
CS 6780 Research Project: Multi-armed Bandits with Dependent Arms

Bangrui Chen
Saul Toscano Palmerin
Zhengdi Shen

BC496@CORNELL.EDU
ST684@CORNELL.EDU
ZS267@CORNELL.EDU

1. Motivation

We are interested in the multi armed bandits problem with correlated arms. Theoretically this problem can be solved using dynamic programming, however it usually suffers from the curse of dimensionality when the dimension of the arm is high. There are two well known heuristic algorithms for this problem which are Exponential Gradient algorithm and the upper confidence bound algorithm. In this project, we hope to investigate the combination of these two different algorithms.

2. Problem Formulation

We have a finite set $\mathcal{U}_r = \{\mathbf{u}_1, \dots, \mathbf{u}_m\} \subset \mathbb{R}^r$ that corresponds to the set of arms, where $r \geq 2$. For any time $t = 1, 2, \dots, T$, we are asked to pick one arm X_t . The reward Y_t of playing arm $X_t \in \mathcal{U}_r$ in period t is given by

$$Y_t = \theta \cdot X_t + \epsilon_t,$$

where $\epsilon_t \sim N(0, \sigma^2)$ is the measurement error with σ known. Here θ is an unknown random vector, which is drawn from a multivariate normal distribution with mean μ and variance Σ . We further assume μ and Σ are known.

For a fixed time period T , the goal of this problem is to find a strategy π to maximize the following expression

$$E^\pi \left[\sum_{t=1}^T Y_t \right]. \quad (1)$$

Or equivalently, we are trying to find a policy that can minimize the Bayes risk under π :

$$\text{Risk}(T, \pi) = E [\text{Regret}(\theta, T, \pi)], \quad (2)$$

where the cumulative regret is defined as the following:

$$\text{Regret}(\theta_0, T, \pi) = \sum_{t=1}^T E \left[\max_{X \in \mathcal{U}_r} X \cdot \theta_0 - X_t \cdot \theta_0 | \theta = \theta_0 \right]. \quad (3)$$

3. PEGE

In *Linearly Parameterized Bandits* written by Paat Rusmevichientong and John N. Tsitsiklis, they proved the following theorem which gives a lower bound for the bayes risk:

Theorem 1. (Lower Bound) Consider a bandit problem where the set of arms is the unit sphere in \mathbb{R}^r , and ϵ_t has a standard normal distribution with mean 0 and variance one for all t and X_t . If θ has a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix \mathbf{I}_r/r , then for all policies π and every $T \geq r^2$,

$$\text{Risk}(T, \pi) \geq 0.006r\sqrt{T}.$$

In their paper, they also provided a simple algorithm called Phased Exploration and Greedy Exploitation (PEGE) (Algorithm 3) that reaches the corresponding lower bound, i.e the bayes risk for PEGE is $O(r\sqrt{T})$. In order for the algorithm to work, it requires the following two assumptions:

- Assumption 1.** • There exists a positive constant σ_0 such that for any $r \geq 2$, $\mathbf{u} \in \mathcal{U}_r$, $t \geq 1$ and $x \in \mathbb{R}$, we have $E[e^{x\epsilon_t}] \leq e^{\frac{x^2\sigma_0^2}{2}}$.
- There exists positive constants \bar{u} and λ_0 such that for any $r \geq 2$,

$$\max_{\mathbf{u} \in \mathcal{U}_r} \|\mathbf{u}\| \leq \bar{u}$$

and the set of arms $\mathcal{U}_r \subset \mathbb{R}^r$ has r linearly independent elements $\mathbf{b}_1, \dots, \mathbf{b}_r$ such that $\lambda_{\min}(\sum_{k=1}^r \mathbf{b}_k \mathbf{b}_k') \geq \lambda_0$.

Assumption 2. We say that a set of arms \mathcal{U}_r satisfies the smooth best arm response with parameter J (SBAR(J), for short) condition if for any nonzero vector $\mathbf{z} \in \mathbb{R}^r \setminus \{\mathbf{0}\}$, there is a unique set best arm $\mathbf{u}^*(\mathbf{z}) \in \mathcal{U}_r$ that gives the maximum expected reward, and for any two unit vectors $\mathbf{z} \in \mathbb{R}^r$ and $\mathbf{y} \in \mathbb{R}^r$ with $\|\mathbf{z}\| = \|\mathbf{y}\| = 1$, we have

$$\|\mathbf{u}^*(\mathbf{z}) - \mathbf{u}^*(\mathbf{y})\| \leq J\|\mathbf{z} - \mathbf{y}\|$$

Under these two conditions, the Bayes risk for the PEGE algorithm is at most $O(r\sqrt{T})$.

