2024 USRP GROUP2 FINAL REPORT

Hui-Ying Shih[1], Yu-Ting Lee[2], Shr-En Chen[3]
Advisor: Pei-Yuan Wu[4]

[1] Department of Mathematics, NTHU
[2] Department of Mathematical Sciences, NCCU
[3] Department of Mathematics, HKUST
[4] Department of Electrical Engineering, NTU

Abstract

This report introduces the fundamentals of reinforcement learning with its mathematical formulation as Markov decision processes. Related analyses of a special case known as the bandit problem are also presented. In particular, we study an algorithm called OFUL and linear bandit problems in section 3, with a pronounced emphasis on regret analysis. Reproducing kernel Hilbert spaces and a sequential optimization problem that can be cast as a Gaussian process bandit are discussed in sections 4 and 5.

Contents

1. Introduction

Reinforcement learning (RL) is a type of problem and a class of solution methods concerned with strategically taking actions in a (possibly) random environment based on previous experience to maximize cumulative reward. Intuitively, RL can be thought of as a decision-making problem. We usually model the interactions between the learner and the environment as a Markov decision process (MDP). It is of great interest whether one may design an algorithm so that a learner may effectively and efficiently learn with mathematical guarantees. One of the major challenges arising in this way is the trade-off between exploration and exploitation. To maximize cumulative reward, a learner should prefer actions that have been tried in the past and found to be lucrative. The learner is said to *exploit* in the case. However, to find such actions, the learner must also *explore* in a strategic way. We discuss two *optimistic* algorithms relevant to bandit problems in sections 3 and 5 and some analyses of *regrets*, a type of performance criterion. To this end, we first give a gentle introduction to the properties of MDPs and policies in section 2.

2. Markov Decision Processes

An infinite-horizon discounted Markov decision process (MDP) $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$ is specified by:

- A state space $\mathcal{S}$. We will assume $\mathcal{S}$ is countable.
- An action space $\mathcal{A}$. We will assume $\mathcal{A}$ is finite.
- A transition probability $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ is the space of probability distributions over $\mathcal{S}$. $P(s'|s, a)$ is the probability of transitioning into state s' after taking action a in state s.
- A reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$. $r(s, a)$ is the immediate reward associated with taking action a in state s.
- A discount factor $\gamma \in [0, 1)$.
- An initial state distribution $\mu \in \Delta(\mathcal{S})$, which specifies how the initial state is generated.

A sequence $\tau_t = (s_0, a_0, r_0, ..., s_t, a_t, r_t)$ recording all interactions up to time t is called a trajectory. A policy π is a mapping from a trajectory to an action. In particular, we say a policy π is stationary if $a_t \sim \pi(\cdot|s_t)$; a policy π is deterministic and stationary if it is of the form $\pi : \mathcal{S} \to \mathcal{A}$.

We now define the V-function and Q-function by:

$$V^\pi(s) = \mathrm{E}\big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|\pi, s_0 = s\big]$$

and

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

The expectation of the V function and Q function is taken w.r.t the randomness of a trajectory, capturing the idea of expected return.

Lemma 2.1. *(Bellman consistency equations). Suppose π is a stationary policy. Then V^π and Q^π satisfy the following Bellman consistency equations: for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$,*

$$\begin{aligned}&V^\pi(s) = Q^\pi(s,\pi(s)),\\ &Q^\pi(s,a) = r(s,a) + \gamma \mathrm{E}_{s'\sim P(\cdot|s,a)}\big[V^\pi(s')\big].\end{aligned}$$

Proof. We see that the value of $V(s)$ is exactly the the sum of all $\pi(a|s)Q^\pi(s,a)$ over every possible $a \sim \pi(\cdot|s)$. Hence,

$$V^\pi(s) = \sum_{a\sim\pi(\cdot|s)} \pi(a|s)Q^\pi(s,a) = Q^\pi(s,\pi(s)).$$

For the other equality, with a direct computation,

$$\begin{aligned}Q^\pi(s,a) &= \mathrm{E}\big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)|\pi, a_0 = a, s_0 = s\big]\\ &= r(s,a) + \mathrm{E}\big[\sum_{t=1}^{\infty}\gamma^t r(s_t,a_t)|\pi, s_1 \sim P(\cdot|s,a)\big]\\ &= r(s,a) + \gamma\mathrm{E}\big[\sum_{t=0}^{\infty}\gamma^t r(s'_t,a'_t)|\pi, s'_0 = s_1\big]\\ &= r(s,a) + \gamma\mathrm{E}_{s'\sim P(\cdot|s,a)}\big[V^\pi(s')\big].\end{aligned}$$

□

Corollary 2.2. *Identifying Q^π, V^π, P^π as matrices, we once again have*

$$\begin{aligned}Q^\pi &= r + \gamma P V^\pi,\\ Q^\pi &= r + \gamma P^\pi Q^\pi.\end{aligned}$$

Proof. Note that

$$[PV^\pi]_{(s,a)} = \sum_{s'} P(s'|s,a)V^\pi(s') = \mathrm{E}_{a\sim\pi(\cdot|s),s'\sim P(\cdot|s,a)}[V^\pi(s')].$$

The first equality follows. On the other hand,

$$\begin{aligned}[P^\pi Q^\pi]_{(s,a)} &= \sum_{(s',a')} P(s'|s,a)\pi(a'|s')Q^\pi(s',a') \\ &= \sum_{s'} P(s'|s,a)V^\pi(s') \\ &= [PV^\pi]_{(s,a)}.\end{aligned}$$ □

While viewing Q^π, V^π, P^π as matrices gives great intuitions, it's a natural question whether one can consider a more general setting. In fact, we may achieve greater generality by defining the following mappings:

$$\begin{aligned} P^\pi : L^\infty(\mathcal{S}\times\mathcal{A}) &\longrightarrow L^\infty(\mathcal{S}\times\mathcal{A}) \\ Q &\longmapsto \sum_{(s',a')} P(s'|\cdot,\cdot)\pi(a'|s')Q(s',a'). \end{aligned}$$

$$\begin{aligned} P : L^\infty(\mathcal{S}) &\longrightarrow L^\infty(\mathcal{S}\times\mathcal{A}) \\ V &\longmapsto \sum_{s'} P(s'|\cdot,\cdot)V(s'). \end{aligned}$$

$$\begin{aligned} B : L^\infty(\mathcal{S}\times\mathcal{A}) &\longrightarrow L^\infty(\mathcal{S}\times\mathcal{A}) \\ Q &\longmapsto r + \gamma P^\pi(Q). \end{aligned}$$

In particular, P and P^π are elements in the dual space. We also call B the *Bellman operator.* It can be shown that B is a continuous contraction, admitting a unique fixed point.

Corollary 2.3. *Suppose π is a stationary policy. One has $Q^\pi = (I - \gamma P^\pi)^{-1}(r)$.*

Proof. We only consider the operator version of the proof. The case for matrices is easily proven.
Note that $P^\pi \in B(L^\infty(\mathcal{S}\times\mathcal{A}))$. Let $\Lambda_n = \sum_{k=0}^n \gamma^k (P^\pi)^k$. It can be verified that (Λ_n) is Cauchy in $B(L^\infty(\mathcal{S}\times\mathcal{A}))$ and we denote by Λ the limit of Λ_n. Now observe that

$$(I-\gamma P^\pi)\Lambda(x) = \lim_{n\to\infty}(I-\gamma P^\pi)\Lambda_n(x) = \lim_{n\to\infty}(I-\gamma^{n+1}(P^\pi)^{n+1}))x = x.$$

Similarly,

$$\Lambda(I-\gamma P^\pi)(x) = \lim_{n\to\infty}(I-\gamma^{n+1}(P^\pi)^{n+1})x = x.$$

This shows $(I-\gamma P^\pi)^{-1} = \Lambda$. □

Lemma 2.4. *We have:*

$$[(1-\gamma)(I-\gamma P^{\pi})^{-1}]_{(s,a),(s',a')} = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P(s_t = s', a_t = a' | s_0 = s, a_0 = a).$$

Proof. Recall that $(1+...+x^n) = (1-x^{n+1})(1-x)^{-1}$. Analogously, one has

$$I + ... + (\gamma P^{\pi})^n = (I - (\gamma P^{\pi})^{n+1})(I-\gamma P^{\pi})^{-1}.$$

Letting $n \to \infty$, we have

$$\sum_{t=0}^{\infty}(\gamma P^{\pi})^t = (I-\gamma P^{\pi})^{-1}.$$

Inductively, for $t > 2$(the equality is obvious when $t = 2$):

$$\begin{aligned}&[(P^{\pi})^t]_{(s,a),(s',a')}\\ &= \sum_{(a_{t-1},s_{t-1})} P(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}|s_0 = s, a_0 = a)P(s'|s_{t-1}, a_{t-1})\pi(a'|s')\\ &= P(s_t = s', a_t = a'|s_0 = s, a_0 = a).\end{aligned}$$

□

We also provide an operator version of lemma 2.4.

Proof. Define $\delta_{(s',a')}(s,a) = \begin{cases} 1 \ \mathit{if}\ s' = s, a' = a, \\ 0, O.W. \end{cases}$ and $\delta^*_{(s,a)}(f) = f(s,a)$. Then,

$$\begin{aligned}\delta^*_{(s,a)}[(I-\gamma P^{\pi})^{-1}(\delta_{(s',a')})] &= \langle (I-\gamma P^{\pi})^{-1}(\delta_{(s',a')}), \delta^*_{(s,a)}\rangle\\ &= \langle \Lambda(\delta_{(s',a')}), \delta^*_{(s,a)}\rangle\\ &= \lim_n \sum_{k=0}^{n}\gamma^k \langle (P^{\pi})^k(\delta_{(s',a')}), \delta^*_{(s,a)}\rangle\\ &= \lim_n \sum_{k=0}^{n}\gamma^k P(s_t = s', a_t = a'|s_0 = s, a_0 = a).\end{aligned}$$

Note that $\langle \cdot, \cdot \rangle$ denotes the duality bracket, not an inner product. □

Theorem 2.5. *Let* $\prod$ *be the set of all policies. Define*

$$V^*(s) = \sup_{\pi\in\prod} V^{\pi}(s), \ \textit{and}\ Q^*(s,a) = \sup_{\pi\in\prod} Q^{\pi}(s,a).$$

The finiteness of V^* *and* Q^* *are guaranteed since* $0 \le V^{\pi}$ *and* $Q^{\pi} \le 1/(1-\gamma)$. *Moreover, there exists a stationary and deterministic* π *such that*

$$V^*(s) = V^{\pi}(s) \ \textit{and}\ Q^*(s,a) = Q^{\pi}(s,a), \textit{for all}\ s \in \mathcal{S} \textit{and}\ a \in \mathcal{A}.$$

Proof. For each $\pi \in \prod$, we define the offset policy

$$\begin{aligned}&\pi_{(s,a,r)}(A_t = a|S_0 = s_0, A_0 = a_0, R_0 = r_0, ..., S_t = s_t)\\ &= \pi(A_tt+1 = a|S_0 = s, A_0 = a, R_0 = r, S_1 = s_0, A_1 = a_0, R_1 = r_0, ..., S_{t+1} = s_t).\end{aligned}$$

Using the Markov property with a change of variable, one has

$$\mathrm{E}\big[\sum_{t=1}^{\infty}\gamma^t r(s_t, a_t)|\pi, s_0 = s, ..., S_1 = s'\big] = \gamma\mathrm{E}\big[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)|\pi_{(s,a,r)}, S_0 = s'\big] = \gamma V^{\pi_{(s,a,r)}}(s').$$

