Stochastic Bandits

CS245: Online Optimization and Learning

Xin Liu
SIST, ShanghaiTech University

Review of Adversarial Bandits

OMD for Adversarial Bandits:

Initialization: $x_1 = [1/K, \cdots, 1/K]$ and η.

For each day $t = 1, \cdots, T$:

- **Learner:** Sample an arm i from x_t.
- **Environment:** Observe the reward of arm i : $r_t(i)$.
- **Reward Estimator:** $\hat{r}_t(i) = r_t(i)/x_t(i)$ and 0 otherwise.
- **Update:** $x_{t+1} = \arg\min_{\mathcal{K}} \ \langle x, -\hat{r}_t \rangle + \frac{1}{\eta} B_\psi(x; x_t)$.

Adversarial bandits:

- Algorithm – OMD with unbiased importance estimator
- Theory – achieve $O(\sqrt{TK \log K})$ regret
- Analysis – reduction from "bandits" to "full info"

From Adversarial Bandits to Stochastic Bandits

Adversarial Bandit problem:

Initialization: K arms.

For each round $t = 1, \cdots, T$:

- **Learner:** Pull an arm $i \in [K]$.
- **Environment:** Observe the reward of the arm $r_t(i)$, which could be arbitrary and adversarial.

Stochastic Bandit problem:

Initialization: K arms.

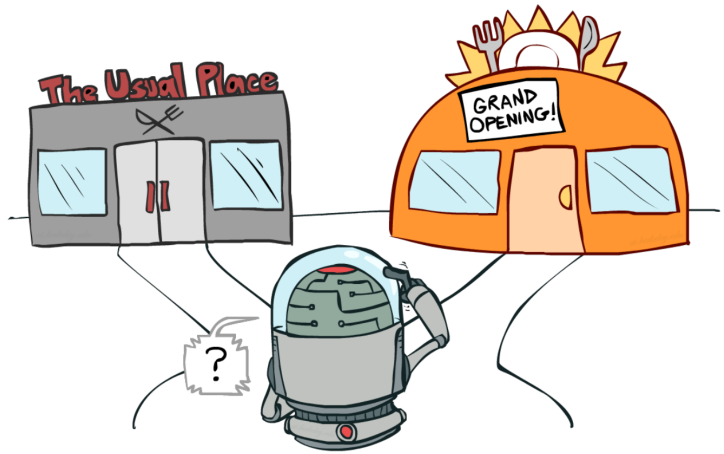
For each round $t = 1, \cdots, T$:

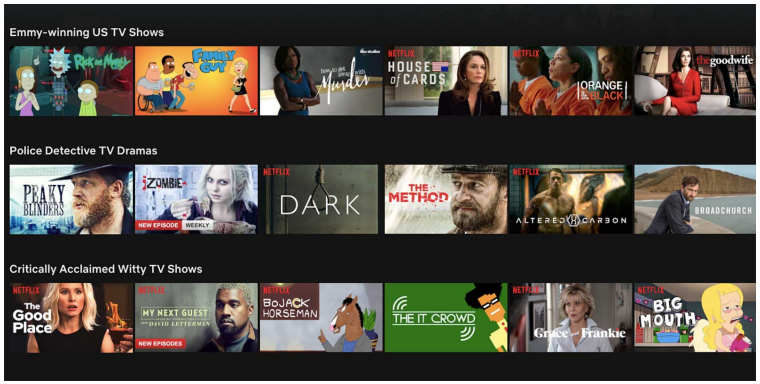
- **Learner:** Pull an arm $i \in [K]$.
- **Environment:** Observe the reward of the arm $r_t(i)$, which is stochastic from some unknown distribution.

Motivation

Bandits in our daily life:

- Restaurant selection.
- Movie recommendation.
- Go game.

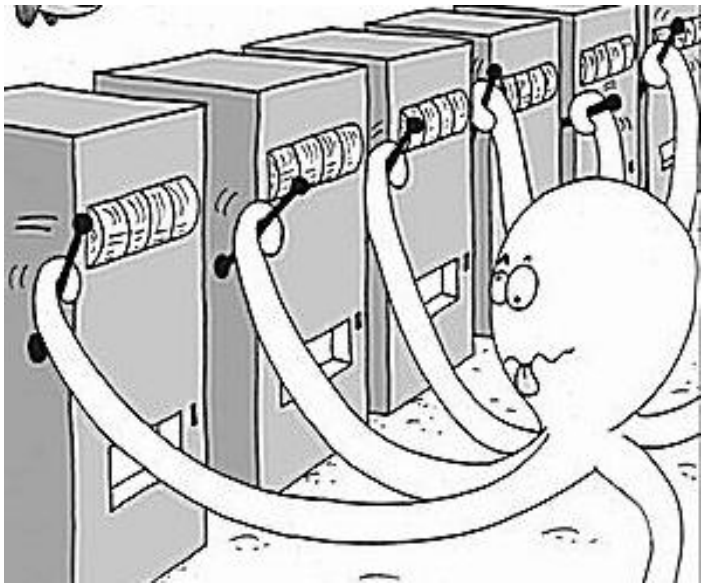




Multi-Armed Bandits - Intro

Multi-Armed Bandits:

- Multiple slot machines with unknown reward probability.
- Play one machine at a time.
- How to earn as much as money as possible?



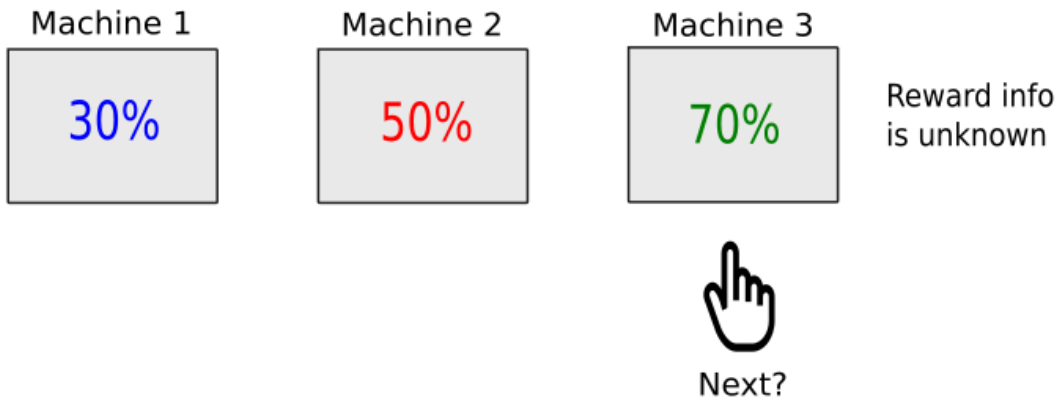
Multi-Armed Bandits - Intro

Multi-Armed Bandits:

- Three slot machines with sampled reward probabilities

$$(3/10,\ 10/20,\ 70/100)$$

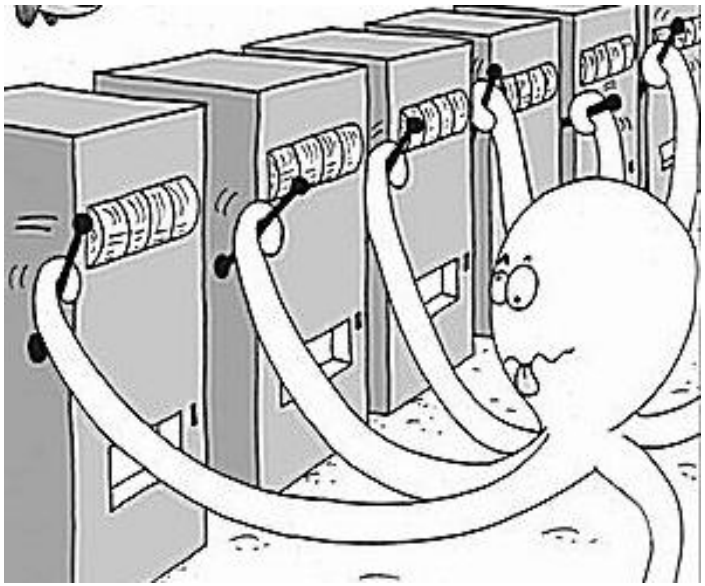
- How to pick next? Stick to Machine 3 or try other machines?
- Exploration and exploitation dilemma.



Multi-Armed Bandits - Model

Multi-Armed Bandits:

- K independent bandits and T rounds.
- At time slot $t \in [T]$, a user chooses an action $a \in [K]$ and receives the reward $R_{t,a} \in \{0, 1\}$, with $\mathbb{E}[R_{t,a}] = r_a$ (Unknown).
- Trajectory is $\{a_1, R_{1,a_1}, a_2, R_{2,a_2}, \cdots, a_T, R_{T,a_T}\}$.



Multi-Armed Bandits - Model

The objective is to design an algorithm/policy to maximize the total expected reward:

$$\mathbb{E}\left[\sum_{t=1}^{T} R_{t,a_t}\right].$$

It is equivalent to minimize regret:

$$\mathcal{R}(T) = Tr_{a^*} - \mathbb{E}\left[\sum_{t=1}^{T} R_{t,a_t}\right],$$

where a^* is the best arm.

Multi-Armed Bandits - Intuition

We want to minimize regret:

$$\mathcal{R}(T) = Tr_{a^*} - \mathbb{E}\left[\sum_{t=1}^{T} R_{t,a_t}\right].$$

Intuition (exploration-exploitation dilemma):

- Find the "best" arm with the minimum time slots and stick to it.
- Select the current "best" arm with high-prob and also explore others.
- Optimism in the face of uncertainty.
- Exploration by "stochasticity".

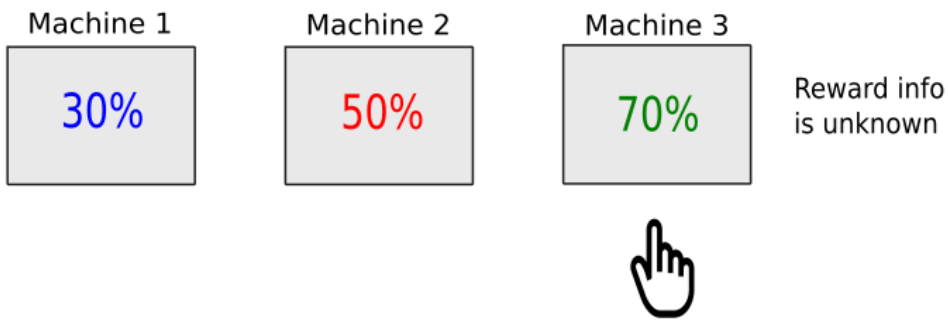
More accurately:

- Estimated reward prob of action a : $Q_t(a)$
- Greedy action: $\hat{a} = \arg\max_a Q_t(a)$
- Exploitation if $a_t = \hat{a}$ and Exploration if $a_t \neq \hat{a}$

Bandits algorithms – Exploration-Then-Exploitation

Exploration-Then-Exploitation (Commit):

- Play each arm N times and compute $Q_{KN}(a)$.
- Play greedy action $\hat{a} = \arg\max_a Q_t(a)$



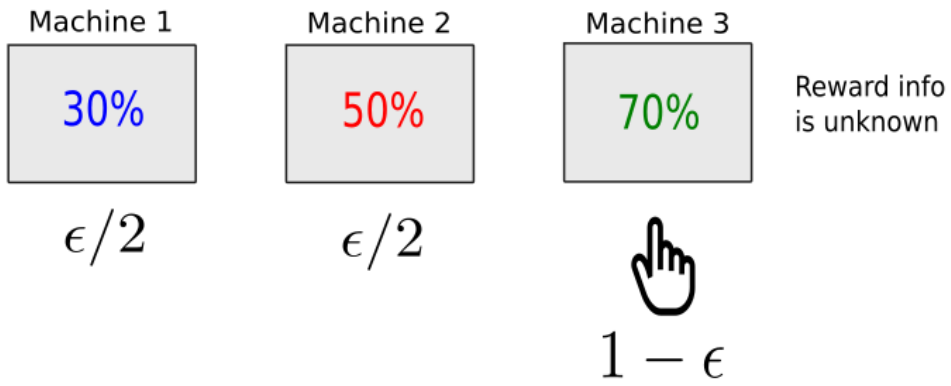
Drawback:

- How to choose N? Large N or small N?
- Greedy action might stuck to sub-optimal arm: $O(T)$

Bandits algorithms – ϵ-Greedy

ϵ-Greedy:

- Play greedy action $\hat{a} = \arg\max_a Q_t(a)$ with probability $1 - \epsilon$.
- Play random action with probability ϵ.