Algorithm 1 Phased Exploration and Greedy Exploitation

Description: For each cycle $c \geq 1$, complete the following two phases:

1. **Exploration (r periods)** For $k = 1, 2, \dots, r$, play arm $\mathbf{b}_k \in \mathcal{U}_r$ given in Assumption 1(b), and observe the reward $Y^{b_k}(c)$. Compute the OLS estimate $\hat{\theta}(c) \in \mathbb{R}^r$, given by

$$\begin{aligned}\hat{\theta}(c) &= \frac{1}{c} \left(\sum_{k=1}^r \mathbf{b}_k \mathbf{b}_k' \right)^{-1} \sum_{s=1}^c \sum_{k=1}^r \mathbf{b}_k Y^{b_k}(s) \\ &= \mathbf{Z} + \frac{1}{c} \left(\sum_{k=1}^r \mathbf{b}_k \mathbf{b}_k' \right)^{-1} \sum_{s=1}^c \sum_{k=1}^r \mathbf{b}_k \epsilon^{b_k}(s)\end{aligned}$$

where for any k , $Y^{b_k}(s)$, and $\epsilon^{b_k}(s)$ denote the observed reward and the error random variable associated with playing arm \mathbf{b}_k in cycle s .

2. **Exploitation (c periods)** Play the greedy arm $\mathbf{G}(c) = \arg \max_{v \in \mathcal{U}^r} \mathbf{v}' \hat{\theta}(c)$ for c periods.
-

Theorem 2. Suppose that assumption 1 holds and that the set \mathcal{U}_r satisfy the SBAR(J) condition. In addition, there exists a constant $M > 0$ such that for every $r \geq 2$ we have $E[\|\theta\|] \leq M$ and $E[1/\|Z\|] \leq M$. Then, there exist a positive constant a_1 that depends only on σ_0 , \bar{u} , λ_0 , J and M , such that for any $T \geq r$,

$$\text{Risk}(T, \text{PEGE}) \leq a_1 r \sqrt{T}$$

The proof of the theorem is you calculate the bayes risk for the exploitation and exploration periods respectively and add them together. In order to prove the above mentioned theorem, it requires the following two lemmas:

Lemma 3. Under Assumption 1, there exists a positive constant h_1 that depends only on σ_0 , \bar{u} and λ_0 such that for any $\mathbf{z} \in \mathbb{R}^r$ and $c \geq 1$,

$$E \left[\|\hat{\theta}(c) - \theta_0\|^2 | \theta = \theta_0 \right] \leq \frac{h_1 r}{c} \quad (4)$$

Lemma 4. Suppose that Assumption 1 holds and the set \mathcal{U}_r satisfy the SBAR(J) condition. Then, there exists a positive constant h_2 that depends only on σ_0 , \bar{u} , λ_0 and J , such that for any $\mathbf{z} \in \mathbb{R}^r$ and $c \geq 1$,

$$\begin{aligned}E \left[\max_{\mathbf{u} \in \mathcal{U}_r} \theta_0'(\mathbf{u} - \mathbf{G}(c)) | \theta = \theta_0 \right] \\ \leq \frac{2}{\|\theta_0\|} E \left[\|\hat{\theta}(c) - \theta_0\|^2 | \theta = \theta_0 \right] \leq \frac{2h_1 r}{c \|\theta_0\|} = \frac{h_2 r}{c \|\theta_0\|} \quad (5)\end{aligned}$$

Regret of PEGE We analyse the convergence order of PEGE based on the above assumptions and lemmas:

In the k th step of an r -period exploration, we have the observation:

$$E \left[\max_{X \in \mathcal{U}_r} X \cdot \theta_0 - \mathbf{b}_k \cdot \theta_0 | \theta = \theta_0 \right] \leq 2\bar{u} \cdot \|\theta_0\|$$

Suppose there are C cycles in the learning (the last cycle may not be completed), then the number of steps is

$$T \geq r(C-1) + C(C-1)/2 \quad (6)$$

Thus, $C \leq \sqrt{2T}$.

$$\begin{aligned}\text{Regret}(\theta_0, T, \text{PEGE}) &\leq \sum_{c=1}^C \left[2\bar{u} \|\theta_0\| \cdot r + \frac{r h_2}{c \|\theta_0\|} \cdot c \right] \\ &= (2\bar{u} \|\theta_0\| + \frac{h_2}{\|\theta_0\|}) r C < \sqrt{2} (2\bar{u} \|\theta_0\| + \frac{h_2}{\|\theta_0\|}) r \sqrt{T} \quad (7)\end{aligned}$$

Although the PEGE algorithm reaches the theoretical bayes risk lower bound, it doesn't perform well in practice. Thus, we proposed the following three modified PEGE algorithms, which still have $O(r\sqrt{T})$ bayes risk, but with a smaller coefficient.