Since for each (s, a, r), the set $\prod = \{\pi_{(s,a,r)}|\pi \in \prod\}$, we have

$$\sup_{\pi\in\prod}\mathrm{E}\big[\sum_{t=1}^{\infty}\gamma^t r(s_t, a_t)|\pi, s_0 = s, ..., S_1 = s'\big] = \gamma\sup_{\pi\in\prod}V^{\pi}(s') = \gamma V^*(s'). \tag{2.5.1}$$

We claim that $\overset{\sim}{\pi}(s) \in \arg\max_{a\in\mathcal{A}}\mathrm{E}[r(s, a)+\gamma V^*(s_1)|S_0 = s, A_0 = a]$ is an optimal deterministic stationary policy. For this, we have that:

$$\begin{aligned}V^*(s_0) &= \sup_{\pi\in\prod}\mathrm{E}\big[r(s_0, a_0) + \sum_{t=1}^{\infty}\gamma^t r(s_t, a_t)\big]\\ &= \sup_{\pi\in\prod}\mathrm{E}\big[r(s_0, a_0) + \mathrm{E}\big[\gamma^t r(s_t, a_t)|\pi, \mathbb{1}_{(S_0,A_0,R_0,S_1)=(s_0,a_0,r_0,s_1)}\big]\big]\\ &\le \sup_{\pi\in\prod}\mathrm{E}\big[r(s_0, a_0) + \sup_{\pi'\in\prod}\mathrm{E}\big[\sum_{t=1}^{\infty}\gamma^t r(s_t, a_t)|\pi', \mathbb{1}_{(S_0,A_0,R_0,S_1)=(s_0,a_0,r_0,s_1)}\big]\big]\\ &\overset{(a)}{=} \sup_{\pi\in\prod}\mathrm{E}\big[r(s_0, a_0) + \gamma V^*(s_1)\big]\\ &= \sup_{a_0\in\mathcal{A}}\mathrm{E}\big[r(s_0, a_0) + \gamma V^*(s_1)\big]\\ &\overset{(b)}{=} \mathrm{E}\big[r(s_0, a_0) + \gamma V^*(s_1)|\overset{\sim}{\pi}\big]\end{aligned}$$

where step (a) uses (2.5.1), and step (b) follows from the definition of $\overset{\sim}{\pi}$.
Recursively applying the above argument, one obtains:

$$V^*(s_0) \le \mathrm{E}\big[r(s_0, a_0) + \gamma V^*(s_1)|\overset{\sim}{\pi}\big] \le ... \le \mathrm{E}\big[r(s_0, a_0) + \gamma r(s_1, a_1) + ...|\overset{\sim}{\pi}\big] \equiv V^{\overset{\sim}{\pi}}(s_0).$$

Consequently, we now get $V^{\widetilde{\pi}}(s) = V^*(s)$ for all s, which proves our first claim.
To prove our second equation, note that $Q^* \geq Q^{\widetilde{\pi}}$. And observe that:

$$\begin{aligned}
Q^*(s_0, a_0) &= \sup_\pi \mathrm{E}\big[r(s,a) + \sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) | \pi, S_0 = s_0, A_0 = a_0, \big] \\
&= r(s,a) + \sup_\pi \mathrm{E}\big[\mathrm{E}\big[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) | \pi, \mathbb{1}_{(S_0,A_0,R_0,S_1)=(s_0,a_0,r_0,s_1)}\big]\big] \\
&\leq r(s,a) + \mathrm{E}\big[\sup_\pi \mathrm{E}\big[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) | \pi, \mathbb{1}_{(S_0,A_0,R_0,S_1)=(s_0,a_0,r_0,s_1)}\big]\big] \\
&= r(s,a) + \gamma \mathrm{E}_{s_1 \sim P(\cdot|s_0,a_0)}\big[V^*(s_1)\big] \\
&= r(s,a) + \gamma \mathrm{E}_{s_1 \sim P(\cdot|s_0,a_0)}\big[V^{\widetilde{\pi}}(s_1)\big] \\
&\overset{(c)}{=} Q^{\widetilde{\pi}}(s,a),
\end{aligned}$$

where (c) uses the fact that $\widetilde{\pi}$ is stationary. □

To aid our proof of Theorem 2.7, we introduce some definitions.
We define the Bellman optimality operator $\mathcal{T} : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \longrightarrow \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ by

$$\mathcal{T}Q = r + \gamma P V_Q,$$

and the greedy policy w.r.t. $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ by

$$\pi_Q(s) \in \arg\max_{a \in \mathcal{A}} Q(s,a).$$

For convenience, we also let

$$V_Q(s) = \max_{a \in \mathcal{A}} Q(s,a).$$

Lemma 2.6. *(contraction). For any pair of vectors Q, $Q' \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$,*

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty.$$

Proof. We denote by π_Q the greedy policy w.r.t. Q and $V_Q(s) = \max_{a \in \mathcal{A}} Q(s,a)$. WLOG, we assume $V_Q(s) \geq V_{Q'}(s)$. Then,

$$\begin{aligned}
|V_Q(s) - V_{Q'}(s)| &= Q(s, \pi_Q(s)) - Q'(s, \pi_{Q'}(s)) \\
&\leq Q(s, \pi_Q(s)) - Q'(s, \pi_Q(s)) \\
&\leq \max_a |Q(s,a) - Q'(s,a)|.
\end{aligned}$$

With the above calculations, we obtain

$$\begin{aligned}\|\mathcal{T}Q-\mathcal{T}Q'\|_\infty &= \gamma\|P(V_Q-P'_Q)\|_\infty\\ &\le \gamma\|V_Q-V_{Q'}\|_\infty\\ &= \gamma\max_s|V_Q(s)-V_{Q'}(s)|\\ &\le \gamma\max_s\max_a|Q(s,a)-Q'(s,a)|\\ &= \gamma\|Q-Q'\|_\infty.\end{aligned}$$

This shows $\mathcal{T}$ is a contraction. □

Theorem 2.7. *(Bellman optimality equation). We say a vector $Q\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ satisfies the Bellman optimality equation if:*

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

Then, for any $Q\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, we have $Q=Q^*$ if and only if Q satisfies the Bellman optimality equation. Furthermore, the deterministic policy defined by $\pi(s)\in\arg\max_{a\in\mathcal{A}}Q^*(s,a)$ is an optimal policy.
More compactly, we may rewrite theorem 2.7 as:

$$Q=Q^*\textit{ if and only if }Q\textit{ is a fixed point of }\mathcal{T}.$$

Proof. Let π^* be a deterministic stationary optimal policy. Notice that

$$V^*(s)=Q^{\pi^*}(s,\pi^*(s))\ge Q^{\pi^*}(s,a)=Q^*(s,a).$$

Also,

$$V^*(s)=Q^{\pi^*}(s,\pi^*(a))\le\max_a Q^{\pi^*}(s,a)=\max_a Q^*(s,a).$$

Thus, $V^*(s)=\max_a Q^*(s,a)$. Now we show that Q^* is a fixed point of $\mathcal{T}$.
For each $a\in\mathcal{A}$, we have:

$$\begin{aligned}Q^*(s,a)&=r(s,a)+\gamma\mathrm{E}_{s'\sim P(\cdot|s,a)}\big[V^{\pi^*}(s')\big]\\ &=r(s,a)+\gamma\mathrm{E}_{s'\sim P(\cdot|s,a)}\big[\max_a Q^*(s,a)\big].\end{aligned}$$

This proves the *only if* part. For the *if* part, recall that $\mathcal{T}$ is a contraction w.r.t.$\|\cdot\|_\infty$, and the uniqueness follows from the fixed point theorem.

We now prove the last assertion that $\pi(s) \in \arg\max_{a\in\mathcal{A}} Q^*(s,a)$ is optimal. We check that:

$$\begin{aligned} Q^\pi(s,a) &= r(s,a) + \mathrm{E}\big[\sum_{t=1}^{\infty} \gamma^t r(s_t,a_t)|\pi, s_0 = s, a_0 = a\big] \\ &\overset{(a)}{=} r(s,a) + \gamma \mathrm{E}_{s'\sim P(\cdot|s,a)}\big[\max_{a'} Q^*(s',a')\big] \end{aligned}$$

where (a) uses the definition of π. Since the Bellman optimality equation is satisfied by Q^π, we see that $Q^\pi \equiv Q^*$ and π is optimal as $V^*(s) = \max_a Q^*(s,a) = V^\pi(s)$. □

Instead of infinite-horizon MDPs, we will work with finite-horizon MDPs in Theorem 2.8. In particular, we **do not** assume the transition probability and reward function to be time-independent. We now define similar notations to those of infinite-horizon MDPs:

$$V_h^\pi(s) = \mathrm{E}\big[\sum_{t=h}^{H-1} r_t(s_t,a_t)|\pi, s_h = s\big], \text{ and } \ Q_h^\pi(s,a) = \mathrm{E}\big[\sum_{t=h}^{H-1} r_t(s_t,a_t)|\pi, s_h = s, a_h = a\big]$$

We also use $V^\pi(s) = V_0^\pi(s)$.

Theorem 2.8. *Suppose $Q_H = 0$. We have: $Q_h = Q_h^*$ for all $h \in 0,1,...,h-1$ if and only if for all $h \in 0,1,...,h-1$,*

$$Q_h(s,a) = r_h(s,a) + \mathrm{E}_{s'\sim P_h(\cdot|(s,a)}\big[\max_{a'} Q_{h+1}(s',a')\big]$$

Furthermore, $\pi(s,h) = \arg\max_a Q_h^*(s,a)$ is an optimal policy.

Proof. We shall first prove the existence of a deterministic optimal policy.
Define $\pi^*(s,h) \in \arg\max_a \mathrm{E}[r_h(s,a) + V_{h+1}^*(s')|S_h = s, A_h = a]$. For each $h \in \{0,1,...,H-1\}$, one has

$$\begin{aligned} V_h^*(s_h) &= \sup_\pi \mathrm{E}\big[\sum_{t=h}^{H-1} r_t(s_t,a_t)|\pi, S_h = s_h\big] \\ &= \sup_\pi \mathrm{E}\big[r_h(s_h,a_h) + \mathrm{E}\big[\sum_{t=h+1}^{H-1} r_t(s_t,a_t)|\pi, \mathbb{1}_{\{S_h=s_h,A_h=a_h,S_{h+1}=s_{h+1}\}}\big]\big] \\ &\le \sup_\pi \mathrm{E}\big[r_h(s_h,a_h) + V_{h+1}^*(s_{h+1})|\pi, S_h = s_h\big] \\ &= \mathrm{E}\big[r_h(s_h,a_h) + V_{h+1}^*(s_{h+1})|\pi^*, S_h = s_h\big]. \end{aligned}$$

Recursively, we now have:

$$\begin{aligned}
V_h^*(s_h) &\leq \mathrm{E}\big[r_h(s_h,a_h)+V_{h+1}^*(s_{h+1})|\pi^*,S_h=s_h\big]\\
&\leq \mathrm{E}\big[r_h(s_h,a_h)+r_{h+1}(s_{h+1},a_{h+1})+V_{h+2}^*(s_{h+2})|\pi^*,S_h=s_h\big]\\
&\vdots\\
&\leq \mathrm{E}\big[r_h(s_h,a_h)+...+r_{H-1}(s_{H-1},a_{H-1})+V_H^*(s_H)|\pi^*,S_h=s_h\big]\\
&= V_h^{\pi^*}(s_h).
\end{aligned}$$

This gives the desired $V_h^*(s)=V^{\pi^*}(s)$, for each $s\in\mathcal{S}$ and $h=0,1,...,H-1$. A similar argument gives $Q^*(s,a)=Q^{\pi^*}(s,a)$, furnishing the optimality of π^*. We proceed to prove the Bellman optimality equation.

