Bandit and RL Reading Notes by Xuanfei

Xuanfei Ren[*] Pan Xu[†]

Abstract

Here are some notes on the papers from my study. I think the only way to remember is to write something while reading.

Contents

[*]University of Science and Technology of China; e-mail: xuanfeir@gmail.com
[†]Duke University; e-mail: pan.xu@duke.edu

1 Finished paper

1. Lifting the Information ratio: [NOPS22] Paper link (See my notes)
2. Contextual IDS: [HLQ22] Paper link (See my notes)
3. Linear bandits (OFUL): [AYPS11] Paper link
4. Bayesian Contextual TS (Information method): [GRGOS23] Paper link
5. LinUCB: [CLRS11] Paper link
6. Generalized linear: [KZS+20]Paper link and [LLZ17] (Leave out the proof) Paper link
7. Perturbed-History Exploration: [KSGB19]Paper link
8. RandUCB: [VMDK19] Paper link
9. Langevin Monte Carlo: [XZM+22] Paper link
10. Frequentist regret of LinTS: (Leave out the proof)[HB20] Paper link
11. ϵ-Exploring TS: (Reading group)Paper link
12. Batched TS: (Leave out the proof)[KO21]Paper link and [KMS21] Paper link
13. Ensemble Sampling: (Leave out the proof)[QWLVR22] Paper link
14. Classical IDS: [RVR14] Paper link
15. Information analysis of TS [RVR16] Paper link
16. Frequentist IDS (Heteroscedastic Noise): [KK18] Paper link
17. Asymptotically Optimal frequentist IDS: [KLVS21] Paper link [Kir21] Thesis link
18. GP-UCB for kernelized bandit: [WWR23] Paper link [SKKS09] Paper link
19. End of Optimism: [LS17] Paper link
20. Langevin Batched TS (bandits and RL) (Leave out the proof):[KKMM23] Paper link
21. Heavy tail (Leave out the proof): [JSS21] Paper link [BCBL13] Paper link
22. DETC: [JXXG21] Paper link
23. Best arm identification (Leave out the proof): [GK16] Paper link and [JP20] Paper link, a pdf and Chapter 33 in *bandit algorithms* for reference.
24. Batched linear bandits: [HZZ+20] Paper link

2 Paper to read

- TS in MAB (technical): [CL15] Paper link
- Variance-Adaptive TS:[SK23] Paper link
- IDS for RL: [CBK+23] Paper link
- Theoretical results for greedy methods: Paper link
- Tabular MDP: Paper link

3 Basic knowledge

3.1 Linear algebra

Proposition 3.1 (Matrix determinant lemma)**.** *Suppose A is an invertible square matrix and u, v are column vectors. Then the matrix determinant lemma states that*

$$\det\left(\mathbf{A}+\mathbf{u}\mathbf{v}^{\mathsf{T}}\right)=\left(1+\mathbf{v}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{u}\right)\det\left(\mathbf{A}\right).$$

Proposition 3.2 (Sherman–Morrison formula)**.** *Suppose $A \in \mathbb{R}^{n\times n}$ is an invertible square matrix and $u, v \in \mathbb{R}^n$ are column vectors. Then $A + uv^T$ is invertible iff $1 + v^T A^{-1} u \neq 0$. In this case,*

$$\left(A+uv^{T}\right)^{-1}=A^{-1}-\frac{A^{-1}uv^{T}A^{-1}}{1+v^{T}A^{-1}u}.$$

Here, uv^T is the outer product of two vectors u and v. The general form shown here is the one published by Bartlett.

3.2 Basic probability

I collect some basic probability results here which I'm not so familiar with.

3.2.1 Bayes

Lemma 3.3 (posterior for sufficient statistic)**.** *$X \sim f(x \mid \theta)$, if $T(X)$ is a sufficient statistic for θ, then*

$$\pi(\theta \mid x) = \tilde{\pi}(\theta \mid t) \tag{3.1}$$

1. $X \sim N(\theta, \sigma_S^2)$, in which σ_S^2 is known. θ has a prior $N(\mu_P, \sigma_p^2)$, then the posterior distribution of θ is:

$$\theta \sim N(\frac{\frac{x}{\sigma_S^2}+\frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_S^2}+\frac{1}{\sigma_P^2}}, (\frac{1}{\sigma_S^2}+\frac{1}{\sigma_P^2})^{-1}), \tag{3.2}$$

which means

$$\begin{aligned}\mu &= \frac{\sigma_P^2}{\sigma_S^2+\sigma_P^2}x+\frac{\sigma_S^2}{\sigma_S^2+\sigma_P^2}\mu_P,\\ \sigma^2 &= \frac{\sigma_S^2\sigma_P^2}{\sigma_S^2+\sigma_P^2}.\end{aligned} \tag{3.3}$$

Proof. $f(x \mid \theta) = \frac{1}{\sqrt{2\pi\sigma_S^2}} \exp\{-\frac{1}{2\sigma_S^2}(x-\theta)^2\}$

$$\begin{aligned}\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{f(x)} &\propto f(x \mid \theta)\pi(\theta)\\ &= \exp\{-\frac{1}{2\sigma_S^2}(x-\theta)^2-\frac{1}{2\sigma_P^2}(\theta-\mu_P)^2\}\\ &\propto \exp\{-\frac{1}{2}(\frac{1}{\sigma_S^2}+\frac{1}{\sigma_P^2})(\theta^2-2\frac{\frac{x}{\sigma_S^2}+\frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_S^2}+\frac{1}{\sigma_P^2}}\theta)\}.\end{aligned} \tag{3.4}$$

□

2. Using the assumption abouve, the posterior of θ after observing $x_1, \cdots, x_n$ is a.s. Gaussian with mean μ_n and variance σ_n^2 given by:

$$\begin{aligned} \mu_n &= \frac{\sigma_S^2/n}{\sigma_S^2/n + \sigma_P^2}\mu_P + \frac{\sigma_P^2}{\sigma_S^2/n + \sigma_P^2}\bar{x}, \\ \sigma_n^2 &= \frac{\sigma_P^2\sigma_S^2}{n\sigma_P^2 + \sigma_S^2}. \end{aligned} \tag{3.5}$$

Proof. Note that $\bar{X} \sim N(\theta, \sigma_S^2/n)$, then subsititute x above with $\bar{x}$.

Or you can calculate the law of $\boldsymbol{x} = (x_1, \cdots, x_n)$ and get the same result. □

3. $X \sim B(n, \theta)$, θ has a prior $Be(a, b)$, then the posterior of θ is $Be(x + a, n - x + b)$.

3.2.2 Beta distribution

Definition 3.4. *We say $X \sim Beta(\alpha, \beta)$, if*

$$f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

Property 1. *$X \sim Beta(\alpha, \beta)$, then* $\mathbb{E}[X] = \dfrac{\alpha}{\alpha+\beta}$, $\mathbb{E}[X^2] = \dfrac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}$,
$\mathrm{Var}[X] = \dfrac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$.

Property 2. *If $Y \sim B(n, p)$, where the prior is $p \sim Beta(\alpha, \beta)$. Then if $Y = k$ is number of successes, the posterior of p is $Beta(\alpha + k, \beta + n - k)$.*

Proof. $X \sim B(n, \theta)$, it has a distribution:

$$f(x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}. \tag{3.6}$$

θ satisfies:

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}. \tag{3.7}$$

then the posterior of θ can be calculated by:

$$\begin{aligned} \pi(\theta \mid x) &\propto f(x \mid \theta)\pi(\theta) \\ &= \theta^{(x+a)-1}(1-\theta)^{(n-x+b)-1}. \end{aligned} \tag{3.8}$$

□

Property 3. $Beta(1, 1) \overset{d}{=} U[0, 1]$.

3.3 Concentration inequalities

This note is a collation of relevant knowledge about measure concentration. Mainly used for our bandit and reinforcement learning study. The most basic part is based on *High-Dimensional Statistics A Non-Asymptotic Viewpoint (2019, Cambridge University Press),* and the rest of the inequalities come from the Internet.

Definition 3.5 (sub-Gaussian)**.** *A random variable X is σ sub-Gaussian if*

$$\mathbb{E}[\exp(\lambda X - \lambda \mathbb{E}[X])] \leq \exp(\sigma^2 \lambda^2 / 2),$$

for all $\lambda \in \mathbb{R}$.

Proposition 3.6 (Bounded random variables)**.** *If $X \in [a, b]$, then X is $\frac{(b-a)^2}{4}$-sub Gaussian r.v., i.e. $\mathbb{E}e^{\lambda X} \leq e^{\frac{(b-a)^2\lambda^2}{8}}$.*

Theorem 3.7 (Equivalent characterizations of sub-Gaussian variables)**.** *Given any zero-mean random variable X, the following properties are equivalent:*

1. *There is a constant $\sigma > 0$ such that*

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2\sigma^2}{2}} \text{ for all } \lambda \in \mathbb{R}.$$

2. *An equivalent definition*

$$\mathbb{P}(|X| \geq t) \leq 2\exp\left(-t^2/C^2\right).$$

3. *There is a constant $c > 0$ and Gaussian random variable $Z \sim N(0, \tau^2)$ such that*

$$\mathbb{P}[|X| > s] \leq c\mathbb{P}[|Z| > s] \text{ for all } s \geq 0.$$

4. *There is a constant $\theta \geq 0$ such that*

$$\mathbb{E}[X^{2k}] \leq \frac{(2k)!}{2^k k!}\theta^{2k} \text{ for all } k = 1, 2, \ldots$$

5. *There is a constant $\sigma > 0$ such that*

$$\mathbb{E}\left[e^{\frac{\lambda X^2}{2\sigma^2}} \leq \frac{1}{\sqrt{1-\lambda}}\right] \text{ for all } \lambda \in [0, 1).$$

Theorem 3.8 (Equivalent characterizations of sub-exponential variables)**.** *For a zeromean random variable X, the following statements are equivalent:*

1. *There are non-negative numbers (v, α) such that*

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{v^2\lambda^2}{2}} \text{for all } |\lambda| < \frac{1}{\alpha}$$

2. *There is a positive number $c_0 > 0$ such that $\mathbb{E}[e^{\lambda X}] < \infty$ for all $|\lambda| \leq c_0$.*

3. There are constants $c_1, c_2 > 0$ such that

$$\mathbb{P}[|X| \geq t] \leq c_1 e^{-c_2 t^2} \text{ for all } t > 0$$

4. The quantity $\gamma := \sup_{k \geq 2} \left[\frac{\mathbb{E}[X^k]}{k!}\right]^{1/k}$ is finite.

First, let's see the simplest concentration inequality: Hoeffding Inequality.

Consider random variables $X_1, X_2, \ldots$ Assume they are mutually independent, but not necessary identially distributed. Let $\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}$ to be the average of the first n random variables, and let $\mu_n = \mathbb{E}[\bar{X}_n]$ be its expectation. We have:

Theorem 3.9 (Hoeffding Inequality)**.**

$$\mathbb{P}\{|\bar{X}_n - \mu_n| \leq \sqrt{\alpha \log T / n}\} \geq 1 - 2T^{-2\alpha}, \alpha > 0. \tag{3.9}$$

Here T is a fixed parameter.

3.3.1 Concentration inequality

I will directly list the common knowledge, easy to consult. The specific learning process and proof details are written in later subsubsections.

Markov's inequality(X:non-negative and a finite mean):

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}, \forall t > 0 \tag{3.10}$$

Chebyshev's inequality($Y = (X - \mu)^2$):

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{var(X)}{t^2}, \forall t > 0 \tag{3.11}$$

Extensions of Markov's inequality (X has a central moment of order k):

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\mathbb{E}|X - \mu|^k}{t^k}, \forall t > 0 \tag{3.12}$$

(Chernoff's bounds)Suppose there is some constant $b > 0$ such that the moment generating function $\phi(\lambda) = \mathbb{E}[e^{\lambda(X-\mu)}]$ exist for all $\lambda \leq |b|$:

$$\log \mathbb{P}(X - \mu \geq t) \leq \inf_{\lambda \in [0,b]} (\log \mathbb{E}[e^{\lambda(X-\mu)}] - \lambda t) \tag{3.13}$$

Proposition 3.10. *Let $Z_1, \ldots, Z_n$ be independent Bernoulli variables where for every i,$\mathbb{P}[Z_i = 1] = p_i$, let $p = \Sigma_{k=1}^n p_i$, $Z = \Sigma_{k=1}^n Z_i$. Then, for any $\delta > 0$,*

$$\begin{aligned} \mathbb{P}[Z > (1+\delta)p] &\leq e^{-h(\delta)p} \quad \leq e^{-p\frac{\delta^2}{2+2\delta/3}}, \\ \mathbb{P}[Z < (1-\delta)p] &\leq e^{-h(-\delta)p} \quad \leq e^{-p\frac{\delta^2}{2+2\delta/3}} \end{aligned} \tag{3.14}$$

where $h(a) = (1+a)\log(1+a) - a$.

(Gaussian tail bounds) Suppose X is any $N(\mu, \sigma^2)$ random variable, then:

$$\mathbb{P}(X \geq \mu + t) \leq e^{-\frac{t^2}{2\sigma^2}}, \forall t > 0 \tag{3.15}$$

Definition 3.11 (Sub-Gaussian). *A random variable X with mean μ is sub-Gaussian if there is a positive number σ such that*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\sigma^2\lambda^2}{2}}, \forall \lambda \in \mathbb{R} \tag{3.16}$$

Remark 3.12. *If $X \in [a, b]$, then it is sub-Gaussian with parameter $\sigma = \frac{b-a}{2}$.*

The sub-Gaussian variable satisfies the concentration inequality (3.15) and

$$\mathbb{P}(|X - \mu| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}, \forall t \in \mathbb{R} \tag{3.17}$$

Theorem 3.13 (Hoeffding bound). *Suppose that the variables $X_i, i = 1, \ldots, n$ are independent and X_i has mean μ_i and sub-Gaussian parameter σ_i, then for all $t > 0$, we have*

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq \exp\{-\frac{t^2}{2\Sigma_{i=1}^{n}\sigma_i^2}\} \tag{3.18}$$

If $X_i \in [a, b]$, then we obtain the bound

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] &\leq \exp\{-\frac{2t^2}{n(b-a)^2}\}, \\
\mathbb{P}\left[\sum_{i=1}^{n}|X_i - \mu_i| \geq t\right] &\leq 2\exp\{-\frac{2t^2}{n(b-a)^2}\}
\end{aligned} \tag{3.19}$$

Definition 3.14 (sub-exponential). *A random variable X with mean $\mu = \mathbb{E}[X]$ is sub-exponential if there are non-negative parameters (v,α) such that*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{v^2\lambda^2}{2}} \text{ for all } |\lambda| < \frac{1}{\alpha} \tag{3.20}$$

Note: Any sub-Gaussian variable is also sub-exponential. However, the converse statement is not true.

(Sub-exponential tail bound) Suppose that X is sub-exponential with parameters (v, α), then

$$\mathbb{P}[X - \mu \geq t] \leq \begin{cases} e^{-\frac{t^2}{2v^2}} & \text{if } 0 \leq t \leq \frac{v^2}{\alpha} \\ e^{-\frac{t}{2\alpha}} & \text{if } t > \frac{v^2}{\alpha}. \end{cases} \tag{3.21}$$

Definition 3.15 (Bernstein's condition). *Given a random variable X with mean μ and variance σ^2, we say that Bernstein's condition with parameter b holds if*

$$|\mathbb{E}[(X - \mu)^k]| \leq \frac{1}{2}k!\sigma^2 b^{k-2} \text{for } k = 2, 3, 4, \ldots. \tag{3.22}$$

Proposition 3.16. *When X satisfies the Bernstein condition, then it is sub-exponential with parameters $(\sqrt{2}\sigma, 2b)$.*

(Bernstein-type bound) For any random variable satisfying the Bernstein's condition, we have

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}} \text{ for all } |\lambda| < \frac{1}{b},$$

and, moreover, the concentration inequality

$$\mathbb{P}[|X-\mu| \geq t] \leq 2e^{-\frac{t^2}{2(\sigma^2+bt)}} \text{ for all } t > 0. \tag{3.23}$$

There are another version of Bennet's and Bernstein's Inequalities.

Lemma 3.17 (Bennet's inequality). *Let $Z_1, \ldots, Z_m$ be independent random variables with zero mean, and assume that $Z_i \leq 1$ with probability 1. Let*

$$\sigma^2 \geq \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[Z_i^2].$$

Then for all $\epsilon > 0$,

$$\mathbb{P}[\sum_{i=1}^{m} Z_i > \epsilon] \leq e^{-m\sigma^2 h(\frac{\epsilon}{m\sigma^2})}, \tag{3.24}$$

where h is the same definition as above.

Theorem 3.18 (Bernstein's inequality). *Same as above, assume $|Z_i| < M$ a.s., then*

$$\mathbb{P}[\sum_{i=1}^{m} Z_i > t] \leq \exp\{-\frac{t^2/2}{\Sigma Z_j^2 + Mt/3}\} \tag{3.25}$$

Corollary 3.19 (The sum of sub-exponential variables). *Consider an independent sequence $\{X_k\}_{k=1}^n$ of random variables, such that X_k has mean μ_k, and is sub-exponential with parameters (v_k, α_k), then*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}(X_k - \mu_k) \geq t\right] \leq \begin{cases} e^{-\frac{nt^2}{2(v_*^2/n)}} & \text{for } 0 \leq t \leq \frac{v_*^2}{n\alpha_*}, \\ e^{-\frac{nt}{2a_*}} & \text{for } t > \frac{v_*^2}{n\alpha_*}, \end{cases} \tag{3.26}$$

Where $\alpha_* := \max_{k=1,\ldots,n} \alpha_k$ and $v_* := \sqrt{\sum_{k=1}^{n} v_k^2}$.

Theorem 3.20 (One-sided Bernstein inequality). *If $X \leq b$ a.s., then*

$$\begin{aligned} \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] &\leq e^{\frac{\lambda^2\mathbb{E}X^2}{1-\lambda b/3}} \forall \lambda \in [0, b/3) \\ \mathbb{P}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X) \geq t] &\leq \exp\{\frac{-nt^2}{2(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i^2) + \frac{bt}{3}}\}. \end{aligned} \tag{3.27}$$

Theorem 3.21 (Slud's Inequality). *Let X be a (m, p) binomial variable and assume that $p = (1-\epsilon)/2$. Then,*

$$\mathbb{P}[X \geq m/2] \geq \frac{1}{2}(1 - \sqrt{1 - \exp\{-m\epsilon^2/(1-\epsilon^2)\}}). \tag{3.28}$$

Definition 3.22 (Martingale, Martingale difference sequence). *Given a sequence $\{Y_k\}_{k=1}^{\infty}$ of random variables adapted to a filtration $\{\mathcal{F}_k\}_{k=1}^{\infty}$, the pair $\{(Y_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ is a martingale, if for all $k \geq 1$,*

$$\mathbb{E}[|Y_k|] < \infty \text{ and } \mathbb{E}[Y_{k+1}|\mathcal{F}_k] = Y_k.$$

A martingale difference sequence is an adapted sequence $\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ such that for all $k \geq 1$,

$$\mathbb{E}[|D_k|] < \infty \text{ and } \mathbb{E}[D_{k+1}|\mathcal{F}_k] = 0.$$

Theorem 3.23 (A general Bernstein-type bound). *Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ be a martingale difference sequence, and suppose that $\mathbb{E}[e^{\lambda D_k}|\mathcal{F}_{k-1}] \leq e^{\lambda^2 v_k^2/2}$ almost surely for any $|\lambda| < 1/\alpha_k$. Then the following hold:*

1. *The sum $\Sigma_{k=1}^n D_k$ is sub-exponential with parameters $(\sqrt{\Sigma_{k=1}^n v_k^2}, \alpha_*)$ where $\alpha_* := \max_{k=1,\ldots,n} \alpha_k$.*
2. *The sum satisfies the concentration inequality:*

$$\mathbb{P}\left[|\sum_{k=1}^n D_k| \geq t\right] \leq \begin{cases} 2e^{-\frac{t^2}{2\Sigma_{k=1}^n v_k^2}} & \text{if } 0 \leq t \leq \frac{\Sigma_{k=1}^n v_k^2}{\alpha_*} \\ 2e^{-\frac{t}{2\alpha_*}} & \text{if } t > \frac{\Sigma_{k=1}^n v_k^2}{\alpha_*}. \end{cases} \tag{3.29}$$

Corollary 3.24. *Let X_i be a sequence of i.i.d. random variables such that $|X_i - E[X_i]| \leq b$. Then, it holds that*

$$\mathbb{P}[\sum_{i=1}^n (X_i - \mathbb{E}X) \geq t] \leq \exp\{\frac{-t^2}{2n\sigma^2 + \frac{2}{3}bt}\}. \tag{3.30}$$

Corollary 3.25 (Azuma-Hoeffding). *Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ be a martingale difference sequence for which there are constants $\{(a_k, b_k)\}_{k=1}^n$ such that $D_k \in [a_k, b_k]$ almost surely for all $k = 1, \ldots, n$. Then, for all $t \geq 0$,*

$$\mathbb{P}\left[|\sum_{k=1}^n D_k| \geq t\right] \leq 2e^{-\frac{2t^2}{\Sigma_{k=1}^n (b_k - a_k)^2}}. \tag{3.31}$$

Theorem 3.26 (One-side Azuma-Hoeffding). *Let $X_i \in \mathcal{F}_i$ and $\mathcal{F}_{k-1} \subseteq \mathcal{F}_k$. If it holds that*

$$\mathbb{E}[X_i - \mathbb{E}X_i|\mathcal{F}_{i-1}] = 0, X_i \leq \mathbb{E}X_i + R_i,$$

then it holds that

$$\mathbb{P}[\sum_{i=1}^n (X_i - \mathbb{E}X) \geq t] \leq 2\exp\{-\frac{2t^2}{\Sigma_{i=1}^n R_i^2}\}.$$

Theorem 3.27 (One-side Azuma-Bernstein). *Let $X_i \in \mathcal{F}_i$ and $\mathcal{F}_{k-1} \subseteq \mathcal{F}_k$. If it holds that*

$$\mathbb{E}[X_i - \mathbb{E}X_i|\mathcal{F}_{i-1}] = 0, X_i \leq \mathbb{E}X_i + R_i, \mathbb{V}[X_i|\mathcal{F}_{i-1}] \leq \sigma_i^2,$$

then it holds that

$$\mathbb{P}[\sum_{i=1}^n (X_i - \mathbb{E}X) \geq t] \leq 2\exp\{-\frac{t^2}{2\Sigma_{i=1}^n \sigma_i^2 + 2/3Rt}\}.$$

Corollary 3.28 (Bounded differences inequality)**.** *Suppose that f satisfies the bounded difference property with parameters $(L_1, \ldots, L_n)$ and that the random vector $X = (X_1, \ldots, X_n)$ has independent components. Then*

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{2t^2}{\Sigma_{k=1}^n L_k^2}} \text{ for all } t \geq 0, \tag{3.32}$$

where the bounded difference property means if you change only the k th component, the value of the function changes at most L_k.

Theorem 3.29 (Lipchitz bound)**.** *Let $X = (X_1, \ldots, X_n)$ be a vector of i.i.d standard Gaussian variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipchitz with respect to the Euclidean norm. Then the variale $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most L, and hence*

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{t^2}{2L^2}} \text{ for all } t \geq 0 \tag{3.33}$$

Using the corollary above we can derive the χ^2-concentration:

$$\mathbb{P}[Y \geq n(1+t)] \leq e^{-\frac{nt^2}{18}} \text{ for all } t \in [0, 3], \tag{3.34}$$

where $Y := \Sigma_{k=1}^n Z_k^2$ follows a χ^2-distribution with n degrees of freedom.

Proposition 3.30. *Let $Z \sim \chi_k^2$, then for all $\epsilon > 0$ we have*

$$\mathbb{P}[Z \leq (1-\epsilon)k] \leq e^{-\frac{\epsilon k^2}{6}}$$

, and for all $\epsilon \in (0, 3)$ we have

$$\mathbb{P}[Z \geq (1+\epsilon)k] \leq e^{-\frac{\epsilon k^2}{6}}.$$

Finally, for all $\epsilon \in (0, 3)$

$$\mathbb{P}[(1-\epsilon)k \leq Z \leq (1+\epsilon)k] \geq 1 - 2e^{-\frac{\epsilon k^2}{6}}.$$

3.4 Bandit algorithms

3.4.1 Simple bandit algorithms

In this part I tend summarize some simple bandit algorithms and their regret bounds.

We consider the basic model with IID rewards, called stochastic bandit. An algorithm has K possible actions to choose from, and there are T rounds, for some known K and T. The mean reward of arm a is $\mu(a) := \mathbb{E}[\mathcal{D}_a]$, in which $\mathcal{D}$ is the reward distribution. The best mean reward is denoted $\mu^* = \max a \in \mathcal{A}\mu(a)$, the difference $\delta(a) := \mu^*(a) - \mu(a)$ describes how bad arm a is compared to μ^* Then we can define our main goal regret by:

Definition 3.31 (Regret)**.**

$$R(T) = \mu^* T - \Sigma_{t=1}^{T} \mu(a_t). \tag{3.35}$$

Then I can begin to summarize some simple algorithms and analysis their regret.

Algorithm 1 Explore-first algorithm

1: Exploration phase: try each arm N times;
2: Select the arm $\hat{a}$ with the highest average reward (break ties arbitrarily);
3: Exploitation phase: play arm $\hat{a}$ in all remaining rounds.

Theorem 3.32. *Explore-first algorithm achieves regret* $\mathbb{E}[R(T)] \leq T^{2/3} \times O(K \log T)^{1/3}$.

