Quick Best Action Identification in Linear Bandit Problems

Jun Geng and Lifeng Lai

1 School of Electrical and Information Engineering, Harbin Institute of Technology, Harbin, China
2 Dept. of Electrical and Computer Engineering, University of California, Davis, CA
Emails: jgeng@hit.edu.cn, llai@ucdavis.edu

Abstract—In this paper, we consider a best action identification problem in the stochastic linear bandit setup with a fixed confident constraint. In the considered best action identification problem, instead of minimizing the accumulative regret as done in existing works, the learner aims to obtain an accurate estimate of the underlying parameter based on his action and reward sequences. To improve the estimation efficiency, the learner is allowed to select his action based on his historical information; hence the whole procedure is designed in a sequential adaptive manner. We first show that the existing algorithms designed to minimize the accumulative regret is not a consistent estimator and hence is not a good policy for our problem. We then characterize a lower bound on the estimation error for any policy. We further design a simple policy and show that the estimation error of the designed policy achieves the same scaling order as that of the derived lower bound.

I. INTRODUCTION

Multi-armed bandit problem is a canonical sequential decision problem that has a wide range of applications [1]–[5]. In the classic multi-armed bandit problem, at each time slot, a decision maker has to choose one of $K$ competing decisions or “arms”, and receives a reward related to certain unknown parameters from his selected decision. Based on the knowledge collected from his past decisions and the corresponding rewards, the decision maker can then carefully decide his future actions according to different goals. The most commonly used goal is to minimize the cumulative regret, which is the cumulative difference between the optimal reward that one can achieve when the underlying parameters are known and the reward of the action taken by the decision maker. This setup nicely captures “exploration versus exploitation” phenomena in sequential decision making, as a crucial tradeoff faced by the decision maker at each round is between “exploitation”, i.e. to choose the decision with the highest estimated expected rewards, and “exploration”, i.e. to choose other decisions so as to obtain better estimates of the expected rewards of these decisions. Recently, another goal named “best arm identification” has received significant attentions [6]–[12]. In the best arm identification problem, instead of minimizing the cumulative regret, the goal is to identify the best arm that provides the highest expected rewards with high probability. This setup is also known as pure exploration since the decision maker now has the freedom to explore all arms without having to worry about regrets incurred in these exploration actions.

A natural generalization of the classic multi-armed bandit problem is so called stochastic linear multi-armed bandit problem [13]. In the stochastic linear multi-armed bandit problem, the decision maker chooses his decision $x_t$ from an $d$-dimensional compact set $D$ and receives a reward $<x_t, \theta^*> + \eta_t$, in which $\theta^*$ is a fixed but unknown parameter and $\eta_t$ is noise. Defining the regret as the difference between the rewards of the best decisions when $\theta^*$ is known and the rewards of the selected decisions, existing works on the stochastic linear multi-armed bandit problem aim to minimize the total regret. For example, [13], [14] have proposed algorithms according to the optimism in the face of uncertainty (OFU) principle, and have shown the proposed algorithms are Hannan consistent.

In this paper, similar to the best arm identification problem studied in the classic multi-armed bandit setup, we consider the best action identification problem in the stochastic linear multi-armed bandit setup. More specifically, instead of aiming to minimize the cumulative regret, we aim to obtain an accurate estimation $\hat{\theta}$ of the unknown parameter $\theta^*$ under a fixed confidence constraint. In particular, the decision maker aims to minimize the total number of actions under the constraint that the estimation error $||\hat{\theta} - \theta^*||_2$ is under control with a large probability. We call this best action identification problem, as the best action $x_t$ should have the same direction as $\theta^*$.

In this paper, we first show that existing algorithms based on the OFU principle lead to inconsistent estimators of $\theta^*$ and hence are not suitable for the best action identification. Intuitively, the OFU algorithm keeps selecting the actions that are close to the current estimation $\hat{\theta}_t$ in each round since it aims to minimize the regret. As a result, all selected actions are concentrated in a small cone around the direction of the true underlying parameter $\theta^*$. The decision maker has to use the rewards of selected actions to estimate $\theta^*$, but the actions with similar directions only bring similar rewards. In other words, it is challenging for the decision...
We express the relationship between decisions and corresponding rewards in the matrix form as

\[ Y_t = X_t \theta^* + \eta_t, \tag{2} \]

in which \( Y_t = [y_1, y_2, \ldots, y_t]^T \), \( \eta_t = [\eta_1, \eta_2, \ldots, \eta_t]^T \) and \( X_t = [x_1^T, x_2^T, \ldots, x_t^T]^T \in \mathbb{R}^{d \times d} \). Denote \( \hat{\theta}_t \) as the estimate of \( \theta^* \) at time \( t \). The decision maker aims to design an efficient algorithm to select decisions \( X_t \) and accurately estimate the unknown parameter \( \theta^* \) based on his sequential information \( \{x_1, \ldots, x_t, y_1, \ldots, y_t\} \). The performance metric is specified as

