Stochastic Linear Bandits with Hidden Low Rank Structure

Sahin Lale [1] **Kamyar Azizzadenesheli** [2] **Animashree Anandkumar** [1] **Babak Hassibi** [1]

Abstract

High-dimensional representations often have a lower dimensional underlying structure. This is particularly the case in many decision making settings. For example, when the representation of actions is generated from a deep neural network, it is reasonable to expect a low-rank structure whereas conventional structures like sparsity are not valid anymore. Subspace recovery methods, such as Principle Component Analysis (PCA) can find the underlying low-rank structures in the feature space and reduce the complexity of the learning tasks. In this work, we propose Projected Stochastic Linear Bandit (PSLB), an algorithm for high dimensional stochastic linear bandits (SLB) when the representation of actions has an underlying low-dimensional subspace structure. PSLB deploys PCA based projection to iteratively find the low rank structure in SLBs. We show that deploying projection methods assures dimensionality reduction and results in a tighter regret upper bound that is in terms of the dimensionality of the subspace and its properties, rather than the dimensionality of the ambient space. We modify the image classification task into the SLB setting and empirically show that, when a pre-trained DNN provides the high dimensional feature representations, deploying PSLB results in significant reduction of regret and faster convergence to an accurate model compared to state-of-art algorithm.

1. INTRODUCTION

Stochastic linear bandit (SLB) is a class of sequential decision-making under uncertainty where an agent sequentially chooses actions from very large action sets. At each round, the agent applies its action, and as a response, the environment emits a stochastic reward whose expected value is an unknown linear function of the action. The agent's goal is to collect as much reward as possible over the course of T interactions.

In SLB, the actions are represented as d-dimensional vectors, and the agent maintains limited information about the unknown linear function of reward. Through the course of interaction, the agent implicitly or explicitly constructs the model of the environment. It dedicates the decisions to not only maximize the current reward but also explore other actions to build a better estimation of the unknown linear function and guarantee higher future rewards. This is known as the *exploration* vs. *exploitation* trade-off.

The lack of oracle knowledge of the true environment model causes the agent to make mistakes by picking sub-optimal actions during the exploration. While the agent examines actions in the decision set, its committed mistakes accumulate. The aim of the agent is to design a strategy to minimize the cumulative cost of these mistakes, known as regret. One promising approach to minimize the regret is through utilizing the *optimism in the face of uncertainty* (OFU) principle first proposed by Lai & Robbins (1985). OFU based algorithms estimate the environment model up to its confidence interval and construct a plausible set of models within that interval. Among those models in the plausible set, they choose the most optimistic one and follow the optimal behavior suggested by the selected model for the next round of decision making.

For general SLB problems, Abbasi-Yadkori et al. (2011) deploy the OFU principle, propose OFUL algorithm, and for d-dimensional SLB, derive a regret upper bound of $\widetilde{\mathcal{O}}\left(d\sqrt{T}\right)$ which matches the lower bound up to a log factor. These regret bounds in high dimensional problems especially when d and T are about the same order are not practically tolerable. Fortunately, real-world problems usually are not arbitrary and may contain hidden low-dimensional structures. For example in classical recommendation systems, each item is represented by a large and highly detailed hand-engineered feature vector; therefore d is intractably large. In these problems, not all the features are helpful for the recommendation task. For instance, the height of goods such as a pen is not a relevant feature for its recommendation while this feature is valuable for furnitures. Therefore the true underlying linear

[1] California Institute of Technology, Pasadena, CA, USA [2] University of California, Irvine, CA, USA. Correspondence to: Sahin Lale <alale@caltech.edu>.

function in SLBs is highly sparse. Abbasi-Yadkori et al. (2012) show how to exploit this additional structure and design a practical algorithm with regret of $\widetilde{\mathcal{O}}\left(\sqrt{sdT}\right)$ where s is the sparsity level of the true underlying linear function. Under slightly stronger assumptions, Carpentier & Munos (2012) show the theory of compressed sensing can provide a tighter bound of $\widetilde{\mathcal{O}}\left(s\sqrt{T}\right)$.

The contemporary success of Deep Neural Networks (DNN) in representation learning enables classical machine learning methods to provide significant advancements in many machine learning problems, e.g., classification and regression tasks (LeCun et al., 1998). DNNs convolve the raw features of the input and construct new feature representations which replace the hand-engineered feature vectors in many real-world sequential decision making applications, e.g., recommendation systems. However, when a DNN provides the feature representations, one cannot see a sparse structure.

Dimension reduction and subspace recovery form the core of unsupervised learning methods and principal component analysis (PCA) is the main technique for linear dimension reduction (Pearson, 1901; Eckart & Young, 1936). At each round of SLB, the agent chooses an action and receives the reward corresponding to that action. Therefore, the chosen action is assigned a supervised reward signal while other actions in the decision set remain unsupervised. Even though the primary motivation in the SLB framework is decision-making within a large and stochastic decision set, the majority of prior works do not exploit possible hidden structures in these sets. For example, Abbasi-Yadkori et al. (2011) only utilizes supervised actions, the actions selected by the algorithm, to construct the environment model. It ignores all other unsupervised actions in the decision set. On the contrary, large number of actions in the decision sets can be useful in reducing the dimension of the problem and simplifying the learning problem.

Contributions: In this paper, we deploy unsupervised subspace recovery using PCA to exploit the massive number of unsupervised actions which are observed in the decision sets of SLB and reduce the dimensionality and the complexity of SLBs. We propose PSLB for SLBs and show that if there exists an m-dimensional subspace structure such that the actions live in a perturbed region around this subspace, deploying PSLB improves the regret upper bound to $min\left\{\widetilde{\mathcal{O}}\left(\Upsilon\sqrt{T}\right), \widetilde{\mathcal{O}}\left(d\sqrt{T}\right)\right\}$. Here Υ represents the difficulty of subspace recovery as a function of the structure of the problem. If learning the subspace is hard, *e.g.*, the eigengap is small to analyze in a reasonable amount of samples, actions are widely distributed in the orthogonal dimensions of the subspace due to perturbation or $m \approx d$, then using projection approaches are not remedial. On the other hand, if underlying subspace is identifiable, *i.e.*, large number of actions are available from the decision sets in each round, the eigengap is significant or $m \ll d$, then using subspace recovery provides faster learning of the underlying linear function; thus, smaller regret.

We adapt the image classification tasks on MNIST (LeCun et al., 1998), CIFAR-10 (Krizhevsky & Hinton, 2009) and ImageNet (Krizhevsky et al., 2012) datasets to the SLB framework and apply both PSLB and OFUL on these datasets. We observe that there exists a low dimensional subspace in the feature space when a pre-trained DNN produces the d-dimensional feature vectors. We empirically show that using subspace recovery PSLB learns the underlying model significantly faster than OFUL and provides orders of magnitude smaller regret in SLBs obtained from MNIST, CIFAR-10, and ImageNet datasets.

2. Preliminaries

For any positive integer n, $[n]$ denotes the set $\{1, \dots, n\}$. The Euclidean norm of a vector x is denoted by $\|x\|_2$. The spectral norm of matrix A is denoted by $\|A\|_2$, *ie.*, $\|A\|_2 := \sup\{\|Ax\| : \|x\|_2 = 1\}$. $A^\dagger$ denotes the Moore-Penrose inverse of matrix A. For any symmetric and positive semi-definite matrix M, let $\|x\|_M$ denote the norm of a vector x defined as $\|x\|_M := \sqrt{x^T M x}$. The j-th eigenvalue of a symmetric matrix A is denoted by $\lambda_j(A)$, where $\lambda_1(A) \geq \lambda_2(A) \geq \dots$. The largest and smallest eigenvalue of A are denoted as $\lambda_{max}(A)$ and $\lambda_{min}(A)$, respectively. I_d denotes $d \times d$ identity matrix. If Y_i is a column vector then $\mathbf{Y}_t$ is a matrix whose columns are $Y_1, \dots, Y_t$ whereas if y_i is a scalar then $\mathbf{y}_t$ is a column vector whose elements are $y_1, \dots, y_t$. $\uplus_{i=1}^t D_i$ defines the multiset summation operation over the sets $D_1, \dots, D_t$.

Model: Let T be the total number of rounds. At each round $t \in [T]$, the agent is given a decision set D_t with K actions, $\hat{x}_{t,1}, \dots, \hat{x}_{t,K} \in \mathbb{R}^d$. Let V be an $d \times m$ orthonormal matrix with $m \leq d$, where $\mathrm{span}(V)$ defines a m-dimensional subspace in $\mathbb{R}^d$. Consider a zero meantrue action vector, $x_{t,i} \in \mathbb{R}^d$, such that $x_{t,i} \in \mathrm{span}(V)$ for all $i \in [K]$ and $t \in [T]$. Let $\psi_{t,i} \in \mathbb{R}^d$ be zero mean random vectors which are uncorrelated with true action vectors, *i.e.*, $\mathbb{E}[x_{t,i}\psi_{t,i}^T] = 0$ for all $i \in [K]$ and $t \in [T]$. Each action $\hat{x}_{t,i}$ is generated as follows,

$$\hat{x}_{t,i} = x_{t,i} + \psi_{t,i}. \tag{1}$$

This model states that each $\hat{x}_{t,i}$ in D_t is a perturbed version of the true underlying $x_{t,i}$. Denote the covariance matrix of $x_{t,i}$ by Σ_x. Notice that Σ_x is rank-m. Perturbation vectors, $\psi_{t,i}$, are assumed to be isotropic, thus covariance matrix $\Sigma_\psi = \sigma^2 I_d$. Let $\lambda_+ := \lambda_1(\Sigma_x)$ and $\lambda_- := \lambda_m(\Sigma_x)$. We will make a boundedness assumption on $x_{t,i}$ and $\psi_{t,i}$.

Assumption 1 (Bounded Action and Perturbation Vectors)**.** *There exists finite constants, d_x and d_ψ, such that for all $t \in [T]$ and $i \in [K]$,*

$$\|x_{t,i}\|_2^2 \le d_x\lambda_+, \quad \|\psi_{t,i}\|_2^2 \le d_\psi\sigma^2.$$

Both d_x and d_ψ can be dependent on m or d and they can be interpreted as the effective dimensions of the corresponding vectors.

At each round t, the agent chooses an action, $\hat{X}_t \in D_t$ and observes a reward r_t such that

$$r_t = (P\hat{X}_t)^T\theta_* + \eta_t \qquad \forall t \in [T] \tag{2}$$

where $P = VV^T$ is the projection matrix for the m-dimensional subspace $\mathrm{span}(V)$, $\theta_* \in \mathrm{span}(V)$ is the unknown parameter vector and η_t is the random noise at round t. Notice that since $\theta_* \in \mathrm{span}(V)$, $(P\hat{X}_t)^T\theta_* = \hat{X}_t^TP\theta_* = \hat{X}_t^T\theta_*$ therefore, $r_t = \hat{X}_t^T\theta_* + \eta_t$.[1] Consider $\{F_t\}_{t=0}^{\infty}$ as any filtration of σ-algebras such that for any $t \ge 1$, $\hat{X}_t$ is F_{t-1} measurable and η_t is F_t measurable.

Assumption 2 (Subgaussian Noise)**.** *For all $t \in [T]$, η_t is conditionally R-sub-Gaussian where $R \ge 0$ is a fixed constant, ie. $\forall\lambda \in \mathbb{R},\ \mathbb{E}[e^{\lambda\eta_t}|F_{t-1}] \le e^{\frac{\lambda^2R^2}{2}}$.*

This implies that $\mathbb{E}\big[\hat{X}_t^T\theta_* + \eta_t|\hat{\mathbf{X}}_t, \eta_{t-1}\big] = \hat{X}_t^T\theta_*$ or equivalently $\mathbb{E}[\eta_t|F_{t-1}] = 0$. The goal of the agent is to maximize the total expected reward accumulated in T rounds, $\sum_{t=1}^{T}\hat{X}_t^T\theta_*$. The oracle's strategy with the knowledge of θ_* at each round t is $\hat{X}_t^* = \arg\max_{x\in D_t} x^T\theta_*$. We evaluate the agent's performance against the oracle performance. Define *regret* as the difference between expected reward of the oracle and the agent,

$$R_T \coloneqq \sum_{t=1}^{T}\hat{X}_t^{*T}\theta_* - \sum_{t=1}^{T}\hat{X}_t^T\theta_* = \sum_{t=1}^{T}(X_t^* - \hat{X}_t)^T\theta_*. \tag{3}$$

The agent aims to minimize this quantity over time. In the setting described above, the agent is assumed to know that there exists a m-dimensional subspace of $\mathbb{R}^d$ in which true action vectors and the unknown parameter vector lie. Finally, we define some quantities about the structure of the problem for all $\delta \in (0, 1)$:

$$g_x = \frac{\lambda_+}{\lambda_-}, \quad g_\psi = \frac{\sigma^2}{\lambda_-}, \quad \Gamma = 2g_\psi + 4\sqrt{g_xg_\psi} \tag{4}$$

$$\alpha = \max(d_x, d_\psi), \ \ n_\delta = 4\alpha\left(\Gamma\sqrt{\log\frac{2d}{\delta}} + \sqrt{2g_x\log\frac{m}{\delta}}\right)^2$$

3. Overview of PSLB

We propose PSLB, a SLB algorithm which employs subspace recovery to extract information from the unsupervised data accumulated in the SLB. During the course of interaction, the agent constructs the confidence set of the underlying model with and without subspace recovery, then takes the intersection of these two sets. Among the plausible models in this set, the agent deploys OFU principle and follows the optimal action of the most optimistic model. The pseudocode of PSLB is given in Algorithm 1. PSLB consists of 4 key elements: warm-up, subspace estimation, creating confidence sets and acting optimistically. In the following, we will discuss each of them briefly.

Algorithm 1 PSLB

```
Input: m, $\lambda_+$, $\lambda_-$, $\sigma^2$, $\alpha$, $\delta$
for t = 1 to $T$ do
    Compute PCA over $\uplus_{i=1}^{t} D_i$
    Create $\hat{P}_t$ with first m eigenvectors
    Construct $\mathcal{C}_{p,t}$, high probability confidence set on $\hat{P}_t$
    Construct $\mathcal{C}_{m,t}$, high probability confidence set for $\theta_*$ using subspace recovery
    Construct $\mathcal{C}_{d,t}$, high probability confidence set for $\theta_*$ without using subspace recovery
    Construct $\mathcal{C}_t = \mathcal{C}_{m,t} \cap \mathcal{C}_{d,t}$
    $(\tilde{P}_t, \hat{X}_t, \tilde{\theta}_t) = \arg\max_{(P',x,\theta)\in\mathcal{C}_{p,t}\times D_t\times\mathcal{C}_t}(P'x)^T\theta$
    Play $\hat{X}_t$ and observe $r_t$
end for
```

3.1. Warm-Up

The decision set at each round i, D_i, has a finite number of actions. The algorithm needs to acquire enough samples of action vectors to reliably estimate the hidden subspace. The process of acquiring sufficient samples is considered as the warm-up period. The duration of the warm-up period, $t_{w,\delta}$, can be chosen in many ways. We set $t_{w,\delta} = \frac{n_\delta}{K}$ based on the theoretical analysis outlined in Section 4.1. The crux of this choice is to provide a theoretical guarantee of convergence to the underlying subspace. In other words, PSLB collects samples until it has some confidence on the recovered subspace. This idea is considered in more detail in Section 3.2. Note that warm-up periods are implicitly assumed in most SLB algorithms since the given bounds are not meaningful for short periods of time.

3.2. Subspace Estimation

At each round, the algorithm predicts the m-dimensional subspace that the true action vectors belong, using the action vectors collected up to that round. In particular, at round t, the algorithm uses PCA over tK action vectors observed so far, $\uplus_{i=1}^{t} D_i$. It calculates $\hat{V}_t$ which is the matrix of top m eigenvectors of $\frac{1}{tK}\sum_{\hat{x}\in\uplus_{i=1}^{t} D_i}\hat{x}\hat{x}^T$, thus $\mathrm{span}(\hat{V}_t)$ is the predicted m-dimensional subspace. Then, $\hat{V}_t$ is used to create the estimated projection matrix associ-

[1] The reward generative model of $r_t = \hat{X}_t^T\theta_* + \eta_t$ is equivalent to $r_t = X_t^T\theta_* + \tilde{\eta}_t$ where $\tilde{\eta}_t$ contains the randomness in η_t as well as the perturbations due to $\psi_{t,i}$.

ated to this subspace, $\hat{P}_t := \hat{V}_t \hat{V}_t^T$.

As the agent observes more action vectors, the estimated projection matrix becomes more accurate. The accuracy of $\hat{P}_t$ is measured by the projection error $\|\hat{P}_t - P\|_2$. As more action vectors are collected, $\|\hat{P}_t - P\|_2$ shrinks. Since P is not known, $\|\hat{P}_t - P\|_2$ cannot be calculated directly. Thus, PSLB calculates a high-probability upper bound on the projection error. Using the derived bound, PSLB deploys confidence in the subspace estimation and construct the set of plausible projection matrices $\mathcal{C}_{p,t}$ where $\hat{P}_t$ and P both lie in with high probability. The construction of $\mathcal{C}_{p,t}$ is reliant on the structural properties of the problem and the number of samples tK. We analyze these properties in Section 4.1.

