Bandits with Partially Observable Confounded Data

Guy Tennenholtz[1] **Uri Shalit**[1] **Shie Mannor**[1,2] **Yonathan Efroni**[1]

[1]Technion, Israel Institute of Technology
[2]Nvidia Research

Abstract

We study linear contextual bandits with access to a large, confounded, offline dataset that was sampled from some fixed policy. We show that this problem is closely related to a variant of the bandit problem with side information. We construct a linear bandit algorithm that takes advantage of the projected information, and prove regret bounds. Our results demonstrate the ability to take advantage of confounded offline data. Particularly, we prove regret bounds that improve current bounds by a factor related to the visible dimensionality of the contexts in the data. Our results indicate that confounded offline data can significantly improve online learning algorithms. Finally, we demonstrate various characteristics of our approach through synthetic simulations.

1 INTRODUCTION

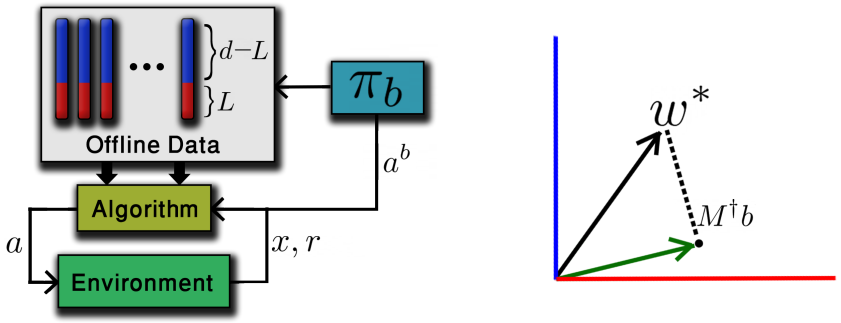
The use of offline data for online control is of practical interest in fields such as autonomous driving, healthcare, dialogue systems, and recommender systems [Mirchevska et al., 2017, Murphy et al., 2001, Li et al., 2016, Covington et al., 2016]. There, an abundant amount of data is readily available, potentially encompassing years of logged experience. This data can greatly reduce the need to interact with the real world, as such interactions may be both costly and unsafe [Amodei et al., 2016]. Nevertheless, as offline data is usually generated in an uncontrolled manner, it poses major challenges, such as unobserved states and actions. Failing to take these into account may result in biased estimates that are confounded by spurious correlation [Gottesman et al., 2019a]. This work focuses on utilizing partially observable offline data in an online bandit setting.

We consider the stochastic linear contextual bandit setting [Auer, 2002, Chu et al., 2011, Zhou et al., 2019]. Here, the context is a vector $x \in \mathbb{R}^d$ encompassing the full state of information. We assume to have additional access to an offline dataset in which only $L < d$ covariates (features) of the context are available. The unobserved covariates in the data are known as unobserved confounding factors in the causal inference literature [Pearl and Mackenzie, 2018], which may cause spurious associations in the data, rendering the data useless unless further assumptions are made [Neuberg, 2003, Shpitser and Pearl, 2012, Bareinboim et al., 2015]. In this work we assume that, when interacting with the online environment, the full context is accessible, and search for methods to combine both sources of information (online and offline) to quickly converge to an optimal solution.

We construct an algorithm that is provably superior to an algorithm which does not utilize the (partially observable) information in the data. We recognize the following fundamental observation: **Confounded offline data can (still) be used to improve online learning**, and specifically, that partially observable offline data can be utilized as linear side information (linear constraints) for the bandit problem.

While the bandit setting with confounded offline data has already been explored, its combination with a fully observable online environment is a new setting with particular challenges and benefits. First, one cannot ensure identification of an optimal policy with confounded offline data (see Section 3). This has implications on safety and applicability of algorithms which are based solely on offline data, e.g., the confounding bias of offline critical care datasets [Johnson et al., 2016]. Second, in contemporary widespread applications, an abundant of offline data is readily available. These application do not necessarily prevent interactions with the real world. On the contrary, countless real-world applications can access the real world. Still, such interactions may be costly, time consuming, or unsafe. It is thus vital to utilize the enormous amounts of previously collected offline data to reduce as much as possible the need for online interactions. We discuss two concrete examples from the healthcare and traffic management domains below.

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).



(a) (b)

Figure 1: **(a)** Block diagram of our setup: an online learner interacting with an environment while utilizing partially observable offline data that was generated by a behavior policy π_b. **(b)** This plot depicts the projection of $Mw^* = b$. We show that partially observable offline data can provide us with approximate linear side information of this form. The online learner must then estimate the orthogonal subspace, attempting to reduce the effective dimensionality of the problem.

Healthcare. Consider the important challenge of cancer chemotherapy control; specifically, optimal drug dosing for cancer chemotherapy [Sbeity and Younes, 2015]. Clinicians usually follow established guidelines for treating each patient, prescribing drug doses according to the stage of the tumor, the weight of the patient, white blood cell levels, concurrent illnesses, and the age of the patient. Suppose we are given access to large amounts of medical records of chemotherapy plans, specifying the frequency and dose of drug administration as well as their effect on the patient. Due to privacy regulations, the patients' socioeconomic characteristics are removed from the data. Nevertheless, these features may have affected the physician's decisions, as well as the outcome of the prescribed treatments. Next, suppose we are able to interact with the world, where the full state of the patients' information is available to us. How would we efficiently construct an algorithm to automate chemotherapy treatment while also utilizing the partially observable, confounded data?

Smart City Traffic Management. Consider the problem of adjusting traffic signals based on real-time traffic conditions using video footage of cameras located over intersections. The development time of the system consists of continual addition of new labels (classes) for the different types of vehicles and pedestrians based on relevant characteristics that may affect traffic congestion. Due to this recurrent process, data that was gathered in previous times may render itself useless, outdated, and even harmless, unless handled properly. This is due to the fact that some of the new information in the state was not previously collected, yet is needed for training future control strategies. How should one use the partially observable historical data for improving the most recent online system?

In this work we show how the confounded information in the data can be utilized for the online bandit problem. Figures 1 and 2 illustrate our basic setup and approach. We show how confounded offline data can be thought of as linear constraints to the online problem. These linear constraints, are not fully known. They are in fact dependent on the cross-correlation matrix of the context vector induced by policy that generated the data (which we denote as the behavior policy, π_b). To learn these constraints and utilize them, we approximate the cross-correlation matrix through online interactions and carefully integrate them into our learning algorithm, decreasing the overall regret.

The contributions of our work are as follows. As a fundamental contribution we propose a framework for combining confounded offline data with online learning. This framework is a gateway between fully confounded offline data to online learning, and encompasses a variety of important problems and applications. While this work only considers the linear bandit setting, it sets the building blocks and insights needed for more complex settings (e.g., reinforcement learning). Our second contribution shows that partially observable confounded data can in fact be realized as linear constraints for the online problem (see Section 3). To the best of our knowledge, this work is the first to show this relation. Finally, we prove that the overall regret can indeed be decreased when using the confounded data. Our proof, too, consists of technical obstacles related to the approximate constraints, which must be learned simultaneously.

2 PROBLEM SETTING

Notations. We use $[n]$ to denote the set $\{1, \ldots, n\}$. We denote by I_m the $m \times m$ identity matrix. Let $y, z \in \mathbb{R}^d$ and $A, B \in \mathbb{R}^{d \times d}$. We use $\|z\|_2$ to denote the ℓ_2-norm and z^T the transpose of z. The inner product is represented as $\langle z, y \rangle$. For A semi-positive definite, the weighted ℓ_2-norm is denoted by $\|z\|_A = \sqrt{z^T A z}$. The minimum and maximum singular values of A are denoted by $\lambda_{min}(A)$ and $\lambda_{max}(A)$ respectively. Furthermore, $A \preceq B$ if $B - A$ is

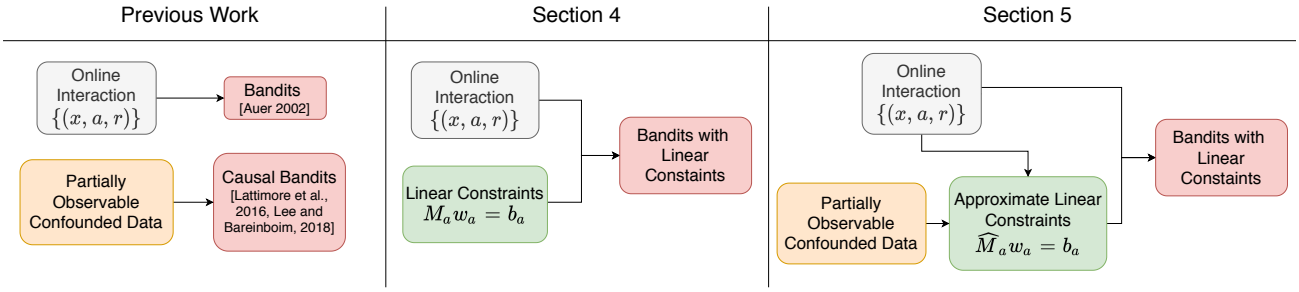


Figure 2: Previous work has dealt with bandits in the online setting. Other work integrated causal information to utilize confounded offline data. This work combines the two through constraints on the online problem. In Section 4 we show how linear constraints can be leveraged to achieve better regret for the bandit problem (Theorem 1). Then, in Section 5 partial linear constraints are estimated from online interactions, and then utilized efficiently by our learning algorithm. Note that b_a is not estimated as it is previously computed from the offline data (see Section 3). Finally, due to fast convergence of the linear constraints, improved performance is still achieved (Theorem 2).

positive semi-definite. The spectral norm of A is denoted by $\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$. The Moore-Penrose inverse of A is denoted by $A^\dagger$. Finally, we use $\mathcal{O}(x)$ to refer to a quantity that depends on x up to a poly-log expression in d, T and δ, and $\widetilde{\mathcal{O}}(x)$ represents the leading dependence of x in d, T and K.

Setup. Our basic framework consists of sequential interactions of a learner with an environment. We assume the following protocol, which proceeds in discrete trials $t = 1, \ldots, T$. At each round $t \in [T]$ the environment outputs a context $x_t \in \mathcal{X} \subseteq \mathbb{R}^d$ sampled from some unknown distribution $\mathcal{P}_x$. We assume that $x_1, \ldots x_T$ are i.i.d. Based on observed payoffs in previous trials, the learner chooses an action $a_t \in \mathcal{A}$, where $\mathcal{A} = [K]$ is the learner's action space. Subsequently, the learner observes a reward $r_t = \langle x_t, w^*_{a_t} \rangle + \eta_t$, where $\left\{w^*_a \in \mathbb{R}^d\right\}_{a \in \mathcal{A}}$ are unknown parameter vectors, and η_t is some conditionally σ-subgaussian random noise, i.e., for some $\sigma > 0$

$$\mathbb{E}\left[e^{\lambda \eta_t} \mid F_{t-1} \right] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right).$$

Here, $\{F_t\}_{t=0}^{\infty}$ is any filtration of σ-algebras such that for any $t \geq 1$, x_t is F_{t-1}-measurable and η_t is F_t-measurable, e.g., the natural σ-algebra $F_{t-1} = \sigma\left((x_1, a_1, \eta_1), \ldots, (x_{t-1}, a_{t-1}, \eta_{t-1}), x_t, a_t\right)$.

The goal of the learner is to maximize the total reward $\sum_{t=1}^{T} \langle x_t, w^*_{a_t} \rangle$ accumulated over the course of T rounds. We evaluate the learner against the optimal strategy, which has knowledge of $\left\{w^*_a \in \mathbb{R}^d\right\}_{a \in \mathcal{A}}$, namely $\pi^*(x) \in \arg\max_{a \in \mathcal{A}} \langle x, w^*_a \rangle$. The difference between the learner and optimal strategy's total reward is known as the regret, and is given by

$$\textit{Regret}\,(T) = \sum_{t=1}^{T} \left\langle x_t, w^*_{\pi^*(x_t)} \right\rangle - \sum_{t=1}^{T} \langle x_t, w^*_{a_t} \rangle.$$

In this work we assume to have additional access to a partially observable offline dataset, consisting of partially observable contexts, actions, and rewards. Specifically, we assume a dataset $\mathcal{D} = \{Q x_i, a_i, r_i\}_{i=1}^N$, in which $\{x_i\}_{i=1}^N$ are i.i.d. samples from $\mathcal{P}_x$, $\{a_i\}_{i=1}^N$ were generated by some fixed behavior policy, denoted by π_b, which is a mapping from contexts $x \in \mathcal{X}$ to a probability over actions, and $\{r_i\}_{i=1}^N$ were generated by the same model described above. Here, we used $Q \in \mathbb{R}^{L \times d}$ to denote the rectangular matrix $Q = \begin{pmatrix} I_L & 0 \end{pmatrix}$. That is, without loss of generality, we assume only the first L features of x_i are visible in the data. Throughout our work we will sometimes use the notation x^{o} and x^{h} to denote the observed and unobserved (hidden) covariates of x, respectively. That is, $x = \left((x^{\text{o}})^T, (x^{\text{h}})^T\right)^T$, where $x^{\text{o}} \in \mathbb{R}^L$, $x^{\text{h}} \in \mathbb{R}^{d-L}$.

Notice that the distribution of $\mathcal{D} = \{x^{\text{o}}_i, a_i, r_i\}_{i=1}^N$, the partially observable dataset, depends on π_b. Any statistic we attempt to draw from the offline data depends on the measure induced by π_b, which we denote by P^{π_b} [1]. Figure 1 depicts a diagram of our basic setup and approach.

3 FROM PARTIALLY OBSERVABLE OFFLINE DATA TO LINEAR SIDE INFORMATION

Consider only having access to the partially observable offline data $\mathcal{D}$. Having access to such data is mostly useless without further assumptions. Particularly, w^*_a may not be identifiable [2]. In fact, it can be shown that for any behavioral policy π_b and induced measure P^{π_b}, $\{w^*_a\}_{a \in \mathcal{A}}$ are not identifiable. More specifically, for all $w^1 = \left\{w^1_a\right\}_{a \in \mathcal{A}}$, exist $w^2 = \left\{w^2_a\right\}_{a \in \mathcal{A}} \neq w^2$ and probability measures P_1, P_2 such that $P_1\left(x^{\text{o}}, a, r; w^1, \pi_b\right) = P_2\left(x^{\text{o}}, a, r; w^2, \pi_b\right)$ and

[1] More precisely, we define the measure P^{π_b} for all Borel sets $R \subseteq [0,1]$, $X \subseteq \mathcal{X}$ and $A \in \mathcal{A}$ $P^{\pi_b}(r \in R, x \in X, a \in A) = P(r \in R | x \in X, a \in A) P(x \in X) \int_{x' \in X, a' \in A} \mathbb{1}_{\{a=a', x=x'\}} d\pi_b$.

[2] We use the notion of identifiability as defined in Definition 2 of Pearl et al. [2009]

$\pi_b\left(a, x; w^1\right) = \pi_b\left(a, x; w^2\right)$. This claim is a standard type of result. A proof is provided in the supplementary material.

To mitigate the identification problem, prior knowledge of characteristics of $\{w_a^*\}_{a\in\mathcal{A}}$ can be leveraged [Cinelli et al., 2019]. Instead, here we consider access to an online environment, where the covariates that were unobserved in the data are supplied, i.e., fully observed. This enables us to deconfound the data and identify $\{w_a^*\}_{a\in\mathcal{A}}$.

Prior to constructing our algorithmic approach, we discuss the relation of confounded offline data to partially known linear constraints. This connection is a principal component of our work which enables us to utilize the (possibly not identifiable) partially observable data.

3.1 LINEAR SIDE INFORMATION

In what follows, we show how partially observable data can be reduced to linear constraints of the form $\{M_a w_a^* = b_a, a \in \mathcal{A}\}$. Nevertheless M_a will not be identifiable solely from the offline data. More specifically, we specify a low dimensional least squares problem under a model mismatch, showing it converges to a solution with unique structural properties. This will become beneficial in our analysis later on, allowing us to project the linear bandit problem to an approximate lower dimensional subspace, improving performance guarantees.

Let us first consider the case of fully-observable offline data, i.e., $x^{\mathrm{o}} = x$. Here, one would be able (with large amounts of data) to closely estimate w_a^* for all $a \in \mathcal{A}$, using, for example, the linear regression estimator

$$\hat{w}_a = \left(\frac{1}{N_a}\sum_{i=1}^{N_a} x_i x_i^T\right)^{-1}\left(\frac{1}{N_a}\sum_{i=1}^{N_a} x_i r_i\right),$$

where we denoted $N_a = \sum_{i=1}^{N} \mathbb{1}_{\{a_i=a\}}$. With $N \to \infty$, under mild assumptions, this estimator would converge to the true weights w_a^* almost surely. It is tempting to try and apply a least square estimator to our partially observable data using a lower dimensional model. Particularly, we might try to solve the optimization problem

$$\min_{b\in\mathbb{R}^L} \sum_{i=1}^{N_a} \left(\langle x_i^{\mathrm{o}}, b\rangle - r_i\right)^2 \quad , \forall a \in \mathcal{A},$$

ignoring the fact that $r_i = x_i^T w_a^* + \eta_i$, i.e., that r_i was generated by a higher dimensional linear model. Solving this problem yields

$$b_a^{LS} = \left(\frac{1}{N_a}\sum_{i=1}^{N_a} \left(x_i^{\mathrm{o}}\right)\left(x_i^{\mathrm{o}}\right)^T\right)^{-1}\left(\frac{1}{N_a}\sum_{i=1}^{N_a} x^{\mathrm{o}} r_i\right). \quad (1)$$

The following proposition establishes our first main result – a relation between the lower-dimension least-square estimator b_a^{LS} and the vector w_a^* in the limit of large data $N \to \infty$ (We discuss the finite data setting in Section 8).

Proposition 1. *[Confoundness = Linear Constraints] Let* $R_{11}(a) = \mathbb{E}^{\pi_b}\left[x^{\mathrm{o}}\left(x^{\mathrm{o}}\right)^T \,\middle|\, a\right]$, $R_{12}(a) = \mathbb{E}^{\pi_b}\left[x^{\mathrm{o}}\left(x^{\mathrm{h}}\right)^T \,\middle|\, a\right]$. *Assume* $R_{11}(a)$ *is invertible for all* $a \in \mathcal{A}$ [3]. *Then, the following holds almost surely for all* $a \in \mathcal{A}$.

$$\lim_{N\to\infty} b_a^{LS} = \Big(I_L, \quad R_{11}^{-1}(a)R_{12}(a)\Big) w_a^*.$$

The proof of the proposition is related to regression analysis with misspecified models (see e.g., Griliches [1957]) and is provided in the supplementary material. It states that, with an infinite amount of data, the low-dimensional least squares estimator in Equation (1) converges to a linear transformation of w_a^*. This linear transformation depends on the auto-correlation matrix of x^{o}, $R_{11}(a)$, and the cross correlation matrix of x^{o} and x^{h}, $R_{12}(a)$. While $R_{11}(a)$ can be estimated from the data, $R_{12}(a)$ depends on unseen features of x, namely x^{h}, as well as the behavior policy π_b, and can thus not be approximated from the given data. As such, we will later assume access to a monotonically non-increasing bound of $R_{12}(a)$ for all $a \in \mathcal{A}$. As we discuss in Section 5, such a bound can be achieved, for example, through queries to π_b (i.e., samples $a \sim \pi_b$).