\[ P(||\hat{\theta}_t - \theta^*||^2 \leq \epsilon) \geq 1 - \delta \tag{3} \]

for some given constant \( \epsilon > 0 \) and \( \delta \in (0, 1) \). That is, the decision maker should have strong confidence on the result that the estimation error is less than a small value \( \epsilon \) when the decision procedure is terminated. Since \( \{\eta_t\} \) is a sequence of sub-Gaussian random variable, we expect that \( \epsilon \) converges to zero and \( \delta \) decays exponentially with respect to \( t \) as \( t \to \infty \).

### III. Algorithms and Performances

A natural estimator for (2) is the ordinary least squares estimator

\[ \hat{\theta}_t = (X_t^T X_t)^{-1} X_t^T Y_t. \tag{4} \]

One difficulty with the above estimator is that \( X_t^T X_t \) is not invertible when its rank is deficient (e.g. \( t \leq d \)). In this paper we focus on the following class of estimators that are slight modification of (4)

\[ \hat{\theta}_t = (X_t^T X_t + W_0)^{-1} X_t^T Y_t, \tag{5} \]

in which \( W_0 \) is a positive definite matrix. This class of estimators are widely used in the regret minimization problems [13, 14]. For notation convenience, we define

\[ W_t := W_0 + X_t^T X_t. \tag{6} \]

It is easy to see that \( W_t \) is always positive definite; hence the inversion in (5) is always valid. We further note that \( W_t \) can be efficiently calculated using the recursive formula

\[ W_t = W_{t-1} + x_t x_t^T. \]

### A. Challenges of Existing Algorithms

The most well known algorithm for the stochastic linear bandit problem is designed according to the optimism in the face of uncertainty principle [14]. The idea of this algorithm is to use observations to construct a confidence set \( C_t \subset \mathbb{R}^d \) that contains the unknown parameter \( \theta^* \) with a high probability. The confidence set \( C_t \) is updated whenever the decision maker obtains a new reward \( y_t \). The algorithm then estimates the unknown parameter by

\[ \hat{\theta}_t = \arg\max_{\theta \in C_t} (\max_{x \in D_t} < x, \theta >) \]

and selects the next decision by solving

\[ x_t = \arg\max_{x \in D_t} < x, \hat{\theta}_t >. \]
In our context, for \( t = 1, 2, \ldots \), the algorithm designed according to the OFU principle can be expressed as:
\[
\hat{\theta}_t = (X_t^T X_t + W_0)^{-1}X_t^T Y_t, \quad (7) \\
C_t = \left\{ \theta \in \mathbb{R}^d : ||\hat{\theta}_t - \theta||_{W_t} \leq \beta \right\}, \quad (8) \\
x_t = \text{argmax}_{x \in D_t} < x, \hat{\theta}_t > - ||\hat{\theta}_t||_2^2. \quad (9)
\]

We note that the confidence region \( C_t \) is an ellipsoid with radius \( \beta \). The value of \( \beta \) is updated at every time slot according to newly obtained information.

Several existing works [13], [14] have shown that, if \( \beta \) is properly designed, the above algorithm has a small cumulative regret. Particularly, let \( x_t^* = \text{argmax}_{x \in D_t} < x, \theta^* > \) be the best decision for \( \theta^* \), let \( r_t = < x_t^*, \theta^* > - < x_t, \theta^* > \) be the regret at time \( t \) for taking decision \( x_t \) and let \( R_n = \sum_{t=1}^n r_t \) be the cumulative regret. [14] proved the following result.

**Theorem 1:** (Theorem 2 and Theorem 3 in [14]) Let \( W_0 = \kappa I, \kappa > 0 \). By setting
\[
\beta_t = \sigma^2 \sqrt{2 \log(\det(W_t)^{1/2}\det(\lambda I)^{-1/2}/\delta)} + \kappa^{1/2} S_t,
\]
then for any \( \delta > 0 \), with probability at least \( 1 - \delta \), \( \theta^* \) lies in the set \( C_t \). Further more, if for all \( t \) and all \( x \in D_t, < x, \theta^* > \in [-1, 1] \), then with probability at least \( 1 - \delta \), the cumulative regret satisfies
\[
\forall n \geq 0, \quad R_n \leq 4 \sqrt{nd \log(n + d/\delta)} + \kappa^{1/2} S_t.
\]

Theorem 1 indicates that the OFU algorithm is Hannan consistent, i.e., \( \lim_{n \to \infty} R_n/n = 0 \). However, in the following, we point out that the OFU algorithm leads to an inconsistent estimator of \( \theta^* \). The result is stated in the following theorem.