3.3. Confidence Set Construction

At each round, PSLB creates two confidence sets for the model parameter θ_*. First, it tries to exploit a possible m-dimensional hidden subspace structure. Thus, it searches for a high probability confidence set, $\mathcal{C}_{m,t}$, that lies around the estimated subspace at round t. Using the history of action-reward pairs, the algorithm solves a regularized least squares problem in the estimated subspace and obtains θ_t, the estimated parameter vector in $\operatorname{span}(\hat{V}_t)$. Then it creates the confidence set $\mathcal{C}_{m,t}$ around θ_t, such that $\theta_* \in \mathcal{C}_{m,t}$ with high probability.

Second, PSLB searches for a high probability confidence set in the ambient space without having subspace recovery. It deploys the confidence set generation subroutine of OFUL by Abbasi-Yadkori et al. (2011). Using the history of action-reward pairs, the algorithm solves another regularized least squares problem but this time in the ambient space and obtains $\hat{\theta}_t$. PSLB then creates the confidence set $\mathcal{C}_{d,t}$ centered around $\hat{\theta}_t$ such that $\theta_* \in \mathcal{C}_{d,t}$ with high probability. Finally, PSLB takes the intersection of constructed confidence sets to create the main confidence set, $\mathcal{C}_t = \mathcal{C}_{m,t} \cap \mathcal{C}_{d,t}$. $\mathcal{C}_t$ still contains θ_* with high probability. With this operation, PSLB provides a new perspective that if there exists an easily recoverable m-dimensional subspace, it exploits that structure to get lower regret than OFUL can solely achieve. If it fails to detect such structure or the confidence set is looser than what OFUL provides, then it still provides the same regret as OFUL.

3.4. Optimistic Action

For the final step in round t, the algorithm chooses an optimistic triplet $(\tilde{P}_t, \hat{X}_t, \tilde{\theta}_t)$ from the confidence sets created and the current decision set which jointly maximizes the reward:

$$(\tilde{P}_t, \hat{X}_t, \tilde{\theta}_t) = \underset{(P',x,\theta)\in\mathcal{C}_{p,t}\times D_t\times\mathcal{C}_t}{\arg\max} (P'x)^T\theta \qquad (5)$$

4. Theoretical Analysis of PSLB

In this section we first state the upper regret bound of PSLB which is the main result of the paper. Then we analyze the components that build up to the result. In order to get a meaningful bound, we assume that the expected rewards are bounded. Recalling the quantities defined in (4), define Υ such that

$$\Upsilon = \mathcal{O}\left(\frac{\alpha\Gamma^2\sqrt{m}}{K(\lambda_- + \sigma^2)}\right). \qquad (6)$$

It represents the difficulty of subspace recovery in terms structural properties of SLB setting, and it is analyzed in Section 4.3. Using Υ, the theorem below states the regret upper bound of PSLB.

Theorem 1 (Regret Upper Bound of PSLB)**.** *Fix any* $\delta \in (0, 1/6)$. *Assume that Assumptions 1 and 2 hold. Also assume that for all* $\hat{x}_{t,i} \in D_t$, $\hat{x}_{t,i}^T\theta_* \in [-1, 1]$. *Then,* $\forall t \geq 1$ *with probability at least* $1 - 6\delta$*, the regret of* PSLB *satisfies*

$$R_t \leq \min\left\{\widetilde{\mathcal{O}}\left(\Upsilon\sqrt{t}\right), \widetilde{\mathcal{O}}\left(d\sqrt{t}\right)\right\}. \qquad (7)$$

The proof of the theorem involves two main pieces: the projection error analysis and the construction of projected confidence sets. They are analyzed in Sections 4.1 and 4.2 respectively. Finally, in Section 4.3 their role in the proof of Theorem 1 is explained and the meaning of the result is discussed.

4.1. Projection Error Analysis

Consider the matrix $\hat{V}_t^T V$ where ith singular value is denoted by $\sigma_i(\hat{V}_t^T V)$, such that $\sigma_1(\hat{V}_t^T V) \geq \ldots \geq \sigma_m(\hat{V}_t^T V)$. Extending the definition of inner products of two vectors to subspaces and using Courant-Fischer-Weyl minimax principle, one can define *ith principal angle* Θ_i between $\operatorname{span}(V)$ and $\operatorname{span}(\hat{V}_t)$ via

$$\cos\Theta_i(\operatorname{span}(V), \operatorname{span}(\hat{V}_t)) = \sigma_i(\hat{V}_t^T V).$$

Using the analysis in Akhiezer & Glazman (2013) it can be seen that:

$$\begin{aligned}\|\hat{P}_t - P\|_2 &= \sqrt{\lambda_{max}\left(I_m - (\hat{V}_t^T V)^T(\hat{V}_t^T V)\right)} \\ &= \sqrt{1 - \sigma_m^2(\hat{V}_t^T V)} = \sin\Theta_m \qquad (8)\end{aligned}$$

where Θ_m is the largest principal angle between the column spans of V and $\hat{V}_t$. Thus, bounding the projection error between two projection matrices is equivalent to bounding the sine of the largest principal angle between the subspaces that they project. In light of this relation, one can use the Davis-Kahan $\sin\Theta$ theorem (Davis & Kahan, 1970) to bound the projection error. The exact theorem statement can be found in Section A in the Supplementary Material. Informally, the theorem considers a symmetric matrix and its' perturbed version and bounds the sine of the largest principal angle caused by this perturbation. Using Davis-Kahan $\sin\Theta$ theorem, following lemma bounds the finite sample projection error.

Lemma 2 (Finite Sample Projection Error)**.** *Fix any $\delta \in (0, 1/3)$. Let $t_{w,\delta} = \frac{n_\delta}{K}$. Suppose Assumption 1 holds. Then with probability at least $1 - 3\delta$, $\forall t \geq t_{w,\delta}$,*

$$\|\hat{P}_t - P\|_2 \leq \frac{\phi_\delta}{\sqrt{t}} \quad , \textit{ where } \phi_\delta = 2\Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}}. \quad (9)$$

The lemma and it's proof are along the same lines of Corollary 2.9 of Vaswani & Narayanamurthy (2017). However, we improve the bound on the projection error by using the Matrix Chernoff Inequality (Tropp, 2015) and provide the precise problem dependent quantities in the bound which are required for defining the minimum number of samples for the warm-up period and the construction of confidence sets for θ_*. Note that as discussed in Section 3.2, (9) defines the confidence set $\mathcal{C}_{p,t}$ for all $t \geq t_{w,\delta}$. The general version of the lemma and the details of the proof are given in Section A of the Supplementary Material, but here we provide a proof sketch.

Up to round t, the agent observes tK action vectors in total within the decision sets. Using PCA, PSLB estimates an m-dimensional subspace spanned by top m eigenvectors of the sample covariance matrix of tK action vectors and obtain the projection matrix $\hat{P}_t$ for that subspace. In order to derive Lemma 2, we first carefully pick two symmetric matrices such that the span of their first m eigenvectors are equivalent to subspaces that P and $\hat{P}_t$ project to. Using Davis-Kahan $\sin\Theta$ theorem with matrix concentration inequalities provided by Tropp (2015), we derive the finite sample projection error bound.

Lemma 2 is key to defining the warm-up period duration. Due to equivalence in (8), $\|\hat{P}_t - P\|_2 \leq 1, \forall t \geq 1$. Therefore, any projection error bound greater than 1 is vacuous. We pick $t_{w,\delta}$ such that with high probability, we obtain theoretically non-trivial bound on projection error. With the given choice of $t_{w,\delta}$, the bound on the projection error in (9) becomes less than 1 when $t \geq t_{w,\delta}$. After $t_{w,\delta}$, PSLB starts to produce non-trivial confidence sets $\mathcal{C}_{p,t}$ around $\hat{P}_t$. However, note that $t_{w,\delta}$ can be significantly big for problems that have structure that is hard to recover, e.g. having α linear in d.

Lemma 2 also brings several important intuitions about the subspace estimation problem in terms of the problem structure. Recalling the definition of Γ in (4), as g_ψ decreases, the projection error shrinks since the underlying subspace becomes more distinguishable. Conversely, as g_x diverges from 1, it becomes harder to recover the underlying m-dimensional subspace. Additionally, since α is the maximum of the effective dimensions of the true action vector and the perturbation vector, having large α makes the subspace recovery harder and the projection error bound looser, whereas observing more action vectors, K in each round produces tighter bound on $\|\hat{P}_t - P\|_2$. The effects of these structural properties on the subspace estimation translate to confidence set construction and ultimately to regret upper bound.

4.2. Projected Confidence Sets

In this section, we analyze the construction of $\mathcal{C}_{m,t}$ and $\mathcal{C}_{d,t}$. For any round $t \geq 1$, define $\hat{\Sigma}_t := \sum_{i=1}^{t} \hat{X}_i \hat{X}_i^T = \hat{\mathbf{X}}_t \hat{\mathbf{X}}_t^T$. At round t, let $A_t := \hat{P}_t(\hat{\Sigma}_{t-1} + \lambda I_d)\hat{P}_t$ for $\lambda > 0$. The rewards obtained up to round t is denoted as $\mathbf{r}_{t-1}$. At round t, after estimating the projection matrix $\hat{P}_t$ associated with the underlying subspace, PSLB tries to find θ_t, an estimate of θ_*, while believing that θ_* lives within the estimated subspace. Therefore, θ_t is the solution to the following Tikhonov-regularized least squares problem with regularization parameters $\lambda > 0$ and $\hat{P}_t$,

$$\theta_t = \arg\min_\theta \|(\hat{P}_t \hat{\mathbf{X}}_{t-1})^T \theta - \mathbf{r}_{t-1}\|_2^2 + \lambda \|\hat{P}_t \theta\|_2^2.$$

Notice that regularization is applied along the estimated subspace. Solving for θ gives $\theta_t = A_t^\dagger\left(\hat{P}_t \hat{\mathbf{X}}_{t-1} \mathbf{r}_{t-1}\right)$. Define L such that for all $t \geq 1$ and $i \in [K]$, $\|\hat{x}_{t,i}\|_2 \leq L$ and let $\gamma = \frac{L^2}{\lambda \log\left(1 + \frac{L^2}{\lambda}\right)}$. The following theorem gives the construction of projected confidence set, $\mathcal{C}_{m,t}$, which is an ellipsoid centered around θ_t which contains θ_* with high probability.

Theorem 3 (Projected Confidence Set Construction)**.** *Fix any $\delta \in (0, 1/4)$. Suppose Assumptions 1 & 2 hold, and $\forall t \geq 1$ and $i \in [K]$, $\|\hat{x}_{t,i}\|_2 \leq L$. If $\|\theta_*\|_2 \leq S$ then, with probability at least $1 - 4\delta$, $\forall t \geq t_{w,\delta}$, θ_* lies in the set*

$$\mathcal{C}_{m,t} = \left\{\theta \in \mathbb{R}^d : \|\theta_t - \theta\|_{A_t} \leq \beta_{t,\delta}\right\}, \textit{ where}$$

$$\beta_{t,\delta} = R\sqrt{2\log\left(\frac{1}{\delta}\right) + m\log\left(1 + \frac{tL^2}{m\lambda}\right)} + LS\phi_\delta\sqrt{\gamma m \log\left(1 + \frac{tL^2}{m\lambda}\right)} + S\sqrt{\lambda}. \quad (10)$$

The detailed proof and a general version of the theorem are given in Section B of the Supplementary Material. We will highlight the key aspects in here. The overall proof follows a similar machinery used by Abbasi-Yadkori et al. (2011). Specifically, the first term of $\beta_{t,\delta}$ in (10) is derived similarly by using the self-normalized tail inequality. However, since at each round PSLB projects the past actions to an estimated m-dimensional subspace to estimate θ_*, d is replaced by m in the bound. While enjoying the benefit of projection, this construction of the confidence set suffers from the finite sample projection error, *i.e.*, uncertainty in the subspace estimation. This effect is observed via second term in (10). The second term involves the confidence bound for the estimated projection matrix, ϕ_δ. This is critical in determining the tightness of the confidence set on θ_*. As discussed in Section 4.1, ϕ_δ reflects the difficulty of subspace recovery of the given problem and it depends on the underlying structure of the problem and SLB. This shows that as estimating the underlying subspace gets more difficult, having a projection based approach in the construction of the confidence sets on θ_* provides looser bounds.

In order to tolerate the possible difficulty of subspace recovery, PSLB also constructs $\mathcal{C}_{d,t}$, which is the confidence set for θ_* without having subspace recovery. The construction of $\mathcal{C}_{d,t}$ follows OFUL by Abbasi-Yadkori et al. (2011). Let $Z_t = \hat{\Sigma}_{t-1} + \lambda I_d$. The algorithm tries to find $\hat{\theta}_t$ which is the ℓ^2-regularized least squares estimate of θ_* in the ambient space. Thus, $\hat{\theta}_t = Z_t^{-1}\hat{\mathbf{X}}_{t-1}\mathbf{r}_{t-1}$. Construction of $\mathcal{C}_{d,t}$ is done under the same assumptions of Theorem 3, such that with probability at least $1-\delta$, θ_* lies in the set

$$\mathcal{C}_{d,t} = \left\{\theta \in \mathbb{R}^d : \|\hat{\theta}_t - \theta\|_{Z_t} \leq \Omega_{t,\delta}\right\} \quad \text{where}$$

$$\Omega_{t,\delta} = R\sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{tL^2}{m\lambda}\right)} + S\sqrt{\lambda}.$$

The search for an optimistic parameter vector happens in the intersection of $\mathcal{C}_{m,t}$ and $\mathcal{C}_{d,t}$. Notice that $\theta_* \in \mathcal{C}_{m,t} \cap \mathcal{C}_{d,t}$ with probability at least $1-5\delta$. Optimistically choosing the triplet, $(\tilde{P}_t, \hat{X}_t, \tilde{\theta}_t)$, within the described confidence sets gives PSLB a way to tolerate the possibility of failure in recovering an underlying structure. If confidence set $\mathcal{C}_{m,t}$ is loose or PSLB is not able to recover an underlying structure, then $\mathcal{C}_{d,t}$ provides the useful confidence set to obtain desirable learning behavior.

4.3. Regret Analysis

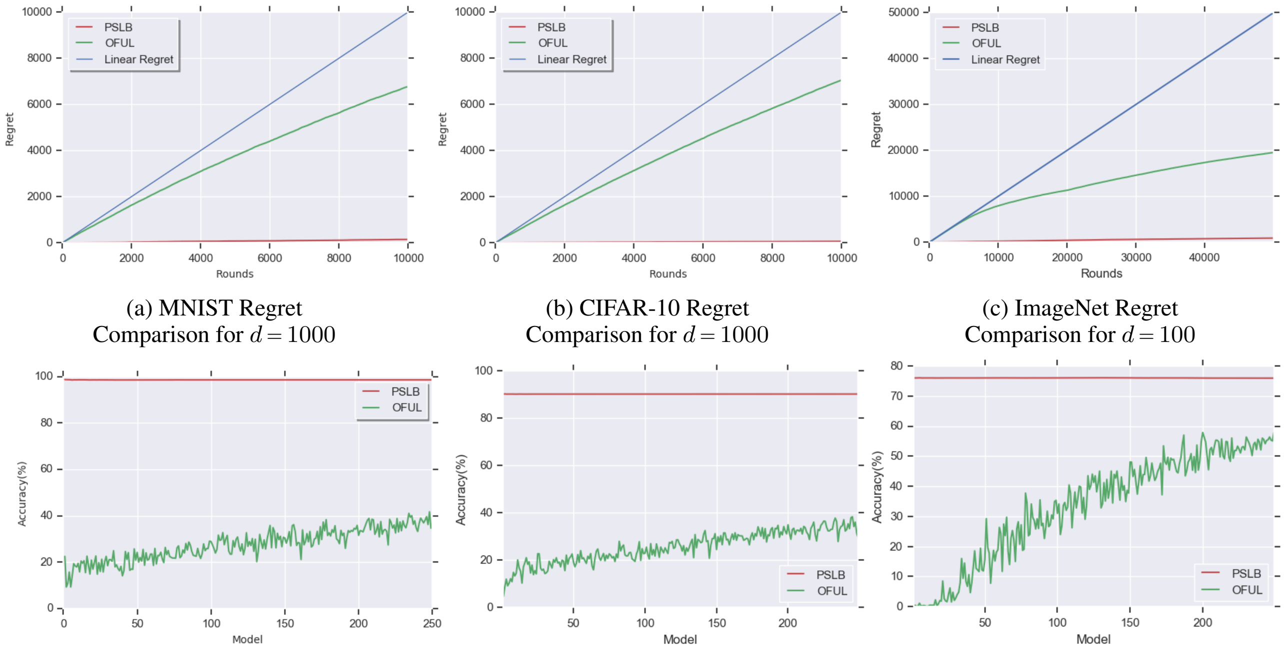
Now that the confidence set constructions and the decision making procedures of PSLB are explained, it only remains to analyze the regret of PSLB. Using the intersection of $\mathcal{C}_{m,t}$ and $\mathcal{C}_{d,t}$ as the confidence set at round t, gives PSLB the ability to obtain the lowest possible instantaneous regret among both confidence sets. Therefore, the regret of PSLB is upper bounded by the minimum of the regret upper bounds on the individual strategies. Using only $\mathcal{C}_{d,t}$ is equivalent to following OFUL and the regret analysis can be found in Abbasi-Yadkori et al. (2011). The regret analysis of using only the projected confidence set $\mathcal{C}_{m,t}$ is the main contribution of this work. It follows the standard regret decomposition into instantaneous regret components. However, due to having different estimated projection matrices in each round, the derivation of the bound uses a different strategy involving the Matrix Chernoff Inequality (Tropp, 2015). The detailed analysis of the regret upper bound and the proof can be found in Section C of the Supplementary Material. Here we elaborate more on the nature of the regret obtained by using projected confidence sets only, *i.e.* first term in Theorem 1, and discuss the effect of Υ in particular.