Algorithm 2 Epsilon-greedy algorithm

1: **for** each round t=1,2,... **do**
2: Toss a coin with success probability ϵ_t;
3: **if** success **then**
4: explore:choose an arm uniformly at random
5: **else**
6: exploit:choose the arm with the highest average reward so far
7: **end if**
8: **end for**

Theorem 3.33. *Epsilon-greedy algorithm with exploration probability* $\epsilon_t = t^{-1/3}(K \log t)^{1/3}$ *achieves regret bound* $\mathbb{E}[R(t)] \leq t^{2/3} O(K \log t)^{1/3}$ *for each round t.*

Remark 3.34. *Explore-first and Epsilon-greedy do not adapt their exploration schedule to the history of the observed rewards. The set of all exploration rounds and the choice of arms therein is fixed before the round 1.*

Definition 3.35 (Confidence interval)**.** *For each arm a and round t,*

$$\begin{aligned} r_t(a) &= \sqrt{2 \log(T)/n_t(a)} \quad \textit{(confidence radius)} \\ UCB_t(a) &= \bar{\mu}_t(a) + r_t(a) \quad \textit{(upper confidence bound)} \\ LCB_t(a) &= \bar{\mu}_t(a) - r_t(a) \quad \textit{(lower confidence bound).} \end{aligned} \tag{3.36}$$

Using concentration inequality we have *confidence interval* $[LCB_t(a), UCB_t(a)]$ and *confidence radius* $r_t(a)$.

Then we have some adaptive exploration algorithms:

Algorithm 3 successive algorithm for two arms

1: Alternate two arms until $UCB_t(a) < LCB_t(a')$ after some even round t;
2: Abandon arm a, and use arm a' forever since.

For multi-armed bandit, we have:

Algorithm 4 Successive Elimination algorithm

1: All arms are initially designed as *active*
2: **loop**
3: play each active arm once
4: deactive all arms a such that, letting t be the current round, $UCB_t(a) < LCB_t(a')$ for some other arm a' deactivation rule
5: **end loop**

Theorem 3.36. *Successive Elimination algorithm achieves regret*

$$\mathbb{E}[R(t)] = O(\sqrt{Kt \log T}) \text{ for all rounds } t \leq T. \tag{3.37}$$

Theorem 3.37. *Successive Elimination algorithms achieves regret*

$$\mathbb{E}[R(t)] \leq O(\log T) \left[\sum_{\text{arms } a \text{ with } \mu(a)<\mu^*(a)} \frac{1}{\mu^*(a) - \mu(a)} \right] \tag{3.38}$$

Note that we can only use UCB_t to determination which arm is better, this is because an arm a can have a large $UCB_t(a)$ for two reasons (or combination thereof): because the average $\bar{mu}_t(a)$ is large, and/or because the confidence radius $r_t(a)$ is large, in which case this arm has not been explored much. So we have the algorithm:

Algorithm 5 Algorithm UCB1

1: Try each arm once
2: **for** each round t=1,. . . ,T **do**
3: pich arm some a which maximizes $UCB_t(a)$.
4: **end for**

Theorem 3.38. *Algorithm UCB1 satisfies regret bounds in 3.37 and 3.38.*

Theorem 3.39 (lower bound)**.** *Fix time horizon T and the number of arms K. For any bandit algorithm, there exists a problem instance such that $\mathbb{E}[R(T)] \geq \Omega(\sqrt{KT})$.*

Consider a simple algorithm for Bayesian bandits, called Thompson Sampling. For each round t and arm a, the algorithm computes the posterior probability that a is the best arm, and samples a with this probability.

Algorithm 6 Thompson Sampling

for each round t=1,2,... **do**
  Observe $H_{t-1} = H$, for some feasible (t-1)-history H;
  Draw arm a_t independently from distribution $p_t(\cdot|H)$, where $p_t(a|H) := \mathbb{P}[a^* = a|H_{t-1} = H]$ for each arm a.
end for

Algorithm 7 Thompson Sampling: equivalent version

for each round t=1,2,... **do**
  Observe $H_{t-1} = H$, for some feasible (t-1)-history H;
  Sample mean reward vector μ_t from the posterior distribution $\mathbb{P}_H$;
  Choose the best arm $\tilde{a}_t$ according to μ_t.
end for

And if we have independent priors, the distribution of μ can be easily calculated by using $\mathbb{P}^a_H$ only.

Then let us analyze Bayesian regret of Thompson Sampling:

Theorem 3.40. *Bayesian Regret of Thompson Sampling is* $BR(T) = O(KT \log T)$.

This is the core theorem of the TS algorithm, and its proof is very subtle, so I want to state it in detail.

First, we can recap the definition of the confidence interval 3.36 defined before. THen we say the key lemma below hold for a more general notion of the confidence bounds, and clearly hold for 3.36. Actually, $U(a, H_t)$ and $L(a, H_t)$ can be arbitrary functions of the arm a and the t-history H_t. There are two properties we want these functions to have, for some $\gamma > 0$ to be specified later:

$$\begin{aligned} \mathbb{E}\big[[U(a, H_t) - \mu(a)]^-\big] &\le \frac{\gamma}{TK} \quad \text{for all arm } a \text{ and rounds } t, \\ \mathbb{E}\big[[\mu(a) - L(a, H_t)]^-\big] &\le \frac{\gamma}{TK} \quad \text{for all arm } a \text{ and rounds } t. \end{aligned} \tag{3.39}$$

The confidence radius can be defined as $r(a, H_t) = \frac{U(a,H_t) - L(a,H_t)}{2}$.

Lemma 3.41. *Assume we have lower and upper bound functions that satisfies properties above, for some parameter* $\gamma > 0$. *Then Bayesian Regret of Thompson Sampling can be bound as follows:*

$$BR(T) \le 2\gamma + 2\Sigma_{t=1}^{T} \mathbb{E}[r(a_t, H_T)].$$

Proof. Fix round t. As two algorithms are equivalent, we have:

$$\mathbb{P}[a_t = a|H_t = H] = \mathbb{P}[a^* = a|H_t = H] \text{for each arm } a. \tag{3.40}$$

It follows that

$$\mathbb{E}[U(a^*, H)|H_t = H] = \mathbb{E}[U(a_t, H)|H_t = H]. \tag{3.41}$$

The Bayesian Regret suffered in round t is

$$\begin{aligned} BR_t &= \mathbb{E}[\mu(a^* - \mu(a_t))] \\ &= \mathop{\mathbb{E}}_{H\sim H_t}[\mathbb{E}[\mu(a^*) - \mu(a_t)|H_t = H]] \\ &= \mathop{\mathbb{E}}_{H\sim H_t}[\mathbb{E}[U(a_t, H) - \mu(a_t) + \mu(a^*) - U(a^*, H)|H_t = H]] \\ &= \mathbb{E}[U(a_t, H_t) - \mu(a_t)] + \mathbb{E}[\mu(a^*) - U(a^*, H_t)]. \end{aligned} \tag{3.42}$$

We will use properties above to bound both summands. Note that we cannot immediately use these properties because they assume a fixed arm a, whereas both a_t and a^* are random varibales.

$$\begin{aligned} \mathbb{E}[\mu(a^*) - U(a^*, H_t)] &\leq \mathbb{E}[(\mu(a^*) - U(a^*, H_t))^+] \\ &\leq \mathbb{E}[\sum_{armsa} (\mu(a^*) - U(a^*, H_t))^+] \\ &= \Sigma_{armsa}\mathbb{E}[(U(a, H_t) - \mu(a))^-] \\ &\leq K\frac{\gamma}{KT} = \frac{\gamma}{T}. \end{aligned} \tag{3.43}$$

$$\mathbb{E}[U(a_t, H_t) - \mu(a_t)] = \mathbb{E}[2r(a_t, H_t) + L((a_t, H_t) - \mu(a_t))] \quad = \mathbb{E}[2r(a_t, H_t)] + \mathbb{E}[L((a_t, H_t) - \mu(a_t))] \tag{3.44}$$

$$\begin{aligned} \mathbb{E}[L((a_t, H_t) - \mu(a_t))] &\leq \mathbb{E}[(L((a_t, H_t) - \mu(a_t)))^+] \\ &\leq \mathop{\mathbb{E}}_{armsa}[(L((a_t, H_t) - \mu(a_t)))^+] \\ &= \mathop{\mathbb{E}}_{armsa}[(\mu(a_t) - L((a_t, H_t)))^-] \\ &\leq K\frac{\gamma}{KT} = \frac{\gamma}{T}. \end{aligned} \tag{3.45}$$

Thus, $BR_t(T) \leq 2\frac{\gamma}{T} + 2\mathbb{E}[r(a_t, H_t)]$, the lemma follows by summing up over all rounds t. □

Now we can proof the main theorem.

Proof. By lemma,

$$BR(T) \leq O(\sqrt{\log T}) \sum_{t=1}^{T} \mathbb{E}\left[\frac{1}{\sqrt{n_t(a_t)}}\right]$$

Moreover,

$$\begin{aligned} \sum_{t=1}^{T} \frac{1}{\sqrt{n_t(a_t)}} &= \sum_{armsa} \sum_{roundst:a_t=a} \frac{1}{\sqrt{n_t(a)}} \\ &= \sum_{armsa} \sum_{j=1}^{n_{T+1}(a)} \frac{1}{\sqrt{j}} = \sum_{armsa} O(\sqrt{n(a)}). \end{aligned}$$

It follows that

$$BR(T) \leq O(\sqrt{\log T}) \sum_{armsa} \sqrt{n(a)} \leq O(\sqrt{logT}) \sqrt{K \sum_{armsa} n(a)} = O(\sqrt{KT \log T})$$

. □

3.4.2 Lipschitz Bandit

In a special case, arms correspond to point in the interval $X = [0, 1]$, and expected rewards obey a Lipschitz condition:

$$|\mu(x) - \mu(y)| \leq L|x - y| \text{for any two arms } x, y \in X. \tag{3.46}$$

We can see this Lipschitz condition describes "similar" arms have similar expected rewards. As this bandit has infinity many arms, a simple solution to this is use finite arms to approximate the whole model. And such Lipschitz condition can guarantee these finite arms have enough information.

Theorem 3.42. *Consider continuum-armed bandits with Lipschitz constant L and time horizon T . Uniform discretization with algorithm ALG satisfying Lipschitz and discretization step $\epsilon = (TL^2/\log T)^{-1/3}$ attains*

$$\mathbb{E}[R(T)] \leq L^{1/3}T^{2/3}(1 + c_{ALG})(\log T)^{1/3} \tag{3.47}$$

The main take-away here is the $\tilde{O}(L^{1/3}T^2/3)$ regret rate. The explicit constant and logarithmic dependence are less important.

Actually, uniform discretization is optimal in the worst case: we have an $\Omega(L^{1/3}T^{2/3})$ lower bound on regret.

Theorem 3.43. *Let* ALG be any algorithm for continuum-armed bandits with time horizon T and Lipschitz constant L. There exists a problem instance $L = L(x^*, \epsilon)$, for some $x^* \in [0, 1]$ and $\epsilon > 0$, such that

$$\mathbb{E}[R(T)|L] \geq (L^{1/3}T^{2/3}). \tag{3.48}$$

In a more general case, the Lipschitz condition can be stated as:

$$|\mu(x) - \mu(y)| \leq \mathcal{D}(x, y) \text{for any arms } x, y \tag{3.49}$$

where $\mathcal{D}$is a metric.

Definition 3.44. *A subset $S \in X$ is called an $\epsilon - mash, \epsilon > 0$, if every point $x \in X$ is within distance ϵ from S, in the sense that $\mathcal{D}(x, y) \leq \epsilon$ for some $y \in S$.*

Then we have regret bound for this more general case:

Theorem 3.45. *Consider Lipschitz bandits with time horizon T. Optimzing over the choice of an $\epsilon - mesh$, uniform discretization with algorithm* ALG attains regret

$$\mathbb{E}[R(T)] \leq \inf_{\epsilon > 0, \epsilon - mesh\ S} \epsilon T + c_{\text{ALG}} \sqrt{|S|T \log T} \tag{3.50}$$

In the $d - dimension$ Eulidean space with l_p metric, we can get a explicit form. And in more general cases, this form can be attained through the definition of "covering dimension". But all of these are about optimizing over the choice of all subset. We leave out these unimportant details to get into our main algorithm quickly.

Now we can describe the algorithm. It contains two parts: activation step and selection step. We consider the confidence ball defined as $B_t(x) = \{y \in X : \mathcal{D}(x, y) \leq r_t(x)\}$. In the activation step, we find an arm which is not in any confidence ball of all the active arms: we can see in intuition that if

some arm lies very close to some active arm, we don't need to explore it because of the Lipschitz condition.

Then in the selection step, we select an arm which is active. The idea is just like UCB1. If an arm x is active at time t, we define

$$index_t(x) = \bar{\mu}_t(x) + 2r_t(x). \tag{3.51}$$

The selcetion rule is very simple: play an active arm with the largest index.

Algorithm 8 Zooming algorithm for adaptive discretization.

```
Initialize: set of active arms S ← ∅
for each round t=1,2,... do
    if some arm y is not covered by the confidence balls of active arms then
        pick any such arm y and "activate" it: S ← S ∪ {y}.
    end if
    Play an active arm x with the largest index_t(x).
end for
```

Definition 3.46. *The smallest number of subsets in an $\epsilon-$covering is called the covering number and denoted $N_\epsilon(X)$.*

For any instance of Lipschitz MAB, the zooming dimension with multiplier $c > 0$ is

$$\inf_{d\geq 0}\{N_{r/3}(X) \leq cr^{-d}\} \tag{3.52}$$

Theorem 3.47. *Consider Lipschitz bandits with time horizon T . Assume that realized rewards take values on a finite set. For any given problem instance and any $c > 0$, the zooming algorithm attains regret*

$$\mathbb{E}[R(T)] \leq O(T^{(d+1)/(d+2)}(c\log T)^{1/(d+2)}), \text{where } d \text{ is the zooming dimension with multiplier } c. \tag{3.53}$$

3.4.3 Adversarial Bandits

To catch up quickly, let's run through this subsection.

First, we need to have a basic knowledge about adversarial cost. It may be influenced by algorithms and mean to "fool" the algorithm. The main set up of this part is at each step, we can choose one arm, suffer the cost and view the cost of every arm. Now we can desctibe two main algorithms and alalysis their effectiveness.

Algorithm 9 Weighted Majority Algorithm

Parameter: $\epsilon \in [0, 1]$
for each round t **do**
  Make predictions using weighted majority vote based on ω.
  for each expert i **do**
    if the i-th expert' s prediction is correct **then**
      ω_i stays the same
    else
      $\omega_i = \omega_i(1 - \epsilon)$
    end if
  end for
end for

Theorem 3.48. *The number of mistakes made by WMA with parameter $\epsilon \in [0, 1]$ is at most*

$$\frac{2}{1-\epsilon} cost^* + \frac{2}{\epsilon} \ln K$$

Algorithm 10 Hedge algorithm for online learning with experts

Initialize the weights as $\omega_1(a) = 1$ for each arm a.
for each round t **do**
  Let $p_t(a) = \frac{\omega_t(a)}{\Sigma_{a'=1}^{K} \omega_t(a')}$.
  Sample an arm a_t from distribution $p_t(\cdot)$.
  Observe cost $c_t(a)$ for each arm a.
  For each arm a, update its weight $\omega_{t+1}(a) = \omega_t(a)(1 - \epsilon)^{c_t(a)}$
end for

Below we analyze Hedge, and prove $O(\sqrt{T \log K})$ bound on expected regret, the best possible for regret.

3.4.4 Contextual Bandits

The problem set of this so-called contextual bandit is every time before we make a choice among arms a_t, algorithm could observe a "context" x_t. As usual, reward $r_t \in [0, 1]$ is realized.

For small number of context, we can just apply algorithms we have developed before. Here is an easy algorithm which describe this process.

Algorithm 11 Contextual bandit algorithm for a small number of contexts

Initialization: For each context x, create an instance ALG_x of algorithm ALG
for each round t **do**
  invoke algorithm ALG_x with $x = x_t$
  "play" action a_t chosen by ALG_x, return reward r_t to ALG_x.
end for

Theorem 3.49. *Algorithm above has regret $\mathbb{E}[R(T)] = O(\sqrt{KT|\mathcal{X}| \ln T})$.*

To handle contextual bandit with a large $|\mathcal{X}|$, we either assume some structure such as Lipschitz condition or the setting of linear contextual bandits, or change the objective.

3.5 Information theory

Definition 3.50 (Entropy)**.** *The entropy of a random variable is the average level of "information", "surprise", or "uncertainty" inherent to the variable's possible outcomes. Given a discrete random variable X, which takes values in the alphabet $\mathcal{X}$ and is distributed according to $p : \mathcal{X} \to [0, 1] p : \mathcal{X} \to [0, 1]$:*

$$\mathrm{H}(X) := -\sum_{x \in \mathcal{X}} p(x) \log p(x) = \mathbb{E}[-\log p(X)],$$

Definition 3.51 (Conditional entropy)**.**

$$H(X|Y) = \mathbb{E}_Y \left[-\sum_{x \in \mathcal{X}} \mathbb{P}(X = x|Y) \log \mathbb{P}(X = x|Y) \right].$$

K-L divergence describes distance between two distributions. It is asymmetric.

Definition 3.52 (Kullback–Leibler divergence)**.** *For discrete probability distributions P and Q defined on the same sample space, $\mathcal{X}$, the relative entropy from Q to P is defined to be*

$$D_{KL}(P \parallel Q) = \int \log \frac{dP}{dQ} dP = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right).$$

Fact 3.53 (Gibbs' inequality)**.**

$$D(P\|Q) \geq 0$$

with equality if and only if $P = Q$ P-a.s.

Property 4 (Chain rule of KL-divergence)**.**

$$D_{KL}\left[p(x, y)|q(x, y)\right] = D_{KL}\left[p(x)|q(x)\right] + D_{KL}\left[p(y|x)|q(y|x)\right].$$

Example 3.54. *For two Bernoulli variable $X_i \sim B(p_i), i = 1, 2$, KL-divergence between them is*

$$D_{KL}(X_1\|X_2) = p_1 \log(\frac{p_1}{p_2}) + (1 - p_1) \log(\frac{1 - p_1}{1 - p_2}).$$

3.5.1 Mutual information

Definition 3.55. *The mutual information is defined as*

$$I(X; Y) = D_{KL}(P_{(X,Y)}\|P_X P_Y).$$

Remark 3.56. *$I(X; Y) \geq 0$ with equality if and only if X and Y are independent.*

It measures how much knowing one of these variables reduces uncertainty about the other. That's how much Y explains the entropy of X.

Property 5. *Mutual information has the following properties:*

- *$I(X; Y) \geq 0$;*

- $I(X;Y) = I(Y;X)$;
- $I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y)$;
- ***(KL-divergence form of mutual information)*** $I(X;Y) = \mathbb{E}_Y[D_{KL}(p_{X|Y}\|p_X)]$, *in some case this can be interpreted by the difference between the prior and the posterior;*
- $I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = \mathbb{E}_Z[D_{KL}(P((X,Y)|Z)\|P(X|Z)P(Y|Z))]$;
- *Chain rule:* $I(X;Y,Z) = I(X;Z) + I(X;Y|Z)$;
- *Data processing inequality: Let three random variables form the Markov chain* $X \to Y \to Z$, *then*

$$I(X;Y) \geq I(X;Z).$$

Proof.

$$\begin{aligned}
\mathrm{I}(X;Y) &= \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p_{(X,Y)}(x,y)\log\left(\frac{p_{(X,Y)}(x,y)}{p_X(x)\,p_Y(y)}\right) \\
&= \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p_{X|Y=y}(x)p_Y(y)\log\frac{p_{X|Y=y}(x)p_Y(y)}{p_X(x)p_Y(y)} \\
&= \sum_{y\in\mathcal{Y}} p_Y(y)\sum_{x\in\mathcal{X}} p_{X|Y=y}(x)\log\frac{p_{X|Y=y}(x)}{p_X(x)} \\
&= \sum_{y\in\mathcal{Y}} p_Y(y)\; D_{\mathrm{KL}}\big(p_{X|Y=y}\parallel p_X\big) \\
&= \mathbb{E}_Y\left[D_{\mathrm{KL}}\big(p_{X|Y}\parallel p_X\big)\right].
\end{aligned}$$

□

Fact 3.57. *If Z is jointly independent of X and Y, then $I(X;Y|Z) = I(X;Y)$.*

Several generalizations of mutual information to more than two random variables have been proposed. I will introduce it when it is useful.

Theorem 3.58 (Pinsker's inequality)**.** *Suppose $P \leq Q$, then*

$$\sqrt{\frac{1}{2}D_{KL}(P\|Q)} \geq \|P-Q\|_{TV} \overset{def}{=} \frac{1}{2}\int_\Omega \left|\frac{dP}{d\mu} - \frac{dQ}{d\mu}\right| d\mu = \sup_A |P(A) - Q(A)|.$$

Property 6 (Application of Pinsker's inequality)**.** *For any distribution P and Q such that P is absolutely continuous w.r.t. Q, any random variable $X : \Omega \to \mathcal{X}$ and any $g : \mathcal{X} \to \mathbb{R}$ such that* $\sup g - \inf g \leq 1$

$$\mathbb{E}_P[g(X)] - E_Q[g(X)] \leq \sqrt{\frac{1}{2}D_{KL}(P\|Q)}.$$

It also holds for sub-Gaussian variables, where we need to bound it with $2\sigma^2$ instead of $\frac{1}{2}$.

This property shows

$$\mathbb{E}[R|A=a] - \mathbb{E}[R] \leq \sqrt{\frac{1}{2}D_{KL}(P_{R|A}\|P_R)}$$

And the proof of last property follows from the variational form of KL-Divergence.

Fact 3.59 (Donsker–Varadhan inequality)**.** *Fix two probability distributions P and Q such that P is absolutely continuous with respect to Q. Then*

$$D_{KL}(P\|Q) = \sup_X \{\mathbb{E}_P[X] - \log \mathbb{E}_Q[e^X]\}.$$

Fact 3.60. *For any matrix $M \in \mathbb{R}^{k\times k}$,*

$$Trace(M) \leq \sqrt{Rank(M)}\|M\|_F.$$

Proof. By the Cauchy-Schwartz inequality, for any vector $x \in \mathbb{R}^n$, $\Sigma_i x_i \leq \sqrt{n}\|x\|_2$. This is just an analogous result for matrices. □

3.6 Dynamic Programming

I read this lecture notes and keep notes to help me memorize.

Basic idea(version 1): What we want to do is take our problem and somehow break it down into a reasonable number of subproblems (where "reasonable" might be something like n^2) in such a way that we can use optimal solutions to the smaller subproblems to give us optimal solutions to the larger ones. Unlike divide-and-conquer (as in mergesort or quicksort) it is OK if our subproblems overlap, so long as there are not too many of them.

Example 3.61. *Longest Common Subsequence.*

Basic idea (version 2): Suppose you have a recursive algorithm for some problem that gives you a really bad recurrence like $T(n) = 2T(n-1) + n$. However, suppose that many of the subproblems you reach as you go down the recursion tree are the same. Then you can hope to get a big savings if you store your computations so that you only compute each different subproblem once. You can store these solutions in an array or hash table. This view of Dynamic Programming is often called memoizing.

Example 3.62 (The Knapsack Problem)**.** *In the* ***knapsack problem*** *we are given a set of n items, where each item i is specified by a size s_i and a value v_i. We are also given a size bound S (the size of our knapsack).*

The goal is to find the subset of items of maximum total value such that sum of their sizes is at most S (they all fit into the knapsack).

Solution Consider whether the n-th item would be chosen. ■

Example 3.63. *Matrix product parenthesization.*

Two properties of problems which can be solved using Dynamic Programming:

1. Optimal solution involves solving a subproblem;
2. There should be only a polynomial number of different subproblems.

3.7 Reproducing kernel Hilbert space

Let X be an arbitrary set and H a Hilbert space of real-valued functions on X, equipped with pointwise addition and pointwise scalar multiplication. The evaluation functional over the Hilbert space of functions H is a linear functional that evaluates each function at a point x,

$$L_x : f \mapsto f(x) \; \forall f \in H.$$

We say that H is a reproducing kernel Hilbert space if, for all x in X, L_x is continuous at every f in H or, equivalently, if L_x is a bounded operator on H, i.e. there exists some $M_x > 0$such that

$$|L_x(f)| := |f(x)| \leq M_x \|f\|_H \qquad \forall f \in H.$$

Although $M_x < \infty$ is assumed for all $x \in X$, it might still be the case that $\sup_x M_x = \infty$.

While this property is the weakest condition that ensures both the existence of an inner product and the evaluation of every function in H at every point in the domain, it does not lend itself to easy application in practice. A more intuitive definition of the RKHS can be obtained by observing that this property guarantees that the evaluation functional can be represented by taking the inner product of f with a function K_x in H. This function is the so-called reproducing kernel for the Hilbert space H from which the RKHS takes its name. More formally, the Riesz representation theorem implies that for all x in X there exists a unique element K_x of H with the reproducing property,

$$f(x) = L_x(f) = \langle f, \; K_x \rangle_H \quad \forall f \in H.$$

Since K_x is itself a function defined on X with values in the field $\mathbb{R}$ (or $\mathbb{C}$ in the case of complex Hilbert spaces) and as K_x is in H we have that

$$K_x(y) = L_y(K_x) = \langle K_x, \; K_y \rangle_H,$$

where $K_y \in H$ is the element in H associated to L_y.

This allows us to define the reproducing kernel of H as a function $K : X \times X \to \mathbb{R}$ by

$$K(x, y) = \langle K_x, \; K_y \rangle_H.$$

From this definition it is easy to see that $K : X \times X \to \mathbb{R}$ (or $\mathbb{C}$ in the complex case) is both symmetric (resp. conjugate symmetric) and positive definite, i.e.