**Theorem 2:** If \( R_n/n \to 0 \) as \( n \to \infty \) with probability at least \( 1 - \delta \), then
\[
\lim_{t \to \infty} P(\|\hat{\theta}_t - \theta^*\|_2^2 \geq \sigma^2) \geq 1 - \delta. \quad (10)
\]

**Proof:** Please see Appendix A

**B. Proposed Algorithm and Performance Analysis**

Motivated by the discussion above, we propose a novel algorithm which leads to a consistent estimator with a fast convergence rate. The proposed algorithm is specified in Algorithm 1. To facilitate the presentation, for \( k = 1, 2, \ldots \), we use the following notations in Algorithm 1:
\[
\begin{align*}
X_{k,d} &= \left[ X_{(k-1)d+1}^T, X_{(k-1)d+2}^T, \ldots, X_{kd}^T \right]^T, \\
Y_{k,d} &= \left[ y_{(k-1)d+1}, y_{(k-1)d+2}, \ldots, y_{kd} \right]^T, \\
\eta_{k,d} &= \left[ \eta_{(k-1)d+1}, \eta_{(k-1)d+2}, \ldots, \eta_{kd} \right]^T.
\end{align*}
\]

The proposed algorithm adopts batch processing. In particular, the proposed algorithm initializes the first \( d \) decisions as a group of standard orthogonal basis. The decision maker updates the estimate \( \hat{\theta}_t \) whenever he collects \( d \) successive rewards. Furthermore, whenever a new estimate \( \hat{\theta}_t \) is calculated, the decision maker chooses next decision \( X_{t+1} \) as the direction of \( \hat{\theta}_t \), and selects another \( d - 1 \) decisions such
that these \(d\) decisions form another group of orthogonal basis. We emphasize that algorithms according to the OFU principle keep taking decisions that maximize the reward \(\langle x, \hat{\theta}_t \rangle\). In our context, the OFU algorithm will always select the decision with the same direction of \(\hat{\theta}_t\). However, in our proposed algorithm, among every successive \(d\) decisions, only one decision is on the direction of \(\hat{\theta}_t\); the rest of \(d-1\) decisions are orthogonal to \(\hat{\theta}_t\). This is the key difference between the OFU algorithm and our algorithm.

**Data:** the adaptively designed decisions \(x_1, \ldots, x_t\) and corresponding rewards \(y_1, \ldots, y_t\)

**Result:** the estimate \(\hat{\theta}_t\)

Initialization: select \(x_1, \ldots, x_d\) as a set of standard orthogonal basis.

for \(k = 1, 2, \ldots, \lceil t/d \rceil\) do

obtain rewards: \(Y_{k,d} = X_{k,d} \theta^* + \eta_{k,d}\);

update matrix: \(W_{kd} = W_{(k-1)d} + X_{k,d}^T X_{k,d}\);

estimate parameter: \(\hat{\theta}_{kd} = W_{kd}^{-1} X_{kd}^T Y_{kd}\);

choose decision: \(x_{kd+1} = \hat{\theta}_{kd}/||\hat{\theta}_{kd}||^2_2\), select \(\{x_{kd+1}, x_{kd+2}, \ldots, x_{(k+1)d}\}\) to be an orthogonal basis;

end

**Algorithm 1:** The Proposed Algorithm

The performance of the proposed algorithm is characterized in the following theorem.

**Theorem 3:** For the proposed algorithm, we have

\[
E[||\hat{\theta}_t - \theta^*||^2_2] \leq \frac{d^2}{t} \sigma^2(1 + o(1)).
\]

Furthermore, if \(\eta_t\) is a sub-Gaussian vector, then

\[
P\left(||\hat{\theta}_t - \theta^*||^2_2 \geq \frac{3\sigma^2 d^2}{t^{1/2}} + O\left(\frac{\sigma^2 d^2}{t}\right)\right) \leq e^{-t}
\]

**Proof:** Please see Appendix B

Theorem 3 characterizes our performance metric (3). In particular, \(d\) decays exponentially as \(t \to \infty\), and the bound of estimation error \(\epsilon\) shrinks to zero on the order \(O(t^{-1/2})\) for the proposed algorithm.

We now provide a lower bound of the mean square estimation error (MSE) for all possible sequential decision selection strategies and show that MSE reduces at most on order \(O(t^{-1})\).