Υ is the reflection of the finite sample projection error at the beginning of the algorithm. It captures the difficulty of subspace recovery based on the structural properties of the problem and determines the regret of deploying projection based methods in SLBs. Recall that α is the maximum of the effective dimensions of the true action vector and the perturbation vector. Depending on the structure of the problem, α can be $\mathcal{O}(d)$, e.g., the perturbation can be effective in many dimensions, which prevents the projection error from shrinking; thus, causes $\Upsilon = \mathcal{O}(d\sqrt{m})$ resulting in $\widetilde{\mathcal{O}}(d\sqrt{mt})$ regret. The eigengap within the true action vectors g_x and the eigengap between the true action vectors and the perturbation vectors g_ψ are critical factors that determine the identifiability of the hidden subspace. As σ^2 increases, the subspace recovery becomes harder since the effect of perturbation increases. Conversely, as λ_- increases, the underlying subspace becomes easier to identify. These effects are significant on the regret of PSLB and they are captured by Γ^2 in Υ. Moreover, having finite samples to estimate the subspace affects the regret bound through Υ. Due to the nature of SLB, this is unavoidable and it scales the final regret by $1/K$. Overall, with all these elements, Υ represents the hardness of using PCA based methods in dimensionality reduction in SLBs.

Theorem 1 states that if the underlying structure is easily recoverable, e.g. $\Upsilon = \mathcal{O}(m)$, then using PCA based dimension reduction and construction of confidence sets provide substantially better regret upper bound for large d. If that is not the case, then due to the best of the both worlds approach provided by PSLB, the agent obtains the best possible regret upper bound. Note that the bound for using only $\mathcal{C}_{m,t}$ is a worst case bound and as we present in Section 5, in practice PSLB can give significantly better results.

5. Experiments

In the experiments, we study MNIST, CIFAR-10 and ImageNet datasets and use them to create the decision sets for the SLB setting. A simple 5-layer CNN, a pre-trained ResNet-18 and a pre-trained ResNet-50 are deployed respectively for MNIST, CIFAR-10 and ImageNet. Before training, we modify the architecture of the representation layer (the layer before the final layer) to make it suitable



(a) MNIST Regret Comparison for $d = 1000$

(b) CIFAR-10 Regret Comparison for $d = 1000$

(c) ImageNet Regret Comparison for $d = 100$

(d) MNIST Model Accuracy Comparison for $d = 1000$

(e) CIFAR-10 Model Accuracy Comparison for $d = 1000$

(f) ImageNet Model Accuracy Comparison for $d = 100$

Figure 1: Regret and Optimistic Model Accuracy Comparisons of PSLB and OFUL on MNIST, CIFAR-10 and ImageNet

Top row: Regret of PSLB vs. Regret of OFUL in SLB setting constructed from image classification tasks. PSLB tries to recover $m = 1$ dimensional subspace which reduces the complexity of SLB and results in very few committed mistakes. Due to lack of additional knowledge besides rewards obtained from chosen actions, OFUL starts with linear regret and commits significant amount of mistakes. Bottom row: Image classification accuracy of periodically sampled optimistic models of PSLB and OFUL over all images in datasets. The ability to reduce the complexity of learning task helps PSLB to learn the best possible underlying linear model just in few rounds whereas OFUL requires more action-reward pairs to get an accurate estimate.

for the SLB study and obtain decision sets for each image.

Consider a standard network whose dimension of the representation layer is d. Therefore, the final layer for K class classification is fully connected and it is a $d \times K$ matrix that outputs K numbers to be used for classification. In this study, instead of having a final layer of $d \times K$ matrix, we construct the final layer as a d-dimensional vector and make the feature representation layer a Kd dimensional vector. We treat this vector as the concatenation of K d-dimensional contexts *i.e.*, $[\hat{x}_1, \ldots, \hat{x}_K]$. The final d-dimensional layer is θ_* of the SLB, where the logit for each class is computed as an inner product of the class context $\hat{x}_i$ and θ_*. We train these architectures for different d values using cross entropy loss. Here we provide results for MNIST and CIFAR-10 with $d = 1000$ and ImageNet with $d = 100$.

Removing the final layer, the resulting trained networks are used to generate the feature representations of each image for each class which produces the decision sets at each time step of SLB. Since MNIST and CIFAR-10 have 10 classes, in each decision set we obtain 10 action vectors where each of them are segments in the representation layer. On the other hand, from the ImageNet dataset we get 1000 actions per decision set due to 1000 classes in the datasets. In the SLB setting, the agent receives a reward of 1 if it chooses the right action, which is the segment in the representation layer corresponding to correct label according to trained network, and 0 otherwise. We apply both PSLB and OFUL on these SLBs. We measure the regret by counting the number of mistakes each algorithm makes. To come up with the optimistic choice of action at each time step, both of these algorithms requires solving an inner optimization problem. To mitigate the burden(CITE Stochastic Linear Optimization under Bandit Feedback

Linear Thompson Sampling Revisited) of these computation costs, we sample many models from the confidence sets and choose the most optimistic model among the sampled ones.

Through computing PCA of the empirical covariance matrix of the action vectors, surprisingly we found that projecting action vectors onto the 1-dimensional subspace defined by the dominant eigenvector is sufficient for these datasets in the SLB setting; thus, $m = 1$. During the experiments PSLB tried to recover a 1-dimensional subspace using the action vectors collected. We present the regrets obtained by PSLB and OFUL for MNIST, CIFAR-10 and ImageNet in Figure 1a, 1b, 1c respectively. With the help of subspace recovery and projection, PSLB provides a massive reduction in the dimensionality of the SLB problem and immediately estimates a fairly accurate model for θ_*. On the other hand, OFUL naively tries to sample from all dimensions in order to learn θ_*. This difference yields orders of magnitude improvement in regret. During the SLB experiment, we also sample the optimistic models that are chosen by PSLB and OFUL. We use these models to test the model accuracy of the algorithms, *i.e.* perform classification over all images in dataset. The optimistic model accuracy comparisons are depicted in Figure 1d, 1e, 1f. These portray the learning behavior of PSLB and OFUL. Using projection, PSLB learns the underlying linear model in the first few rounds, whereas OFUL suffers from high-dimension of SLB framework and lack of knowledge besides chosen action-reward pairs. We extend these experiments for settings with $d = 100, 500, 1000$ and $m = 1, 2, 4, 8, 16$ which can be found in Section D.

6. Related Work

The primary class of partial information problems is the multi-arm bandit (MAB). Robbins (1985) introduces the standard stochastic MAB and Lai & Robbins (1985) studies the asymptotic property of learning algorithms on this class. Stochastic MABs are a special case of SLB when the arms representations are orthogonal to each other. For finite sample regime, Auer et al. (2002) deploys the principle of OFU and provide finite sample guarantee for MABs. Auer (2002) deploys the same principle to provide regret guarantee for MABs with the linear pay-off. This principle is realized as the primary approach even for more general problems such as Linear Quadratic systems (Abbasi-Yadkori & Szepesvári, 2011) and Markov Decision Processes (Jaksch et al., 2010).

The study of linear bandit problems extends to various algorithms and environment settings (Dani et al., 2008; Rusmevichientong & Tsitsiklis, 2010; Li et al., 2010). Kleinberg et al. (2010) studies the class of problems when the decision set changes time to time, while Dani et al. (2008) studies this problem when the decision set provides a set of fixed actions. Further analysis in the area extend these approaches to classes where there are more structures in the problem setup. In traditional decision-making problems, where hand engineered feature representations are provided, sparsity in the linear function is a valid structure. Sparsity, as the key in high-dimensional conventional structured linear bandits, conveys series of successes in classical settings (Abbasi-Yadkori et al., 2012; Carpentier & Munos, 2012). In recommendations systems, where a set of users and items are given, Gopalan et al. (2016) consider the low-rank structure of the user-item preference matrix and provide an algorithm which exploits this further structure.

To the best of our knowledge, there are no hidden low-dimensional subspace assumptions on actions and/or unknown weight vector in literature for SLB. On the other hand, subspace recovery and dimension reduction problems are well studied in the literature. Several linear and nonlinear dimension reduction methods have been proposed such as PCA (Pearson, 1901), independent component analysis (Hyvärinen & Oja, 2000), random projections (Candes & Tao, 2006) and non-convex robust PCA (Netrapalli et al., 2014). Among the linear dimension reduction techniques, PCA is the simplest, yet most widely used method. Analysis of PCA based methods mostly focus on the asymptotic results (Anderson et al., 1963; Jain et al., 2016). However, in the settings like SLB with finite number of arms, it is necessary to have finite sample guarantees for the application of PCA. In the literature, among few finite sample PCA works, Nadler (2008) provides finite sample guarantees for one-dimensional PCA, whereas Vaswani & Narayanamurthy (2017) extends it to larger dimensions with various noise models.

7. Conclusion

In this paper, we study a linear subspace structure in the action set of an SLB problem. We deploy PCA based projection to exploit the immense number of unsupervised actions in the decision sets and learn the underlying subspace. We proposed PSLB, a SLB algorithm which utilizes the subspace estimated through PCA to improve the regret upper bound of SLB problems. If such structure does not exist or is hard to recover, then the PSLB reduces to the standard SLB algorithm, OFUL. We empirically study MNIST, CIFAR-10 and ImageNet datasets to create SLB framework from image classification tasks. We test the performance of PSLB versus OFUL in the SLB setting created. We show that when DNNs produce features of the actions, a significantly low dimensional structure is observed. Due to this structure, we showed that PSLB substantially outperforms OFUL and converges to an accurate model while OFUL still struggles to sample in high dimensions to learn the underlying parameter vector.

In this work, we studied the class of linear subspace structures. In the future work, we plan to extend this line of study to the general class of low dimensional manifold structured problems. Bora et al. (2017) peruse a similar approach for compression problems. While optimism is the primary approach in the theoretical analyses of SLBs, it mainly poses a computationally intractable internal optimization problem. An alternative method is Thompson sampling, a practical algorithm for SLBs. In future work, we plan to deploy Thompson sampling and mitigate the computational complexity of PSLB.

Acknowledgement

K. Azizzadenesheli is supported in part by NSF Career Award CCF-1254106 and Air Force FA9550-15-1-0221. This research has been conducted when the second author was a visiting researcher at Stanford University and Caltech. A. Anandkumar is supported in part by Microsoft Faculty Fellowship, Google faculty award, Adobe grant, NSF Career Award CCF-1254106, and AFOSR YIP FA9550-15-1-0221. All the experimental study have been done using Caltech AWS credits grant.

References

Abbasi-Yadkori, Y. and Szepesvári, C. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 1–26, 2011.

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pp. 1–9, 2012.

Akhiezer, N. I. and Glazman, I. M. *Theory of linear operators in Hilbert space*. Courier Corporation, 2013.

Anderson, T. W. et al. Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, 34 (1):122–148, 1963.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Bora, A., Jalal, A., Price, E., and Dimakis, A. G. Compressed sensing using generative models. *arXiv preprint arXiv:1703.03208*, 2017.

Candes, E. J. and Tao, T. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006.

Carpentier, A. and Munos, R. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *Artificial Intelligence and Statistics*, pp. 190–198, 2012.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. 2008.

Davis, C. and Kahan, W. M. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Freedman, D. A. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Gopalan, A., Maillard, O.-A., and Zaki, M. Low-rank bandits with latent mixtures. *arXiv preprint arXiv:1609.01508*, 2016.

Hyvärinen, A. and Oja, E. Independent component analysis: algorithms and applications. *Neural networks*, 13 (4-5):411–430, 2000.

Jain, P., Jin, C., Kakade, S. M., Netrapalli, P., and Sidford, A. Streaming pca: Matching matrix bernstein and near-optimal finite sample guarantees for ojas algorithm. In *Conference on Learning Theory*, pp. 1147–1164, 2016.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Kleinberg, R., Niculescu-Mizil, A., and Sharma, Y. Regret bounds for sleeping experts and bandits. *Machine learning*, 80(2-3):245–272, 2010.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670. ACM, 2010.

Nadler, B. Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6):2791–2817, 2008.

Netrapalli, P., Niranjan, U., Sanghavi, S., Anandkumar, A., and Jain, P. Non-convex robust pca. In *Advances in Neural Information Processing Systems*, pp. 1107–1115, 2014.

Pearson, K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11): 559–572, 1901.

Robbins, H. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, pp. 169–177. Springer, 1985.

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Tropp, J. A. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Vaswani, N. and Narayanamurthy, P. Finite sample guarantees for pca in non-isotropic and data-dependent noise. In *Communication, Control, and Computing (Allerton), 2017 55th Annual Allerton Conference on*, pp. 783–789. IEEE, 2017.

A. Projection Error Analysis, Proof of Lemma 2

In this section, we provide the general version of Lemma 2 with the proof details. As stated in the main text, in order to bound the projection error, we will use Davis-Kahan $\sin\Theta$ theorem which states the following:

Theorem 4 ((Davis & Kahan, 1970)). *Let $S, H \in \mathbb{R}^{d\times d}$ be symmetric matrices, such that $\hat{S} = S + H$. The eigenvalues of S and $\hat{S}$ are $\lambda_1 \geq \ldots \geq \lambda_m \geq \ldots \geq \lambda_d$ and $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_m \geq \ldots \geq \hat{\lambda}_d$ respectively. Define the eigendecompositions of S and $\hat{S}$:*

$$S = [U \quad U_o]\begin{bmatrix} \Lambda & 0 \\ 0 & \Lambda_o \end{bmatrix}[U \quad U_o]^T$$

$$\hat{S} = [\hat{U} \quad \hat{U}_o]\begin{bmatrix} \hat{\Lambda} & 0 \\ 0 & \hat{\Lambda}_o \end{bmatrix}[\hat{U} \quad \hat{U}_o]^T$$

where Λ and $\hat{\Lambda}$ are diagonal matrices with first m eigenvalues of S and $\hat{S}$ respectively. $U = (u_1, \ldots, u_m) \in \mathbb{R}^{d\times m}$ and $\hat{U} = (\hat{u}_1, \ldots, \hat{u}_m) \in \mathbb{R}^{d\times m}$ denote the corresponding eigenvectors. Define

$$\delta \coloneqq \inf\{|\hat{\lambda} - \lambda| : \lambda \in [\lambda_m, \lambda_1], \hat{\lambda} \in (-\infty, \hat{\lambda}_{m+1}]\}.$$

If $\delta > 0$, then $\sin\Theta_m$, sine of the largest principal angle between the column spans of U and $\hat{U}$, can be upper bounded as

$$\sin\Theta_m \leq \frac{\|\hat{S}U - U\Lambda\|_2}{\delta} = \frac{\|\hat{S}U - U\Lambda\|_2}{|\lambda_m - \hat{\lambda}_{m+1}|}. \tag{11}$$

Notice that in order to use Davis-Kahan $\sin\Theta$ theorem in our setting, we need to pick 2 symmetric matrices S and $\hat{S}$ such that their first m eigenvectors has the same span with the subspaces that P and $\hat{P}$ project to. Followed by these choices, in order to get a non-trivial bound we require a significant eigengap between λ_m and $\hat{\lambda}_{m+1}$, due to denominator in (11). We use the following matrix concentration inequalities to maintain an eigengap with high probability.

Theorem 5 (Matrix Chernoff Inequality; (Tropp, 2015)). *Consider a finite sequence $\{X_k\}$ of independent, random, symmetric matrices in $\mathbb{R}^{d\times d}$. Assume that $\lambda_{min}(X_k) \geq 0$ and $\lambda_{max}(X_k) \leq L$ for each index k. Introduce the random matrix $Y = \sum_k X_k$. Let μ_{min} denote the minimum eigenvalue of the expectation $\mathbb{E}[Y]$,*

$$\mu_{min} = \lambda_{min}\big(\mathbb{E}[Y]\big) = \lambda_{min}\bigg(\sum_k \mathbb{E}[X_k]\bigg).$$

Then,

$$\Pr\bigg[\lambda_{min}(Y) \leq \epsilon\mu_{min}\bigg] \leq d\exp\big(-(1-\epsilon)^2\frac{\mu_{min}}{2L}\big) \qquad \textit{for } \epsilon \in [0,1).$$

Theorem 6 (Corollary of Matrix Bernstein; (Tropp, 2015)). *Consider a set of n i.i.d. realization of a $d_1 \times d_2$ random matrix R, as $\{R_1, \ldots, R_n\}$. If $\mathbb{E}[R]$ is bounded, $\|R\|_2 \leq L$ almost surely, with second moment of*

$$m_2(R) = \max\bigg\{\|\mathbb{E}[RR^T]\|_2, \|\mathbb{E}[R^TR]\|_2\bigg\}.$$

Then, for all $t \geq 0$,

$$\Pr\bigg[\|\frac{1}{n}\sum_{i=1}^n R_i - \mathbb{E}[R]\|_2 \geq t\bigg] \leq (d_1 + d_2)\exp\bigg(\frac{-nt^2/2}{m_2(R) + 2Lt/3}\bigg)$$

Define $t_{\min,\delta} = \left(\sqrt{\frac{2d_x g_x}{K}\log\frac{m}{\delta}} + \Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}}\right)^2$. Now that we have the required machinery, we present general version of Lemma 2.