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) = \sum_{i=1}^{n} c_i \left\langle K_{x_i}, \sum_{j=1}^{n} c_j K_{x_j} \right\rangle_H = \left\langle \sum_{i=1}^{n} c_i K_{x_i}, \sum_{j=1}^{n} c_j K_{x_j} \right\rangle_H = \left\| \sum_{i=1}^{n} c_i K_{x_i} \right\|_H^2 \geq 0$$

for every $n \in \mathbb{N}, x_1, \ldots, x_n \in X,$ and $c_1, \ldots, c_n \in \mathbb{R}$. The Moore–Aronszajn theorem is a sort of converse to this: if a function K satisfies these conditions then there is a Hilbert space of functions on X for which it is a reproducing kernel.

4 RL reading notes

4.1 Markov decision process

In a given MDP $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$, the agent interacts with the environment according to the following protocol: the agent starts at some state $s_0 \sim \mu$; at each time step $t = 0, 1, 2, \ldots$, the agent takes an action at $a_t \in \mathcal{A}$, obtains the immediate reward $r_t = r(s_t, a_t)$, and observes the next state s_{t+1} sampled according to $s_{t+1} \sim P(\cdot|s_t, a_t)$. The interaction record at time t,

$$\tau_t = (s_0, a_0, r_0, \ldots, s_t, a_t, r_t),$$

is called a *trajectory*, which include the observed state at time t.

A *policy* is a mapping from a trajectory to an action.

A *value function* is the discoundtd sum of future rewards

$$V_M^\pi(s) = \mathbb{E}\bigg[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|\pi, s_0 = s\bigg].$$

The *action-value (or Q-value) function* $Q_M^\pi : \mathcal{S} \times \mathcal{A} \leftarrow \mathbb{R}$ is defined as

$$Q_M^\pi(s, a) = \mathbb{E}\bigg[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|\pi, s_0 = s, a_0 = a\bigg].$$

Since $r(s, a)$ is bounded between 0 and 1, the value functions are bounded by $1/(1-\gamma)$.

The goal is:

$$\max_\pi V_M^\pi(s).$$

Lemma 4.1 (**Bellman Consistency Equations for Stationary Policies**). *Suppose that π is a stationary policy. Then V^π and Q^π satisfy the following Bellman consistency equations: for all $s \in \mathcal{S}, a \in \mathcal{A}$,*

$$V^\pi(s) = Q^\pi(s, \pi(s))$$
$$Q^\pi(s, a) = r(s, a) + \gamma \mathbb{E}_{a \sim \pi(\cdot|s), s' \sim P(\cdot|s,a)}[V^\pi(s')].$$

Lemma 4.2 (**Induced distribution**). $[(1-\gamma)(I - \gamma P^\pi)^{-1}]_{(s,a),(s',a')} = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}^\pi(s_t = s', a_t = a'|s_0 = s, a_0 = a)$ *is a induced distribution.*

There exists a stationary and deterministic policy that simultaneously maximizes $V^\pi(s)$ for all $s \in \mathcal{S}$.

Theorem 4.3 (Best policy is stationary and determinstic). *Let Π be the set of all non-stationary and randomized policies. There exists a stationary and determinstic policy π such that for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$*

$$V^\pi(s) = V^*(s) := \sup_{\pi \in \Pi} V^\pi(s),$$
$$Q^\pi(s, a) = Q^*(s, a) := \sup_{\pi \in \Pi} Q^\pi(s, a).$$

Theorem 4.4 (**Bellman optimality equations**). *We say that a vector $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ satisfies the Bellman optimality equations if:*

$$Q(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} Q(s', a') \right].$$

For any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, we have that $Q = Q^*$ iff Q satisfies the Bellman optimality equations. Furthermore, the deterministic policy defined by $\pi(s) \in \arg\max_{a \in \mathcal{A}} Q^*(s,a)$ is an optimality policy.

This theorem tells us why we can use Q-function to find the best policy. And so-called Q-learning is based on this.

And for finnite-horizon MDP, we defined

$$V_h^\pi(s) = \mathbb{E}[\sum_{t=h}^{H-1} r_h(s_t, a_t) | \pi, s_h = s],$$
$$Q_h^\pi(s,a) = \mathbb{E}[\sum_{t=h}^{H-1} r_h(s_t, a_t) | \pi . s_h = s, a_h = a].$$

Use the same method we can prove **Bellman optimality equations** for finite horizon MDP.

Theorem 4.5. *Define*

$$Q_h^*(s,a) = \sup_{\pi \in \Pi} Q_h^\pi(s,a),$$

we have that $Q_h = Q_h^*$ forall h iff for all h,

$$Q_h(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} Q_{h+1}(s', a') \right].$$

Next we discuss two simple algorithm. We suppose the MDP $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ is known.

Value iteration:

$$Q \leftarrow \mathcal{T} Q.$$

This will convergerce to Q^* since Q^* is the stationary point of operator $\mathcal{T}$. Then we get the policy π by $\pi(s) = \pi_Q(s) = \operatorname{argmax}_{a \in \mathcal{A}} Q(s,a)$.

Policy iteration:

1. Policy evaluation: Compute $Q^{\pi_k} = (I - \gamma P^\pi)^{-1} r$ by Bellman consistency equations;
2. policy improvement: Update the policy: $\pi_{k+1} = \pi_{Q^{\pi_k}}$.

Next we give a technical lemma, which describes difference between value functions with respect to two policies.

The *advantage* $A\pi(s,a)$ of a policy π is defined as

$$A^\pi(s,a) := Q^\pi(s,a) - V^\pi(s).$$

Lemma 4.6 (Performance difference lemma)**.**

4.2 Sample complexity with a generative model

This chapter talks about how many samples do we need if we want to get value accuracy with high probability.

4.3 Linear Bellman Completeness

The sample complexity of "tabular" MDPs scaled polynomially in the size of the state and action spaces. Now we seek methods which are applicable to cases where number of states and actions is large.

We will work with a feature mapping $\phi : \mathcal{A} \times \mathcal{A} \to \mathbb{R}^d$.The main idea is to assume $Q^*(s, a)$ is a linear function of $\phi(s, a)$.

Definition 4.7 (Linear Bellman Completeness)**.** *There exist* $\omega \in \mathbb{R}^d$ *such that*

$$w^T \phi(s, a) = r(s, a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} \theta^T \phi(s', a'),$$

Note that we already know

$$Q_h^*(s, a) = r(s, a) + E_{s' \sim P_h(s,a)} \max_{a'} \theta^T \phi(s', a').$$

Then for "completeness" $\Rightarrow$ "realizability", I have a question. By definition we know if we have a Q_{h+1} satisfies the realizability condition, then we get Q_t satisfies the same realizability. But how can we get the first Q function? Do we need to use $Q_H = 0$ is clearly a linear function of ϕ, then can we get the conclusion by this?

The answer is yes!

Least square value iteration:

1. $\hat{\theta}_h = \arg\min_{\theta} \sum_{s,a,r,s' \in \mathcal{D}_h} \left(\theta^T \phi(s, a) - r - \max_{a \in \mathcal{A}} \theta_{h+1}^T \phi(s', a) \right)^2$;
2. $\hat{\pi}_h(s) := \arg\max_a \left(\theta_h^T \phi(s, a) \right)$.

4.4 Fitted Dynamic Programming Methods: Fitted Q Iteration

Definition 4.8 (Bellman operator)**.** $\mathcal{T} f(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in \mathcal{A}} f(s', a')$.

Given dataset $D = \{(s_i, a_i, r_i, s_i')\}$ and a function class $\mathcal{F}$, define FQI algorithm

$$f_t \in \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} (f(s_i, a_i) - r_i - \gamma \max_{a' \in \mathcal{A}} f_{t-1}(s_i', a_i))^2.$$

To understand this algorithm, we can think it as a stationary point problem. Since $\mathcal{T} Q^* = Q^*$, we want to find the stationary point of the operator $\mathcal{T}$. For example, start with some $f_0 \in \mathcal{F}$ and use iteration $f_t = \mathcal{T} f_{t-1}$. But we don't know the MDP and $\mathcal{T}$. But given f_{t-1} we know

$$f_t(s, a) := \mathcal{T} f_{t-1}(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in \mathcal{A}} f_{t-1}(s', a')$$

$$= \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s,a) + \gamma \max_{a' \in \mathcal{A}} f_{t-1}(s', a') \right]$$

Actually for some (s_i, a_i), observed value

$$f_t(s_i, a_i) = r_i(s_i, a_i) + \gamma \max_{a' \in \mathcal{A}} f_{t-1}(s_i', a') + noise,$$

so we can just define the loss function as square loss and minimize the loss function.

After k many iterations, we output a policy $\pi^k(s) := \arg\max_a f_k(s,a), \forall s$.

This method has performance guarantee:

Theorem 4.9 (FQI guarantee)**.** *Fix $K \in \mathbb{N}^+$. Fitted Q Iteration guarantees that with probability $1-\delta$,*

$$V^* - V^{\pi^K} \leq \frac{1}{(1-\gamma)^2} \left(\sqrt{\frac{22 C V_{max}^2 \ln(|\mathcal{F}|^2 K/\delta)}{n}} + \sqrt{20 C \epsilon_{approx,\nu}} \right) + \frac{\gamma^K V_{max}}{(1-\gamma)}.$$

4.5 Multi-Armed & Linear Bandits

This chapter introduces basic UCB algorithms of multi-armed and linear bandits. See bandit reading notes.

4.6 Exploration: UCB value iteration for Tabular MDPs and linear MDPs

Algorithm 12 UCBVI

for $n = 1, \ldots, N$ **do**
  Learn transition model $\{\hat{P}_h^n\}_{h=0}^{H-1}$from all previous data;
  Design reward bonus b_h^n;
  Plan: π^{n+1} =Value-Iteration$\left(\{\hat{P}^n\}_h, \{r_h + b_h^n\} \right)$;
  Execute π^{n+1} for H steps.
end for

Algorithm 13 LSVI-UCB

for $k = 1, \ldots, K$ **do**
  Receive the initial state x_1^k
  for step h=H,. . . ,1 **do**
    $\Lambda_h \leftarrow \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau) \phi(x_h^\tau, a_h^\tau)^T + \lambda I.$
    $w_h \leftarrow \Lambda_h^{-1} \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau) \left[r_h(x_h^\tau, a_h^\tau) + \max_a Q_{h+1}(x_{h+1}^\tau, a) \right].$
    $Q_h(\cdot,\cdot) \leftarrow \min \left\{ w_h^T \phi(\cdot,\cdot) + \beta [\phi(\cdot,\cdot^T \Lambda_h^{-1} \phi(\cdot,\cdot))]^{1/2}, H \right\}.$
  end for
  for step h=1,. . . ,H **do**
    Take action $a_h^k \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} Q_h(x_h^\tau, a)$, and observe x_{h+1}^k.
  end for
end for

In tabular setting, we use an empirical model to estimate $P_h^k(s'|s,a)$. While, in low-rank linear MDPs, we use ridge linear regression:

$$\hat{\mu}_h^n = \underset{\mu \in \mathbb{R}^{|\mathcal{S}| \times d}}{\operatorname{argmin}} \sum_{i=0}^{n-1} \|\mu\phi(s_h^i, a_h^i) - \delta(s_{h+1}^i)\|_2^2 + \lambda \|\mu\|_F^2.$$

We do this because $\mathbb{E}[P(\cdot|s,a) - \delta(s_{h+1}^i)|\mathcal{H}_h^i] = 0$.

Ridge linear regression has the following closed-form solution:

$$\hat{\mu}_h^n = \sum_{i=1}^{n} \delta(s_{h+1}^i)\phi(s_h^i, a_h^i)^T (\Lambda_h^n)^{-1}.$$

Note that $\mu_h^n \in \mathbb{R}^{|\mathcal{S}| \times d}$, so we never want to explicitly store it. Note that in our value iteration we just care about $\hat{P}_h^n(\cdot|s,a) \cdot V := \phi(s,a)^T \hat{\mu_h^n}^T V$, which can be re-written as:

$$\phi(s,a)^T \sum_{i=0}^{n-1} (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i) V(s_{h+1}^i),$$

where we use the fact that $\delta(s)^T V = V(s)$. Thus the operator $\hat{P}_h^n(\cdot|s,a) \cdot V$ simply requires storing all data and can be computed via simple linear algebra and the computation complexity is simply $poly(d,n)$-no poly dependency on $|\mathcal{S}|$.

There is a very important lemma in the decomposition of regret.

Lemma 4.10 (Simulation lemma)**.**

$$\hat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a \sim d_n^{\pi^n}} \left[b_h^n(s,a) + (\hat{P}_h^n(\cdot|s,a) - P_h(\cdot|s,a)) \cdot \hat{V}_{h+1}^n \right].$$

Proof.

$$\begin{aligned}
\hat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) &= \hat{Q}_0^n(s_0, \pi^n(s_0)) - Q_0^{\pi^n}(s_0, \pi^n(s_0)) \\
&\leq r_0(s_0, \pi^n(s_0)) + b_h^n(s_0, \pi^n(s_0)) + \hat{P}_0^n(\cdot|s_0, \pi^n(s_0)) \cdot V_1^n - r_0(s_0, \pi^n(s_0)) - P_0^n(\cdot|s_0, \pi^n(s_0)) \cdot V_1^{\pi^n} \\
&= b_h^n(s_0, \pi^n(s_0)) + \hat{P}_0^n(\cdot|s_0, \pi^n(s_0)) \cdot \hat{V}_1^n - P_0^n(\cdot|s_0, \pi^n(s_0)) \cdot V_1^{\pi^n} \\
&= b_h^n(s_0, \pi^n(s_0)) + \left(\hat{P}_0^n(\cdot|s_0, \pi^n(s_0)) - P_0^n(\cdot|s_0, \pi^n(s_0)) \right) \cdot \hat{V}_1^n + P_0^n(\cdot|s_0, \pi^n(s_0)) \cdot \left(\hat{V}_1^n - V_1^{\pi^n} \right) \\
&\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a \sim d_n^{\pi^n}} \left[b_h^n(s,a) + (\hat{P}_h^n(\cdot|s,a) - P_h(\cdot|s,a)) \cdot \hat{V}_{h+1}^n \right],
\end{aligned}$$

where we use induction in the last step. □

Theorem 4.11. *For tabular MDPs, UCBVI has regret bound* $\tilde{O}(H^2\sqrt{S^2AN})$.

Theorem 4.12. *For linear MDPs, UCBVI has regret bound* $\tilde{O}(H^2 d^{1.5}\sqrt{N})$.

4.7 Learning in Large Scale MDPs (Bellman rank)

Obtaining sample size results which are independent of the size of the state space (and possibly the action space) is essentially a question of *generalization*, which is the focus of this chapter.

Definition 4.13 (Bellman rank)**.** *Q-Bell man rank: related to the Bellman error of a Q function estimate g:*

$$\mathcal{E}(g; f, h) = \mathbb{E}_{s_h, a_h \sim d_h^{\pi_f}} \left[g(s_h, a_h) - r(s_h, a_h) - \mathbb{E}_{s_{h+1} \sim P(\cdot|s_h, a_h)} \left[\max_{a \in \mathcal{A}} g(s_{h+1}, a) \right] \right]$$

V-Bellman rank: related to the Bellman error to a V function estimate

$$\mathcal{E}(g; f, h) = \mathbb{E}_{s_h \sim d_h^{\pi_f}} \left[V_g(s_h) - r(s_h, \pi_g(s_h) - \mathbb{E}_{s_{h+1} \sim P(\cdot|s_h, \pi_g(s_h))} \left[V_g(s_{h+1}) \right] \right]$$

Many models (more in the book chapter) indeed have low-Q or V Bellman rank.

Algorithm 14 BLin-UCB

1: **Input:** number of iteration T, batch size m, confidence radius R
2: **for** $t = 1, \ldots, T$ **do**
3: Select $f_t = \underset{g \in \mathcal{F}}{\operatorname{argmax}} V_g(s_0)$ s.t.

$$\forall h : \sum_{i=0}^{t-1} (\mathbb{E}_{\mathcal{D}_{h,i}}[\ell(s_h, a_h, s_{h+1}, g)])^2 \leq R^2$$

4: For all h, create $\mathcal{D}_{h,t} = \{s_h, a_h, s_{h+1}\}$ with m triples
5: **end for**
6: **return:** $\underset{\pi \in \pi_{f_0}, \ldots, \pi_{f_{T-1}}}{\operatorname{argmax}} V^{\pi}$.

An interesting lemma used in the proof:

Lemma 4.14.

$$V_{f_t}(s_0) - V^{\pi_{f_t}}(s_0) = \sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim d_h^{\pi_{f_t}}} \left[f_t(s_h, a_h) - r(s_h, a_h) - \mathbb{E}_{s_{h+1} \sim P_h(s_h, a_h)} \max_{a'} f_t(s_{h+1}, a') \right].$$

Proof. Use telescoping. Left hand side in the equation can be split into H terms and most of them cancel out. □

The main theorem in this chapter is:

Theorem 4.15 (Analysis of BLin-UCB)**.** *After running BLin-UCB for $T = \tilde{O}(HD)$ many iterations, there exists a policy among T many policies, such that:*

$$V^*(s_0) - V^{\pi}(s_0) \leq \tilde{O}(\varepsilon_{gen}(m, \mathcal{F}, \delta/(TH)) \cdot \sqrt{dH^3}).$$

, and the number of trajectories we used is mHT.

For discrete (but maybe large) hypothesis class $\mathcal{F}$ for Q-Bellman rank, we have:

Corollary 4.16. *With probability at least* $1-\delta$*, BLin-UCB learns a policy with* $V^* - V^\pi \leq \varepsilon$*, with the number of trajectories* $\tilde{O}\left(\frac{H^6 d^2 \ln(|\mathcal{F}|/\delta)}{\varepsilon^2}\right)$.

5 Paper notes: bandit

5.1 Lifting the Information Ratio: An Information-Theoretic Analysis of Thompson Sampling for Contextual Bandits

5.1.1 Abstract

The paper adapt the information-theoretic perspective of [RVR16] to the contextual setting by introducing a new concept of information ratio based on the mutual information between the unknown model parameter and the observed loss. And the main goal is to bound the regret in terms of the entropy of the prior distribution through a remarkably simple proof, and with no structural assumptions on the likelihood or the prior. After proving the general results, it is mentioned in this paper that several specific binary loss bandits and linear gaussian loss bandit have good regret bounds.

5.1.2 Questions to ask

Question 5.1. *What is the difference between the information ratio based TS and the original TS algorithm?*

Actually, information ratio just consider a new way to analysis the regret of TS algorithm, the algorithm is the same. Using information ratio can bound the regret in terms of the entropy of the prior distribution and the upper bound of so-called information ratio.

Question 5.2. *What is the benefit of using this new notion?*

I think it provides a new framework for algorithmic regret analysis. In the future if we want to prove a regret bound for TS, we can first think if we can find upper bound for this information ratio and maybe the entropy of the prior distribution.

In the main section, I use red color to denote some things we may improve this method. Just my personal thinking, due to limited knowledge may not be the right idea

5.1.3 Preliminaries

We consider a parametric class of contextual bandits with parameters space Θ, context space $\mathcal{X}$, and K actions. To each parameter $\theta \in \Theta$ there corresponds a contextual bandit with loss distribution $P_{\theta,x,a}$ for each context $x \in \mathcal{X}$ and action $a \in \mathcal{A}$, with the mean loss of the distribution denoted by $l(\theta, x, a)$.

We study the problem of regret minimization in the Bayesian setting. In this setting, the environment secretly samples a parameter θ^* from a known prior distribution Q_1 over Θ. We assume that the agent has full knowledge of the prior and the likelihood model $P_{\theta,x,a}$. The goal of the agent is to minimize the expected sum of losses. In the Bayesian setting, this is equivalent to minimizing the Bayesian regret, defined as follows:

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T}(l(\theta^*, X_t, A_t) - l(\theta^*, X_t, A_t^*))\right],$$

where A_t^* is the optimal action for round t.

Furthermore, let $\mathcal{F}_t = \sigma(X_1, A_1, L_1, \ldots, X_t, A_t, L_t)$. We use Q_t to denote the distribution of the unknown parameter θ^* conditional on the past history $\mathcal{F}_{t-1}$. We denoted by $\pi(\cdot|X_t)$ the distribution over the agent's actions conditional on X_t and $\mathcal{F}_{t-1}$, and call it agent's policy. Finally, $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_{t-1}, X_t]$ and $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot|\mathcal{F}_{t-1}, X_t]$.

In Thompson Sampling, an important fact is $\mathbb{P}_t[A_t = a] = \mathbb{P}_t[A_t^* = a]$ and $(\theta, A_t) \overset{\mathrm{d}}{=} (\theta^*, A_t^*)$.

The Shannon entropy of X is defined as

$$H(X) = -\sum_{x \in \mathcal{X}} \mathbb{P}(X = x) \log \mathbb{P}(X = x).$$

Fact 5.3. $0 \leq H(X) \leq \log(|\mathcal{X}|)$.

For two probability measures P and Q, if P is absolutely continuous with respect to Q, the *Kullback-Leibler divergence* between them is

$$D(P||Q) = \int \log \frac{dP}{dQ} dP = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

Fact 5.4. *$D(P||Q) \geq 0$ with equality if and only if $P = Q$ P-a.s.*

Mutual information is defined by

$$I(X;Y) = D(P(X,Y)||P(X)P(Y)) = \mathbb{E}_X[D(P(Y|X)||P(Y))].$$

Fact 5.5 (Data processing inequality)**.** *$I(X;Y) \geq I(X;g(Y))$.*

Definition 5.6 (Information ratio)**.** *Informally, the information ratio measures the trade-off between achieving low regret and gaining information about the identity of the optimal action A^* (which is a deterministic function of θ^* in the standard multi-armed bandit setting). The formal definition is given by*

$$\rho_t^* = \frac{(\mathbb{E}_t[l(\theta^*, A_t) - l(\theta^*, A^*)])^2}{I_t(A^*; (A_t, L_t))}.$$

But in contextual bandit, the optimal action A_t^* changes from round to round, influenced by the context X_t, so information about A^* is useless.

Definition 5.7 (Lifted information ratio)**.**

$$\rho_t = \frac{(\mathbb{E}_t[l(\theta^*, A_t) - l(\theta^*, A^*)])^2}{I_t(\theta^*; L_t)}.$$

We can use

$$\theta \to A_t \to L_t$$

to describe the relationship between θ, A_t and L_t, the arrow means given A_t, θ and L_t are conditional independent. Then the data processing inequality implies that the information gain about θ^* is always smaller than that about A_t^* , which in turn implies that ρ_t is greater than ρ_t^*.

As our analysis will establish, a bounded lifted information ratio guarantees low regret, and we will show that the ratio itself can be bounded reasonably under conditions similar to the ones required by the analysis of [RVR16].

5.1.4 Main results

Theorem 5.8. *assume Q_1 is supported on the countable set $\Theta_1 \subseteq \Theta$ and that the lifted information ratio for all rounds t satisfies $\rho_t \leq \rho$ for some $\rho > 0$. Then, the Bayesian regret of TS after T rounds can be bounded as*

$$R_T \leq \sqrt{\rho T H(\theta^*)}.$$

Using this theorem, if we can find upper bounds for ρ and $H(\theta^*)$, then we find upper bound for the regret.

Lemma 5.9. *Suppose that the losses are binary and $|\mathcal{A}| = K$. Then, the lifted information ratio of Thompson sampling satisfies $\rho_t \leq 2K$ for all $t \geq 1$.*

We now instantiate our bounds in two well-studied settings for Bernoulli bandits. We start from the fully unstructured case, assuming finite actions and finitely supported prior. The following regret bound follows direst from Theorem 5.1.4 and Lemma 5.9.

Theorem 5.10. *Consider a contextual bandit with K actions and binary losses, and suppose Θ_1, the support of Q_1, is finite with $|\Theta_1| = N$. Then, the Bayesian regret of TS satisfies:*

$$R_T \leq \sqrt{2KT \log N}.$$

Unfortunately, the Shannon entropy can be unbounded for distributions with infinite support, which is in fact the typical situation that one encounters in practice. To address this concern, we develop a more general result, that holds for a broader family of distributions.

In the following, (Θ, ϱ) is a metric space with metric $\varrho : \Theta^2 \to \mathbb{R}$. We make the following regularity assumption on the likelihood function $P_{\theta,x,a}$:

Assumption 5.11 (log-Lipschitz)**.** *There exists a constant $C > 0$ such that for any $\theta, \theta' \in \Theta_1$, $|\log P_{\theta,x,a} - \log P_{\theta',x,a}| \leq C\varrho(\theta, \theta')$ holds for all $x \in \mathcal{X}$, $a \in \mathcal{A}$, and $L \in \{0, 1\}$.*

Under this assumption, we can state a variant of Theorem 1 that applies to metric parameter spaces:

Theorem 5.12. *assume (Θ, ϱ) is a metric space, and Q_1 is supported on $\Theta_1 \subseteq \Theta$ with ϵ-covering number $\mathcal{N}_\epsilon(\Theta_1, \varrho)$. Let assumption hold, and assume the lifted information ratio for all round t satisfies $\rho_t \leq \rho$ for some $\rho > 0$. Then, the Bayesian regret of TS after T rounds can be bounded as*

$$R_T \leq \sqrt{\rho T \min_\epsilon (\log \mathcal{N}_\epsilon(\Theta_1, \varrho) + 2\epsilon CT)}$$

When the Shannon entropy is bounded, we can use it to bound the regret. And the article find one way to deal with unbounded entropy problem by adding Lipschitz assumption.