Modification 1. In the first modified algorithm, instead of doing c periods of exploitation in cycle c , we now do kc periods of exploitation, where k is a constant (Algorithm 1).

Algorithm 2 PEGE Modified 1

Description: For each cycle $c \geq 1$, complete the following two phases:

1. **Exploration (r periods)** For $k = 1, 2, \dots, r$, play arm $\mathbf{b}_k \in \mathcal{U}_r$ given in Assumption 1(b), and observe the reward $Y^{b_k}(c)$. Compute the OLS estimate $\hat{\theta}(c) \in \mathbb{R}^r$, given by

$$\begin{aligned}\hat{\theta}(c) &= \frac{1}{c} \left(\sum_{k=1}^r \mathbf{b}_k \mathbf{b}_k' \right)^{-1} \sum_{s=1}^c \sum_{k=1}^r \mathbf{b}_k Y^{b_k}(s) \\ &= \mathbf{Z} + \frac{1}{c} \left(\sum_{k=1}^r \mathbf{b}_k \mathbf{b}_k' \right)^{-1} \sum_{s=1}^c \sum_{k=1}^r \mathbf{b}_k \epsilon^{b_k}(s)\end{aligned}$$

where for any k , $Y^{b_k}(s)$, and $\epsilon^{b_k}(s)$ denote the observed reward and the error random variable associated with playing arm \mathbf{b}_k in cycle s .

2. **Exploitation (kc periods)** Play the greedy arm $\mathbf{G}(c) = \arg \max_{v \in \mathcal{U}^r} \mathbf{v}' \hat{\theta}(c)$ for kc periods.
-

Regret of Modification 1: In this case, we modify the regret bound in equation (6), (7), and get

$$T \geq r(C-1) + kC(C-1)/2 \quad (8)$$

When $k < 2r$, $C \leq \sqrt{\frac{2T}{k}}$. Therefore

$$\begin{aligned} \text{Regret}(\theta_0, T, \text{PEGE1}) &\leq (2\bar{u}\|\theta_0\| + k \frac{h_2}{\|\theta_0\|})rC \\ &\leq \sqrt{2}(2\bar{u}\|\theta_0\|/\sqrt{k} + \sqrt{k} \frac{h_2}{\|\theta_0\|})r\sqrt{T} \end{aligned} \quad (9)$$

A suitable k which minimises this upper bound is

$$k^* = \frac{2\bar{u}\|\theta_0\|^2}{h_2}$$

However, since we do not know $\|\theta_0\|$ in advance, we need to estimate it base on the prior distribution of θ and may also adjust our estimation adaptively based on the results we get in each step.

Modification 2. In the second modified algorithm (Algorithm 2), instead of doing exploration and exploitation in turns, we stop the exploration if

$$\|\hat{\theta}(c) - \hat{\theta}(c+1)\| < \delta.$$

Algorithm 3 PEGE Modified 2

Description: For each cycle $c \geq 1$, complete the following two phases:

1. **Exploration (r periods)** For $k = 1, 2, \dots, r$, play arm $\mathbf{b}_k \in \mathcal{U}_r$ given in Assumption 1(b), and observe the reward $Y^{b_k}(c)$. Compute the OLS estimate $\hat{\theta}(c) \in \mathbb{R}^r$, given by

$$\begin{aligned} \hat{\theta}(c) &= \frac{1}{c} \left(\sum_{k=1}^r \mathbf{b}_k \mathbf{b}_k' \right)^{-1} \sum_{s=1}^c \sum_{k=1}^r \mathbf{b}_k Y^{b_k}(s) \\ &= \theta + \frac{1}{c} \left(\sum_{k=1}^r \mathbf{b}_k \mathbf{b}_k' \right)^{-1} \sum_{s=1}^c \sum_{k=1}^r \mathbf{b}_k \epsilon^{b_k}(s) \end{aligned}$$

where for any k , $Y^{b_k}(s)$, and $\epsilon^{b_k}(s)$ denote the observed reward and the error random variable associated with playing arm \mathbf{b}_k in cycle s .