Lemma 7. *Fix any* $\delta \in (0, 1/3)$. *Suppose that Assumption 1 holds. Then with probability at least* $1 - 3\delta$,

$$\|\hat{P}_t - P\|_2 \leq \Phi_{t,\delta}, \qquad \forall t \geq t_{w,\delta},$$

where

$$\Phi_{t,\delta} = \frac{\Gamma\sqrt{\frac{\alpha}{tK}\log\frac{2d}{\delta}}}{1 - \sqrt{\frac{2d_x g_x}{tK}\log\frac{m}{\delta}} - \Gamma\sqrt{\frac{\alpha}{tK}\log\frac{2d}{\delta}}}. \tag{12}$$

Proof. We set $\hat{S} = \frac{1}{n}\sum_{i=1}^n \hat{x}_i\hat{x}_i^T$ and $S = \frac{1}{n}\sum_{i=1}^n x_i x_i^T + VV^T\Sigma_\psi VV^T$ where $n = tK$. Let U be the top m eigenvectors of S. Notice that $\mathrm{span}(U) = \mathrm{span}(V)$ and $\hat{V}_t$ is the matrix of top m eigenvectors of $\hat{S}$. Therefore, one can apply Theorem 4 with given choices of S and $\hat{S}$, to bound $\|\hat{P}_t - P\|_2$. Since $\|\hat{S}U - U\Lambda\|_2 = \|(\hat{S} - S)V\|_2$,

$$\|\hat{P}_t - P\|_2 \leq \frac{\|(\hat{S}-S)V\|_2}{\lambda_m(S) - \lambda_{m+1}(\hat{S})} \overset{(1)}{\leq} \frac{\|(\hat{S}-S)V\|_2}{\lambda_m(S) - \|\hat{S}-S\|_2} \overset{(2)}{\leq} \frac{\|\mathbb{E}[\hat{S}-S]V\|_2 + \|\hat{S}-S-\mathbb{E}[\hat{S}-S]\|_2}{\lambda_m(S) - \|\mathbb{E}[\hat{S}-S]\|_2 - \|\hat{S}-S-\mathbb{E}[\hat{S}-S]\|_2}$$

where (1) follows from Weyl's inequality and the fact that S is rank m, $\lambda_{m+1}(S) = \ldots = \lambda_d = 0$, and (2) is due to triangle inequality. With the given choices of S and $\hat{S}$ and Assumption 1, we have the following:

$$\begin{aligned}
&\lambda_m(S) \geq \lambda_m(\frac{1}{n}\sum_{i=1}^n x_i x_i^T) + \lambda_{\min}(V^T\Sigma_\psi V) = \lambda_m(\frac{1}{n}\sum_{i=1}^n x_i x_i^T) + \lambda_{\min}(\sigma^2 I_m) = \lambda_m(\frac{1}{n}\sum_{i=1}^n x_i x_i^T) + \sigma^2 \\
&\hat{S} - S = \frac{1}{n}\sum_{i=1}^n \psi_i\psi_i^T + \frac{1}{n}\sum_{i=1}^n x_i\psi_i^T + \frac{1}{n}\sum_{i=1}^n \psi_i x_i^T - VV^T\Sigma_\psi VV^T \\
&\|\mathbb{E}[\hat{S}-S]\|_2 = \|\sigma^2 I_d - \sigma^2 P\|_2 = \sigma^2 \qquad \mathbb{E}[\hat{S}-S]V = \Sigma_\psi V - VV^T\Sigma_\psi V = V_\perp V_\perp^T\Sigma_\psi V = 0 \\
&\hat{S} - S - \mathbb{E}[\hat{S}-S] = \frac{1}{n}\sum_{i=1}^n \psi_i\psi_i^T - \Sigma_\psi + \frac{1}{n}\sum_{i=1}^n x_i\psi_i^T + \frac{1}{n}\sum_{i=1}^n \psi_i x_i^T.
\end{aligned}$$

Inserting these expressions we get,

$$\|\hat{P}_t - P\|_2 \leq \frac{\|\frac{1}{n}\sum_{i=1}^n \psi_i\psi_i^T - \Sigma_\psi + \frac{1}{n}\sum_{i=1}^n x_i\psi_i^T + \frac{1}{n}\sum_{i=1}^n \psi_i x_i^T\|_2}{\lambda_m(\frac{1}{n}\sum_{i=1}^n x_i x_i^T) - \|\frac{1}{n}\sum_{i=1}^n \psi_i\psi_i^T - \Sigma_\psi + \frac{1}{n}\sum_{i=1}^n x_i\psi_i^T + \frac{1}{n}\sum_{i=1}^n \psi_i x_i^T\|_2} \tag{13}$$

We first bound $\lambda_m(\frac{1}{n}\sum_{i=1}^n x_i x_i^T)$. From Assumption 1, $\lambda_{max}\big(x_i x_i^T\big) \leq d_x\lambda_+$ for all $i \in [n]$ and from the model properties, $\lambda_m\big(\sum_{i=1}^n \mathbb{E}[x_i x_i^T]\big) = n\lambda_-$. Using Theorem 5, one can get that

$$\Pr\left[\lambda_m\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^T\right) \leq \lambda_-\left(1 - \sqrt{\frac{2d_x g_x}{n}\log\frac{m}{\delta}}\right)\right] \leq \delta. \tag{14}$$

Now we consider $\|\frac{1}{n}\sum_{i=1}^n \psi_i\psi_i^T - \Sigma_\psi + \frac{1}{n}\sum_{i=1}^n x_i\psi_i^T + \frac{1}{n}\sum_{i=1}^n \psi_i x_i^T\|_2$. From triangle inequality we have,

$$\left\|\frac{1}{n}\sum_{i=1}^n \psi_i\psi_i^T - \Sigma_\psi + \frac{1}{n}\sum_{i=1}^n x_i\psi_i^T + \frac{1}{n}\sum_{i=1}^n \psi_i x_i^T\right\|_2 \leq \left\|\frac{1}{n}\sum_{i=1}^n \psi_i\psi_i^T - \Sigma_\psi\right\|_2 + 2\left\|\frac{1}{n}\sum_{i=1}^n x_i\psi_i^T\right\|_2$$

We will consider each term on the right hand side separately. If Assumption 1 holds, then we have:

$$\begin{aligned}
\mathbb{E}[\psi_i\psi_i^T] &= \Sigma_\psi \\
\|\psi_i\psi_i^T\|_2 &\leq d_\psi\sigma^2 \\
\|\mathbb{E}[\psi_i\psi_i^T\psi_i\psi_i^T]\|_2 &\leq d_\psi\sigma^2\|\mathbb{E}[\psi_i\psi_i^T]\|_2 = d_\psi\sigma^4
\end{aligned}$$

Applying Theorem 6, we get

$$\Pr\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\psi_i\psi_i^T-\Sigma_\psi\right\|_2 \geq 2\sigma^2\sqrt{\frac{d_\psi}{n}\log\frac{2d}{\delta}}\right]\leq\delta \quad \text{for } 2\sqrt{\frac{d_\psi}{n}\log\frac{2d}{\delta}}\leq 1.5. \tag{15}$$

Under the same assumption for the second term we have:

$$\begin{aligned}\mathbb{E}[x_i\psi_i^T]&=0\\ \|x_i\psi_i^T\|_2&=\sqrt{\lambda_{max}(\psi_i x_i^T x_i\psi_i^T)}\leq\sqrt{d_x\lambda_+ d_\psi\sigma^2}\\ \|\mathbb{E}[x_i\psi_i^T\psi_i x_i^T]\|_2&\leq d_\psi\sigma^2\|\mathbb{E}[x_i x_i^T]\|_2=d_\psi\lambda_+\sigma^2\\ \|\mathbb{E}[\psi_i x_i^T x_i\psi_i^T]\|_2&\leq d_x\lambda_+\|\mathbb{E}[\psi_i\psi_i^T]\|_2\leq d_x\lambda_+\sigma^2\end{aligned}$$

Once again applying Theorem 6,

$$\Pr\left[\left\|\frac{1}{n}\sum_{i=1}^{n}x_i\psi_i^T\right\|_2\geq 2\sqrt{\lambda_+\sigma^2}\sqrt{\frac{\alpha}{n}\log\frac{2d}{\delta}}\right]\leq\delta \quad \text{for } 2\sqrt{\frac{\alpha}{n}\log\frac{2d}{\delta}}\leq 1.5. \tag{16}$$

Finally, combining (14), (15), (16) and using union bound, for any round $t\geq t_{\min,\delta}$, we get:

$$\|\hat{P}_t-P\|_2\leq\min\left(\frac{\Gamma\sqrt{\frac{\alpha}{tK}\log\frac{2d}{\delta}}}{1-\sqrt{\frac{2d_x g_x}{tK}\log\frac{m}{\delta}}-\Gamma\sqrt{\frac{\alpha}{tK}\log\frac{2d}{\delta}}},1\right) \quad \text{w.p. } 1-3\delta.$$

As explained in the main text, due to equivalence between the projection error and the sine of the largest angle between the subspaces, the projection error is always bounded by 1. Thus, in our bound we impose that constraint. Notice that lower bound on t is to satisfy that concentration inequalities provide meaningful results. In other words, $Kt_{\min,\delta}$ is the number of samples required to have non-negative denominator to use Davis-Kahan $\sin\Theta$ theorem. However, observe that we need $Kt_{w,\delta}$ samples to obtain high probability error bound which is non-trivial, *i.e.* less than 1 and $t_{w,\delta}=4t_{\min,\delta}$. Therefore, for any $t\geq t_{w,\delta}$ the stated bound (12) in the lemma holds with high probability and for any $1\leq t\leq t_{w,\delta}$ we bound the projection error by 1.

Only step remaining to show that lemma holds $\forall t\geq t_{w,\delta}$. This requires an argument which shows that this bound is valid uniformly over all rounds. To this end, we use stopping time construction, which goes back at least to Freedman (1975).

Define the bad event,

$$E_\tau(\delta)=\left\{\|\hat{P}_\tau-P\|_2>\frac{\Gamma\sqrt{\frac{\alpha}{\tau K}\log\frac{2d}{\delta}}}{1-\sqrt{\frac{2d_x g_x}{\tau K}\log\frac{m}{\delta}}-\Gamma\sqrt{\frac{\alpha}{\tau K}\log\frac{2d}{\delta}}}\right\}.$$

We are interested in the probability of $\bigcup_{t\geq t_{w,\delta}}E_t(\delta)$. Define $\tau(\omega)=\min\{t\geq t_{w,\delta}:\omega\in E_t(\delta)\}$, with the convention that $\min\emptyset=\infty$. Then, τ is a stopping time. Thus, $\bigcup_{t\geq t_{w,\delta}}E_t(\delta)=\{\omega:\tau(\omega)<\infty\}$. The Lemma 7 can be obtained as follows:

$$\begin{aligned}\Pr\left[\bigcup_{t\geq t_{w,\delta}}E_t(\delta)\right]=\Pr[\tau<\infty]&=\Pr\left[\|\hat{P}_\tau-P\|_2>\frac{\Gamma\sqrt{\frac{\alpha}{\tau K}\log\frac{2d}{\delta}}}{1-\sqrt{\frac{2d_x g_x}{\tau K}\log\frac{m}{\delta}}-\Gamma\sqrt{\frac{\alpha}{\tau K}\log\frac{2d}{\delta}}},\tau<\infty\right]\\ &=\Pr\left[\|\hat{P}_\tau-P\|_2>\frac{\Gamma\sqrt{\frac{\alpha}{\tau K}\log\frac{2d}{\delta}}}{1-\sqrt{\frac{2d_x g_x}{\tau K}\log\frac{m}{\delta}}-\Gamma\sqrt{\frac{\alpha}{\tau K}\log\frac{2d}{\delta}}}\right]\leq 3\delta.\end{aligned}$$

Finally, notice that Lemma 2 presented in the main text is direct consequence of having denominator at (12) greater than $\frac{1}{2}$ for all $t\geq t_{w,\delta}$.

□

B. Confidence Set Construction Analysis, Proof of Theorem 3

In this section, we state the general version of Theorem 3 and provide the proof details. First, recall that $A_t = \hat{P}_t(\hat{\Sigma}_{t-1} + \lambda I_d)\hat{P}_t$. Let B_t be a symmetric matrix such that $A_t = \hat{V}_t B_t \hat{V}_t^T$. Notice that B_t is a full rank $m \times m$ matrix. Also define $\bar{A}_t = A_t - \lambda \hat{P}_t = \hat{P}_t \hat{\Sigma}_{t-1} \hat{P}_t = \hat{V}_t \hat{V}_t^T \hat{\Sigma}_{t-1} \hat{V}_t \hat{V}_t^T = \hat{V}_t \bar{B}_t \hat{V}_t^T$ where $\bar{B}_t = \hat{V}_t^T \hat{\Sigma}_{t-1} \hat{V}_t = B_t - \lambda I_m$. Using these definitions we can now state the general version of Theorem 3 in which also provides the worst case bound presented in the main text as (10).

Theorem 8. *Fix any $\delta \in (0, 1/4)$. Suppose Assumption 2 holds. If $\|\theta_*\|_2 \leq S$ then, with probability at least $1 - 4\delta$, $\forall t \geq 1$, θ_* lies in the set*

$$\mathcal{C}_{m,t} = \left\{ \theta \in \mathbb{R}^d : \|\theta_t - \theta\|_{A_t} \leq \beta_{t,\delta} \right\},$$

where

$$\beta_{t,\delta} = R\sqrt{2 \log \frac{\det(B_t)^{1/2} \det(\lambda I_m)^{-1/2}}{\delta}} + S\Phi_{t,\delta} \|(A_t^\dagger)^{1/2} \hat{P}_t \hat{\Sigma}_{t-1}\|_2 + S\sqrt{\lambda}. \tag{17}$$

If Assumptions 1 also holds, then with probability at least $1 - 4\delta$, $\forall t \geq t_{w,\delta}$, θ_* lies in the same set with

$$\beta_{t,\delta} = R\sqrt{2 \log\left(\frac{1}{\delta}\right) + m \log\left(1 + \frac{tL^2}{m\lambda}\right)} + 2\Gamma S L \sqrt{\frac{\alpha}{K} \log \frac{2d}{\delta}} \sqrt{\gamma m \log\left(1 + \frac{tL^2}{m\lambda}\right)} + S\sqrt{\lambda}. \tag{18}$$

Proof. Let $S_t \coloneqq \sum_{i=1}^t \hat{P}_t \hat{X}_{i-1} \eta_{i-1} = \hat{P}_t \mathbf{X}_{t-1} \boldsymbol{\eta}_{t-1}$. From the definition of θ_t and r_t, we get the following:

$$\begin{aligned}
\theta_t &= A_t^\dagger S_t + A_t^\dagger \hat{P}_t \hat{\Sigma}_{t-1} P \theta_* \quad \text{since } \theta_* \in \operatorname{span}(V) \\
&= A_t^\dagger S_t + A_t^\dagger \big(\hat{P}_t \hat{\Sigma}_{t-1}(\hat{P}_t + P - \hat{P}_t) + \lambda \hat{P}_t - \lambda \hat{P}_t\big)\theta_* \\
&= A_t^\dagger S_t + \hat{P}_t \theta_* + A_t^\dagger (\hat{P}_t \hat{\Sigma}_{t-1}(P - \hat{P}_t))\theta_* - \lambda A_t^\dagger \theta_*.
\end{aligned}$$

Using this, we derive the following for $x = A_t(\theta_t - \theta_*)$:

$$\begin{aligned}
x^T\theta_t - x^T\theta_* &= x^T A_t^\dagger S_t + x^T A_t^\dagger (\hat{P}_t \hat{\Sigma}_{t-1}(P - \hat{P}_t))\theta_* - \lambda x^T A_t^\dagger \theta_* \\
&= \langle x, S_t \rangle_{A_t^\dagger} + \langle x, \hat{P}_t \hat{\Sigma}_{t-1}(P - \hat{P}_t)\theta_* \rangle_{A_t^\dagger} - \lambda \langle x, \theta_* \rangle_{A_t^\dagger}.
\end{aligned}$$

Using Cauchy-Schwarz inequality, we can upper bound the magnitude of the difference as follows:

$$\begin{aligned}
|x^T\theta_t - x^T\theta_*| &\leq \|x\|_{A_t^\dagger} \big(\|S_t\|_{A_t^\dagger} + \|\hat{P}_t \hat{\Sigma}_{t-1}(P - \hat{P}_t)\theta_*\|_{A_t^\dagger} + \lambda \|\theta_*\|_{A_t^\dagger}\big) \\
&\leq \|x\|_{A_t^\dagger} \big(\|S_t\|_{A_t^\dagger} + \|(A_t^\dagger)^{1/2} \hat{P}_t \hat{\Sigma}_{t-1}(P - \hat{P}_t)\theta_*\|_2 + \sqrt{\lambda}\|\theta_*\|_2\big) \\
&\leq \|x\|_{A_t^\dagger} \big(\|S_t\|_{A_t^\dagger} + \|(A_t^\dagger)^{1/2} \hat{P}_t \hat{\Sigma}_{t-1}\|_2 \|P - \hat{P}_t\|_2 \|\theta_*\|_2 + \sqrt{\lambda}\|\theta_*\|_2\big) \quad \text{Using C.S. again.}
\end{aligned} \tag{19}$$

Plugging in $x = A_t(\theta_t - \theta_*)$, we get

$$\|\theta_t - \theta_*\|_{A_t}^2 \leq \|A_t(\theta_t - \theta_*)\|_{A_t^\dagger} \left(\|S_t\|_{A_t^\dagger} + \|(A_t^\dagger)^{1/2} \hat{P}_t \hat{\Sigma}_{t-1}\|_2 \|(P - \hat{P}_t)\|_2 \|\theta_*\|_2 + \sqrt{\lambda}\|\theta_*\|_2 \right).$$

Since $\|A_t(\theta_t - \theta_*)\|_{A_t^\dagger} = \|\theta_t - \theta_*\|_{A_t}$, dividing both sides with $\|\theta_t - \theta_*\|_{A_t}$ gives and using the fact that $\|\theta_*\| \leq S$,

$$\|\theta_t - \theta_*\|_{A_t} \leq \|S_t\|_{A_t^\dagger} + S\|(A_t^\dagger)^{1/2} \hat{P}_t \hat{\Sigma}_{t-1}\|_2 \|(P - \hat{P}_t)\|_2 + S\sqrt{\lambda} \tag{20}$$

We will now bound each term in the (20) separately. The first term is projected version of Theorem 1 in (Abbasi-Yadkori et al., 2011) and second term is the additional term appearing in the confidence interval construction due to non-zero projection error. As it can be seen with the knowledge of true projection matrix the confidence interval reduces to the one in (Abbasi-Yadkori et al., 2011) with replacement of d with m. We will first provide the theorem that bounds $\|S_t\|_{A_t^\dagger}$ followed by its proof.