Proposition 1 provides us with a structural dependency between w_a^* and the low-order least squares estimator b_a^{LS} that can be calculated from the offline data. Specifically, every w_a^* is constrained to a set $\left\{w \in \mathbb{R}^d : Mw = b\right\}$, for some full row rank matrix $M \in \mathbb{R}^{L\times d}$ and vector $b \in \mathbb{R}^L$. A natural question arises: How can such linear side information be used? In the next section we show that we can decrease the effective dimensionality of our problem using such linear side information whenever M and b are known exactly. Then, in Section 5, we expand this result using estimates of the linear relation in Proposition 1. We provide improved regret bounds on the linear contextual bandit problem, consequently exploiting the confounded information present in the partially observable data.

4 LINEAR CONTEXTUAL BANDITS WITH LINEAR SIDE INFORMATION

In the previous section we showed how partially observable data can be reduced to linear constraints. Before diving into the subtleties of utilizing the specific structural properties of the linear relations in Proposition 1, we form a general

[3] The invertibility assumption on R_{11} can be verified, since R_{11} can be estimated by statistics of the observable covariates, x^{o}. If it does not hold, other covariates of x^{o} can be chosen to satisfy this assumption.

result for linear bandits under linear side information when both M and b are given. Particularly, we show that linear side information can be used to improve performance by decreasing the effective dimensionality of the underlying problem.

Assume we are given linear side information

$$M_a w_a^* = b_a \qquad , a \in \mathcal{A}. \tag{2}$$

In this section we assume $M_a \in \mathbb{R}^{L\times d}, b_a \in \mathbb{R}^L$ are *known*, and don't assume any structural characteristics. Without loss of generality assume that $\{M_a\}_{a\in\mathcal{A}}$ are full row rank [4]. One way of using the relations in Equation (2) is by constraining an online learning algorithm to a lower dimensional space. Particularly, notice that for all $a \in \mathcal{A}$,

$$w_a^* \in \left\{w \in \mathbb{R}^d : w = M_a^\dagger b_a + P_a w\right\}, \tag{3}$$

where P_a is the orthogonal projection onto the kernel of M_a, and is given by $P_a = I - M_a^\dagger M_a$. Equation (3) suggests that knowledge of the linear relation in Equation (2) may allow us to reduce the estimation problem to that of the projected vector, $P_a w_a^*$. Indeed, we may attempt to solve the following corrected, low order ridge regression problem

$$\min_{w\in\mathbb{R}^d}\left\{\sum_{i=1}^{t-1}\left(\langle x_i, P_a w\rangle - y_{a,i}\right)^2 + \lambda\left\|P_a w\right\|_2^2\right\}, \tag{4}$$

where $y_{a,i} = r_i - \left\langle x_i, M_{a_i}^\dagger b_{a_i}\right\rangle$. Taking its smallest norm solution yields

$$\hat{w}_{t,a}^{P_a} = \left(P_a\left(\lambda I + \sum_{i=1}^{t-1} x_i x_i^T\right)P_a\right)^\dagger \times \left(\sum_{i=1}^{t-1} r_i x_i - \sum_{i=1}^{t-1} x_i x_i^T M_a^\dagger b_a\right). \tag{5}$$

Perhaps intuitively, this least squares estimator is in fact equivalent to one in a lower dimensional space $\mathbb{R}^m$, the rank of P_a. Indeed, letting $P_a = UU^T$, where $U \in \mathbb{R}^{d\times m}$ is a matrix with orthonormal columns[5], we have that (see supplementary material for full derivation)

$$U^T\hat{w}_{t,a}^{P_a} = \left(\lambda I_m + \sum_{i=1}^{t-1}\left(U^T x_i\right)\left(U^T x_i\right)^T\right)^{-1} \times \left(\sum_{i=1}^{t-1} y_{a,i}\left(U^T x_i\right)\right).$$

[4] If M_a is not full row rank, we remove dependent rows. In fact, we assume L to be the rank of M_a.

[5] As orthogonal projection matrices have eigenvalues which are either 0 or 1, any projection matrix can be decomposed into $P = UU^T$, where U is a matrix with $\mathrm{rank}(P)$ orthonormal columns.

Algorithm 1 OFUL with Linear Side Information

1: **input:** $\alpha > 0, M_a \in \mathbb{R}^{L\times d}, b_a \in \mathbb{R}^L, \delta > 0$
2: **init:** $V_a = \lambda I_d, Y_a = 0, \forall a \in \mathcal{A}$
3: **for** $t = 1, \ldots$ **do**
4: Receive context x_t
5: $\hat{w}_{t,a}^{P_a} = \left(P_a V_a P_a\right)^\dagger\left(Y_a - \left(V_a - \lambda I_d\right)M_a^\dagger b_a\right)$
6: $\hat{y}_{t,a} = \left\langle x_t, M_a^\dagger b_a\right\rangle + \left\langle x_t, \hat{w}_{t,a}^{P_a}\right\rangle$
7: $\mathrm{UCB}_{t,a} = \sqrt{\beta_t(\delta)}\left\|x_t\right\|_{(P_a V_a P_a)^\dagger}$
8: $a_t \in \arg\max_{a\in\mathcal{A}}\{\hat{y}_{t,a} + \alpha\mathrm{UCB}_{t,a}\}$
9: Play action a_t and receive reward r_t
10: $V_{a_t} = V_{a_t} + x_t x_t^T, Y_{a_t} = Y_{a_t} + x_t r_t$
11: **end for**

That is, $U^T\hat{w}_{t,a}^{P_a}$ is a least squares estimator in $\mathbb{R}^m$.

We are now ready to construct a least squares variant for w_a^*, which utilizes the information in Equation (2). Having an estimation for $P_a w_a^*$, we make use of the set defined in Equation (3) to construct our final estimator $\hat{w}_{a,t} = M_a^\dagger b_a + \hat{w}_{t,a}^{P_a}$, where $\hat{w}_{t,a}^{P_a}$ is given by Equation (5). Then, estimation of $\hat{w}_{a,t}$ will depend on the rank of P_a, i.e., $\mathrm{rank}(P_a) = d - L$. In what follows we will show how this projected estimator can be integrated into a linear bandit algorithm, reducing its effective dimensionality to that of the rank of P_a, i.e., $d - L$.

Algorithm 1 describes the reduction of the OFUL algorithm [Abbasi-Yadkori et al., 2011] to its projected variant, in which linear side information is leveraged by means of low order ridge regression (Equations (4)) to decrease the effective dimensionality of the problem. In Line 5 of the algorithm, the estimator of Equation (5) for $P_a w_a^*$ is used. This becomes useful in Line 7, as the confidence set around w_a^* is reduced to a lower dimension, i.e., $d - L$.

For all $a \in \mathcal{A}$, assume $\left\|P_a x_i\right\|_2 \le S_{x,o}$ almost surely and $\left\|P_a w_a^*\right\|_2 \le S_{w,o}$. Letting $\sqrt{\beta_t(\delta)} = \lambda^{1/2}S_{w,o} + \sigma\sqrt{(d-L)\log\left(\frac{K(1+tS_{x,o}^2/\lambda)}{\delta}\right)}$, the following theorem provides the improved regret of Algorithm 1. Its proof is given in the supplementary material, and is based on a reduction of the linear bandit problem to a lower dimensional space, based on Equation (5).

Theorem 1. *For all $T \ge 0$, with probability at least $1-\delta$, the regret of Algorithm 1 is bounded by*

$$RegretT \le \widetilde{\mathcal{O}}\left((d-L)\sqrt{KT}\right).$$

Indeed, by Theorem 1, linear relations of rank L reduce the linear bandit problem to a lower dimensional problem, with regret guarantees that are equivalent to those of a linear bandit problem of dimension $d - L$. However, these results hold

Algorithm 2 OFUL with Partially Observable Offline Data

1: **input:** $\alpha > 0, \delta > 0, T, b_a \in \mathbb{R}^L$ (from dataset)
2: **for** $n = 0, \ldots, \log_2(T) - 1$ **do**
3: Use 2^n previous samples from π_b to update the estimate of $\hat{M}_{2^n,a}, \forall a \in \mathcal{A}$
4: Calculate $\hat{M}^{\dagger}_{2^n,a}, \hat{P}_{2^n,a}, \forall a \in \mathcal{A}$
5: Run Algorithm 1 for 2^n time steps with bonus $\sqrt{\beta_{n,t}(\delta)}$ and $\hat{M}_{2^n,a}, b_a$
6: **end for**

only for M_a, b_a that are fully known. When $\{M_a\}_{a\in\mathcal{A}}$ are unknown, we must rely on estimations of M_a. The accuracy of our estimation as well as its rate of convergence would highly affect the applicability of such constraints. As we will see next, the linear transformation of Proposition 1 can be efficiently estimated whenever R_{12} can be efficiently estimated. Such an assumption will allow us to achieve similar regret guarantees under mild conditions.

5 DECONFOUNDING PARTIALLY OBSERVABLE DATA

This section builds upon the observations collected in the previous sections in order to construct our second main result: an algorithm that leverages large, partially observable, offline data in the online linear bandit setting. While Proposition 1 seemingly provides us with linear side information in the form of linear equalities $M_a w^* = b_a$, the matrix M_a cannot be obtained from the partially observable offline data, since $R_{12}(a)$ depends on the unobserved covariates x^{h}, as well as the behavior policy π_b. Nevertheless $M_a = \begin{pmatrix} I_L, & R_{11}^{-1}(a)R_{12}(a) \end{pmatrix}$ can be efficiently estimated whenever $R_{12}(a) = \mathbb{E}^{\pi_b}\left[x^{\mathrm{o}} \left(x^{\mathrm{h}}\right)^T \,\middle|\, a \right]$ can be efficiently estimated. Particularly we make the following assumption.

Assumption 1. *We assume for every $t > 0$ we can approximate $R_{12}(a), \forall a \in .\mathcal{A}$ such that*

$$\left\| R_{12}(a) - \hat{R}_{12}(a,t) \right\|_2 \leq \frac{g(d,L)}{\sqrt{t}} \quad \textit{w.h.p.}$$

5.1 CASE STUDY: QUERIES TO π_b

Consider the problem of identifying the statistic $R_{12}(a)$. Due to its dependence on π_b, this may be impossible without access to π_b or other information on its induced measure, P^{π_b}. As such, we assume that during online interactions, the online learner can query π_b, i.e., sample an action $a^b \sim \pi_b(x)$.

Having access to queries from π_b, we can construct an online estimator for the cross-correlation matrix $R_{12}(a)$. More specifically, at each round $t \in [T]$, we observe a context x_t and query π_b by sampling $a_t^b \sim \pi_b(x_t)$. We then estimate $R_{12}(a)$ using the empirical estimator[6]

$$\hat{R}_{12}(a,t) = \frac{1}{t}\sum_{i=1}^{t} \frac{\mathbb{1}_{\{a_i = a\}}}{P^{\pi_b}(a)} \left(x_i^{\mathrm{o}}\right)\left(x_i^{\mathrm{h}}\right)^T,$$

where $P^{\pi_b}(a)$ is known due to the offline data. Assuming $\|x^{\mathrm{o}}\|_2 \leq S_1$ and $\left\|x^{\mathrm{h}}\right\|_2 \leq S_2$ a.s., it can be shown that with probability at least $1-\delta$ (see supplementary material, Lemma 8, for proof)

$$\left\| R_{12}(a) - \hat{R}_{12}(a,t) \right\|_2 \leq$$
$$\mathcal{O}\left(S_1 S_2 \sqrt{\frac{1}{t}\left(\frac{\sqrt{\mathrm{trace}\left(R_{11}\right)\mathrm{trace}\left(R_{22}\right)}}{S_1 S_2} \right) \log\left(\frac{d}{\delta}\right)} \right),$$

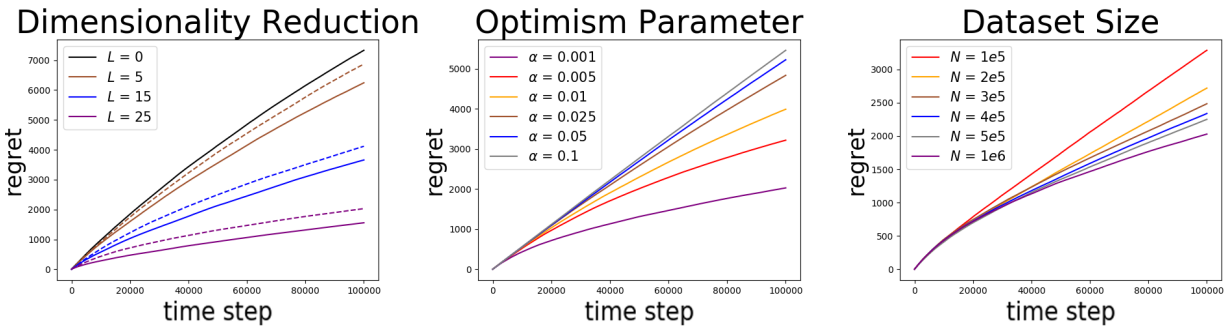
indeed, satisfying Assumption 1. We can now naturally construct an estimator for M_a. Its estimator is given by

$$\hat{M}_{t,a} = \left(I_L, \quad R_{11}^{-1}(a)\hat{R}_{12}(a,t) \right). \tag{6}$$

A natural question arises: can the estimated linear constraints $\hat{M}_a w_a^* = b_a$ be used as linear side information while still maintaining the regret guarantees of Theorem 1, i.e., decrease the effective dimensionality of the problems from d to $d-L$? Specifically, we wish to construct a variant of Algorithm 1 in which $\hat{M}_{t,a}$ are used as linear side information. In this setting the estimated projection matrix $\hat{P}_{t,a}$ and the estimated Moore-Pensore Inverse $\hat{M}^{\dagger}_{t,a}$ are directly calculated from $\hat{M}_{t,a}$, i.e., these matrices are approximate.

Algorithm 2 describes the linear bandit variant with partially observable confounded data. Note that, unlike Algorithm 1, Algorithm 2 is not an anytime algorithm, but rather acts knowing the horizon T. Assuming $\|x_i\|_2 \leq S_x$ a.s. and $\|w_a^*\|_2 \leq S_w$ for all $a \in \mathcal{A}$, the algorithm uses an augmented confidence, given by $\sqrt{\beta_{n,t}(\delta)} = \lambda^{1/2} S_w + (\sigma + S_x S_w f_n)\sqrt{(d-L)\log\left(\frac{1+tS_x^2/\lambda}{\delta/2\log(T)K}\right)}$, where $f_n = f_{B1} + f_{B2}2^{-n/2}$, $f_{B1} = \widetilde{\mathcal{O}}\Big(\max_a \frac{\lambda_{\min}(R_{11}(a))^{-1}}{P^{\pi_b}(a)} S_x (\mathrm{trace}\left(R_{11}(a)\right)\mathrm{trace}\left(R_{22}(a)\right))^{1/4}\Big)$ and $f_{B2} = \widetilde{\mathcal{O}}\Big(\max_a \frac{\lambda_{\min}(R_{11}(a))^{-1}}{P^{\pi_b}(a)} S_x^2\Big)$. At every time step $t \in [T]$, the learner uses the estimate $\hat{M}_{t,a}$ and subsequently considers it to be linear side information, as in Algorithm 1. The following theorem provides regret guarantees for Algorithm 2, proving partially observable data can be beneficial for online learning.

[6] In fact, we can construct a tighter estimator for $R_{12}(a)$ using our knowledge of $\mathbb{E}^{\pi_b}[\, x^{\mathrm{o}} \mid a\,]$, which can be estimated exactly from the offline data. We leave its analysis out for clarity.



(a) (b) (c)

Figure 3: All experiments were conducted with the same vectors w_a^* of dimension $d = 30$ and $K = 30$ arms. **(a)** Plot compares effect of L when R_{12} is known (solid lines) vs. estimated (dashed lines). For $L = 0$ (i.e., no side information) we executed Algorithm 1 without using the dataset. **(b)** Comparison of different values of α using an offilne dataset and $L = 25$. **(c)** Effect of dataset size on performance for $L = 25$.

Theorem 2. *For any $T > 0$, with probability at least $1 - \delta$, the regret of Algorithm 2 with the estimator given in Equation* (6) *is bounded by*

$$\textit{Regret}(T) \leq 3\sqrt{T(d-L)K\log\left(\lambda + \frac{TS_x^2}{d-L}\right)} \times \left(\sigma_\epsilon\sqrt{(d-L)\log\left(\frac{1+TS_x^2/\lambda}{\delta/(2K\log(T))}\right)} + \lambda^{1/2}S_w\right) + \mathcal{O}\left((d-L)\sqrt{K}S_xS_wf_{B2}\right),$$

where $\epsilon = S_xS_wf_{B1}$ and $\sigma_\epsilon = \sigma + \epsilon$. This leads to, Regret $(T) \leq \widetilde{\mathcal{O}}\left((1+f_{B_1})(d-L)\sqrt{KT}\right)$.

Notice that, unlike Theorem 1, the regret of Algorithm 2 is worsened asymptotically by a factor relating to f_{B_1}. This function can also scale with d, due to its dependence on $\left(\text{trace}\left(R_{11}(a)\right)\right)^{1/4}$ and $\left(\text{trace}\left(R_{22}(a)\right)\right)^{1/4}$. Specifically, a worst case dependence yields $f_{B_1} \leq \widetilde{\mathcal{O}}\left(\max_a \frac{(L(d-L))^{1/4}}{P^{\pi_b}(a)}\right)$, where here $\max_a \frac{1}{P^{\pi_b}(a)} \geq K$. That is, f_{B_1} is a factor indicating how hard it is to approximate the linear constraints, dependent on the amount of information in x as well as the support of the behavior policy, π_b. Still, in settings in which d and T are prominent over K, a significant improvement in performance is achieved.

The proof of the theorem is provided in the supplementary material. Unlike in Theorem 1, we do not have access to the true matrices $M_a^\dagger, P_a$, but to increasingly more accurate estimates of these matrices. To deal with this more challenging situation we use the doubling trick. The algorithm acts in exponentially increasing episodes. In each such episode, we fix the estimation of M_a, i.e., we use the estimate of M_a available at the beginning of the episode. The analysis of this algorithm amounts to study the performance of the exact algorithm (as in Theorem 1) up to a fixed, approximated, M_a, which induces errors in the used $M_a^\dagger, P_a$. Finally, summing the regret on each episode, we obtain the result.

The proof heavily relies on the convergence properties of $P_a, M^\dagger$, which are shown to converge at a rate of $O(T^{-1/2})$. These convergence rates are due to the special structure of M_a. Specifically, we prove that $\left\|P_a - \hat{P}_{t,a}\right\| \leq 2\left\|M_a - \hat{M}_{t,a}\right\|$ and $\left\|M_a^\dagger - \hat{M}_{t,a}^\dagger\right\| \leq 2\left\|M_a - \hat{M}_{t,a}\right\|$, meaning, the convergence of $\hat{P}_{t,a}$ and $\hat{M}_{t,a}^\dagger$ is well controlled by the convergence of $\hat{M}_{t,a}$. This property does not hold for general matrices. In fact, for a general matrix A, $A^\dagger$ is not even continuous w.r.t. perturbations in A (see e.g., Stewart 1969). Thus, the structure of M_a establishes convergence rates of $\hat{P}_{t,a}, \hat{M}_{t,a}^\dagger$ sufficient to achieve the desired regret.

Algorithm 2 is highly wasteful w.r.t. the information gathered through time. Specifically, it discards all information upon updates of $\hat{M}_{t,a}$. In a practical setting, we expect the algorithm to achieve similar performance guarantees even when information is not discarded. Moreover, as we show empirically in the next section, significant improvement can still be achieved without applying the doubling trick, i.e., by running Algorithm, 1 with the approximated $\hat{M}_{t,a}$.