Suppose $Q_h=Q_h^*$ and let π^* be a deterministic optimal policy. First, we have

$$V_h^*(s)=V_h^{\pi^*}(s)=Q_h^{\pi^*}(s,\pi^*(s))\geq Q_h^{\pi^*}(s,a)=Q_h^*(s,a).$$

And

$$V_h^*(s)=V_h^{\pi^*}(s)=Q_h^{\pi^*}(s,\pi^*(s))\leq \max_a Q_h^{\pi^*}(s,a)=\max_a Q_h^*(s,a).$$

Combining the above inequality, we compute:

$$\begin{aligned}
Q_h^*(s,a)&=\mathrm{E}\big[r_h(s,a)+\sum_{t=h+1}^{H-1}r_t(s_t,a_t)|\pi^*,s_h=s,a_h=a\big]\\
&=r_h(s,a)+\mathrm{E}_{s'\sim P_h(\cdot|s,a)}\big[V_{h+1}^*(s')\big]\\
&=r_h(s,a)+\mathrm{E}_{s'\sim P_h(\cdot|s,a)}\big[\max_{a'}Q_{h+1}^*(s',a')\big].
\end{aligned}$$

This is the *only if* part. On the other hand, suppose $Q_h(s,a)$ satisfies the optimality equation. Define $\pi_Q^{(h)}(s,h)\in\arg\max_a Q_h(s,a)$, Notice that $Q_H(s,\pi_Q^{(H)}(s,H))=0=\max_a Q_H^*(s,a)$. Inductively, we suppose $Q_t(s,\pi_Q^{(t)}(s,t))=\max_a Q_t^*(s,a)$ for $t=h+1,...,H$. Observe that:

$$\begin{aligned}
Q_h(s,a)&=r_h(s,a)+\mathrm{E}_{s'\sim P_h(\cdot|s,a)}\big[Q_{h+1}^{\pi_Q^{(h+1)}}(s',\pi_Q^{(h+1)}(s',h+1))\big]\\
&=r_h(s,a)+\mathrm{E}_{s'\sim P_h(\cdot|s,a)}\big[\max_{a'}Q_{h+1}^*(s',a')\big]\\
&=Q_h^*(s,a).
\end{aligned}$$

This completes the *if* part. □

Lemma 2.9. *(Q-Error Amplification). For any vector $Q\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$,*

$$V^{\pi_Q}\geq V^*-\frac{2\|Q-Q^*\|_\infty}{1-\gamma}\mathbb{1},$$

where $\mathbb{1}$ denotes a column of all ones.

Proof. Let $s \in \mathcal{S}$ and $a = \pi_Q(s)$. We have

$$\begin{aligned} V^*(s) - V^{\pi_Q}(s) &= Q^*(s, \pi^*(s)) - Q^*(s,a) + Q^*(s,a) - Q^{\pi_Q}(s,a) \\ &= Q^*(s, \pi^*(s)) - Q^*(s,a) + \gamma \mathrm{E}_{a' \sim P(\cdot|s,a)}\big[V^*(s') - V^{\pi_Q}(s')\big] \\ &\leq Q^*(s, \pi^*(s)) - Q^*(s,a) + Q(s,a) - Q(s, \pi^*(s)) \\ &\qquad + \gamma \mathrm{E}_{a' \sim P(\cdot|s,a)}\big[V^*(s') - V^{\pi_Q}(s')\big] \\ &\leq 2\|Q^* - Q\|_\infty + \gamma \|V^* - V\|_\infty. \end{aligned}$$

Therefore,

$$\|V^* - V\|_\infty \leq 2\|Q^* - Q\|_\infty + \gamma \|V^* - V\|_\infty. \qquad \square$$

Theorem 2.10. *(Q-value iteration convergence). Set $Q^{(0)} = 0$ and for $k \geq 0$, we define*

$$Q^{(k+1)} = \mathcal{T} Q^{(k)}.$$

Let $\pi^{(k)} = \pi_{Q^{(k)}}$. For $k \geq \frac{\log \frac{(1-\gamma)^2 \epsilon}{2}}{\log \gamma}$,

$$V^{(\pi^{(k)})} \geq V^* - \epsilon \mathbb{1}.$$

Proof. By Theorem 2.7, Q^* satisfies Bellman optimality equation i.e. $Q^* = \mathcal{T} Q^*$ and recall that $\mathcal{T}$ is a contraction. Hence,

$$\begin{aligned} \|Q^{(k)} - Q^*\|_\infty &= \|\mathcal{T}^{(k)} Q^{(k)} - \mathcal{T}^{(k)} Q^*\|_\infty \\ &\leq \gamma^k \|Q^{(0)} - Q^*\|_\infty = \gamma^k \|Q^*\|_\infty \\ &\leq \frac{\gamma^k}{1-\gamma}. \end{aligned}$$

Note that $\pi^{(k)}$ is a stationary, deterministic policy. Using Lemma 2.9,

$$\|V^* - V^{\pi^{(k)}}\|_\infty \leq \frac{2\|Q^* - Q^{(k)}\|_\infty}{1-\gamma} \leq \frac{2\gamma^k}{(1-\gamma)^2} < \epsilon.$$

This completes the proof with our choice of k. $\square$

Lemma 2.11. *We consider the policy iteration algorithm. Starting from an arbitrary π_0, we repeat the following procedure: for each k= 0, 1, 2,...*

(1) Policy evaluation. Compute Q^{π_k}.

(2) Policy improvement. Update the policy $\pi_{k+1} = \pi_{Q^{\pi_k}}$.

Then, we have the following inequalities:

(a) $Q^{\pi_{k+1}} \leq \mathcal{T} Q^{\pi_k} \leq Q^{\pi_k}$,

(b) $\|Q^* - Q^{\pi_{k+1}}\| \leq \gamma \|Q^* - Q^{\pi_k}\|$.

Proof. Note that

$$\mathcal{T}Q^{\pi_k}(s,a) = r(s,a) + \gamma \mathrm{E}_{a' \sim P(\cdot|s,a)}\big[\max_{a'} Q^{\pi_k}(s',a')\big] \geq Q^{\pi_k}(s,a).$$

Now, we show that $Q^{\pi_k}(s, \pi_k(s) \leq Q^{\pi_{k+1}}(s, \pi_{k+1}(s))$.

$$\begin{aligned} Q^{\pi_k}(s,\pi_k(s)) \leq Q^{\pi_k}(s,\pi_{k+1}(s)) \\ &= \mathrm{E}\big[r(s,\pi_{k+1}(s)) + \gamma \mathrm{E}_{s' \sim P(\cdot|s,\pi_{k+1}(s))}\big[Q^{\pi_k}(s',\pi_k(s'))\big]\big] \\ &\leq \mathrm{E}\big[r(s,\pi_{k+1}(s)) + \gamma \mathrm{E}_{s' \sim P(\cdot|s,\pi_{k+1}(s))}\big[Q^{\pi_k}(s',\pi_{k+1}(s'))\big]\big] \\ &= \mathrm{E}\big[r(s,\pi_{k+1}(s)) + \gamma r(s',\pi_{k+1}(s')) + \gamma^2 \mathrm{E}_{s'' \sim P(\cdot|s',\pi_{k+1}(s'))}\big[Q^{\pi_k}(s'',\pi_k(s''))\big]\big] \\ &\vdots \\ &\leq \mathrm{E}\big[r(s,\pi_{k+1}(s)) + \gamma r(s',\pi_{k+1}(s')) + ...\big] = Q^{\pi_{k+1}}(s,\pi_{k+1}(s)). \end{aligned}$$

Finally,

$$\begin{aligned} Q^{\pi_{k+1}}(s,a) &= r(s,a) + \gamma \mathrm{E}_{s' \sim P(\cdot|s,a)}\big[Q^{\pi_{k+1}}(s',\pi_{k+1}(s'))\big] \\ &\geq r(s,a) + \gamma \mathrm{E}_{s' \sim P(\cdot|s,a)}\big[Q^{\pi_k}(s',\pi_k(s'))\big] \\ &= Q^{\pi_k}(s,a) \end{aligned}$$

To prove the last assertion, we have

$$\|Q^* - Q^{\pi_{k+1}}\|_\infty \leq \|Q^* - \mathcal{T}Q^{\pi_k}\|_\infty = \|\mathcal{T}Q^* - \mathcal{T}Q^{\pi_k}\|_\infty \leq \gamma \|Q^* - Q^{\pi_k}\|_\infty. \qquad \square$$

Theorem 2.12. *(Policy iteration convergence). Let* π_0 *be an initial policy. For* $k \geq \frac{\log \frac{1}{(1-\gamma)\epsilon}}{\log \frac{1}{\gamma}}$, *the k-th policy in policy iteration has the following performance bound*

$$Q^{\pi_k} \geq Q^* - \epsilon \mathbb{1}.$$

Proof. By Lemma 2.11, we have

$$\begin{aligned} \|Q^{\pi_k} - Q^*\|_\infty &\leq \gamma \|Q^{\pi_{k-1}} - Q^*\|_\infty \\ &\leq \gamma \|\mathcal{T}^{k-1}Q^{\pi_0} - \mathcal{T}^{k-1}Q^*\|_\infty \\ &\leq \gamma^2 \|\mathcal{T}^{k-2}Q^{\pi_0} - \mathcal{T}^{k-2}Q^*\|_\infty \\ &\leq \cdots \leq \gamma^k \|Q^{\pi_0} - Q^*\|_\infty \leq \gamma^k \cdot \frac{1}{1-\gamma}. \end{aligned}$$

Therefore, for all $k \geq \frac{\log \frac{1}{(1-\gamma)\epsilon}}{\log \frac{1}{\gamma}}$, we have $Q^{\pi_k} \geq Q^* - \epsilon \mathbb{1}$. □

Finally, we introduce a lemma that in helpful is the analysis of RL algorithms. We first define the *advantage* $A^\pi(s,a)$ of a policy π as

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

Also, we consider the state-action visitation distribution and a visitation measure:

$$d^\pi_{s_0}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P^\pi(s_t = s|s_0),$$

and we write

$$d^\pi_\mu(s) = \mathrm{E}_{s_0\sim\mu}[d^\pi_{s_0}(s)].$$

Lemma 2.13. *(The performance difference lemma). For all stationary policies π, π' and distributions μ over $\mathcal{S}$,*

$$V^\pi(\mu) - V^{\pi'}(\mu) = \frac{1}{1-\gamma}\mathrm{E}_{s'\sim d^\pi_\mu}\mathrm{E}_{a'\sim\pi(\cdot|s')}\big[A^{\pi'}(s',a')\big].$$