Theorem 9. *For any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 1$,*

$$\|S_t\|^2_{A_t^\dagger} \leq 2R^2 \log\left(\frac{\det(B_t)^{1/2}\det(\lambda I_m)^{-1/2}}{\delta}\right).$$

Proof. Without loss of generality, assume that $R = 1$ since by appropriately scaling S_t, this can be achieved. Let $\lambda \in \mathbb{R}^d$ be a Gaussian random vector which is independent of all the other random variables and has covariance matrix $C^{-1} = \frac{1}{\lambda} I_d$. Consider for any $t \geq 0$,

$$M_t^\lambda = \exp\left(\lambda^T S_t - \frac{1}{2}\big(\lambda^T \sum_{i=1}^{t} \hat{P}_t \hat{X}_{i-1}\big)^2\right)$$

Define

$$M_t = \mathbb{E}_\lambda[M_t^\lambda | F_\infty]$$

where F_∞ is the tail σ-algebra of the filtration, *i.e.* the σ-algebra generated by the union of the all the events in the filtration. Thus,

$$M_t = \int_{\mathbb{R}^d} \exp\left(\lambda^T S_t - \frac{1}{2}\lambda^T \hat{P}_t \hat{\Sigma}_{t-1} \hat{P}_t \lambda\right) f(\lambda) d\lambda$$

where $f(\lambda)$ is the pdf of λ. The following lemma will be crucial in proving the theorem.

Lemma 10. $\mathbb{E}[M_t] \leq 1$ *for all* $t \geq 1$.

Proof.

$$\begin{aligned}
\mathbb{E}[M_t] &= \mathbb{E}\bigg[\int_{\mathbb{R}^d} \exp\left(\lambda^T S_t - \frac{1}{2}\lambda^T \hat{P}_t \hat{\Sigma}_{t-1} \hat{P}_t \lambda\right) f(\lambda) d\lambda\bigg] \\
\mathbb{E}[M_t] &= \int_{\mathbb{R}^d} \mathbb{E}\bigg[\exp\left(\lambda^T S_t - \frac{1}{2}\lambda^T \hat{P}_t \hat{\Sigma}_{t-1} \hat{P}_t \lambda\right)\bigg] f(\lambda) d\lambda
\end{aligned}$$

If one can show that $\mathbb{E}\bigg[\exp\left(\lambda^T S_t - \frac{1}{2}\lambda^T \hat{P}_t \hat{\Sigma}_{t-1} \hat{P}_t \lambda\right)\bigg] \leq 1$, then the claim follows. In the following, we use the law of total expectation.

$$\begin{aligned}
\mathbb{E}\bigg[\exp\left(\lambda^T S_t - \frac{1}{2}\lambda^T \hat{P}_t \hat{\Sigma}_{t-1} \hat{P}_t \lambda\right)\bigg] &= \mathbb{E}\bigg[\mathbb{E}_{\eta_{t-1}}\bigg[\exp\left(\lambda^T \sum_{i=1}^{t} \hat{P}_t \hat{X}_{i-1}\eta_{i-1} - \frac{1}{2}\lambda^T \hat{P}_t \big(\sum_{i=1}^{t} \hat{X}_{i-1}\hat{X}_{i-1}^T\big)\hat{P}_t \lambda\right)\bigg| F_{t-1}\bigg]\bigg] \\
&\leq \mathbb{E}\bigg[\exp\left(\lambda^T \sum_{i=1}^{t-1} \hat{P}_t \hat{X}_{i-1}\eta_{i-1} - \frac{1}{2}\lambda^T \hat{P}_t \big(\sum_{i=1}^{t-1} \hat{X}_{i-1}\hat{X}_{i-1}^T\big)\hat{P}_t \lambda\right)\bigg] \qquad (21) \\
&= \mathbb{E}\bigg[\mathbb{E}_{\eta_{t-2}}\bigg[\exp\left(\lambda^T \sum_{i=1}^{t-1} \hat{P}_t \hat{X}_{i-1}\eta_{i-1} - \frac{1}{2}\lambda^T \hat{P}_t \big(\sum_{i=1}^{t-1} \hat{X}_{i-1}\hat{X}_{i-1}^T\big)\hat{P}_t \lambda\right)\bigg| F_{t-2}\bigg]\bigg] \\
&\vdots \\
&\leq 1.
\end{aligned}$$

where 21 follows from the assumption that η_t is conditionally R-sub-gaussian. □

We will use Lemma 10 shortly but we first calculate M_t. For a positive definite matrix K, define $g(K) := \sqrt{(2\pi)^m/\det(K)} = \int_{\mathbb{R}^m} \exp(-\frac{1}{2}x^T K x)dx$. One can calculate M_t as follows,

$$
\begin{aligned}
M_t &= \int_{\mathbb{R}^d} \exp\left(\lambda^T S_t - \frac{1}{2}\lambda^T \bar{A}_t \lambda\right) f(\lambda) d\lambda \\
&= \int_{\mathbb{R}^m} \exp\left(\bar{\lambda}^T \hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t - \frac{1}{2}\bar{\lambda}^T \hat{V}_t^T \mathbf{X}_t \mathbf{X}_t^T \hat{V}_t \bar{\lambda}\right) f(\bar{\lambda}) d\bar{\lambda} \qquad \text{change of integration with } \bar{\lambda} = \hat{V}_t^T \lambda \\
&= \int_{\mathbb{R}^m} \exp\left(- \frac{1}{2}\|\bar{\lambda} - \bar{B}_t^{-1}\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t} + \frac{1}{2}\|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t^{-1}}\right) f(\bar{\lambda}) d\bar{\lambda} \\
&= \frac{\exp\big(\frac{1}{2}\|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t^{-1}}\big)}{g(\bar{C})} \int_{\mathbb{R}^m} \exp\left(- \frac{1}{2}\big(\|\bar{\lambda} - \bar{B}_t^{-1}\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t} + \|\bar{\lambda}\|^2_{\bar{C}}\big)\right) d\bar{\lambda} \qquad \text{where } \bar{C} = \hat{V}_t^T C \hat{V}_t \qquad (22) \\
&= \frac{\exp\big(\frac{1}{2}\|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t^{-1}}\big)}{g(\bar{C})} \int_{\mathbb{R}^m} \exp\left(- \frac{1}{2}\big(\|\bar{\lambda} - (\bar{C} + \bar{B}_t)^{-1}\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{C}+\bar{B}_t} + \|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t^{-1}} - \|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{(\bar{C}+\bar{B}_t)^{-1}}\big)\right) d\bar{\lambda} \qquad (23) \\
&= \frac{\exp\big(\frac{1}{2}\|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{(\bar{C}+\bar{B}_t)^{-1}}\big)}{g(\bar{C})} \int_{\mathbb{R}^m} \exp\left(- \frac{1}{2}\big(\|\bar{\lambda} - (\bar{C} + \bar{B}_t)^{-1}\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{C}+\bar{B}_t}\big)\right) d\bar{\lambda} \\
&= \frac{\exp\big(\frac{1}{2}\|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{(\bar{C}+\bar{B}_t)^{-1}}\big)}{g(\bar{C})} g(\bar{C} + \bar{B}_t) = \left(\frac{\det(\bar{C})}{\det(\bar{C} + \bar{B}_t)}\right)^{1/2} \exp\big(\frac{1}{2}\|S_t\|^2_{(C+\bar{A}_t)^\dagger}\big),
\end{aligned}
$$

where (22) follows from the fact that $f(\bar{\lambda}) = \frac{\exp(-\frac{1}{2}\bar{\lambda}^T \bar{C}\bar{\lambda})}{\sqrt{(2\pi)^m \det(\bar{C}^{-1})}}$ and (23) follows since

$$
\begin{aligned}
\|\bar{\lambda} - \bar{B}_t^{-1}\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t} + \|\bar{\lambda}\|^2_{\bar{C}} &= \|\bar{\lambda} - (\bar{C} + \bar{B}_t)^{-1}\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{C}+\bar{B}_t} + \|\bar{B}_t^{-1}\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t} - \|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{(\bar{C}+\bar{B}_t)^{-1}} \\
&= \|\bar{\lambda} - (\bar{C} + \bar{B}_t)^{-1}\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{C}+\bar{B}_t} + \|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t^{-1}} - \|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{(\bar{C}+\bar{B}_t)^{-1}}.
\end{aligned}
$$

Consider the following equivalence:

$$
\begin{aligned}
\Pr\left[\|S_t\|^2_{(C+\bar{A}_t)^\dagger} > 2\log\left(\frac{\det(\bar{C} + \bar{B}_t)^{1/2}}{\delta \det(\bar{C})^{1/2}}\right)\right] &= \Pr\left[\frac{\exp\big(\frac{1}{2}\|S_t\|^2_{(C+\bar{A}_t)^\dagger}\big)\delta}{\big(\frac{\det(\bar{C}+\bar{B}_t)}{\det(\bar{C})}\big)^{1/2}} > 1\right] \\
&\leq \mathbb{E}\left[\frac{\exp\big(\frac{1}{2}\|S_t\|^2_{(C+\bar{A}_t)^\dagger}\big)\delta}{\big(\frac{\det(\bar{C}+\bar{B}_t)}{\det(\bar{C})}\big)^{1/2}}\right] \qquad (24) \\
&= \mathbb{E}_{F_t}[M_t]\delta \leq \delta \qquad (25)
\end{aligned}
$$

where 24 follows from Markov's inequality and 25 is due to Lemma 10. Notice that, $A_t = \bar{A}_t + C$ and $B_t = \bar{B}_t + \bar{C}$. We will once again use a stopping time construction. Define the bad event,

$$
E_t(\delta) = \left\{\|S_t\|^2_{A_t^\dagger} > 2R^2 \log\left(\frac{\det(B_t)^{1/2}}{\delta \det(\bar{C})^{1/2}}\right)\right\}.
$$

We are interested in the probability of $\bigcup_{t\geq 0} E_t(\delta)$. Define $\tau(\omega) = \min\{t \geq 0 : \omega \in E_t(\delta)\}$, with the convention that $\min \emptyset = \infty$. Then, τ is a stopping time. Thus, $\bigcup_{t\geq 0} E_t(\delta) = \{\omega : \tau(\omega) < \infty\}$. The Theorem 9 can be obtained as follows:

$$
\begin{aligned}
\Pr\left[\bigcup_{t\geq 0} E_t(\delta)\right] = \Pr[\tau < \infty] &= \Pr\left[\|S_\tau\|^2_{A_\tau^\dagger} > 2R^2 \log\left(\frac{\det(B_\tau)^{1/2}\det(\bar{C})^{-1/2}}{\delta}\right), \tau < \infty\right] \\
&\leq \Pr\left[\|S_\tau\|^2_{A_\tau^\dagger} > 2R^2 \log\left(\frac{\det(B_\tau)^{1/2}\det(\bar{C})^{-1/2}}{\delta}\right),\right] \leq \delta.
\end{aligned}
$$

Since $C = \lambda I_d$, inserting $\bar{C} = \lambda I_m$ proves the theorem. □

Combining Theorem 9 with (20) and Lemma 12, we obtain the first statement (17) of Theorem 8:

$$
\|\theta_t - \theta_*\|_{A_t} \leq R\sqrt{2\log\frac{\det(B_t)^{1/2}\det(\lambda I_m)^{-1/2}}{\delta}} + S\Phi_{t,\delta}\|(A_t^\dagger)^{1/2}\hat{P}_t\hat{\Sigma}_{t-1}\|_2 + S\sqrt{\lambda} \qquad (26)
$$

To prove the second statement of the theorem, we need to bound $\|(A_t^\dagger)^{1/2}\hat{P}_t\hat{\Sigma}_{t-1}\|_2$ with the help of Assumptions 1 and 2. Define $B_{t,s} = \hat{V}_t^T(\hat{\Sigma}_{s-1} + \lambda I_d)\hat{V}_t$. Note that $B_{t,t} = B_t$. Now consider the following lemmas which will be used to bound $\|(A_t^\dagger)^{1/2}\hat{P}_t\hat{\Sigma}_{t-1}\|_2$.

Lemma 11. *Suppose Assumptions 1 and 2 hold. Then,* $\det(B_t) \leq \left(\lambda + \frac{tL^2}{m}\right)^m$

Proof. $\det(B_t) = \det(\hat{V}_t^T\hat{\Sigma}_{t-1}\hat{V}_t + \lambda I_m) = \alpha_1\alpha_2\cdots\alpha_m$ where α_is are the eigenvalues of B_t. Notice that

$$\sum_{i=1}^m \alpha_i = m\lambda + \operatorname{tr}\left(\hat{V}_t^T(\sum_{i=1}^t \hat{X}_{i-1}\hat{X}_{i-1}^T)\hat{V}_t\right) = m\lambda + \sum_{i=1}^t \operatorname{tr}\left(\hat{V}_t^T\hat{X}_{i-1}\hat{X}_{i-1}^T\hat{V}_t\right) \leq m\lambda + \sum_{i=1}^t \|\hat{X}_{i-1}\|_2^2 \leq m\lambda + tL^2$$

from Assumptions 1 and 2. Using AM-GM inequality, *i.e,* $\sqrt[m]{\alpha_1\alpha_2\cdots\alpha_m} \leq \frac{1}{m}\sum_{i=1}^m \alpha_i$, we get

$$\alpha_1\alpha_2\cdots\alpha_m \leq \left(\lambda + \frac{tL^2}{m}\right)^m.$$

□

Lemma 12. *Suppose Assumptions 1 and 2 hold. Then*

$$\sum_{i=1}^t \left\|\hat{V}_t^T\hat{X}_{i-1}\right\|^2_{B^{-1}_{t,i-1}} \leq \gamma m \log\left(1 + \frac{tL^2}{m\lambda}\right)$$

Proof. Analyzing $\det(B_t)$ at round t, we get the following:

$$\begin{aligned}\det\left(B_{t,t}\right) &= \det\left(B_{t,t-1} + \hat{V}_t^T\hat{X}_{t-1}\hat{X}_{t-1}^T\hat{V}_t\right) = \det\left(B_{t,t-1}^{1/2}(I_m + B_{t,t-1}^{-1/2}\hat{V}_t^T\hat{X}_{t-1}\hat{X}_{t-1}^T\hat{V}_tB_{t,t-1}^{-1/2})B_{t,t-1}^{1/2}\right)\\ &= \det\left(B_{t,t-1}\right)\left(1 + \|\hat{V}_t^T\hat{X}_{t-1}\|^2_{B^{-1}_{t,t-1}}\right) = \lambda^m\prod_{i=1}^t\left(1 + \|\hat{V}_t^T\hat{X}_{i-1}\|^2_{B^{-1}_{t,i-1}}\right)\end{aligned}$$

Thus, $\sum_{i=1}^t \log(1+\|\hat{V}_t^T\hat{X}_{i-1}\|^2_{B^{-1}_{t,i-1}}) = \log\frac{\det(B_t)}{\lambda^m} \leq m\log\left(1+\frac{tL^2}{m\lambda}\right)$ where inequality follows from Lemma 11. Recall the definition of $\gamma = \frac{L^2}{\lambda\log\left(1+\frac{L^2}{\lambda}\right)}$. Since Assumption 1 and 2 hold, $\|\hat{V}_t^T\hat{X}_{i-1}\|^2_{B^{-1}_{t,i-1}} \leq \frac{L^2}{\lambda}$. Using $\|\hat{V}_t^T\hat{X}_{i-1}\|^2_{B^{-1}_{t,i-1}} \leq \gamma\log(1 + \|\hat{V}_t^T\hat{X}_{i-1}\|^2_{B^{-1}_{t,i-1}})$, which is true for $\|\hat{V}_t^T\hat{X}_{i-1}\|^2_{B^{-1}_{t,i-1}} \leq \frac{L^2}{\lambda}$, we get

$$\sum_{i=1}^t \left\|\hat{V}_t^T\hat{X}_{i-1}\right\|^2_{B^{-1}_{t,i-1}} \leq \gamma\sum_{i=1}^t \log(1 + \|\hat{V}_t^T\hat{X}_{i-1}\|^2_{B^{-1}_{t,i-1}})$$

The lemma follows immediately. □

Finally, we provide the bound on $\|(A_t^\dagger)^{1/2}\hat{P}_t\hat{\Sigma}_{t-1}\|_2$ as follows,

Lemma 13. *Suppose Assumptions 1 and 2 hold. Then,* $\|(A_t^\dagger)^{1/2}\hat{P}_t\hat{\Sigma}_{t-1}\|_2 \leq L\sqrt{t}\sqrt{\gamma m}\sqrt{\log\left(1 + \frac{tL^2}{m\lambda}\right)}$.