**Theorem 4:** (Lower Bounds on MSE) Let \(\eta_t\) be a sub-Gaussian random variable with variance proxy \(\sigma^2\). If estimator (5) is adopted, then for any adaptively selected decision sequence \(\{x_i, i = 1, 2, \ldots, t\}\), we have

\[
E[||\hat{\theta}_t - \theta^*||^2_2] \geq \frac{1}{t} \sigma^2 + o\left(\frac{1}{t}\right).
\]

**Proof:** Please see Appendix C

Theorem 3 indicates that the convergence rate of MSE for the proposed algorithm is on order \(O(t^{-1})\), while Theorem 4 shows that the convergence rate of MSE cannot be faster than \(O(t^{-1})\). Hence, the proposed algorithm is order optimal.

IV. NUMERICAL SIMULATION

In this section, we provide a numerical example to illustrate the results obtained in this paper. In this numerical example, we set \(d = 5\), and compare the performance of the OFU algorithm and our proposed algorithm. In particular, the MSE of each algorithm is calculated by Monte Carlo method. In the simulation, the estimation procedure proceeds 3000 rounds; hence, for each trial, the decision maker has to adaptively make 3000 decisions. For each algorithm, we conduct \(10^5\) trials with randomly created underlying parameter \(\theta^*\), and record the estimation error at each round of decision. Then, the logarithm of MSE, which is estimated by the average of estimation error at each trial, at each decision round is illustrated in Figure 2.

In Figure 2, the blue solid line is the performance of the OFU algorithm and the red dash line is the performance of the proposed algorithm. The simulation result shows that the error of the OFU algorithm tends to be a constant as the number of decisions goes large; hence, the corresponding MSE also tends to a constant. However, the error of the proposed algorithm decays when the number of decisions grows, which indicates the estimation error tends to zero as the number of decisions goes to infinity. Hence, the proposed estimator is consistent.

V. CONCLUSION

In this paper, we have studied the problem of identifying the best action in the stochastic linear bandit setup with a fixed confidence constraint. We have shown that the existing OFU algorithm is an inconsistent estimator for the unknown parameter \(\theta^*\). We have proposed and analyzed a novel algorithm. We have shown that the proposed algorithm is consistent and that its mean square estimation error reduces on order \(O(t^{-1})\). Furthermore, we have shown that the probability that the estimation error is larger than \(t^{-1/2}\) decays exponentially with respect to \(t\).
Appendix A
Proof of the Theorem

Recall that the decision set is
\[ D_t = \{ x \in \mathbb{R}^d : ||x||_2^2 \leq 1 \}, \]
hence, for the OFU algorithm, it is easy to see that the decision selected by the decision maker is
\[ x_{t+1} = \arg\max_{x \in D_t} < x, \hat{\theta}_t > = \hat{\theta}_t/||\hat{\theta}_t||_2^2, \]
and the optimal decision with known \( \theta^* \) is
\[ x^* = \arg\max_{x \in D_t} < x, \theta^* > = \theta^*/||\theta^*||_2^2. \]
For notation convenience, we denote \( \hat{\theta}_t := \hat{\theta}_t/||\hat{\theta}_t||_2^2 \) and \( \theta = \theta^*/||\theta^*||_2^2 \). Then the cumulative regret can be written as
\[ R_n = \sum_{t=1}^{n} < x^*_t - x_t, \theta^* > = ||\theta^*||_2^2 \sum_{t=1}^{n} < \theta - \hat{\theta}_{t-1}, \theta >, \]
and the assumption \( \lim_{n \to \infty} R_n/n = 0 \) indicates
\[ \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} < \theta - \hat{\theta}_{t-1}, \theta > = 0. \]  

In the following, we calculate the lower bound of the estimation error. Since
\[ \hat{\theta}_t = W_t^{-1}X_t^T Y_t, \]
we have
\[ \hat{\theta}_t - \theta^* = W_t^{-1}X_t^T (X_t^T \theta^* + \eta_t) - \theta^* = -W_t^{-1}W_0 \theta^* + W_t^{-1}X_t^T \eta_t, \]  
where we have used (6).

Therefore,
\[ W_t^{1/2} (\hat{\theta}_t - \theta^*) = W_t^{-1/2} (X_t^T \eta_t - W_0 \theta^*), \]
and we have
\[ (\hat{\theta}_t - \theta^*)^T W_t (\hat{\theta}_t - \theta^*) = ||X_t^T \eta_t - W_0 \theta^*||^2_{W_t^{-1}}. \]