6 EXPERIMENTS

In this section we demonstrate the effectiveness of using offline data in a synthetic environment. Our environment consisted of $K = 30$ arms and vectors $w_a^* \in \mathbb{R}^{30}$ uniformly sampled in $\left[0, \frac{1}{d}\right]^d$ and fixed across all experiments. Contexts were sampled from a uniform distribution in $[0,1]^d$ and normalized to have norm 1. The behavioral policy π_b was chosen to follow a softmax distribution $\pi_b(a,x) \propto \exp\left(\phi_a^T x\right)$, where $\phi_a \in \mathbb{R}^d$ were randomly chosen and fixed across all experiments.

Figure 3a illustrates the effectiveness of using partially observable data. We used a dataset of 1 million examples to simulate a sufficiently large dataset. Solid lines depict regret when $R_{12}(a)$ are known in advance, allowing us to apply Algorithm 1 without estimations (Section 4). Dashed lines depict regret for the estimated case using queries to π_b, i.e., M_a were estimated at every iteration using an estimate of $R_{12}(a)$ (see Section 5). Note that $L = 0$ corresponds to the linear bandit problem with no side information, i.e., the original OFUL algorithm. It is evident that utilizing the partially observable data can significantly improve performance, even when using approximate projections. We note that the experiments were run under constant updates of $\hat{M}_{t,a}$, i.e., without epoch schedules.

Figure 3b depicts the effect of the optimism parameter α (see Algorithm 1) on overall performance when utilizing a dataset with $L = 25$ observed features. A gap is evident between the proposed theoretical confidence and the practical results, as very small values of α showed best performance. This gap is most likely due to worst case scenarios that were not imposed by our simulated environments.

Finally, Figure 3c depicts experiments with varying amount of data. While the number of examples has an effect, it does not significantly deteriorate overall performance, suggesting that partially observable offline data can be used even with finite datasets, as long as they are sufficiently large.

7 RELATED WORK

The linear bandits problem, first introduced by Auer [2002], has been extensively investigated in the pure online setting [Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010, Abbasi-Yadkori et al., 2011], with numerous variants and extensions [Agrawal and Devanur, 2016, Kazerouni et al., 2017, Amani et al., 2019].

The offline (logged) bandit setting usually assumes the algorithm must learn a policy from a batch of fully observable data [Shivaswamy and Joachims, 2012, Swaminathan and Joachims, 2015, Joachims et al., 2018]. The use of offline data has also been investigated under the reinforcement learning framework, including batch-mode off-policy reinforcement learning and off-policy evaluation [Ernst et al., 2005, Lizotte et al., 2012, Fonteneau et al., 2013, Precup, 2000, Thomas and Brunskill, 2016, Gottesman et al., 2019b]

More related to our work are attempts to establish unbiased estimates or control schemes from confounded offline data [Lattimore et al., 2016, Oberst and Sontag, 2019, Tennenholtz et al., 2020]. Other work in which partially observable data is used usually consider the standard confounded setting (e.g., identification of $P(r|\text{do}(a))$) [Zhang and Bareinboim, 2019, Ye et al., 2020]. Wang et al. [2016] also consider hidden features, where biases are accounted for under assumptions on the hidden features. In these works the unobserved features (confounders) are never disclosed to the learner. Prior knowledge is thereby usually assumed over their support (e.g., known bounds). When such priors are unknown, these methods may thus fail. Moreover, they are sub-optimal in settings of fully observable interactions, where unobserved confounders become observed covariates.

In this work we view the problem from an online learner's perspective, where *offline data is used as side information*. Specifically we project the given information, reducing our problem to its orthogonal subspace. Projections have been previously used in the bandit setting for reducing time complexity and dimensionality [Yu et al., 2017]. Other work consider bandits under constraints [Agrawal and Devanur, 2019]. Finally, Djolonga et al. [2013] consider subspace-learning by combining Gaussian Process UCB sampling and low-rank matrix recovery techniques.

8 DISCUSSION AND FUTURE WORK

In this work we showed that partially observable confounded data can be efficiently utilized in the linear bandit setting. In this section we further discuss two central assumptions made in our work; namely, infinite data and bounding the cross correlation matrix R_{12}.

Finite Data. Throughout our work we assumed the limit of infinite sized data. From a technical perspective, the use of finite data would introduce an error in the least squares estimator [Krikheli and Leshem, 2018]. A straightforward analysis would propagate this error as additional linear penalty to the regret that is dependent on the number of samples in the data. More involved techniques may combine optimistic bounds on the finite samples in the data. We chose to leave its derivation out to focus on the topic of missing covariates in the data. Finally, our experiments demonstrate that the number of samples does not greatly affect performance, as long as they are sufficiently large, i.e., when the error is small relative to T.

Bounding R_{12}. Being able to estimate $R_{12}(a)$ is an essential requirement for deconfounding the partially observable data. Nevertheless, $R_{12}(a)$ is dependent on π_b, raising the question, can $R_{12}(a)$ be estimated without knowledge of π_b? In our work we showed how one can estimate it using queries to π_b. In fact, we did not require knowledge of π_b, nor did we require interactions of π_b with the environment (i.e., we do not act according to π_b), but rather, only view samples from π_b. While such an assumption may be strict in some settings, it is reasonable in others. For instance, when π_b was controlled by us when the data was recorded. Other settings for estimating $R_{12}(a)$ are also possible, e.g., having access to additional fully observable datasets that were generated by π_b [Kallus et al., 2018].

Consider the examples of the healthcare and traffic management settings presented in Section 1. In the medical setting,

quering π_b would amount to asking the clinician that induced the data what she would have done in a provided situation. In this scenario, cooperation of the clinician is needed to deconfound the data. Nevertheless, note that this approach is not limited by the amount of confounding bias inherent in the data, allowing us identify *optimal* control policies. Unlike the medical example, in the traffic management example we have access to the behavior policy that generated the data. In this scenario, the querying assumption is insignificant.

Future Work. While this work assumed a monotonically vanishing error of $\hat{R}_{12}(a)$ (i.e., asymptotic identifiability), future work can consider looser bounds on the estimate. It is also interesting to understand the contextual bandit algorithms, both in the linear as well as the general function class settings. It is also interesting to generalize our results to the reinforcement learning setting.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. In *Advances in Neural Information Processing Systems*, pages 3450–3458, 2016.

Shipra Agrawal and Nikhil R Devanur. Bandits with global convex constraints and objective. *Operations Research*, 67(5):1486–1502, 2019.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems*, pages 9252–9262, 2019.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

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

João Carlos Alves Barata and Mahir Saleh Hussein. The moore–penrose pseudoinverse: A tutorial review of the theory. *Brazilian Journal of Physics*, 42(1-2):146–165, 2012.

Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, pages 1342–1350, 2015.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Yan Mei Chen, Xiao Shan Chen, and Wen Li. On perturbation bounds for orthogonal projections. *Numerical Algorithms*, 73(2):433–444, 2016.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

Carlos Cinelli, Daniel Kumor, Bryant Chen, Judea Pearl, and Elias Bareinboim. Sensitivity analysis of linear structural causal models. In *International Conference on Machine Learning*, pages 1252–1261, 2019.

Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

Josip Djolonga, Andreas Krause, and Volkan Cevher. High-dimensional gaussian process bandits. In *Advances in Neural Information Processing Systems*, pages 1025–1033, 2013.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(Apr):503–556, 2005.

Raphael Fonteneau, Susan A Murphy, Louis Wehenkel, and Damien Ernst. Batch mode reinforcement learning based on the synthesis of artificial trajectories. *Annals of operations research*, 208(1):383–416, 2013.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nat Med*, 25(1):16–18, 2019a.

Omer Gottesman, Yao Liu, Scott Sussex, Emma Brunskill, and Finale Doshi-Velez. Combining parametric and nonparametric models for off-policy evaluation. *arXiv preprint arXiv:1905.05787*, 2019b.

Zvi Griliches. Specification bias in estimates of production functions. *Journal of farm economics*, 39(1):8–20, 1957.

Thorsten Joachims, Adith Swaminathan, and Maarten de Rijke. Deep learning with logged bandit feedback. 2018.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems*, pages 10888–10897, 2018.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems*, pages 3910–3919, 2017.

Michael Krikheli and Amir Leshem. Finite sample performance of linear least squares estimators under subgaussian martingale difference noise. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4444–4448. IEEE, 2018.

Sahin Lale, Kamyar Azizzadenesheli, Anima Anandkumar, and Babak Hassibi. Stochastic linear bandits with hidden low rank structure. *arXiv preprint arXiv:1901.09490*, 2019.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems*, pages 1181–1189, 2016.

Jiwei Li, Alexander H Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. Learning through dialogue interactions by asking questions. *arXiv preprint arXiv:1612.04936*, 2016.

Daniel J Lizotte, Michael Bowling, and Susan A Murphy. Linear fitted-q iteration with multiple reward functions. *Journal of Machine Learning Research*, 13(Nov):3253–3295, 2012.

Branka Mirchevska, Manuel Blum, Lawrence Louis, Joschka Boedecker, and Moritz Werling. Reinforcement learning for autonomous maneuvering in highway scenarios. In *Workshop for Driving Assistance Systems and Autonomous Driving*, pages 32–41, 2017.

Susan A Murphy, Mark J van der Laan, James M Robins, and Conduct Problems Prevention Research Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

Leland Gerson Neuberg. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory*, 19(4):675–685, 2003.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. *arXiv preprint arXiv:1905.05824*, 2019.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

M Planitz. 3. inconsistent systems of linear equations. *The Mathematical Gazette*, 63(425):181–185, 1979.

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

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

Hoda Sbeity and Rafic Younes. Review of optimization methods for cancer chemotherapy treatment planning. *Journal of Computer Science & Systems Biology*, 8(2):74, 2015.

Pannagadatta Shivaswamy and Thorsten Joachims. Multi-armed bandit problems with history. In *Artificial Intelligence and Statistics*, pages 1046–1054, 2012.

Ilya Shpitser and Judea Pearl. What counterfactuals can be tested. *arXiv preprint arXiv:1206.5294*, 2012.

Gilbert W Stewart. On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM review*, 19(4):634–662, 1977.

GW Stewart. On the continuity of the generalized inverse. *SIAM Journal on Applied Mathematics*, 17(1):33–45, 1969.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755, 2015.

Guy Tennenholtz, Shie Mannor, and Uri Shalit. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

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

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Huazheng Wang, Qingyun Wu, and Hongning Wang. Learning hidden features for contextual bandits. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1633–1642, 2016.

Per-Åke Wedin. Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, 13(2):217–232, 1973.

Li Ye, Yishi Lin, Hong Xie, and John Lui. Combining offline causal inference and online bandit learning for data driven decisions. *arXiv preprint arXiv:2001.05699*, 2020.

Xiaotian Yu, Michael R Lyu, and Irwin King. Cbrap: Contextual bandits with random projection. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In *Advances in Neural Information Processing Systems*, pages 13401–13411, 2019.

Zhengyuan Zhou, Renyuan Xu, and Jose Blanchet. Learning in generalized linear contextual bandits with stochastic delays. In *Advances in Neural Information Processing Systems*, pages 5198–5209, 2019.

APPENDICES

Appendix A Overview **13**

Appendix B Orthogonal Projections and the Moore–Penrose Inverse **14**

B.1 Useful Perturbation Bounds 14

Appendix C Unidentifiability Problem of Partially Observable Data **15**

Appendix D Proof of Proposition 1 **16**

Appendix E Linear Regression with Side Information **18**

Appendix F OFUL with Linear Side Information **20**

F.1 Equivalences of Update Rule 20

F.2 Proof of Theorem 1 21

F.3 Optimism of OFUL 22

Appendix G OFUL with Learned Subspace **24**

G.1 Notations and the Good Event of Theorem 2 24

G.2 Proof of Theorem 2 26

G.3 Optimism of OFUL with Learned Subspace 29

Appendix H Convergence of $M, M^\dagger, P$ **32**

Appendix I Useful Results **36**

APPENDIX A OVERVIEW

The appendix is orginized as follows:

- Appendix B provides a short background on orthogonal projections and the Moore-Penrose inverse, a fundamental tool in our work. We also state some well known pertrubations bounds that will be useful for achieving well-behaved convergence rates in the proof of Theorem 2.
- In Appendix C we discuss the unidentifiability problem of using partially observable data, with focus to our setting. Specifically, we show that $\{w_a^*\}_{a\in\mathcal{A}}$ are not identifiable, and provide further motivation for this work.
- Appendix D provides a proof of Proposition 1, namely, showing that the structure of the transformation obtained from the offline data is given by $M_a = (I_L, R_{11}(a)^{-1}R_{12}(a))$.
- In Appendix E we discuss Projected Ridge Regression (P-RR) with side information, as presented in Section 4. Specifically, we state and prove the equivalence of P-RR to a low-order ridge regression problem with a correction term (see Equation (4)).
- Appendix F provides a proof of Theorem 1, showing that Algorithm 1 reduces the OFUL algorithm to the lower-dimensional projected subspace of known linear transformation $\{M_a w = b_a\}$, decreasing the regret from $\widetilde{\mathcal{O}}\left(d\sqrt{KT}\right)$ to $\widetilde{\mathcal{O}}\left((d-L)\sqrt{KT}\right)$. Its proof is a direct consequence of the P-RR formulation.
- Based on the above, we formulate our final result, proving Theorem 2 in Appendix G. We begin by showing the approximate linear transformations obtained by estimating $R_{12}(a)$ from online samples are well behaved. Specifically, we show that the projection and pseudo-inverse operators converge at the same rate as $\hat{M}_{t,a}$, i.e., $\left\|P_a - \hat{P}_{t,a}\right\| \leq 2\left\|M_a - \hat{M}_{t,a}\right\|$ and $\left\|M_a^\dagger - \hat{M}_{t,a}^\dagger\right\| \leq 2\left\|M_a - \hat{M}_{t,a}\right\|$. We then leverage these important properties, together with a doubling trick approach, showing that similar regret guarantees can be achieved as in the exact case.

APPENDIX B ORTHOGONAL PROJECTIONS AND THE MOORE–PENROSE INVERSE

In this section we give a short review of the Moore–Penrose inverse [Barata and Hussein, 2012] and its corresponding orthogonal projection. We state some well known properties that will be useful in our analysis.

For a matrix $M \in \mathbb{R}^{L\times d}$ we denote its Moore–Penrose inverse by $M^\dagger$. Let $P^\parallel = M^\dagger M$ be the orthogonal projection onto the range of M and $P^\perp = I - P^\parallel$ be the orthogonal projection onto the kernel of M. We have the following well known properties.

Property 1. *If M has independent rows, then $M^\dagger$ can be computed as $M^\dagger = M^T(MM^T)^{-1}$.*

Property 2. *If $Mw = b$ then $P^\parallel w = M^\dagger b$.*

Property 3 (Planitz 1979)**.** *The vector $x = M^\dagger b$ is the vector with the smallest L_2 norm which satisfies $Mx = b$.*

Property 4. *If $w \in \mathbb{R}^d$ satisfies $Mw = b$ then w can be written as*

$$w = P^\perp w + M^\dagger b.$$

Proof. Using the fact $I = P^\parallel w + P^\perp$ and Property 2 we get

$$w = P^\perp w + P^\parallel w = P^\perp w + M^\dagger b.$$

□

Remark 1 (Notation)**.** *For brevity, we denote $P = P^\perp$ as the orthogonal projection to the kernel of M and $I - P = I - P^\perp = P^\parallel$ as the orthogonal projection to the range of M.*

B.1 USEFUL PERTURBATION BOUNDS

The following two results are standard in perturbation theory. They bound the L_2 difference between some matrix A, and its perturbed counterpart $B = A + E$, where E a perturbation (i.e., error) matrix.

Theorem 3 (E.g., Chen et al. 2016, Corollary 2.7)**.** *For any matrices let $A, B, E \in \mathbb{R}^{d\times d}$ and $E = B - A$. Let P_A and P_B be the orthogonal projection on the raw space of A and B, respectively. Assume $\operatorname{rank}(A) = \operatorname{rank}(B)$. Then,*

$$\|P_A - P_B\|_2 \leq \min\left(\left\|A^\dagger\right\|_2, \left\|B^\dagger\right\|_2\right) \|E\|_2,$$

where $P, \hat{P}$ are the orthogonal projections into the row space of $M, \hat{M}$, respectively,

Theorem 4 (Stewart 1977, Theorem 3.3)**.** *For any matrices let $A, B, E \in \mathbb{R}^{d\times d}$ and $E = B - A$. Then,*

$$\left\|B^\dagger - A^\dagger\right\|_2 \leq 2\max\left(\left\|A^\dagger\right\|_2^2, \left\|B^\dagger\right\|_2^2\right) \|E\|_2.$$

Remark 2. *Note that Theorem 3 assumes $\operatorname{rank}(A) = \operatorname{rank}(B)$. While other perturbation bounds exist for the case $\operatorname{rank}(A) \neq \operatorname{rank}(B)$, they do not provide sufficient guarantees for our analysis (e.g., $\left\|A^\dagger - B^\dagger\right\|$ may diverge Stewart 1969). Luckily, due to the special structure of M, i.e., $M = (I_L, R_{11}^{-1}R_{12})$, the perturbed version of M as defined in Section 5 will always have rank L, ensuring this assumption holds.*

APPENDIX C UNIDENTIFIABILITY PROBLEM OF PARTIALLY OBSERVABLE DATA

Having access to partially observable offline data may not be enough to obtain the optimal policy, even if $N \to \infty$. The following is a standard identifiability result, showing that $\{w_a^*\}_{a\in\mathcal{A}}$ are not identifiable in the setting of partially observable data and no online interaction.

Proposition 2. *For any behavioral policy π_b and induced measure P^{π_b}, $\{w_a^*\}_{a\in\mathcal{A}}$ are not identifiable. More specifically, for all $w^1 = \{w_a^1\}_{a\in\mathcal{A}}$ exist $w^2 = \{w_a^2\}_{a\in\mathcal{A}} \neq w^2$ and probability measures P_1, P_2 such that $P_1\left(x^{\mathrm{o}}, a, r; w^1, \pi_b\right) = P_2\left(x^{\mathrm{o}}, a, r; w^2, \pi_b\right)$ and $\pi_b\left(a, x; w^1\right) = \pi_b\left(a, x; w^2\right)$.*

Proof. Denote by $w_a^o \in \mathbb{R}^L, w_a^h \in \mathbb{R}^{d-L}$, the vectors corresponding to the observed and hidden parts of x, namely, x^{o} and x^{h}, respectively. That is, $w_a = \begin{pmatrix} w_a^o \\ w_a^h \end{pmatrix}$ and

$$r_i = x_i^T w_a + \eta_i = (x^{\mathrm{o}})^T w_a^o + (x^{\mathrm{h}})^T w_a^h + \eta_i.$$

Let P_1, P_2 be two measures such that $P_1((x^{\mathrm{h}})^T w_a^{1o} \leq \alpha | a, x^{\mathrm{o}}) = P_2((x^{\mathrm{h}})^T w_a^{2o} \leq \alpha | a, x^{\mathrm{o}})$ for all $\alpha \in \mathbb{R}$. Note that such measures always exist. For example letting x^{h} be a random vector with i.i.d. elements independent of x^{o}, such that

$$\begin{cases} P_1\left((x^{\mathrm{h}})_k = \frac{1}{(w_a^{1o})_k}\right) = \frac{1}{2} \\ P_1\left((x^{\mathrm{h}})_k = \frac{1}{(w_a^{2o})_k}\right) = 0 \\ P_1\left((x^{\mathrm{h}})_k = 0\right) = \frac{1}{2} \end{cases} \qquad \begin{cases} P_2\left((x^{\mathrm{h}})_k = \frac{1}{(w_a^{1o})_k}\right) = 0 \\ P_2\left((x^{\mathrm{h}})_k = \frac{1}{(w_a^{2o})_k}\right) = \frac{1}{2} \\ P_2\left((x^{\mathrm{h}})_k = 0\right) = \frac{1}{2} \end{cases}$$

where without loss of generality we assumed $(w_a^{1o})_k \neq 0, (w_a^{2o})_k \neq 0, \forall 1 \leq k \leq d - L$ [7]. Then it follows that $P_1((x^{\mathrm{h}})^T w_a^{1o} \leq \alpha | a, x^{\mathrm{o}}) = P_2((x^{\mathrm{h}})^T w_a^{2o} \leq \alpha | a, x^{\mathrm{o}})$.