2. **Exploitation (c periods)** Play the greedy arm $\mathbf{G}(c) = \arg \max_{v \in \mathcal{U}^r} \mathbf{v}' \hat{\theta}(c)$ for c periods.
-

Regret of Modification 2: Now we discuss the optimal δ and the regret:

Theorem 5. Suppose $E \left[\|\hat{\theta}(c_0) - \theta_0\|^2 \mid \theta = \theta_0 \right] < \epsilon$, then doing pure exploitation thereafter will also give us $O(r\sqrt{T})$ regret.

Proof. Denote $N = T - (c_0 r + \frac{c_0(c_0+1)}{2})$ and suppose we can do K more cycles in the rest of the N periods. In order

to show the new algorithm reaches $O(r\sqrt{T})$ regret bound, we just need to show

$$\frac{2\epsilon N}{\|\theta_0\|} \leq (2\bar{u}\|\theta_0\| + \frac{h_2}{\|\theta_0\|})rK$$

This is equivalent to show

$$\epsilon \leq \frac{(2\bar{u}\|\theta_0\|^2 + h_2)rK}{2N}.$$

□

However, since we can not observe $E \left[\|\hat{\theta}(c_0) - \theta_0\|^2 \mid \theta = \theta_0 \right]$ during the forwarding process, the following theorem shows us how to use $\|\hat{\theta}(c) - \hat{\theta}(c+1)\| = \delta$ to construct a upper bound.

Theorem 6. *sda*

Proof. We first consider the case where T is fixed. For a fixed θ_0 , we consider the regret over one sample path S . Suppose we do pure exploitation after $\|\hat{\theta}(c_0(S)) - \hat{\theta}(c_0(S)+1)\|^2 < \delta$, then we know the total regret for this sample path is

$$(2\bar{u}\|\theta_0\| + \frac{h_2}{\|\theta_0\|})c_0(S)r + \frac{2\|\hat{\theta}(c(S)) - \theta_0\|^2 N(S)}{\|\theta_0\|}, \quad (10)$$

where $N(S)$ denote the remaining steps after $c_0(S)$ cycles. We first take the expectation with respect to different θ_0 , since the sample path S is the same, we know $c_0(S)$ will remain the same (This is because in the OLS, the ϵ^{b_k} will remain the same). Without loss of generality, we assume $\max\{E[\|\theta\|], E[\frac{1}{\|\theta\|}]\} \leq M$. Thus, the expectation of the regret over different θ is less or equal to

$$M(2\bar{u} + h_2)c_0(S)r + 2ME_\theta \|\hat{\theta}(c(S)) - \theta_0\|^2 N(S).$$

Now we take the expectation with respect to different sample paths, notice that

$$\begin{aligned} &E_\theta \|\hat{\theta}(c_0(S)) - \theta_0\|^2 \\ &= \frac{1}{c_0(S)} \sum_{s=1}^{c_0(S)} \sum_{k=1}^r (\epsilon^{b_k}(s))^2 b_k' B^2 b_k, \end{aligned}$$

thus the bayes risk is

Bayes Risk

$$\begin{aligned} &\leq M(2\bar{u} + h_2)E_S[c_0(S)]r \\ &\quad + 2ME_{S,\theta} \|\hat{\theta}(c(S)) - \theta_0\|^2 N(S) \\ &\leq M(2\bar{u} + h_2)r\sqrt{2T} \\ &\quad + 2MN(S)E_S \left[\frac{1}{c_0(S)} \sum_{s=1}^{c_0(S)} \sum_{k=1}^r (\epsilon^{b_k}(s))^2 b_k' B^2 b_k \right] \end{aligned}$$

Since

$$\begin{aligned}
 & E_S \left[\frac{1}{c_0(S)} \sum_{s=1}^{c_0(S)} \sum_{k=1}^r (\epsilon^{b_k}(s))^2 b'_k B^2 b_k \right] \\
 &= E \left[E \left[\frac{1}{c_0(S)} \sum_{s=1}^{c_0(S)} \sum_{k=1}^r (\epsilon^{b_k}(s))^2 b'_k B^2 b_k \mid c_0(S) \right] \right] \\
 &= E \left[\frac{1}{c_0(S)} \sum_{s=1}^{c_0(S)} \sum_{k=1}^r E[(\epsilon^{b_k}(s))^2 \mid c_0(S)] b'_k B^2 b_k \right] \\
 &\leq E \left[\frac{1}{c} \sum_{s=1}^c \sum_{k=1}^r E[(\epsilon^{b_k}(s))^2] b'_k B^2 b_k \right] \\
 &\leq E[(c_0(S) + 1) \left(\frac{1}{c_0(S)} - \frac{1}{c_0(S) + 1} \right) \\
 &\quad \sum_{s=1}^{c_0(S)} \sum_{k=1}^r E[(\epsilon^{b_k}(s))^2] b'_k B^2 b_k] \\
 &\leq E_S[(c_0(S) + 1) \|\theta(c_0(S)) - \theta(c_0(\hat{S}) + 1)\|] \\
 &\leq \delta \sqrt{2T}
 \end{aligned}$$