As a result, we have
\[ \left[ (\hat{\theta}_t - \theta^*)^T W_t (\hat{\theta}_t - \theta^*) \right]^{1/2} \leq ||X_t^T \eta_t||_{W_t^{-1}} - ||W_0 \theta^*||_{W_t^{-1}} \]
\[ \geq ||X_t^T \eta_t||_{W_t^{-1}} - \frac{\lambda_{\min}(W_0)}{\lambda_{\min}(W_t)} ||\theta^*||_2, \]  
where the last step is true since
\[ ||W_0 \theta^*||_{W_t^{-1}}^2 \leq \frac{1}{\lambda_{\min}(W_t)} ||W_0 \theta^*||_2^2 \leq \frac{\lambda_{\max}(W_0)}{\lambda_{\min}(W_0)} ||\theta^*||_2. \]
We note that
\[ W_t - t\theta\theta^T = W_0 + \sum_{i=1}^{t-1} x_i x_i^T - t\theta\theta^T = W_0 + (x_1 x_1^T - \theta\theta^T) + \sum_{i=1}^{t-1} (\hat{\theta} - \theta)(\hat{\theta} - \theta)^T. \]

Then, we have
\[ W_t = W_0 + (x_1 x_1^T - \theta\theta^T) + t\theta\theta^T + \sum_{i=1}^{t-1} 2\theta(\hat{\theta} - \theta) + \sum_{i=1}^{t-1} (\hat{\theta} - \theta)(\hat{\theta} - \theta)^T. \]

Therefore, for any \( t \rightarrow \infty \), the first item on the right-hand side of the equality
\[ \frac{1}{t-1} W_t \theta = \frac{1}{t-1}(W_0 + x_1 x_1^T)\theta + \theta\theta^T \theta \]

because \( x_1 \) and \( \theta \) have finite norms. The third term
\[ \frac{1}{t-1} \sum_{i=1}^{t-1} 2\theta(\hat{\theta} - \theta)^T \theta \]

is
\[ = 2\theta \left[ \frac{1}{t-1} \sum_{i=1}^{t-1} (\hat{\theta} - \theta) \right] - 0 \]

and the forth item
\[ = \frac{1}{t-1} \sum_{i=1}^{t-1} (\hat{\theta}_i - \theta) \leq \frac{1}{t-1} \sum_{i=1}^{t-1} \|\hat{\theta}_i - \theta\| \leq 2 \frac{1}{t-1} \sum_{i=1}^{t-1} \|\hat{\theta}_i - \theta\| \rightarrow 0. \]

because of (12). Since (12) holds with probability at least \( 1 - \delta \), then as \( t \rightarrow \infty \),
\[ \frac{1}{t-1} W_t \theta = \theta \]

holds with probability \( 1 - \delta \). That is, \( \theta \) is the eigenvector associated with eigenvalue 1 for matrix \( \frac{1}{t-1} W_t \). As \( t \rightarrow \infty \), we further have, with probability at least \( 1 - \delta \),
\[ \frac{1}{t-1} \eta_i^T X_i W_t^{-1} X_i^T \eta_i \]
\[ = \frac{1}{(t-1)^2} \eta_i^T X_t \left( \frac{1}{t-1} W_t \right)^{-1} X_i^T \eta_i \]
\[ = \frac{1}{(t-1)^2} (\eta_i^T X_t \theta)^2 \]
\[ = \frac{1}{(t-1)^2} \text{trace}(X_t \theta \eta_i^T X_i^T \eta_i) \]
\[ = \frac{1}{t-1} \text{trace} \left( X_t \theta \left( \frac{1}{t-1} \eta_i^T \eta_i \right) \theta^T X_i^T \right) \]
\[ = \frac{\sigma^2}{t-1} \text{trace}(X_t \theta \theta^T X_i^T X_i \theta) \]
\[ = \sigma^2. \] 

In above derivations, \( \frac{1}{t-1} W_t \) is a positive definite matrix, then \( \left( \frac{1}{t-1} W_t \right)^{-1} \) is a positive definite matrix sharing the same eigenvectors with \( \frac{1}{t-1} W_t \). Hence (a) holds because \( \frac{1}{t-1} W_t \) is a positive definite matrix. (b) is true, because for a rank 1 matrix \( A \), we have \( \text{trace}[A A^T] = \lambda(A)^2 = \text{trace}[A]^2 \). (c) is true because \( \lim_{t \rightarrow \infty} \eta_i^T \eta_i = \frac{\sigma^2}{t} \). (d) holds almost surely under the strong law of large number. (d) is true because
\[ \theta^T \frac{1}{t-1} X_i^T X_i \theta = \frac{1}{t-1} (W_t - W_0) \theta \]
\[ = \frac{1}{t-1} W_t \theta - \theta \frac{1}{t-1} W_0 \theta \]
\[ = \theta^T \theta - \frac{1}{t-1} \theta^T W_0 \theta = 1. \]