We have that

$$\begin{aligned} P_1(r \leq \beta | a = a, x^{\mathrm{o}} = c; w^1, \pi_b) &= P_1((x^{\mathrm{h}})^T w_a^o + \eta \leq \beta - c^T w_a^o | a = a, x^{\mathrm{o}} = c; w^1, \pi_b) \\ &= P_2(r \leq \beta | a = a, x^{\mathrm{o}} = c; w^2, \pi_b). \end{aligned}$$

Therefore,

$$\begin{aligned} P_1(r, a, x^{\mathrm{o}}; w^1, \pi_b) &= P_1(r | a, x^{\mathrm{o}}; w^1, \pi_b) P^{\pi_b}(a, x^{\mathrm{o}}) \\ &= P_2(r | a, x^{\mathrm{o}}; w^2, \pi_b) P^{\pi_b}(a, x^{\mathrm{o}}) \\ &= P_2(r, a, x^{\mathrm{o}}; w^2, \pi_b). \end{aligned}$$

□

The above proposition tells us that partially observable offline data cannot be used unless further assumptions are made. This is true even in the linear model, which is the focus of this work. To mitigate this problem, prior knowledge and assumptions over the hidden variables can be used. Nevertheless, such assumptions may be inaccurate and impossible to validate. Moreover, most concurrent assumptions (such as bounding the "amount of confoundness"), do not completely resolve this issue, but rather mitigate it so that perhaps better policies can be found.

This work considers an alternative assumption, namely, access to online interactions with the environment. This has several benefits over assuming prior knowledge:

1. There is no problem of validating this assumption, i.e., if we do not have access to an online environment, we will know it.
2. The online regime allows us to achieve an optimal policy. Specifically, since online interactions reveal the hidden context, an optimal policy is always identifiable.
3. Partially observable data cannot hurt us, but rather only improve our performance. Looking at the problem from an online learner's point of view, the offline data is only used to improve its performance. Therefore, the offline data does not bias our results, but only decreases the number of online interactions.

[7] If one of them is zero, simply choose $P_1\left((x^{\mathrm{h}})_k = 0\right) = P_2\left((x^{\mathrm{h}})_k = 0\right) = 1$

APPENDIX D PROOF OF PROPOSITION 1

We recall Proposition 1

Proposition 1. *[Confoundness = Linear Constraints]*
Let $R_{11}(a) = \mathbb{E}^{\pi_b}\left[x^{\mathrm{o}}\left(x^{\mathrm{o}}\right)^T \,\middle|\, a\right]$, $R_{12}(a) = \mathbb{E}^{\pi_b}\left[x^{\mathrm{o}}\left(x^{\mathrm{h}}\right)^T \,\middle|\, a\right]$. *Assume* $R_{11}(a)$ *is invertible for all* $a \in \mathcal{A}$ [8]. *Then, the following holds almost surely for all* $a \in \mathcal{A}$.

$$\lim_{N\to\infty} b_a^{LS} = \begin{pmatrix} I_L, & R_{11}^{-1}(a)R_{12}(a) \end{pmatrix} w_a^*.$$

Before proving the proposition, we remind the reader of the continuous mapping theorem, which states that continuous functions preserve limits even if their arguments are sequences of random variables.

Theorem 5 (Continuous Mapping, e.g., [Van der Vaart, 2000])**.** *Let* $\{X_n\}$ *be a set of real random variables such* $X_n \in \mathbb{R}^k$ *and* $X_n \xrightarrow{a.s} X$. *Let* $d(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *be a distance function that generates the usual topology. Let* $g : \mathbb{R}^k \to \mathbb{R}^m$ *be a continuous function at the point* $X \in \mathbb{R}^k$. *Then* $g(X_n) \xrightarrow{a.s} g(X)$.

Proof of Proposition 1. Fix $a \in \mathcal{A}$. By definition, b_a^{LS} is also given by (1),

$$b_a^{LS} = \left(\frac{1}{N_a}\sum_{n\in[N]} \mathbb{1}_{\{a_n=a\}} x_n^{\mathrm{o}}\left(x_n^{\mathrm{o}}\right)^T\right)^{-1}\left(\frac{1}{N_a}\sum_{n\in[N]} \mathbb{1}_{\{a_n=a\}} x_n^{\mathrm{o}} r_n\right), \tag{7}$$

Define $\left(Q^{\mathrm{o}}\right)^T = \begin{pmatrix} I_L & 0 \end{pmatrix} \in \mathbb{R}^{L\times d}$ and $\left(Q^{\mathrm{c}}\right)^T = \begin{pmatrix} 0 & I_{d-L} \end{pmatrix} \in \mathbb{R}^{d-L\times d}$. Observe that $x^{\mathrm{o}} = Q^{\mathrm{o}}x$, $x^{\mathrm{h}} = Q^{\mathrm{c}}x$ for $x \in \mathbb{R}^d$, and that $I_d = Q^{\mathrm{o}}\left(Q^{\mathrm{o}}\right)^T + Q^{\mathrm{c}}\left(Q^{\mathrm{c}}\right)^T$. Due to the model assumption (see Section 2), we can write

$$\begin{aligned} r(x_n,a) &= \langle x_n, w_a^*\rangle + \eta_n \\ &= \left\langle Q^{\mathrm{o}}\left(Q^{\mathrm{o}}\right)^T x_n, w_a^*\right\rangle + \left\langle Q^{\mathrm{c}}\left(Q^{\mathrm{c}}\right)^T x_n, w_a^*\right\rangle + \eta_n \qquad \left(Q^{\mathrm{o}}\left(Q^{\mathrm{o}}\right)^T + Q^{\mathrm{c}}\left(Q^{\mathrm{c}}\right)^T = I\right) \\ &= \left\langle x_n^{\mathrm{o}}, (Q^{\mathrm{o}})^T w_a^*\right\rangle + \left\langle x_n^{\mathrm{h}}, (Q^{\mathrm{c}})^T w_a^*\right\rangle + \eta_n \end{aligned}$$

Plugging this relation into (1) we get

$$\begin{aligned} b_a^{LS} =& \left(\frac{1}{N_a}\sum_{n\in[N]} \mathbb{1}_{\{a_n=a\}} x_n^{\mathrm{o}}\left(x_n^{\mathrm{o}}\right)^T\right)^{-1}\left(\frac{1}{N_a}\sum_{n\in[N]} \mathbb{1}_{\{a_n=a\}} x_n^{\mathrm{o}}\left(x_n^{\mathrm{o}}\right)^T\right)\left(Q^{\mathrm{o}}\right)^T w_a^* \\ &+ \left(\frac{1}{N_a}\sum_{n\in[N]} \mathbb{1}_{\{a_n=a\}} x_n^{\mathrm{o}}\left(x_n^{\mathrm{o}}\right)^T\right)^{-1}\left(\frac{1}{N_a}\sum_{n\in[N]} \mathbb{1}_{\{a_n=a\}} x_n^{\mathrm{o}}\left(x_n^{\mathrm{h}}\right)^T\right)\left(Q^{\mathrm{c}}\right)^T w_a^* \\ &+ \left(\frac{1}{N_a}\sum_{n\in[N]} \mathbb{1}_{\{a_n=a\}} x_n^{\mathrm{o}}\left(x_n^{\mathrm{o}}\right)^T\right)^{-1}\left(\frac{1}{N_a}\sum_{n\in[N]} \mathbb{1}_{\{a_n=a\}} x_n^{\mathrm{o}}\eta_n\right). \end{aligned} \tag{8}$$

We now analyze the three terms and apply the continuous mapping theorem iteratively (Theorem 5) to prove the lemma. Let

$$Y_N^{(i)} \coloneqq \frac{1}{N_a}\sum_{n\in[N]} \mathbb{1}_{\{a_n=a\}} x_n^{\mathrm{o}}\left(x_n^{\mathrm{o}}\right)^T = \frac{1}{N_a/N}\frac{1}{N}\sum_{n\in[N]} \mathbb{1}_{\{a_n=a\}} x_n^{\mathrm{o}}\left(x_n^{\mathrm{o}}\right)^T = \frac{Z_N^{(2)}}{Z_N^{(1)}}, \tag{9}$$

where $Z_N^{(1)} \coloneqq N_a/N = \frac{1}{N}\sum_{n\in[N]} \mathbb{1}_{\{a_n=a\}}$ and $Z_N^{(2)} \coloneqq \frac{1}{N}\sum_{n\in[N]} \mathbb{1}_{\{a_n=a\}} x_n^{\mathrm{o}}\left(x_n^{\mathrm{o}}\right)^T$.

[8] The invertibility assumption on R_{11} can be verified, since R_{11} can be estimated by statistics of the observable covariates, x^{o}. If it does not hold, other covariates of x^{o} can be chosen to satisfy this assumption.

By the strong law of large numbers (SLLN)

$$Z_N^{(1)} \overset{a.s}{\to} P^{\pi_b}(a)$$
$$Z_N^{(2)} \overset{a.s}{\to} P^{\pi_b}(a)\mathbb{E}^{\pi_b}\Big[x^{\mathrm{o}}\left(x^{\mathrm{o}}\right)^T \mid a\Big],$$

for $N \to \infty$ since both are empirical averages of N i.i.d. random variables and, thus, converge to their means a.s..

The mean of a random variable in the empirical average $Z_N^{(1)}$ is simply given by

$$\mathbb{E}^{\pi_b}\left[\mathbb{1}_{\{a_n=a\}}\right] = P^{\pi_b}(a),$$

since all the random variables are i.i.d..

The mean of a random variable in the empirical average of $Z_N^{(2)}$ is given by

$$\mathbb{E}^{\pi_b}\Big[\mathbb{1}_{\{a_n=a\}}x_n^{\mathrm{o}}\left(x_n^{\mathrm{o}}\right)^T\Big] = P^{\pi_b}(a)\mathbb{E}^{\pi_b}\Big[x_n^{\mathrm{o}}\left(x_n^{\mathrm{o}}\right)^T \mid a\Big],$$

since $\mathbb{E}[X \mid A] = \frac{\mathbb{E}\left[\mathbb{1}_{\{A\}}X\right]}{P(A)}$.

By the continuous mapping theorem (Theorem 5) we get that

$$Y_N^{(i)} \overset{a.s}{\to} \frac{P^{\pi_b}(a)\mathbb{E}^{\pi_b}\Big[x^{\mathrm{o}}\left(x^{\mathrm{o}}\right)^T \mid a\Big]}{P^{\pi_b}(a)} = \mathbb{E}^{\pi_b}\Big[x^{\mathrm{o}}\left(x^{\mathrm{o}}\right)^T \mid a\Big],$$

which is valid since $g(a,b) = \frac{a}{b}$ is continuous at $b > 0$ and we assume that for all $a \in \mathcal{A}$ $P^{\pi_b}(a) > 0$.

Similar reasoning, leads to the following convergence

$$Y_N^{(ii)} \coloneqq \frac{1}{N_a}\sum_{n\in[N]}\mathbb{1}_{\{a_n=a\}}x_n^{\mathrm{o}}\left(x_n^{\mathrm{h}}\right)^T \overset{a.s}{\to} \mathbb{E}^{\pi_b}\Big[x^{\mathrm{o}}\left(x^{\mathrm{h}}\right)^T \mid a\Big], \tag{10}$$

and

$$Y_N^{(iii)} \coloneqq \frac{1}{N_a}\sum_{n\in[N]}\mathbb{1}_{\{a_n=a\}}x_n^{\mathrm{o}}\eta_n \overset{a.s}{\to} 0, \tag{11}$$

where in the last relation we also use the fact that η_n is zero mean random variable, $\mathbb{E}^{\pi_b}[\eta_n] = 0$, and is independent of x_n, a_n and thus η_n is also independent of $x_n^{\mathrm{o}} = Q^{\mathrm{o}}x_n, a_n$.

By (8) and by definitions (9), (10), (11),

$$b_a^{LS} = (Y_N^{(i)})^{-1}Y_N^{(i)}Q^{\mathrm{o}}w_a^* + (Y_N^{(i)})^{-1}Y_N^{(ii)}Q^{\mathrm{c}}w_a^* + (Y_N^{(i)})^{-1}Y_N^{(iii)}.$$

We now apply the continuous mapping theorem (Theorem 5) on each of the three terms. Notice that the conditions of this theorem are satisfied since the limit of $Y_N^{(i)}$, $\mathbb{E}^{\pi_b}\Big[x^{\mathrm{o}}\left(x^{\mathrm{o}}\right)^T \mid a\Big]$, has an inverse by the assumption which implies that the limit point is continuous. Thus, we get that for $N \to \infty$ it holds a.s. that

$$b_a^{LS} = \left(Q^{\mathrm{o}}\right)^T w_a^* + R_{11}(a)^{-1}R_{12}(a)\left(Q^{\mathrm{c}}\right)^T w_a^*$$

where $R_{11}(a) = \mathbb{E}^{\pi_b}\Big[x^{\mathrm{o}}\left(x^{\mathrm{o}}\right)^T \Big| a\Big]$, $R_{12}(a) = \mathbb{E}^{\pi_b}\Big[x^{\mathrm{o}}\left(x^{\mathrm{h}}\right)^T \Big| a\Big]$. By taking union bound on all $a \in \mathcal{A}$ we conclude the proof. □

APPENDIX E LINEAR REGRESSION WITH SIDE INFORMATION

We wish to construct a least squares variant for w_a^*, which utilizes the information in Equation (2). In Section 4 we considered the problem of linear regression under the linear model $y = \langle x, Pw^* \rangle + \eta$, where $P \in \mathbb{R}^{d \times d}$ is a projection matrix of rank m and η is some centered random noise. One way to solve this problem is through ridge regression on the full euclidean space $\mathbb{R}^d$, under projection of w. In Section 4 we constructed a projected variant of ridge regression (P-RR), which attempts to solve the regularized problem in Equation 4, i.e.,

$$\min_{w \in \mathbb{R}^d} \left\{ \sum_{i=1}^{t-1} \left(\langle x_i, P_a w \rangle - y_{a,i} \right)^2 + \lambda \left\| P_a w \right\|_2^2 \right\},$$

We then took its smallest norm solution

$$\hat{w}_{t,a}^{P_a} = \left(P_a \left(\lambda I_d + \sum_{i=1}^{t-1} x_i x_i^T \right) P_a \right)^{\dagger} \left(\sum_{i=1}^{t-1} y_i x_i \right). \tag{12}$$

where $y_{a,i} = r_i - \left\langle x_i, M_{a_i}^{\dagger} b_{a_i} \right\rangle$.

Let us now take a closer look at Equation (12). We wish to show that this solution is closely related to a ridge regression problem in a lower dimension. First, notice that taking the pseudoinverse of $P_a \left(\lambda I_d + \sum_{i=1}^{t-1} x_i x_i^T \right) P_a$ can be written in an equivalent method using the inverse operator. This is formalized generally in the following proposition.

Proposition 3. *Let $V \in \mathbb{R}^{d \times d}$ be a PD matrix and $P = UU^T$ a projection matrix of rank l, where and $U \in \mathbb{R}^{d \times l}$ is a matrix with orthonormal columns. Then,*

$$(PVP)^{\dagger} = U(U^T V U)^{-1} U^T.$$

Proof. Observe that

$$PVP = U(U^T V U)U^T \tag{13}$$

Let $U_{\mathrm{Ext}} \in \mathbb{R}^{d \times d}$ be a unitary matrix with its first l columns U and rest $d - l$ columns be arbitrary orthonormal vectors (such that U_{Ext} is unitary), i.e., an extension of U to the entire space $\mathbb{R}^d$. Using this notation (13) can be written as

$$PVP = U_{\mathrm{Ext}} \begin{pmatrix} U^T V U & 0 \\ 0 & 0 \end{pmatrix} U_{\mathrm{Ext}}^T. \tag{14}$$

Next, recall that for any unitary matrix $\bar{U}$ and any matrix A it holds that $(\bar{U} A \bar{U}^T)^{\dagger} = \bar{U} A^{\dagger} \bar{U}^T$. Then, by Equation (14)

$$\begin{aligned}
(PVP)^{\dagger} &= \left(U_{\mathrm{Ext}} \begin{pmatrix} U^T V U & 0 \\ 0 & 0 \end{pmatrix} U_{\mathrm{Ext}}^T \right)^{\dagger} \\
&= U_{\mathrm{Ext}} \begin{pmatrix} U^T V U & 0 \\ 0 & 0 \end{pmatrix}^{\dagger} U_{\mathrm{Ext}}^T && ((UAU^T)^{\dagger} = UA^{\dagger}U^T) \\
&= U_{\mathrm{Ext}} \begin{pmatrix} (U^T V U)^{\dagger} & 0 \\ 0 & 0 \end{pmatrix} U_{\mathrm{Ext}}^T \\
&= U_{\mathrm{Ext}} \begin{pmatrix} (U^T V U)^{-1} & 0 \\ 0 & 0 \end{pmatrix} U_{\mathrm{Ext}}^T = U(U^T V U)^{-1} U^T, && (15)
\end{aligned}$$

where the third relation holds since $U^T V U$ is full rank (since V is PD so is $U^T V U$). □

Using Proposition 3 we now show that P-RR is in fact equivalent to solving a ridge regression problem in a lower dimension. This will become useful in our proof of Theorem 1, as it will allow us to reduce the linear bandit problem to a lower dimensional variant of the same problem. The following proposition proves this equivalence.

Proposition 4 (Equivalent forms of P-RR)**.** *Let $w_{\mathrm{PRR}} \in \mathbb{R}^d$ be the least L_2-norm solution of the following P-RR for $\lambda > 0$*

$$\arg\min_{w \in \mathbb{R}^d} \left(\sum_{n=1}^{t} \left(\langle P x_n, w \rangle - y_n \right)^2 + \lambda \left\| P w \right\|_2^2 \right),$$

where $x_n \in \mathbb{R}^d, y_n \in \mathbb{R}$, $P = UU^T$ is a projection matrix of rank l and $U \in \mathbb{R}^{d \times l}$ is a matrix with orthonormal columns. Define $X \in \mathbb{R}^{t \times d}$ and $\tilde{X} \in \mathbb{R}^{t \times l}$ as matrices with $\left\{ x_n^T \right\}_{n=1}^t$ and $\left\{ (U^T x_n)^T \right\}_{n=1}^t$ in their rows, respectively. Define $Y \in \mathbb{R}^t$ as the vector of $\{y_n\}_{n=1}^t$. Then, w_{PRR} satisfies the following relations.

1. $w_{\mathrm{PRR}} = (P(X^T X + \lambda I_d)P)^\dagger P X^T Y$.
2. $U^T w_{\mathrm{PRR}} = (\tilde{X}^T \tilde{X} + \lambda I_l)^{-1} \tilde{X}^T Y$.

Proof. Claim (1). The P-RR problem can we recast as

$$\arg\min_{w \in \mathbb{R}^d} \left(w^T (P(X^T X + \lambda I_d)P) w + 2 w^T P X^T Y \right).$$

The smallest L_2 norm solution of this optimization problem is given by

$$w_{\mathrm{PRR}} = (P(X^T X + \lambda I_d)P)^\dagger P X^T Y \tag{16}$$

which establishes the first claim.

Claim (2).