Theorem 7.

$$\begin{aligned}
 & E \left[\|\hat{\theta}(c) - \theta_0\|^2 \mid \|\hat{\theta}(c) - \hat{\theta}(c+1)\| = \delta, \theta = \theta_0 \right] \\
 &= \frac{\delta^2}{c^2(c+1)^2} + \frac{1}{c+1} \text{tr} \left(\left(\sum_{k=1}^r \mathbf{b}_k \mathbf{b}'_k \right)^{-1} \right)
 \end{aligned}$$

Proof. Let $W(c) = (\sum_{k=1}^r \mathbf{b}_k \mathbf{b}'_k)^{-1} \sum_{k=1}^r \mathbf{b}_k \epsilon_{c_k}$, for which ϵ_{c_k} is the measurement error of the k th period of exploration in the c th cycle. Then $E[W(c)] = \mathbf{0}$, $\text{Cov}(W(c)) = (\sum_{k=1}^r \mathbf{b}_k \mathbf{b}'_k)^{-1}$. We construct a r -dimensional Brownian motion $B(t)$ such that:

$$\begin{aligned}
 E[dB]/dt &= 0 \\
 E[d(BB')]/dt &= \left(\sum_{k=1}^r \mathbf{b}_k \mathbf{b}'_k \right)^{-1}
 \end{aligned}$$

Then it is easy to prove that

$$(B(1), B(2), \dots, B(c), B(c^2 + c))$$

has the same distribution as

$$(\hat{\theta}(1) - \theta_0, 2(\hat{\theta}(2) - \theta_0), \dots, c(\hat{\theta}(c) - \theta_0), c(c+1)(\hat{\theta}(c) - \hat{\theta}(c+1))).$$

According to the result from Brownian bridge, we know that

$$\begin{aligned}
 E[B(c) \mid B(c^2 + c)] &= \frac{1}{c+1} B(c^2 + c) \\
 \text{Cov}[B(c) \mid B(c^2 + c)] &= \frac{c^2}{c+1} \left(\sum_{k=1}^r \mathbf{b}_k \mathbf{b}'_k \right)^{-1}
 \end{aligned}$$

Then

$$\begin{aligned}
 & E[\|B(c)\|^2 \mid B(c^2 + c)] \\
 &= (E[B(c) \mid B(c^2 + c)])^2 + \text{tr}(\text{Cov}[B(c) \mid B(c^2 + c)]) \\
 &= \left(\frac{B(c^2 + c)}{c+1} \right)^2 + \frac{c^2}{c+1} \text{tr} \left(\left(\sum_{k=1}^r \mathbf{b}_k \mathbf{b}'_k \right)^{-1} \right)
 \end{aligned}$$

Similarly, we have

$$\begin{aligned}
 & E \left[\|\hat{\theta}(c) - \theta_0\|^2 \mid \|\hat{\theta}(c) - \hat{\theta}(c+1)\| = \delta \right] \\
 &= \frac{1}{c^2} \left(\frac{\delta^2}{(c+1)^2} + \frac{c^2}{c+1} \text{tr} \left(\left(\sum_{k=1}^r \mathbf{b}_k \mathbf{b}'_k \right)^{-1} \right) \right) \\
 &= \frac{\delta^2}{c^2(c+1)^2} + \frac{1}{c+1} \text{tr} \left(\left(\sum_{k=1}^r \mathbf{b}_k \mathbf{b}'_k \right)^{-1} \right)
 \end{aligned}$$