Proof. Recall the definition of $\hat{\Sigma}_{t-1} = \sum_{i=1}^{t-1} \hat{X}_i \hat{X}_i^T$. Using this, we get

$$\begin{aligned}
\|(A_t^\dagger)^{1/2}\hat{P}_t\hat{\Sigma}_{t-1}\|_2 &= \left\| \sum_{i=1}^{t} (A_t^\dagger)^{1/2}\hat{P}_t\hat{X}_{i-1}\hat{X}_{i-1}^T \right\|_2 \\
&\leq \sum_{i=1}^{t} \left\|(A_t^\dagger)^{1/2}\hat{P}_t\hat{X}_{i-1}\hat{X}_{i-1}^T\right\|_2 \quad \text{Using Weyl's inequality for singular values} \\
&\leq \sum_{i=1}^{t} \left\|(A_t^\dagger)^{1/2}\hat{P}_t\hat{X}_{i-1}\right\|_2 \|\hat{X}_{i-1}\|_2 \quad \text{From Cauchy Schwarz} \\
&\leq L\sum_{i=1}^{t} \left\|(A_t^\dagger)^{1/2}\hat{P}_t\hat{X}_{i-1}\right\|_2 \quad \text{From Assumption 1} \\
&= L\sum_{i=1}^{t} \left\|\hat{P}_t\hat{X}_{i-1}\right\|_{A_t^\dagger} \\
&= L\sum_{i=1}^{t} \left\|\hat{V}_t^T\hat{X}_{i-1}\right\|_{B_t^{-1}} \quad \text{From the equality that } \hat{X}_{i-1}^T\hat{V}_tB_t^{-1}\hat{V}_t^T\hat{X}_{i-1} = \hat{X}_{i-1}^T\hat{P}_tA_t^\dagger\hat{P}_t\hat{X}_{i-1} \\
&\leq L\sum_{i=1}^{t} \left\|\hat{V}_t^T\hat{X}_{i-1}\right\|_{B_{t,i-1}^{-1}} \quad \text{Since at round t, } B_{t,i} = B_{t,i-1} + \hat{V}_t^T\hat{X}_i\hat{X}_i^T\hat{V}_t \\
&\leq L\sqrt{t}\sqrt{\sum_{i=1}^{t} \left\|\hat{V}_t^T\hat{X}_{i-1}\right\|_{B_{t,i-1}^{-1}}^2} \leq L\sqrt{\gamma m t}\sqrt{\log\left(1+\frac{tL^2}{m\lambda}\right)} \quad \text{From Lemma 12}
\end{aligned}$$

□

Now that we obtain bounds on every term at (26), we can obtain the second statement of Theorem 8 directly. For the described setting in the theorem, using Lemma 11 and Lemma 13, we get the following

$$\begin{aligned}
\|\theta_t - \theta_*\|_{A_t} &\leq R\sqrt{2\log\left(\frac{1}{\delta}\right) + m\log\left(1+\frac{tL^2}{m\lambda}\right)} + SL\sqrt{\gamma m t}\sqrt{\log\left(1+\frac{tL^2}{m\lambda}\right)}\|(P-\hat{P}_t)\|_2 + S\sqrt{\lambda} \\
&\leq R\sqrt{2\log\left(\frac{1}{\delta}\right) + m\log\left(1+\frac{tL^2}{m\lambda}\right)} + 2\Gamma SL\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}}\sqrt{\gamma m\log\left(1+\frac{tL^2}{m\lambda}\right)} + S\sqrt{\lambda}
\end{aligned}$$

where the last inequality gives (10) due to Lemma 2. □

C. Regret Analysis, Proof of Theorem 1

First consider the following lemma.

Lemma 14. *At round k, let $\hat{x} \in D_k$. If $\nu \in C_k$, then*

$$|(\hat{P}_k\hat{x})^T(\nu - \theta_k)| \leq \beta_{k,\delta}\|\hat{x}\|_{A_k^\dagger}.$$

Proof.

$$\begin{aligned}
|(\hat{P}_k\hat{x})^T(\nu-\theta_k)| &= |(\hat{P}_k\hat{x})^T(A_k^\dagger)^{1/2}A_k^{1/2}(\nu-\theta_k)| \qquad \text{since } (A_k^\dagger)^{1/2}A_k^{1/2} = \hat{P}_k \\
&= |(A_k^\dagger)^{1/2}\hat{P}_k\hat{x})^TA_k^{1/2}(\nu-\theta_k)| \\
&\leq \|(A_k^\dagger)^{1/2}\hat{P}_k\hat{x}\|_2\|A_k^{1/2}(\nu-\theta_k)\|_2 \quad \text{by C.S.} \\
&\leq \beta_{k,\delta}\|\hat{P}_k\hat{x}\|_{A_k^\dagger} \qquad \text{since } \nu \in C_k. \\
&= \beta_{k,\delta}\|\hat{V}_k^T\hat{x}\|_{B_k^{-1}} = \beta_{k,\delta}\|\hat{x}\|_{A_k^\dagger}
\end{aligned}$$

□

Before providing the proof of Theorem 1, consider the following lemmas:

Lemma 15. *For all $t \geq t_{w,\delta}$, with probability at least $1-\delta$*

$$\lambda_m(\hat{P}_t\hat{\Sigma}_{t-1}\hat{P}_t) \geq (t-1)(\lambda_- + \sigma^2) - \sqrt{t-1}\left(4L^2\Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}} + \sqrt{2L(\lambda_- + \sigma^2)\log\frac{m}{\delta}}\right) \tag{27}$$

Define $t_{r,\delta} = 1 + \left(\frac{8L^2\Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}}+\sqrt{8L(\lambda_-+\sigma^2)\log\frac{m}{\delta}}}{\lambda_-+\sigma^2}\right)^2$. Then for all $t \geq t_{r,\delta}$, with probability at least $1-\delta$,

$$\lambda_m(\hat{P}_t\hat{\Sigma}_{t-1}\hat{P}_t) \geq \frac{(\lambda_- + \sigma^2)}{2}(t-1). \tag{28}$$

Proof.

$$\begin{aligned}\lambda_m(\hat{P}_t\hat{\Sigma}_{t-1}\hat{P}_t) &= \lambda_m\big((\hat{P}_t - P)\hat{\Sigma}_{t-1}\hat{P}_t + P\hat{\Sigma}_{t-1}(\hat{P}_t - P) + P\hat{\Sigma}_{t-1}P\big)\\ &\geq \lambda_m(P\hat{\Sigma}_{t-1}P) - 2(t-1)L^2\|\hat{P}_t - P\|_2\\ &\geq \lambda_{\min}(V^T\hat{\Sigma}_{t-1}V) - 4L^2\Gamma\sqrt{\frac{\alpha(t-1)}{K}\log\frac{2d}{\delta}} \quad \text{from Lemma 2}\end{aligned}$$

We also have that:

$$\lambda_{\max}(V^T\hat{X}_j\hat{X}_j^TV) \leq L \quad \forall j \in \{1,\ldots,i-1\}$$
$$\lambda_{\min}\left(\mathbb{E}\big[\sum_{j=1}^{t-1}V^T\hat{X}_j\hat{X}_j^TV\big]\right) = (t-1)(\lambda_- + \sigma^2).$$

Applying Theorem 5,

$$\Pr\left[\lambda_{\min}(V^T\hat{\Sigma}_tV) \leq (t-1)(\lambda_- + \sigma^2) - \sqrt{2L(t-1)(\lambda_- + \sigma^2)\log\frac{m}{\delta}}\right] \leq \delta.$$

Combining these with similar stopping time construction as described in previous sections we derive the first statement of lemma. Now for second statement with a constant C, observe that, $(t-1)(\lambda_- + \sigma^2) - \sqrt{t-1}\left(4L^2\Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}} + \sqrt{2L(\lambda_- + \sigma^2)\log\frac{m}{\delta}}\right) \geq C(t-1)$ holds if and only if $t \geq 1 + \left(\frac{4L^2\Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}}+\sqrt{2L(\lambda_-+\sigma^2)\log\frac{m}{\delta}}}{\lambda_-+\sigma^2-C}\right)^2$. Choosing $C = \frac{\lambda_-+\sigma^2}{2}$ proves the bound. □

Finally, we state one more lemma which will help us derive the regret bound.

Lemma 16.

$$2\sqrt{t+1} - 2 \leq \sum_{i=1}^{t}\frac{1}{\sqrt{i}} \leq 2\sqrt{t} - 1 \qquad\qquad \log(t+1) \leq \sum_{i=1}^{t}\frac{1}{i} \leq 1 + \log(t)$$

Proof. First one can be obtained using integral estimates and the second one is due harmonic sums. □

Proof of Theorem 1. The instantaneous regret, l_i of the algorithm at ith round can be decomposed as follows:

$$\begin{aligned}l_i &= \hat{X}_i^{*T}\theta_* - \hat{X}_i^T\theta_*\\ &\leq (\tilde{P}_i\hat{X}_i)^T\tilde{\theta}_i - (P\hat{X}_i)^T\theta_* \qquad \text{since } (\tilde{P}_i, \hat{X}_i, \tilde{\theta}_i) \text{ is optimistic}\\ &= \hat{X}_i^T(\tilde{P}_i - \hat{P}_i + \hat{P}_i)\tilde{\theta}_i - \hat{X}_i^T(\hat{P}_i + P - \hat{P}_i)\theta_*\\ &= (\hat{P}_i\hat{X}_i)^T(\tilde{\theta}_i - \theta_i) + (\hat{P}_i\hat{X}_i)^T(\theta_i - \theta_*) + ((\hat{P}_i - P)\hat{X}_i)^T\theta_* + ((\tilde{P}_i - \hat{P}_i)\hat{X}_i)^T\tilde{\theta}_i\\ &\leq 2\beta_{i,\delta}\|\hat{X}_i\|_{A_i^\dagger} + 2LS\|\hat{P}_i - P\|_2 \quad \text{holds } \forall i \text{ w.p. } 1-4\delta \text{ due to Lemma 14 and Theorem 3.}\end{aligned}$$

Combining this decomposition with the fact that $l_i \leq 2$, we get

$$\begin{aligned}l_i &\leq 2\min\left(\beta_{i,\delta}\|\hat{X}_i\|_{A_i^\dagger} + LS\|\hat{P}_i - P\|_2, \quad 1\right)\\ &\leq 2\min(\beta_{i,\delta}\|\hat{X}_i\|_{A_i^\dagger}, 1) + 2LS\min(\|\hat{P}_i - P\|_2, 1)\\ &\leq 2\beta_{i,\delta}\min(\|\hat{X}_i\|_{A_i^\dagger}, 1) + 2LS\|\hat{P}_i - P\|_2\end{aligned} \tag{29}$$

where the last inequality is due to considering the regret of the algorithm after warm-up period which provides that $\|\hat{P}_i - P\|_2 < 1$. Now we can provide an upper bound on the regret. For all $t \geq 1$, with probability at least $1 - 5\delta$,

$$
\begin{aligned}
R_t &\leq \sum_{i=1}^{t} 2\beta_{i,\delta} \min(\|\hat{X}_i\|_{A_i^\dagger}, 1) + 2LS\|\hat{P}_i - P\|_2 \\
&= 2LS \sum_{i=1}^{t} \|\hat{P}_i - P\|_2 + \sum_{i=1}^{t} 2\beta_{i,\delta} \min(\|\hat{X}_i\|_{A_i^\dagger}, 1) \\
&\leq 2LS \sum_{i=1}^{t} \|\hat{P}_i - P\|_2 + 2\beta_{t,\delta} \sum_{i=1}^{t} \min(\|\hat{X}_i\|_{A_i^\dagger}, 1) && (30) \\
&\leq 2LS \sum_{i=1}^{t} \|\hat{P}_i - P\|_2 + 2\beta_{t,\delta} \sqrt{t \sum_{i=1}^{t} \min(\|\hat{X}_i\|^2_{A_i^\dagger}, 1)} \\
&\leq 2LS \sum_{i=1}^{t} \|\hat{P}_i - P\|_2 + 2\sqrt{t}\beta_{t,\delta} \sqrt{\sum_{i=1}^{t} \min\left(\lambda_{\max}(A_i^\dagger) L^2, 1\right)} && (31) \\
&\leq 2LS \sum_{i=1}^{t} \|\hat{P}_i - P\|_2 + 2\sqrt{t}\beta_{t,\delta} \sqrt{\sum_{i=1}^{t} \min\left(\frac{L^2}{\lambda + \lambda_m(\hat{P}_i \hat{\Sigma}_{i-1} \hat{P}_i)}, 1\right)} && (32) \\
&\leq 2LS\Bigg(t_{w,\delta} + 2\Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}} \sum_{i=t_{w,\delta}}^{t} \frac{1}{\sqrt{i}}\Bigg) && (33) \\
&\quad + 2\sqrt{t}\beta_{t,\delta} \sqrt{\sum_{i=1}^{t} \min\left(\frac{L^2}{\lambda + \max\Big((i-1)(\lambda_- + \sigma^2) - \sqrt{i-1}\big(4L^2\Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}} + \sqrt{2L(\lambda_- + \sigma^2)\log\frac{m}{\delta}}\big), 0\Big)}, 1\right)} \\
&\leq 2LS\Bigg(t_{w,\delta} + 2\Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}} \sum_{i=t_{w,\delta}}^{t} \frac{1}{\sqrt{i}}\Bigg) + 2L\sqrt{t}\beta_{t,\delta}\sqrt{\frac{t_{r,\delta}}{\lambda} + \frac{2}{\lambda_- + \sigma^2}\sum_{i=t_{r,\delta}}^{t}\frac{1}{i}} && (34) \\
&\leq 2LSt_{w,\delta} + 4LS\Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}}(2\sqrt{t} - 2\sqrt{t_{w,\delta} + 1} + 1) + 2L\sqrt{t}\beta_{t,\delta}\sqrt{\frac{t_{r,\delta}}{\lambda} + \frac{2 + 2\log t - 2\log(t_{r,\delta} + 1)}{\lambda_- + \sigma^2}} && (35)
\end{aligned}
$$

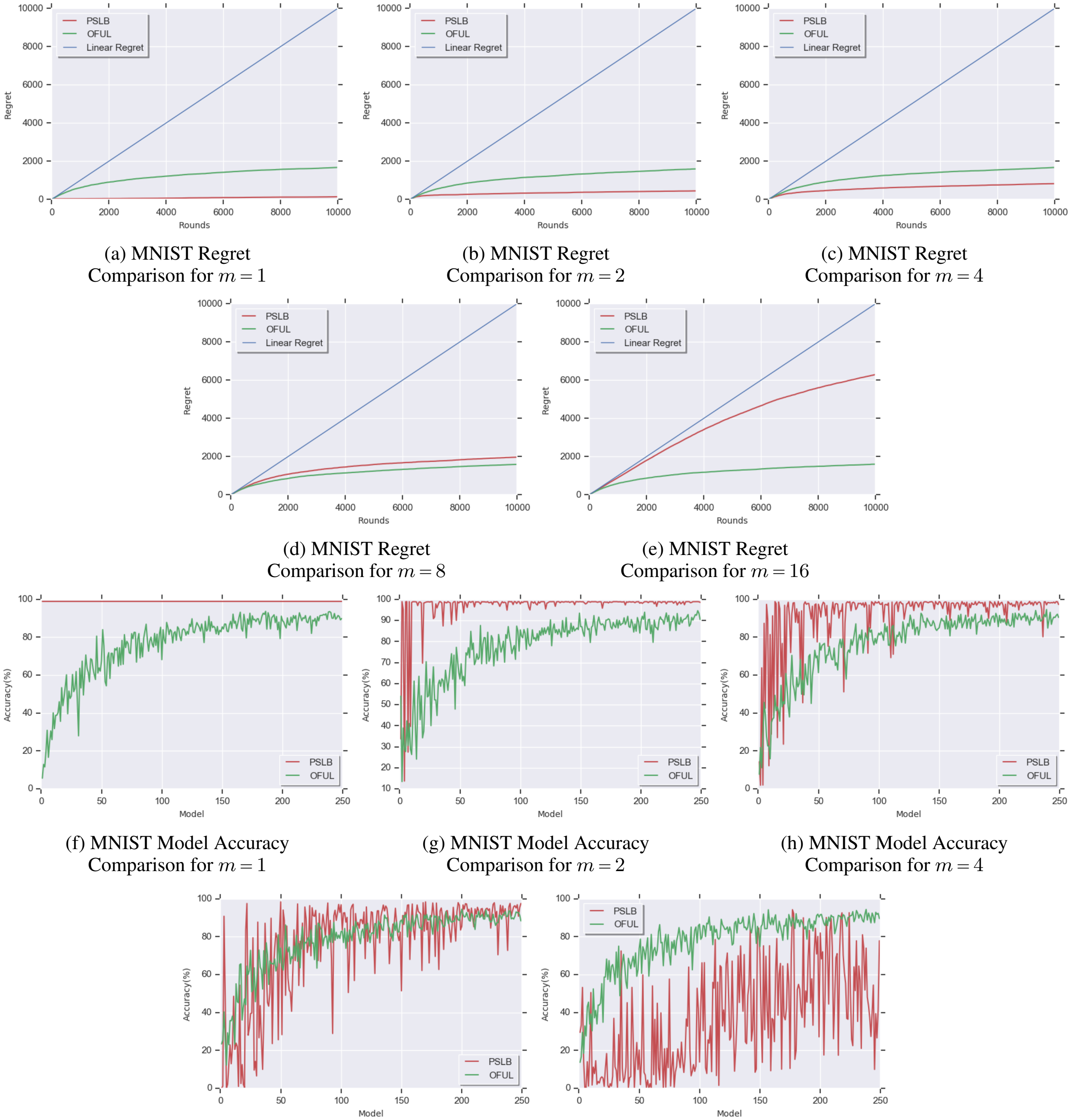
where (30) follows from the fact that $\beta_{1,\delta} \leq \cdots \leq \beta_{t,\delta}$, (31) follows since $\|x\|_M \leq \lambda_{\max}(M)\|x\|_2$. Maximum eigenvalue of $A_t^\dagger$ is equivalent to mth eigenvalue of A_t, thus (32) is obtained. Using Lemma 2 with Lemma 15 we get (33). Using the second statement of Lemma 15 gives (34). Finally, Lemma 16 provides the bound on regret shown in (35).