One potential direction of improvement is to find new ways to bound $\sum_{t=1}^{T} I_t(\theta^*, L_t)$.

In this model, the losses are generated by a Bernoulli distribution as $L_t(\theta, x, a) \sim Ber(\sigma(f_\theta(x, a)))$, where $\sigma(z) = 1/(1 + e^{-z})$ is sigmoid function.

See Theorem 4 and Corollary 1 in [NOPS22] for the results.

We can consider two types of linear bandits. And the regret analysis is similar. The first one supposes the losses are binary, and the expected losses are linear functions of the form $l(\theta, x, a) =< \theta, \phi(x, a) >$, see Lemma 2 in [NOPS22].

Another type is linear bandits with Gaussian noise. $L_t \sim \mathcal{N}(l(\theta^*, X_t, A_t), \sigma^2)$. See Lemma 3 in [NOPS22].

Another potential direction is apply the method to more types of bandits. For one thing, we can use this lifted information ratio to deal with other types of contextual bandit. For another, we can develop new theory about information ratio to deal with bandits beyond basic and contextual bandits. For example, we may find another type of information ratio.

I don't have so much knowledge about types of bandits beyond these, and we can have a discussion here.

5.2 Contextual Information-Directed Sampling

Paper notes

5.3 Improved Algorithms for Linear Stochastic Bandits

This is a very classic paper on linear bandit. Familiarity with this technique is very helpful in understanding linear bandit.

I'd like to summarize the main results and the sketch of proofs.

In each round t, the learner is given a decision set $D_t \subseteq \mathbb{R}^d$ from which he has to choose an action X_t. Subsequently he observes reward $Y_t = \langle X_t, \theta_t \rangle + \eta_*$ where $\theta_* \in \mathbb{R}^d$ is an unknown parameter and η_t is a random noise satisfying $\mathbb{E}[\eta_t | X_{1:t}, \eta_{1:t-1}] = 0$ and some tail-constraints, to be specified soon.

The goal of the learner is to maximize his total reward $\Sigma_{t=1}^n \langle X_t, \theta_* \rangle$ accumulated over the course of n rounds.

5.3.1 OFU: Optimism in the face of uncertainty

In each round t, the learner is given a decision set $D_t \in \mathbb{R}^d$, and the algorithm maintains a confidence set $C_{t-1} \subseteq \mathbb{R}^d$.

The algorithm choose the pair

$$(X_t, \tilde{\theta}_t) = \underset{(x,\theta) \in D_t \times C_{t-1}}{\operatorname{argmax}} \langle x, \theta \rangle. \tag{5.1}$$

5.3.2 Self-normalized tail inequality for vector-valued martingales

Theorem 5.13 (Self-Normalized Bound for Vector-Valued Martingales)**.** *Let $\{F_t\}_{t=0}^\infty$ be a filtration. Let $\{\eta_t\}_{t=1}^\infty$ be a real-valued stochastic process such that η_t is F_t-measurable and η_t is conditionally R-sub-Gaussian for some $R > 0$ i.e.*

$$\forall \lambda \in \mathbb{R} \qquad \mathbb{E}[e^{\lambda \eta_t} | F_{t-1}] \leq \exp \frac{\lambda^2 R^2}{2}$$

Let $\{X_t\}_{t=1}^\infty$ be a $\mathbb{R}^d$-valued stochastic process such that X_t is F_{t-1}-measurable. Assume that V is a $d \times d$ positive definite matrix. For any $t > 0$, define

$$\bar{V}_t = V + \sum_{s=1}^t X_s X_s^T \qquad S_t = \sum_{s=1}^t \eta_s X_s.$$

Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $t > 0$

$$\|S_t\|^2_{\bar{V}_t^{-1}} \leq 2R^2 \log(\frac{\det(\bar{V}_t)^{1/2}\det(V_t)^{-1/2}}{\delta}).$$

Proof. Use the property of sub-Gaussian variable to construct a supermartingale.
Then prove the event concerning the stopped process has the high-probability bound.
Finally claim the event we care can be described using this stopped process. □

5.3.3 Construction of confidence sets

Let $\hat{\theta}_t$ be the ℓ^2-regulized least-squares estimate of θ_* with regularization parameter $\lambda > 0$:

$$\hat{\theta}_t = \left(X_{1:t}^T X_{1:t} + \lambda I\right)^{-1} X_{1:t}^T Y_{1:t} = \left(\sum_{i=1}^{t} X_i X_i^T + \lambda I\right)^{-1} (X_1, \ldots, X_t) \begin{pmatrix} Y_1^T \\ \vdots \\ Y_t^T \end{pmatrix} \tag{5.2}$$

The following theorem shows that $theta_*$ lies with high probability in an ellipsod with center at $\hat{\theta}_t$.

Lemma 5.14 (Determinant-Trace Inequality)**.** *Suppose $X_1, \ldots, X_t \in \mathbb{R}^d$ and for any $1 \leq s \leq t$, $\|X_s\|_2 \leq L$. Let $\hat{V}_t = \lambda I + \Sigma_{s=1}^t X_s X_s^T$ for some $\lambda > 0$. Then,*

$$\det(\hat{V}_t) \leq (\lambda + tL^2/d)^d.$$

Theorem 5.15 (Confidence Ellipsoid)**.** *Assume the same condition. Let $V = I\lambda, \lambda > 0$, define $Y_t = \langle X_t, \theta_* \rangle + \eta_t$ and assume that $\|\theta_*\|_2 \leq S$. Then, for any delta > 0, with probability at least $1-\delta$. for all $t \geq 0$,$theta_*$ lies in the set*

$$C_t = \left\{ \theta \in \mathbb{R}^d : \|\hat{\theta}_t - \theta\|_{\bar{V}_t} \leq R\sqrt{2\log\left(\frac{\det(\hat{V}_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2} S \right\}.$$

Furthermore, if for all $t \geq 1$, $\|X_s\|_2 \leq L$ then with probability at least $1-\delta$, for all $t \geq 0$, θ_* lies in the set

$$C_t = \left\{ \theta \in \mathbb{R}^d : \|\hat{\theta}_t - \theta\|_{\bar{V}_t} \leq R\sqrt{d\log\left(\frac{1 + tL^2/\lambda}{\delta}\right)} + \lambda^{1/2} S \right\}.$$

Proof. Using the Cauchy-Schwarz Inequality, we get

$$|x^T\hat{\theta}_t - x^T\theta_*| \leq \|x\|_{\bar{V}_t^{-1}} \left(\|\boldsymbol{X}^T\eta\|_{\bar{V}_t^{-1}} + \lambda^{1/2}\|\theta_*\|_2 \right),$$

By 5.13 we have a high probability bound for $\|\boldsymbol{X}^T\eta\|_{\bar{V}_t^{-1}}$. Assuming θ_* is bounded, we can get the final confidence ellipsoid for θ_* by plugging in $x = \bar{V}_t(\hat{\theta}_t - \theta_*)$. □

5.3.4 Regret analysis of OFUL algorithm

Recall the OFUL algorithm maintains a confidence set $C_{t-1} \subseteq \mathbb{R}^d$, and choose the pair

$$(X_t, \tilde{\theta}_t) = \underset{(x,\theta)\in D_t \times C_{t=1}}{\operatorname{argmax}} \langle x, \theta \rangle \tag{5.3}$$

at each rounds.

Theorem 5.16 (Regret of OFUL)**.** *Assume that for all t and all $x \in D_t$, $\langle x, \theta_* \rangle \in [-1, 1]$ and let $\lambda \geq 1$. Then, with probability at least $1 - \delta$, the regret of the OFUL algorithm satisfies*

$$\forall n \geq 0, \quad R_n \leq 4\sqrt{nd \log(\lambda + nL/d)} \left(\lambda^{1/2} S + R\sqrt{2\log(1/\delta) + d\log(1 + nL/(\lambda d))} \right).$$

Remark 5.17. *The notation β in the proof is strange. Its first time appearance is about confidence bound from other paper. And there are two confusing inequality in the proof.*

Since we have

$$\begin{aligned}
\det(\bar{V}_n) &= \det(\bar{V}_{n-1} + X_n X_n^T) \\
&= \det(\bar{V}_{n-1}) \det(I + \bar{V}_{n-1}^{-1/2} X_n (\bar{V}_{n-1}^{-1/2} X_n)^T) \\
&= \det(\bar{V}_{n-1})(1 + \|X_{n-1}\|_{\bar{V}_{n-1}}) \\
&= \det(V) \prod_{t=1}^{n} \left((1 + \|X_{n-1}\|_{\bar{V}_{n-1}} \right),
\end{aligned}$$

we can recompute θ_t whenever $\det(\bar{V}_t)$ increases by a constant factor $1 + C$. We call the resulting algorithm the RARELY SWITCHING OFUL algorithm.

And theorem 4 in paper proves a regret bound for this algorithm.

Then the paper discuss problem dependent bound.

Theorem 5.18. *Assume that $\lambda \geq 1$ and $\|\theta_*\|_2 \leq S$ where $S \geq 1$. With probability at least $1 - \delta$, for all $n \geq 1$, the regret of OFUL satisfies*

$$\begin{aligned}
R_n \leq \frac{16R^2\lambda S^2}{\bar{\Delta}_n} (\log(Ln) + (d-1)\log\frac{64R^2\lambda S^2 L}{\bar{\Delta}_n} \\
+ 2(d-1)\log(d\log\frac{d\lambda + nL^2}{D} + 2\log(1/\delta) + 2\log(1/\delta))^2.
\end{aligned}$$

5.3.5 Multi-armed bandit problem

They can use the method in this paper to analysis multi-armed bandit. Just let $d = 1$ and get confidence intervals, then get regret of the algorithm $UCB(\delta)$.

Lemma 5.19 (Confidence Intervals)**.** *Assuming that the noise η_t is conditionally 1-sub-Gaussian. With probability at least $1 - \delta$,*

$$\forall i \in \{1, 2, \ldots, d\}, \forall t \geq 0, \quad |\bar{X}_{i,t} - \mu_i| \leq c_{i,t},$$

where

$$c_{i,t} = \sqrt{\frac{1 + N_{i,t}}{N_{i,t}^2} \left(1 + 2\log(\frac{d(1 + N_{i,t})^{1/2}}{\delta}) \right)}.$$

Theorem 5.20 (Regret of UCB(δ))**.** *Assume that the noise η_t is conditionally 1-sub-Gaussian, with probability at least $1-\delta$, the total regret of the $UCB(\delta)$ is bounded as*

$$R_n \leq \sum_{i:\Delta_i>0} \left(3\Delta_i + \frac{16}{\Delta_i} \log \frac{2d}{\Delta_i \delta}\right).$$

5.4 Thompson Sampling Regret Bounds for Contextual Bandits with sub-Gaussian rewards

The proof in this paper is quite well written! It's very comfortable to read!

5.4.1 Genearl theoretic results

In this paper, we extend the results from [NOPS22] to contextual bandits with sub-Gaussian rewards.

The first theorem is just the same as the one in [NOPS22], but slightly change the form of the notion.

Theorem 5.21. *Assume that the average of the lifted information ratios is bounded* $\frac{1}{T}\Sigma_{t=1}^{T}\mathbb{E}[\Gamma_t] \leq \Gamma$ *for some* $\Gamma > 0$*. Then, the TS cumulative regret is bounded as*

$$\begin{aligned} Regret &\leq \sqrt{\Gamma T I(\Theta; \hat{H}_{T+1})} \\ &= \sqrt{\Gamma T \mathbb{E}[D_{KL}(\mathbb{P}_{\Theta|\hat{H}^{T+1}} \| \mathbb{P}_{\Theta})]} \end{aligned}$$

Proof. The proof follows by the chain rule of the mutual information.

$$I(X; Y, Z) = I(X; Z) + I(X; Y|Z).$$

□

In the next theorem, they want to research on problems when $\theta \in \mathbb{R}^d$. It is based a weaker regularity condition. And the requirement of binary rewards is unnecessary.

Assumption 5.22 (Lipschitz process)**.** *There is a random variable* $C > 0$ *that can be depend only on* R_t, X_t *and* $\hat{A}_t$ *such that*

$$|\log f_{R_t|X_t,\hat{A}_t,\Theta=\theta}(R_t) - \log f_{R_t|X_t,\hat{A}_t,\Theta=\theta'}(R_t)| \leq C\rho(\theta, \theta') \quad a.s.\ for\ all\ \theta, \theta' \in \mathcal{O}.$$

Note that the smallest cardinality of an ϵ-net for $(\mathcal{O}, \rho)$ is called the ϵ-covering number.

Theorem 5.23. *Assume that the parameters' space is a metric space* $(\mathcal{O}, \rho)$ *and let* $|\mathcal{N}(\mathcal{O}, \rho, \epsilon)|$ *be the* ϵ*-covering number of this space for any* $\epsilon > 0$*. Assume as well that the log-likelihood is a Lipschitz process according to assumption above and that the average of the lifted information ratio is bounded like the first theorem. Then, the TS cumulative regret is bounded as*

$$Regret \leq \sqrt{\Gamma T \min_{\epsilon>0}\{\epsilon\mathbb{E}[C]T + \log|\mathcal{N}(\mathcal{O}, \rho, \epsilon)|\}}.$$

Proof. The mutual information can be written as

$$I(\Theta; R_t|\hat{H}_t, X_t, A_t) = \mathbb{E}\left[\int_{\mathcal{O}} f_{\Theta|R_t,\hat{H}_t,X_t,A_t}(\theta)\left(\log\frac{f_{\Theta,\hat{H}_t,X_t,A_t(R_t)}}{f_{\pi(\Theta),\hat{H}_t,X_t,A_t}(R_t)} + \log\frac{f_{\pi(\Theta),\hat{H}_t,X_t,A_t(R_t)}}{f_{R_t|\hat{H}_t,X_t,A_t}(R_t)}\right)d\theta\right],$$

where the first term can be bounded by log-Lipschitz condition, and the second term is $I(\Theta_\pi; R_t|\hat{H}_t, X_t, A_t))$, Θ_π is a discrete version of Θ and it is finite since we can find a finite ϵ-net. So the second term can be bounded by the ϵ-covering number just like the first theorem. □

5.4.2 Bounding the lifted information ratio

In this section, they prove two lemmata. The first is for finite action set just like Lemma 1 in [NOPS22]. The second is for finite action set and $\theta \in \mathbb{R}^d$ setting. In this paper the authors deal with sub-Gaussian rewards instead of binary rewards.

Lemma 5.24. *Assume the number of actions $|\mathcal{A}|$ is finite and the random rewards R_t are σ^2-sub-Gaussian, then $\Gamma_t \leq 2\sigma^2|\mathcal{A}|$.*

Lemma 5.25. *Assume the number of actions $|\mathcal{A}|$ is finite, the expectation of the rewards is $\mathbb{E}[R(x,a,\theta)] = \langle \theta, m(x,a)\rangle$ for some feature map m. The random rewards R_t are σ^2-sub-Gaussian. Then $\Gamma_t \leq 2\sigma^2 d$.*

This lemma is useful in cases where the dimension is smaller than the number of actions $d < |A|$.

5.4.3 Application for special bandits

The first result is a corollary of the first theorem and the first lemma. They close the gap on the regret of the TS algorithm showing that it is in $O(\sqrt{|\mathcal{A}|T\log|\mathcal{O}|})$ for sub-Gaussian rewards, and thus for bounded ones.

Corollary 5.26. *Assume that the rewards are bounded in $[0, L]$. Then, for any contextual bandit problem Φ, the TS cumulative regret after T rounds is bounded as*

$$Regret \leq \sqrt{\frac{L^2|\mathcal{A}|TH(\Theta)}{2}}.$$

The next corollary gives a general result for bandits with Laplace likelihoods. This setting considers rewards with a likelihood proportional to $\exp\left(-\frac{|r-f_\theta(x,a)}{\beta}\right)$ for some $\beta > 0$. In addition, this setting assumes that the random variable $f_\theta(X, A)$ is a Lipschitz process with respect to θ with random variable $C = C(X, A)$. This ensure the Lipschitz log-likelihood.

Corollary 5.27. *Assume that $\mathcal{O} \in \mathbb{R}^d$ with $diag(\mathcal{O}) \leq S$. Consider a contextual bandit problem Φ with Laplace likelihood and rewards bounded in $[0, L]$. Then, the TS cumulative regret after T rounds is bounded as*

$$Regret \leq \sqrt{\frac{L^2|\mathcal{A}|Td}{2}\left(1 + \log(\frac{3S\mathbb{E}[C]T}{d\beta})\right)}.$$

Then they deal with bernoulli bandits with rewards distributed as $Ber(g \circ f_\Theta(X_t, \hat{A}_t))$, where g is a binomial like function and f is a linear function. This part just like the section about Lipschitz bandits in [NOPS22].

Finally they talked about bounded linear contextual bandits. In this setting [CLRS11] showed that $LinUCB$ has a regret bound in $O(\sqrt{dT\log^3(|\mathcal{A}|T\log(t)/\delta)})$ with probability no smaller than $1-\delta$. This corollary shows that TS has a regret bound in $O\left(\sqrt{d^2T\log(\frac{3}{\epsilon})}\right)$.

5.5 Contextual Bandits with Linear Payoff Functions

In this paper they prove an $O(\sqrt{Td\ln^3(KT\ln(T)/\delta)})$ regret bound that holds with probability $1-\delta$ for the simplest known UCB algorithm. They also prove a lower bound of $\Omega(\sqrt{Td})$ for this setting, matching the upper bound up to logarithmic factors.

Let T be the number of rounds and K the number of possible actions. Let $r_{t,a} \in [0,1]$ be the reward of action a on round t.

On each round t, for each action a the learner observes K feature vectors $x_{t,a} \in \mathbb{R}^d$, with $\|x_{t,a}\|$ where $\|\cdot\|$ denotes the $\ell 2$-norm. This is the meaning of "contextual bandits".

Linear realizability assumption:

$$\mathbb{E}[r_{t,a}|x_{t,a}] = x_{t,a}^T\theta^*,$$

where $\|\theta^*\| \leq 1$.

5.5.1 LinUCB

For convenience, define:

$$\begin{aligned}
s_{t,a} &= \sqrt{x_{t,a}^T A^{-1} x_{t,a}} \in \mathbb{R}_+ \\
D_t &= [x_{\tau,a_\tau}^T]_{\tau\in\Psi_t} \in \mathbb{R}^{|\Psi_t|\times d} \\
y_t &= [r_{\tau,a_\tau}]_{\tau\in\Psi_t} \in \mathbb{R}^{|\Psi_t|\times 1} \\
A_t &= I_d + D_t^T D_t \\
b_t &= D_t^T y_t \\
\lambda_{t,j} &= \text{the eigenvalue of} \quad A_t
\end{aligned}$$

First I will talk about the classic LinUCB algorithm.

The method to estimate the unknown parameter θ is Ridge Regression. It calculate the estimator in this way: $\hat{\theta}_t = \left(\Sigma x_i x_i^T\right)\Sigma x_i y_i$, where x_i denotes the feature of one arm, y_i denotes the reward. And in each round, the algorithm choose the action according to its upper confidence bound:

$$a_t = \arg\max \qquad \hat{\theta}_t^T x_{t,a} + \alpha\sqrt{x_{t,a}^T A^{-1} x_{t,a}}.$$

While experiments show LinUCB is probably sufficient in practice, there is technical difficulty in alalyzing it. Because predictions in later rounds are made using previous outcomes. To handle this problem, they modify the algorithm into BaseLinUCB which assumes statistical independence among the samples, and use as master algorithm SupLinUCB to ensure the assumption holds, in order to apply Azuma/Hoeffding inequality.

Actually, this problem was solved later, see the first paper notes concerning linear bandits.

Now, we discuss the method in this paper.

Algorithm 15 BaseLinUCB

1: Inputs: $\alpha \in \mathbb{R}_+, \Psi_t \subset \{1,2,\ldots,t-1\}$

2: Calculate $A_t, b_t, \theta_t, w_{t,a} = \alpha\sqrt{x_{t,a}^T A^{-1} x_{t,a}}, \hat{r}_{t,a} = \theta_t^T x_{t,a}$

Lemma 5.28 (Confidence bound)**.** *Suppose the input index set Ψ_t in BaseLinUCB is constructed so that for fixed x_{τ,a_τ} with $\tau \in \Psi_t$, the random rewards r_{τ,a_τ} are independent random variables with mean $\mathbb{E}[r_{\tau,a_\tau}] = x_{\tau,a_\tau}^T \theta^*$. Then, with probability at least $1 - \frac{\delta}{T}$, we have for all $a \in [K]$ that*

$$|\hat{r}_{t,a} - x_{t,a}^T \theta^*| \leq (\alpha + 1) s_{t,a}.$$

Proof. The proof of this lemma based on Azuma-Hoeffding bound:
If $X_i \in [a, b]$, then we obtain the bound

$$\mathbb{P}[\sum_{i=1}^{n} |X_i - \mu_i|] \leq 2 \exp\{-\frac{2t^2}{n(b-a)^2}\}.$$

□

Algorithm 16 SupLinUCB

Inputs: $T \in N$
$S \leftarrow \ln(T)$
Index set: $\Psi_t^s \leftarrow \varnothing$ for all $s \in [T]$
for t=1,2,...,T **do**
  repeat
    Use BaseLinUCB with Ψ_t^s to calculate the width $w_{t,a}^s$, and upper confidence bound $\hat{r}_{t,a}^s + w_{t,a}^s$ for all $a \in \hat{A}_s$.
    if $w_{t,a}^s \leq 1/\sqrt{T}$ for all $a \in \hat{A}_s$ **then**
      choose a_t
      Keep the same index set at all levels
    else if $w_{t,a}^s \leq 2^{-s}$ for all $a \in \hat{A}_s$ **then**
      Update the action set $\hat{A}_{s+1}$
      $s \leftarrow s + 1$
    else
      Choose action a_t.
      Update the index set at all levels: keep this step
    end if
  until an action a_t is found.
end for

Theorem 5.29. *If SupLinUCB is run with*

$$\alpha = \sqrt{\frac{1}{2} \ln \frac{2TK}{\delta}},$$

then with probability at least $1 - \delta$, the regret of the algorithm is

$$O\left(\sqrt{Td \ln^3 \left(KT \ln(T)/\delta\right)}\right).$$

5.5.2 Lower bound

In this section, they prove the following lower bound that matches the upper bound up to logarithmic factors.

Theorem 5.30. *For the contextual bandit problem with linear payoff function, for any number of trials T and K actions (where $T \geq K \geq 2$), for any algorithm $\mathcal{A}$ choosing action a_t at time t, there is a constant $\gamma > 0$, for $d^2 \leq T$ a sequence of a d-dimensional vectors $x_{t,a}$, such that*

$$\mathbb{E}\left[\sum_{t=1}^{T} \max_a x_{t,a}^T \theta^* - \sum_{t=1}^{T} r_{t,a_t}\right] \geq \gamma\sqrt{Td}.$$

Proof. We can reduce the problem to the classic multi-armed bandit problem, and find a useful setting according to the results there. □

5.6 Generalized linear bandits

A generalized linear model is a probabilistic model where observation Y conditioned on feature vector $x \in \mathbb{R}^d$ has an exponential-family distribution with mean $\mu(x^T\theta)$, where θ is the unknown parameter.

Assume the probability density function of Y conditioned on $x.\theta$ is

$$p(y|x,\theta) = \exp\{y(x^T\theta) + c(y) - b(x^T\theta)\},$$

then we know the mean value of Y is $\mu = \dot{b}$. Because

$$\begin{aligned}
\int f = 1 \Rightarrow & \int \exp\{y(x^T\theta) + c(y)\} = e^{b(x^T\theta)} \\
& \Rightarrow \int y \exp\{y(x^T\theta) + c(y)\} = \dot{b}(x^T\theta) e^{b(x^T\theta)} \\
& \Rightarrow \int y \exp\{y(x^T\theta) + c(y) - b(x^T\theta)\} = \dot{b}(x^T\theta) \\
& \Rightarrow \mathbb{E}[Y|x,\theta] = \mathbb{E}[Y|x^T\theta] = \dot{b}(x^T\theta)
\end{aligned}$$

In the same way we can show $\mathrm{Var}(Y|x,\theta) = \ddot{b}(x^T\theta)$.

The negative log likelihood function of $\mathcal{D} = \{(x_\ell, y_\ell)_{\ell=1}^n\}$ is

$$L(\mathcal{D};\theta) = \sum_{\ell=1}^{|\mathcal{D}|} b(x_\ell^T\theta) - y_\ell x_\ell^T\theta - c(y_\ell).$$

The gradient and Hessian of $L(\mathcal{D},\theta)$ with respect to θ are

$$\begin{aligned}
\nabla L(\mathcal{D};\theta) &= \sum_{\ell=1}^{|\mathcal{D}|} (\mu(x_\ell^T\theta - y_\ell)) x_\ell, \\
\nabla^2 L(\mathcal{D};\theta) &= \sum_{\ell=1}^{|\mathcal{D}|} \dot{\mu}(x_\ell^T\theta) x_\ell x_\ell^T.
\end{aligned}$$

The mean function μ is increasing and therefore its derivative $\dot{\mu}$ is positive. The MLE of model parameters is a vector $\theta \in \mathbb{R}^d$ such that $\nabla(L(\mathcal{D};\theta)) = 0$.

5.6.1 GLM-TSL and GLM-FPL algorithms

Algorithm 17 General randomized exploration in a generalized linear bandit.