By Proposition 3 it holds that

$$\begin{aligned}
(P(X^T X + \lambda I_d)P)^\dagger &= U(U^T(X^T X + \lambda I_d)U)^{-1} U^T && \text{(Prop. 3)} \\
&= U(U^T X^T X U + \lambda I_l)^{-1} U^T && (U^T U = I_l) \\
&= U(\tilde{X}^T \tilde{X} + \lambda I_l)^{-1} U^T, && (17)
\end{aligned}$$

by defining $\tilde{X} = XU$, which is a matrix with $\left\{ U^T x_n \right\}$ in its rows. To conclude the proof, observe the following holds

$$\begin{aligned}
U^T w_{\mathrm{PRR}} &= U^T \left(U(\tilde{X}^T \tilde{X} + \lambda I_l)^{-1} U^T \right) P X^T Y && \text{(By (16) and (17))} \\
&= U^T U (\tilde{X}^T \tilde{X} + \lambda I_l)^{-1} U^T U \tilde{X}^T Y \\
&= (\tilde{X}^T \tilde{X} + \lambda I_l)^{-1} \tilde{X}^T Y. && (U^T U = I_l)
\end{aligned}$$

□

APPENDIX F OFUL WITH LINEAR SIDE INFORMATION

Algorithm 3 OFUL with Linear Side Information (Equivalent Form)

1: **input:** $\alpha > 0, M_a \in \mathbb{R}^{L\times d}, b_a \in \mathbb{R}^L$

2: **init:** $V_a = \lambda I, Y_a = 0, \forall a \in \mathcal{A}$

3: **for** $t = 1, \ldots, T$ **do**

4: Receive context x_t

5: $\hat{w}_{t,a}^{P_a} = (P_a V_a P_a)^\dagger \left(Y_a - (V_a - \lambda I) M_a^\dagger b_a\right)$

6: $\sqrt{\beta_t(\delta)} = \sigma\sqrt{(d-L)\log\left(\frac{K(1+tS_x^2/\lambda)}{\delta}\right)} + \lambda^{1/2} S_w$

7: $a_t, \tilde{w}_{t,a_t} \in \arg\max_{a\in\mathcal{A}, w_a \in C_{t,a}} \left(\langle x_t, P_a w_a\rangle + \left\langle x_t, M_a^\dagger b_a\right\rangle\right)$

8: Play action a_t and receive r_t

9: $V_{a_t} = V_{a_t} + x_t x_t^T, Y_{a_t} = Y_{a_t} + x_t r_t$

10: **end for**

In this section we supply regret guarantees for OFUL [Abbasi-Yadkori et al., 2011] with linear side information of the form $M_a w_a = b_a$ for every $a \in \mathcal{A}$. In Appendix G, building on this analysis, we study the case M_a is estimated in an online manner.

F.1 EQUIVALENCES OF UPDATE RULE

The optimistic estimation of the reward of each arm has the form (Algorithm 1)

$$\langle x_t, P_a \hat{w}_{t,a}\rangle + \|x\|_{(P_a V_{t-1,a} P_a)^\dagger} + \left\langle x_t, M_a^\dagger b_a\right\rangle.$$

In this section we establish that this update rule is equivalent to the update rule written in Algorithm 3. For computational purposes, it is easier to consider the version given in Algorithm 1, whereas in terms of analysis, the equivalent form given in Algorithm 3. The following proposition proves this equivalence by providing an analytical solution to the optimisitic optimization problem in Line 7 of Algorithm 3.

Proposition 5 (Equivalent Forms of OFUL's Optimistic Estimation)**.** *Let $P = UU^T$ be an orthogonal projection matrix, $U \in \mathbb{R}^{d\times l}$ a matrix with orthonormal columns and $V \in \mathbb{R}^{d\times d}$ a PD symmetric matrix. Fix $\hat{w} \in \mathbb{R}^d, \beta \in \mathbb{R}$ and let*

$$\mathcal{C} = \{w : \left\|U^T w - U^T \hat{w}\right\|_{\tilde{V}} \le \sqrt{\beta}\},$$

where $\tilde{V} = U^T V U \in \mathbb{R}^{l\times l}$. Then, for any fixed $x \in \mathbb{R}^d$

$$\max_{w\in\mathcal{C}}(\langle x, Pw\rangle) = \langle x, P\hat{w}\rangle + \sqrt{\beta}\,\|x\|_{(PVP)^\dagger}$$

Proof. We have that

$$\begin{aligned}\max_{w\in\mathcal{C}}(\langle x, Pw\rangle) &= \langle x, P\hat{w}\rangle + \max_{w\in\mathcal{C}}(\langle x, P\,(w-\hat{w})\rangle) \\ &= \langle x, P\hat{w}\rangle + \max_{w\in\mathcal{C}}\left(\left\langle U^T x, U^T w - U^T\hat{w}\right\rangle\right). \end{aligned} \tag{18}$$

Next,

$$\begin{aligned}\left\langle U^T x, U^T w - U^T\hat{w}\right\rangle &\le \left\|U^T x\right\|_{\tilde{V}^{-1}} \left\|U^T w - U^T\hat{w}\right\|_{\tilde{V}} \\ &\le \sqrt{\beta}\left\|U^T x\right\|_{\tilde{V}^{-1}},\end{aligned}$$

where the first inequality follows by Cauchy-Schwartz and $\tilde{V} = U^T V U$ is PD, and the second inequality by the assumption $\left\|U^T w - U^T \hat{w}\right\|_{\tilde{V}} \leq \sqrt{\beta}$ for any $w \in \mathcal{C}$. Moreover, the inequality is attained with equality for $\tilde{V}(U^T w - U^T \hat{w}) = \frac{\sqrt{\beta}}{\|U^T x\|_{\tilde{V}^{-1}}} \tilde{V}^{-1} U^T x$ (such w is indeed contained in $\mathcal{C}$). Thus, we get that

$$\max_{w \in \mathcal{C}} \left(\left\langle U^T x, U^T w - U^T \hat{w} \right\rangle \right) = \sqrt{\beta} \left\| U^T x \right\|_{\tilde{V}^{-1}}.$$

Plugging the above into Equation (18) and we get

$$\begin{aligned}
\max_{w \in \mathcal{C}} (\langle x, Pw \rangle) &= \langle x, P\hat{w} \rangle + \sqrt{\beta} \left\| U^T x \right\|_{\tilde{V}^{-1}} \\
&= \langle x, P\hat{w} \rangle + \sqrt{\beta} \|x\|_{U\tilde{V}^{-1}U^T} \\
&= \langle x, P\hat{w} \rangle + \sqrt{\beta} \|x\|_{(PVP)^\dagger}. && \text{(Proposition 3)}
\end{aligned}$$

□

F.2 PROOF OF THEOREM 1

We are now ready to prove Theorem 1. We first provide a sketch of the proof. Using the linear side observation we can deduce $(I - P_a) w_a^* = M_a^\dagger b_a$. Thus, we should only estimate the part of w_a^* not given by the linear side observation, $P_a w_a^*$. To this end, we use the P-RR estimator (see Section 4 and Appendix E)

$$\hat{w}_{t+1,a} \in \arg \min_{w \in \mathbb{R}^d} \left\{ \sum_{n=1}^{t} \mathbb{1}_{\{a_n = a\}} \left(\langle P_a x_n, w \rangle - \left(y_n - \left\langle x_n, M_a^\dagger b_a \right\rangle \right) \right)^2 + \lambda \|P_a w\|_2^2 \right\}, \tag{19}$$

Proof. We start by defining the good event $\mathcal{G}$ as the event $\left\{ U_a^T w_a^* \in C_{t,a} \forall t \geq 0, \forall a \in \mathcal{A} \right\}$. By Lemma 1, $P(\mathcal{G}) > 1 - \delta$. Denote by $I_k(a)$ the k^{th} time action a was chosen in the sequence $x_1, a_1, \ldots, x_t, a_t$, that is,

$$I_k(a) = \min \left\{ t : \sum_{j=1}^{t} \mathbb{1}_{\{a_j = a\}} = k \right\},$$

and denote by $N_t(a)$ the total number of times action a was chosen by time t. Also, denote the PD matrix

$$\tilde{V}_{t-1,a} = \lambda I_{d-L} + \sum_{i=1}^{t-1} \mathbb{1}_{\{a_i = a\}} (U_a^T x_i)(U_a^T x_i)^T.$$

Then the following relations hold for every $t \geq 0$ conditioning on the good event.

$$\begin{aligned}
r_t =& \left\langle x_t, w^*_{a^*(x_t)} \right\rangle - \left\langle x_t, w^*_{a_t} \right\rangle \\
&\leq \langle P_{a_t} x_t, \tilde{w}_{a_t} \rangle + \left\langle x_t, M^\dagger_{a_t} b_{a_t} \right\rangle - \left\langle x_t, w^*_{a_t} \right\rangle && \text{(Corollary (1))} \\
&= \langle P_{a_t} x_t, \tilde{w}_{a_t} \rangle - \left\langle P_{a_t} x_t, w^*_{a_t} \right\rangle && (w_a = P_a w_a + M_a^\dagger b_a, \text{Property 4}) \\
&= \left\langle U^T_{a_t} x_t, \left(U^T_{a_t} \tilde{w}_{a_t} - U^T_{a_t} w^*_{a_t} \right) \right\rangle && (P_a = U_a U_a^T) \\
&\leq \left\| U^T_{a_t} x_t \right\|_{\tilde{V}^{-1}_{t-1,a_t}} \left\| U^T_{a_t} \hat{w}_{a_t} - U^T_{a_t} w^*_{a_t} \right\|_{\tilde{V}_{t-1,a_t}} + \left\| U^T_{a_t} x_t \right\|_{\tilde{V}^{-1}_{t-1,a_t}} \left\| U^T_{a_t} \hat{w}_{a_t} - U^T_{a_t} \tilde{w}_{a_t} \right\|_{\tilde{V}_{t-1,a_t}} \\
&\leq 2\sqrt{\beta_{t-1}(\delta)} \left\| U^T_{a_t} x_t \right\|_{\tilde{V}^{-1}_{t-1,a_t}}, && \text{(Lemma 1)}
\end{aligned}$$

Next, notice that $r_t \leq 2$, since $\langle x_t, w_a \rangle \in [-1, 1]$. Then, using the above we get that

$$r_t \leq 2\sqrt{\beta_{t-1}(\delta)} \min \left(\left\| U^T_{a_t} x_t \right\|_{\tilde{V}^{-1}_{t-1,a_t}}, 1 \right). \tag{20}$$

Combining the above and conditioning on the good event we get that for all $t \geq 0$

$$
\begin{aligned}
\textit{Regret}\,(t) &\leq \sqrt{t\sum_{i=1}^{t} r_t^2} \\
&\leq 2\sqrt{t\beta_t(\delta)\sum_{i=1}^{t}\min\left(\left\|U_{a_i}^T x_i\right\|^2_{\tilde{V}^{-1}_{i-1,a_i}}, 1\right)} && \text{((20) and } \beta_t(\delta) \geq \beta_{t'}(\delta) \text{ for } t \geq t') \\
&= 2\sqrt{t\beta_t(\delta)\sum_{a\in\mathcal{A}}\sum_{i=1}^{N_t(a)}\min\left(\left\|U_a^T x_{I_i(a)}\right\|^2_{\tilde{V}^{-1}_{I_i(a),a}}, 1\right)} \\
&\leq 2\sqrt{t}\left(\lambda^{1/2}S_w + \sigma\sqrt{(d-L)\log\left(\frac{K(1+tS_x^2/\lambda))}{\delta}\right)}\right)\sqrt{(d-L)\sum_{a\in\mathcal{A}}\log\left(\lambda + \frac{N_t(a)S_x^2}{d-L}\right)} && \text{(Lemma 9)} \\
&\leq 2\sqrt{t}\left(\lambda^{1/2}S_w + \sigma\sqrt{(d-L)\log\left(\frac{K(1+tS_x^2/\lambda))}{\delta}\right)}\right)\sqrt{(d-L)K\log\left(\lambda + \frac{tS_x^2}{(d-L)K}\right)}, \\
& && \text{(Jensen's Inequality \& } \textstyle\sum_{a\in[K]} N_t(a) = t)
\end{aligned}
$$

which concludes the proof.

□

F.3 OPTIMISM OF OFUL

The next result establishes that with high probability $U_a^T w_a^* \in C_{t,a}$, i.e., the true vector $U_a^T w_a^*$ is contained in the set $C_{t,a}$, which is a ball around the rotated PRR estimator $\hat{w}_{t,a}$. We prove this result via reduction of the PRR estimator to lower dimensional ridge regression (see Proposition 4).

Lemma 1 (Projected Subspace Optimism)**.** *Let $\hat{w}_{t+1,a}$ be the PRR estimator of w with $P_a = U_a U_a^T$ and $U_a \in \mathbb{R}^{d\times d-L}$. Let $\|P_a w_a^*\| \leq S_w, \|P_a x_n\| \leq S_x$. Then, for all $t \geq 0, a \in \mathcal{A}$ and for any $\delta > 0$ it holds that $U_a^T w_a^* \in C_{t,a}$ with probability greater than $1-\delta$, where*

$$C_{t,a} = \left\{w \in \mathbb{R}^d : \left\|U_a^T\hat{w}_{t+1,a} - U_a^T w\right\|_{\tilde{V}_{t,a}} \leq \sqrt{\beta_t(\delta)}\right\},$$

where $\tilde{V}_{t,a} = \lambda I_{d-L} + \sum_{n=1}^{t}\mathbb{1}_{\{a_n=a\}}(U_a^T x_n)(U_a^T x_n)^T$, and

$$\sqrt{\beta_t(\delta)} = \sigma\sqrt{(d-L)\log\left(\frac{K(1+tS_x^2/\lambda))}{\delta}\right)} + \lambda^{1/2}S_w.$$

Proof. Fix $a \in \mathcal{A}$. Since $\hat{w}_{t+1,a}$ is the smallest norm solution of the PRR, by proposition 4, $U_a^T\hat{w}_{t+1,a}$ has the following form

$$U_a^T\hat{w}_{t+1,a} = (\tilde{X}_{t,a}^T\tilde{X}_{t,a} + \lambda I_{d-L})^{-1}\tilde{X}_{t,a}^T Y_{t,a}. \tag{21}$$

$\tilde{X}_{t,a}$ is a matrix with $\{\tilde{x}_{i,a}\}_{i=1}^t$ in its rows where $\tilde{x}_{i,a} = \mathbb{1}_{\{a_i=a\}}(U_a^T x_n)^T$, and $Y_{t,a} = (y_{1,a}, .., y_{t,a})$ such that

$$y_{i,a} = \mathbb{1}_{\{a_i=a\}}\left(\left\langle U_a^T x_i, U_a^T w_a^*\right\rangle + \eta_i\right) = \left\langle \mathbb{1}_{\{a_i=a\}}U_a^T x_i, U_a^T w_a^*\right\rangle + \mathbb{1}_{\{a_i=a\}}\eta_i.$$

We now recast (21) such that we are able to apply Theorem 7 in a smaller dimension $d-L$. Since a_i, x_i, U_a are $\mathcal{F}_{i-1}$ measurable it holds that $\tilde{x}_{i,a}$ is $\mathcal{F}_{i-1}$ measurable. Define $\tilde{\eta}_i = \mathbb{1}_{\{a_i=a\}}\eta_i$. Since $\eta_i \in \mathcal{F}_n$ it holds that $\tilde{\eta}_i$ is $\mathcal{F}_i$ measurable. Furthermore, $\tilde{\eta}_i$ is σ sub Gaussian since η_i is σ sub Gaussian. Define $\tilde{Y}_{t,a} = (\tilde{y}_{1,a}, .., \tilde{y}_{t,a})$ where $\tilde{y}_{i,a} = \left\langle\tilde{x}_i, U_a^T w_a^*\right\rangle + \tilde{\eta}_i$. Thus, (21) can be written as

$$U_a^T\hat{w}_{t+1,a} = (\tilde{X}_{t,a}^T\tilde{X}_{t,a} + \lambda I_{d-L})^{-1}\tilde{X}_{t,a}^T\tilde{Y}_{t,a}.$$

We now employ Theorem 7 in dimension $d-L$ since $\tilde{x}_i, U_a^T w_a^* \in \mathbb{R}^{d-L}$. Furthermore, $\left\|U_a^T x_i\right\| = \|Px_i\| \leq S_x$ and similarly $\|Pw_a^*\| \leq S_w$. Taking a union bound on $a \in \mathcal{A}$ concludes the claim. □

Corollary 1 (Update Rule Optimism)**.** *Assume $U_a w_a^* \in C_{t,a}$ for any $t \geq 0$ and for all $a \in \mathcal{A}$. Then,*

$$\langle x_t, P_a \tilde{w}_{a,t} \rangle + \left\langle x_t, M_a^\dagger b_a \right\rangle \geq \langle x_t, w_a^* \rangle.$$

Thus, $\left\langle x_t, w^*_{a^*(x_t)} \right\rangle \leq \langle x_t, P_{a_t} \tilde{w}_{a_t,t} \rangle + \left\langle x_t, M^\dagger_{a_t} b_{a_t} \right\rangle$.

Proof. For all $a \in \mathcal{A}$, by the update rule definition

$$\begin{aligned}
&\langle x_t, P_a \tilde{w}_{a,t} \rangle + \left\langle x_t, M_a^\dagger b_a \right\rangle \\
&= \max_{U_a^T w \in C_{t-1,a}} \langle x_t, P_a w \rangle + \left\langle x_t, M_a^\dagger b_a \right\rangle \\
&= \max_{U_a^T w \in C_{t-1,a}} \left\langle U_a^T x_t, U_a^T w \right\rangle + \left\langle x_t, M_a^\dagger b_a \right\rangle \\
&\geq \left\langle U_a^T x_t, U_a^T w_a^* \right\rangle + \left\langle x_t, M_a^\dagger b_a \right\rangle && \text{(Lemma 1)} \\
&= \langle x_t, P_a w_a^* \rangle + \left\langle x_t, M_a^\dagger b_a \right\rangle = \langle x_t, w_a^* \rangle. && (P_a = U_a U_a^T \text{ \& Property 4})
\end{aligned}$$

Since the latter holds for all $a \in \mathcal{A}$ it also holds for the maximizer, i.e.,

$$\begin{aligned}
&\left\langle x_t, w_{a^*(x_t)} \right\rangle \leq \left\langle x_t, P_a \tilde{w}_{a^*(x_t),t} \right\rangle + \left\langle x_t, M^\dagger_{a^*(x_t)} b_{a^*(x_t)} \right\rangle \\
&\leq \max_{a \in \mathcal{A}} \langle x_t, P_a \tilde{w}_{a,t} \rangle + \left\langle x_t, M_a^\dagger b_a \right\rangle = \langle x_t, P_{a_t} \tilde{w}_{a_t,t} \rangle + \left\langle x_t, M^\dagger_{a_t} b_{a_t} \right\rangle.
\end{aligned}$$

□

APPENDIX G OFUL WITH LEARNED SUBSPACE

Algorithm 4 OFUL with Partially Observable Offline Data (Equivalent Form)

input: $\alpha > 0, M_a \in \mathbb{R}^{L\times d}, b_a \in \mathbb{R}^L, \delta > 0, T$

init: $V_a = \lambda I, Y_a = 0, \forall a \in \mathcal{A}$

for $n = 1, \ldots, \log_2(T) - 1$ **do**

  $\forall a \in \mathcal{A}$, estimate $\hat{M}_{n,a}$ and calculate $\hat{M}_{n,a}, \hat{P}_{n,a}$

  $\forall a \in \mathcal{A}\ V_a = \lambda I_d$ and $Y_a = 0$

  for $t = 0, \ldots, 2^n - 1$ **do**

    Receive context x_t

    $\hat{w}_{t,a} = \left(\hat{P}_{n,a} V_a \hat{P}_{n,a}\right)^\dagger \left(Y_a - (V_a - \lambda I)\, \hat{M}^\dagger_{n,a} b_a\right)$

    $\sqrt{\beta_t(\delta)} = \left(\sigma + S_x S_w (C_{B1}(\delta) + C_{B2}(\delta)\bar{t}_n^{-1/2})\right)\sqrt{(d-L)\log\left(\frac{2\log(T)K(1+tS_x^2/\lambda)}{\delta}\right)} + \lambda^{1/2} S_w$

    $a_t, \tilde{w}_{t,a_t} \in \arg\max_{a\in\mathcal{A}, w_a \in C_{t,a}} \left(\left\langle x_t, \hat{P}_{n,a} w_a \right\rangle + \left\langle x_t, \hat{M}^\dagger_{n,a} b_a \right\rangle\right)$

    Play action a_t and receive r_t

    $V_{a_t} = V_{a_t} + x_t x_t^T, Y_{a_t} = Y_{a_t} + x_t r_t$

  end for

end for

Before supplying the proof we define some useful notations. We denote $\hat{M}_{t,a}$ as the estimated matrix M_a at time step t (see (6)), $\hat{P}_{t,a}, \hat{M}^\dagger_{t,a}$ as the orthogonal projection on the kernel of $\hat{M}_{t,a}$ and the pseudo-inverse of $\hat{M}_{t,a}$, respectively. We also denote $\hat{P}_{t,a} = \hat{U}_{t,a}\hat{U}^T_{t,a}$, where $\hat{U}_{t,a} \in \mathbb{R}^{d-L\times d}$ as a matrix orthonormal vectors in its rows that span the kernel of $\hat{M}_{t,a}$. Let $\hat{b}_{t,a} = \hat{M}_{t,a} w^*_a$ be the result of the linear transformation of the true w^*_a by the estimated $\hat{M}_{t,a}$. Although this quantity is unknown it will be very useful in our analysis. Furthermore, by Property 4 it holds that $w^*_a = \hat{P}_{t,a} w^*_a + \hat{M}^\dagger_{t,a}\hat{b}_{t,a}$. Thus, for any x it holds that

$$\langle x, w^*_a\rangle = \left\langle x, \hat{P}_{t,a}\right\rangle + \left\langle x, \hat{M}^\dagger_{t,a}\hat{b}_{t,a}\right\rangle. \tag{22}$$

The proof follows the same line of analysis as in the exact case, i.e., of Theorem 1. Unlike in Theorem 1, we do not have access to the true matrices $M^\dagger_a, P_a$, but to increasingly more accurate estimates of these matrices.

