambda))$$

$V_t^{-\frac{1}{2}}(\lambda) \cdot m_t \quad \overset{d}{\approx} N(0, I_d)$

$$\| V_t^{-\frac{1}{2}}(\lambda) \, m_t \|_2^2 \overset{\text{whp.}}{\leq} O\left(d + \ln \frac{1}{\delta}\right).$$

problem with this intuition: ϕ_s's are not fixed.

ϕ_s depends on $\phi_1 \varepsilon_1 \cdots \phi_{s-1} \varepsilon_{s-1}$.

Q1 summary:

$\forall t$. define $\Theta_t = \{ \theta : \| \theta - \hat{\theta}^t(1) \|_{V_{t-1}(1)} \leq \beta_t(1) \}$

$$P(\forall t, \ \theta^* \in \Theta_t) \geq 1 - \frac{1}{T}.$$

Algorithm induced by this Θ_t construction:

$$a_t = \arg\max_{a \in A} \left[\max_{\theta \in \Theta_t} \langle \theta, \phi(x_t, a) \rangle \right] \rightarrow \text{Interpretation:} \ =: UCB_t(x_t, a)$$

$$= \max_{x : \|x\|_{V_{t-1}(1)} \leq \beta_t(1)} \langle \hat{\theta}^t(1) + x, \phi(x_t, a) \rangle$$

$\theta^* \in \Theta_t \Rightarrow$ this is $\geq \langle \theta^*, \phi(x_t, a) \rangle = f^*(x_t, a)$

|| Exercise

$$\langle \hat{\theta}^t(1), \phi(x_t, a) \rangle + \beta_t(1) \| \phi(x_t, a) \|_{V_{t-1}(1)^{-1}}$$

↑ Estimated reward

$b_t(a)$: uncertainty / exploration bonus

$\phi(z_0, a) = e_a$

$V_{t-1}(1) = \sum_{s=1}^{t-1} \phi(z_0, a_s) \phi(z_0, a_s)^T + I$

$= (N_{t-1}(1) + 1$

For MAB: $\sqrt{K}$ $\quad \sqrt{\frac{1}{N_{t-1}(a) + 1}}$

$N_{t-1}(A)+1$)

Q2 Regret analysis

Thm: on E $(P(E) \geq 1-\frac{1}{T})$, $PReg(T) \leq \tilde{O}(d\sqrt{T})$.

$(\Rightarrow Reg(T) \leq \tilde{O}(d\sqrt{T}).)$

(For MAB, what regret bound do we get ?)

Pf: Recall $PReg(T) = \sum_{t=1}^{T} reg(t)$, where

$$reg(t) = \max_a f^*(x_t, a) - f^*(x_t, a_t)$$

How to bound? Similar to MAB analysis

Step 1: $reg(t) \leq 2\, b_t(a_t)$

Why:

$$reg(t) = \max_a f^*(x_t, a) - f^*(x_t, a_t)$$

$$\leq \max_a UCB_t(a) - f^*(x_t, a_t)$$

$$= UCB_t(a_t) - f^*(x_t, a_t)$$

$$= \langle \hat{\theta}^t(1), -\theta^* , \phi(x_t, a_t) \rangle + b_t(a_t)$$

$$\leq \underbrace{\| \hat{\theta}^t(1) - \theta^* \|_{V_{t-1}(1)}}_{\leq \beta_t(1)} \| \phi(x_t, a_t) \|_{V_{t-1}(1)^{-1}} + b_t(a_t) \leq b_t(a_t)$$

$$= 2\, b_t(a_t)$$

Step 2: bounding $\sum_t b_t(a_t)$

how:

$$\sum_t b_t(a_t) = \sum_t \beta_t(1) \| \phi(x_t, a_t) \|_{V_{t-1}(1)^{-1}}$$

$$\leq \underbrace{\left(\max_t \beta_t(1) \right)}_{\tilde{O}(\sqrt{d})} \cdot \underbrace{\left(\sum_{t=1}^{T} \| \phi_t \|_{V_{t-1}(1)^{-1}} \right)}_{(*)}$$

$$(*) = \sum_{t=1}^{T} \| \phi_t \|_{V_{t-1}(1)^{-1}}$$

$$\leq \sqrt{\sum_{t=1}^{T} \| \phi_t \|^2_{V_{t-1}(1)^{-1}} \cdot T}$$

$$= \sqrt{\sum_{t=1}^{T} \phi_t^T V_{t-1}(1)^{-1} \phi_t \cdot T}$$

Note: $V_t(1) \leq 2 V_{t-1}(1) \Rightarrow V_{t-1}(1)^{-1} \leq 2 V_t(1)^{-1}$

special case: $d=1$.