Inputs: Number of exploration rounds τ
for t=1,. . . ,n **do**
  if $t > \tau$ **then**
    $\tilde{\theta} \leftarrow$ Randomized MLE on $\{(x_\ell, y_\ell)_{\ell=1}^{t-1}$
    $I_t \leftarrow \text{argmax}_{i \in [K]} x_i^T \tilde{\theta}_t$
  else
    Choose I_t based on $\{X_\ell\}_{\ell=1}^{t=1}$
  end if
  Pull arm I_t and get reward $Y_{I_t,t}$
  $X_t \leftarrow x_{I_t}, Y_t \leftarrow Y_{I_t,t}$
end for

Two algorithm follows the general template in Algorithm 17.

The first algorithm, GLM-TSL, is Thompson sampling where the posterior of θ is approximated by Laplace approximation. The randomized parameter vector is sampled from the Laplace approximation as

$$\tilde{\theta}_t \sim \mathcal{N}(\bar{\theta}_t, a^2 H_t^{-1}),$$

where

$$\bar{\theta}_t = \underset{\theta \in \mathbb{R}^d}{\text{argmin}}\, L(\{(X_\ell, Y_\ell)\}_{\ell=1}^{t-1}; \theta)$$
$$H_t = \sum_{\ell=1}^{t-1} \dot{\mu}(X_\ell^T \bar{\theta}_t) X_\ell X_\ell^T,$$

and $a > 0$ a tunable parameter.

In GLM-FPL, the randomized parameter vector is the MLE on the past $t-1$ rewards perturbed with Gaussian noise,

$$\tilde{\theta}_t = \underset{\theta \in \mathbb{R}^d}{\text{argmin}}\, L(\{(X_\ell, Y_\ell + Z_\ell)\}_{\ell=1}^{t-1}; \theta),$$

where $Z_\ell \sim \mathcal{N}(0, a^2)$ are normal random variables that are resampled in each round, independently of each other and history.

Posterior sampling and perturbations by Gaussian noise in linear bandits are equivalence, when both the prior of θ and rewards are normally distributed. But this equivalence no longer holds in generalied linear models. Thus GLM-TSL and GLM-FPL are two different algorithms.

5.6.2 Theoretic analysis

Let θ_* be the unknown parameter vector, $\bar{\theta}_t$ be its maximum likelyhood estimate in round t, and $\tilde{\theta}_t$ be the randomized solution in the round t. Let $G_t = \Sigma_{\ell=1}^{t-1} X_\ell X_\ell^T$.

We define

$$E_{1,t} = \left\{ \forall i \in [K] : |x_i^T \bar{\theta}_t - x_i^T \theta_*| \le c_1 \|x_i\|_{G_i^{-1}} \right\},$$

$$E_{2,t} = \left\{\forall i \in [K] : |x_i^T \tilde{\theta}_t - x_i^T \bar{\theta}| \leq c_2 \|x_i\|_{G_i^{-1}}\right\}, \text{(I think the "absolute value" should be deleted)}$$

$$E_{3,t} = \left\{\forall i \in [K] : x_i^T \tilde{\theta}_t - x_i^T \bar{\theta} > c_1 \|x_i\|_{G_i^{-1}}\right\},$$

The central part of the analysis is an upper bound on the expected per-round regret of any randomized algorithm that chooses the perturbed solution in round t as a function of its history. The corresponding lemma is stated below.

Lemma 5.31. *Let $p_2 \geq \mathbb{P}_t(\bar{E}_{2,t}), p_3 \leq \mathbb{P}_t(E_{3,t})$ and $p_3 > p_2$. Then on event $E_{1,t}$,*

$$\mathbb{E}_t[\Delta_{I_t}] \leq \dot{\mu}_{max}(c_1 + c_2)\big(1 + \frac{2}{p_3 - p_2}\big) \times \mathbb{E}_t[\|x_{I_t}\|_{G_t^{-1}}] + \Delta_{max} p_2.$$

This is a general result. In specific algorithm, we need to bound p_2 and p_3. Then we finally ready to analyze GLM-TSL and GLM-FPL.

For GLM-TSL, the regret bound is $\tilde{O}(d\sqrt{n \log K})$.

Theorem 5.32. *Assume the noise $\eta_{i,t} = Y_{i,t} - \mu(x_i^T \theta_*)$ is σ^2-sub-Gaussian.The n-round regret of GLM-TSL is bounded as*

$$Regret(n) \leq \dot{\mu}_{max}(c_1 + c_2)(1 + \frac{2}{0.15 - 1/n}) \times \sqrt{2dn \log(2n/d)} + (3n + \tau)\Delta_{max},$$

where

$$\begin{aligned} a &= c_1 \sqrt{\dot{\mu}_{max}}, \\ c_1 &= \sigma \dot{\mu}_{min}^{-1} \sqrt{d \log(n/d) + 2 \log n}, \\ c_2 &= c_1 \sqrt{2 \dot{\mu}_{min}^{-1} \dot{\mu}_{max} \log(Kn)}, \end{aligned}$$

and the number of initial exploration round τ is chosen such that

$$\lambda_{min}(G_\tau) \geq \max\{\sigma^2 \dot{\mu}_{min}^{-2}(d \log(n/d) + 2 \log n), 1\}.$$

Proof. First, we bound the probability of event $\bar{E}_{1,t}$.

Second, we bound the probabilities of events $\bar{E}_{2,t}$ and $E_{3,t}$ from above and below.

Finally, we choose the number of initial exploration rounds τ such that $\|\bar{\theta}_t - \theta_*\|_2 \leq 1$ is likely in any round $t \geq \tau$. □

And the theorem concerning GLM-FPL is almost identical to this theorem. The differences are finding a and c_2.

Theorem 5.33. *Assume the noise $\eta_{i,t} = Y_{i,t} - \mu(x_i^T \theta_*)$ is σ^2-sub-Gaussian.The n-round regret of GLM-FPL is bounded as*

$$Regret(n) \leq \dot{\mu}_{max}(c_1 + c_2)(1 + \frac{2}{0.15 - 1/n}) \times \sqrt{2dn \log(2n/d)} + (4n + \tau)\Delta_{max},$$

where

$$
\begin{aligned}
a &= c_1 \dot{\mu}_{max}, \\
c_1 &= \sigma \dot{\mu}_{min}^{-1} \sqrt{d \log(n/d) + 2 \log n}, \\
c_2 &= c_1 \dot{\mu}_{min}^{-1} \dot{\mu}_{max} \sqrt{2 \log(Kn)},
\end{aligned}
$$

and the number of initial exploration round τ is chosen such that

$$
\lambda_{min}(G_\tau) \geq \max\{4\sigma^2 \dot{\mu}_{min}^{-2}(d \log(n/d) + 2 \log n), 8a^2 \dot{\mu}_{min}^{-2} \log n, 1\}.
$$

However, the analysis is under the assumption that all feature vectors x_i have at most one non-zero entry. I think this assumption is so strong that it reduce this problem to a multi-armed bandit since different arms are irrelevant.

The proofs of technical lemmas in this paper is confusing. It uses lots of consequences from other papers but makes a little change. And the author don't write carefully why it is true. What's impressive is that it quotes martingale's theorem from linear bandit.

Another paper about generalized linear bandit proposed an algorithm just like Lin-UCB in [CLRS11]. I've only skimmed over it.

5.7 Perturbed-History Exploration in Stochastic Linear Bandits

In last paper we study GLM-FPL, but the proof is not good enough, for it assumes **feature vectors** have at most one non-zero item. So I read this paper. But don't find the answer.

Perturbed-History Exploration means when we estimate the model parameter θ^*, we don't use true history $\{(X_i, Y_i)\}_{i=1}^{t-1}$ but a mixture of history $\{(X_i, Y_i + Z_i)\}_{i=1}^{t-1}$. Note that in linear bandit setting, when Z_i is gaussian random variable, this estimate is just the same as Thompson Sampling. But here we can use other $r.v.$ such as Bernoulli. Therefore, it cannot justify their perturbation scheme as a form of posterior sampling.

Question 5.34. *Why we need Perturbed-History Exploration in case we already have OFU and TS?*

Because these designs do not extend easily to complex problems. For instance, in generalized linear bandits, OFU algorithms use approximate high-probability confidence sets, which are loop and statistically suboptimal. And the main problem is we don't have a closed form of the estimator of model parameters in complex settings such as generalized linear bandit. The posterior sampling TS don't have a closed form and need to be approximated. This is computationally.

Then this paper analyse the theory of $LinPHE$ algorithm in the **Bernoulli setting** and evaluate $LogPHE$ in a logistic model empirically without proof. The details are just like last paper so I just describe the algorithm and leave out the others.

Algorithm 18 Perturbed-history exploration in a linear bandit($LinPHE$) with [0,1] rewards.

Inputs: Integer perturbation scale $a > 0$
Regularization parameter $\lambda > 0$
for t=1,. . . ,n **do**
  if $t > d$ **then**
    Generate $(Z_{j,\ell})_{j\in[a],\ell\in[t-1]} \sim Ber(1/2)$
    $G_t \leftarrow (a+1)\sum_{\ell=1}^{t-1} X_\ell X_\ell^T + \lambda(a+1)I_d$
    $\tilde{\theta} \leftarrow G_t^{-1}\sum_{\ell=1}^{t-1} X_\ell\left[Y_\ell + \sum_{j=1}^{a} Z_{j,\ell}\right]$
    $I_t \leftarrow \text{argmax}_{i\in[K]}\, x_i^T\tilde{\theta}_t$
  else
    Choose $I_t \leftarrow K - t + 1$
  end if
  Pull arm I_t and get reward $Y_{I_t,t}$
  $X_t \leftarrow x_{I_t}, Y_t \leftarrow Y_{I_t,t}$
end for

Actually, $\tilde{\theta}_t$ is a regularized least-squares solution on the past $t-1$ rewards and $a(t-1)$ $i.i.d$ *pseudo-rewards*.

In linear bandit setting, their regret bound $\tilde{O}(d\sqrt{n})$ scales with d and n better than $LinTS$ which proves to be $\tilde{O}(d^{3/2}\sqrt{n})$. (Bayesian regret for LinTS has upper bound $\tilde{O}(d\sqrt{n})$.) Their bound does not improve over those of OFU designs, such as $LinUCB$. The improvement is in practical performance.

5.8 Old Dog Learns New Tricks: Randomized UCB for Bandit Problems

In search of potential solution to the analysis of GLM-FPL, I read this paper [VMDK19].

5.8.1 Summary of some classic and randomized strategies

In this section, blue color means advantages, and red color means drawbacks.

First, I summarize some **classic algorithms** in multi-armed and structured bandit settings.

- ε-greedy
  - Simple, widely used in practice
  - Statistically sub-optimal, does not explore in a problem dependent manner, sensitive to hyper-parameter tuning
- TS
  - When the posterior has a closed form, as in the Bernoulli or Gaussian MAB or linear bandits, it is possible to sample exactly from it. In these cases, TS is computationally efficient and have good empirical performance; near-optimal for MAB
  - Not practical in settings where there is no closed form posterior [Pan: this can be solved by Langevin Monte Carlo [XZM+22]]
  - Sub-optimal dependence on the feature dimension for structured bandits, for example $\tilde{O}(d^{3/2}\sqrt{T})$ or $\tilde{O}(d\sqrt{T\log N})$
- OFU, UCB, GLM-UCB, UCB-GLM
  - Theoretically optimal in many bandit settings, including MAB and linear bandits: $\tilde{O}(d\sqrt{T})$
  - Not practical: since these confidence sets are constructed to obtain good worst-case performance, they often have poor empirical performance on typical problem instances; in non-linear setting, confidence sets are often too conservative in practice
  - Computationally inefficient: inverting a $d \times d$ matrix; at step t, MLE is computed using $\Theta(t)$ samples, meaning that the per-step complexity grows at least linearly with t for a straightforward implementation of the algorithms

Next, I summarize some improvements and limitations of **randomized algorithms**. I focus on their application on generalized linear bandits.

Randomized algorithms: Perturbed history exploration, GLM-FPL, random-UCB ...

- Do not rely on closed form posterior distributions like TS, but they "sample" from an implicit distribution.(PHE, GLM-FPL, rand-UCB); whereas still need to solve MLE
- Near-optimal regret bounds in the general MAB setting; however, the degree of exploration is difficult to control, complicating their proofs.
- Linear bandits: closely follows that of TS and inherits its sub-optimality in the feature dimension $\tilde{O}(d^{3/2}\sqrt{T})$ or $\tilde{O}(d\sqrt{T\log K})$. (GLM-FPL, GLM-TSL)

- Proving regret bounds for the generalized linear case requires additional assumptions. (GLM-FPL)
- MLE problem cannot be solved in an efficient online manner while preserving regret guarantees, approximations do not have rigorous theoretical guarantees and add another layer of complexity to the algorithm design.(For all perturbed history algorithms)
- RandUCB match the empirical performance of TS (and often ourperforms it) and yet attains the theoretic optimal regret bounds of OFU-based algorithms, thus achieving the best of both worlds.

Both LinUCB and and RandUCB required to compute spectral norms of all actions $\|x\|_{V_{t,\lambda}^{-1}}$ in every round so that they cannot be efficiently implemented with an infinite set of arms.

Remark 5.35 (exploration in UCB)**.** *UCB- based algorithms tend to choose arms which have big* $\|x\|_{V_{t,\lambda}^{-1}}$*. This means* x *lies in the direction of eigenvectors of small eigenvalues, named* ***exploration****.*

And in order to implement UCB we need to estimate the model parameter θ^*, in that way we need to solve the MLE problem which is not efficient. All algorithms need to do this so this is not a critical problem. Or can we think this rand-UCB is the best one? I hope to answer this question in the final step.

5.8.2 The RandUCB Meta-Algorithm

When arm $i \in \mathcal{A}$ is pulled, reward distribution with mean μ_i and support $[0, 1]$. The learner's objective is to maximize its expected cumulative reward arcoss T rounds.

OFU-based strategies have the same general form: in round t, they choose the arm

$$i_t = \underset{i \in \mathcal{A}}{\operatorname{argmax}} \{\hat{\mu}_i(t) + \beta \mathcal{C}_i(t)\},$$

where $\mathcal{C}_i(t)$ is the size of the confidence interval.

As a simple modification, RandUCB randomizes the confidence intervals and chooses the arm

$$i_t = \underset{i \in \mathcal{A}}{\operatorname{argmax}} \{\hat{\mu}_i(t) + Z_t \mathcal{C}_i(t)\}.$$

Here $Z_1, \ldots, Z_T$ are *i.i.d.* samples from the *sampling distribution*. This distribution is discrete. Suppose $Z_1, \ldots, Z_T \sim Z$. $Z \in [L, U]$. Let $\alpha_1 = L, \ldots, \alpha_M = U$ denote equally spaced points in $[L, U]$, and define $p_m := \mathbb{P}(Z = \alpha_m)$.

To obtain optimal theoretical guarantees, the probabilities $p_1, \ldots, p_M$ in RandUCB must be chosen in a way that ensures $P(Z \geq \beta) > 0$.

By only considering positive values for Z (by setting $L = 0$), we maintain the OFU principle of the corresponding OFU-based algorithm.

5.8.3 Instantiating RandUCB

For multi-armed bandit, RandUCB begins by pulling each arm once and in each subsequent round $t > K$, selects

$$i_t = \underset{i}{\operatorname{argmax}} \left\{ \hat{\mu}_i(t) + Z_t \sqrt{\frac{1}{s_i(t)}} \right\}.$$

Note that if we want to prove the optimal regret bound, we need choose Z such that $\mathbb{P}(Z > \beta) > c > 0$, where OFU-based algorithm sets the constant $\beta = \sqrt{2\ln T}$. For randUCB, we choose $L = 0$ and $U = 2\sqrt{\ln T}$. I think we inflate the confidence interval to get a better exploration.

Theorem 5.36 (Minimax regret of RandUCB with coupled arms for MAB)**.** *Let* $c_1 := 1+\sqrt{\ln(KT^2)}$ *and* $c_3 := 2K\ln(1+\frac{T}{K})$*. For any* $c_2 > c_1$*, the regret* $R(T)$ *of RandUCB for MAB is bounded by*

$$(c_1+c_2)\left(1+\frac{2}{\mathbb{P}(Z>c_1)-\mathbb{P}(|Z|>c_2)}\right)\times\sqrt{c_3T}+T\mathbb{P}(|Z|>c_2)+K+1.$$

When $\mathbb{P}(Z > c_1) > c > 0$ and $|Z| \le c_2$ a.s. the regret of RandUCB can be bounded by $O(\sqrt{KT\ln(KT)})$, which is minimax-optimal up to logarithmic factors.

There is another result of MAB about Instance-dependent regret:

Theorem 5.37 (Instance-dependent regret of uncoupled RandUCB for MAB)**.** *Uncoupled RandUCB can be bouned as* $O(\Sigma_{\Delta_i>0}\Delta_i^{-1})\times\left(\frac{M}{p_M}+Te^{-2\alpha_M^2}+\alpha_M^2\right)$.

I find this not so interesting and leave out the proof.

For structured bandit setting where each arm is associated with a d-dimensional feature vector, we first consider linear bandits:

Theorem 5.38 (RandUCB for linear bandits)**.** *Let* $c_1 := \sqrt{\lambda}+\frac{1}{2}\sqrt{d\ln(T+T^2/d\lambda)}$ *and* $c_3 := 2d\ln(1+\frac{T}{d\lambda})$*. For any* $c_2 > c_1$*, the regret of RandUCB for linear bandits is bounded by*

$$(c_1+c_2)\left(1+\frac{2}{\mathbb{P}(Z>c_1)-\mathbb{P}(|Z|>c_2)}\right)\times\sqrt{c_3T}+T\mathbb{P}(|Z|>c_2)+1.$$

We next consider structured bandits where the feature to reward mapping is a generalized linear model. We have $\mathbb{E}[Y_t|i_t = i] = g(\langle x_i, \theta^*\rangle) \in [0,1]$, where g is a known, strictly increasing, differentiable function called the *link* function or the *mean* function.

RandUCB for GLB starts by pulling each of the v_i(bases for action space) for $O(d\ln(T)/\mu^2\rho)$ many times such that $\|\hat{\theta}_t - \theta\| \le 1$ with probability at least $1 - 1/T$. This is because $\sum_{j=1}^{d} v_j v_j^T \ge \rho I$. While in linear bandits setting, we use regularization $M_t := \lambda I_d + \sum_{\ell=1}^{t-1} X_\ell X_\ell^T$.

Next theorem gives the promised $\tilde{O}(d\sqrt{T})$ regret bound by choosing $c_2 = 3\sqrt{\mathcal{L}}c_1$.

Theorem 5.39 (RandUCB for GLB)**.** *Let* $c_1 = \sqrt{d\ln(T/d)+2\ln)(T)}/2\mu, c_3 = 2d\ln(1+\frac{T}{d})$*. For any* $c_2 > c_1$*, the regret* $R(T)$ *of RandUCB for generalized linear bandits is bounded by*

$$(c_1+c_2/\sqrt{\mu})\left(1+\frac{2}{\mathbb{P}(Z>c_1\sqrt{\mathcal{L}})-\mathbb{P}(|Z|>c_2)}\right)\times\mathcal{L}\sqrt{c_3T}+T\mathbb{P}(|Z|>c_2)+O(d^2\ln(T)/\mu^2\rho).$$

5.8.4 Conclusion

RandUCB matches the empirical performance of TS (and often outperforms it) and yet attains the theoretically optimal regret bounds of OFU-based algorithms, thus achieving the best of both worlds.

5.9 Langevin Monte Carlo for Contextual Bandits

5.9.1 Laplace Approximation Thompson Sampling

After $t-1$ rounds of the bandit problem, assume we have collected data $\{x_1, r_1, x_2, r_2, \ldots, x_{t-1}, r_{t-1}\}$. Define the following quantities based on historical data:

$$V_t = \lambda I + \sum_{s=1}^{t-1} x_s x_s^T, b_t = \sum_{s=1}^{t-1} r_s x_s,$$

where $\lambda > 0$ is a regularization parameter.

Denote $\hat{\theta}_t = V_t^{-1} b_t$. At round t, the agent receives an action set $\mathcal{X}_t \subseteq \mathbb{R}^d$ which consists of feature vectors of candidate actions at round t.Then LinTS samples a parameter $\tilde{\theta}_t$ from distribution $\mathcal{N}(\hat{\theta}_t, v_t V_t^{-1})$ and then choose the arm as follows

$$x_t = \underset{x \in \mathcal{X}_t}{\operatorname{argmax}}\, x_t^T \tilde{\theta}_t.$$

Remark 5.40. *To understand this **Laplace Approximation** [AG13] method, we need some deduction.*

*First, assume the distribution of r_t given choosing arm x_i and parameter is $\mathcal{N}(x_i^T\theta, v^2)$. Then, if given **prior** for θ at time t is $\mathcal{N}(\hat{\theta}_t, v^2 V_t^{-1})$. To be more specific, the primary prior at time 0 is $\mathcal{N}(0, v^2\lambda^{-1} I)$. Then we can compute the **posterior** distribution at time $t+1$ as $\mathcal{N}(\hat{\theta}_{t+1}, v^2 V_{t+1}^{-1})$.*

There are two problems when using classical LinTS algorithms:

1. The Laplace approximation(Gaussian distribution) is not a good estimation for the posterior distribution when the reward distribution has more general forms than linearity;
2. Sampling from a Gaussian distribution with general covariance matrix in high dimensional problems is computationally inefficient.

Remark 5.41. *The second one is easy to understand. For the first claim, we can see later that LMS-TS approxiamtely samples from the real posterior distribution.*

5.9.2 LMC-TS algorithm

Advantages of LMC-TS:

1. Approximately samples from the true posterior distribution;
2. Computationally efficient due to
   - it only needs to sample from isotropic Gaussian $N(0, I)$;
   - it only needs to perform noisy gradient descent updates.

Algorithm 19 Langevin Monte Carlo Thompson Sampling(LMC-TS)

Input: step size $\{\eta_t > 0\}_{t\geq 1}$, inverse temperature parameters $\{\beta_t\}_{t\geq 1}$, loss function $L_t(\theta)$, and reward model function $f(x, \theta)$. $\theta_{1,0} = 0, K_0 = 0$.
for $t = 1, 2, \ldots$ **do**
  $\theta_{t,0} = \theta_{t-1,K-1}$
  for $k = 1, \ldots, K_t$ **do**
    sample a standard normal vector $\epsilon_{t,k} \sim \mathcal{N}(0, I)$
    $\theta_{t,k} = \theta_{t,k-1} - \eta_t \nabla L_t(\theta_{t,k-1}) + \sqrt{2\eta_t \beta_t^{-1}}\epsilon_{t,k}$
  end for
  Play arm $x_t = \text{argmax}_{x\in\mathcal{X}_t} f(x, \theta_{t,K_t})$ and observe reward r_t
end for

5.9.3 Implication

For *linear contextual bandits*, we can show the LMC-TS algorithm generates samples approximately from the Gaussian posterior distribution. And the algorithm is similar to TS, but is more computationally efficient.

Proposition 5.42. *If the epoch length K_t in Algorithm 5.9.2 is sufficiently large, the distribution of K_t convergences to Gaussian distribution $\mathcal{N}(V_t^{-1}b_t, \beta_t^{-1}V_t^{-1})$ up to an arbitrary accuracy.*

Proof. Use an existing result that the distribution convergence to $\pi_t \propto \exp(-\beta_t L_t(\theta))$. □

Remark 5.43. *Note that if L is the negative log-likelihood function, this $\exp(-L)$ is the likelihood. With $\pi(\theta|r) \propto \pi(\theta)p(r|\theta)$ we can see if $\pi(\theta) = $ constant, i.e. we have no prior knowledge, the posterior $\pi(\theta|r) \propto p(r|\theta) = \exp(-L)$. This tells us the distribution convergences to the real posterior.*

For *generalized linear bandits* model, recall the negative likelihood (or density) function is

$$L(\mathcal{D}; \theta) = \sum_{\ell=1}^{|\mathcal{D}|} m(x_\ell^T\theta) - y_\ell x_\ell^T\theta - c(y_\ell),$$

where $\dot{m} = \mu$, and μ is the link function in GLM. And we can calculate the gradient and Hessian of $L(\mathcal{D}; \theta)$ with respect to θ:

$$\nabla L(\mathcal{D}; \theta) = \sum_{\ell=1}^{|\mathcal{D}|} (\mu(x_\ell^T\theta) - y_\ell)x_\ell;$$
$$\nabla^2 L(\mathcal{D}; \theta) = \sum_{\ell=1}^{|\mathcal{D}|} \dot{\mu}(x_\ell^T\theta)x_\ell x_\ell^T.$$

So the posterior of θ has the form $\pi_t \propto \exp(-\beta_t L_t(\theta))$. Similarly, the distribution of iterates θ_{t,K_t} in Algorithm 5.9.2 convergences to π_t. In other words, it convergences to the true posterior distribution if the epoch length K_t of the inner loop is sufficiently large.

For *neural contextual bandits* or *deep bandits*, where reward function $f(x, \theta^*)$ is a neural network with x as its input and θ^* as the collection of all weight matrices.

We can choose flexible loss functions based on the belief in the prior and posterior distributions to boost the empirical performance. One possible choice is

$$L_t(\theta) = \sum_{i=1}^{t-1}(f(x_i,\theta) - r_i)^2 + \lambda\|\theta\|^2,$$

where $\lambda > 0$.