To deal with this more challenging situation we use the doubling trick. The algorithm acts in exponentially increasing episodes. In each such episode, we fix the estimation of M, i.e., we use the estimate of M available in the beginning of the episode.

The analysis of this algorithm amounts to study the performance of the exact algorithm (as in Theorem 1) up to a fixed, approximated, M_a, which induces errors in the used $M^\dagger_a, P_a$. Finally, summing the regret on each episode, we obtain the final result.

The proof heavily relies on the convergence properties of $P_a, M^\dagger$. These are shown to converge at a rate of $O(T^{-1/2})$ in Appendix H. These convergence rates are due to the special structure of M_a, as provided by Proposition 1 and proven in Appendix D.

G.1 NOTATIONS AND THE GOOD EVENT OF THEOREM 2

Before supplying the proof, we define some notations. We define the good event $\mathcal{G}$ and prove it holds with high probability. We denote $n \in \{0, .., \log(T) - 1\}$ as the episode index, $\bar{t}_n = 2^n$ as the number of rounds performed at the beginning at the

start of the n^{th} episode, and

$$t \in \{0, .., 2^n - 1\} = \{0, .., \bar{t}_n - 1\} = \{0, .., \bar{t}_{n+1} - \bar{t}_n\},$$

as the number of rounds at the n^{th} episode.

Let

$$\Delta M_n(\delta) = \frac{C_{B1}(\delta)}{\sqrt{\bar{t}_n}} + \frac{C_{B2}(\delta)}{\bar{t}_n} = \frac{C_{B1}(\delta)}{\sqrt{2^n}} + \frac{C_{B2}(\delta)}{2^n}, \tag{23}$$

characterize the convergence of the estimation of M_a (see Corollary 2), and

$$C_{t,a,n}(\delta') = \left\{ w \in \mathbb{R}^d : \left\| \hat{U}_{a,n}^T \hat{w}_{t+1,a} - \hat{U}_{a,n}^T w \right\|_{\tilde{V}_{t,a}(\hat{U}_{a,n})} \leq \sqrt{\beta_{t,n}(\delta')} \right\},$$

where $\sqrt{\beta_{t,n}(\delta')}$ is defined in Lemma 4 and given by

$$\sqrt{\beta_{t,n}(\delta')} = (\sigma + S_x S_w C_n(\delta)) \sqrt{(d-L) \log\left(\frac{K(1 + tS_x^2/\lambda)}{\delta'} \right)} + \lambda^{1/2} S_w$$

where we used $\Delta M_n(\delta)\sqrt{t} \leq \Delta M_n(\delta)\sqrt{\bar{t}_n} \equiv C_n(\delta)$, since $t \leq \bar{t}_n$ and by the definition of ΔM_n.

Furthermore, denote $C_{t,a,n} \equiv C_{t,a,n}\left(\frac{\delta}{2\log(T)}\right)$, $\Delta M_n \equiv \Delta M_n\left(\frac{\delta}{2\log(T)}\right)$, and define the following failure events.

$$F_n^{CI} = \left\{ \exists t \in \{0, .., \bar{t}_{n+1} - \bar{t}_n\}, a \in \mathcal{A} : \hat{U}_{n,a}^T w_a^* \notin C_{t,a,n} \right\},$$
$$F_n^{M} = \left\{ \exists n \in [\log(T)], a \in \mathcal{A} : \left\| M_a - \hat{M}_{a,n} \right\| \geq \Delta M_n \right\},$$

where $\hat{M}_{a,n}$ is the estimated M matrix at the beginning of the n^{th} episode based on $\bar{t}_n$ samples gathered so far. Recall that $\hat{P}_{t,a} = \hat{U}_{n,a}\hat{U}_{n,a}^T$. The first event has to do with the rotated weights lying in the confidence ellipsoid $C_{t,a,n}$. The second event ensures the error in approximation of M_a is small enough, i.e., converges at a rate given by Equation (23).

Lemma 2 (Good Event Holds with High Probability)**.** *Let the bad event be* $\cup_{n=1}^{\log(T)} F_n^{CI} \cup_{n=1}^{\log(T)} F_n^{M}$ *and the good event,* $\mathcal{G}$*, its complement. Then,* $\Pr(\mathcal{G}) \geq 1 - \delta$.

Proof. We prove $\Pr\left(\cup_{n=1}^{\log(T)} F_n^{CI}\right) \leq \delta$ and $\Pr\left(\cup_{n=1}^{\log(T)} F_n^{M}\right) \leq \delta$. Then, applying the union bound and re-scaling $\delta \leftarrow \delta/2$ we conclude the proof.

The failure event $\cup_{n=1}^{\log(T)} F_n^{CI}$. Fix episode $n \in [\log(T)]$. Define the in-episode filtration $\{F_t^n\}_{t=0}^{\bar{t}_n}$, where $F_t^n = F_{\bar{t}_n + t}$ where $F_{\bar{t}_n+t}$ is the natural filtration (see Section 2). Observe that $\hat{M}_n$ is measurable w.r.t. to all $\{F_t^n\}_{t=0}^{\bar{t}_n}$ since it is determined at the beginning of the n^{th} episode. Thus, we can apply Lemma 4 which holds uniformly for all $t \in \{0, .., \bar{t}_n\}$, and get

$$\Pr\left(F_n^{CI}\right) \leq \delta'.$$

By taking a union bound on all $n \in [\log(T)]$ and setting $\delta' = \frac{\delta}{\log(T)}$ we get that $\Pr\left(\cup_{n=1}^{\log(T)} F_n^{CI}\right) \leq \delta$.

The failure event $\cup_{n=1}^{\log(T)} F_n^{M}$. By Corollary 2, for any fixed $\bar{t}_n \in [\log(T)]$ the event holds with high probability. Applying the union bound and re-scaling $\delta \leftarrow \frac{\delta}{\log(T)}$ establishes that $\Pr\left(\cup_{n=1}^{\log(T)} F_n^{M}\right) \leq \delta$. □

Remark 3 (The failure event $\cup_{n=1}^{\log(T)} F_n^{M}$)**.** *The probability for this failure event can be bounded using a stopping time argument, without resorting to applying a union bound. However, for brevity, and since it does not improve the final result by much (due to the union bound applied in the failure event* $\cup_{n=1}^{\log(T)} F_n^{CI}$*) we use simpler union bound arguments to prove this failure event does not occur with high probability.*

G.2 PROOF OF THEOREM 2

Proof. We start by conditioning on the good event $\mathcal{G}$. By Lemma 2 it occurs with probability greater than $1-\delta$.

Observe that the cumulative regret at round T (assuming $\log(T) \in \mathbb{N}$) is also given by the sum of episodic regret,

$$\textit{Regret}(T) = \sum_{n=0}^{\log_2(T)-1} \textit{Regret}_n, \tag{24}$$

where the episodic regret is given by

$$\textit{Regret}_n = \sum_{t=\bar{t}_n}^{\bar{t}_{n+1}-1} r_t \quad \text{where} \quad \bar{t}_n = 2^n.$$

We now bound the episodic regret for any $n \in [\log(T)]$. Let $t \in \{0, .., \bar{t}_n - 1\}$ be a time index of the n^{th} episode.

$$\begin{aligned}
r_t =& \left\langle x_t, w_{a^*(x_t)} \right\rangle - \langle x_t, w_{a_t} \rangle \\
&\leq \left\langle \hat{P}_{n,a_t} x_t, \tilde{w}_{a_t} \right\rangle + \left\langle x_t, \hat{M}^{\dagger}_{n,a_t} b_{a_t} \right\rangle + 2S_x S_w \Delta M_n - \langle x_t, w_{a_t} \rangle && \text{(Lemma 3)} \\
&= \left\langle \hat{P}_{n,a_t} x_t, \tilde{w}_{a_t} \right\rangle - \left\langle \hat{P}_{n,a_t} x_t, w_{a_t} \right\rangle + 2S_x S_w \Delta M_n + \left\langle x_t, \hat{M}^{\dagger}_{n,a_t} \Delta b_{n,a_t} \right\rangle. && \text{(By (22))}
\end{aligned}$$

Conditioning on the good event, the last term is bounded by

$$\left\langle x_t, \hat{M}^{\dagger}_{n,a_t} \Delta b_{a_t} \right\rangle \leq \|x_t\| \left\| \hat{M}^{\dagger}_{n,a_t} \right\| \left\| b_{a_t} - \hat{b}_{t,a_t} \right\| \leq S_x S_w \Delta M_n,$$

since $\|x\| \leq S$, $\left\| \hat{M}^{\dagger}_{n,a_t} \right\| \leq 1$ (By Lemma 7) and $\left\| b_{a_t} - \hat{b}_{n,a_t} \right\| \leq S_w \Delta M_n$ (By Lemma 6, error ΔM_n in the estimation of M).

Plugging this bound and setting $\hat{P}_{n,a} = \hat{U}_{n,a} \hat{U}^T_{n,a}$ we get

$$\begin{aligned}
r_t &\leq \left\langle \hat{U}^T_{n,a_t} x_t, \left(\hat{U}^T_{n,a_t} \tilde{w}_{a_t} - \hat{U}^T_{n,a_t} w_{a_t} \right) \right\rangle + 3S_x S_w \Delta M_n \\
&\leq \left\| \hat{U}^T_{n,a_t} x_t \right\|_{\tilde{V}_{t-1,a_t}(\hat{U}_n)^{-1}} \left\| \hat{U}^T_{n,a_t} \hat{w}_{a_t} - \hat{U}^T_{n,a_t} w_{a_t} \right\|_{\tilde{V}_{t-1,a_t}(\hat{U}_n)} \\
&\quad + \left\| \hat{U}^T_{n,a_t} x_t \right\|_{\tilde{V}_{t-1,a_t}(\hat{U}_n)^{-1}} \left\| \hat{U}^T_{n,a_t} \hat{w}_{a_t} - \hat{U}^T_{n,a_t} \tilde{w}_{a_t} \right\|_{\tilde{V}_{t-1,a_t}(\hat{U}_n)} + 3S_x S_w \Delta M_n && \text{(C.S. Inequality)} \\
&\leq 2\sqrt{\beta_{t-1}(\delta)} \left\| \hat{U}^T_{n,a_t} x_t \right\|_{\tilde{V}_{t-1,a_t}(\hat{U}_n)^{-1}} + 3S_x S_w \Delta M_n, && \text{(Lemma 4)}
\end{aligned}$$

where we define the PD matrix

$$\tilde{V}_{t-1,a}(\hat{U}_n) = \lambda I_{d-L} + \sum_{i=1}^{t-1} \mathbb{1}_{\{a_i = a\}} (\hat{U}^T_{n,a} x_i)(\hat{U}^T_{n,a} x_i)^T$$

Using the fact $r_t \leq 2$ since $\langle x_t, w_a \rangle \in [-1, 1]$ for any $a \in \mathcal{A}$, $\min(a+b, 1) \leq \min(a, 1) + \min(b, 1)$ and by the above we get

$$r_t \leq 2\sqrt{\beta_{t-1}(\delta)} \min\left(\left\| U^T_{a_t} x_t \right\|_{\tilde{V}^{-1}_{t-1,a_t}}, 1 \right) + 3S_x S_w \Delta M_n. \tag{25}$$

Combining the above and using Cauchy-Schwartz inequality (as in the proof of Theorem 1) we get

$$\begin{aligned}\textit{Regret}_n \leq & 2\sqrt{\bar{t}_n \sum_{i=0}^{\bar{t}_n-1} \beta_i(\delta) \min\left(\left\|\hat{U}_{n,a_i}^T x_i\right\|^2_{\tilde{V}_{i-1,a_i}(\hat{U}_n)^{-1}}, 1\right)} + 3S_x S_w \bar{t}_n \Delta M_n \\ \leq & 2\sqrt{\bar{t}_n \beta_{\bar{t}_{n+1}}(\delta) \sum_{i=0}^{\bar{t}_n-1} \min\left(\left\|\hat{U}_{n,a_i}^T x_i\right\|^2_{\tilde{V}_{i-1,a_i}(\hat{U}_n)^{-1}}, 1\right)} + 3S_x S_w \bar{t}_n \Delta M_n, \end{aligned} \tag{26}$$

where the last relation holds since $\beta_t(\delta)$ is increasing with t (see its definition (30)). We now bound the first term of (26) with similar technique as in the proof of Theorem 1.

Define $\tilde{x}_{i,a}^{(n)} = \mathbb{1}_{\{a_i=a\}} \hat{U}_{n,a} x_i$ and $V_{i,a}^{(n)} = \lambda I_{d-L} + \sum_{j=1}^{i} \tilde{x}_{i,a}^{(n)} \left(\tilde{x}_{i,a}^{(n)}\right)^T$. Importantly, since $\hat{U}_n$ is fixed for the entire n^{th} episode it holds that

$$V_{i,a}^{(n)} = V_{i-1,a}^{(n)} + \tilde{x}_{i,a}^{(n)} \left(\tilde{x}_{i,a}^{(n)}\right)^T. \tag{27}$$

Furthermore, denote by $I_k(n,a)$ the k^{th} time action a was chosen at the n^{th} episode,

$$I_k(a,n) = \min\left\{ t \in \{0,..,\bar{t}_n - 1\} : \sum_{j=0}^{t} \mathbb{1}_{\{a_j=a\}} = k \right\},$$

and denote by $N_{\bar{t}_n-1}(a)$ the total number of times action a was chosen by the end on the n^{th} episode. By these definitions the following relations hold.

$$\begin{aligned} & \sum_{i=0}^{\bar{t}_n-1} \min\left(\left\|\hat{U}_{n,a_i}^T x_i\right\|^2_{\tilde{V}_{i-1,a_i}(\hat{U}_n)^{-1}}, 1\right) \\ & = \sum_{a\in\mathcal{A}} \sum_{i=0}^{N_{\bar{t}_n-1}(a)} \min\left(\left\|\tilde{x}^{(n)}_{I_i(a,n),a}\right\|^2_{\left(V^{(n)}_{i-1,a}\right)^{-1}}, 1\right) \\ & \leq 2(d-L) \sum_{a\in\mathcal{A}} \log\left(\lambda + \frac{N_{\bar{t}_n-1}(a) S_x^2}{d-L}\right) && \text{(Lemma 9)} \\ & \leq 2(d-L)K \log\left(\lambda + \frac{(\bar{t}_n - 1)S_x^2}{d-L}\right). && \text{(Jensen's Ineq. \& } \textstyle\sum_a N_{\bar{t}_n-1}(a) = \bar{t}_n - 1) \end{aligned}$$

Plugging this back into (26), denote $\delta_T = \delta/(2\log(T))$, we bound $\textit{Regret}_n$ as follows.

$$\begin{aligned} & \textit{Regret}_n \leq (26) \leq 2\sqrt{\bar{t}_n \beta_{\bar{t}_{n+1}}(\delta_T)(d-L)K \log\left(\lambda + \frac{\bar{t}_n S_x^2}{d-L}\right)} + 3S_x S_w \bar{t}_n \Delta M_n \\ & \leq 2\left(\left(\sigma + S_x S_w \Delta M_n \sqrt{\bar{t}_n}\right)\sqrt{(d-L)\log\left(\frac{K(1+\bar{t}_n S_x^2/\lambda)}{\delta_T}\right)} + \lambda^{1/2} S_w\right)\sqrt{\bar{t}_n (d-L)K \log\left(\lambda + \frac{\bar{t}_n S_x^2}{d-L}\right)} \\ & + 3S_x S_w \bar{t}_n \Delta M_n \\ & \leq 2\left((\sigma + S_x S_w C_{B1}(\delta_T))\sqrt{(d-L)\log\left(\frac{K(1+\bar{t}_n S_x^2/\lambda)}{\delta_T}\right)} + \lambda^{1/2} S_w\right)\sqrt{\bar{t}_n (d-L)K \log\left(\lambda + \frac{\bar{t}_n S_x^2}{d-L}\right)} \\ & \quad + 2\left(S_x S_w C_{B2}(\delta_T) \bar{t}_n^{-1/2} \sqrt{(d-L)\log\left(\frac{K(1+\bar{t}_n S_x^2/\lambda)}{\delta_T}\right)} + \lambda^{1/2} S_w\right)\sqrt{(d-L)K \log\left(\lambda + \frac{\bar{t}_n S_x^2}{d-L}\right)} \\ & \quad + 3S_x S_w C_{B1}(\delta_T)\sqrt{\bar{t}_n} + 3S_x S_w C_{B2}(\delta_T). && \left(\Delta M_n = \tfrac{C_{B1}(\delta_T)}{\sqrt{\bar{t}_n}} + \tfrac{C_{B2}(\delta_T)}{\bar{t}_n} \text{ conditioned on } \mathcal{G} \;\&\; \tfrac{t}{\bar{t}_n} \leq 1\right) \end{aligned}$$