Recall that $\beta_{t,\delta} = \mathcal{O}\left(\Gamma\sqrt{\frac{\alpha m}{K}\log t}\right)$. Therefore, last term dominates the asymptotic upper bound on regret. Using the definition of $t_{r,\delta}$ we get that the regret of the algorithm is

$$
R_t \leq \mathcal{O}\left(\frac{\alpha\Gamma^2\sqrt{m}}{K(\lambda_- + \sigma^2)}\sqrt{t}\log t\right) \qquad (36)
$$

From the definition of Υ and the fact that the PSLB uses the confidence set of $\mathcal{C}_t = \mathcal{C}_{m,t} \cap \mathcal{C}_{d,t}$, the theorem follows. □

D. Additional Experiment Results



(a) MNIST Regret Comparison for $m = 1$

(b) MNIST Regret Comparison for $m = 2$

(c) MNIST Regret Comparison for $m = 4$

(d) MNIST Regret Comparison for $m = 8$

(e) MNIST Regret Comparison for $m = 16$

(f) MNIST Model Accuracy Comparison for $m = 1$

(g) MNIST Model Accuracy Comparison for $m = 2$

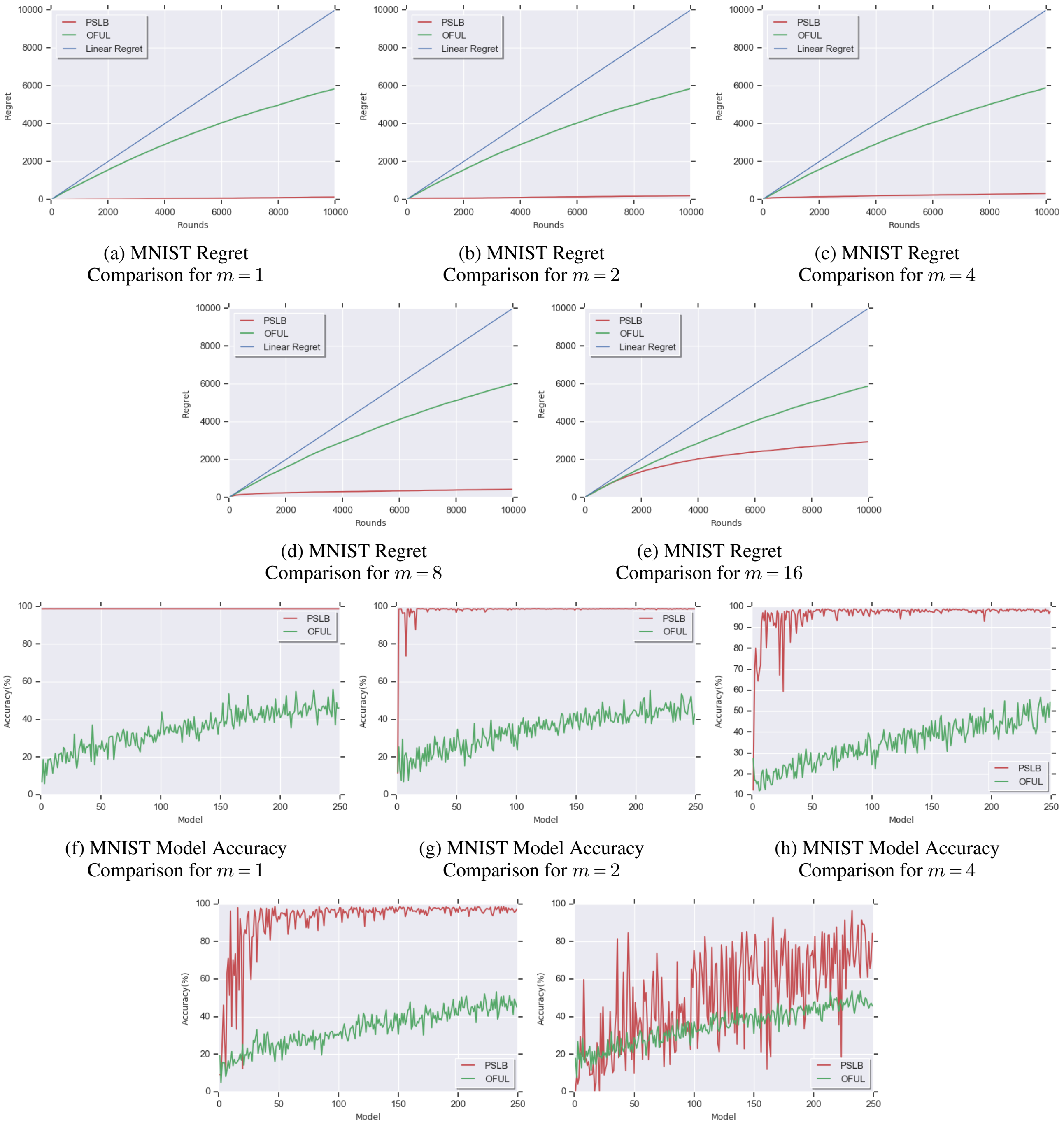
(h) MNIST Model Accuracy Comparison for $m = 4$

(i) MNIST Model Accuracy Comparison for $m = 8$

(j) MNIST Model Accuracy Comparison for $m = 16$

Figure 2: Regret and Optimistic Model Accuracy Comparisons of PSLB and OFUL on MNIST with $d = 100$

Throughout Section D, while running PSLB, only projected confidence sets are used in choosing optimistic actions. This way, we show the effect of subspace recovery problem on the regret of PSLB explicitly. Figure 2 provides the regret and the accuracy of optimistically chosen parameters of PSLB and OFUL in SLB constructed from MNIST with $d = 100$. Figures 2a, 2b, 2c, 2d, 2e show the regrets obtained while PSLB tries to recover $m = 1, 2, 4, 8, 16$ dimensional subspaces respectively. Since the feature space is only 100-dimensional PSLB is not as superior over OFUL as in high dimensional cases like $d = 500, 1000$. Note that as we search for a higher dimensional subspace, the subspace becomes less identifiable and finite sample projection error starts to dominate the regret. For 100-dimensional MNIST SLB setting, when PSLB tries to recover a 8-dimensional or bigger subspace, OFUL starts to dominate PSLB. Fortunately, by using the intersection of confidence sets approach, PSLB tolerates this and performs at least as good as OFUL.



(a) MNIST Regret Comparison for $m = 1$

(b) MNIST Regret Comparison for $m = 2$

(c) MNIST Regret Comparison for $m = 4$

(d) MNIST Regret Comparison for $m = 8$

(e) MNIST Regret Comparison for $m = 16$

(f) MNIST Model Accuracy Comparison for $m = 1$

(g) MNIST Model Accuracy Comparison for $m = 2$

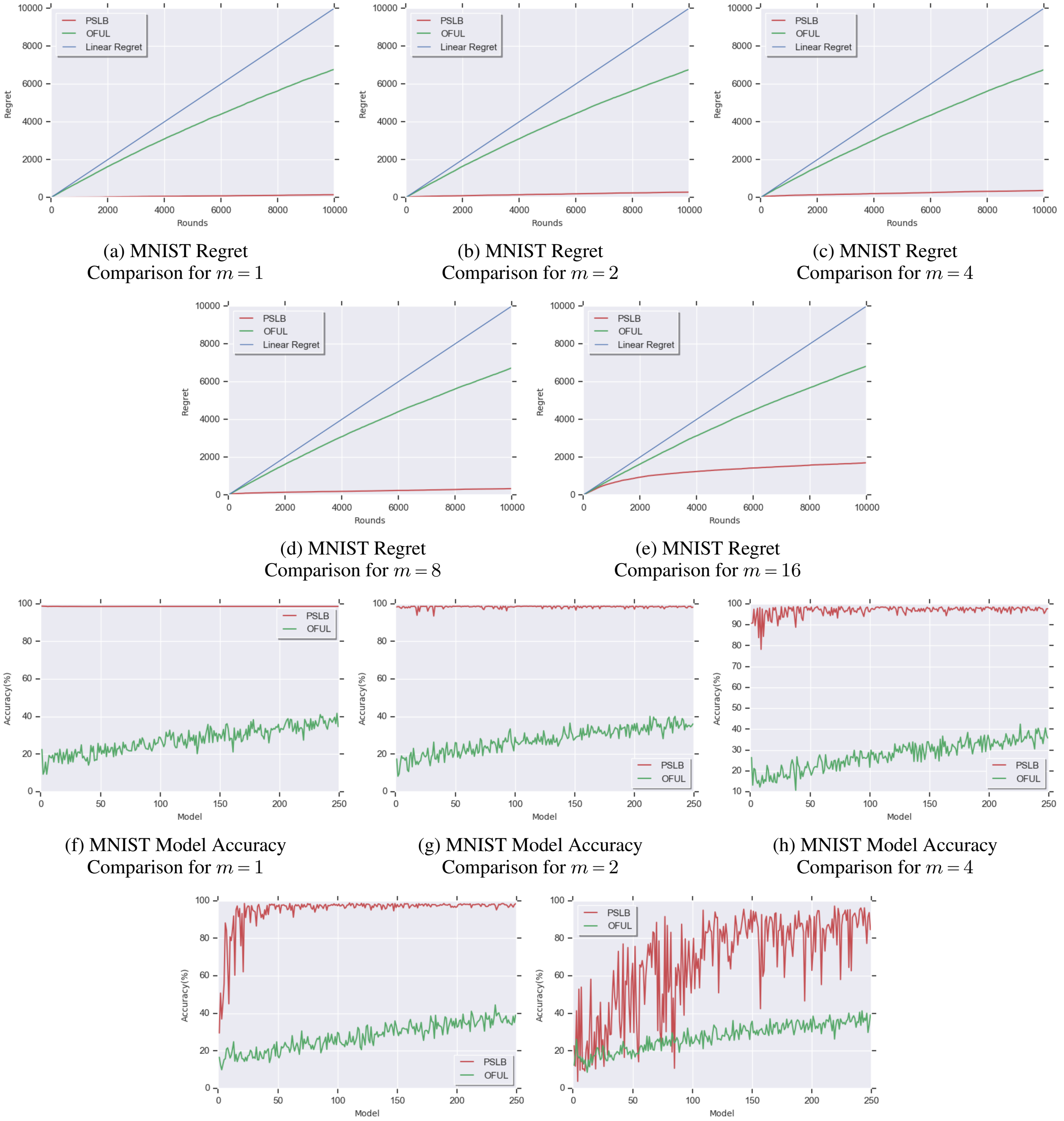
(h) MNIST Model Accuracy Comparison for $m = 4$

(i) MNIST Model Accuracy Comparison for $m = 8$

(j) MNIST Model Accuracy Comparison for $m = 16$

Figure 3: Regret and Optimistic Model Accuracy Comparisons of PSLB and OFUL on MNIST with $d = 500$

Figure 2 provides the regret and the accuracy of optimistically chosen parameters of PSLB and OFUL in SLB constructed from MNIST with $d = 500$. As we go in higher dimensional representations, the benefit of subspace recovery on regret becomes more apparent. In Figure 3, it can be seen that PSLB has smaller regret for each choice of m. With the PCA based subspace recovery, even for recovering higher dimensional subspaces like 8 dimensions, PSLB performs well. It explores some in the beginning and as the subspace estimation gets more accurate it converges to accurate model. This behavior can be seen in Figure 3i.



(a) MNIST Regret Comparison for $m = 1$

(b) MNIST Regret Comparison for $m = 2$

(c) MNIST Regret Comparison for $m = 4$

(d) MNIST Regret Comparison for $m = 8$

(e) MNIST Regret Comparison for $m = 16$

(f) MNIST Model Accuracy Comparison for $m = 1$

(g) MNIST Model Accuracy Comparison for $m = 2$

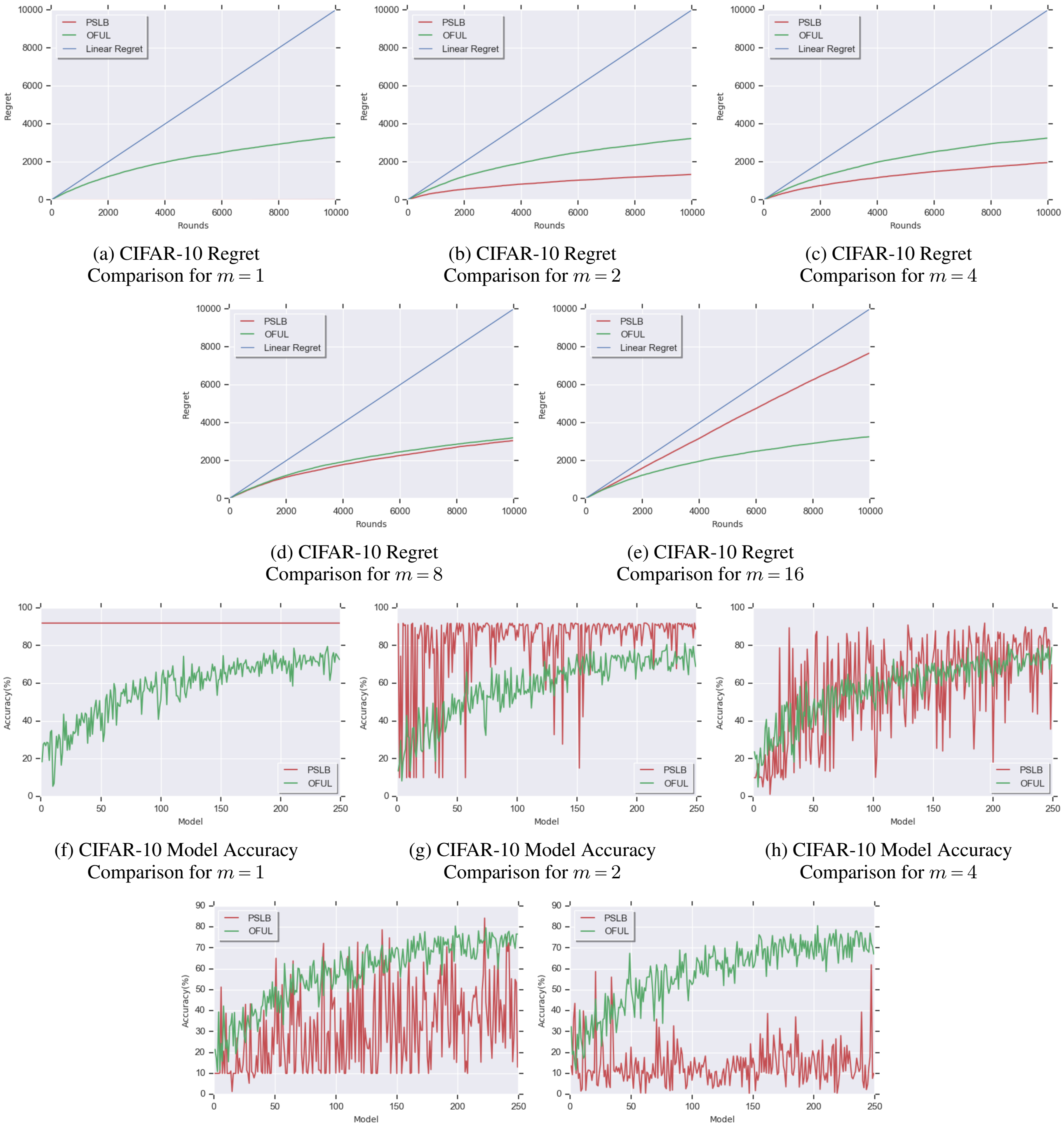
(h) MNIST Model Accuracy Comparison for $m = 4$

(i) MNIST Model Accuracy Comparison for $m = 8$

(j) MNIST Model Accuracy Comparison for $m = 16$

Figure 4: Regret and Optimistic Model Accuracy Comparisons of PSLB and OFUL on MNIST with $d = 1000$

Figure 4 provides the regret and the accuracy of optimistically chosen parameters of PSLB and OFUL in SLB constructed from MNIST with $d = 1000$. This is the setting where PSLB becomes significantly superior to OFUL. In all choices of m, PSLB learns the underlying model accurately and starts exploiting this information. However, OFUL still continues to explore in each dimension to figure out the underlying model. Therefore, it needs significantly more samples to achieve the classification performance of PSLB and during that time it continues to make mistakes and accumulate regret.