Remark 5.44. *Just like regression analysis, we can use linear, log-linear, logistic, neural network ... to fit the real function. Here we can choose the $f(x,\theta)$ as a neural network to fit.*

5.9.4 Theoretical Analysis

Assumption 5.45. *There is an unknown parameter $\theta^* \in \mathbb{R}^d$ such that for any arm $x \in \mathcal{X} \subseteq \mathbb{R}^d$, the reward is $r(x) = x^T\theta^* + \xi$, where ξ is assumed to be a R-subGaussian random variable for some constant $R > 0$.*

Theorem 5.46. *Let $\delta \in (0,1)$, choose algorithm parameter as:*

$$\begin{aligned} K_j &= \kappa_j \log(3R\sqrt{2dT\log(T^3/\delta)}), \\ \beta_j^{-1} &= 4(R\sqrt{d\log(T^3/\delta)}), \\ \kappa_j &= \lambda_{max}(V_j)/\lambda_{min}(V_j). \end{aligned}$$

Then with probability $1-\delta$, it holds that

$$R(T) \le CRd\log(1/\delta)\sqrt{dT\log^3(1+T/(\lambda d))},$$

where $C > 0$ is an absolute constant that is independent of the problem.

5.10 On Frequentist Regret of Linear Thompson Sampling

I leave out proofs in this paper since I think the proof details of counter-examples is not so useful. And I find them a little boring. I'd like to read something new and interesting first.

Recall that linear Thompson sampling samples $\theta_t \sim \mathcal{N}(\hat{\theta}_t, rV_t^{-1})$, where $r = \Theta(d)$ a constant, for example $r = R\sqrt{9d\ln(T/\delta)}$ if the time horizon T is known or $r = R\sqrt{9d\ln(T/\delta)}$ if T is unknown. And

$$V_t = I + \sum_{s=1}^{t-1} x_s x_s^T$$
$$\hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t-1} r_s x_s.$$

Then $x_t = \text{argmax}_{x \in \mathcal{X}_t} \langle \theta_t, x \rangle$.

This corresponds to assuming the noise Gaussian with variance r and choosing prior $Q = \mathcal{N}(0, rI)$, see last notes or the paper [AG13] for details.

The frequentist regret of this algorithm is $\tilde{O}(d\sqrt{dn})$, which is worse than LinUCB by a factor $\sqrt{d}$. **The increase regret is caused by the choice of noise and prior distribution model, which assumes the variance is $r = \Theta(d)$ rather than $r = 1$.** The reason to do this comes from the analysis, which works by showing the algorithm is 'optimistic' with reasonable probability. Examples in this paper shows without this blowup(i.e. inflation) of variance, the regret may be linear.

We first give a sufficient condition for sub-linear regret:

Theorem 5.47. *If Lin-TS where the posterior distribution is inflated by a positive parameter ι, i.e. sample $\tilde{\Theta}_t \sim \mathcal{N}(\hat{\Theta}_t, \iota^2 \Sigma_t)$, satisfies*

$$\mathbb{P}(\sup_{A \in \mathcal{A}_t} \langle A, \tilde{\Theta}_t \rangle \geq \sup_{A \in \mathcal{A}_t} \langle A, \Theta^* \rangle | \Theta^*, \mathcal{F}_t) \geq p, \tag{5.4}$$

whenever $\|\hat{\Theta}_t - \Theta^*\|_{\Sigma_t^{-1}} \leq \rho$, we then have

$$Regret(T, \pi^{LinTS}) \leq \tilde{O}(\frac{\rho\iota}{p}\sqrt{dT}).$$

Recall

$$\sup_{A \in \mathcal{A}_t} \langle A, \tilde{\Theta}_t \rangle - \sup_{A \in \mathcal{A}_t} \langle A, \Theta^* \rangle \geq \langle A_t^*, \tilde{\Theta}_t - \Theta^* \rangle.$$

Therefore a **sufficient condition** for Eq.(5.4) is that

$$\mathbb{P}(\langle A_t^*, \tilde{\Theta}_t - \Theta^* \rangle \geq 0 | \Theta^*, \mathcal{F}_t) \geq p$$

whenever whenever $\|\hat{\Theta}_t - \Theta^*\|_{\Sigma_t^{-1}} \leq \rho$.

Remark 5.48. *This condition assumes the best arm would be over-estimated with probability larger than a constant, which would cause it be chosen. If the best arm is under-estimated, i.e. the estimated reward of the best arm is not big enough, we will choose sub-optimal arm with high probability.*

So a counter-example in which LinTS would fail must violate this sufficient condition which gives us intuition on how to construct counter-examples.

Then the paper gives two examples to show if we choose inflation parameter $\iota = 1$, i.e. choose unit variance of prior and noise, Lin-TS will suffer a linear regret.

Then the paper propose an algorithm which adapts the inflation parameter in each round. Define

$$\rho_t := \sqrt{2\log\left(\frac{\det(\Sigma_t^{-1})\det(0.1I_d)^{-\frac{1}{2}}}{0.0001}\right)} + \sqrt{d}. \tag{5.5}$$

Algorithm 20 Thompson Sampling with Adaptive Inflation (TS-AI)

Require: Inflation parameter ι and thinness threshold Ψ.
Initialize $\Sigma_1 \leftarrow \lambda I$ and $\hat{\Theta}_1 \leftarrow 0$
for $t = 1, 2 \cdots$ **do**
  Observe $\mathcal{A}_t$
  if $\phi(\Sigma_t) > \Psi$ **then**
    Sample $\tilde{\Theta}_t \sim \mathcal{N}(\hat{\Theta}_t, \rho_t^2 \Sigma_t)$ where ρ_t is defined in Eq.(5.5)
  else
    Sample $\tilde{\Theta}_t \sim \mathcal{N}(\hat{\Theta}_t, \iota^2 \Sigma_t)$
  end if
  $\tilde{A}_t \leftarrow \text{argmax}_{A \in \mathcal{A}_t} \langle A, \tilde{\Theta}_t \rangle$
  Observe reward Y_t
  $\Sigma_{t+1}^{-1} \leftarrow \Sigma_t^{-1} + \tilde{A}_t \tilde{A}_t^T$
  $\hat{\Theta}_{t+1} \leftarrow \Sigma_{t+1}(\Sigma_t^{-1}\hat{\Theta}_t + \tilde{A}_t Y_t)$
end for

This algorithm is proved to have $\tilde{O}(d\sqrt{T})$ regret. While analyzing performance of TS, they even improve OFUL algorithms by use smaller confidence sets. This do not improve the optimal regret bound but lead to improved empirical performance of the OFUL algorithm.

5.11 Thompson Sampling with Less Exploration is Fast and Optimal

First I'd like to record two concepts: **asymptotically optimal** and **minimax optimal**.

For a fixed multi-armed bandit instance, an agent has a set of K arms to play with, where each arm $i \in [K]$ is associated with a reward distribution with an unknown mean value μ_i. When T goes to infinity, the regret of any algorithm is at least

$$C(\mu)\log(T)(1-o(1)))$$

for some constant

$$C(\mu) = \sum_{i>1} \frac{\Delta_i}{KL(\mu_i, \mu_1)}.$$

A bandit algorithm is said to be *asymptotically optimal* if its regret can be uppper bounded by $C(\mu)\log(T)(1-o(1)))$ for some constant $C(\mu)$.

When the time horizon T is fixed, no algorithm can achieve a worst-case regret lower than $C\sqrt{KT}$ for some universal constant C. Here the *worst-case regret* is defined as the maximum regret of the algorithm on any possible bandit instance. A bandit algorithm that achieve the worst-case regret $O(\sqrt{KT})$ is said to be *minimax optimal*.

Introduction part in this paper is well-written. It describes pros and cons of different TS-based algorithms used in multi-armed bandit problems. Then it proposes a new algorithm:ϵ-TS, which is fast and optimal.

Algorithm 21 ϵ-Exploring Thompson Sampling

Initialize the prior distributions.

For all $i \in [K], \hat{\mu}_i(1) = 0$ and $T_i(1) = 0$.

for $t = 1, 2, \cdots, T$ **do**

For all $i \in [K]$, update the posterior, and obtain

$$a_i(t) = \begin{cases} \theta_i(t) \sim \ P^i_{Posterior} & \text{with prob.}\epsilon \\ \hat{\mu}_i(t) & \text{with prob.}1-\epsilon. \end{cases}$$

Pull the arm $A_t = \text{argmax}_{i\in[K]}\, a_i(t)$, and observe the corresponding reward r_t;

For all $i \in [K]$, $T_{i+1}(t) = T_i(t) + \mathbb{1}\{i = A_t\}$, $\hat{\mu}_i(t+1) = \dfrac{T_i(t)\hat{\mu}_i(t) + r_t\mathbb{1}\{i = A_t\}}{T_i(t+1)}$.

end for

Theorem 5.49. *For Gaussian, Bernoulli, Poisson, and Gamma reward distributions, and $\epsilon \in [1/K, 1]$, there exists a universal constant $C > 0$ such that the regret of ϵ-TS is bounded as follows:*

$$R_\mu(T) \le C(\sqrt{VKT\log(eK\epsilon)}) + 2\sum_{i>1}\Delta_i,$$

where $V = \sigma^2$ for Gaussian, $V = 1/4$ for Bernoulli, $V = \mu_1$ for Poisson and $V = \mu_1^2$ for Gamma. Moreover,

$$\lim_{T\leftarrow\infty} \frac{R_\mu(T)}{\log T} = \sum_{i>1} \frac{\Delta_i}{KL(\mu_i, \mu_1)}.$$

Remark 5.50. *For $\epsilon = 1$, this implies TS is minimax optimal up to a factor of $\sqrt{\log K}$ and is asymptotically optimal. For $\epsilon = 1/K$, this implies that $1/K$-TS is simultaneously minimax and asymptotically optimal.*

5.12 Parallelizing Thompson Sampling

This batched algorithm is intuitive. For instance, if we want to do a vaccine testing experiment, in each batch we collect some action-reward pairs then update the model. We cannot update the model parameters every time before we make a decision because of the high cost.

In batched TS algorithm, we use TS but update $\theta(a), \forall a \in [N]$ only if $k_a = 2^{l_a}$ for some arm $a \in [N]$.

Algorithm 22 Batched Thompson Sampling

Initialize: $k_a = 0, l_a = 0, \forall a \in [N]$, batch $\leftarrow \varnothing$
for $t = 1, 2, \cdots, T$ **do**
  $\theta_a(t) \sim D_a(t), \forall a \in [N]$
  $a(t) := \text{argmax}_{a \in [N]} \theta_a(t)$.
  $k_{a(t)} \leftarrow k_{a(t)} + 1$
  if $k_{a(t)} < 2^{l_{a(t)}}$ **then**
    batch $\leftarrow$ batch $\cup \{a(t)\}$
  else
    $l_{a(t)} = l_{a(t)} + 1$
    Query(batch) and recieive rewards
    Update $D_a(t)$ $\forall a \in$ batch
    batch $\leftarrow \varnothing$
  end if
end for

Theorem 5.51. *The total number of batches carried out by B-TS is at most $O(N \log T)$.*

Proof. For each arm $a \in [N]$, the number of batches related to this arm (meaning the batch ends because of $k_a = 2^{l_a}$) is at most $O(\log T)$. Now we have N arms, so the number of batches is at most $O(N \log T)$. □

Theorem 5.52 (Regret Bounds with Beta Priors)**.** *B-TS achieves the problem-dependent asymptotic optimal regret and worst-case regret $R(T) = O(\sqrt{NT \ln T})$ with $O(N \log T)$ batches.*

As we instantiate B-TS with Gaussian priors, the regret bound slightly improves.

Theorem 5.53 (Regret Bounds with Gaussian Priors.)**.** *B-TS achieves $\mathbb{E}[R(T)] = O(\sqrt{NT \ln T})$ with $O(N \log T)$ batch queries.*

Batch Minimax Optimal Thompson Sampling (B-MOTS) achieves the optimal minimax regret bound of $O(\sqrt{NT})$, as well as the asymptotic optimal regret bound for Gaussian rewards, with only $O(N \log Tbatches)$.

Remark 5.54. *Compared with TS, MOTS just clip the sample by a confidence range $(-\infty, \tau_a(t))$, where*

$$\tau_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{\alpha}{k_a(B(t))} \log^+(\frac{T}{N k_a(B(t))})}.$$

To understand this, we know TS with less exploration can be optimal.

Theorem 5.55. *B-MOTS is minimax optimal and matches the problem-dependent lower bound* $\log(T)\Sigma_{a:\Delta>0}\frac{1}{\Delta_a}$ *up to a multiplicative factor* $1/\rho$.

For linear contextual bandits, they design a batched algorithm.

Theorem 5.56. *The B-TS-C algorithm achieves the total regret of*

$$R(T) = O(d^{3/2}\sqrt{T}(\ln(T) + \sqrt{ln(T)\ln(1/\delta)})$$

with probability $1-\delta$ *with* $O(N \log T)$ *batch queries.*

5.13 An Analysis of Ensemble Sampling

For complex models, exact posterior may be computationally intractable. Ensemble sampling (ES) can serve as a practice approximation to TS.

Remark 5.57. *Langevin TS solves the problem that we might may be able to calculate the posterior but the sample step is hard to implement. It samples approximately from the real posterior. ES solves the problem that we cannot even calculate the posterior.*

This is a general framework of Ensemble Sampling:

Algorithm 23 Ensemble Sampling

Input: number of models M and prior $\mathbb{P}(\theta \in \cdot)$
Sample: $\tilde{\theta}_{0,1}, \ldots, \tilde{\theta}_{0,M} \sim \mathbb{P}(\theta \in \cdot)$
for $t = 0, 1, \ldots$ **do**
  Sample: $m_t \sim \text{unif}\{1, \ldots, M\}$
  Execute: $A_t \sim \left\{ \underset{a \in \mathcal{A}}{\text{argmax}}\, a^T \tilde{\theta}_{t,m_t} \right\}$
  Observe: R_{t+1,A_t}
  Update: $\tilde{\theta}_{t,m} \to \tilde{\theta}_{t+1,m} \forall m \in [M]$
end for

For linear bandits (though we don't need ensemble sampling for this setting because TS for linear bandit is efficient), we begin by sampling M model parameter $\tilde{\theta}_{0,1}, \ldots, \tilde{\theta}_{0,M}$ i.i.d. $\sim \mathcal{N}(\mu_0, \Sigma_0)$. (It could natural here to let $\mu_0 = 0$ and $\Sigma_0 = \sigma_0^2 I$.) Then we sample one model uniformly from M models and choose an action which maximize the expected reward under this model:

$$A_t = \underset{a \in \mathcal{A}}{\text{argmax}}\, \theta_t^T A_t$$

and observe reward R_t.

Finally, we update M parameters independently. One possible procedure is maintain a covariance matrix V_t and statistics b_t:

$$\begin{aligned} V_t &= \Sigma_0 + \sum_{i=1}^{t-1} A_t A_t^T / \sigma^2 \\ b_t &= \sum_{i=1}^{t-1} R_t A_t, \\ V_{t+1} &= V_t + A_t A_t^T / \sigma^2, \end{aligned}$$

Recall that in LinTS we use ridge regression estimator $\hat{\theta}_t = V_t^{-1} b_t$. Then sample $\tilde{\theta}_t \sim \mathcal{N}(\hat{\theta}_t, v V_t^{-1})$.

But we generate model parameters incrementally according to

$$\tilde{\theta}_{t,m} = V_t^{-1} \left(V_t \tilde{\theta}_{t-1,m} + A_t (R_t + W_{t,m}) / \sigma^2 \right),$$

for $m = 1, \ldots, M$, where $(W_{t,m})$ are m independent $\mathcal{N}(0, \sigma^2)$ random variables.

This is intuitive for linear setting, as it seems like LinTS. But for complex model we cannot derive closed form solution for $\hat{\theta}_t$. So there is another approach for understanding.

It is easy to verify the resulting parameter satisfy:

$$\tilde{\theta}_{t,m} = \operatorname*{argmin}_{\nu} \left(\frac{1}{\sigma^2} \sum_{i=1}^{t-1} (R_t + W_{t,m} - A_t^T \nu)^2 + (\nu - \tilde{\theta}_{0,m})^T \Sigma_0^{-1} (\nu - \tilde{\theta}_{0,m}) \right),$$

which admits an intuitive interpretation: each $\tilde{\theta}_{t,m}$ is a model fit to random perturbed observations (the first term) a randomly perturbed prior (the second term).

We can compare it with ridge regression here:

$$\tilde{\theta}_{t,m} = \operatorname*{argmin}_{\nu} \left(\frac{1}{\sigma^2} \sum_{i=1}^{t-1} (R_t - A_t^T \nu)^2 + \lambda \|\nu\|^2 \right)$$

Theorem 5.58. *Under ensemble sampling,*

$$Regret(T) \leq \iota \sqrt{dT\mathbb{H}(A^*)} + \eta T \sqrt{\frac{K \log(6TM)}{M}},$$

where $\iota = \sqrt{2 \left(\max_{a \in \mathcal{A}} a^T \Sigma_0 a + \sigma^2 \right)} = O(1)$ *and* $\eta = 2\sqrt{\mathbb{E}\left[\max_{a \in \mathcal{A}} (a^T \theta)^2 \right] + \sigma^2} = O(\sqrt{\min\{d, \log K\}})$.

So

$$Regret(T) \leq O\left(\sqrt{dT\mathbb{H}(A^*)} + T \sqrt{\frac{\min\{d, \log K\} K \log(6TM)}{M}} \right).$$

5.14 Learning to Optimize via Information-Directed Sampling

This paper is where IDS born. It leaves some open problems. The most important is it requires significantly more compute time.

It is worth noting that we refer to IDS as a *design principle* rather than an *algorithm*. The reason is that IDS does not specify steps to be carried out in terms of basic computational operations but only an abstract objective to be optimized.

Some notations:

$$\begin{aligned} I_t(X_1; X_2) &= D_{KL}(\mathbb{P}((X_1, X_2) \in \cdot|\mathcal{F}_t)\|\mathbb{P}(X_1 \in \cdot|\mathcal{F}_t)\mathbb{P}(X_2 \in \cdot|\mathcal{F}_t)) \\ g_t(a) &= I_t(A^*; Y_{t,a}) \\ \Delta_t(a) &= \mathbb{E}[R_{t,A^*} - R_{t,a}|\mathcal{F}_t] \\ g_t(\pi) &= \Sigma_{a\in\mathcal{A}}\pi(a)g_t(a) \\ \Delta_t(\pi) &= \Sigma_{a\in\mathcal{A}}\pi(a)\Delta_t(a). \end{aligned}$$

Note that they are random variables due to their dependence on the conditional probability measure $\mathbb{P}(\cdot|\mathcal{F}_t)$. Let $\mathcal{D}(\mathcal{A})$ denote the set of probability distributions over $\mathcal{A}$.

The policy $\pi^{IDS} = (\pi_1^{IDS}, \pi_2^{IDS}, \ldots)$ is defined by

$$\pi_t^{IDS} = \underset{\pi\in\mathcal{D}(\mathcal{A})}{\operatorname{argmin}} \left\{ \Psi_t(\pi) := \frac{\Delta_t(\pi)^2}{g_t(\pi)} \right\}. \tag{5.6}$$

Remark 5.59. *Consider TS algorithm:* $\mathbb{P}(A_t \in \cdot|\mathcal{F}_t) = \mathbb{P}(A^* \in \cdot|\mathcal{F}_t)$. *This is trying to minimize*

$$\Delta_t(\pi) = \mathbb{E}_{a\sim\pi}\mathbb{E}[R_{t,A^*} - R_{t,a}|\mathcal{F}_t]$$

by choosing $A_t \overset{d}{=} A^*|\mathcal{F}_t$.

Remark 5.60. *Given history* $\mathcal{F}_t$, $I_t(A^*; Y_{t,a}) = H(A^*|\mathcal{F}_t) - H(A^*|\mathcal{F}_t, Y_{t,a})$. *So maximize this mutual information, is to select an action such that minimize the entropy of posterior distribution of* A^* *after observing* $Y_{t,a}$ *(i.e.* $H(A^*|\mathcal{F}_t, Y_{t,a}) = H(A^*|\mathcal{F}_{t+1})$*).*

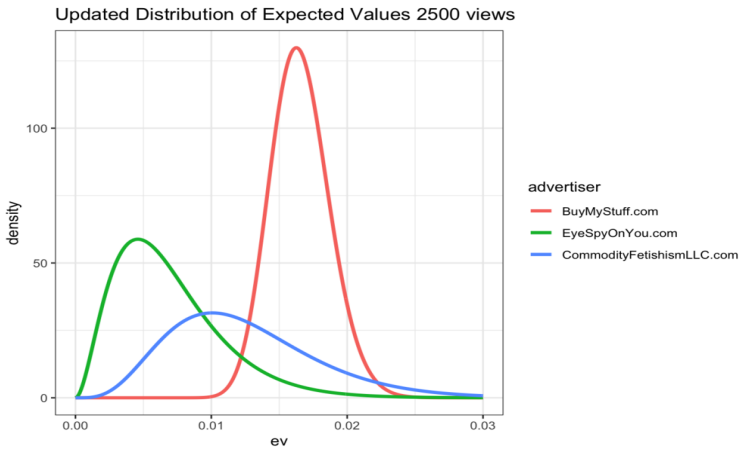

Figure 1: Reduce the uncertainty

Example 5.61 (a revealing action)**.** *Let $\mathcal{A} = \{0, 1, \ldots, K\}$ consists of $K + 1$ actions and suppose that θ is drawn uniformly at random from a finite set $\Theta = \{1, \ldots, K\}$ of K possible values. Consider a problem with bandit-feedback $Y_{t,a} = R_{t,a}$. Under θ, the reward of action a is*

$$R_{t,a} = \begin{cases} 1 & \theta = a \\ 0 & \theta \neq a, a \neq 0 \\ \dfrac{1}{2\theta} & a = 0 \end{cases}$$

Example 5.62 (Sparse linear model)**.**

Then we can establish regret bounds for information-directed sampling.
General bound:

Proposition 5.63. *For any policy $\pi = (\pi_1, \pi_2, \ldots)$ and time $T \in \mathbb{N}$,*

$$\mathbb{E}[Regret(T, \pi)] \leq \sqrt{\bar{\Psi}_T(\pi) H(A^*) T},$$

where

$$\bar{\Psi}_T(\pi) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_\pi[\Psi_t(\pi_t)].$$

Corollary 5.64. *For a deterministic $\lambda \in \mathbb{R}$ such that $\Psi_t(\pi_t) \leq \lambda$ a.e. for each $t \in \{1, \ldots, T\}$, then*

$$\mathbb{E}[Regret(T, \pi)] \leq \sqrt{\lambda H(A^*) T}.$$

Since $\Psi_t(\pi_t^{IDS}) \leq \Psi_t(\pi_t^{TS})$, where π^{TS} is the Thompson sampling policy, it is enough to bound $\Psi_t(\pi_t^{TS})$. So the implied bounds are the same as those established for TS. But IDS outperforms TS in simulation, and it is sometimes provably much more information efficient.

Assumption 5.65 (Uniformly bounded rewards)**.**

$$\sup_{y \in \mathcal{Y}} R(y) - \inf_{y \in \mathcal{Y}} R(y) \leq 1.$$

Here are some results that we can get from information theoretic analysis of TS.

Proposition 5.66 (Worst-case bound)**.** *There is always an action sampling distribution $\pi \in \mathcal{D}(\mathcal{A})$ such that $\Delta_t(\pi)^2 \leq (|\mathcal{A}|/2)\, g_t(\pi)$. In that way, $\Psi_t(\pi_t^{IDS}) \leq |\mathcal{A}|/2$ a.s. Then,*

$$\mathbb{E}[Regret(T, \pi^{IDS})] \leq \sqrt{\frac{1}{2} |\mathcal{A}| H(A^*) T}.$$

Proposition 5.67 (Full information)**.** *Under full information, $\Psi_t(\pi_t^{IDS}) \leq \dfrac{1}{2}$.*

Proposition 5.68 (Linear feedback)**.** *If $\mathcal{A} \subset \mathbb{R}^d, \Theta \subset \mathbb{R}^d$, and $\mathbb{E}[R_{t,a}|\theta] = a^T \theta$ for each action $a \in \mathcal{A}$, then $\Psi_t(\pi_t^{IDS}) \leq d/2$ a.s. .*

5.14.1 Computation methods

For concrete settings, we need some **computation methods** to implement our abstract algorithm. We will focus on the problem of generating an action A_t given the posterior distribution over θ at time t.

To implement IDS-algorithm, we need two steps:

- **Evaluating/approximate the information ratio**: some efficient algorithms can be devised to implement IDS for various problem classes, but some applications call for extremely fast computation. So we need some approximation methods.
- **Optimizing the information ratio**: choose the policy $\pi_t = \operatorname{argmin}_{\pi} \frac{(\pi^T \Delta)^2}{\pi^T g}$ and action $a_t \sim \pi_t$.

The following result shows this optimization problem is a convex optimization problem and surprisingly, has an optimal solution with at most two non-zero components.

Proposition 5.69. *For all $\Delta, g \in \mathbb{R}_+^K$ such that $g \neq 0$, the function $\pi \mapsto (\pi^T \Delta)^2 / \pi^T g$ is convex. Moreover, this function is minimized by some π^* with at most two non-zero components.*

Proof. Proof of this proposition relies on basic convex optimization and some interesting tricks. □

Algorithm 24 finiteIR(L, K, N, R, p, q)

$\Theta_a \leftarrow \{\theta | a = \operatorname{argmax}_{a'} \Sigma_y q_{\theta,a'} R(y)\}$
$p(a^*) \leftarrow \Sigma_{\theta \in \Theta_{a*}} p(\theta)$
$p_a(y) \leftarrow \Sigma_\theta p(\theta) q_{\theta,a}(y)$
$p_a(a^*, y) \leftarrow \Sigma_{\theta \in \Theta_{a*}} p(\theta) q_{\theta,a}(y)$
$R^* \leftarrow \Sigma_a \Sigma_{\theta \in \Theta_{a*}} \Sigma_y p(\theta) q_{\theta,a}(y) R(y)$
$g_a \leftarrow \Sigma_{a^*,y} p_a(a^*, y) \log \frac{p_a(a^*, y)}{p(a^*) p_a(y)}$
$\Delta_a \leftarrow R^* - \Sigma_\theta p(\theta) \Sigma_y q_{\theta,a}(y) R(y)$
Return: Δ, g

Then we discuss some useful approximation concepts. One approach to addressing this challenge is to replace integrals with sample-based estimates.