Finally, using

$$\sum_{n=0}^{\log_2(T)-1}\sqrt{\bar{t}_n}=\sum_{n=0}^{\log_2(T)-1}2^{n/2}\leq\sqrt{2}2^{\log_2(T)/2}=\sqrt{2T},$$
$$\sum_{n=0}^{\log_2(T)-1}1\leq\log_2(T),\quad\text{and}\quad\sum_{n=0}^{\log_2(T)-1}\bar{t}_n^{1/2}\leq 3\sqrt{2},$$

we bound the regret by

$$\begin{aligned}
\textit{Regret}(T)&\leq\sum_{n=0}^{\log_2(T)-1}\textit{Regret}_n\\
&\leq 3\left((\sigma+S_xS_wC_{B1}(\delta_T))\sqrt{(d-L)\log\left(\frac{K(1+tS_x^2/\lambda)}{\delta_T}\right)}+\lambda^{1/2}S_w\right)\sqrt{T(d-L)K\log\left(\lambda+\frac{TS_x^2}{d-L}\right)}\\
&\quad+3S_xS_wC_{B1}(\delta_T)\sqrt{T}+3\sqrt{2}S_xS_wC_{B2}(\delta_T)(d-L)\sqrt{K\log\left(\lambda+\frac{TS_x^2}{d-L}\right)\log\left(\frac{K(1+tS_x^2/\lambda)}{\delta_T}\right)}\\
&\quad+3S_xS_wC_{B2}(\delta_T).
\end{aligned}$$

□

Remark 4 (Why use the Doubling Trick?)**.** *Importantly, since $\hat{U}_{n,a}$ is fixed for all $a\in\mathcal{A}$ for the entire n^{th} episode we can apply the elliptical potential lemma 9, as* (27) *holds. If we would change the estimate of $\hat{P}=\hat{U}\hat{U}^T$ at every round, a relation such as* (27) *would not hold. We believe this problem might be alleviated by combining optimism w.r.t. the approximated subspace. We leave such study to future work.*

Lemma 3 (Update Rule Approximate Optimism)**.** *Let $\left\|M_a-\hat{M}_{t,a}\right\|\leq\Delta M$. Then, conditioning on the good event,*

$$\left\langle x_t,w^*_{a^*(x_t)}\right\rangle\leq\left\langle x_t,\hat{P}_{t,a_t}\tilde{w}_{a_t}\right\rangle+\left\langle x_t,\hat{M}^\dagger_{t,a_t}b_{a_t}\right\rangle+2S_xS_w\Delta M.$$

Proof. Conditioning on the good event, for all $a\in\mathcal{A}$ and $t\geq 1$ it holds that

$$\begin{aligned}
\left\langle x_t,\hat{P}_{t,a}\tilde{w}_{t,a}\right\rangle&=\left\langle\hat{U}^T_{t,a}x_t,\hat{U}^T_{t,a}\tilde{w}_{t,a}\right\rangle\\
&=\max_{w\in C_{t-1,a}}\left\langle\hat{U}^T_{t,a}x_t,\hat{U}^T_{t,a}\tilde{w}_{t,a}\right\rangle\\
&\geq\left\langle\hat{U}^T_{t,a}x_t,\hat{U}^T_{t,a}w^*_a\right\rangle&\text{(Conditioning on }\mathcal{G}\text{, Lemma 4)}\\
&=\left\langle x_t,\hat{P}^T_{t,a}w^*_a\right\rangle.&(28)
\end{aligned}$$

Applying this, we get the following relations hold for all $a\in\mathcal{A}$.

$$\begin{aligned}
&\left\langle x_t,\hat{P}_{t,a}\tilde{w}_{t,a}\right\rangle+\left\langle x_t,\hat{M}^\dagger_{t,a}b_a\right\rangle\geq\left\langle x_t,\hat{P}^T_{t,a}w^*_a\right\rangle+\left\langle x_t,\hat{M}^\dagger_{t,a}b_a\right\rangle&\text{(By (28))}\\
&=\left\langle x_t,\hat{P}^T_{t,a}w^*_a\right\rangle+\left\langle x_t,\hat{M}^\dagger_{t,a}\hat{b}_{t,a}\right\rangle+\left\langle x_t,\hat{M}^\dagger_{t,a}(b_a-\hat{b}_{t,a})\right\rangle\\
&=\langle x_t,w^*_a\rangle+\left\langle x_t,\hat{M}^\dagger_{t,a}(b_a-\hat{b}_{t,a})\right\rangle&\text{(By (22))}\\
&\geq\langle x_t,w^*_a\rangle-\|x_t\|\left\|\hat{M}^\dagger_{t,a}\right\|\left\|b_a-\hat{b}_{t,a}\right\|&\text{(C.S. Inequality)}\\
&\geq\langle x_t,w^*_a\rangle-S_xS_w\Delta M,&(29)
\end{aligned}$$

where the last relation holds by bounding $\|x_t\|\leq S_x$, $\left\|\hat{M}^\dagger_{t,a}\right\|\leq 1$ (Lemma 7) and the bound on $\left\|b_a-\hat{b}_{t,a}\right\|$ follows from Lemma 6, third claim.

Thus,

$$\begin{aligned}
&\Big\langle x_t, w^*_{a^*(x_t)} \Big\rangle = \max_a \langle x_t, w^*_a \rangle \\
&\leq \max_a \Big(\Big\langle x_t, \hat{P}_{t,a} \tilde{w}_{t,a} \Big\rangle + \Big\langle x_t, \hat{M}^\dagger_{t,a} b_a \Big\rangle \Big) + 2 S_x S_w \Delta M \\
&= \Big\langle x_t, \hat{P}_{t,a_t} \tilde{w}_{t,a_t} \Big\rangle + \Big\langle x_t, \hat{M}^\dagger_{t,a_t} b_{a_t} \Big\rangle + 2 S_x S_w \Delta M.
\end{aligned}$$

□

G.3 OPTIMISM OF OFUL WITH LEARNED SUBSPACE

Lemma 4 (Projected Subspace Optimism with Subspace Error)**.** *Assume $\hat{M}_a$ is measurable w.r.t. the filtration $\{F_t\}_{t=0}^\infty$. Assume for all $a \in \mathcal{A}$ $\left\|\hat{M}_a - M_a\right\| \leq \Delta M$, let $\hat{P}_a$ be the orthogonal projection on the kernel of $\hat{M}_a$ and $\hat{M}_a^\dagger$ its psuedo-inverse. Let $\hat{w}_{t+1,a}$ be the PRR estimator w.r.t. the projection matrix $\hat{P}_a$ (see Eq. (19)). Let $\hat{P}_a = \hat{U}_a \hat{U}_a^T$ where $\hat{U}_a \in \mathbb{R}^{d \times d-L}$. Assume $\left\|\hat{P}_a w_a^*\right\| \leq S_w, \left\|\hat{P}_a x_n\right\| \leq S$. Then, for all $t > 0, a \in \mathcal{A}$ and for any $\delta \in (0,1)$ it holds that $\hat{U}_a^T w_a^* \in C_{t,a}$ with probability greater than $1-\delta$, where*

$$C_{t,a} = \left\{ w \in \mathbb{R}^d : \left\| \hat{U}_a^T \hat{w}_{t+1,a} - \hat{U}_a^T w \right\|_{\tilde{V}_{t,a}(\hat{U})} \leq \sqrt{\beta_t(\delta)} \right\},$$

where $\tilde{V}_{t,a}(\hat{U}) = \lambda I_{d-L} + \sum_{i=1}^t \mathbb{1}_{\{a_i=a\}} (\hat{U}_a^T x_i)(\hat{U}_a^T x_i)^T$, and

$$\sqrt{\beta_t(\delta)} = \left(\sigma + S_x S_w \Delta M \sqrt{t}\right) \sqrt{(d-L) \log\left(\frac{K(1+tS_x^2/\lambda)}{\delta}\right)} + \lambda^{1/2} S_w. \tag{30}$$

Proof. Fix $a \in [K]$. Remember that $y_{i,a}$ is given by

$$y_{i,a} = \mathbb{1}_{\{a_i=a\}} \Big(\langle x_i, w_a \rangle + \eta_i - \Big\langle x_i, \hat{M}_a^\dagger b_a \Big\rangle \Big).$$

Plugging $\langle x_i, w_a \rangle = \Big\langle x_i, \hat{P}_a w_a \Big\rangle + \Big\langle x_i, \hat{M}_a^\dagger \hat{b}_a \Big\rangle$ (see Property 4) setting $\hat{P}_a = \hat{U}_a \hat{U}_a^T$ we get

$$y_{i,a} = \mathbb{1}_{\{a_i=a\}} \Big(\Big\langle \hat{U}_a^T x_i, \hat{U}_a^T w_a \Big\rangle + \eta_i - \Big\langle x_i, \hat{M}_a^\dagger \Delta b_a \Big\rangle \Big), \tag{31}$$

where $\Delta b_a = b_a - \hat{b}_a$.

Let $\tilde{X}_{t,a} \in \mathbb{R}^{d-L \times t}$ be the matrix with $\left\{ \mathbb{1}_{\{a_i=a\}} \hat{U}_a x_i \right\}_{i=1}^t$ in its rows, and let $X_{t,a} \in \mathbb{R}^{d \times t}$ be the matrix with $\left\{ \mathbb{1}_{\{a_i=a\}} x_i \right\}_{i=1}^t$ in its rows. The PRR estimator is thus given by

$$\begin{aligned}
&\hat{U}_a^T w_{t+1,a} \\
&= (\tilde{V}_{t,a})^{-1} \tilde{X}_{t,a}^T \tilde{X}_{t,a} \hat{U}_{t,a}^T w_a + (\tilde{V}_{t,a})^{-1} \tilde{X}_{t,a}^T \tilde{\eta}_t + (\tilde{V}_{t,a})^{-1} \tilde{X}_{t,a}^T X_{t,a} \hat{M}_{t,a}^\dagger \Delta b_{t,a} \\
&= \hat{U}_{t,a}^T w_a - \lambda (\tilde{V}_{t,a})^{-1} \hat{U}_{t,a}^T w_a + (\tilde{V}_{t,a})^{-1} \tilde{X}_{t,a}^T \tilde{\eta}_t + (\tilde{V}_{t,a})^{-1} \tilde{X}_{t,a}^T X_{t,a} \hat{M}_{t,a}^\dagger \Delta b_{t,a}.
\end{aligned}$$

Rearranging and multiplying by x both sides we get

$$\begin{aligned}
&x^T \left(\hat{U}_a^T w_{t+1,a} - \hat{U}_a^T w_a \right) \\
&= -\lambda x^T (\tilde{V}_{t,a})^{-1} \hat{U}_a^T w_a + x^T (\tilde{V}_{t,a})^{-1} \tilde{X}_{t,a}^T \tilde{\eta}_t + x^T (\tilde{V}_{t,a})^{-1} \tilde{X}_{t,a}^T X_{t,a} \hat{M}_a^\dagger \Delta b_a \\
&\leq \|x\|_{(\tilde{V}_{t,a})^{-1}} \left(\lambda \left\| \hat{U}_a^T w_a \right\|_{(\tilde{V}_{t,a})^{-1}} + \left\| \tilde{X}_{t,a}^T \tilde{\eta}_t \right\|_{(\tilde{V}_{t,a})^{-1}} + \left\| \tilde{X}_{t,a}^T X_{t,a} \hat{M}_a^\dagger \Delta b_a \right\|_{(\tilde{V}_{t,a})^{-1}} \right)
\end{aligned}$$

Setting $x = \tilde{V}_{t,a}\left(\hat{U}_a^T w_{t+1,a} - \hat{U}_a^T w_a\right)$, which implies that

$$\|x\|_{\tilde{V}_{t,a}^{-1}} = \left\|\left(\hat{U}_a^T w_{t+1,a} - \hat{U}_a^T w_a\right)\right\|_{\tilde{V}_{t,a}},$$

and dividing both sides by $\left\|\left(\hat{U}_a^T w_{t+1,a} - \hat{U}_a^T w_a\right)\right\|_{\tilde{V}_{t,a}}$ we get

$$\left\|\left(\hat{U}_a^T w_{t+1,a} - \hat{U}_{t,a}^T w_a\right)\right\|_{\tilde{V}_{t,a}} \leq \lambda \left\|\hat{U}_a^T w_a\right\|_{(\tilde{V}_{t,a})^{-1}} + \left\|\tilde{X}_{t,a}^T \tilde{\eta}_t\right\|_{(\tilde{V}_{t,a})^{-1}} + \left\|\tilde{X}_{t,a}^T X_{t,a} \hat{M}_a^\dagger \Delta b_{t,a}\right\|_{(\tilde{V}_{t,a})^{-1}}. \tag{32}$$

The first term of (32) is bound by

$$\lambda \left\|\hat{U}_a^T w_a\right\|_{(\tilde{V}_{t,a})^{-1}} \leq \lambda^{1/2} \left\|\hat{U}_a^T w_a\right\| = \lambda^{1/2} \left\|\hat{P}_a^T w_a\right\| \leq \lambda^{1/2} S_w.$$

The second term of (32) bound by applying Theorem 6 and Lemma 9,

$$\left\|\tilde{X}_{t,a}^T \tilde{\eta}_t\right\|_{(\tilde{V}_{t,a})^{-1}} \leq \sigma \sqrt{(d-L)\log\left(\frac{1 + tS_x^2/\lambda}{\delta}\right)}$$

Theorem 6 is applicable by verifying its assumptions. First, $\tilde{X}_{t,a}$ is a matrix with $\mathbb{1}_{\{a_i=a\}}\hat{U}_a x_i \in \mathbb{R}^{d-L}$ in its rows (which are F_{t-1} measurable by the fact $\hat{U}_a, x_i, \mathbb{1}_{\{a_i=a\}}$ are F_{t-1} measurable). The vector $\tilde{\eta}_t$ is a vector with $\mathbb{1}_{\{a_i=a\}}\eta_i$ in its entries. Since η_i is F_{t-1} measurable and η_i is F_t measurable, η_i is F_t measurable. Furthermore, it is easy to verify that $\mathbb{1}_{\{a_i=a\}}\eta_i$ is conditionally σ-sub-Gaussian w.r.t. F_{t-1}.

Lastly, the third term of (32) is bounded by applying the elliptical potential lemma and the assumption $\left\|M_a - \hat{M}_a\right\| \leq \Delta M$ which implies by Lemma 6

$$\Delta b_a \leq S_w \Delta M. \tag{33}$$

We have that

$$\begin{aligned}
&\left\|\tilde{X}_{t,a}^T X_{t,a} \hat{M}_a^\dagger \Delta b_a\right\|_{(\tilde{V}_{t,a})^{-1}} = \left\|(\tilde{V}_{t,a})^{-1/2} \tilde{X}_{t,a}^T X_{t,a} \hat{M}_a^\dagger \Delta b_a\right\| \\
&\leq \left\|(\tilde{V}_{t,a})^{-1/2} \tilde{X}_{t,a}^T X_{t,a}\right\| \left\|\hat{M}_a^\dagger\right\| \|\Delta b_a\| && \text{(Norm is submultiplicative)} \\
&\leq \left\|(\tilde{V}_{t,a})^{-1/2} \tilde{X}_{t,a}^T X_{t,a}\right\| \|\Delta b_a\| && (\left\|M_a^\dagger\right\| \leq 1, \text{Lemma 7}) \\
&\leq S_w \Delta M \left\|(\tilde{V}_{t,a})^{-1/2} \tilde{X}_{t,a}^T X_{t,a}\right\| && \text{(By (33))} \\
&= S_w \Delta M \left\|\tilde{X}_{t,a}^T X_{t,a}\right\|_{(\tilde{V}_{t,a})^{-1}}.
\end{aligned}$$

By Lemma 5 we have

$$\left\|\tilde{X}_{t,a}^T X_{t,a}\right\|_{(\tilde{V}_{t,a})^{-1}} \leq S_x \sqrt{t(d-L)\log\left(1 + S_x^2 t/\lambda\right)}.$$

From which we get that the third term of (32) is also bounded by ($\delta \in (0,1)$)

$$\left\|\tilde{X}_{t,a}^T X_{t,a} \hat{M}_{t,a}^\dagger \Delta b_{t,a}\right\|_{(\tilde{V}_{t,a})^{-1}} \leq S_x S_w \Delta M \sqrt{t} \sqrt{(d-L)\log\left(\frac{1 + S_x^2 t/\lambda}{\delta}\right)}.$$

Combining the above and taking union bound on $a \in [K]$. □

The following lemma is based on Lemma 13 of Lale et al. 2019, and relies on Lemma 10.

Lemma 5 (Deterministic Bound on Cumulative Visitation)**.**

$$\left\|\tilde{X}_{t,a}^T X_{t,a}\right\|_{(\tilde{V}_{t,a})^{-1}} \leq S_x\sqrt{t}\sqrt{(d-L)\log\left(1+tS_x^2/\lambda\right)}.$$

Proof. The following relations hold.

$$\begin{aligned}
\left\|\tilde{X}_{t,a}^T X_{t,a}\right\|_{(\tilde{V}_{t,a})^{-1}} &= \left\|\sum_{i=1}^{t}(\tilde{V}_{t,a})^{-1/2}\tilde{x}_{i,a}^{(t)}x_i^T\right\| \\
&\leq \sum_{i=1}^{t}\left\|(\tilde{V}_{t,a})^{-1/2}\tilde{x}_{i,a}^{(t)}x_i^T\right\| && \text{(Triangle Inequality)} \\
&\leq S_x\sum_{i=1}^{t}\left\|(\tilde{V}_{t,a})^{-1/2}\tilde{x}_{i,a}^{(t)}\right\| && \text{(Norm is submultiplicative, \& } \|x\| \leq S_x\text{)} \\
&\leq S_x\sum_{i=1}^{t}\left\|(\tilde{V}_{i,a})^{-1/2}x_{i,a}^{(t)}\right\| && ((\tilde{V}_{i,a})^{-1} \succeq (\tilde{V}_{j,a})^{-1/2} \text{ for } j \geq i) \\
&\leq S_x\sqrt{t}\sqrt{\sum_{i=1}^{t}\left\|\tilde{x}_{i,a}^{(t)}\right\|_{(\tilde{V}_{i,a})^{-1}}^2} && \text{(C.S. Inequality)} \\
&\leq S_x\sqrt{t}\sqrt{(d-L)\log\left(1+tS_x^2/\lambda\right)},, && \text{(Lemma 10)}
\end{aligned}$$

where Lemma 10 is applied with $d-L$ (the dimension of the vectors $\tilde{x}_{t,a}^{(t)}$). This concludes the proof. □

APPENDIX H CONVERGENCE OF $M, M^\dagger, P$

Proposition 1 establishes that from a partially observable data one is able to obtain b_a^{LS} which is related to w_a^* through the following linear transformation

$$b_a^{LS} = M_a w_a^* \text{ where } M_a = \left(I_L, \quad R_{11}^{-1}(a)R_{12}(a)\right).$$

Although we cannot recover w_a^* from this relation we can recover $(I-P)w^*a = M_a^\dagger b_a^{LS}$, i.e., the projection of w^*a on the row space spanned by M_a is $M_a^\dagger b_a^{LS}$. Unfortunately, M_a itself depends on statistics of x^{h}, $R_{12}(a)$, which does not exist in the offline data. For brevity, we denote $b_a = b_a^{LS}$.

In this section we supply finite sample guarantees on the estimation of M_a based on samples. First, observe that the only unknown part of M_a is $R_{12}(a)$ (since $R_{11}(a)$ can be evaluated from the offline data). Thus, estimating M_a is reduced to equivalent to estimating $R_{12}(a)$, i.e., estimating a sub-matrix of the full covariance matrix.

We assume access to t samples of the form $\{x_i, a_i\}_{i=1}^t$ where $x_i \sim \mathcal{P}_x$ and $a_i \sim \pi_b(\cdot \mid x_i)$. Using this data, which can be gathered in an online manner, we prove finite convergence guarantees for an unbiased estimate of $R_{12}(a)$, i.e.,

$$\hat{R}_{12,t}(a) = \frac{1}{t-1}\sum_{i=1}^{t} \frac{\mathbb{1}_{\{a_i=a\}}}{P^{\pi_b}(a)} x_i^{\mathrm{o}} \left(x_i^{\mathrm{h}}\right)^T,$$

for $t \geq 2$. Notice that indeed $\mathbb{E}\left[\hat{R}_{12,t}(a)\right] = R_{12}(a) = \mathbb{E}^{\pi_b}\left[x^{\mathrm{o}}\left(x^{\mathrm{h}}\right)^T \mid a\right]$.

