Psuedo-Data Injection for CLO Bandit Problems

Jad Soucar*
Department of Industrial & Systems Engineering
University of Southern California
Los Angeles, CA 91214
soucar@usc.edu

Ryan Edmonds*
Department of Computer Science
University of Southern California
Los Angeles, CA 91214
ryanedmo@usc.edu

Abstract

Contextual linear optimization (CLO) with bandit feedback is a class of CLO problems where only the costs of historical actions are observable. Finding an optimal decision making policy in this setting suffers from the fundamental challenge that real-world data often lacks coverage over the action space, making the full cost vector unidentifiable with the data available. A common remedy is to apply regularization to ensures stability of the learning problem. We show that this approach admits an alternative interpretation as a specific form of pseudo-data injection where synthetic data is added to induce coverage. This perspective suggests a broader question regarding how arbitrary pseudo-data can be injected when prior beliefs about the environment or data collection process are available. We propose two methods of pseudo-data injection that reflect structured beliefs about the underlying cost distribution or the data collection process, and show that regularization is a special case. We provide regret bounds for policies learned using post-injection datasets, which highlights how data injection strategies to repair coverage impact downstream decision quality. We conclude with numerical experiments demonstrating how policy regret behaves depending on veracity of prior beliefs, and under what regimes our methods outperform regularization.

1 Introduction

Contextual linear optimization (CLO) uses contextual information to find effective decision making policies under uncertainty [6, 7, 8, 18]. In most cases the problem of CLO can be reduced to finding a policy $\pi(x)$ that solves

$$\min_{\pi(\cdot)} \mathbb{E}_{(X,Y)}[\langle \pi(X), Y \rangle] \quad \text{s.t} \quad \pi(x) \in \mathcal{Z}, \tag{1}$$

where $X \in \mathbb{R}^c$ is a random variable representing contextual information available to the decision maker, $Y \in \mathbb{R}^d$ is a random variable representing cost coefficients, and $\mathcal{Z} \subseteq \mathbb{R}^d$ is the set of feasible actions. We assume throughout that (X, Y) is fully described by a latent joint probability distribution and that $\mathcal{Z}$ is finite and satisfies $\sup_{z \in \mathcal{Z}} \|z\|_2 \leq B$. The expectation in problem (1) is taken over this joint distribution. In practice, the problem is solved by observing i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^n$ drawn from the joint distribution, then finding an optimal policy $\pi(\cdot)$ in some policy class Π that minimizes the sample average approximation of problem (1). At decision time, a practitioner will observe some context X, then use the obtained policy to make a decision $\pi(X)$. Only after the action is taken, can the practitioner observe Y which they can use to refine their policy by resolving problem (1).

In the full-feedback setting, the full state of the world is observable and recordable - i.e., the full vector Y is available [9]. For example, consider the stochastic shortest path setting where a driver must choose a path on a road network $G = (V, E)$ given some contextual information X_i which

*Equal contribution authors.

encodes features such as the weather or road closures. In this case, actions $\pi(X_i)$ can be represented as a binary vector $Z_i \in \mathcal{Z} \subseteq \{0,1\}^{|E|}$ where $\pi_j(X_i) = 1$ corresponds to the driver using the j^{th} edge and $Y_i \in \mathbb{R}^{|E|}$ represents the cost of every road in the network. However, in many cases the decision maker does not have access to the full-feedback cost vector Y_i. For instance, it may be more reasonable to assume that after taking an action $\pi(X_i)$, the driver only observes the total cost of the route taken $C_i = \pi(X_i)^\top Y_i \in \mathbb{R}$ while the full cost coefficient vector remains hidden. Indeed many practical implementations of stochastic shortest path problems have access only to this partial information [13, 14, 21]. This setting, where only information about the decision taken is observed, is known as CLO with *bandit feedback*.