□

Modification 2* At the c th cycle, conditional on $\hat{\theta}(1), \dots, \hat{\theta}(c-1), \hat{\theta}(c)$, we can get a more delicate estimation $E[\|\hat{\theta}(c) - \theta_0\|^2 \mid \hat{\theta}(1), \dots, \hat{\theta}(c-1), \hat{\theta}(c)]$ instead of $E[\|\hat{\theta}(c) - \theta_0\|]$. As we have showed, $\hat{\theta}(c) - \theta_0 = \frac{1}{c} \sum_{s=1}^c W(s)$, where $W(s)$ are i.i.d. r -dimensional normal variables. Suppose $W(i) \sim \mathcal{N}(0, \Sigma)$. Thus, we know:

$$\begin{aligned}
 \frac{1}{2}W(1) - \frac{1}{2}W(2) &= \hat{\theta}(1) - \hat{\theta}(2) \\
 \frac{2}{3}W(1) - \frac{1}{3}W(2) - \frac{1}{3}W(3) &= \hat{\theta}(1) - \hat{\theta}(3) \\
 \dots &\dots \\
 \frac{c-1}{c}W(1) - \frac{1}{c}W(2) - \dots - \frac{1}{c}W(c) &= \hat{\theta}(1) - \hat{\theta}(c)
 \end{aligned}$$

We can represent it as a linear system:

$$\begin{aligned}
 A\vec{W} &= \vec{b} \\
 A &= \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \frac{c-1}{c} & -\frac{1}{c} & -\frac{1}{c} & \dots & -\frac{1}{c} \end{pmatrix} \\
 \vec{W} &= \begin{pmatrix} W'(1) \\ \dots \\ W'(c) \end{pmatrix} \\
 \vec{b} &= \begin{pmatrix} \hat{\theta}'(1) - \hat{\theta}'(2) \\ \hat{\theta}'(1) - \hat{\theta}'(3) \\ \dots \\ \hat{\theta}'(1) - \hat{\theta}'(c) \end{pmatrix}
 \end{aligned}$$

The dimension of A is $(c-1)$ by c . Using QR decomposition, we can decompose $A = LQ$, where L is $(c-1)$ by c lower triangular matrix, Q is c by c orthogonal matrix. Let $\vec{V} = Q\vec{W}$. Then

$$L\vec{V} = \vec{b}$$

We can solve the first $(c-1)$ rows of \vec{V} while the last row of \vec{V} remains unknown. Because Q is orthogonal,

$V(1), \dots, V(c)$ are i.i.d. normal variables with the same distribution as $W(1), \dots, W(c)$ have. Let

$$L = (L_1, \vec{0})$$

where L_1 is $(c-1)$ by $(c-1)$ invertible lower triangular matrix. Therefore,

$$(V(1), \dots, V(c-1))' = L_1^{-1} \vec{b}$$

Since $V(c)$ is independent from $V(1), \dots, V(c-1)$, $E[V(c)] = 0$, $Cov(V(c)) = \Sigma$.

$$\begin{aligned} & E[\|\hat{\theta}(c) - \theta_0\|^2 | \hat{\theta}(1), \dots, \hat{\theta}(c)] \\ &= \frac{1}{c^2} E[\|\sum_{s=1}^c W(s)\|^2 | \hat{\theta}(1), \dots, \hat{\theta}(c)] \\ &= \frac{1}{c^2} E[\|\sum_{s=1}^c V(s)\|^2 | \hat{\theta}(1), \dots, \hat{\theta}(c)] \\ &= \frac{1}{c^2} (\|(1, 1, \dots, 1)L_1^{-1}\vec{b}\|^2 + tr(\Sigma)) \end{aligned}$$

In each cycle, we can compute $L_1^{-1}\vec{b}$. If $\|(1, 1, \dots, 1)L_1^{-1}\vec{b}\| \leq \delta$, then

Modification 3. In the third modified algorithm, it further modified the second algorithm: during the exploitation period, it also updates the ordinary least estimate each time (Algorithm 3).

Algorithm 4 PEGE Modified 3

Description: For each cycle $c \geq 1$, complete the following two phases:

1. **Exploration (r periods)** For $k = 1, 2, \dots, r$, play arm $\mathbf{b}_k \in \mathcal{U}_r$ given in Assumption 1(b), and observe the reward $Y^{b_k}(c)$. Compute the OLS estimate $\hat{\theta}(c) \in \mathbb{R}^r$, given by

$$\begin{aligned} \hat{\theta}(c) &= \frac{1}{c} \left(\sum_{k=1}^r \mathbf{b}_k \mathbf{b}_k' \right)^{-1} \sum_{s=1}^c \sum_{k=1}^r \mathbf{b}_k Y^{b_k}(s) \\ &= \mathbf{Z} + \frac{1}{c} \left(\sum_{k=1}^r \mathbf{b}_k \mathbf{b}_k' \right)^{-1} \sum_{s=1}^c \sum_{k=1}^r \mathbf{b}_k \epsilon^{b_k}(s) \end{aligned}$$

where for any k , $Y^{b_k}(s)$, and $\epsilon^{b_k}(s)$ denote the observed reward and the error random variable associated with playing arm \mathbf{b}_k in cycle s .