Proof. Let $P(\cdot|\pi, s_0)$ denote the distribution of trajectory τ when we start at s_0 and follows policy π. Then,

$$\begin{aligned}
V^\pi(s_0) - V^{\pi'}(s_0) &= \mathrm{E}_{\tau\sim P^\pi(\cdot|\pi,s_0)}\big[r(s_0,a_0) + \gamma r(s_1,a_1) + ... - V^{\pi'}(s_0)\big] \\
&= \mathrm{E}_{\tau\sim P^\pi(\cdot|\pi,s_0)}\big[Q^{\pi'}(s_0,a_0) - \gamma\mathrm{E}\big[V^{\pi'}(s_1)|s_0,a_0\big] + \gamma r(s_1,a_1) + ... - V^{\pi'}(s_0)\big] \\
&= \mathrm{E}_{\tau\sim P^\pi(\cdot|\pi,s_0)}\big[Q^{\pi'}(s_0,a_0) - \gamma V^{\pi'}(s_1) + \gamma r(s_1,a_1) + ... - V^{\pi'}(s_0)\big] \\
&= \mathrm{E}_{\tau\sim P^\pi(\cdot|\pi,s_0)}\big[A^{\pi'}(s_0,a_0) + \gamma\big[r(s_1,a_1) + \gamma r(s_2,a_2) + ... - V^{\pi'}(s_1)\big]\big] \\
&\vdots \\
&= \mathrm{E}_{\tau\sim P^\pi(\cdot|\pi,s_0)}\big[A^{\pi'}(s_0,a_0) + \gamma A^{\pi'}(s_1,a_1) + ...\big] \\
&\overset{(a)}{=} \sum_{t=0}^{\infty}\gamma^t\sum_{s,a}P(s_t = s, a_t = a|\pi, s_0)A^{\pi'}(s,a) \\
&= \sum_{t=0}^{\infty}\gamma^t\sum_{s}P(s_t = s|\pi, s_0)\pi(a_t = a|s)A^{\pi'}(s,a) \\
&\overset{(b)}{=} \frac{1}{1-\gamma}\mathrm{E}_{s\sim d^\pi_{s_0}}\mathrm{E}_{a\sim\pi(\cdot|s)}\big[A^{\pi'}(s,a)\big]
\end{aligned}$$

where (a) and (b) are justified by Fubini's theorem for L^1 functions. □

3. Bandit Problems and OFUL

In this section, we look at bandit problems, which are concerned with sequential selections of experiments. We consider a general setting, usually referred to as the linear bandits problem. We describe a bandit by a vector $x \in \mathbb{R}^d$ (a feature vector). In addition, we assume the reward r is given by the inner product with some unknown vector μ_* plus a noise term

η, i.e.,

$$r_t = \langle \mu_*, x_t \rangle + \eta_t \in [-1, 1].$$

For mathematical convenience, we also assume with $\|\mu_*\|_2 \leq M$ for some $M > 0$.

At every time-step, the agent picks a x_t from the decision set D_t, which may vary at different time-steps, and maintain a confidence set C_t such that $\mu^* \in C_t$ with high probability. One natural way to perform such task can be done by an algorithm called *optimism in the face of uncertainty* (OFU). We abbreviate "optimism in the face of uncertainty linear bandit algorithm" as OFUL.

In fact, OFUL is quite conceptually easy. In each round, the agent picks the pair $(x_t, \hat{\mu_t}) \in D_t \times C_{t-1}$ which maximizes the expected reward. After observing reward r_t, the agent updates a new confidence set C_t. We summarize this process in the next figure.

Algorithm 1 OFUL

```
for $t \leftarrow 1$ to $T$ do
    $(x_t, \widetilde{\mu_t}) \leftarrow \arg\max_{(x,\mu) \in D_t \times C_{t-1}} \langle x, \mu \rangle$
    Play $x_t$ and observe reward $r_t$
    Update $C_t$
end for
```

Although conceptually convincing, we have not yet given a mathematical guarantee that a confidence set C_t which contains μ_* with high probability can be constructed. We now demonstrate one possible way to build C_t with the promised property.

Let $\{F_t\}_{t \geq 0}$ be a filtration, that is, an increasing sequence of $\sigma-$algebras. Suppose $\{\eta_t\}_{t \geq 1}$ is a real-valued adaptive process such that η_t is conditionally σ^2-sub-Gaussian for some $\sigma > 0$, i.e.,

$$\mathrm{E}[e^{\lambda \eta_t} | F_{t-1}] \leq \exp(\frac{\lambda^2 \sigma^2}{2}), \text{ for each } \ \lambda \in \mathbb{R}. \tag{3.1}$$

Furthermore, suppose $\{X_t\}_{t \geq 1}$ is a $\mathbb{R}^d$-valued predictable process and $\Sigma_0 \in M^{d \times d}$ is a positive definite matrix. Define

$$\Sigma_t = \Sigma_0 + \sum_{i=1}^{t} X_i X_i^\top, V_t = \sum_{i=1}^{t} X_i X_i^\top, S_t = \sum_{i=1}^{t} \eta_i X_i, \text{ and } \|v\|_\Sigma = v^\top \Sigma v,$$

where Σ is a positive definite matrix. The following result is adopted from [1].

Theorem 3.1. *(Self-Normalized Bound for Vector-Valued Martingales). For any $\delta > 0$, with probability at least $1 - \delta$, it holds for all $t \geq 0$ that*

$$\|S_t\|_{\Sigma_t^{-1}} \leq \sigma \sqrt{\log \frac{\det \Sigma_t}{\det \Sigma_0} + \log \frac{1}{\delta^2}}.$$

Before proving Theorem 3.1, we give two lemmas.

LEMMA 3.2. *Let $\lambda \in \mathbb{R}^d$ be arbitrary and for any $t \geq 0$, define*

$$M_t^\lambda = \exp\big(\sum_{i=1}^{t}\big[\frac{\eta_i\langle X_i,\lambda\rangle}{\sigma} - \frac{1}{2}\langle X_i,\lambda\rangle^2\big]\big).$$

Let τ be a stopping time w.r.t. the filtration $\{F_t\}_{t\geq 0}$. Then M_τ^λ is almost surely well-defined with $\mathrm{E}[M_\tau^\lambda] \leq 1$.

Proof. First, we show that M_t^λ is a super-martingale, and as a consequence, $\mathrm{E}[M_t^\lambda] \leq 1$ for any $t \geq 0$.
Define $D_t^\lambda = \exp\big(\frac{\eta_t\langle X_t,\lambda\rangle}{\sigma} - \frac{1}{2}\langle X_t,\lambda\rangle^2\big)$. Observe that

$$\begin{aligned}\mathrm{E}\big[M_t^\lambda|F_{t-1}\big] &= \mathrm{E}\big[M_{t-1}^\lambda \cdot D_t^\lambda|F_{t-1}\big]\\ &= M_{t-1}^\lambda\mathrm{E}\big[D_t^\lambda|F_{t-1}\big]\\ &= M_{t-1}^\lambda \exp(-\frac{1}{2}\langle X_t,\lambda\rangle^2)\mathrm{E}\big[\exp(\eta_t \cdot \frac{\langle X_t,\lambda\rangle}{\sigma})|F_{t-1}\big]\\ &\leq M_{t-1}^\lambda \exp(-\frac{1}{2}\langle X_t,\lambda\rangle^2)\cdot\exp(\frac{1}{2}\langle X_t,\lambda\rangle^2)\\ &= M_{t-1}^\lambda.\end{aligned}$$

The first equality comes from the definition of M_t^λ. The second and the third are due to measurability. The last inequality is by (3.1). It is now obvious that

$$\mathrm{E}\big[M_t^\lambda\big] = \mathrm{E}\big[\mathrm{E}\big[M_t^\lambda|\mathcal{F}_{t-1}\big]\big] \leq \mathrm{E}\big[M_{t-1}^\lambda\big] = \cdots \leq \mathrm{E}\big[M_0^\lambda\big] = 1$$

Let $M_{\tau\wedge n}^\lambda$ be a stopped version of the super-martingale, which is still a super-martingale with $\mathrm{E}[M_{\tau\wedge n}^\lambda] \leq 1$, for all $n \geq 0$.
By Fatou's Lemma,

$$\mathrm{E}\big[M_\tau^\lambda\big] = \mathrm{E}\big[\liminf_{n\to\infty} M_{\tau\wedge n}^\lambda\big] \leq \liminf_{n\to\infty}\mathrm{E}\big[M_{\tau\wedge n}^\lambda\big] \leq 1. \qquad \square$$

LEMMA 3.3. *Let τ be a stopping time w.r.t. the filtration $(F_t)_{t\geq 0}$. Then, for $\delta > 0$, with probability at least* $1-\delta$,

$$\|S_\tau\|_{\Sigma_\tau^{-1}} \leq \sigma\sqrt{\log\frac{\det\Sigma_\tau}{\det\Sigma_0} + \log\frac{1}{\delta^2}}.$$

(Notice the difference between Theorem 3.1 and Lemma 3.3.)

Proof. Without loss of generality, we assume $\sigma = 1$.
Let $\Lambda \sim N(0,\Sigma_0^{-1})$ be a Gaussian random variable independent of $F_\infty = \sigma(\bigcup_{t\geq 0} F_t)$. Define

$$M_t = \mathrm{E}[M_t^\Lambda|F_\infty].$$

We have $\mathrm{E}\big[M_\tau\big] = \mathrm{E}\big[\mathrm{E}\big[M_\tau^\Lambda|F_\infty\big]\big] = \mathrm{E}\big[\mathrm{E}\big[M_\tau^\Lambda|\Lambda\big]\big] \leq 1$ by lemma 3.2.
Let's calculate M_t. Let f be the density of Λ and for a positive definite matrix P we define $c(P) = \sqrt{(2\pi)^d / \det(P)}$. Then,

$$\begin{aligned} M_t &= \int_{\mathbb{R}^d} \exp(\langle \lambda, S_t\rangle - \frac{1}{2}\|\lambda\|_{V_t}^2) f(\lambda) d\lambda \\ &= \int_{\mathbb{R}^d} \exp\big(- \frac{1}{2}\|\lambda - V_t^{-1}S_t\|_{V_t}^2 + \frac{1}{2}\|S_t\|_{V_t^{-1}}^2\big) f(\lambda) d\lambda \\ &= \frac{1}{c(\Sigma_0)} \exp\big(\frac{1}{2}\|S_t\|_{V_t^{-1}}^2\big) \int_{\mathbb{R}^d} \exp\big(- \frac{1}{2}\|\lambda - V_t^{-1}S_t\|_{V_t}^2 - \frac{1}{2}\|\lambda\|_{\Sigma_0}^2\big) d\lambda. \end{aligned}$$

Notice that

$$\|x-a\|_P^2 = \|x\|_Q^2 = \|x-(P+Q)^{-1}Pa\|_{P+Q}^2 + \|a\|_P^2 - \|Pa\|_{(P+Q)^{-1}}^2.$$

Then,

$$\|\lambda - V_t^{-1}S_t\|_{V_t}^2 + \|\lambda\|_{\Sigma_0}^2 = \|\lambda - \Sigma_t^{-1}S_t\|_{\Sigma_t}^2 + \|S_t\|_{V_t^{-1}}^2 - \|S_t\|_{\Sigma_t^{-1}}^2.$$

Hence,

$$\begin{aligned} M_t &= \frac{1}{c(\Sigma_0)} \exp\big(\frac{1}{2}\|S_t\|_{\Sigma_t^{-1}}^2\big) \int_{\mathbb{R}^d} \exp\big(- \frac{1}{2}\|\lambda - \Sigma_t^{-1}S_t\|_{\Sigma_t}^2\big) d\lambda \\ &= \frac{c(\Sigma_t)}{c(\Sigma_0)} \exp\big(\frac{1}{2}\|S_t\|_{\Sigma_t^{-1}}^2\big) \underbrace{\int_{\mathbb{R}^d} \frac{1}{c(\Sigma_t)} \exp\big(- \frac{1}{2}\|\lambda - \Sigma_t^{-1}S_t\|_{\Sigma_t}^2\big) d\lambda}_{=1} \\ &= \frac{\det(\Sigma_0)^{\frac{1}{2}}}{\det(\Sigma_t)^{\frac{1}{2}}} \exp\big(\frac{1}{2}\|S_t\|_{\Sigma_t^{-1}}^2\big). \end{aligned}$$