Drawback:

- How to choose ϵ? Large ϵ or small ϵ? Decaying ϵ_t?
- Fixed ϵ: $O(T)$ and Decaying ϵ_t : $o(T)$ but hard to schedule.

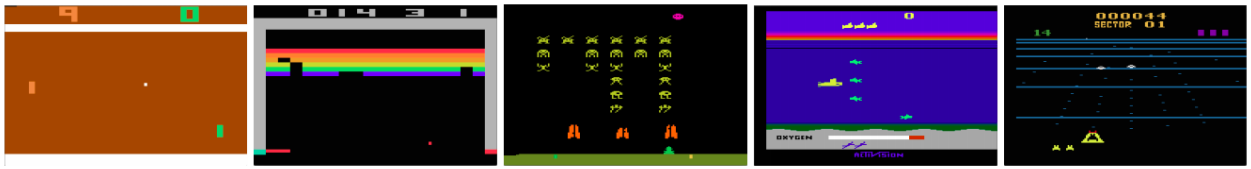
ϵ-Greedy in video games

DeepMind › Research › Playing Atari with Deep Reinforcement Learning

PUBLICATIONS

SHARE

Playing Atari with Deep Reinforcement Learning

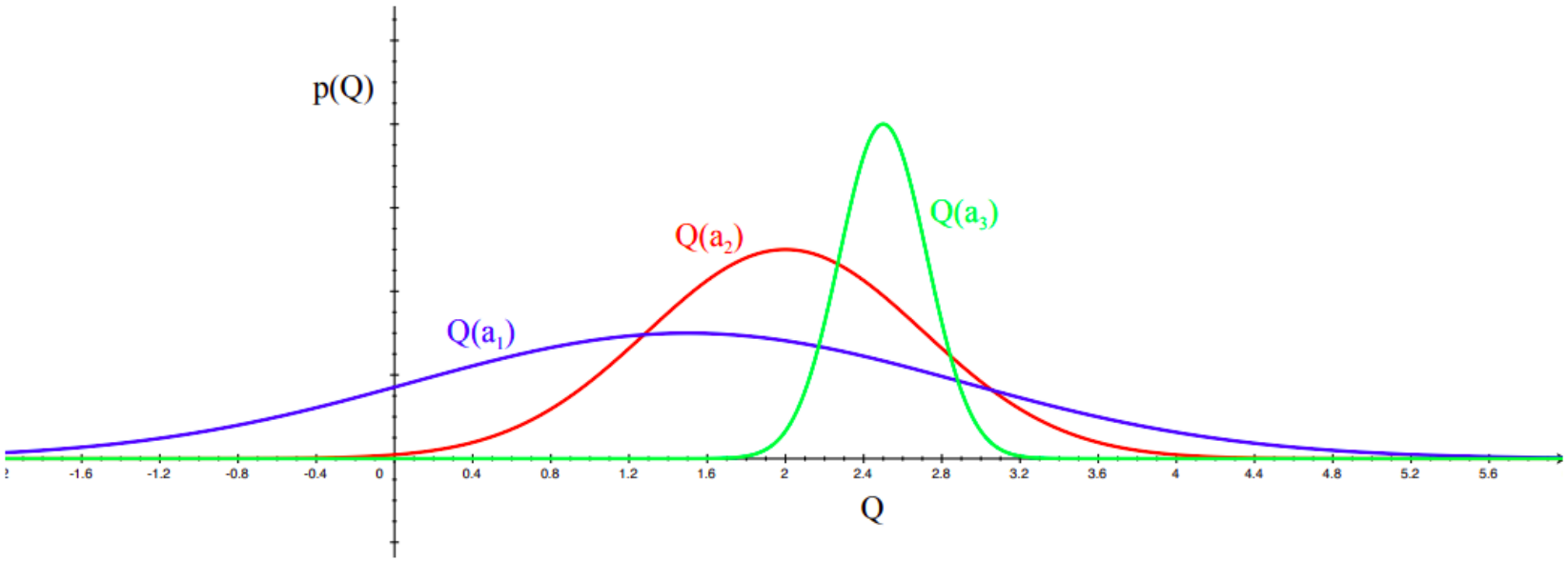


Algorithm 1 Deep Q-learning with Experience Replay

Initialize replay memory $\mathcal{D}$ to capacity N
Initialize action-value function Q with random weights
for episode $= 1, M$ **do**
  Initialise sequence $s_1 = \{x_1\}$ and preprocessed sequenced $\phi_1 = \phi(s_1)$
  for $t = 1, T$ **do**
    With probability ϵ select a random action a_t
    otherwise select $a_t = \max_a Q^*(\phi(s_t), a; \theta)$
    Execute action a_t in emulator and observe reward r_t and image x_{t+1}
    Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$
    Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in $\mathcal{D}$
    Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from $\mathcal{D}$
    Set $y_j = \begin{cases} r_j & \text{for terminal } \phi_{j+1} \\ r_j + \gamma \max_{a'} Q(\phi_{j+1}, a'; \theta) & \text{for non-terminal } \phi_{j+1} \end{cases}$
    Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ according to equation 3
  end for
end for

With probability ϵ select a random action a_t
otherwise select $a_t = \max_a Q^*(\phi(s_t), a; \theta)$