2. **Exploitation (c periods)** Play the greedy arm $\mathbf{G}(c) = \arg \max_{v \in \mathcal{U}^r} \mathbf{v}' \hat{\theta}(c)$ for c periods.
-

Regret of Modification 3: Since during the exploration periods, it will play the same arms as in algorithm 2, we just

need to show the regret during the exploitation periods is smaller than the regret of algorithm 2. If we denote $\hat{\theta}(c, t)$ as the OLS estimator of θ during the t_{th} exploitation period in the c_{th} cycle, then to show algorithm 3 has smaller regret during exploitation is equivalent to show

$$E[\|\hat{\theta}(c, t) - \theta_0\|^2 | \theta = \theta_0] < E[\|\hat{\theta}(c) - \theta_0\|^2 | \theta = \theta_0]. \quad (11)$$

Since we assume $\epsilon \sim N(0, \sigma^2)$, thus the OLS estimator $\hat{\beta}$ of $Y = X\beta + \epsilon$ follows a normal distribution $\hat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1})$. Thus, if we denote $X(c, t)$ is the X vector for the $\theta(\hat{c}, t)$ and $X(c)$ is the X vector for the $\hat{\theta}(c)$, then $X(c, t)$ contains more rows than $X(c)$. Thus, $X(c)^T X(c) < X(c, t)^T X(c, t)$ elementwise. Thus, $\hat{\theta}(c, t) - \theta_0$ has a smaller covariance matrix than $\hat{\theta}(c)$, which implies (11) is true. ???????????

4. Heuristic Algorithms

There are two well known algorithms for this problem, which are Exponential Gradient algorithm and the upper confidence bound algorithm.

Notations:

- x_i : feature of the recommended restaurant at step i , binary vector
- y_i : rate given by the user at step i , range is $(0, 5)$
- θ : user's preference
- μ_0 : prior knowledge of $E[\theta]$
- Σ_0 : prior knowledge of $Cov(\theta)$

At each step, we assume the user's rating $y_i = \theta \cdot x_i + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

4.1. Upper Confidence Bound

At step $t + 1$, we have knowledge $\mu_0, \Sigma_0, (x_1, y_1), \dots, (x_t, y_t)$. The distribution

$$\theta | [\mu_0, \Sigma_0, (x_1, y_1), \dots, (x_t, y_t)] \sim \mathcal{N}(\mu_t, \Sigma_t)$$

where

$$\Sigma_t^{-1} = \Sigma_0^{-1} + \frac{1}{\sigma^2} X_t^T X_t \quad = \Sigma_t^{-1} + \frac{1}{\sigma^2} x_t x_t^T \quad (12)$$

$$\mu_t = \Sigma_t \left(\Sigma_0^{-1} \mu_0 + \frac{1}{\sigma^2} X_t^T y \right) = \Sigma_t \left(\Sigma_{t-1}^{-1} \mu_{t-1} + \frac{1}{\sigma^2} x_t y_t \right) \quad (13)$$

where

$$X_t = \begin{pmatrix} x_1^T \\ \vdots \\ x_t^T \end{pmatrix}$$

Then, for each restaurant r , we suppose its feature vector is $x^{(r)}$. And its expected rating and variance are

$$E[x^{(r)} \cdot \theta] = x^{(r)} \cdot \mu_t, \quad \text{Var}(x^{(r)} \cdot \theta) = (x^{(r)})^T \Sigma_t x^{(r)}$$

The restaurant we will recommend at step $t + 1$ is

$$r_t = \arg \max_r E[x^{(r)} \cdot \theta] + 1.96 \sqrt{\text{Var}(x^{(r)} \cdot \theta)}$$

4.2. Upper Confidence Bound Algorithms

Upper confidence bound algorithms (UCB) are widely used in multiarmed bandit problems because these algorithms usually have good empirical performance. However, they are not as fast as other methods like the EXP3 algorithm, and this may be a problem when the number of stages is very big.

UCB algorithms follow two steps. First, at time t , for each arm u we compute an upper confidence bound $U_t(u)$. Then, at time t we choose the arm u^* such that $u^* \in \arg \max_u U_t(u)$, i.e. u^* has the maximal upper confidence bound.