We also have
\[ (\hat{\theta}_t - \theta^*)^T \frac{1}{t-1} W_t (\hat{\theta}_t - \theta^*) \]
\[ \leq \lambda_{\max} \left( \frac{1}{t-1} W_t \right) (\hat{\theta}_t - \theta^*)^T(\hat{\theta}_t - \theta^*) \]
\[ \leq \text{trace} \left( \frac{1}{t-1} W_t (\hat{\theta}_t - \theta^*)^T(\hat{\theta}_t - \theta^*) \right) \leq \| \hat{\theta}_t - \theta^* \|^2. \]
Therefore we have

\[ \| \hat{\theta}_t - \theta^* \|_2^2 \geq (\hat{\theta}_t - \theta^*)^T \frac{1}{t-1} W_t (\hat{\theta}_t - \theta^*) \]

\[ \geq \frac{1}{t-1} \left[ \| X^T _t \eta \| W_t^{-1} - \frac{\lambda_{\max}(W_0)}{\lambda_{\min}(W_0)} \| \theta^* \|_2 \right]^2 \]

\[ \overset{(a)}{=} \frac{1}{t-1} \eta^T _t X_t W_t^{-1} X^T _t \eta_t = \sigma^2 \]

with probability at least \( 1 - \delta \). In (17), (a) and (b) are due to (14) and (16) respectively.

APPENDIX B
PROOF OF THEOREM 3

In this appendix, we show Theorem 3. Recall that

\[ X_{k,d} = \left[ x^T _{(k-1)d+1} x^T _{(k-1)d+2} \ldots x^T _{kd} \right]^T, \]

\[ Y_{k,d} = [y_{d(k-1)+1}, y_{d(k-1)+2}, \ldots, y_{dk}]^T, \]

\[ \eta_{k,d} = [\eta_{d(k-1)+1}, \eta_{d(k-1)+2}, \ldots, \eta_{dk}]^T. \]

Since \( x_{(k-1)d+1}, \ldots, x_{kd} \) are selected as orthogonal basis, then we have \( X_{k,d} X^T _{k,d} = I \). Furthermore, it is easy to see that \( X_{k,d} \) is independent of \( \eta_{k,d} \).

Let \( t = ld \), then

\[ W_t = W_0 + \sum_{k=1}^{l} X_{k,d} X^T _{k,d} = W_0 + lI. \]

Since \( \hat{\theta}_t - \theta^* = -W_{t-1} W_0 \theta^* + W_{t-1} X^T _t \eta_t \), we have

\[ \| \hat{\theta}_t - \theta^* \|_2 \leq \left( -W_{t-1} W_0 \theta^* + W_{t-1} X^T _t \eta_t \right)^T \]

\[ \left( -W_{t-1} W_0 \theta^* + W_{t-1} X^T _t \eta_t \right) \]

\[ = \theta^T W_0 W_{t-2} W_0 \theta^* + \eta^T _t X_t W_{t-2} X^T _t \eta_t \]

\[ -2 \theta^T W_0 W_{t-2} X^T _t \eta_t \]

\[ \overset{(a)}{\leq} \frac{1}{l^2} \theta^T W_0 W_0 \theta^* + \frac{1}{l^2} \eta^T _t X_t X^T _t \eta_t \]

\[ -2 \theta^T W_0 W_{t-2} X^T _t \eta_t, \]

\[ \overset{(b)}{= \lambda_{\max}(W_0)} = \sigma^2 \]

in which (a) is because of \( W_{t-2} \leq l^{-2} I \). To see this, note that \( W_t - I = W_0 \geq 0 \), hence we can obtain \( W_t \geq I \), which further indicates \( W_{t-1} \leq l^{-1} I \).

We then calculate the expectations of the three items in the right hand side of (18). We note that the first item \( \theta^T W_0 W_0 \theta^* \) is a constant. For the second item, we have

\[ \frac{1}{l^2} \mathbb{E}[\eta^T _t X_t X^T _t \eta_t] \]

\[ = \frac{1}{l^2} \mathbb{E} \left[ \sum_{i=1}^{l} \eta^T _{i,d} X_{i,d} \sum_{j=1}^{l} X^T _{j,d} \eta_{j,d} \right] \]

\[ = \frac{1}{l^2} \sum_{i=1}^{l} \sum_{j=1}^{l} \mathbb{E} [\eta^T _{i,d} X_{i,d} X^T _{j,d} \eta_{j,d}] \]

\[ = \frac{1}{l^2} \sum_{i=1}^{l} \sum_{j=1}^{l} \sum_{k=1}^{l} \mathbb{E} [\theta^T X_{j,d} \eta_{j,d} X^T _{i,d} X_{i,d}] \]

\[ = \frac{1}{l^2} \sum_{i=1}^{l} \sum_{j=1}^{l} \mathbb{E} [\theta^T X_{j,d} \eta_{j,d} X_{i,d} X_{i,d}] \]

\[ = \frac{1}{l^2} \sum_{i=1}^{l} \sigma^2 \text{trace}[I] = \frac{d^2}{t} \sigma^2, \]

