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 [Mirchevsk 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 [Amoder 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.

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).
we are able to interact with the world, where the full state of the
we are given access to large amounts of medical records yet is needed for training future control strategies. How and where
illustrate our basic setup and approach. We show how data can be utilized for the online bandit problem. Figures 1
In this work we show how the confounded information in the
improving the most recent online system?
should one use the partially observable historical data for
new information in the state was not previously collected, less
handled properly. This is due to the fact that some of the
recurrent process, data that was gathered in previous times
acteristics that may affect traffic congestion. Due to this
conditions using video footage of cameras located over inter-
sections. The development time of the system consists of
continual addition of new labels (classes) for the different
acteristics 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, $\pi_b$). 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 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, $\pi_b$). 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 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, $\pi_b$). 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 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, $\pi_b$). 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 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, $\pi_b$). 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.

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 $\ell_2$-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 $\ell_2$-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 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 $\pi_b$. (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.
The goal of the learner is to maximize the total reward (Theorem 2). Our basic framework consists of sequential interactively observable offline dataset, consisting of partially observable data. We evaluate the learner against the optimal strategy, $F_t$ for any $t$. Here, $\sigma$, $\eta$, unknown parameter vectors, and $a$ reward $\xi$ learner’s action space. Subsequently, the learner observes $\xi_t \in A$ i.i.d. Based on observed payoffs in previous trials, the unknown distribution $P_t$ is denoted as $\xi_t \in A$, namely $F_t$ is the learner’s action space. Subsequently, the learner observes a reward $r_t = \langle x_t, w^*_a \rangle + \eta_t$, where $\{w^*_a \in \mathbb{R}^d\}_{a \in A}$ are unknown parameter vectors, and $\eta_t$ is some conditionally $\sigma$-subgaussian random noise, i.e., for some $\sigma > 0$

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

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

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

$$\text{Regret}(T) = \sum_{t=1}^T \langle x_t, w^*_a(x_t) \rangle - \sum_{t=1}^T \langle x_t, w^*_a \rangle.$$

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 $D = \{x_i, a_i, r_i\}_{i=1}^N$, in which $\{x_i\}_{i=1}^N$ are i.i.d. samples from $P_x$, $\{a_i\}_{i=1}^N$ were generated by some fixed behavior policy, denoted by $\pi_b$, 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 $Q = (I_L \ 0)$. That is, without loss of generality, we assume only the first $L$ features of $x$ are visible in the data.

Notice that the distribution of $D = \{x_i, a_i, r_i\}_{i=1}^N$ is partially observable dataset, depends on $\pi_b$. Any statistic we attempt to draw from the offline data depends on the measure induced by $\pi_b$, which we denote by $P^{\pi_b}$. 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 $D$. Having access to such data is mostly useless without further assumptions. Particularly, $w^*_a$ may not be identifiable\footnote{More precisely, we define the measure $P^{\pi_b}_\eta$ for all Borel sets $R \subseteq [0,1]$, $X \subseteq \mathcal{X}$ and $A \subseteq A$ $P^{\pi_b}_\eta(r \in R, x \in X, a \in A) = P(r \in R | x \in X, a \in A)P(x \in X) \int_{x \in X, a \in A} \mathbb{1}_{\{a=a', x=x'\}} d\pi_b$.} In fact, it can be shown that for any behavioral policy $\pi_b$, and induced measure $P^{\pi_b}$, $\{w^*_a\}_{a \in A}$ are not identifiable. More specifically, for all $w^1 = \{w^1_a\}_{a \in A}$ exist $w^2 = \{w^2_a\}_{a \in A} \neq w^2$ and probability measures $P_1, P_2$ such that $P_1(\{x, a, r; w^1, \pi_b\}) = P_2(\{x, a, r; w^2, \pi_b\})$ and $P_1(\{x, a, r; w^1, \pi_b\}) = P_2(\{x, a, r; w^2, \pi_b\})$ and

\footnote{We use the notion of identifiability as defined in Definition 2 of Pearl et al. [2009].}.\footnote{We use the notion of identifiability as defined in Definition 2 of Pearl et al. [2009].}
\[ \pi_b(a, x; w^t) = \pi_b(a, x; w^2) \] 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 \( \{w_a^*\}_{a \in A} \) can be leveraged \cite{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 \( \{w_a^*\}_{a \in A} \).

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 \( \{M_aw_a^* = b_a, a \in A\} \). 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., \( x^o = x \). Here, one would be able (with large amounts of data) to closely estimate \( w_a^* \) for all \( a \in A \), using, for example, the linear regression estimator

\[
\hat{w}_a = \left( \frac{1}{N_o} \sum_{i=1}^{N_o} x^o_i x_i^T \right)^{-1} \left( \frac{1}{N_o} \sum_{i=1}^{N_o} x^o_i r_i \right)
\]

where we denoted \( N_o = \sum_{i=1}^{N_o} 1_{(a=i)} \). 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

\[
\min_{b \in \mathbb{R}^d} \sum_{i=1}^{N_a} (x^o_i, b - r_i)^2, \forall a \in A,
\]