(a) CIFAR-10 Regret Comparison for $m = 1$

(b) CIFAR-10 Regret Comparison for $m = 2$

(c) CIFAR-10 Regret Comparison for $m = 4$

(d) CIFAR-10 Regret Comparison for $m = 8$

(e) CIFAR-10 Regret Comparison for $m = 16$

(f) CIFAR-10 Model Accuracy Comparison for $m = 1$

(g) CIFAR-10 Model Accuracy Comparison for $m = 2$

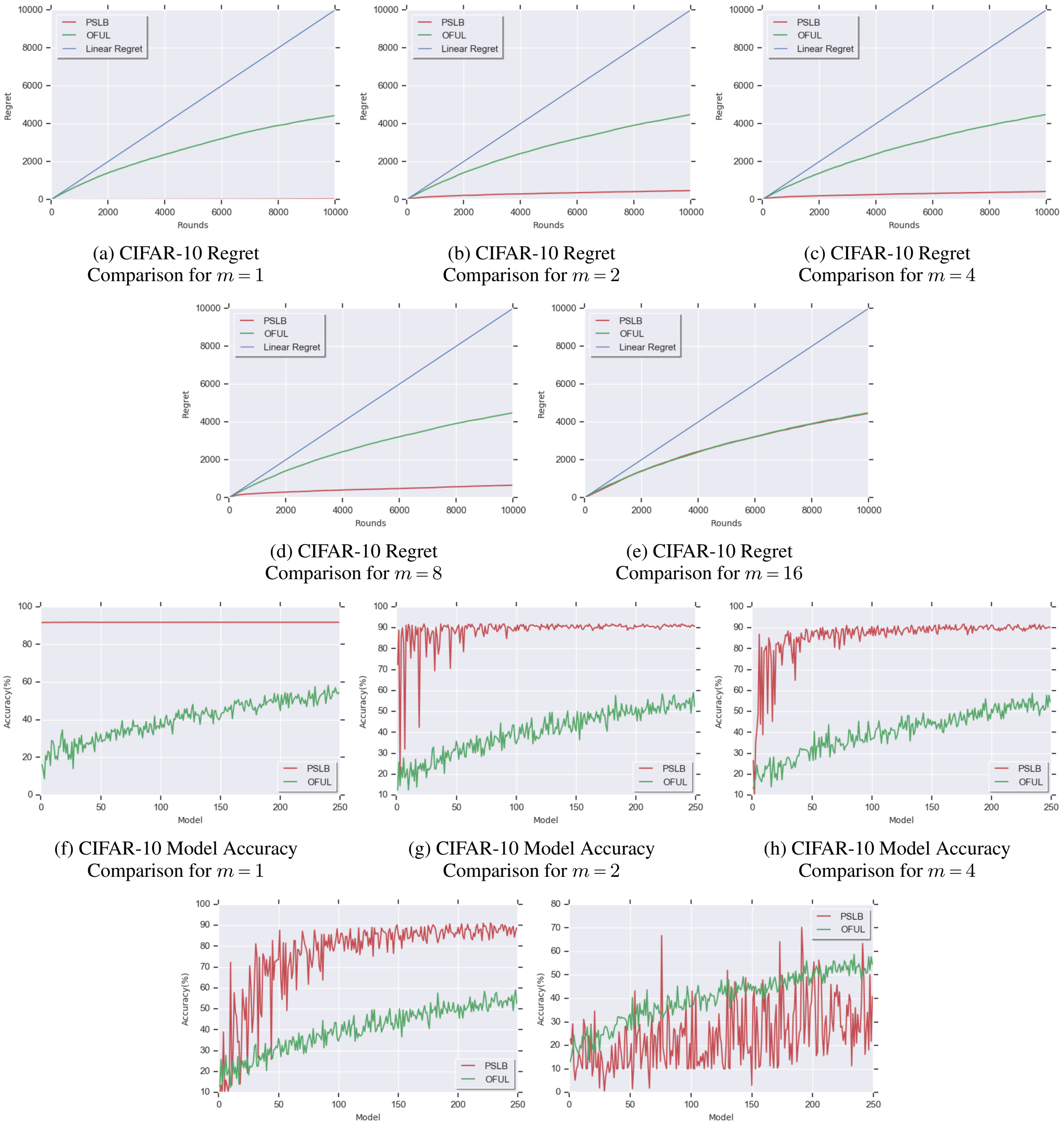
(h) CIFAR-10 Model Accuracy Comparison for $m = 4$

(i) CIFAR-10 Model Accuracy Comparison for $m = 8$

(j) CIFAR-10 Model Accuracy Comparison for $m = 16$

Figure 5: Regret and Optimistic Model Accuracy Comparisons of PSLB and OFUL on CIFAR-10 with $d = 100$

Figure 5 provides the regret and the accuracy of optimistically chosen parameters of PSLB and OFUL in SLB constructed from CIFAR-10 with $d = 100$. Similar to MNIST, due to difficulty of subspace recovery for high-dimensional subspaces, using projected confidence sets doesn't provide substantial benefit compared to OFUL except $m = 1, 2$ and 4. However, the best of both algorithms approach of PSLB bounds our regret with the regret of OFUL which performs well under low dimensional ambient spaces. Moreover, we should note that by projecting the decision set onto 1-dimensional subspace, PSLB makes almost no mistakes during the course of interaction, Fig 5a.



(a) CIFAR-10 Regret Comparison for $m = 1$

(b) CIFAR-10 Regret Comparison for $m = 2$

(c) CIFAR-10 Regret Comparison for $m = 4$

(d) CIFAR-10 Regret Comparison for $m = 8$

(e) CIFAR-10 Regret Comparison for $m = 16$

(f) CIFAR-10 Model Accuracy Comparison for $m = 1$

(g) CIFAR-10 Model Accuracy Comparison for $m = 2$

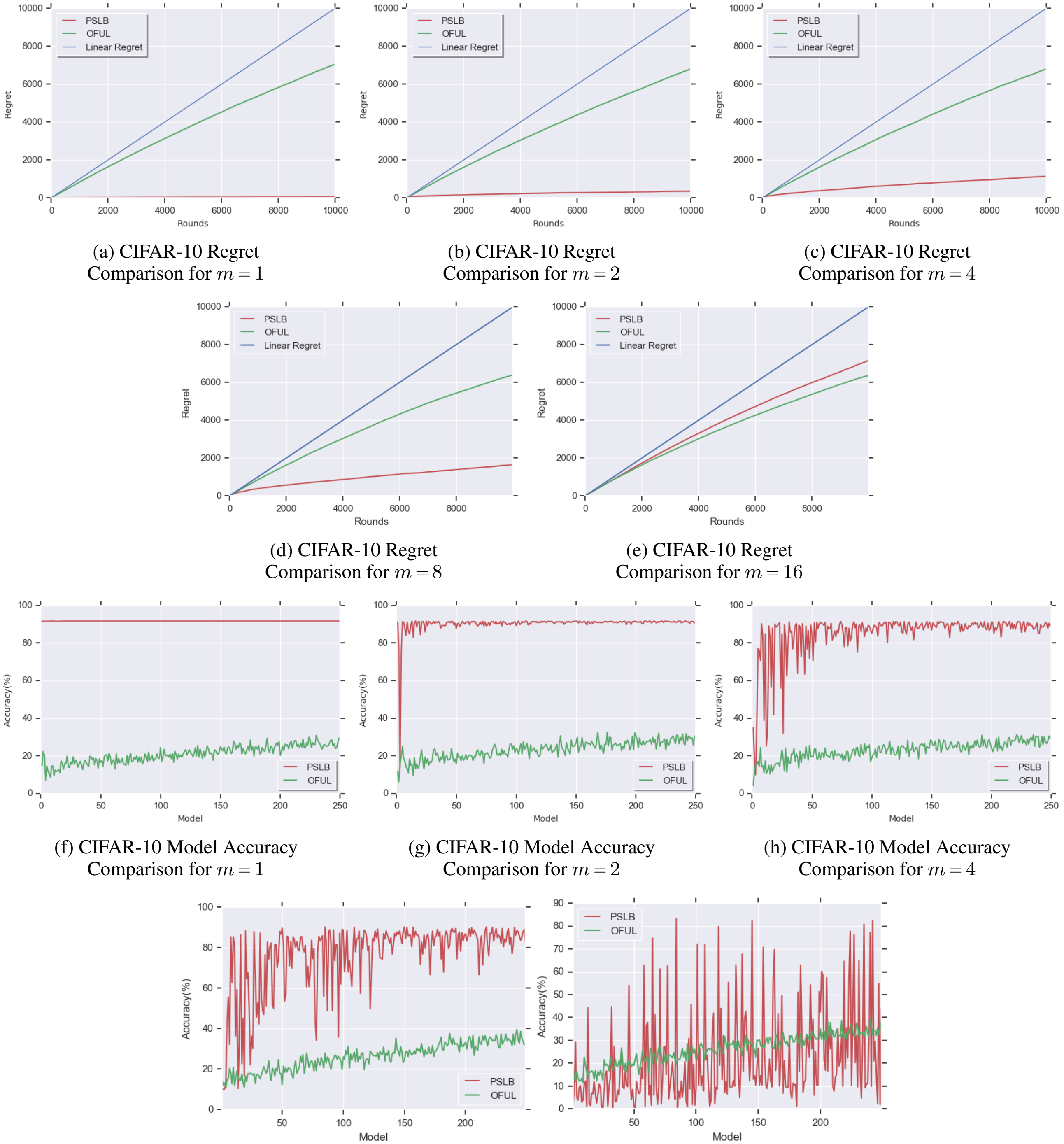
(h) CIFAR-10 Model Accuracy Comparison for $m = 4$

(i) CIFAR-10 Model Accuracy Comparison for $m = 8$

(j) CIFAR-10 Model Accuracy Comparison for $m = 16$

Figure 6: Regret and Optimistic Model Accuracy Comparisons of PSLB and OFUL on CIFAR-10 with $d = 500$

Figure 6 provides the regret and the accuracy of optimistically chosen parameters of PSLB and OFUL in SLB constructed from CIFAR-10 with $d = 500$. Similar to $d = 100$ setting, PSLB makes very few mistakes when it tries to recover and project action vectors onto a 1-dimensional subspace.



(a) CIFAR-10 Regret Comparison for $m = 1$

(b) CIFAR-10 Regret Comparison for $m = 2$

(c) CIFAR-10 Regret Comparison for $m = 4$

(d) CIFAR-10 Regret Comparison for $m = 8$

(e) CIFAR-10 Regret Comparison for $m = 16$

(f) CIFAR-10 Model Accuracy Comparison for $m = 1$

(g) CIFAR-10 Model Accuracy Comparison for $m = 2$

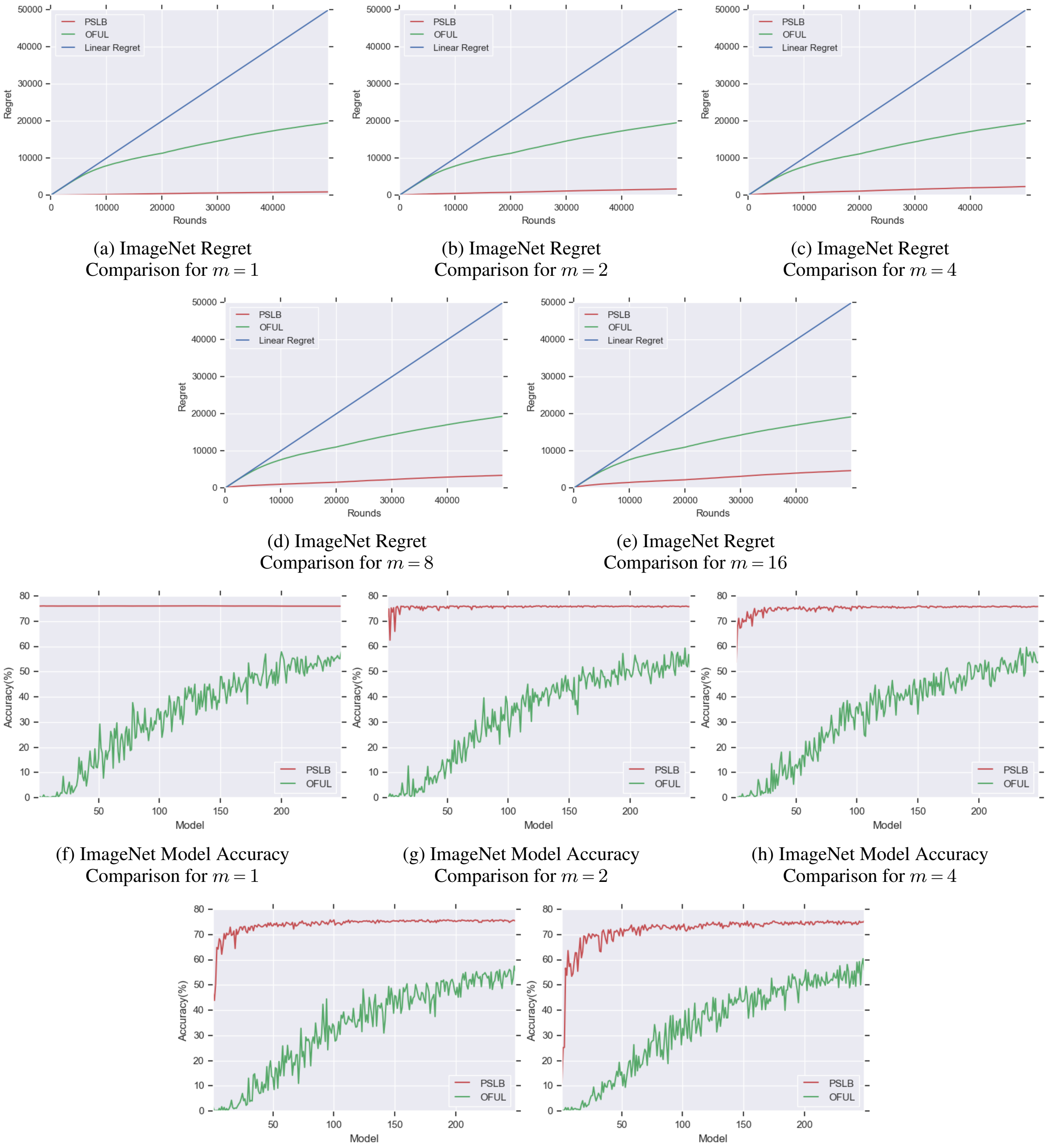
(h) CIFAR-10 Model Accuracy Comparison for $m = 4$

(i) CIFAR-10 Model Accuracy Comparison for $m = 8$

(j) CIFAR-10 Model Accuracy Comparison for $m = 16$

Figure 7: Regret and Optimistic Model Accuracy Comparisons of PSLB and OFUL on CIFAR-10 with $d = 1000$

Figure 7 provides the regret and the accuracy of optimistically chosen parameters of PSLB and OFUL in SLB constructed from CIFAR-10 with $d = 1000$.



(a) ImageNet Regret Comparison for $m = 1$

(b) ImageNet Regret Comparison for $m = 2$

(c) ImageNet Regret Comparison for $m = 4$

(d) ImageNet Regret Comparison for $m = 8$

(e) ImageNet Regret Comparison for $m = 16$

(f) ImageNet Model Accuracy Comparison for $m = 1$

(g) ImageNet Model Accuracy Comparison for $m = 2$

(h) ImageNet Model Accuracy Comparison for $m = 4$

(i) ImageNet Model Accuracy Comparison for $m = 8$

(j) ImageNet Model Accuracy Comparison for $m = 16$

Figure 8: Regret and Optimistic Model Accuracy Comparisons of PSLB and OFUL On ImageNet with $d = 100$

Figure 8 provides the regret and the accuracy of optimistically chosen parameters of PSLB and OFUL in SLB constructed from ImageNet with $d = 100$. Since there are 1000 different classes in the dataset, SLB framework synthesized from ImageNet dataset has 1000 actions in each decision set. Therefore, even if d=100 is not a fairly high dimensional feature space, having 1000 actions makes the learning task harder. Thus, SLB algorithms are expected to have higher regrets and slower convergence to underlying model. However, large number of actions is key to having lower regret for PSLB. Instead of ignoring actions that are not chosen at the current round, PSLB uses them to get an idea about the structure of the action vectors. This setting clearly points out the advantage of PSLB over OFUL. While OFUL obtains linear regret in the beginning and struggles to construct a meaningful confidence set, PSLB uses hidden information in the massive number of action vectors and reduce the dimensionality of the SLB framework. Then it exploits this information and converges to the accurate model without committing many mistakes.