In the CLO with bandit feedback setting, introduced by Hu et. al. [9], the observable dataset is now given by $\mathcal{D} = \{(X_i, Z_i, C_i)\}_{i=1}^n$, where (X_i, Y_i) is drawn from the latent joint distribution, $Z_i = \pi_{\log}(X_i)$ is generated by the historical *logging policy*, and $C_i = Z_i^\top Y_i$ encodes the total cost of the decision made. This induces a distribution P on (X, Z, C). The learning task remains to find an optimal policy $\pi : \mathbb{R}^c \to \mathcal{Z}$. However, depending on the logging policy, the dataset may contain little to no information about the costs of alternative actions for a given context X_i. Consequently, if certain actions were rarely or never taken, then their associated costs cannot be reliably inferred. This creates an inherent difficulty for learning, since the optimal policy may depend on actions that are poorly represented in the data. We refer to this problem as a *lack of coverage*. Developing methods to mitigate the coverage problem is the central focus of this work.

1.1 Background: CLO with bandit feedback & the problem of coverage

To solve the CLO with bandit feedback problem, Hu et. al [9] propose a two step learning framework designed to find a policy $\pi : \mathbb{R}^c \to \mathcal{Z}$ that has low regret

$$\mathrm{Reg}(\pi) = \mathbb{E}_P[f_0(X)^\top \pi(X) - \min_{z \in \mathcal{Z}} f_0(X)^\top z], \tag{2}$$

where $f_0(x) = \mathbb{E}_{(X,Y)}[Y|X = x]$. Since Y is not available, the first step is to estimate $f_0(x)$ and $\Sigma_0(x) = \mathbb{E}_P[ZZ^\top|X]$ by $\hat{f}$ and $\hat{\Sigma}$ respectively using the data available. These estimators are typically referred to as nuisance functions. The nuisance function $\hat{f}$ is estimated by solving

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}^N} \frac{1}{n} \sum_{i=1}^n (C_i - Z_i^\top f(X_i)), \tag{3}$$

where $\mathcal{F}^N$ is a user specified hypothesis class for the nuisance function. The nuisance function $\Sigma_0(x)$ can be estimated by $\sum_{z \in \mathcal{Z}} zz^\top e(z|x)$ where $e(z|x)$ is the estimated propensity score of an action z given context x under the bandit data generation process. We assume that $\hat{\Sigma}(x)$ is constructed using propensity scores throughout. The authors then introduce a score function θ which satisfies

$$\mathbb{E}_P[\theta(X, Z, C; f_0, \Sigma_0)^\top \pi(X)] = \mathbb{E}_P[f_0(X)^\top \pi(X)] \quad \forall \pi(x) \in \Pi_{\mathcal{F}}. \tag{4}$$

The authors propose the following choices for $\theta(x, z, c; f, \Sigma)$:

1. (Direct Method) $f(x)$,
2. (Inverse Spectral Method) $\Sigma^\dagger(x)zc$,
3. (Doubly Robust Method) $f(x) + \Sigma^\dagger(x)z(c - z^\top f(x))$,

where $\dagger$ denotes the Moore-Penrose inverse. Each of these choices of θ satisfies equation (4) [9, Proposition 1]. We assume these functional forms on θ for the remainder of this work. In the second stage, the nuisance functions $\hat{f}$ and $\hat{\Sigma}$ are used to obtain an optimal policy by solving

$$\hat{\pi}(x) = \operatorname*{argmin}_{\pi \in \Pi_{\mathcal{F}}} \sum_{i \in \mathcal{D}} \theta(X_i, Z_i, C_i; \hat{f}, \hat{\Sigma})^\top \pi(X_i). \tag{5}$$

Here we denote the policy class as $\Pi_{\mathcal{F}} = \{\pi_f(x) = \mathrm{argmin}_{z \in \mathcal{Z}} z^\top f(x) : f \in \mathcal{F}\}$ since the class is induced by a choice of some hypothesis class, which need not be equivalent to $\mathcal{F}^N$. Critically, this process relies on certain assumptions on P.

Assumption 1.1 (Coverage & Ignorability)**.** *The bandit data generating process satisfies the following properties:*

1. *(Ignorability)* $\mathbb{E}_P[C|Z,X] = Z^\top f_0(X) \iff Z \perp Y|X$
2. *(Coverage)* $\inf_{\substack{z\in\mathrm{span}(\mathcal{Z})\\ \|z\|_2=1}} \mathbb{E}_P[(Z^\top z)^2|X] > 0$*,* $\mathbb{E}_P[ZZ^\top|X]$ *is invertible on* $\mathrm{span}(\mathcal{Z})$.