(19)

in which (a) is true due to the fact that \( X_{k,d} \) is independent of \( \eta_{k,d} \). For the last item, we have

\[ \mathbb{E}[\theta^T W_0 W_0 \eta^T _t X^T _t \eta_t] \]

\[ = \theta^T W_0 (W_0 + lI)^{-1} \mathbb{E}[X^T _t \eta_t] \]

\[ = \theta^T W_0 (W_0 + lI)^{-1} \mathbb{E} \left[ \sum_{k=1}^{l} X_{k,d} \eta_{k,d} \right] = 0. \]

As a result,

\[ \mathbb{E}[\| \hat{\theta}_t - \theta^* \|_2^2] \leq \frac{1}{l^2} \theta^T W_0 \theta^* + \frac{d^2}{t} \sigma^2 \]

\[ \leq \frac{d^2}{l^2} \theta^T W_0 \theta^* + \frac{d^2}{t} \sigma^2. \]

Therefore, the first part of Theorem 3 is established.

We then show the second part of Theorem 3. To this end, we use the concentration inequality in Theorem 2.1 [17]. In particular, let \( A \in \mathbb{R}^{n \times n} \) be a square matrix and let \( x \in \mathbb{R}^{n \times 1} \) be a sub-Gaussian random vector with zero mean and proxy \( \sigma^2 \). It has been shown that

\[ \| Ax \|_2 \leq \sqrt{\text{trace}(AA^T) + 2 \sqrt{\text{trace}(AA^T^2)} t + \| A \|_2^2} \]

holds with probability at least \( 1 - e^{-t} \).

It is easy to find that

\[ \text{trace}(X^T _t X_t) = ld, \]

\[ \text{trace}(X^T _t X_t^2) = l^2d, \]

\[ \| X_t \|_2 = \lambda_{\max}(X^T _t X_t)^{1/2} = l^{1/2}. \]
Hence, using above tail inequality, we have
\[ P \left( \eta_t^T X_t X_t^T \eta_t \leq \sigma^2 (ld + 2\sqrt{ld} dt + \sqrt{lt^2}) \right) \geq 1 - e^{-t}. \quad (21) \]

Recall (18), we have
\[
\|\hat{\theta}_t - \theta^*\|^2 \leq \frac{1}{l^2} \theta^{*T} W_0 W_0 \theta^* + \frac{1}{l^2} \eta_t^T X_t X_t^T \eta_t
+ \frac{2\|\theta^*\|}{l} \theta^{T} W_0 \eta_t^2 X_t^T \eta_t. 
\]

For the last item on the right hand of the inequality, we have
\[
\|\theta^* W_0 \eta_t^2 X_t^T \eta_t\|
\leq \lambda_{\max}(W_0) \lambda_{\max}(W_t^{-2}) \|\theta^* \| X_t^T \eta_t
\leq \lambda_{\max}(W_0) \|\theta^* \| \|\eta_t^T X_t X_t^T \eta_t\|^{1/2}
\leq \lambda_{\max}(W_0) \|\theta^* \| \left( \frac{1}{l^2} \right) \|\eta_t^T X_t X_t^T \eta_t\|^{1/2}.
\]

Therefore
\[
\|\hat{\theta}_t - \theta^*\|^2 \leq \frac{1}{l^2} \theta^{*T} W_0 W_0 \theta^* + \frac{1}{l^2} \eta_t^T X_t X_t^T \eta_t
+ \frac{\lambda_{\max}(W_0) \|\theta^* \|}{l} \left( \frac{1}{l^2} \right) \|\eta_t^T X_t X_t^T \eta_t\|^{1/2}.
\]

Then the event
\[
\left\{ \eta_t^T X_t X_t^T \eta_t \leq \sigma^2 (ld + 2\sqrt{ld} dt + \sqrt{lt^2}) \right\}
\]
holds with probability at least 1 $- e^{-t}$ indicates that the event
\[
\left\{ \|\hat{\theta}_t - \theta^*\|^2 \leq \frac{1}{l^2} \theta^{*T} W_0 W_0 \theta^* + \frac{\sigma^2}{l^2} (ld + 2\sqrt{ld} dt + \sqrt{lt})
+ \lambda_{\max}(W_0) \|\theta^* \| \left( \frac{1}{l^2} \right) (ld + 2\sqrt{ld} dt + \sqrt{lt})^{1/2} \right\}
\]
holds with probability at least 1 $- e^{-t}$. Since $t = ld$, we then can obtain
\[
P \left( \|\hat{\theta}_t - \theta^*\|^2 \geq \frac{d^2}{t^2} \theta^{*T} W_0^2 \theta^* + \sigma^2 \left( \frac{d^2}{t} + 3 \sqrt{\frac{d^2}{t}} \right)
+ \lambda_{\max}(W_0) \|\theta^* \| \left( \frac{d^2}{t} + 3 \sqrt{\frac{d^2}{t}} \right)^{1/2} \right)
= P \left( \|\hat{\theta}_t - \theta^*\|^2 \geq \frac{3\sigma^2 d^{3/2}}{t^{1/2}} + \frac{\sigma^2 d^2}{t} \right)
\leq e^{-t}.
\]