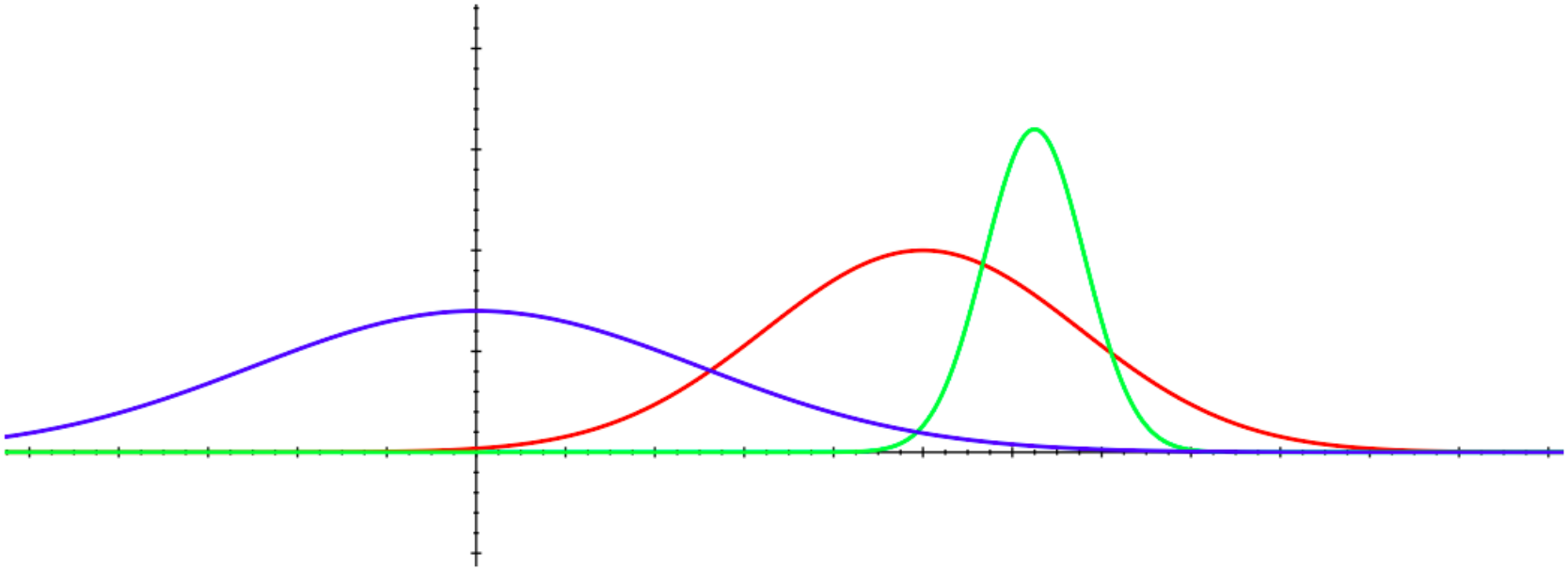
Optimism in the Face of Uncertainty



Optimism in the Face of Uncertainty:

- Which action should we pick?
- The more uncertain we are about an action-value; the more important it is to explore that action
- It has potential to be the best action

Optimism in the Face of Uncertainty



Optimism in the Face of Uncertainty:

- After picking blue action, we are less uncertain about the value
- And more likely to pick another action
- Eventually we could find the best action

Bandits algorithms – Upper Confidence Bound (UCB)

UCB: Optimism in the Face of Uncertainty

– Construct an upper bound $U_t(a)$ for each action a.
– Such that $r_a \leq Q_t(a) + U_t(a)$ with high probability.
– This depends on the number of times $N_t(a)$ action a has been selected
  - Small $N_t(a) \to$ large $U_t(a)$ (estimated value is uncertain)
  - Large $N_t(a) \to$ small $U_t(a)$ (estimated value is accurate)
– Select action maximising Upper Confidence Bound (UCB):

$$a_t = \arg\max_{a \in [K]} \; Q_t(a) + U_t(a)$$

The key to construct adaptive UCB terms $U_t(a)$. To make it rigorous, we detour a bit and study some probability inequalities.

Review of Probability Inequality

Empirical Mean, Variance, and Tail Probability

Let $X, X_1, \cdots, X_s$ be i.i.d. random variables with mean $\mu = \mathbb{E}[X]$ and $\mathbb{V}[X] = \sigma$. Let $\hat{X}_n = \sum_{s=1}^{n} X_s/n$ be the sample mean. Then

$$\mathbb{E}[\hat{X}_n] = \mu \quad \text{and} \quad \mathbb{V}[\hat{X}_n] = \frac{\sigma^2}{n}.$$

and the tail probability w.r.t. the quantity ϵ

$$\mathbb{P}(\hat{X}_n - \mu \geq \epsilon) \quad \text{and} \quad \mathbb{P}(\hat{X}_n - \mu \leq -\epsilon).$$

Markov Inequality and Chebyshev Inequality

For any random variable X and positive $\epsilon > 0$, the following holds:

$$\mathbb{P}(|X| \geq \epsilon) \leq \frac{\mathbb{E}[|X|]}{\epsilon},$$

$$\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{\mathbb{V}[X]}{\epsilon^2}.$$

Review of Probability Inequality

Central Limit Theorem

Let $X_1, \cdots, X_s$ be i.i.d. random variables with mean $\mu = \mathbb{E}[X]$ and $\mathbb{V}[X] = 1$. Let $S_n = \sum_{s=1}^{n}(X_s - \mu)/\sqrt{n}$. Then we have as $n \to \infty$

$$S_n \to Z \sim \mathcal{N}(0, 1).$$

With CLT, we can establish the tail probability as follows

$$\begin{aligned}
\mathbb{P}(\hat{X}_n - \mu \geq \epsilon) =& \mathbb{P}(S_n \geq \sqrt{n}\epsilon) \\
\approx& \mathbb{P}(Z \geq \sqrt{n}\epsilon) \\
=& \int_{\sqrt{n}\epsilon}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\
\leq& \sqrt{\frac{1}{2\pi n\epsilon^2}} e^{-\frac{n\epsilon^2}{2}}.
\end{aligned}$$

Review of Probability Inequality

Subgaussian Random Variables

A random variable X is σ-subgaussian if $\mathbb{E}[X] = 0$ and its moment generating function satisfies

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2 \sigma^2}{2}}, \ \lambda \in \mathcal{R}.$$

Examples of subgaussian random variables

- If X is $\mathcal{N}(0, 1)$, then X is 1-subgaussian.
- If X is zero-mean and $X \in [a, b]$ almost surely, then X is $(b - a)/2$-subgaussian (e.g. Rademacher variables).