Finally, one has

$$\begin{aligned} M_\tau > \frac{1}{\delta} &\iff \frac{\det(\Sigma_0)^{\frac{1}{2}}}{\det(\Sigma_t)^{\frac{1}{2}}} \exp\big(\frac{1}{2}\|S_t\|_{\Sigma_t^{-1}}^2\big) > \frac{1}{\delta} \\ &\iff \|S_t\|_{\Sigma_t^{-1}} > \sqrt{\log\frac{1}{\delta^2} + \log\frac{\det\Sigma_t}{\det\Sigma_0}}. \end{aligned}$$

Using $\mathrm{E}[M_\tau] \leq 1$ and Markov's inequality

$$\Pr(M_\tau > \frac{1}{\delta}) \leq \mathrm{E}[M_\tau]\delta \leq \delta.$$

That is, with probability at least $1-\delta$,

$$\|S_\tau\|_{\Sigma_\tau^{-1}} \leq \sqrt{\log\frac{\det\Sigma_\tau}{\det\Sigma_0} + \log\frac{1}{\delta^2}}. \qquad \square$$

Proof of Theorem 1. Define the bad event

$$B_t(\delta) = \left\{\omega \in \Omega : \|S_t\|_{\Sigma_t^{-1}} > \sigma\sqrt{\log\frac{\det\Sigma_t}{\det\Sigma_0} + \log\frac{1}{\delta^2}}\right\}.$$

Consider the stopping time $\tau(\omega) = \min\{t \geq 0 : \omega \in B_t(\delta)\}$ such that

$$\bigcup_t B_t(\delta) = \{\omega : \tau(\omega) < \infty\}.$$

Thus, by Lemma 3.3,

$$\begin{aligned}
\Pr\ (\bigcup_{t\geq 0} B_t(\delta)) &= \Pr\ [\tau < \infty] \\
&= \Pr\ \left[\|S_\tau\|_{\Sigma_\tau^{-1}} > \sigma\sqrt{\log\frac{\det\Sigma_\tau}{\det\Sigma_0} + \log\frac{1}{\delta^2}}, \tau < \infty\right] \\
&\leq \Pr\ \left[\|S_\tau\|_{\Sigma_\tau^{-1}} > \sigma\sqrt{\log\frac{\det\Sigma_\tau}{\det\Sigma_0} + \log\frac{1}{\delta^2}}\right] < \delta.
\end{aligned}$$

□

Continuing the construction of C_t, we consider the solution $\hat{\mu_t}$ of the following regularized least squares problem:

$$\begin{aligned}
\hat{\mu_t} &= \arg\min_\mu \frac{1}{2}\sum_{i=1}^t (\langle X_i, \mu\rangle - r_i)^2 + \frac{1}{2}\|\mu\|_{\Sigma_0}^2 \\
&= \Sigma_t^{-1}\sum_{i=1}^t r_i X_i \\
&= \Sigma_t^{-1}(\Sigma_t - \Sigma_0)\mu_* + \Sigma_t^{-1}(\sum_{i=1}^t X_i\eta_i).
\end{aligned}$$

By a direct calculation, one obtains

$$\begin{aligned}
\|\hat{\mu_t} - \mu_*\|_{\Sigma_t} &\leq \|\Sigma_t^{-1}S_t\|_{\Sigma_t} + \|\Sigma^{-1}\Sigma_0\mu_*\|_{\Sigma_t} = \|S_t\|_{\Sigma_t^{-1}} + \|\Sigma_t^{-\frac{1}{2}}\Sigma_0\mu_*\|_2 \\
\Rightarrow \|\hat{\mu_t} - \mu_*\|_{\Sigma_t} &\leq \sigma\sqrt{\log\frac{\det(\Sigma_t)}{\det(\Sigma_0)} + \log\frac{1}{\delta^2}} + \|\Sigma_t^{-\frac{1}{2}}\Sigma_0\|\|\mu_*\|_2 \\
&\leq \sigma\sqrt{\log\frac{\det(\Sigma_t)}{\det(\Sigma_0)} + \log\frac{1}{\delta^2}} + M\|\Sigma_t^{-\frac{1}{2}}\Sigma_0\|
\end{aligned}$$

with probability at least $1 - \delta$. This suggests we let $C_t = \{\mu | \|\hat{\mu_t} - \mu_*\|_{\Sigma_t} \leq \beta_t\}$, where $\beta_t = \sigma\sqrt{\log\frac{\det(\Sigma_t)}{\det(\Sigma_0)} + \log\frac{1}{\delta^2}} + M\|\Sigma_t^{-\frac{1}{2}}\Sigma_0\|$. This construction is often referred to as the linear UCB algorithm, summarized as in the following.

Algorithm 2 LinUCB

```
$b_0 \leftarrow 0$
for $t \leftarrow 1$ to $T$ do
    $\hat{\mu}_{t-1} \leftarrow \Sigma_{t-1}^{-1} b_{t-1}$
    $\beta_{t-1} \leftarrow \sigma\sqrt{\log \frac{\det(\Sigma_t)}{\det(\Sigma_0)} + \log \frac{1}{\delta^2}} + \|\Sigma_t^{-\frac{1}{2}}\Sigma_0\| M$
    $x_t \leftarrow \arg\max_{x \in D_t} \left(\langle x, \hat{\mu}_{t-1}\rangle + \beta_{t-1}\sqrt{x^\top \Sigma_{t-1}^{-1} x}\right)$
    Play $x_t$ and observe $r_t$
    $\Sigma_t \leftarrow \Sigma_{t-1} x_t x_t^\top$
    $b_t \leftarrow b_{t-1} + r_t x_t$
end for
```

We now discuss the efficiency of the linUCB algorithm. We use the concept of *cummulative regret* to measure the performance of the algorithm.

Definition 3.4. The cummulative regret R_T at time T is defined as:

$$R_T = \sum_{t=1}^{T} \langle x_t^*, \mu_* \rangle - \langle x_t, \mu_* \rangle,$$

where $x_t^* = \arg\max_{x \in D_t} \langle x, \mu_* \rangle$.

Definition 3.5. The instantaneous regret at time t is defind as:

$$regret_t = \langle x_t^*, \mu_* \rangle - \langle x_t, \mu_* \rangle.$$

We start with the following technical lemma.

Lemma 3.6. *Let $A \in \mathbb{R}^{d \times d}$ be a positive semi-definite matrix. Then*

$$\log\det(A) \le d \log\left(\frac{1}{d} Tr(A)\right).$$

Proof. Let $\lambda_1, ..., \lambda_d$ be the eigenvalues of A. Recall that the determinant (trace) of a matrix is the product (sum) of the eigenvalues. It suffices to show

$$\frac{1}{d}\sum_{i=1}^{d} \log(\lambda_i) \le \log(\frac{1}{d}\sum_{i=1}^{d} \lambda_i).$$

The inequality holds by concavity. □

Corollary 3.7.

$$\log \frac{\det \Sigma_T}{\det \Sigma_0} \le d \log\left(1 + \frac{1}{d}\Sigma_{t=1}^{T} x_t^\top \Sigma_0^{-1} x_t\right).$$

Therefore, if $\Sigma_0 = \lambda I$ and that $\|x_t\|_2 \le B$ for all t, then

$$\log \frac{\det \Sigma_T}{\det \Sigma_0} \le d \log \Big(1 + \frac{TB^2}{d\lambda}\Big) \le \frac{TB^2}{\lambda}.$$

Proof. Put $v_t = \Sigma_0^{-\frac{1}{2}} x_t$. Observe that

$$\begin{aligned} \log \frac{\det \Sigma_T}{\det \Sigma_0} &= \log \Big(\det(\Sigma_0^{-\frac{1}{2}} \Sigma_T \Sigma_0^{-\frac{1}{2}})\Big) \\ &= \log\det \Big(I + \sum_{t=1}^{T} v_t v_t^\top\Big) \\ &\le d \log \Big(\frac{1}{d} \mathrm{Tr}\Big(I + \sum_{t=1}^{T} v_t v_t^\top\Big)\Big) \\ &= d \log \Big(1 + \frac{1}{d} \sum_{t=1}^{T} \|v_t\|_2^2\Big). \end{aligned}$$

The last equality holds because $v_t v_t^\top$ is a rank-1 matrix and v_t is the eigenvector with the eigenvalue $\|v_t\|_2^2$. □

Proposition 3.8. *Denote $\sqrt{x_t^\top \Sigma_{t-1}^{-1} x_t}$ by w_t. Then*

(a) $\det(\Sigma_T) = \det(\Sigma_0) \prod_{t=1}^{T} (1 + w_t^2)$.

Further suppoe that $\mu_* \in C_t$ for all t, where C_t and β_t are defined as in LinUCB. Then

(b) $regret_t \le 2\beta_{t-1} w_t$

(c) If $|\langle x, \mu_* \rangle|$ for $x \in \bigcup_t D_t$, then $\sum_{t=1}^{T} \log(1 + \beta_{t-1}^2) \cdot regret_t^2 \le 4 \log \frac{\det \Sigma_T}{\det \Sigma_0}$.

Proof. (a) Put $v_t = \Sigma_{t=1}^{-\frac{1}{2}} x_t$ so that $\|v_t\|^2 = w_t^2$. For any $t > 0$

$$\det(\Sigma_t) = \det(\Sigma_{t-1} + x_t x_t^\top) = \det\Big(\Sigma_{t-1}^{\frac{1}{2}}\Big(I + \Sigma_{t-1}^{-\frac{1}{2}} x_t x_t^\top \Sigma_{t-1}^{-\frac{1}{2}}\Big)\Sigma_{t-1}^{\frac{1}{2}}\Big) = \det(\Sigma_{t-1})(1 + \|v_t\|_2^2).$$

The claim follows easily by induction.

(b) By the choice of x_t, there exists a $\tilde{\mu_t} \in C_{t-1}$ for which

$$\langle x_t, \tilde{\mu_t} \rangle = \max_{(x,\mu) \in D_t \times C_{t-1}} \langle x, \mu \rangle \ge \langle x_t^*, \mu_t \rangle.$$

Thus,

$$\mathrm{regret}_t = \langle x_t^*, \mu_* \rangle - \langle x_t, \mu_* \rangle \le \langle x_t, \tilde{\mu_t} - \mu_* \rangle = \langle x_t, \tilde{\mu_t} - \hat{\mu_{t-1}} \rangle + \langle x_t, \hat{\mu_{t-1}} - \mu_* \rangle \le 2\beta_{t-1} w_t.$$

(c) By (b) and the additional assumption, $\mathrm{regret}_t \le 2 \min(\beta_{t-1} w_t, 1)$. Let $\tilde{w_t} = \min(w_t, \beta_{t-1}^{-1})$. Notice that $\tilde{w_t}^2 \in [0, \beta_{t-1}^{-2}]$, which implies

$$\mathrm{regret}_t^2 \le 4\beta_{t-1}^2 \tilde{w_t}^2 \le 4\beta_{t-1}^2 \frac{\beta_{t-1}^2}{\log(1 + \beta_{t-1}^2)} \log(1 + w_t^2) = 4 \frac{\log(1 + w_t^2)}{\log(1 + \beta_{t-1}^2)}.$$

The second inequality holds by considering the slope of $\log(1+x)$ and concavity. □
Combining with (a), we get the the following theorem.