It can sometimes be helpful to replace the information ratio with alternative information measures that adequately address these issues for more specialized classes of problems.

Variance-based information ratio, which is suitable for some problems with bandit feedback, satisfies our regret bounds for such problems, and can facilitate design of more efficient numerical methods.

Use data-processing inequality and Pinsker's inequality to get

$$g_t(a) \geq 2 \operatorname{Var}(\mathbb{E}_t[R_{t,a} | A^*]) := 2 v_t(a).$$

So we can define variance-based information ratio by

$$\frac{(\pi^T \Delta)^2}{\pi^T v}.$$

Algorithm 25 SampleIR($K, q, R, M, \theta^1, \ldots, \theta^M$)

$\hat{\Theta}_a \leftarrow \{m | a = \text{argmax}_{a'} \Sigma_y q_{a',\theta^m} R(y)\}$
$\hat{p}(a^*) \leftarrow |\hat{\Theta}_{a^*}|/M$
$\hat{p}_a(y) \leftarrow \Sigma_m q_{\theta^m,a}(y)/M$
$\hat{p}_a(a^*, y) \leftarrow \Sigma_{m \in \hat{\Theta}_{a*}} q_{a,\theta^m}(y)/M$
$\hat{R}^* \leftarrow \Sigma_{a,y} \hat{p}_a(a, y) R(y)$
$g_a \leftarrow \Sigma_{a^*,y} \hat{p}_a(a^*, y) \log \frac{\hat{p}_a(a^*, y)}{\hat{p}(a^*)\hat{p}_a(y)}$
$\Delta_a \leftarrow R^* - M^{-1} \Sigma_m \Sigma_y q_{\theta^m,a}(y) R(y)$
Return: Δ, g

And actions with high variance $v_t(a)$ must yield substantial information about which action is optimal.

The next proposition establishes that variance-based IDS satisfies the bounds on the information ratio given before.

Proposition 5.70. *Suppose* $\sup_y R(y) - \inf_y R(y) \leq 1$ *and*

$$\pi_t \in \underset{\pi}{\text{argmin}} \frac{\Delta_t(\pi)^2}{v_t(\pi)},$$

then $\Psi_t(\pi_t) \leq |\mathcal{A}|/2$. *And for linear setting with feature dimension* d, $\Psi_t(\pi_t) \leq d/2$.

Drawbacks:

- Computationally demanding;
- Whether IDS attains the lower bound;
- Understanding of information complexity, whether information is a right form;
- Derive lower bounds use information theoretic methods.

5.14.2 Extensions

There are some extension problems where IDS can be used.

- Pure exploration: it seems strange to minimize $\mathbb{E}[\min_{a \in \mathcal{A}} \Delta_T(a)]$;
- Use information gain about θ;
- A tunable version of IDS.

5.15 An Information-Theoretic Analysis of Thompson Sampling

In order to better understand this set of methods, I think I must read this paper that originally proposed the information ratio.

Definition 5.71 (Information ratio)**.**

$$\Gamma_t = \frac{\mathbb{E}_t[R(Y_{t,A^*}) - R(Y_{t,A_t})]^2}{I_t(A^*; (A_t, Y_{t,A_t}))}$$

.

Proposition 5.72 (General regret bound)**.** *For any $T \in \mathbb{N}$, if $\Gamma_t \leq \bar{\Gamma}$ a.s. for each $t \in \{1, \ldots, T\}$, then*

$$\mathbb{E}[Regret(T, \pi^{TS})] \leq \sqrt{\bar{\Gamma} H(A^*) T}.$$

5.16 Frequentist IDS algorithms

I sort out frequentist Information-Directed Sampling methods through these papers [KK18, KLVS21] and this thesis [Kir21].

Here are examples where classical UCB and TS algorithms fail to gain information.

Example 5.73 (Choice between low and high noise.)**.** *Action space $\mathcal{A} = \mathcal{S} \times \{\rho_1, \rho_2\}$, where $0 < \rho_1 < \rho_2 < \infty$. Just copy any action to get a low noise version and a high noise version. In this setting, choosing (s, ρ_1) over (s, ρ_2) yields a more efficient solution. But UCB and TS fain to do this.*

Example 5.74 (Extreme linear bandits)**.** *Take a linear bandit problem in $\mathbb{R}^d$, and add a set of d basis vectors to the action set, such that observations from these actions have no (or infinitesimal small) observation noise. By scaling the new basis vectors, one can also achieve that none of the new actions is optimal. Still, if these actions are played first, one can have vanishing regret after d steps, but again UCB and Thompson Sampling fail to take these actions.*

Next we introduce a frequentist framework:

$$\forall \lambda \in \mathbb{R},\ \mathbb{E}[e^{\lambda \epsilon_t} | \mathcal{F}_{t-1}, x_t] \le \exp(\frac{\lambda^2 \rho_t^2}{2})$$

$$\gamma_T = \operatorname*{ess\,sup}_{\mathcal{F}_T} \sum_{t=1}^{T} I_t(x_t)$$

$$\Psi_t(\mu) = \frac{\mathbb{E}_\mu[\Delta(x)|\mathcal{F}_{t-1}]^2}{\mathbb{E}_\mu[I_t(x)|\mathcal{F}_{t-1}]}$$

For a fixed policy π with sampling distribution π_t, we write $\Psi_t = \Psi(\pi_t)$, and with slight abuse of notation, we define $\Psi_t(x) = \Psi_t(\delta_x)$.

Theorem 5.75 (Regret bound for randomized policies)**.** *Let $S = \max_{x \in \mathcal{X}} \Delta(x)$, and let Ψ_t, γ_t as defined above. Then, for any policy, with probability at least $1 - \delta$, at any time $T \ge 1$, it holds that*

$$R_T \le \frac{5}{4} \sqrt{\sum_{t=1}^{T} \Psi_t \left(2\gamma_T + 4\gamma_T \log \frac{2}{\delta} + 8\gamma_T \log(4\gamma_T) + 1 \right)} + 4S \log \left(\frac{8\pi^2 T^2}{3\delta} (\log(T) + 1) \right).$$

If also $I_t(x_t) \le 1$ holds for all $t \ge 1$, then with probability at least $1 - \delta$, at any time $T \ge 1$,

$$R_T \le \frac{5}{4} \sqrt{\sum_{t=1}^{T} \Psi_t \left(2\gamma_T + 4 \log \frac{2}{\delta} + 8 \log(4) + 1 \right)} + 4S \log \left(\frac{8\pi^2 T^2}{3\delta} (\log(T) + 1) \right).$$

Proof. We can rewrite the regret

$$R_T = \sum_{t=1}^{T} \Delta_t = \sum_{t=1}^{T} \mathbb{E}[\Delta_t | \mathcal{F}_{t-1}] + \sum_{t=1}^{T} (\Delta_t - \mathbb{E}[\Delta_t | \mathcal{F}_{t-1}]).$$

Then bound the first term like before in Bayesian setting, and the second term use a new martingale difference sequence concentration inequality.

For the first part, note that

$$\sum_{t=1}^{T} \mathbb{E}[\Delta_t|\mathcal{F}_{t-1}] = \sum_{t=1}^{T} \sqrt{\Psi_t}\sqrt{\mathbb{E}[I_t|\mathcal{F}_{t-1}]} \leq \sqrt{\sum_{t=1}^{T} \Psi_t \sum_{t=1}^{T} \mathbb{E}[I_t|\mathcal{F}_{t-1}]}$$

.

Again, we need to bound a martingale difference sequence $\mathbb{E}[I_t|\mathcal{F}_{t-1}] - I_t$ such that we can make use of $\sum_{t=1}^{T} I_t \leq \gamma_T$ to bound the term $\mathbb{E}[I_t|\mathcal{F}_{t-1}]$. □

So we need the following lemma to show that, for any non-negative stochastic process X_t, with high probability the sum of conditional means $\sum_{t=1}^{T} \mathbb{E}[X_t|\mathcal{F}_{t-1}]$ is not much larger than $\sum_{t=1}^{T} X_t$.

Lemma 5.76 (Concentration on conditional mean)**.** *Let X_t be any non-negative stochastic process adapted to a filtration $\{\mathcal{F}_T\}$, and define $m_t = \mathbb{E}[X_t|\mathcal{F}_{t-1}]$ and $M_T = \sum_{t=1}^{T} m_t$. Further assume that $X_t \leq b_t$ for a fixed, non-decreasing sequence $(b_t)_{t\geq 1}$ and let $(l_t)_{t\geq 1}$ be any fixed, positive sequence. Then, with probability at least $1-\delta$, for any $T \geq 1$,*

$$\sum_{t=1}^{T} m_t - X_t \leq \sqrt{2(b_T M_T + l_T)\log(\frac{1}{\delta}\frac{(b_T M_T + l_T)^{1/2}}{l_T^{1/2}})}.$$

Further, if $b_T \geq 1$, with probability at least $1-\delta$ for any $T \geq 1$ it holds that,

$$\sum_{t=1}^{T} m_t \leq 2\sum_{t=1}^{T} X_t + 4b_T \log\frac{1}{\delta} + 8b_T \log(4b_T) + 1$$

.

Theorem 5.77 (Regret bound for deterministic policies)**.** *Let the sequence $(x_t)_{t=1}^{T}$ be generated by any deterministic policy. Then, at any time $T \geq 1$, the regret is bounded by $R_T \leq \sqrt{\Sigma_{t=1}^{T} \Psi_t \gamma_T}$.*

Proof. Since $\Psi_t = \dfrac{\Delta(x_t)^2}{I_t(x_t)}$, we can simply apply Cauchy-Schwarz inequality. □

Frequentist Information Directed Sampling We look for policies such that the regret-information ratio is as small as possible. But in frequentist framework, $\Delta(x_t)$ can not be directly calculated. However, if a confidence band $l_t(x), u_t(x)$ is available, containing the true function values $f(x)$ with probability $1-\delta$, one can construct an upper bound $\Delta_t^+(x) = \max_{x'} u_t(x') - l_t(x)$, such that $\Delta(x) \leq \Delta_t^+$ also holds with probability $1-\delta$. Then we define a *surrogate of the regret-information ratio*

$$\Psi_t^+(\mu) = \frac{\mathbb{E}_\mu[\Delta_t^+(x)|\mathcal{F}_{t-1}]^2}{\mathbb{E}_\mu[I_t(x)|\mathcal{F}_{t-1}]}.$$

We define *Information Directed Sampling*(IDS) to be a policy π^{IDS}, which depends on the choice of information functions $(I_t)_{t\geq 1}$, such that at any time t,

$$\pi_t^{IDS} \in \operatorname*{argmin}_{\mu\in\mathcal{P}(\mathcal{X})} \Psi_t^+(\mu).$$

Define *Deterministic Information Directed Sampling*(DIDS) to be a deterministic policy which at tiem t choose an action

$$x_t^{DIDS} \in \operatorname*{argmin}_{\mu \in \mathcal{X}} \Psi_t^+(\mu).$$

We can choose information function as follows:

- $I_t^F = \log\left(1 + \frac{\sigma_t(x)^2}{\rho(x)^2}\right)$, where $\sigma_t(x)$ is the confidence widths of the reward $f(x)$. This is motivated by the Bayesian setting with Gaussian prior and likelihood, where I_t^F is up to a constant factor equal to the *conditional mutual information* $\mathbb{I}(f; x|\mathcal{F}_{t-1})$.
- $I_t^{UCB} = \log\left(\frac{\sigma_t(x_t^{UCB})^2}{\sigma_t(x_t^{UCB}|x)^2}\right)$, where we define $\sigma_t(x_t^{UCB}|x)$ as the confidence width at x_t^{UCB} after x has been evaluated.
- $I_t^{TS}(x) = \log\left(\frac{\sigma_t(x_t^{TS})^2}{\sigma_t(x_t^{TS}|x)^2}\right)$.
- $I_t^E = \frac{1}{m}\sum_{i=1}^{m} \log\left(\frac{\sigma_t(x_{t,i}^{TS})^2}{\sigma_t(x_{t,i}^{TS}|x)^2}\right)$.

Corollary 5.78. *With probability at least $1-\delta$, the regret of IDS-F and IDS-UCB is bounded by $\mathcal{O}\left(R\beta_T^\delta\sqrt{T(\gamma_T + \log\frac{1}{\delta})}\right)$, and DIDS by $R_T = \mathcal{O}\left(R\beta_T^\delta\sqrt{T\gamma_T}\right)$.*

About discussion about this regret bound, see the table in page 11 of [KK18].

5.17 Asymptotically Optimal Information-Directed Sampling

Surprisingly, results show that no algorithm based on optimism or Thompson sampling will ever achieve the optimal rate, and indeed, can be arbitrarily far from optimal, even in very simple cases. This is a disturbing result because these techniques are standard tools that are widely used for sequential optimisation.

In paper *The end of optimism*, there is a theorem stating that

Theorem 5.79.

$$\limsup_{n\to\infty} \frac{R^{UCB}}{\log(n)} = \sum_{x \in \mathcal{A}:\Delta_x > 0} \frac{2}{\Delta_x}.$$

For actions $x, z \in \mathcal{X}$, we denote by $\mathcal{H}_x^z = \{\nu \in \mathcal{M} : \langle x - z, \nu\rangle \geq 0\}$ the convex set of parameters where the reward of x is at least the reward of z. The set of *alternative parameters* is $C^* = \bigcup_{x \neq x^*} \mathcal{H}_x^{x^*}$.

Theorem 5.80 (Asymptotical lower bound)**.** *Any consistent algorithm π for the linear bandit setting with Gaussian noise has regret $R_n(\theta^*, \pi)$ at least*

$$\liminf_{n\to\infty} \frac{R_n(\theta^*, \pi)}{\log(n)} \geq c^*(\theta^*),$$

where c^* is the solution to the following convex program,

$$c^* = \inf_{\alpha} \sum_{x \in \mathcal{X}} \alpha(x) \langle x^* - x, \theta^* \rangle \ \ s.t. \ \ \min_{\nu \in C^*} \frac{1}{2} \|\nu - \theta^*\|^2_{V(\alpha)} \geq 1. \tag{5.7}$$

α is an allocation over actions, and $V(\alpha) = \sum_{x \in \mathcal{X}} \alpha(x) x x^T$.

IDS design principle

$$\mu_s = \operatorname*{argmin}_{\mu \in \mathcal{P}(\mathcal{X})} \left\{ \Psi_s(\mu) \stackrel{def}{=} \frac{\hat{\Delta}_s(\mu)^2}{I_s(\mu)} \right\}.$$

Algorithm 26 Asymptotically Optimal Information-Directed Sampling

$s \leftarrow 1$
for t=1,2,... **do**
  $V_s \leftarrow \Sigma_{i=1}^{s-1} x_i x_i^T + \mathbf{1}_d$
  $\hat{\theta}_s \leftarrow V_s^{-1} \Sigma_{i=1}^{s-1} x_i y_i$
  $\hat{x}_s \leftarrow \text{argmax}_{x \in \mathcal{X}} \langle x, \hat{\theta}_s \rangle$
  $\beta_{\delta, 1/\delta} \leftarrow (\sqrt{2 \log \delta^{-1} + \log \det(V_s)} + 1)^2$
  $\hat{\Delta}_s(x) \leftarrow \max_{z \in \mathcal{X}} \left(\langle z, \hat{\theta}_s \rangle + \beta^{1/2}_{s,s^2} \|z\|_{V_s^{-1}} \right) - \langle x, \hat{\theta}_s \rangle$
  $\hat{\nu}_s(z) \leftarrow \text{argmin}_{\nu \in \mathcal{H}_z^{\hat{x}_s}} \|\nu - \hat{\theta}_s\|^2_{V_s}$
  $m_s \leftarrow \min_{z \neq \hat{x}_s} \frac{1}{2} \|\hat{\nu}_s(z) - \hat{\theta}_s\|^2_{V_s}$
  $\eta_s \leftarrow \min_{l \leq s} m_l^{-1/2} \log(k)$
  $q_s(z) \leftarrow \exp(-\eta_s \|\hat{\nu}_s(z) - \hat{\theta}_s\|^2_{V_s})$
  $I_s(x) \leftarrow \frac{1}{2} \Sigma_{z \neq \hat{x}_s} q_s(z) \left(|\langle \hat{\nu}_s(z) - \hat{\theta}_s, x \rangle| + \beta^{1/2}_{s,s^2} \|x\|_{V_s^{-1}} \right)^2$
  if $m_s \geq \frac{1}{2} \beta_{s, t \log(t)}$ **then**
    Choose $\hat{x}_s$
  else
    $\mu_s \leftarrow \text{argmin}_{\mu \in \mathcal{P}(\mathcal{X})} \frac{\hat{\Delta}_s(\mu)^2}{I_s(\mu)}$
    Sample $x_s \sim \mu_s$, observe $y_s = \langle x_s, \theta^* \rangle + \epsilon_s$
    $s \leftarrow s + 1$
  end if
end for

Gap Estimates

All estimated quantities are defined using data collected in exploration rounds, whereas observation data from exploitation rounds is discarded. Ignoring data from exploitation rounds leads to a much more balanced data set.

Let $\hat{\theta}_s = V_s^{-1} \Sigma_{i=1}^{s-1} x_i y_i$ be the regularized least squares estimator with covariance matrix $V_s = \Sigma_{i=1}^{s-1} + \mathbf{1}_d$.

Concentration coefficient:

$$\beta^{1/2}_{s,1/\delta} = \sqrt{2 \log \delta^{-1} + \log \det(V_s)} + 1.$$

Let $\hat{\nu}_s(z) \leftarrow \text{argmin}_{\nu \in \mathcal{H}_z^{\hat{x}_s}} \|\nu - \hat{\theta}_s\|_{V_s}^2$ be the closest parameter to $\hat{\theta}_s$ in V_s-norm for which z is better than $\hat{x}_s$.

Exploitation condition:

$$m_s = \frac{1}{2} \min_{z \neq \hat{x}_s} \|\hat{\nu}_s(z) - \hat{\theta}_s\|_{V_s}^2 \geq \frac{1}{2} \beta_{s,t \log(t)}.$$

The gap estimate:

$$\hat{\Delta}_s(x) \leftarrow \max_{z \in \mathcal{X}} \left(\langle z, \hat{\theta}_s \rangle + \beta_{s,s^2}^{1/2} \|z\|_{V_s^{-1}} \right) - \langle x, \hat{\theta}_s \rangle.$$

Then with high probability, $\Delta(x) \leq 2\hat{\Delta}_s(x)$. Let $\delta_s := \hat{\Delta}_s(\hat{x}_s)$, then $\hat{\Delta}_s(x) = \langle \hat{x}_s - x, \hat{\theta}_s \rangle + \delta_s$. We also refer to δ_s as the *estimation error*. The UCB action is $x_s^{UCB} := \text{argmax}_{x \in \mathcal{X}} \langle x, \hat{\theta}_s \rangle + \beta_{s,s^2}^{1/2} \|x\|_{V_s^{-1}}$.

Information Gain

The *information gain* is set to

$$I_s(x) := \frac{1}{2} \Sigma_{z \neq \hat{x}_s} q_s(z) \left(|\langle \hat{\nu}_s(z) - \hat{\theta}_s, x \rangle| + \beta_{s,s^2}^{1/2} \|x\|_{V_s^{-1}} \right)^2,$$

where the mixing distribution $q_s \in \mathcal{P}(\mathcal{X})$ is defined so that

$$q_s(z) \propto \begin{cases} 0 & \text{if } z = \hat{x}_s \\ \exp\left(-\frac{\eta_s}{2} \|\hat{\nu}_s(z) - \hat{\theta}_s\|_{V_s}^2 \right) & \text{otherwise} \end{cases}$$

The learning rate is $\eta_s \leftarrow \min_{l \leq s} m_l^{-1/2} \log(k)$, where $m_s \leftarrow \min_{z \neq \hat{x}_s} \frac{1}{2} \|\hat{\nu}_s(z) - \hat{\theta}_s\|_{V_s}^2$.

Regret Bounds

Theorem 5.81 (Worst-case regret). *The regret of Algorithm 26 is bounded by*

$$R_n \leq \mathcal{O}(d\sqrt{n} \log)(n).$$

Theorem 5.82 (Gap-dependent bound). *The regret of Algorithm 26 is bounded by*

$$R_n \leq \mathcal{O}(\Delta_{min}^{-1} d^3 \log(n)^2).$$

Theorem 5.83 (Asymptotic regret). *Algorithm 26 is asymptotically optimal,*

$$\lim_{n \to \infty} \frac{R_n}{\log(n)} = c^*,$$

where c^* is the solution to the lower bound 5.8 and we assume that $\|x^*\| > 0$.

5.18 Improved Self-Normalized Concentration in Hilbert Spaces: Sublinear Regret for GP-UCB

In this problem, there is some unknown function $f^* : \mathcal{X} \to \mathbb{R}$ of bounded norm in H, where $\mathcal{X} \in \mathbb{R}^d$ is a bounded set. Feedback $Y_t = f^*(X_t) + \epsilon_t$.

Assumption 5.84. *There is some constant D such that $\|f^*\|_H \leq D$ and for every $t \geq 1$, ϵ_t is σ-subGaussian condition on $\sigma(Y_{1:t-1}, X_{1:t})$.*

Fact 5.85. *Let M_t be a non-negative supermartingale with respect to some filtration. Suppose $\mathbb{E}M_0 = 1$. Then, for any $\delta \in (0,1)$, we have*

$$\mathbb{P}(\exists t \geq 0 : M_t \geq \frac{1}{\delta}) \leq \delta.$$

An identical result holds in infinite-dimensional Hilbert spaces:

Theorem 5.86 (Self-normalized concentration in Hilbert spaces)**.** *Defining $S_t := \Sigma_{s=1}^t \epsilon_s f_s$ and $V_t := \Sigma_{s=1}^t f_s f_s^T$, we have that for any $\rho > 0$, the process $(M_t)_{t\geq 0}$ defined by*

$$M_t := \frac{1}{\sqrt{\det(id_H + \rho^{-1}V_t)}} \exp\left\{\frac{1}{2}\|(\rho id_H + V_t)^{-1/2} S_t/\sigma\|_H^2\right\}$$

is a nonnegative supermartingale with respect to $(\mathcal{F}_t)_{t\geq 0}$. Consequently, by the last Fact, for any $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for all $t \geq 1$, we have

$$\|(V_t + \rho id_H)^{-1/2} S_t\|_H \leq \sigma\sqrt{2\log\left(\frac{1}{\delta}\sqrt{\det(id_H + \rho^{-1}V_t)}\right)}.$$

Sketch of proof. In infinite dimension Euclidean spaces, the result holds. Then use methods of taking the limit.

□

Algorithm 27 Gaussian Process Upper Confidence Bound (GP-UCB)

Input: regularization parameter $\rho > 0$, norm bound D, confidence bounds $(U_t)_{t\geq 1}$, and time horizon T.
$V_0 = \rho id_H, f_0 = 0, \mathcal{E}_0 = \{f \in H : \|f\|_H \leq D\}$
for $t = 1, \ldots, T$ **do**
  Let $(X_t, \tilde{f}_t) := \text{argmax}_{x\in\mathcal{X}, f\in\mathcal{E}_{t-1}} (f, k(\cdot, x))_H$
  Play action X_t and observe reward $Y_t := f^*(X_t) + \epsilon_t$
  Set $V_t = V_{t-1} + k(\cdot, X_t)k(\cdot, X_t)^T$ and $f_t = (V_t + \rho id_H)^{-1}\Phi_t^T Y_{1:t}$
  Set $\mathcal{E}_t = \{f \in H; \|(V_t + \rho id_H)^{1/2}(f_t - f)\|_H \leq U_t\}$
end for

Theorem 5.87. *With probability at least $1-\delta$, the regret of Algorithm above satisfies*

$$R_T = \mathcal{O}\left(\gamma_T(\rho)\sqrt{T} + \sqrt{\rho\gamma_T(\rho)T}\right).$$

If the kernel k experiences (C,β)-polynomial eigendecay for some $C > 0$ and $\beta > 1$, taking $\rho = T^{\frac{1}{1+\beta}}$ yields $R_T = \tilde{O}(T^{\frac{3+\beta}{2+2\beta}})$, which is always sub-linear in T.

Corollary 5.88. *Matérn kernel with smoothness $\nu > 1/2$ in dimension d experiences $\left(C, \frac{2\nu + d}{d}\right)$-eigendecay, for some constant $C > 0$. Thus, GP-UCB obtains a regret rate of $R_T = \tilde{O}\left(T^{\frac{\nu+2d}{2\nu+2d}}\right)$.*

Contributions:

- First, we show how to extend self-normalized concentration inequalities for finite-dimensional, Euclidean spaces directly to the case of Hilbert spaces through carefully making a truncation argument.
- Second, we demonstrate the importance of regularization in the kernelized bandit problem.

5.19 The End of Optimism? An Asymptotic Analysis of Finite-Armed Linear Bandits

We analyse the asymptotic regret and show matching upper and lower bounds on what is achievable. Surprisingly, our results show that no algorithm based on optimism or Thompson sampling will ever achieve the optimal rate. In fact, they can be arbitrarily far from optimal, even in very simple cases. This is a disturbing result because these techniques are standard tools that are widely used for sequential optimisation, for example, generalised linear bandits and reinforcement learning.

This paper mentions that frequentist IDS may be a promising candidate to overcome the failures of optimism and Thompson sampling.