Review of Probability Inequality

Lemma (Tail Prob of Subgaussian)

If a random variable X is σ-subgaussian, for any $\epsilon > 0$, its tail probability satisfies

$$\mathbb{P}(X \geq \epsilon) \leq e^{-\frac{\epsilon^2}{2\sigma^2}} \quad \textit{and} \quad \mathbb{P}(X \leq -\epsilon) \leq e^{-\frac{\epsilon^2}{2\sigma^2}}.$$

Review of Probability Inequality

Lemma (Hoeffding's Inequality)

Let $\{X_1, \cdots, X_n\}$ be i.i.d. random variables with X_i being σ_i-subgaussian random variables. Then for any ϵ, it satisfies

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mu) > \epsilon\right) \leq e^{-\frac{\epsilon^2}{2\sum_{i=1}^{n}\sigma_i}},$$

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mu) < -\epsilon\right) \leq e^{-\frac{\epsilon^2}{2\sum_{i=1}^{n}\sigma_i}}.$$

Bandits algorithms – Upper Confidence Bound (UCB)

Apply Hoeffding's inequality, the true mean are in the interval with probability $1 - 2p$ s.t.

$$\mu \in \left[\hat{X}_n - \sqrt{\frac{\log 1/p}{n}}, \hat{X}_n + \sqrt{\frac{\log 1/p}{n}}\right]$$

Let $p = T^{-2}$, we have with probability $1 - 2/T^2$

$$\mu \in \left[\hat{X}_n - \sqrt{\frac{2\log T}{n}}, \hat{X}_n + \sqrt{\frac{2\log T}{n}}\right]$$

UCB bonus term is

$$U_t(a) = \sqrt{\frac{2\log T}{N_t(a)}}.$$

Bandits algorithms – Upper Confidence Bound (UCB)

UCB Algorithm

Construct a confidence interval based on history. For any $a \in [K]$, let

$$\mathsf{UCB}_t(a) = Q_t(a) + \sqrt{\frac{2 \log T}{N_t(a)}}.$$

Choose the action such that

$$a_t = \arg\max_a \ \mathsf{UCB}_t(a).$$

Observe reward X_{t,a_t} and update the sample mean of selected arm

$$Q_{t+1}(a_t) = \frac{Q_t(a_t) \times N_t(a_t) + X_{t,a_t}}{N_t(a) + 1}.$$

Regret Analysis of UCB

Theorem (Problem dependent regret)

Assuming K i.i.d arms and the reward is in $[0,1]$, the regret under UCB algorithm for T rounds will have

$$\mathcal{R}(T) := Tr_{a^*} - \mathbb{E}\left[\sum_{t=1}^{T} X_{t,a_t}\right] \leq O(K \log T/\Delta_{\min}).$$

Sketch of Proof

- Regret decomposition

$$\mathcal{R}(T) = \sum_{a=1}^{K} \Delta_a \mathbb{E}[N_T(a)].$$

- Bounded number of suboptimal arms pulled

$$\mathbb{E}[N_T(a)] = O\left(\frac{\log T}{\Delta_a^2}\right).$$

Regret Analysis of UCB

Regret Analysis of UCB

Theorem (Problem independent regret)

Assuming K i.i.d arms and the reward is in $[0, 1]$, the regret under UCB algorithm for T rounds will have

$$\mathcal{R}(T) := Tr_{a^*} - \mathbb{E}\left[\sum_{t=1}^{T} X_{t,a_t}\right] \leq O(\sqrt{KT \log T}).$$

UCB in AlphGo

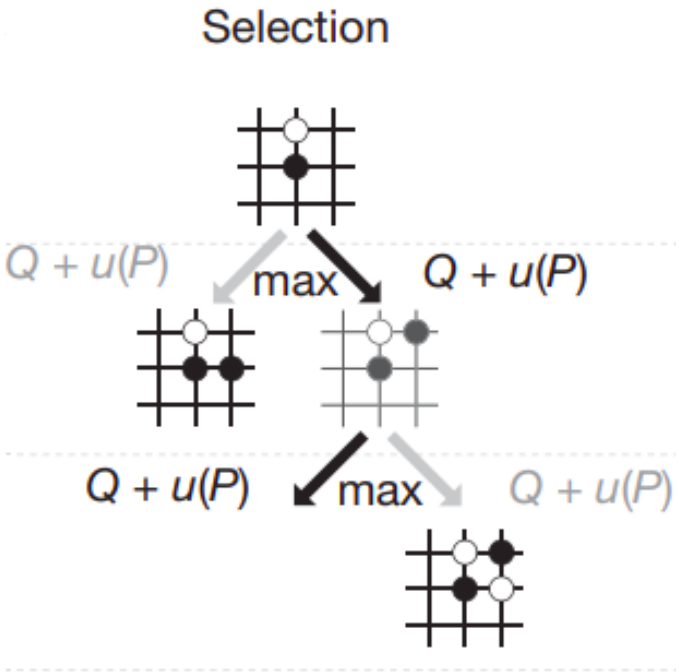
nature

Explore our content ⌄ Journal information ⌄ Subscribe

nature › articles › article

Published: 27 January 2016

Mastering the game of Go with deep neural networks and tree search



Thompson Sampling

Goals in MABs:

- Identify the arm with the largest mean.
- Learn the unknown mean of each arm with high confidence.

A Bayesian learning view in (Bernoulli) bandits:

- For each arm, maintain a belief about its mean $[0, 1]$ and start with the uniform distribution.
- Update our belief to left if observe 0 and to right if we observe 1.
- Choose the arm we believe it is the best arm.

This is exactly Thompson Sampling [Thompson 1933], which comes back mainly due to "An Empirical Evaluation of Thompson Sampling" NIPS, 2011 by O. Chapelle and L. Li.

Thompson Sampling

Thompson sampling in bandits:

- For each arm, maintain a belief about its mean $[0, 1]$ and start with the uniform distribution.
- Update our belief (distribution) according to our feedback $(\mathcal{H})$ with Bayesian law:

$$\mathbb{P}(\mu|\mathcal{H}) \propto \mathbb{P}(\mathcal{H}|\mu)\mathbb{P}(\mu).$$

- Sample $\hat{\mu}$ from $\mathbb{P}(\mu|\mathcal{H})$ and choose the arm with maximum value that is

$$a = \arg\max_{i \in [K]} \hat{\mu}(i)$$

The key step is how we update our belief according to our feedback or history.