- $A \leq B \Leftrightarrow A - B \geq 0$. positive semi definite order of matrices
- property: $A \leq B \Rightarrow \forall x.\ x^T A x \leq x^T B x$

  $A \leq B \Rightarrow B^{-1} \leq A^{-1}$.

$$\leq \sqrt{\sum_{t=1}^{T} \phi_t^T V_t(1)^{-1} \phi_t \cdot 2T}$$

The elliptic potential lemma: (EPL)

suppose $u_1, \dots, u_T \in \mathbb{R}^d$. $A_t = \mu I + \sum_{s=1}^{t} u_s u_s^T$, then ($A_0 = \mu I$)

$$\sum_{t=1}^{T} \| u_t \|^2_{A_t^{-1}} \leq \ln \frac{\det(A_T)}{\det(A_0)}$$

if all $\|u_t\| \le D$

$$\le d \ln\left(1 + \frac{D^2 T}{d\mu}\right) = \tilde{O}(d)$$

$$\le \tilde{O}(\sqrt{dT})$$

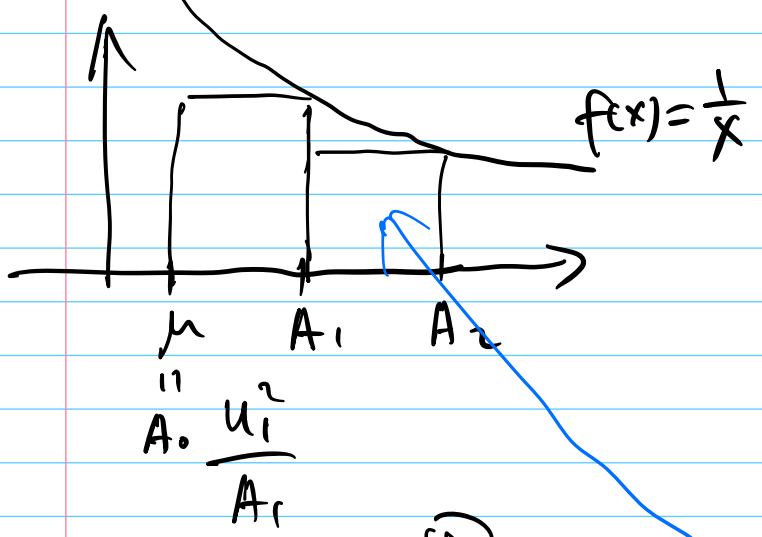
$$\Rightarrow \sum_{t=1}^{T} b_t(a_t) \le \tilde{O}(\sqrt{d}) \cdot \tilde{O}(\sqrt{dT}) = \tilde{O}(d\sqrt{T})$$

∎

Intuition of EPL : when $d=1$.

$$A_t = \mu + \sum_{s=1}^{t} u_s^2$$

$$\sum_{t=1}^{T} \frac{u_t^2}{A_t} \le \int_{\mu}^{A_T} \frac{1}{x}\, dx = \ln \frac{A_T}{\mu}$$



pf idea of EPL:

① Show that $\|u_t\|^2_{A_t^{-1}} \le \ln \frac{\det(A_t)}{\det(A_{t-1})}$

(1-d: $\frac{u_t^2}{A_t} \le \ln \frac{A_t}{A_{t-1}}$)

Note: $\ln \frac{\det(A_t)}{\det(A_{t-1})}$

$$= \ln\left(\det\left(A_{t-1}^{-\frac{1}{2}} A_t A_{t-1}^{-\frac{1}{2}}\right)\right)$$

$$= \ln\left(\det\left(I + A_{t-1}^{-\frac{1}{2}} u_t u_t^T A_t^{-\frac{1}{2}}\right)\right)$$

Exercise (Hint: $\det(M) = \prod_{i=1}^{d} \lambda_i(M)$)

$$= \ln\left(1 + \|u_t\|^2_{A_{t-1}^{-1}}\right)$$

$$\geq \frac{\|u_t\|^2_{A_{t-1}^{-1}}}{1 + \|u_t\|^2_{A_{t-1}^{-1}}} \qquad (\forall x \geq 0.\ \ln(1+x) \geq \frac{x}{1+x})$$

$$= \|u_t\|^2_{A_t^{-1}} \qquad (A_t^{-1} = (A_t + u_t u_t^T)^{-1}$$

Sherman-Morrison

$$= A_{t-1}^{-1} - \frac{A_{t-1}^{-1} u_t u_t^T A_{t-1}^{-1}}{1 + u_t^T A_{t-1}^{-1} u_t}$$

② Show $\ln \frac{\det(A_T)}{\det(A_0)} = \tilde{O}(d)$

Fact: for p.s.d matrices $A = U \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix} U^T$, $\det(A) = \prod_{i=1}^{d} \lambda_i$

For $A_0 = \begin{pmatrix} \mu & & \\ & \ddots & \\ & & \mu \end{pmatrix} \Rightarrow \det(A_0) = \mu^d$.

For $A_T = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix} \Rightarrow \det(A_T) = \prod_{i=1}^{d} \lambda_i \overset{\text{AM-GM}}{\leq} \left(\frac{1}{d}\sum_{i=1}^{d} \lambda_i\right)^d$

$$= \left(\frac{1}{d}\operatorname{tr}(A_T)\right)^d$$

$$\overset{\text{Ex}}{\leq} \left(\frac{1}{d}(d\mu + TD^2)\right)^d. \qquad \text{✗}$$