Theorem 3.9. *(Regret Analysis of LinUCB). Suppose that $\|\mu_*\|_2 \leq M, \|x\|_2 \leq B$ and that $|\langle x, \mu_* \rangle| \leq 1$ for $x \in \bigcup_t D_t$. If one sets $\Sigma_0 = \lambda I$, then w.p. at least $1-\delta$, it holds that*

$$R_T^2 \leq 8d \log\Big(1 + \frac{TB^2}{d\lambda}\Big) \frac{\gamma^{-2}}{\log(1+\gamma^{-2})} \Big(\sigma^2 d\Big(\Big(T + \frac{d\lambda}{B^2}\Big) \log\Big(1 + \frac{TB^2}{d\lambda}\Big) - T\Big) + 2T\sigma^2 \log(\frac{1}{\delta}) + T\lambda M^2\Big),$$

where $\gamma = \sigma\sqrt{\log(\frac{1}{\delta^2})}$.

Proof. By theorem 3.1, the event $\{\mu_* \in C_t \text{ for all } t\}$ happens w.p. at least $1-\delta$ by the construction of C_t under LinUCB. By Hölder's inequality and proposition 3.8 (c),

$$R_T^2 = \Big(\sum_{t=1}^{T} \text{regret}_t\Big)^2 \leq \Big(\sum_{t=1}^{T} \log(1+\beta_{t-1}^{-2})\text{regret}_t^2\Big)\Big(\sum_{t=1}^{T} \frac{1}{\log(1+\beta_{t-1}^{-2})}\Big) \leq 4 \log \frac{\det \Sigma_T}{\det \Sigma_0} \sum_{t=1}^{T} \frac{1}{\log(1+\beta_{t-1}^{-2})}.$$

Since $\beta_{t-1} \geq \gamma$, one has $\log(1+\beta_{t-1}^{-2}) \geq \frac{\log(1+\gamma^{-2})}{\gamma^{-2}} \beta_{t-1}^{-2}$. Hence,

$$\begin{aligned} R_T^2 &\leq 4 \log \frac{\det \Sigma_T}{\det \Sigma_0} \cdot \frac{\gamma^{-2}}{\log(1+\gamma^{-2})} \sum_{t=1}^{T} \beta_{t-1}^2 \\ &\leq 8 \log \frac{\det \Sigma_T}{\det \Sigma_0} \cdot \frac{\gamma^{-2}}{\log(1+\gamma^{-2})} \sum_{t=0}^{T-1} \Big(\sigma^2 \log \frac{\det \Sigma_t}{\delta^2 \det \Sigma_0} + \lambda M^2\Big) \\ &\leq 8d \log\Big(1 + \frac{TB^2}{d\lambda}\Big) \frac{\gamma^{-2}}{\log(1+\gamma^{-2})} \sum_{t=0}^{T-1} \Big(\sigma^2\Big(d \log\Big(1 + \frac{tB^2}{d\lambda}\Big) + \log(\frac{1}{\delta^2})\Big) + \lambda M^2\Big). \end{aligned}$$

The second inequality is by an application AM-GM inequality with the definition of β_t. The last inequality is from Corollary 3.7.
We complete the proof by noting the following fact:

$$\begin{aligned} \sum_{t=0}^{T-1} \log\Big(1 + \frac{tB^2}{d\lambda}\Big) \leq \frac{d\lambda}{B^2} \int_0^{\frac{TB^2}{0}} \log(1+x)dx &= \frac{d\lambda}{B^2}\Big((1+x)\log(1+x) - x\Big)\Big|_0^{\frac{TB^2}{d\lambda}} \\ &= \Big(T + \frac{d\lambda}{B^2}\Big) \log\Big(1 + \frac{TB^2}{d\lambda}\Big) - T \end{aligned}$$ □

4. Kernels and RKHSs

Before our problem statement, we give a brief introduction to RKHSs. We start with the following definitions. Let Ω be a non-empty set.

Definition 4.1. We call $k : \Omega \times \Omega \to \mathbb{R}$ a kernel if k satisfies:

(a) $k(x,y) = k(y,x)$, for all $x, y \in \Omega$,

(b) $\sum_{i,j=1}^{n} c_i k(x_i, x_j) c_j \geq 0,$ for any $(x_i)_{i=1}^{n} \subset \Omega$ and $(c_i)_{i=1}^{n} \subset \mathbb{R}$.

Suppose H is a Hilbert space of functions on Ω. We denote by L_x the evaluation functional at point $x \in \Omega$.

Definition 4.2. We say H is a reproducing kernel Hilbert space (RKHS) if $L_x \in H^*$ for all $x \in \Omega$, where H^* denotes the dual space.

Definition 4.3. We say $k : \Omega \times \Omega \to \mathbb{R}$ is a reproducing kernel (RK) if k satisfies:

(a) $k(x, \cdot) \in H,$ for all $x \in \Omega$,

(b) $\langle f, k(x, \cdot)\rangle_H = f(x),$ for all $x \in \Omega$ and $f \in H$.

It's obvious that a RK is indeed a kernel. We now draw the connection between a RKHS and a RK.

Theorem 4.4. *A Hilbert space of functions H is a RKHS if and only if H is endowed with a RK k.*

Proof. We first prove the *only if* part. Suppose H is a RKHS. By Riesz representation theorem, we identify each evaluation functional L_x by a $\Phi_x \in H$. Define a function $k : \Omega \times \Omega \to \mathbb{R}$ by $k(x, y) = \langle \Phi_x, \Phi_y\rangle_H$. Clearly, k is symmetric. For any $(c_i)_{i=1}^{n} \subset \mathbb{R}$ and $(x_i)_{i=1}^{n} \subset \Omega$,

$$\sum_{i,j=1}^{n} c_i k(x_i, x_j) c_j = \Big\langle \sum_{i=1}^{n} c_i \Phi_{x_i}, \sum_{j=1}^{n} c_j \Phi_{x_j} \Big\rangle_H = \|\sum_{i=1}^{n} c_i \Phi_{x_i}\|_H^2 \geq 0.$$

Hence, k is a kernel with $\langle f, k(x, \cdot)\rangle_H = \langle f, \Phi_x\rangle_H = L_x(f) = f(x)$. I.e., k is a RK.
Now, suppose k is a RK on H. For each $x \in \Omega$, one has

$$|L_x(f)| = |f(x)| = |\langle f, k(x, \cdot)\rangle_H| \leq \|f\|_H \|k(x, \cdot)\|_H.$$

That is, $L_x \in H^*$ and H is a RKHS. □

Theorem 4.5. *(Moore-Aronszajn). If k is a kernel on Ω, then there exists a unique H, which is a Hilbert space of functions on Ω, such that k is a RK for H. Furthermore, span$\{k(x, \cdot)|x \in \Omega\}$ is dense in H.*

Proof. The result is well-known, so we only give a sketch of the proof. Intuitively, the idea of the proof is similar to the completion of a metric space.
Step 1: Define $H_0 = \text{span}\{k(x, \cdot)|x \in \Omega\}$. For each $f \in H_0$, we write $f = \sum_{i=1}^{n} a_i \phi_{x_i}$, where $\phi_{x_i} = k(x_i, \cdot)$. Then, define $\langle f, g\rangle_{H_0} = \sum_{i,j=1}^{n} a_i b_j k(x_i, x_j)$. It can be verified that $\langle f, g\rangle_{H_0}$ does not depend on the representations of f and g, and is indeed an inner product.
Step 2: Note that $(f_n(x))$ is Cauchy on $\mathbb{R}$, for any Cauchy sequence $(f_n) \subset H_0$, at any $x \in \Omega$.

Now, one can show that $\lim_{n\to\infty}\|f_n\|_{H_0}$ if and only if $f_n \to 0$, for any Cauchy $(f_n) \subset H_0$.
Step 3: Let $H = \{f|f$ is a pointwise limit of some Cauchy sequence in $H_0\}$ and define $\langle f, g\rangle_H = \lim_{n\to\infty}\langle f_n, g_n\rangle_{H_0}$, where f_n and g_n are Cauchy sequences in H_0 that converge to f and g, respectively. It can be shown that $\langle f, g\rangle_H$ doe not depend on the choice of converging Cauchy sequences. Following that, $\langle\cdot,\cdot\rangle_H$ is a well-defined inner product. With this setting, we can show the density result as well as the completeness of H.
Step 4: Finally, we may first prove that any Hilbert space H' on which k is a RK is an extension of H. From that, using an orthogonal decomposition argument for an arbitrary element $f \in H'$, we have $f = f_H \in H$, where $f = f_H + f_{H^\perp}$ is the corresponding orthogonal decomposition. □

Theorem 4.6. *(The Representer Theorem). Let k be a kernel on Ω with a corresponding RKHS H. Suppose g is a non-decreasing function on $[0,\infty)$. For any non-empty $S \subset \Omega$ and a real-valued $L : S \to \mathbb{R}$, we have*

$$\inf_{f\in H:\|f\|_H\le R}(L(f|_S) + g(\|f\|_H)) = \inf_{f\in M:\|f\|_H\le R}(L(f|_S) + g(\|f\|_H)),$$

where $R \in (0,\infty)$ and $M = \overline{span(k(x,\cdot|x \in S))}$.

Proof. Let $f \in H$. Since M is closed, we may write $f = f_M + f_{M^\perp}$. In particular, $f_{M^\perp}(x) = \langle f_{M^\perp}, k(x,\cdot)\rangle_H = 0$ for all $x \in S$. Hence, $f_{M^\perp}|_S = 0$ and $f|_S = f_M|_S \in M$. The theorem now follows as $g(\|f\|_H) \ge g(\|f|_S\|_H)$ and $\|f\|_H \ge \|f|_S\|_H$. □

Intuitively, one may consider S to be the set of all collected data. The representer theorem states that the minimization problem over H can be done simply over the linear span of the data collected. This is in fact a very powerful theorem, reducing computationally infeasible problems into more feasible ones.

5. Simple Regret for Gaussian Process Bandits

5.1. Problem Statement.

We formulate our sequential optimization problem as follows: Let an unknown $f : \mathcal{X} \to \mathbb{R}$ be the objective function, where $\mathcal{X} \subset \mathbb{R}^d$ is a convex and compact domain. Consider an optimal $x^* \in \arg\max_{x\in\mathcal{X}} f(x)$. An algorithm sequentially chooses observations $\{x_n \in \mathcal{X}\}_{n\in\mathbb{N}}$ and observes perturbed objective values $\{y_n = f(x_n) + \epsilon_n\}_{n\in\mathbb{N}}$. For simplicity of notations, we adopt $X_n = [x_1, ..., x_n]^\top$, $E_n = [\epsilon_1, ..., \epsilon_n]^\top$ and $F_n = [f(x_1), ..., f(x_n)]^\top$.
We will focus on the simple regret of N tries:

$$r_N = f(x^*) - f(\hat{x}^*_N)$$

where $\hat{x}^*_N$ is a candidate maximizer, and N may be unknown *a priori*. Throughout the rest of the section, our analysis follows [2].