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

\[
b_a^{LS} = \left( \frac{1}{N_a} \sum_{i=1}^{N_a} (x_i^o) (x_i^o)^T \right)^{-1} \left( \frac{1}{N_a} \sum_{i=1}^{N_a} x_i^o r_i \right), \quad (1)
\]

The following proposition establishes our first main result – a relation between the lower-dimensional least-squares 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

\[
R_{11}(a) = \mathbb{E}_{\pi_b} \left[ x^o (x^o)^T \right] \quad \text{and} \quad R_{12}(a) = \mathbb{E}_{\pi_b} \left[ x^o (x^h)^T \right] \quad \text{for all } a \in A.
\]

Assume \( R_{11}(a) \) is invertible for all \( a \in A \). Then, the following holds almost surely for all \( a \in A \).

\[
\lim_{N \to \infty} b_a^{LS} = \left( I_{L_s} - R_{11}^{-1}(a) R_{12}(a) \right) w_a^*.
\]

The proof of the proposition is related to regression analysis with misspecified models (see e.g., \cite{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 autocorrelation 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 \( x^o \), namely \( x^h \), as well as the behavior policy \( \pi_b \), 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 \( a \in A \). As we discussed in Section 5, such a bound can be achieved, for example, through queries to \( \pi_o \) (i.e., samples \( a \sim \pi_o \)).

**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 \( \{w \in \mathbb{R}^d : Mw = b\} \), 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
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

\[
M_a w_a^* = b_a, \quad a \in A. \tag{2}
\]

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 \( \{ M_a \}_{a \in A} \) are full row rank\(^4\) One way of using the relations in Equation (2) is by constraining \( P_L \) equivalent to one in a lower dimensional space \( R \). Perhaps intuitively, this least squares estimator is in fact a matrix with orthonormal columns\(^5\) we have that (see supplementary material for full derivation)

\[
\hat{w}_{t,a}^P = \left( M_a^T M_a + \sum_{i=1}^{L-1} x_i x_i^T \right) \left( M_a \right)^{\dagger} \times \\
\left( \sum_{i=1}^{L-1} y_{t,i} x_i - \sum_{i=1}^{L-1} x_i x_i^T M_a b_a \right)^{-1}. \tag{5}
\]

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

\[
U^T \hat{w}_{t,a}^P = \left( L_m + \sum_{i=1}^{L-1} (U^T x_i) (U^T x_i)^T \right)^{-1} \times \\
\left( \sum_{i=1}^{L-1} y_{t,i} (U^T x_i) \right).
\]

That is, \( U^T \hat{w}_{t,a}^P \) 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 (2) to construct our final estimator \( \hat{w}_{t,a} = M_a^T b_a + \hat{w}_{t,a}^P \) where \( \hat{w}_{t,a}^P \) is given by Equation (5). Then, estimation of \( \hat{w}_{t,a} \) will depend on the rank of \( P_a \), i.e., \( \text{rank}(P_a) = d - L \). 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 \( a \in A \), assume \( \| P_a x_i \|_2 \leq S_{x,o} \) almost surely and \( \| P_a w_a^* \|_2 \leq S_{w,o} \). Letting

\[
\sqrt{b_0(\delta)} = \lambda^{1/2} S_{w,o} + \sqrt{(d - L) \log \left( \frac{K(1 + S_{x,o}^2/\delta)}{\delta} \right)}
\]

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 \( T \geq 0 \), with probability at least \( 1 - \delta \), the regret of Algorithm 1 is bounded by

\[
\text{Regret}_T \leq \tilde{O}\left( (d - L) \sqrt{KT} \right).
\]

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

---

\(^4\) 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 rank(\( P \)) orthonormal columns.
Algorithm 2 OFUL with Partially Observable Offline Data

1: **input**: $\alpha > 0$, $\delta > 0$, $T$, $b_0 \in \mathbb{R}^k$ (from dataset)
2: **for** $n = 0, \ldots, \log_2(T) - 1$ **do**
3:  Use $2^n$ previous samples from $\pi_b$ to update the estimate of $\hat{M}_{2^n,a}$, $\forall a \in A$
4:  Calculate $\hat{M}_{2^n,a}, \hat{P}_{2^n,a}$, $\forall a \in A$
5: **Run Algorithm 1** for $2^n$ time steps with bonus $\sqrt{\beta_{n+1}(\delta)}$ and $M_{2^n,a}, b_a$
6: **end for**

only for $M_a, b_a$ that are fully known. When $\{M_a\}_{a \in A}$ 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 $x^b$, as well as the behavior policy $\pi_b$. Nevertheless $M_a = (I_L, \ R_{11}^{-1}(a) R_{12}(a))$ can be efficiently estimated whenever $R_{12}(a) = \mathbb{E}^{\pi_b} [x^a (x^h)^T | a]$ can be efficiently estimated. Particularly we make the following assumption.