Thompson Sampling in Bernoulli Bandits

Thompson Sampling in (Bernoulli) bandits:

- For each arm, start with Beta$(1, 1)$, i.e., uniform distribution.
- Sample $\hat{\mu}_t(i)$ from Beta$(\alpha_t(i), \beta_t(i))$ and choose the best arm that

$$a_t = \arg\max_{i \in [K]} \hat{\mu}_t(i)$$

- Update our belief (distribution) according to the feedback (a_t, X_{t,a_t}):

$$\text{Beta}(\alpha_t(a_t) + X_{t,a_t}, \beta_t(a_t) + 1 - X_{t,a_t}).$$

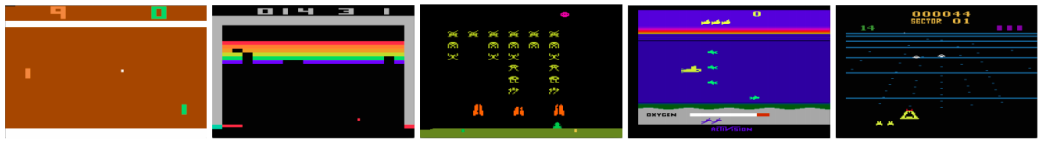
The updating rule is simple with the beta distribution.

Thompson Sampling in Bernoulli Bandits

Why does Thompson sampling work? How does it solve exploitation-exploration dilemma?

- The Bayesian (distributional) view takes into account the level of uncertainty about the means.
- For a less explored arm, the uncertainty is higher and encourages exploration. The uncertainty reduces as the arm is explored enough and tends to exploitation.

Thompson sampling utilizes the "stochasticity" to explore! This idea has been used to reinforcement learning to achieve SOTA in Atari game[1].



[1] Marc G. Bellemare et al. "A Distributional Perspective on Reinforcement Learning". ICML 2017.

Regret Analysis of Thompson Sampling

Theorem (Problem dependent regret)

Assuming K i.i.d arms and the reward is in $[0, 1]$, the regret under TS algorithm for T rounds will have

$$\mathcal{R}(T) = O(K \log T / \Delta_{\min}).$$

Theorem (Problem independent regret)

Assuming K i.i.d arms and the reward is in $[0, 1]$, the regret under TS algorithm for T rounds will have

$$\mathcal{R}(T) = O(\sqrt{KT \log T}).$$

TS algorithm achieves $O(\log(T))$-type instance dependent regret and $O(\sqrt{T})$-type independent regret, which is similar with UCB.

Best-of-Two-Worlds for Stochastic and Adversarial Bandits

Can we design algorithms to achieve optimal regret for stochastic and adversarial bandits (best-of-two-worlds)?

- UCB or TS for stochastic bandits
- EXP3 for adversarial bandits
- Design a detector and apply our classical algorithms?
- .....

Intuitively, we need to design "very adaptive" algorithms to achieve $O(\log T/\Delta_{\min})$ for stochastic bandits and $O(\sqrt{TK\log T})$ for adversarial bandits without knowing which world we are living in.

Best-of-Two-Worlds for Stochastic and Adversarial Bandits

Can we design algorithms to achieve optimal regret for stochastic and adversarial bandits (best-of-two-worlds)?

- UCB or TS for stochastic bandits
- EXP3 for adversarial bandits
- Design a detector and apply our classical algorithms?
- .....

Intuitively, we need to design "very adaptive" algorithms to achieve $O(\log T/\Delta_{\min})$ for stochastic bandits and $O(\sqrt{TK\log T})$ for adversarial bandits without knowing which world we are living in.

A bit surprising, "OMD + proper regularization" can achieve the best-of-two- worlds[2]!!!

[2] J. Zimmert et al. "Tsallis-INF: An Optimal Algorithm for Stochastic and Adversarial Bandits". JMLR 2021.

Online Mirrored Descent for the Best-of-Two-Worlds

Let $\psi(x) = -\sum_{i=1}^{K} \sqrt{x_i}$ in Bregman divergence and $\psi(x)$ is called Tsallis entropy.

Online Mirrored Descent for the Best-of-Two-Worlds:

Initialization: $x_1 = [1/K, \cdots, 1/K]$ and η.
For each day $t = 1, \cdots, T$:

- **Learner:** Sample an arm i from x_t.
- **Environment:** Observe the reward of arm i : $r_t(i)$.
- **Reward Estimator:** $\hat{r}_t(i) = r_t(i)/x_t(i)$ and 0 otherwise.
- **Update:** $x_{t+1} = \arg\min_{\mathcal{K}} \ \langle x, -\hat{r}_t \rangle + \frac{1}{\eta_t} B_\psi(x; x_t)$.

"OMD + Tsallis entropy" can achieve the best-of-two-worlds:

- $O(\sqrt{TK \log K})$ regret for adversarial bandits similar as EXP3.
- $O(\log T/\Delta_{\min})$ regret for stochastic bandits!!!

OMD + Tsallis entropy for Stochastic Bandits

The key to establishing $O(\log T/\Delta_{\min})$ regret for stochastic bandits is called "self-bounding" techniques. OMD + Tsallis entropy achieves the regret bound

$$\sum_{t=1}^{T}\sum_{i\neq i^*}\Delta_i x_{t,i} \leq C\sum_{t=1}^{T}\sum_{i\neq i^*}\sqrt{\frac{x_{t,i}}{t}}.$$

By combining simple yet elegant algebra

$$C\sqrt{\frac{x_{t,i}}{t}} \leq \frac{1}{2}\left(\frac{C^2}{t\Delta_i} + \Delta_i x_{t,i}\right),$$

we have

$$\sum_{t=1}^{T}\sum_{i\neq i^*}\Delta_i x_{t,i} \leq \frac{1}{2}\sum_{t=1}^{T}\sum_{i\neq i^*}\Delta_i x_{t,i} + \frac{1}{2}\sum_{t=1}^{T}\sum_{i\neq i^*}\frac{C^2}{t\Delta_i}.$$

Contextual Bandits

Contextual Bandits generalize the classical MABs by making use of side information (features or structures).

Figure: Yahoo Front Page[3]

User 1 : [Gender: Male, Age: 23, History: NBA, NFL]
User 2 : [Gender: Female, Age: 60, History: Travel, Handicraft]

[3] Lihong Li, et al. "A Contextual-Bandit Approach to Personalized News Article Recommendation." WWW 2010.

Stochastic Contextual Bandits

Stochastic Contextual Bandits:

- Context set $\mathcal{C}$ and K bandits.
- At time slot $t \in [T]$, a user with context $C_t \in \mathcal{C}$ arrives.
- Choose an action $k \in [K]$ and receive the reward:

$$X_t = r(C_t, k) + \eta_t,$$

where $r : \mathcal{C} \times [K] \to \mathbb{R}$ is the reward function and η_t is the noise.