5.2. **Gaussian Processes.**

Recall that a random process $\{\hat{f}(x)\}_{x\in\mathcal{X}}$ is said to be a Gaussian Process if its finite-dimensional distributions are multivariate Gaussian. More explicitly, given any finite sequence $(x_i)_{i=1}^n \subset \mathcal{X}$, $[\hat{f}(x_1), ..., \hat{f}(x_n)]^\top$ must be a multivariate Gaussian random vector. In particular, the GP $\{\hat{f}(x)\}_{x\in\mathcal{X}}$ is uniquely determined by its mean function and covariance function:

$$\mu(x) = \mathrm{E}\big[\hat{f}(x)\big], \text{ and } k(x,x') = \mathrm{E}\big[(\hat{f}(x) - \mu(x))(\hat{f}(x') - \mu(x'))\big].$$

Suppose our observations $[\hat{f}(x_1), ..., \hat{f}(x_n)]^\top$ are corrupted by i.i.d. noise terms $\epsilon_i \sim N(0, \lambda^2)$. By simple linear algebra, the posterior means and variances conditioned on noisy observations $Y_n = [y_1, y_2, \dots, y_n]^\top = [\hat{f}(x_1) + \epsilon_1, ..., \hat{f}(x_n) + \epsilon_n]^\top$ can be shown to be:

$$\begin{aligned} \mu_n(x) &= k^\top(x, X_n)(k(X_n, X_n) + \lambda^2 I_n)^{-1} Y_n = Z_n(x)^\top Y_n, \\ k_n(x,x) &= k(x,x) - k^\top(x, X_n)(k(X_n, X_n) + \lambda^2 I_n)^{-1} k(x, X_n) \\ &= k(x,x) - Z_n(x)^\top k(x, X_n) := \sigma_n^2(x), \end{aligned}$$

where $k(x, X_n) = [k(x, x_1), ..., k(x, x_n)]^\top$, $k(X_n, X_n)$ is the covariance matrix, and $Z_n(x)^\top = k^\top(x, X_n)(k(X_n, X_n) + \lambda^2 I_n)^{-1}$.

5.3. **Regularity Assumptions.**

We consider different cases regarding the regularity assumptions.

Assumption 5.1. *The objective function f is assumed to live in the RKHS H corresponding to a positive definite kernel k. In particular, $\|f\|_H \le B$, for some $B \ge 0$.*

Assumption 5.2. *(Sub-Gaussian Noise)*
The noise terms ϵ_n are i.i.d.over n. In addition, $\forall h \in \mathbb{R}$, $\forall n \in \mathbb{N}$, $\mathrm{E}[e^{h\epsilon_n}] \le \exp(\frac{h^2 R^2}{2})$, for some $R > 0$.

Assumption 5.3. *(Light-Tailed Noise)*
The noise terms ϵ_n are i.i.d. zero mean random variables over n. In addition, $\forall h \le h_0$, $\forall n \in \mathbb{N}$, $\mathrm{E}[e^{h\epsilon_n}] \le \exp(\frac{h^2 \xi_0}{2})$, for some $\xi_0 > 0$.

Assumption 5.4. *For each given $n \in \mathbb{N}$ and $f \in H$ with $\|f\|_H \le B$, there exists a discretization $\mathcal{D}_n$ of $\mathcal{X}$ such that $f(x) - f([x]_n) \le \frac{1}{\sqrt{n}}$, where $[x]_n = \arg\min_{x'\in\mathcal{D}_n} \|x' - x\|_{l^2}$ is the closest point in $\mathcal{D}_n$ to x, and $|\mathcal{D}_n| \le CB^d n^{d/2}$, where C is a constant independent of n and B.*

5.4. **Confidence Intervals Under Two Types of Noise.**

We first build two types of confidence intervals that are paramount in our regret analysis, corresponding to the two cases of noise, respectively.

Lemma 5.5. *Given a RKHS H with a reproducing kernel k, any $(c_i)_{i=1}^n \subset \mathbb{R}$ and $(x_i)_{i=1}^n \subset \mathcal{X}$,*

$$\|\sum_{i=1}^n c_i k(\cdot, x_i)\|_H = \sup_{\|f\|_H \le 1} \big(\sum_{i=1}^n c_i f(x_i)\big).$$

Proof. If $\|f\|_H \le 1$, then

$$\|\sum_{i=1}^n c_i k(\cdot, x_i)\|_H \ge \|\sum_{i=1}^n c_i k(\cdot, x_i)\|_H \|f\|_H \ge \big\langle \sum_{i=1}^n c_i k(\cdot, x_i), f\big\rangle_H = \sum_{i=1}^n c_i f(x_i).$$

On the other hand, we let $g := \sum_{i=1}^n c_i k(\cdot, x_i)/\|\sum_{i=1}^n c_i k(\cdot, x_i)\|_H$. Then,

$$\sup_{\|f\|_H \le 1} \big(\sum_{i=1}^n c_i f(x_i)\big) \ge \big\langle \sum_{i=1}^n c_i k(\cdot, x_i), g\big\rangle_H = \|\sum_{i=1}^n c_i k(\cdot, x_i)\|_H.$$ □

Corollary 5.6. *For a positive semi-definite kernel k and its corresponding RKHS H, it holds that*

$$\sup_{\|f\|_H \le 1} \big(f(x) - \sum_{i=1}^n \zeta_i(x) f(x_i)\big)^2 = \left\|k(\cdot, x) - \sum_{i=1}^n \zeta_i(x) k(\cdot, x_i)\right\|_H^2.$$

where $\zeta_i(x) = [Z_n(x)]_i$.

Proposition 5.7. *Let $\sigma_n^2(x) := k(x,x) - Z_n(x)^\top k(x, X_n)$. Then*

$$\sigma_n^2(x) = \sup_{f:\|f\|_{\mathcal{H}_k} \le 1} \underbrace{(f(x) - Z_n^\top(x) F_n)^2}_{\textit{noise free}} + \underbrace{\lambda^2 \|Z_n(x)\|^2.}_{\textit{noise term}}$$

Proof. Expanding the RKHS norm in the right hand side of *Corollary* 5.6 through an algebraic manipulation, we get

$$\left\|k(\cdot, x) - \sum_{i=1}^n \zeta_i(x) k(\cdot, x_i)\right\|_H^2 = \sigma_n^2(x) - \lambda^2 \|Z_n(x)\|^2.$$

Hence, by *Corollary* 5.6

$$\sigma_n^2(x) = \sup_{f:\|f\|_{\mathcal{H}_k} \le 1} (f(x) - Z_n^\top(x) F_n)^2 + \lambda^2 \|Z_n(x)\|^2.$$ □

Notice that the first term $f(x) - Z_n^\top(x) F_n$ captures the maximum prediction error from noise free observations. The second term captures the effect of noise.

Theorem 5.8. *(Confidence Interval for Sub-Gaussian Noise)*
Assume Assumption 5.1 and 5.2 hold. Provided n noisy observations Y_n from f, let $\mu_n := Z_n(x)^\top Y_n$ and $\sigma_n^2 := k(x,x) - Z_n(x)^\top k(x, X_n.)$ Assume X_n and E_n are independent. For a fixed $x \in \mathcal{X}$, define the upper and lower confidence bounds, respectively,

$$U_n^\delta(x) := \mu_n(x) + (B + \beta(\delta))\sigma_n(x), \text{ and } L_n^\delta(x) := \mu_n(x) - (B + \beta(\delta))\sigma_n(x)$$

with $\beta(\delta) = \frac{R}{\lambda}\sqrt{2\log(\frac{1}{\delta})}$, where $\delta \in (0,1)$, and B, R are the parameters specified in Assumption 5.1 and 5.2. We have

$$f(x) \leq U_n^\delta(x) \text{ w.p. at least } 1-\delta, \text{ and } f(x) \geq L_n^\delta(x) \text{ w.p. at least } 1-\delta.$$

Proof. First, we split $\mu_n(x)$ into noise-free term and noise term, i.e.

$$f(x) - \mu_n(x) = f(x) - Z_n^T(x)F_n - Z_n^T(x)E_n.$$

For the noise-free term, we let $\tilde{f}(\cdot) = f(\cdot)/B$, and $\tilde{F}_n = [\tilde{f}(x_1), \tilde{f}(x_2), ..., \tilde{f}(x_n)]^\top$, then

$$|f(x) - Z_n^T(x)F_n| = B|\tilde{f}(x) - Z_n^T(x)\tilde{F}_n| \leq B\sigma_n(x).$$

For the noise term, by *Assumption* 5.2, we can show that $Z_n^T(x)E_n$ is a R^2-Sub-Guassian random variables whose moment generating function is bounded.

$$\mathrm{E}\left[\exp(Z_n^T(x)E_n)\right] = \prod_{i=1}^n \exp(\zeta_i(x)\epsilon_i) \leq \prod_{i=1}^n \exp\left(\frac{R^2(\zeta_i(x))^2}{2}\right) \leq \exp\left(\frac{R^2\sigma_n^2(x)}{2\lambda^2}\right),$$

where the first equation is a result of *Assumption* 5.2 and the independence between X_n and E_n. The last inequality holds by *Proposition* 5.7.
With the inequality above, we can utilize a Chernoff bound to get:

$$\begin{aligned} Z_n(x)E_n &\geq -\tfrac{\sigma_n(x)R}{\lambda}\sqrt{2\log(\tfrac{1}{\delta})}, \\ Z_n(x)E_n &\leq \tfrac{\sigma_n(x)R}{\lambda}\sqrt{2\log(\tfrac{1}{\delta})}, \end{aligned}$$

with probability at least $1-\delta$.
Putting together the bounds for $f(x) - Z_n^T(x)F_n$ and $Z_n^T(x)E_n$, the proof is done. □

Theorem 5.9. *(Confidence Interval for Light-Tailed Noise)*
Assume Assumption 5.1 and 5.3 hold. Provided n noisy observations Y_n from f, let μ_n and σ_n be defined as in Theorem 5.8. Assume X_n are independent of E_n. For a fixed $x \in \mathcal{X}$, define the upper and lower confidence bounds, respectively,

$$U_n^\delta(x) := \mu_n(x) + (B + \beta(\delta))\sigma_n(x), \text{ and } L_n^\delta(x) := \mu_n(x) - (B + \beta(\delta))\sigma_n(x)$$

with $\beta(\delta) = \frac{1}{\lambda}\sqrt{2(\xi_0 \vee \frac{2\log(\frac{1}{\delta})}{h_0^2})\log(\frac{1}{\delta})}$, *where* $\delta \in (0,1)$, *and* B, ξ_0, h_0 *are the parameters specified in Assumption 5.1 and 5.3. We have*

$$f(x) \leq U_n^\delta(x) \text{ w.p. at least } 1-\delta, \text{ and } f(x) \geq L_n^\delta(x) \text{ w.p. at least } 1-\delta.$$

Proof. For the simplicity of notations, we use

$$\tau = \|Z_n(x)\|\sqrt{2(\xi_0 \vee \frac{2\log(1/\delta)}{h_0^2})\log(\frac{1}{\delta})} \text{ and } \xi = \xi_0 \vee \frac{2\log(1/\delta)}{h_0^2}.$$

For $\theta = \frac{\tau}{\xi\|Z_n(x)\|^2}$, we have

$$\begin{aligned}
\Pr[Z_n^T(x)E_n \geq \tau] &= \Pr\left[\exp(\theta Z_n^T(x)E_n) \geq \exp(\theta\tau)\right] \\
&\leq \exp(-\theta\tau)\mathrm{E}\left[\exp(\theta Z_n^T(x)E_n)\right] \\
&\leq \exp(-\theta\tau)\prod_{i=1}^{n}\exp\left(\frac{1}{2}\xi_0\theta^2(\zeta_i(x))^2\right) \\
&\leq \exp\left(-\frac{\tau^2}{2\xi\|Z_n(x)\|^2}\right) = \delta.
\end{aligned}$$

The first inequality follows from Markov's Inequality. The second inequality is due to *Assumption 5.3.* and $\zeta_i(x) = [Z_n(x)]_i$. The last inequality is resulted from $\xi_0 \leq \xi$ and some algebraic manipulations. By the above, we have $\Pr[Z_n^T(x)E_n \leq \tau] \geq 1-\delta$.
From *Proposition* 5.7, we have

$$\sigma_n^2(x) \geq \lambda^2\|Z_n(x)\|_2^2 \implies \|Z_n(x)\|_2 \leq \frac{\sigma_n(x)}{\lambda}.$$

Then

$$\tau \leq \frac{\sigma_n(x)}{\lambda}\sqrt{2(\xi_0 \vee \frac{2\log(1/\delta)}{h_0^2})\log(\frac{1}{\delta})} =: \tau^*,$$

and

$$\Pr[Z_n^T E_n \leq \tau^*] \geq 1-\delta.$$

Hence, with probability at least $1-\delta$, we have

$$f(x) - \mu_n(x) \leq B\sigma_n(x) + \tau^*.$$

We can obtain the lower bound similarly. □

5.5. **MVR and Simple Regret.**

In this subsection, we discuss an exploration algorithm referred to as Maximum Variance Reduction (MVR). MVR relies on the principle of reducing the maximum uncertainty where the uncertainty is measured by the posterior variance of the surrogate GP model. After N exploration trials, MVR returns a candidate maximizer according to the prediction provided by the learnt GP model. A pseudo-code is provided below.