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

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

5.1 CASE STUDY: QUERIES TO $\pi_b$

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

Having access to queries from $\pi_b$, 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 $\pi_b$ by sampling $a^b_t \sim \pi_b(x_t)$. We then estimate $R_{12}(a)$ using the empirical estimator\footnote{In fact, we can construct a tighter estimator for $R_{12}(a)$ using our knowledge of $\mathbb{E}^{\pi_b} [x^a | a]$, which can be estimated exactly from the offline data. We leave its analysis out for clarity.}

$$\hat{R}_{12}(a,t) = \frac{1}{t} \sum_{i=1}^t \mathbb{I}_{\{a_t = a\}} P^{\pi_b}(a) (x_i^a) (x_i^b)^T,$$

where $P^{\pi}(a)$ is known due to the offline data. Assuming $\|x^a\|_2 \leq S_2$ and $\|x^b\|_2 \leq S_2$ a.s., it can be shown that with probability at least $1 - \delta$ (see supplementary material, Lemma 3, for proof)

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

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

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

A natural question arises: can the estimated linear constraints $M_{a,b} w^* = b_a$ 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 $\hat{M}_{t,a}$ 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 $\|x_i\|_2 \leq S_2$ a.s. and $\|w^*_a\|_2 \leq S_2$ for all $a \in A$, the algorithm uses an augmented confidence, given by

$$\sqrt{\beta_{n+1}(\delta)} = \lambda^{1/2} S_w + (\sigma + S_x S_w f_n) \sqrt{(d - L) \log \left( \frac{1 + \delta_2}{\lambda^{1/2} \log(T) R} \right)},$$

where

- $f_n = f_{B1} + f_{B2} 2^{-n/2}$,
- $f_{B1} = \mathcal{O}(\max_a \lambda_{\min}(R_{11}(a))^{-1} S_x (\text{trace}(R_{11}(a)) (\text{trace}(R_{22}(a)))^{1/4})$,
- $f_{B2} = \mathcal{O}(\max_a \lambda_{\min}(R_{11}(a))^{-1} S_2^2).$

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.
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

$$\text{Regret}(T) \leq 3\sqrt{T(d - L)K\log\left(\lambda + \frac{TS^2}{d - L}\right)} \times$$

$$\left(\sigma_\epsilon \sqrt{(d - L)\log\left(\frac{1 + TS^2/\delta}{(2K\log(T))}\right) + \lambda^{1/2}S_w}\right) +$$

$$O\left((d - L)\sqrt{KS_wS_xf_{B2}}\right),$$

where $\epsilon = S_xS_wf_{B1}$ and $\sigma_\epsilon = \sigma + \epsilon$. This leads to,

$$\text{Regret}(T) \leq \tilde{O}\left((1 + f_{B1})(d - L)\sqrt{KT}\right).$$

Notice that, unlike Theorem 1, the regret of Algorithm 2 is worsened asymptotically by a factor relating to $f_{B1}$. This function can also scale with $d$, due to its dependence on $(\text{trace}(R_{11}(a)))^{1/4}$ and $(\text{trace}(R_{22}(a)))^{1/4}$. Specifically, a worst case dependence yields $f_{B1} \leq \tilde{O}\left(\max_a \frac{L(d - L)^{1/4}}{(P + 1)^{1/4}}\right)$, where here $\max_a \frac{1}{P+1} \geq K$. That is, $f_{B1}$ 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, $\pi_3$. 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 $M^\dagger_a$, $P_a$, 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 $M_a$, 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 $M^\dagger_a$, $P_a$. 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 $1 \leq \text{Regret}(T) \leq \tilde{O}\left((1 + f_{B1})(d - L)\sqrt{KT}\right)$.

Notice that, unlike Theorem 1, the regret of Algorithm 2 is worsened asymptotically by a factor relating to $f_{B1}$. This function can also scale with $d$, due to its dependence on $(\text{trace}(R_{11}(a)))^{1/4}$ and $(\text{trace}(R_{22}(a)))^{1/4}$. Specifically, a worst case dependence yields $f_{B1} \leq \tilde{O}\left(\max_a \frac{L(d - L)^{1/4}}{(P + 1)^{1/4}}\right)$, where here $\max_a \frac{1}{P+1} \geq K$. That is, $f_{B1}$ 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, $\pi_3$. 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 $M^\dagger_a$, $P_a$, 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 $M_a$, 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 $M^\dagger_a$, $P_a$. 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 $1 \leq \text{Regret}(T) \leq \tilde{O}\left((1 + f_{B1})(d - L)\sqrt{KT}\right)$.

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 $\omega^*_a \in \mathbb{R}^{30}$ uniformly sampled in $[0, \frac{1}{2}]^d$ 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 $\pi_b$ was chosen to follow a softmax distribution $\pi_b(a, x) \propto \exp(\phi^T_\omega x)$, where $\phi_\omega \in \mathbb{R}^d$ were randomly chosen and fixed across all experiments.
Figure 3 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 approximate projections. We note that the experiments were run under constant updates of $M_{t,a}$, i.e., without epoch schedules.

Figure 3 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 3 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 bands 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; Kazeroni 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., 2019].

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 $P(r|do(a))$) [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 infinite 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,
 querying \( \pi_b \) 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.

**Acknowledgements**

This research was partially supported by the ISF under contract 2199/20.

**References**