Definition (Regret)

$$\mathcal{R}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \max_{k \in [K]} r(C_t, k) - \sum_{t=1}^{T} X_t\right].$$

Stochastic Contextual Bandits

Naive UCB Algorithm for Contextual Bandits

For each "arm" (context, action), we do the UCB algorithm and the regret is

$$O\left(\sqrt{|\mathcal{C}|KT\log T}\right),$$

where $|\mathcal{C}|$ is the number of contexts (could be huge).

The structure (e.g. Lipschitz):

$$c \text{ and } c' \text{ are "close" implies } r(c,\cdot) \text{ and } r(c',\cdot) \text{ are "close".}$$

Utilize the structure:

$$r(c,k) = \langle \theta_*, \phi(c,k) \rangle + \eta,$$

where $\phi(c,k)$ is a feature related to the context c and k.

Stochastic Linear Bandits

Stochastic Linear Bandits:

- Context set $\mathcal{C}$ and K bandits.
- At time slot t, a user with context $C_t \in \mathcal{C}$ arrives.
- Choose an action $A_t \in \mathcal{A}_t \subset R^d$ and receive the reward:

$$X_t = \langle \theta_*, A_t \rangle + \eta_t,$$

where η_t is 1-subgaussian given the history.

Definition (Regret)

$$\mathcal{R}(T) := \mathbb{E}[\hat{\mathcal{R}}(T)] := \mathbb{E}\left[\sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} \langle \theta_*, a \rangle - \sum_{t=1}^{T} X_t\right].$$

UCB for Linear Bandits

LinUCB

Construct a confidence set $\mathcal{B}_t$ based on $(A_1, X_1, \cdots, A_{t-1}, X_{t-1})$ such that $\theta_* \in \mathcal{B}_t$ with a high-probability. For any action $a \in \mathbb{R}^d$, let

$$\mathsf{UCB}_t(a) = \max_{\theta \in \mathcal{B}_t} \langle \theta, a \rangle.$$

Choose the action such that

$$A_t = \mathsf{argmax}_{a \in \mathcal{A}_t} \ \mathsf{UCB}_t(a).$$

The two conflicting properties of $\mathcal{B}_t$:

- $\mathcal{B}_t$ should contain θ_* with high probability.
- $\mathcal{B}_t$ should be as small as possible.

LinUCB

(Regularised) Least-squares estimator

Given the history $(A_1, X_1, \cdots, A_t, X_t)$ and a regularization factor $\lambda > 0$:

$$\text{argmax}_{\theta \in \mathbb{R}^d} \left(\sum_{s=1}^{t} (X_s - \langle \theta, A_s \rangle)^2 + \lambda ||\theta||^2 \right).$$

Least-squares estimator is

$$\hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} A_s X_s \text{ with } V_t = \lambda I + \sum_{s=1}^{t} A_s A_s^{\dagger}.$$

Confident Set

Given the estimator $\hat{\theta}_{t-1}$ and carefully choose β_t:

$$\mathcal{B}_t = \left\{ \theta \in \mathbb{R}^d : ||\theta - \hat{\theta}_{t-1}||^2_{V_{t-1}} \leq \beta_t \right\}.$$

Regret Analysis of LinUCB

Assumptions

- $\max_t \sup_{a,b \in \mathcal{A}_t} \langle \theta_*, a - b \rangle \leq 1.$
- $||a||_2 \leq 1, a \in \mathcal{A}_t, \forall t.$
- $1 \leq \beta_1 \leq \beta_2 \leq \cdots \beta_n.$
- There exits a $\delta \in (0,1)$ such that $\theta_* \in \mathcal{B}_t$ holds for all t with probability $1 - \delta$.

Theorem

Under Assumptions above, with prob $1 - \delta$, *the regret of LinUCB satisfies*

$$\hat{\mathcal{R}}(T) \leq \sqrt{8T \log\left(\frac{det(V_T)}{det(V_0)}\right)} \leq \sqrt{8dT\beta_T \log\left(\frac{d\lambda + T}{d\lambda}\right)},$$

where $\sqrt{\beta_T} = \sqrt{\lambda}||\theta_*|| + \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log(\frac{d\lambda+T}{d\lambda})}.$

Regret Analysis of LinUCB

Corollary

Under Assumptions above, let $\delta = 1/T$ the regret of LinUCB satisfies

$$\mathcal{R}(T) \leq Cd\sqrt{T}\log(T),$$

where $C > 0$ is a large universal constant.

Proof.

By choosing $\delta = 1/T$, we have

$$\begin{aligned}\mathcal{R}(T) \leq & \delta \cdot T + \mathbb{E}[\hat{\mathcal{R}}(T) \mid \text{Theorem}] \\ \leq & 1 + \sqrt{8dT\beta_T \log\left(\frac{d\lambda + T}{d\lambda}\right)} \\ \leq & 1 + C'd\sqrt{T}\log(T)\end{aligned}$$

Regret Analysis of LinUCB

Proof.

Let $A_t^* = \text{argmax}_{a \in \mathcal{A}_t} \langle \theta_*, a \rangle$, and the regret at time slot t is

$$\mathcal{R}_t = \langle \theta_*, A_t^* - A_t \rangle.$$

Let $\tilde{\theta}_t$ be the parameter such that $\langle \theta, A_t \rangle = \text{UCB}_t(A_t)$, then we have

$$\langle \theta_*, A_t^* \rangle \leq \text{UCB}_t(A_t^*) \leq \text{UCB}_t(A_t) = \langle \tilde{\theta}_t, A_t \rangle.$$

Further, we have

$$\begin{aligned}
\langle \theta_*, A_t^* - A_t \rangle \leq & \langle \tilde{\theta}_t - \theta_*, A_t \rangle \leq ||A_t||_{V_{t-1}^{-1}} ||\tilde{\theta}_t - \theta_*||_{V_{t-1}} \\
\leq & 2\sqrt{\beta_t} ||A_t||_{V_{t-1}^{-1}} \leq 2 \wedge 2\sqrt{\beta_t} ||A_t||_{V_{t-1}^{-1}} \\
\leq & 2\sqrt{\beta_t} \left(1 \wedge ||A_t||_{V_{t-1}^{-1}}\right)
\end{aligned}$$

Regret Analysis of LinUCB

Proof.