Algorithm 3 Maximum Variance Reduction (MVR)

```
Initialization: $k$, $\mathcal{X}$, $f$, $\sigma_0^2(x) = k(x, x)$.
for $n = 1, 2, \dots, N$ do
    $x_n = \arg\max_{x \in \mathcal{X}} \sigma_{n-1}^2(x)$, where a tie is broken arbitrarily.
    Update $\sigma_n^2(\cdot) := k(\cdot, \cdot) - Z_n(\cdot)^\top k(\cdot, X_n)$.
end for
Update $\mu_N(\cdot) := Z_n(x)^\top Y_n$.
return $\hat{x}_N^* = \arg\max_{x \in \mathcal{X}} \mu_N(x)$, where a tie is broken arbitrarily.
```

We consider several lemmas that are used in our analysis of the MVR algorithm.

LEMMA 5.10. *Let* $\mathcal{I}(Y_n; \hat{f}) = \frac{1}{2} \log \left(\det(I_n + \frac{1}{\lambda^2} k(X_n, X_n)) \right)$. *We have*

$$\sum_{n=1}^{N} \sigma_{n-1}^2(x_n) \leq \frac{2}{\log(1 + \frac{1}{\lambda^2})} \mathcal{I}(Y_n; \hat{f}).$$

A proof can be found in [4]

LEMMA 5.11. *Conditioned on a set of noisy observation* Y_n *from* f, *we have*

$$\mu_n = \arg\min_{g \in H} \left(\lambda^2 \|g\|_H^2 + \sum_{i=1}^{n} (g(x_i) - y_i)^2 \right).$$

A proof can be found in [3]

LEMMA 5.12. *Conditioned on noisy observations* Y_n *from* f *with* $\|f\|_H \leq B$, *the RKHS norm of* $\mu_n(x)$ *with probability at least* $1 - \delta$ *satisfies*

$$\|\mu_n\|_H \leq B + \sqrt{n}\beta(2\delta/n).$$

For R^2*-Sub-Gaussian noise,* $\beta(\delta) = \frac{R}{\lambda}\sqrt{2\log(\frac{1}{\delta})}$,

For Light-tailed noise, $\beta(\delta) = \frac{1}{\lambda}\sqrt{2\left(\xi_0 \vee \frac{2\log(1/\delta)}{h_0^2}\right)\log(\frac{1}{\delta})}$.

Proof. Note that

$$\|\mu_n\|_H = \|Z_n^T(x)F_n + Z_n^T(x)E_n\|_H \leq \|Z_n^T(x)F_n\|_H + \|Z_n^T(x)E_n\|_H.$$

From *Lemma* 5.11, we have

$$\lambda^2 \|Z_n^T(.)F_n\|_H^2 + \sum_{i=1}^{n} (Z_n^T(x_i)F_n - f(x_i)^2 \leq \lambda^2 \|f\|_H^2 + \sum_{i=1}^{n} (f(x_i) - f(x_i))^2.$$

Thus,

$$\|Z_n^T(.)F_n\|_H \leq \|f\|_H \leq B.$$

By the reproducing property and some algebraic manipulations, one can show that

$$\|Z_n^T(x)E_n\|_H \leq \frac{1}{\lambda^2}\|E_n\|_{l_2}^2.$$

For R^2-Sub-Gaussian noise, by Chernoff-Hoeffding inequality, with probability at least $1-\delta'$, we have

$$\epsilon_i^2 \leq 2R^2 \log(\frac{1}{2\delta'}).$$

With a union bound over $i = 1, 2, ..., n$, with probability at least $1 - \delta'$

$$\frac{1}{\lambda^2}\|E_n\|_{l_2}^2 \leq \frac{2nR^2}{\lambda^2} \log(\frac{n}{2\delta}),$$

where $\delta' = \frac{\delta}{n}$.
For the light-tailed noise, we use the fact that $\Pr[Z_n^T(x)E_n \geq \tau] \leq \delta$, which is proved *Theorem* 5.9. Taking $n = 1,\ Z_n = 1$,

$$\epsilon_i^2 \leq 2\left(\xi_0 \vee \frac{2\log(1/\delta)}{h_0^2}\right) \log(\frac{1}{2\delta'}) \text{ w.p. at least } 1 - \delta'.$$

Similarly, we get a bound for $\frac{1}{\lambda^2}\|E_n\|_{l_2}^2$ through an union bound of the above over $i = 1, 2, ..., n$. Combining the above bounds for $\|Z_n^T(x)F_n\|_H$ and $\|Z_n^T(x)E_n\|_H$, the lemma is proven. □

We close this section with our foremost theorem. That is, a bound for the simple regret of MVR.

Theorem 5.13. *Under Assumptions 5.1 and 5.4 with* $\gamma_n := \sup_{X_N \in \mathcal{X}} \mathcal{I}(Y_n; \hat{f})$, B, R, h_0, ξ_0 *and* C *being constants, for* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$, *MVR satisfies*

$$r_N^{MVR} \leq \sqrt{\frac{2\gamma_N}{\log(1 + \frac{1}{\lambda^2})N}} \left(2B + \beta\left(\frac{\delta}{3}\right) + \beta\left(\frac{\delta}{3C(B + \sqrt{N}\beta(2\delta/3N)^d N^{\frac{d}{2}}}\right)\right).$$

For R^2-Sub-Gaussian noise, $\beta(\delta) = \frac{R}{\lambda}\sqrt{2\log(\frac{1}{\delta})}$.

For Light-tailed noise, $\beta(\delta) = \frac{1}{\lambda}\sqrt{2\left(\xi_0 \vee \frac{2\log(1/\delta)}{h_0^2}\right)\log(\frac{1}{\delta})}$.

Proof. The proof of this theorem could be mainly separated into four steps:

Step 1: Consider the event $\mathcal{E} := \{\|\mu_N\|_H \leq B_0(\delta/3) := B + \sqrt{N}\beta(2\delta/3N)\}$, with $P(\mathcal{E}) \geq 1 - \frac{\delta}{3}$.

Under the event $\mathcal{E}$, with *Assumption* 5.4, there exists a discretization $\mathcal{D}_n(\delta)$ of $\mathcal{X}$ such that

$$f(x) - f([x]_N) \leq \frac{1}{\sqrt{N}},\ \mu_N(x) - \mu_N([x]_N) \leq \frac{1}{\sqrt{N}},\ \textit{and}\ |\mathcal{D_N}(\delta)| \leq CB_0^d(\delta/3)N^{\frac{d}{2}}.$$

Moreover, under the event $\mathcal{E}$, one sees that

$$r_N^{MVR} = f(x^*) - f(\hat{x}_N^*) \leq f(x^*) - \mu_N(x^*) + \mu_N([\hat{x}_N^*]_N) - f([\hat{x}_N^*]_N) + \frac{2}{\sqrt{N}}.$$

Step 2: By *Theorem* 5.8 (or 5.9) and an union bound, for each $x \in \mathcal{D}_N(\delta)$, we have

$$f(x) \geq \mu_n(x) - \left(B + \beta(\frac{\delta}{3|\mathcal{D}_N(\delta)|})\right)\sigma_n(x),\ \textit{w.p. at least}\ 1 - \frac{\delta}{3}.$$

Then, under the event $\mathcal{E}$,

$$f(x^*) - f(\hat{x}_N^*) \leq \underbrace{\left(B + \beta(\frac{\delta}{3})\right)\sigma_N(x^*)}_{\text{first part}} + \underbrace{\left(B + \beta(\frac{\delta}{3|\mathcal{D}_N(\delta)|})\right)\sigma_N([\hat{x}_N^*]_N)}_{\text{second part}} + \frac{2}{\sqrt{N}}.$$

The first part results from *Theorem* 5.8 (or 5.9) with probability at least $1 - \frac{\delta}{3}$. The second part holds with probability at least $1 - \frac{\delta}{3}$, from an union bound over all $x \in \mathcal{X}$. Hence, we see that this inequality holds with probability at least $1 - \delta$.

Step 3: Since k is positive semi-definite and $x_n = \arg\max_{x \in \mathcal{X}} \sigma_{n-1}^2(x)$,

$$\sigma_N^2(x) \leq \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2(x_n).$$

Therefore, by *Lemma* 5.10, we have a bound for $\sigma_N^2(x)$, i.e.,

$$\sigma_N^2(x) \leq \frac{2\mathcal{I}(Y_n;\hat{f})}{\log(1+\frac{1}{\lambda^2})N} \leq \frac{2\gamma_N}{\log(1+\frac{1}{\lambda^2})N}.$$

Step 4: Using the upper bounds for $\sigma_N(x^*)$, $\sigma_N([\hat{x}_N^*]_N)$, and that $|\mathcal{D_N}(\delta)| \leq CB_0^d(\delta/3)N^{\frac{d}{2}}$,

$$f(x^*) - f(\hat{x}_N^*) \leq \sqrt{\frac{2\gamma_N}{\log(1+\frac{1}{\lambda^2})N}}\left(2B + \beta\left(\frac{\delta}{3}\right) + \beta\left(\frac{\delta}{3C(B+\sqrt{N}\beta(2\delta/3N))^d N^{\frac{d}{2}}}\right)\right),$$

with probability at least $1 - \delta$, where $f(x^*) - f(\hat{x}_N^*) = r_N^{MVR}$. □

6. Acknowledgement

The authors sincerely thank Prof. Pei-Yuan Wu for his dedicated teaching and our teaching assistants. We also thank National Center for Theoretical Sciences (NCTS) for organizing this spectacular event as well as holding teatime on every Wednesday.

7. References

[1] Abbasi-yadkori, Yasin, Dàvid Pal, and Csàba Szepesvàri. 2011. "Improved Algorithms for Linear Stochastic Bandits." In *Advances in Neural Information Processing Systems*. Vol. 24. Curran Associates, Inc.

[2] Vakili, Sattar, Nacime Bouziani, Sepehr Jalali, Alberto Bernacchia, and Da-shan Shiu. 2021. "Optimal Order Simple Regret for Gaussian Process Bandits." In *Advances in Neural Information Processing Systems*, 34:21202–21215. Curran Associates, Inc.

[3] Kanagawa, Motonobu, Philipp Hennig, D. Sejdinovic and Bharath K. Sriperumbudur. "Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences." *ArXiv abs/1807.02582* (2018)

[4] Srinivas, Niranjan, Andreas Krause, Sham M. Kakade, and Matthias W Seeger. 2010. "Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design." In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 1015–22. ICML.