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

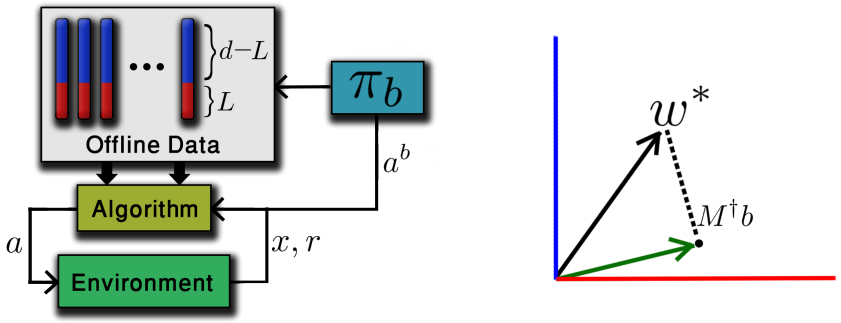
The use of offline data for online control is of practical interest in fields such as autonomous driving, healthcare, dialogue systems, and recommender systems [Mirchevska et al., 2017, Murphy et al., 2001, Li et al., 2016, Covington et al., 2016]. There, an abundant amount of data is readily available, 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.

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:430–439.



(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

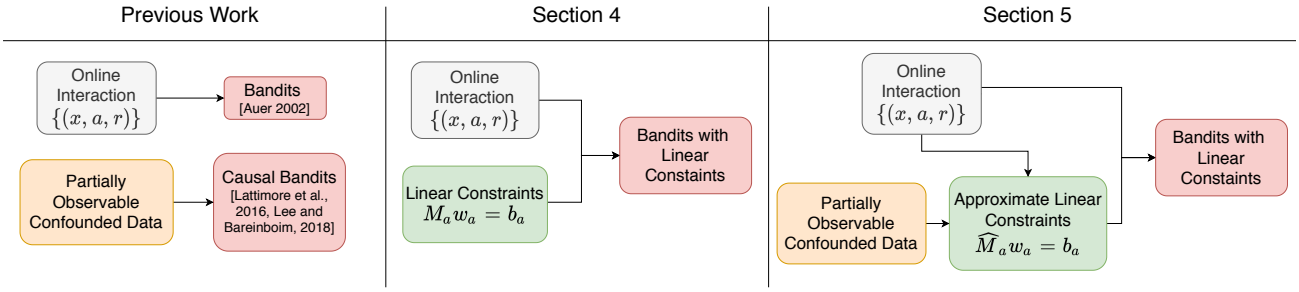


Figure 2: Previous work has dealt with bandits in the online setting. Other work integrated causal information to utilize confounded offline data. This work combines the two through constraints on the online problem. In Section 4 we show how linear constraints can be leveraged to achieve better regret 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, $x = \left((x^{\mathrm{o}})^T, (x^{\mathrm{h}})^T\right)^T$, where $x^{\mathrm{o}} \in \mathbb{R}^L$, $x^{\mathrm{h}} \in \mathbb{R}^{d-L}$.

Notice that the distribution of $\mathcal{D} = \{x^{\mathrm{o}}_i, a_i, r_i\}_{i=1}^N$, 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 $P_1\left(x^{\mathrm{o}}, a, r; w^1, \pi_b\right) = P_2\left(x^{\mathrm{o}}, a, r; w^2, \pi_b\right)$ 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* $a \in \mathcal{A}$.

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

The proof of the proposition is related to regression analysis with misspecified models (see e.g., 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)

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

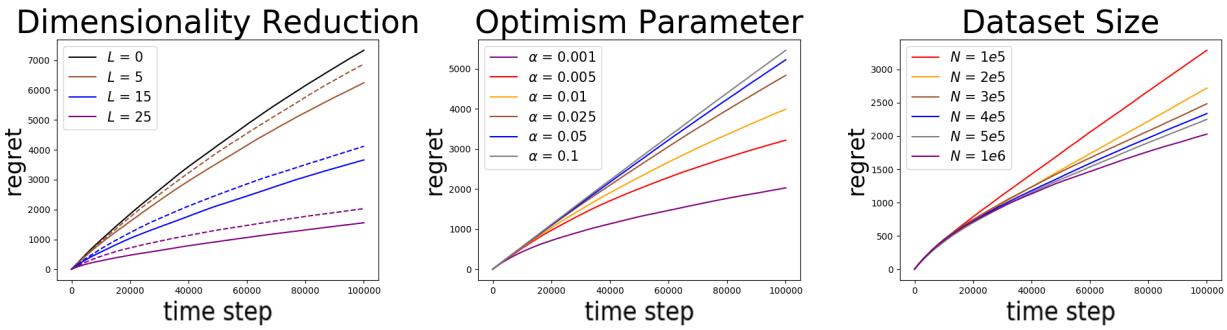
indeed, satisfying Assumption 1. We can now naturally construct an estimator for M_a. Its estimator is given by

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

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 $\sqrt{\beta_{n,t}(\delta)} = \lambda^{1/2} S_w + (\sigma + S_x S_w f_n) \sqrt{(d-L) \log\left(\frac{1 + t S_x^2/\lambda}{\delta/2\log(T)K} \right)}$, where $f_n = f_{B1} + f_{B2} 2^{-n/2}$, $f_{B1} = \widetilde{\mathcal{O}}\Big(\max_a \frac{\lambda_{\min}(R_{11}(a))^{-1}}{P^{\pi_b}(a)} S_x (\operatorname{trace}(R_{11}(a)) \operatorname{trace}(R_{22}(a)))^{1/4} \Big)$ 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.

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.

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 sub-gaussian martingale difference noise. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4444–4448. IEEE, 2018.

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.

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.

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.

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.

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.