Definition 5.89 (Consistent)**.** *A policy is consistent if for all θ and $p > 0$ it holds that $R_\theta^\pi(n) = O(n^p)$ or $o(n^p)$.*

Assuming consistency is required to rule out policies that are defined to always play a fixed action x^*, which incurs zero regret when x^* is indeed optimal, but linear regret on other instances.

Theorem 5.90 (Asymptotically lower bound)**.** *Let π be a consistent policy, $\theta \in \mathbb{R}^d$ such that there is a unique optimal arm in $\mathcal{A}$. Let*

$$\bar{G}_n^{-1} = \mathbb{E}\left[\sum_{t=1}^n A_t A_t^T\right].$$

Then

$$\limsup_{n\to\infty} \log(n)\|x\|^2_{\bar{G}_n^{-1}} \leq \frac{\Delta_x^2}{2},$$

and also

$$\liminf_{n\to\infty} \frac{R_\theta^\pi(n)}{\log(n)} \geq c(\mathcal{A}, \theta),$$

where $c(\mathcal{A}, \theta)$ is the solution to the optimisation problem:

$$\inf_{\alpha \in [0,\infty)^{\mathcal{A}}} \sum_{x \in \mathcal{A}^-} \alpha(x)\Delta_x \text{ subject to}$$
$$\|x\|^2_{H^{-1}(\alpha)} \leq \frac{\Delta_x^2}{2}, \forall x \in \mathcal{A}^-,$$

where $H(\alpha) = \Sigma_{x\in\mathcal{A}}\alpha(x)xx^T$ and $\mathcal{A}^- = \mathcal{A} - \{x^*\}$.

Remark 5.91. *The intuition underlying the optimisation problem is that no consistent strategy can escape allocating samples so that the gaps of all suboptimal actions are identified with high confidence, while a good strategy will also minimise the regret subject to the identifiability condition.*

In the former paper, c^* is the solution to the following convex program,

$$c^* = \inf_\alpha \sum_{x\in\mathcal{X}} \alpha(x)\langle x^* - x, \theta^*\rangle \text{ s.t. } \min_{\nu \in C^*} \frac{1}{2}\|\nu - \theta^*\|^2_{V(\alpha)} \geq 1. \tag{5.8}$$

α is an allocation over actions, and $V(\alpha) = \sum_{x\in\mathcal{X}} \alpha(x)xx^T$.

Remark 5.92. *This form has the intuition that choose arms to discriminate the true parameter and alternative parameter.*

Example 5.93 (Finite armed bandits)**.** *Suppose $k = d$ and $\mathcal{A} = \{e_1, \dots, e_k\}$ be the standard basis vectors. Then*

$$c(\mathcal{A}, \theta) = \sum_{x \in \mathcal{A}: \Delta_x > 0} \frac{2}{\Delta_x},$$

which recovers the lower bound for MAB.

Example 5.94. *Let $\alpha > 1$ and $d = 2$ and $\mathcal{A} = \{x_1, x_2, x_3\}$ with $x_1 = (1, 0)$ and $x_2 = (0, 1)$ and $x_3 = (1 - \epsilon, \alpha\epsilon)$ and $\theta = (1, 0)$. Then $c(\mathcal{A}, \theta) = 2\alpha^2$ for all sufficiently small ϵ. And $c(\mathcal{A} - \{x_2\}, \theta) = 2\epsilon^{-1} \gg c(\mathcal{A}, \theta)$. Determining which of x_1 and x_3 is optimal is easy by playing x_2.*

5.20 Langevin Thompson Sampling with Logarithmic Communication: Bandits and Reinforcement Learning

This paper use Langvin Monte Carlo in TS, and batched methods to solve bandits and RL problems.

5.21 Bandits with heavy tail/ sub-exponential rewards

5.21.1 Bandits with heavy tail

The vast majority of authors assume that the unknown distributions are sub-Gaussian, i.e.

$$\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \leq e^{\frac{v\lambda^2}{2}}.$$

A paper before assumes that there exists a convex function ψ such that for all $\lambda \geq 0$,

$$\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \leq e^{\psi(\lambda)}.$$

Then one can show the so-called ψ-**UCB** strategy satisfies the following regret guarantee:

$$R_n \leq \sum_{i:\Delta_i>0} \left(\frac{4\Delta_i}{\psi^*(\Delta_i/2)\ln n + 2} \right),$$

where ψ^* is the Legendre-Fenchel transform of ψ, defined by $\psi^*(\varepsilon) = \sup_{\lambda\in\mathbb{R}}(\lambda\varepsilon - \psi(\lambda))$.

In fact, in this paper, the author shows this bound is sub-optimal when the tails are heavier than sub-Gaussian. Then they show when the distributions are heavy tailed, under weak assumptions, regret bounds of the same form as in the sub-Gaussian case may be achieved.

To be more specific, when the reward distributions have a finite moments of order $1+\varepsilon$ for some $\varepsilon > 0$, they derive a strategy that satisfies

$$R_n \leq \sum_{i:\Delta_i>0} \left(8\left(\frac{4}{\Delta_i}\right)^{\frac{1}{\varepsilon}} \log n + 5\Delta_i \right).$$

They also prove matching lower bounds that shows that the proposed strategies are optimal up to constant factors.

For each estimator of reward mean they describe their performance in terms of *concentration to the mean* and deduce the corresponding regret bound.

We need to find new mean estimators rather than empirical mean, in order to get the concentration property in the following assumption.

Assumption 5.95. *Let $\varepsilon \in (0,1]$ be a positive parameter and let c, v be positive constants. Let $X_1, \ldots, X_n$ be i.i.d. random variables with finite mean μ. Suppose for all $\delta \in (0,1)$, there exists an estimator $\hat{\mu} = \hat{\mu}(n,\delta)$ such that with probability at least $1-\delta$,*

$$\hat{\mu} \leq \mu + v^{1/(1+\varepsilon)} \left(\frac{c\log(1/\delta)}{n} \right)^{\varepsilon/(1+\varepsilon)},$$

and also with probability at least $1-\delta$,

$$\mu \leq \hat{\mu} + v^{1/(1+\varepsilon)} \left(\frac{c\log(1/\delta)}{n} \right)^{\varepsilon/(1+\varepsilon)}.$$

Then we can propose the algorithm, **Robust UCB**:

Define the index

$$B_{i,s,t} = \hat{\mu}_{i,s,t} + v^{1/(1+\varepsilon)} \left(\frac{c\log t^2}{s} \right)^{\varepsilon/(1+\varepsilon)},$$

for $s, t \geq 1$ and $B_{i,0,t} = \infty$.

At time t, draw an arm maximizing $B_{i,T_i(t-1),t}$.

Remark 5.96. *We just change the estimator and the way to get this similar concentration result, then use UCB.*

Theorem 5.97. *The regret of* ***Robust UCB*** *satisfies*

$$R_n \leq \sum_{i:\Delta_i>0} \left(2c \left(\frac{v}{\Delta_i} \right)^{\frac{1}{\varepsilon}} \log n + 5\Delta_i \right).$$

Also, if n is sufficiently large,

$$R_n \leq n^{\frac{1}{1+\varepsilon}} \left(4Kc \log n \right)^{\frac{\varepsilon}{1+\varepsilon}} v^{1/(1+\varepsilon)}.$$

Choices of estimators:

1. Truncated empirical mean (Bernstein);
2. Median of means (Hoeffding);
3. Catoni's M estimator.

5.21.2 MAB with sub-exponential rewards

Propose new algorithms to improve UCB.

5.22 Double Explore-then-Commit: Asymptotic Optimality and Beyond

Consider *asymptotic lower bound* for multi-armed bandits. If gaps $\Delta_i (i = 1, 2, \ldots, K)$ are known to the decision maker in advance, the lower bound is

$$\liminf_{T\to\infty} \frac{R_\mu(T)}{\log T} \geq \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}.$$

When gaps $\Delta_i (i = 1, 2, \ldots, K)$ are unknown to the decision maker in advance, the asymptotic lower bound turns to

$$\liminf_{T\to\infty} \frac{R_\mu(T)}{\log T} \geq \sum_{i:\Delta_i>0} \frac{1}{2\Delta_i}.$$

For two-armed bandits, the minimax optimal rate is $\tilde{O}(\sqrt{T})$ and the instance dependent optimal rate is $\tilde{O}(\Delta + \log(T\Delta^2)/\Delta)$.

The algorithms *Double explore then commit* has four-stages in common:

1. **Explore:** Pull all arms for τ_1 rounds;
2. **Exploit:** Commit to the arm with the largest average reward;
3. **Explore:** Explore the unchosen arm in Stage 2;
4. **Exploit:** Commit to the arm with the largest average reward.

Algorithm 1, 2 in this paper deal with known/unknown gap respectively, and they can be shown to have low round complexity. Since Algorithm 2 can just guarantee asymptotically optimality, the authors combine it with a finite time optimal algorithm to ensure simutaneously(finite and infinite time) optimal.

I learn the idea that two optimal algorithm can be combined to be simultaneously minimax/instance dependent and asymptotically optimal from this paper.

To get this, one need to design an asymptotically optimal algorithm, then modify it to ensure the finite time optimality. But the combined algorithm need to be the same as the original asymptotically optimal algorithm with high probability as $T \to \infty$.

Then they convert these DETC algorithms into batched versions in known/unknown gap setting. They prove that they not only achieve the asymptotically optimal regret bound but also enjoy $\tilde{O}(1)$ round complexity.

Remark 5.98. *In this section they only focus on deriving the asymptotically optimality along with a constant round complexity in the batched bandits setting. For minimax and instance dependent regret bounds, it is proved that any algorithm achieving the minimax optimality or instance dependent optimality will cost at least* $\Omega(\log\log T)$ *or* $\Omega(\log T/\log\log T)$ *rounds respectively. How to extending simultaneously optimal Algorithm 3 to the batched bandit setting is an interesting open question.*

5.23 Best Arm Identification

Three settings about pure exploration have been considered in the literature:

1. Simple regret;
2. Best-Arm Identification with a Fixed Confidence;
3. Best-Arm Identification with a Budget.

Definition 5.99 (ε-optimal arm)**.** *An arm a is called ϵ-optimal if its expected reward is larger than the highest expected reward minus ϵ.*

Definition 5.100. *An algorithm is a (ε, δ)-PAC algorithm for the multi armed bandit with sample complexity T, if it outputs an ε-optimal arm a', with probability at least $1-\delta$, when it terminates, and the number of time steps the algorithm performs until it terminates is bounded by T .*

In the former paper: The *Naive algorithm* is a (ε, δ)-PAC algorithm with sample complexity

$$\tilde{O}\big((\frac{n}{\varepsilon^2})\log(\frac{n}{\delta})\big).$$

The *Median elimination algorithm* is a (ε, δ)-PAC algorithm with sample complexity

$$\tilde{O}\big((\frac{n}{\varepsilon^2})\log(\frac{1}{\delta})\big).$$

Successive elimination algorithm with known/unknown bias choose best arm with probability at least $1-\delta$ with sample complexity

$$\tilde{O}\bigg(\log(\frac{n}{\delta})\sum_{i=2}^{n}\frac{1}{\Delta_i^2}\bigg)$$

and

$$\tilde{O}\bigg(\sum_{i=1}^{n}\frac{\log(\frac{n}{\delta\Delta_i})}{\Delta_i^2}\bigg)$$

respectively.

Definition 5.101 (Lower bound)**.** *Let $\delta \in (0,1)$. For any δ-PAC strategy and any bandit model $\mu \in \mathcal{S}$,*

$$\mathbb{E}_\mu[\tau_\delta] \geq T^*(\mu)\log(\frac{1}{4\delta}),$$

where

$$T^*(\mu)^{-1} := \sup_{\omega}\inf_{\lambda \in Alt(\mu)}\bigg(\sum_{a=1}^{K}\omega_a KL(\mu_a, \lambda_a)\bigg).$$

Algorithm 28 Track-and-Stop

Input: δ and $\beta_t(\delta)$
Initialize: Choose each arm once and set $t = k$
while $Z_t(\delta) < \beta_t(\delta)$ **do**
  if $\min_{t\in[k]} T_i(t) \leq \sqrt{t}$ **then**
    Choose $A_{t+1} = \text{argmin}_{t\in[k]} T_i(t)$
  else
    Choose $A_{t+1} = \text{argmax}(t\hat{\alpha}_i^*(t) - T_i(t))$
  end if
  Observe reward X_{t+1}, update statistics and increment t.
end while
Return $\psi = i^*(\hat{\nu}(t))$,$\tau = t$.

Theorem 5.102. *Track-and-stop policy is sound(δ-PAC) asymptotically optimal with respect to sampling complexity.*

5.24 Sequential Batch Learning in Finite-Action Linear Contextual Bandits

5.24.1 Learning with Adversarial Contexts

In this section, authors propose a sequential batch UCB algorithm. It just merges LinUCB and batched algorithms. To validate the conditional independence assumption, it uses a master algorithm called SupSBUCB and a regression algorithm BaseSBUCB.

Remark 5.103. *The analysis of the algorithm needs a assumption that assumes for a fixed sequence of selected contexts $\{x_{t,a_t}\}_{t\in[t_m]}$ up to time t_m, the random rewards $\{r_{t,a_t}\}_{t\in[t_m]}$ are independent. However, this assumption does not hold in the vanilla version of the algorithm. This is because a future selected action a_t and hence the chosen context x_{t,a_t} depends on the previous rewards. Consequently, by conditioning on x_{t,a_t}, previous rewards, say $r_{\tau_1}, r_{\tau_2}(\tau_1, \tau_2 < t)$ can become dependent. Note the somewhat subtle issue here on the dependence of the rewards: when conditioning on x_{t,a_t}, the corresponding reward r_t becomes independent of all the past rewards $r_\tau \tau < t$. Despite this, when a future $x_{t',a_{t'}}$ is revealed $(t' > t)$, these rewards (i.e. r_t and all the rewards prior to r_t) become coupled again: what was known about r_t now reveals information about the previous rewards $\{r_\tau\}_{\tau<t}$, because r_t itself would not determine the selection of $x_{t',a_{t'}}$: all those rewards have influence over $x_{t',a_{t'}}$. Consequently, a complicated dependence structure is thus created when conditioning on $\{x_{t,a_t}\}_{t\in[t_m]}$. This lack of independence issue will be handled with a master algorithm. Using the master algorithm to decouple dependencies is a standard technique in contextual bandits that was used in [CLRS11].*

Assumption 5.104. *$K = O(poly(d))$ and $T \geq d^2$.*

Theorem 5.105. *Let T, M and d be the learning horizon, number of batches and each context's dimension, respectively. Denote by $polylog(T)$ all the poly-logarithmic factors in T.*

1. *Under assumption above, there exists a sequential batch learning algorithm $Alg = (\mathcal{T}, \pi)$, where $\mathcal{T}$ is a uniform grid defined by $t_m = \lfloor \frac{mT}{M} \rfloor$ and π is explicit defined in SupSBUCB algorithm, such that*

$$\sup_{\theta^*:\|\theta^*\|_2\leq 1} \mathbb{E}_{\theta^*}[R_T(Alg)] \leq polylog(T) \cdot \left(\sqrt{dT} + \frac{dT}{M}\right).$$

2. *Conversely, for $K = 2$ and any sequential batch learning algorithm, we have:*

$$\sup_{\theta^*:\|\theta^*\|_2\leq 1} \mathbb{E}_{\theta^*}[R_T(Alg)] \geq c \cdot \left(\sqrt{dT} + \left(\frac{T\sqrt{d}}{M} \wedge \frac{T}{\sqrt{M}}\right)\right),$$

   where $c > 0$ is a universal constant independent of (T, M, d).

Corollary 5.106. *Under adversarial contexts, $\Theta(\sqrt{dT})$ batches acieve the fully online regret $\tilde{\Theta}(\sqrt{dT})$.*

5.24.2 Learning with Stochastic Contexts

Assumption 5.107. *At each time $t \in [T]$, each context $x_{t,a}$ is drawn from $N(0, \Sigma)$, with a possible unknown covariance matrix Σ. The covariance matrix Σ satisfies $\frac{\kappa}{d} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \frac{1}{d}$ for some numerical constant $\kappa > 0$, where $\lambda_{\min}(\Sigma), \lambda_{\max}(\Sigma)$ denote the smallest and the largest eigenvalues of Σ, respectively.*

Theorem 5.108. *Let $T, M = O(\log\log T)$ and d be the learning horizon, number of batches and each context's dimension, respectively. Denote by $polylog(T)$ all the poly-logarithmic factors in T.*

1. *Under assumptions above, there exists a sequential batch learning algorithm $Alg = (\mathcal{T}, \pi)$ such that:*

$$\sup_{\theta^*:\|\theta^*\|_2\leq 1} \mathbb{E}_{\theta^*}[R_T(Alg)] \leq polylog(T)\cdot\sqrt{\frac{dT}{\kappa}}\left(\frac{T}{d^2}\right)^{\frac{1}{2(2^M-1)}}.$$

2. *Conversely, even when $K = 2$ and contexts $x_{t,a} \sim \mathcal{N}(0, I_d/d)$ are independent over all $a \in [K], t \in [T]$, for any $M \leq T$ and any sequential batch learning algorithm, we have:*

$$\sup_{\theta^*:\|\theta^*\|_2\leq 1} \mathbb{E}_{\theta^*}[R_T(Alg)] \geq c\cdot\sqrt{dT}\left(\frac{T}{d^2}\right)^{\frac{1}{2(2^M-1)}}.$$

Corollary 5.109. *Under stochastic contexts, it is necessary and sufficient to have $\Theta(\log\log(T/d^2))$ batches to achieve the fully online regret $\tilde{\Theta}(\sqrt{dT})$.*

Algorithm 29 Sequential Batch Pure-exploitation

Input: Time horizon T; context dimension d; number of batches M.
Set: $a = \Theta\left(\sqrt{T}(\frac{T}{d^2})^{\frac{1}{2(2^M-1)}}\right)$.
Grid choice: $\mathcal{T} = \{t_1, \cdots, t_M\}$, with $t_1 = ad, t_m = \lfloor a\sqrt{t_{m-1}}\rfloor$, $m = 2, 3, \cdots, M$.
Initialization: $A = \mathbf{0} \in \mathbb{R}^{d\times d}, \hat{\theta} = \mathbf{0} \in \mathbb{R}^{d\times d}$.
for $m = 1$ **to** M **do**
  for $t = t_{m-1} + 1$ **to** t_m **do**
    choose $a_t = \text{argmax}_{a\in[K]}\, x_{t,a}^T\hat{\theta}$.
    receive reward r_{t,a_t}.
  end for
  $A = A + \sum_{t=t_{m-1}+1}^{t_m} x_{t,a_t}x_{t,a_t}^T$.
  $\hat{\theta} = A^{-1}\sum_{t=t_{m-1}+1}^{t_m} r_{t,a_t}x_{t,a_t}$.
end for

To handle with the dependency of rewards, we need to design another master algorithm. But this is much easier than SupLinUCB. Instead of using all past contexts and rewards before t_m, we only use the past observations inside the time frame $T^{(m)} \subsetneq [t_m]$ to construct the estimator. The key property of the time frame is the disjointness, i.e., $T^{(1)}, \cdots, T^{(M)}$ are pairwise disjoint. The the conditional independency condition holds within each time frame $t^{(m)}$.

Algorithm 30 Batched Pure-exploitation (with sample splitting)

Input: Time horizon T; context dimension d; number of batches M.

Set: $a = \Theta\left(\sqrt{T}(\frac{T}{d^2})^{\frac{1}{2(2^M-1)}}\right)$.

Grid choice: $\mathcal{T} = \{t_1, \cdots, t_M\}$, with $t_1 = ad, t_m = \lfloor a\sqrt{t_{m-1}} \rfloor$, $m = 2, 3, \cdots, M$.

Initialization: Partition each batch into M intervals evenly, i.e., $(t_m, t_{m+1}] = \bigcup_{j=1}^{M} T_m^{(j)}$.

for $m = 1$ **to** M **do**

  if m=1 **then**

    choose $a_t = 1$ and receives reward r_{r,a_t} for any $t \in [1, t_1]$.

  else

    for $t = t_{m-1} + 1$ **to** t_m **do**

      choose $a_t = \text{argmax}_{a \in [K]} x_{t,a}^T \hat{\theta}$.

      receive reward r_{t,a_t}.

    end for

  end if

  $T^{(m)} = \bigcup_{m'=1}^{m} T_{m'}^{(m)}$.

  $A = A + \sum_{t=t_{m-1}+1}^{t_m} x_{t,a_t} x_{t,a_t}^T$.

  $\hat{\theta} = A^{-1} \sum_{t=t_{m-1}+1}^{t_m} r_{t,a_t} x_{t,a_t}$.

end for

Output: resulting policy $\pi = (a_1, \cdots, a_T)$.

References

[AG13] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013. 53, 56

[AYPS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011. 3

[BCBL13] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013. 3

[CBK+23] Souradip Chakraborty, Amrit Singh Bedi, Alec Koppel, Mengdi Wang, Furong Huang, and Dinesh Manocha. Steering: Stein information directed exploration for model-based reinforcement learning. *arXiv preprint arXiv:2301.12038*, 2023. 4

[CL15] Lijie Chen and Jian Li. On the optimal sample complexity for best arm identification. *arXiv preprint arXiv:1511.03774*, 2015. 4

[CLRS11] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011. 3, 41, 48, 84

[GK16] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016. 3

[GRGOS23] Amaury Gouverneur, Borja Rodríguez-Gálvez, Tobias J Oechtering, and Mikael Skoglund. Thompson sampling regret bounds for contextual bandits with sub-gaussian rewards. *arXiv preprint arXiv:2304.13593*, 2023. 3

[HB20] Nima Hamidi and Mohsen Bayati. On worst-case regret of linear thompson sampling. *arXiv preprint arXiv:2006.06790*, 2020. 3

[HLQ22] Botao Hao, Tor Lattimore, and Chao Qin. Contextual information-directed sampling. In *International Conference on Machine Learning*, pages 8446–8464. PMLR, 2022. 3

[HZZ+20] Yanjun Han, Zhengqing Zhou, Zhengyuan Zhou, Jose Blanchet, Peter W Glynn, and Yinyu Ye. Sequential batch learning in finite-action linear contextual bandits. *arXiv preprint arXiv:2004.06321*, 2020. 3

[JP20] Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017, 2020. 3

[JSS21] Huiwen Jia, Cong Shi, and Siqian Shen. Multi-armed bandit with sub-exponential rewards. *Operations Research Letters*, 49(5):728–733, 2021. 3

[JXXG21] Tianyuan Jin, Pan Xu, Xiaokui Xiao, and Quanquan Gu. Double explore-then-commit: Asymptotic optimality and beyond. In *Conference on Learning Theory*, pages 2584–2633. PMLR, 2021. 3, 92

[Kir21] Johannes Kirschner. *Information-Directed Sampling-Frequentist Analysis and Applications*. PhD thesis, ETH Zurich, 2021. 3, 69

[KK18] Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pages 358–384. PMLR, 2018. 3, 69, 71, 91

[KKMM23] Amin Karbasi, Nikki Lijing Kuang, Yi-An Ma, and Siddharth Mitra. Langevin thompson sampling with logarithmic communication: Bandits and reinforcement learning. *arXiv preprint arXiv:2306.08803*, 2023. 3, 91

[KLVS21] Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvári. Asymptotically optimal information-directed sampling. In *Conference on Learning Theory*, pages 2777–2821. PMLR, 2021. 3, 69, 91

[KMS21] Amin Karbasi, Vahab Mirrokni, and Mohammad Shadravan. Parallelizing thompson sampling. *Advances in Neural Information Processing Systems*, 34:10535–10548, 2021. 3

[KO21] Cem Kalkanli and Ayfer Ozgur. Batched thompson sampling. *Advances in Neural Information Processing Systems*, 34:29984–29994, 2021. 3

[KSGB19] Branislav Kveton, Csaba Szepesvari, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic linear bandits. *arXiv preprint arXiv:1903.09132*, 2019. 3

[KZS+20] Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2066–2076. PMLR, 2020. 3

[LLZ17] Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017. 3

[LS17] Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017. 3

[NOPS22] Gergely Neu, Iuliia Olkhovskaia, Matteo Papini, and Ludovic Schwartz. Lifting the information ratio: An information-theoretic analysis of thompson sampling for contextual bandits. *Advances in Neural Information Processing Systems*, 35:9486–9498, 2022. 3, 35, 36, 40, 41

[QWLVR22] Chao Qin, Zheng Wen, Xiuyuan Lu, and Benjamin Van Roy. An analysis of ensemble sampling. *Advances in Neural Information Processing Systems*, 35:21602–21614, 2022. 3

[RVR14] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Advances in Neural Information Processing Systems*, 27, 2014. 3, 93

[RVR16] Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016. 3, 33, 34

[SK23] Aadirupa Saha and Branislav Kveton. Only pay for what is uncertain: Variance-adaptive thompson sampling. *arXiv preprint arXiv:2303.09033*, 2023. 4

[SKKS09] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009. 3

[VMDK19] Sharan Vaswani, Abbas Mehrabian, Audrey Durand, and Branislav Kveton. Old dog learns new tricks: Randomized ucb for bandit problems. *arXiv preprint arXiv:1910.04928*, 2019. 3, 50

[WWR23] Justin Whitehouse, Steven Wu, and Aaditya Ramdas. Improved self-normalized concentration in hilbert spaces: Sublinear regret for gp-ucb. *arXiv preprint arXiv:2307.07539*, 2023. 3, 91

[XZM+22] Pan Xu, Hongkai Zheng, Eric V Mazumdar, Kamyar Azizzadenesheli, and Animashree Anandkumar. Langevin monte carlo for contextual bandits. In *International Conference on Machine Learning*, pages 24830–24850. PMLR, 2022. 3, 50