Our estimator for $R_{12}(a)$ given t samples is then given by

$$\hat{R}_{12,t}(a) = \hat{\Sigma}_{12}(a) + \mu_{1|a}\hat{\mu}_{\mathrm{h}|a}^T,$$

and, naturally, the estimator for M_a given t samples is then

$$\hat{M}_{t,a} = \left(I_L \quad R_{11}(a)^{-1}\hat{R}_{12,t}(a).\right)$$

Our approach requires access to $M_a^\dagger$ and P_a (defined as the orthogonal projection on the kernel of M_a). We use the plug-in estimator to obtain them both from the empirical estimator of M_a. Meaning

$$\hat{M}_{t,a}^\dagger = \hat{M}_{t,a}^T(\hat{M}_{t,a}\hat{M}_{t,a}^T)^{-1} \quad \text{and} \quad \hat{P}_{t,a} = I - \hat{M}_{t,a}^\dagger \hat{M}_{t,a}.$$

To establish finite sample convergence guarantees for $\hat{M}_{t,a}^\dagger$ and $\hat{P}_{t,a}$ we need to use important properties (see Lemma 7) of $\hat{M}_{t,a}$ and M_a, which holds due to their very special structure,

$$\operatorname{rank}\left(\hat{M}_t\right) = \operatorname{rank}(M) = L, \text{ and } \left\|M_a^\dagger\right\|, \left\|\hat{M}_{t,a}^\dagger\right\| \leq 1.$$

These properties are crucial to derive the convergence of the plug-in estimator of P_a and $M_a^\dagger$ from the convergence of M_a [Wedin, 1973].

In Corollary 2 we characterize the finite-sample convergence of the estimates of M_a. The following lemma shows that approximation errors of M_a leads to well controlled approximation errors in the approximations of $P_a, M_a^\dagger$ and b_a as a result of the special structure of M_a.

Lemma 6 (Deconfounder Matrix Error Propagation). *Denote by $\left\|M_a - \hat{M}_{t,a}\right\|$ as the estimation error of $\hat{M}_{t,a}$ relatively to M_a. Then,*

1. $\left\|P_a - \hat{P}_{t,a}\right\| \leq 2\left\|M_a - \hat{M}_{t,a}\right\|$.
2. $\left\|M_a^\dagger - \hat{M}_{t,a}^\dagger\right\| \leq 2\left\|M_a - \hat{M}_{t,a}\right\|$.
3. *Assuming* $\|w_a^*\| \leq S_w$, $\left\|b_a - \hat{b}_{t,a}\right\| \leq R\left\|M_a - \hat{M}_{t,a}\right\|$.

Proof. Claim (1). The second claim is a direct application of Theorem 3, which requires that $\operatorname{rank}\left(\hat{M}_{t,a}\right) = \operatorname{rank}(M_a)$. Indeed, by the first claim of Lemma 7 this condition is satisfies (for any t and a).

Claim (2). The third claim follows by applying Theorem 4, by which

$$\left\|M_a^\dagger - \hat{M}_{t,a}^\dagger\right\| \leq 2\max\left\{\left\|M_a^\dagger\right\|^2, \left\|\hat{M}_{t,a}^\dagger\right\|^2\right\}\left\|M_a - \hat{M}_{t,a}\right\|.$$

Since both matrices $M_a, \hat{M}_{t,a}$ are of the form $\begin{pmatrix} I_L & B \end{pmatrix}$, for some B, by the second claim of Lemma 7 it holds that $\left\|M_a^\dagger\right\| \leq 1, \left\|\hat{M}_{t,a}^\dagger\right\| \leq 1$ which implies that $\left\|M_a^\dagger - \hat{M}_{t,a}^\dagger\right\| \leq 2\left\|M_a - \hat{M}_{t,a}\right\|$.

Claim (3). Denote $b_a = b_a^{LS}$. Observe that

$$M_a w_a^* = b_a$$
$$\hat{M}_{t,a} w_a^* = \hat{b}_{t,a}.$$

Decreasing the two equations and taking the L_2 norm we get

$$\left\|b_a - \hat{b}_{t,a}\right\| = \left\|(M_a - \hat{M}_{t,a})w_a^*\right\| \leq \left\|M_a - \hat{M}_{t,a}\right\|\|w_a^*\| \leq S_w\left\|M_a - \hat{M}_{t,a}\right\|.$$

□

Lemma 7 (Properties of M)**.** *Let $L \leq d$ and let $M \in \mathbb{R}^{L\times d}$ be the matrix defined by*

$$M = \begin{pmatrix} I_L & B \end{pmatrix},$$

where $B \in \mathbb{R}^{L\times(d-L)}$. Then, the following claims hold for any B.

1. $\operatorname{rank}(M) = L$.
2. $\left\|M^\dagger\right\| \leq 1$.

Proof. We prove that M has L non-zero singular values $\{\sigma_i\}_{i=1}^L$ such that $\sigma_i \geq 1$ for all $i \in [L]$. This follows by lower bounding the minimal eigenvalue of MM^T. We show it is lower bounded by 1. We have that

$$\begin{aligned}\lambda_{\min}(MM^T) &= \min_{x\in\mathbb{R}^L:\|x\|=1}\left(x^T MM^T x\right)\\ &= \min_{x\in\mathbb{R}^L:\|x\|=1}\left(\|x\| + x^T BB^T x\right)\\ &= 1 + \min_{x\in\mathbb{R}^L:\|x\|=1}\left(x^T BB^T x\right) \geq 1,\end{aligned}$$

since $x^T BB^T x = \left\|B^T x\right\|^2 \geq 0$ for any x. Thus, $\lambda_{\min}(MM^T) \geq 1$ which implies that $MM^T \in \mathbb{R}^{L\times L}$ has L eigenvalues $\{\lambda_i\}_{i=1}^L$ greater than 1. The latter implies that M has exactly L non-zero singular-values, $\{\sigma_i\}_{i=1}^L$, greater than 1, since $\sigma_i = \sqrt{\lambda_i} \geq 1$.

Claim (1). Since M has L non-zero singular values, the rank of M is L, since the rank of M is also the total number of non-zero singular values.

Claim (2). Let $M = U\Sigma V^T$ be the SVD decomposition of M. Observe that the pseudo-inverse of M is also given by $M^\dagger = U\Sigma^+ V^T$ where $(\Sigma^+)_{ii} = \frac{1}{\sigma_i}$ for non-zero σ_i and zero otherwise. By the first claim $\sigma_i \geq 1$ for all non-zero σ_i, which implies that $\left\|M^\dagger\right\| = \max_{i:\sigma_i\neq 0}\frac{1}{\sigma_i} \leq 1$. □

Lemma 8 (Masked Cross Correlation Estimation)**.** *Let x, y be random vectors in $R^{d_1}, \mathbb{R}^{d_2}$, respectively, with $d_1, d_2 \geq 2$. Assume that for some $S_1, S_2 \geq 1$*

$$\|x\|_2 \leq S_1 \text{ and } \|y\|_2 \leq S_2 \text{ almost surely}$$

Denote $R = \mathbb{E}\left[xy^T\right], R_x = \mathbb{E}\left[xx^T\right], R_y = \mathbb{E}\left[yy^T\right]$. *For any* $t \geq 1$ *define* $\hat{R}_t = \frac{1}{t}\sum_{i=1}^{t} x_i y_i^T$. *Then with probability at least* $1-\delta$

$$\left\|\hat{R}_t - R\right\|_2 \leq S_1 S_2 \left(\sqrt{\frac{2}{t}\left(\frac{\sqrt{trace\left(R_x\right) trace\left(R_y\right)}}{S_1 S_2}\right)\log\left(\frac{d_1+d_2}{\delta}\right)} + \frac{4}{3t}\log\left(\frac{d_1+d_2}{\delta}\right)\right),$$

Proof. Denote $A_i = x_i y_i - R$ and notice that $\mathbb{E}\left[\frac{1}{t}\left(A_i - R\right)\right] = 0$. Then applying Lemma 11 we have that with probability at least $1-\delta$

$$\begin{aligned}\left\|\hat{R}_t - R\right\|_2 &= \left\|\sum_{i=1}^{t} A_i\right\|_2 \\ &\leq \sqrt{\frac{2V}{t}\log\left(\frac{d_1+d_2}{\delta}\right)} + \frac{2}{3t}C\log\left(\frac{d_1+d_2}{\delta}\right),\end{aligned}$$

where

$$V = \max\left\{\left\|\mathbb{E}\left[\left(x_i y_i^T - R\right)\left(x_i y_i^T - R\right)^T\right]\right\|_2, \left\|\mathbb{E}\left[\left(x_i y_i^T - R\right)^T\left(x_i y_i^T - R\right)\right]\right\|_2\right\}$$

and C is a constant chosen such that $\left\|x_i y_i^T - R\right\|_2 \leq C, a.s.$

We start by bounding V and next bounding C. We have that

$$\begin{aligned}\mathbb{E}\left[\left(x_i y_i^T - R\right)\left(x_i y_i^T - R\right)^T\right] &= \mathbb{E}\left[\left(x_i y_i^T - R\right) x_i y_i^T\right] \\ &= \mathbb{E}\left[x_i x_i^T \left\|y_i\right\|_2^2\right] - RR^T \\ &\preceq \mathbb{E}\left[x_i x_i^T \left\|y_i\right\|_2^2\right].\end{aligned}$$

Then, using the fact that $\mathbb{E}\left[\left(x_i y_i^T - R\right)\left(x_i y_i^T - R\right)^T\right]$ and $\mathbb{E}\left[x_i x_i^T \left\|y_i\right\|_2^2\right]$ are both PSD, we have that

$$\left\|\mathbb{E}\left[\left(x_i y_i^T - R\right)\left(x_i y_i^T - R\right)^T\right]\right\|_2 \leq \left\|\mathbb{E}\left[x_i x_i^T \left\|y_i\right\|_2^2\right]\right\|_2.$$

Next, by Jensen's inequality

$$\begin{aligned}\left\|\mathbb{E}\left[x_i x_i^T \left\|y_i\right\|_2^2\right]\right\|_2 &\leq \mathbb{E}\left[\left\|x_i x_i^T\right\|_2 \left\|y_i\right\|_2^2\right] \\ &= \mathbb{E}\left[\left\|x_i\right\|_2^2 \left\|y_i\right\|_2^2\right] && \left(\left\|zz^T\right\|_2 = \left\|z\right\|_2^2\right) \\ &\leq \sqrt{\mathbb{E}\left[\left\|x_i\right\|_2^4\right]\mathbb{E}\left[\left\|y_i\right\|_2^4\right]} && \text{(C.S.)} \\ &\leq S_1 S_2 \sqrt{\mathbb{E}\left[\left\|x_i\right\|_2^2\right]\mathbb{E}\left[\left\|y_i\right\|_2^2\right]} \\ &= S_1 S_2 \sqrt{\text{trace}\left(R_x\right)\text{trace}\left(R_y\right)}\end{aligned}$$

Combining the above we have that

$$\left\|\mathbb{E}\left[\left(x_i y_i^T - R\right)\left(x_i y_i^T - R\right)^T\right]\right\|_2 \leq S_1 S_2 \sqrt{\text{trace}\left(R_x\right)\text{trace}\left(R_y\right)}.$$

Similarly,

$$\left\|\mathbb{E}\left[\left(x_i y_i^T - R\right)^T\left(x_i y_i^T - R\right)\right]\right\|_2 \leq S_1 S_2 \sqrt{\text{trace}\left(R_x\right)\text{trace}\left(R_y\right)}.$$

We therefore have that

$$V \leq S_1 S_2 \sqrt{\text{trace}\left(R_x\right)\text{trace}\left(R_y\right)}.$$

Finally we find a bound for C. Indeed,

$$\begin{aligned}\left\|x_i y_i^T - R\right\|_2 &\leq \left\|x_i y_i^T\right\|_2 + \|R\|_2 \\ &\leq S_1 S_2 + \|R\|_2 \\ &\leq 2 S_1 S_2 \\ &= C.\end{aligned}$$

This completes the proof. □

Corollary 2 (Finite-Sample Analysis of M_a Estimation)**.** *For any $a \in \mathcal{A}$, let $\hat{M}_{t,a}$ be the estimation of M_a based on t samples (see* (6)*), and $\delta > 0$. Then, with probability greater than $1 - \delta$,*

$$\left\|M_a - \hat{M}_{t,a}\right\| \leq \frac{C_{B1}}{\sqrt{t}} + \frac{C_{B2}(\delta)}{t},$$

where

$$\begin{aligned}C_{B1}(\delta) &= \max_a \left(\frac{\lambda_{\min}(R_{11}(a))^{-1}}{P^{\pi_b}(a)} \right) \sqrt{2 S_1 S_2 \sqrt{\textit{trace}\,(R_{11})\,\textit{trace}\,(R_{22})} \log\left(\frac{d}{\delta}\right)}, \\ C_{B2}(\delta) &= \frac{3}{4} \max_a \left(\frac{\lambda_{\min}(R_{11}(a))^{-1}}{P^{\pi_b}(a)} \right) S_1 S_2 \log\left(\frac{d}{\delta}\right).\end{aligned}$$

Proof. Fix $a \in \mathcal{A}$. See that

$$\begin{aligned}M_a &= \begin{pmatrix} I_L & R_{11}(a)^{-1} R_{12}(a) \end{pmatrix}, \\ \hat{M}_{t,a} &= \begin{pmatrix} I_L & R_{11}(a)^{-1} \hat{R}_{t,12}(a) \end{pmatrix}. \qquad \text{(By (6))}\end{aligned}$$

The following relations holds.

$$\left\|M_a - \hat{M}_{t,a}\right\| \leq \left\|R_{11}(a)^{-1}\right\| \left\|R_{12}(a) - \hat{R}_{t,12}(a)\right\|. \qquad \text{(Norm is sub-multiplicative)}$$

By Lemma 8 we have that with probability at least $1 - \delta$

$$\left\|R_{12}(a) - \hat{R}_{t,12}(a)\right\| \leq \frac{S_1 S_2}{P^{\pi_b}(a)} \left(\sqrt{\frac{2}{t} \left(\frac{\sqrt{\text{trace}\,(R_{11})\,\text{trace}\,(R_{22})}}{S_1 S_2} \right) \log\left(\frac{d}{\delta}\right)} + \frac{4}{3t} \log\left(\frac{d}{\delta}\right) \right).$$

Plugging back into Equation (Norm is sub-multiplicative), using $\left\|R_{11}(a)^{-1}\right\|_2 = \lambda_{\min}(R_{11}(a))^{-1}$, and applying the union bound on $a \in \mathcal{A}$ and $t \in [T]$ we conclude the first claim. □

APPENDIX I USEFUL RESULTS

We restate several very useful lemmas from Abbasi-Yadkori et al. 2011 and Cesa-Bianchi and Lugosi 2006.

Theorem 6 (Abbasi-Yadkori et al. 2011, Theorem 1)**.** *Let $\{F_t\}_{t=1}^{\infty}$ be a filtration. Let η_t be a real-values stochastic process such that η_t is F_t measurable and conditionally σ-sub-Gaussian w.r.t. F_{t-1}. Let $\{x_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that x_t is F_{t-1} measurable. Assume V is a $d \times d$ PD matrix. For any $t \geq 0$, define*

$$V_t = V + \sum_{i=1}^{t} x_i x_i^T \qquad S_t = \sum_{i=1}^{t} \eta_i x_i.$$

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

$$\|S_t\|^2_{V_t^{-1}} \leq 2\sigma^2 \log\left(\frac{det\,(V_t)^{1/2}\, det\,(V)^{-1/2}}{\delta}\right).$$

Theorem 7 (Abbasi-Yadkori et al. 2011, Theorem 2)**.** *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=0}^{\infty}$ be a real-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable and η_t is conditionally σ-sub-Gaussian for $\sigma \geq 0$. Let $\{x_t\}_{t=0}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process s.t. X_t is $\mathcal{F}_{t-1}$-measurable and $\|x_t\| \leq S_x$. Define $y_t = \langle x_t, w\rangle + \eta_t$ and assume that $\|w\| \leq S_w$ and $\lambda > 0$. Let*

$$\hat{w}_t = (X_t^T X_t + \lambda I_d)^{-1} X_t^T Y_t,$$

where X_t is the matrix whose rows are $x_1^T, .., x_t^T$ and $Y_t = (y_1, .., y_t)^T$. Then, for any $\delta > 0$ with probability at least $1-\delta$ for all, $t \geq 0$ w lies in the set

$$\left\{w \in \mathbb{R}^d : \|\hat{w}_t - w\|_{V_t} \leq \sigma\sqrt{d\log\left(\frac{1+tS_x^2/\lambda}{\delta}\right)} + \lambda^{1/2} S_w\right\}.$$

Lemma 9 (Elliptical Potential Lemma, Abbasi-Yadkori et al. 2011, Lemma 11)**.** *Let $\{x_t\}_{t=1}^{\infty}$ be a sequence in $\mathbb{R}^d$ and $V_t = V + \sum_{i=1}^{t} x_i x_i^T$. Assume $\|x_t\| \leq S_x$ for all t. Then,*

$$\sum_{i=1}^{t} \min\left(\|x_i\|^2_{V_{i-1}^{-1}}, 1\right) \leq 2\log\left(\frac{det\,(V_t)}{det\,(V)}\right) \leq 2d\log\left(\frac{trace\,(V) + tS_x^2}{d}\right) - 2\log\left(det\,(V)\right).$$

Lemma 10 (E.g, Cesa-Bianchi and Lugosi 2006, Lemma 11.11 and Theorem 11.7)**.** *Let $x_1, .., x_t$ be a sequence of vectors in $\mathbb{R}^d$ and $\lambda > 0$. Let $V_i = \lambda I_d + \sum_{i=j}^{i} x_j x_j^T$ and assume $\|x_i\| \leq S_x$. Then,*

$$\sum_{i=1}^{t} \|x_i\|^2_{V_i^{-1}} \leq d\log\left(1 + tS_x^2/\lambda\right).$$

Lemma 11 (Matrix Bernstein, Tropp et al. 2015, Theorem 6.1.1)**.** *Consider a finite sequence $\{A_k\}$ of independent, random matrices with common dimension $d_1 \times d_2$. Assume that for all k*

$$\mathbb{E}[A_k] = 0 \text{ and } \|A_k\|_2 \leq S \quad \text{almost surely.}$$

Denote $Z = \sum_k A_k$ and

$$V(Z) = \max\left\{\left\|\mathbb{E}\left[ZZ^T\right]\right\|_2, \left\|\mathbb{E}\left[Z^T Z\right]\right\|_2\right\} = \max\left\{\left\|\sum_k \mathbb{E}\left[A_k A_k^T\right]\right\|_2, \left\|\sum_k \mathbb{E}\left[A_k^T A_k\right]\right\|_2\right\}.$$

Then for all $\epsilon \geq 0$,

$$P\left(\|Z\|_2 \geq \epsilon\right) \leq (d_1 + d_2)\exp\left\{-\frac{\epsilon^2/2}{V(Z) + S\epsilon/3}\right\}.$$

Thus, with probability at least $1-\delta$,

$$\|Z\|_2 \leq \sqrt{2V(Z)\log\left(\frac{d_1+d_2}{\delta}\right)} + \frac{2}{3}S\log\left(\frac{d_1+d_2}{\delta}\right).$$