To sum up, we have

$$\hat{\mathcal{R}}(T) = \sum_{t=1}^{T} \mathcal{R}_t \leq \sqrt{T \sum_{t=1}^{T} \mathcal{R}_t^2} \leq 2\sqrt{T\beta_T \sum_{t=1}^{T} \left(1 \wedge ||A_t||^2_{V_{t-1}^{-1}}\right)}.$$

Lemma

Let V_0 be positive definite and $a_1, a_2, \cdots, a_T \in \mathbb{R}^d$ with $||a_t|| \leq L, \forall t$, $V_t = V_0 + \sum_{s \leq t} a_s a_s^\dagger$. Then

$$\sum_{t=1}^{T} \left(1 \wedge ||a_t||^2_{V_{t-1}^{-1}}\right) \leq 2 \log \left(\frac{det(V_T)}{det(V_0)}\right) \leq 2d \log \left(\frac{Tr(V_0) + T}{d \cdot det(V_0)^{1/d}}\right).$$

Confidence Bounds

Theorem

Let $\delta \in (0,1)$. The following inequality holds for all t with probability $1-\delta$ that

$$||\hat{\theta}_t - \theta_*||_{V_t} < \sqrt{\lambda}||\theta_*||_2 + \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{det(V_t)}{\lambda^d}\right)}$$

Proof.

Recall $\hat{\theta}_t = V_t^{-1}\sum_{s=1}^t A_s X_s$ with $V_t = \lambda I + \sum_{s=1}^t A_s A_s^\dagger$ and let $S_t = \sum_{s=1}^t \eta_s A_s$.

$$\begin{aligned}
||\hat{\theta}_t - \theta_*||_{V_t} =& ||V_t^{-1}S_t + (V_t^{-1}V_t(0) - I)\theta_*||_{V_t} \\
\leq& ||S_t||_{V_t^{-1}} + (\theta_*^\dagger(V_t^{-1}V_t(0) - I)V_t(V_t^{-1}V_t(0) - I)\theta_*)^{1/2} \\
=& ||S_t||_{V_t^{-1}} + \lambda^{1/2}(\theta_*^T(I - V_t^{-1}V_t(0))\theta_*)^{1/2} \\
\leq& ||S_t||_{V_t^{-1}} + \sqrt{\lambda}||\theta_*||_2.
\end{aligned}$$

□

Confidence Bounds

Theorem

For all $\lambda > 0$ *and* $\delta \in (0, 1)$,

$$\mathbb{P}\left(\textit{exists } t : ||S_t||^2_{V_t^{-1}} \geq 2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\textit{det}(V_t)}{\lambda^d}\right)\right) \leq \delta.$$

Proof.

Let $M_t(x) = \exp(\langle x, S_t\rangle - \frac{1}{2}||x||^2_{V_t(0)})$, we have

$$\begin{aligned}\mathbb{P}\left(\frac{1}{2}||S_t||^2_{V_t(0)^{-1}} \geq \log(1/\delta)\right) =& \mathbb{P}\left(\max_{x\in\mathbb{R}^d} M_t(x) \geq 1/\delta\right)\\ \approx& \mathbb{P}\left(\bar{M}_t(x) \geq 1/\delta\right)\\ \leq& \delta\mathbb{E}[\bar{M}_0(x)] \leq \delta.\end{aligned}$$

where $\max_x M_t(x) \approx \bar{M}_t := \int_x M_t(x)dh(x)$ (Laplace's approx). □

Confidence Bounds

Proof.

Let $H = \lambda I$ and $h \sim \mathcal{N}(0, H^{-1})$. Compute Laplace's approximation that

$$\begin{aligned}
&\int_x M_t(x) dh(x) \\
=&\frac{1}{\sqrt{(2\pi)^d \det(H^{-1})}} \int \exp(\langle x, S_t \rangle - \frac{1}{2}||x||^2_{V_t(0)} - \frac{1}{2}||x||^2_H) dx \\
=&\frac{1}{\sqrt{(2\pi)^d \det(H^{-1})}} \int \exp\left(\frac{1}{2}||S_t||^2_{V_t^{-1}} - \frac{1}{2}||x - V_t^{-1} S_t||^2_{V_t}\right) dx \\
=&\left(\frac{\det(H)}{\det(V_t)}\right)^{1/2} \exp\left(\frac{1}{2}||S_t||^2_{V_t^{-1}}\right).
\end{aligned}$$

Thompson Sampling in Stochastic Linear Bandits

Thompson Sampling in Stochastic Linear bandits:

- Start with $\mathcal{N}(0, \beta I)$, i.e., Gaussian distribution.
- Sample an estimator $\tilde{\theta}_t$ from $\mathcal{N}(\hat{\theta}_t, \beta V_t^{-1})$, and choose the best action that

$$A_t = \arg\max_{a \in \mathcal{A}_t} \langle \tilde{\theta}_t, a \rangle$$

- Update our belief (distribution) according to the feedback

$$\hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} A_s X_s \text{ with } V_t = \lambda I + \sum_{s=1}^{t} A_s A_s^{\dagger}.$$

The updating rule similar with LinUCB except the step to obtain $\tilde{\theta}_t$ by using "sampling".

Regret Analysis

Theorem

Under Assumptions as in LinUCB, with prob $1-\delta$, the regret of LinTS with $\beta = O(\sqrt{d\log(T/\delta)})$ satisfies

$$\hat{\mathcal{R}}(T) = O\left(d\sqrt{T}(\log T + \sqrt{\log T \log(1/\delta)}\right).$$

Corollary

Under Assumptions as in LinUCB, let $\delta = 1/T$ the regret of LinTS satisfies

$$\mathcal{R}(T) = O(d\sqrt{T}\log T).$$

The regret bound is similar as that in LinUCB.

Conclusion

- Basic model for stochastic bandits problems.
- Intuition to deal with Exploitation-Exploration dilemma (e.g. optimism in the face of uncertainty).
- Four MAB algorithms: 1) Exploration-Then-Exploitation; 2) ϵ-Greedy; 3) Upper Confidence Bound; 4) Thompson Sampling.
- LinUCB and LinTS algorithms for stochastic linear bandits.
- Applications: ϵ-Greedy in video game, UCB in AlphaGo and Thompson Sampling in recommendation system.