**APPENDIX C**

**PROOF OF THEOREM**

For the estimator $\hat{\theta}_t = (X_t^T X_t + W_0)^{-1} X_t^T Y_t$, we have
\[
\hat{\theta}_t - \theta^* = -W_0^{-1} W_0 \theta^* + W_0^{-1} X_t^T \eta_t.
\]

Therefore
\[
(\hat{\theta}_t - \theta^*)^T W_0^2 (\hat{\theta}_t - \theta^*)
= (\theta^* \theta^* W_0 + \eta_t^2 X_t (-W_0 \theta^* + X_t^T \eta_t)
= \theta^* W_0^2 \theta^* - 2 \theta^* W_0 \theta^* + W_0^2 \eta_t^2 X_t^T \eta_t + \eta_t^2 X_t X_t^T \eta_t.
\]

We then analyze the expectation of the three items on the right hand side of the equality one by one. We note that the first item $\theta^* W_0^2 \theta^*$ is a constant; for the second item, we have
\[
E \left[ \theta^* W_0 X_t^T \eta_t \right] = \theta^* W_0 E \left[ \sum_{i=1}^t x_i \eta_i \right] = 0. \quad (23)
\]

For the third item, we have
\[
E[\eta_t^T X_t X_t^T \eta_t] = E \left[ \sum_{i=1}^t x_i \eta_i \right] = \left( \sum_{i=1}^t x_i \right)^2
\leq \sum_{i=1}^t \eta_t^2 X_t^T \eta_t + 2 \sum_{i=1}^t \sum_{j=i+1}^t \eta_t^2 x_j x_j
\leq \sum_{i=1}^t E \left[ \eta_t^2 x_i^2 x_i \right] + 2 \sum_{i=1}^t \sum_{j=i+1}^t E \left[ \eta_t^2 x_i x_j \right] = t \sigma^2 = t \sigma^2.
\]

Since $\eta_t$ is independent of $x_i$, $\eta_j$ is independent of $\eta_t^2 x_j x_j$ for $j > i$ and $\|x_i^2 \|^2 = 1$, we have
\[
E \left[ \eta_t^2 x_i x_i \right] = E \left[ x_i^2 x_i \right] E \left[ \eta_t^2 \right] = \sigma^2,
E \left[ \eta_t^2 x_i x_j \right] = E \left[ \eta_t^2 x_i \right] E \left[ \eta_j \right] = 0.
\]

Therefore
\[
E[\eta_t^T X_t X_t^T \eta_t] = \sum_{i=1}^t \sigma^2 = t \sigma^2. \quad (24)
\]

As a result
\[
E[(\hat{\theta}_t - \theta^*)^T W_0^2 (\hat{\theta}_t - \theta^*)] = \theta^* W_0^2 \theta^* + t \sigma^2.
\]

Since $W_t = W_0 + X_t^T X_t$, we have
\[
\lambda_{\max}(W_t) \leq \text{trace}(W_t) = \text{trace}(W_0) + \sum_{i=1}^t x_t^2 x_i
= \text{trace}(W_0) + t,
\]
\[
\lambda_{\max}(W_t^2) = \lambda_{\max}(W_t)^2 \leq (\text{trace}(W_0) + t)^2.
\]

Therefore,
\[
E[\|\hat{\theta}_t - \theta^*\|^2] \geq \frac{1}{\lambda_{\max}(W_t^2)} E[(\hat{\theta}_t - \theta^*)^T W_0^2 (\hat{\theta}_t - \theta^*)]
\geq \frac{\theta^* W_0^2 \theta^* + t \sigma^2}{(\text{trace}(W_0) + t)^2}. \quad (25)
\]
in which the first inequality is because of
\[(\hat{\theta}_t - \theta^*)^T W_t^2 (\hat{\theta}_t - \theta^*) \leq \lambda_{\max}(W_t^2) \| (\hat{\theta}_t - \theta^*) \|_2^2.\]

Then, the result of Theorem 4 can be obtained by taking \( t \to \infty \).