In our project, we start supposing that $\theta \sim N(\mu_0, \Sigma_0)$ where μ_0 and Σ_0 are computed using the available data. We have that $\theta \sim N(\mu_t, \Sigma_t)$ at step $t + 1$, where μ_t, Σ_t are the parameters of the posterior distribution of θ after we have seen $x_1, y_1, \dots, x_t, y_t$. Specifically,

$$\begin{aligned} \mu_t &= \Sigma_t \left(\Sigma_0^{-1} \mu_0 + \frac{1}{\sigma^2} X_t^T y \right) \\ &= \Sigma_t \left(\Sigma_{t-1} \mu_{t-1} + \frac{1}{\sigma^2} x_t y_t \right) \\ \Sigma_t^{-1} &= \Sigma_0^{-1} + \frac{1}{\sigma^2} X_t^T X_t = \Sigma_t^{-1} + \frac{1}{\sigma^2} x_t x_t^T \end{aligned}$$

where

$$X_t = \begin{pmatrix} x_1^T \\ \vdots \\ x_t^T \end{pmatrix}.$$

Consequently an upper bound of the .95—confidence interval of the distribution of $y(u)$ is

$$E_{\mu_t, \Sigma_t} [x^{(u)} \cdot \theta] + 1.96 \sqrt{\text{Var}_{\mu_t, \Sigma_t} (x^{(u)} \cdot \theta)}$$

and then we will choose the arm $u^* = \arg \max_u \left(E_{\mu_t, \Sigma_t} [x^{(u)} \cdot \theta] + 1.96 \sqrt{\text{Var}_{\mu_t, \Sigma_t} (x^{(u)} \cdot \theta)} \right)$ at step $t + 1$. In terms of our example, the arms are restaurants and the vector $x^{(u)}$ is a binary vector that represents the categories of the restaurant u .

4.3. Exponentiated Gradient Algorithm for Bandit Setting (EXP3)

This is a randomized algorithm. It keeps a vector of probabilities for each of the arms, and it chooses an arm according to this vector. The weight of the arm chosen is increased when the loss is small and decreased when the loss is high. Specifically, the algorithm is

1. Given $\gamma \in [0, 1]$, initialize the weights $w_u(1) = 1$ for each arm u .
2. At each step t :
 - (a) Set $p_u(t) = (1 - \gamma) \frac{w_u(t)}{\sum_j w_j(t)} + \frac{\gamma}{K}$ for each arm u .
 - (b) Draw the arm u_t chosen according to the distribution $p_{u_t}(t)$.
 - (c) Observe the loss $y_{u_t}(t)$.
 - (d) Set $w_{u_t}(t+1) = w_{u_t}(t) \exp\left(\gamma \frac{y_{u_t}(t)}{p_{u_t}(t)m}\right)$, and $w_j(t+1) = w_j(t)$ for all other arms.

This algorithm is usually fast, but if there are too many arms, the weights may be zero in most of the cases.

5. Numerical Experiment

In this simulation, we use the yelp academic dataset. The goal of this simulation is to find the favorite restaurant categories for a new user. There are 4596 restaurants in the dataset and each restaurant belongs to one or multiple categories. We first find the top twenty categories that has most restaurants, which are Pizza, Sandwiches, Food etc, and use those 20 categories as our feature. For each restaurant, if it belongs to certain category, then the corresponding element of its feature vector is 1 and 0 otherwise. So the feature vector of each restaurant is a 20 dimensional binary vector.

For each user, we calculated his user preference vector based on his rating and the restaurants' feature vectors that he rated using ridge regression (since there are not too many ratings, ordinary linear regression doesn't work here due to singularity). Then we calculated the sample mean and the sample variance of all users' preference vector and denote them as (μ, Σ) . We further assume that for each user's preference vector $\theta \sim N(\mu, \Sigma)$ and generate new user from this distribution.

6. Possible Applications

Recommender systems want to find the preference that a user would give to a subset of a finite set of items. They're widely applied to different problems. For example, they're used in Netflix where there are thousands of movies and

TV episodes. The biggest challenge of these problems is that there are millions of objects and hundreds of million of users, and so it's necessary to find a model that performs well and be sufficiently fast.

References

- 1] X Zhao, P Frazier, *Exploration vs. Exploitation in the Information Filtering Problem* [2] Paat Rusmevichientong, John N. Tsitsiklis, *Linearly Parameterized Bandits*