Coverage can be best understood geometrically. Specifically, coverage requires that for any observable context X, the historical logging policy explores every direction of the action space $\mathcal{Z}$. To understand why coverage is mechanically important for solving (5), consider an example where the data generation process does not satisfy coverage. Specifically, suppose there exists a direction $v \in \mathrm{span}(\mathcal{Z})$ for which $Z^\top v = 0$ almost surely. This means that $Z^\top(f(x) + \alpha v) = Z^\top f(x)$ almost surely, which in turn implies that if $\hat{f}$ is a minimizer of problem (3), then $f + \alpha v$ is also a minimizer for any $\alpha \in \mathbb{R}$. This poses a challenge to identifying an optimal policy, since using $\hat{f}(x)$ versus some other minimizer $\hat{f}(x) + \alpha v$ may result in different policies learned. For example, consider two actions $z_1, z_2 \in \mathcal{Z}$ with $\hat{f}(x)^\top z_1 < \hat{f}(x)^\top z_2$ and without loss of generality assume $v^\top z_1 < v^\top z_2$. If the nuisance function estimator returns $\hat{f}$, we can conclude that z_1 is a "better" decision then z_2. If instead the nuisance function estimator returns the minimizer $\hat{f}(x) + \alpha v$ with $\alpha > (\hat{f}(x)^\top z_2 - \hat{f}(x)^\top z_1)/(v^\top z_1 - v^\top z_2)$ then we have that $(\hat{f}(x) + \alpha v)^\top z_2 > (\hat{f}(x) + \alpha v)^\top z_1$. In this case z_2 is a "better" decision than z_1. Hence, without coverage, the optimal policy is not identifiable from the observed data and can be arbitrarily different depending on the choice of a non-unique optimal nuisance function.

Despite the importance of assuming coverage for reliably obtaining a policy, a bandit data generation process that creates systematic lack of coverage is not unimaginable in real-world settings. For example, in a road network, historical factors such as redlining and uneven infrastructure investment can make certain regions less desirable to drivers. One might consider a problem within a healthcare setting where specific treatments are rarely or never used for certain classes of patients, resulting in a similar lack of coverage. Such real-world considerations might make it difficult for a data generating process to satisfy the coverage assumption. Importantly, even if a data generating process satisfies coverage, finite sample datasets can still lead to practical concerns. In particular, there may be directions $v \in \mathrm{span}(\mathcal{Z})$ that have low probability of occurring under the chosen logging policy. As a result the nuisance function $\hat{\Sigma}(x)$ may be singular over the span of $\mathcal{Z}$ for certain contexts x or can become poorly conditioned resulting in unstable inverse or pseudo inverses, which are essential to solving problem (5) under the doubly robust and inverse spectral choices of θ. In short, lack of coverage can arise in low data regimes or from systemic real-world decision biases intrinsic to the data generation process. This leads us to the paper's core question: **can we intelligently repair coverage in historical datasets**?

We answer this question as follows. In section 2, we examine how lack of coverage is typically addressed via regularization and show its connection to *pseudo-data injection* and prior imposition. In section 3, we study how a practitioner can repair the coverage of a dataset using pseudo-data injection based on prior information about the logging policy or the data-generating distribution. We also analyze how the quality of a prior-driven data injection impacts policy regret through theoretical bounds. Finally, in section 4, we present numerical experiments demonstrating that incorporating prior information to induce coverage can substantially improve decision quality relative to standard coverage repair techniques.

1.2 Background: warm-starting & data-injection for multi-armed bandits

A special case, emphasized by Hu et al [9], is the simplex setting $\mathcal{Z} = \Delta_d$, which corresponds with the classical d-armed contextual bandit problem. Much of the contextual bandit literature focuses on this setting and develops algorithms for learning low-regret policies from partial feedback data [2, 12, 19, 22, 25]. A related line of work studies how to warm-start multi-armed and contextual bandit learners by incorporating distributional priors into classical contextual bandit algorithms [1, 4, 15]. More recent work has also explored the use of pseudo-observations, generated from auxiliary sources including supplementary datasets and large language models, to remedy the cold-start problem [16, 20, 24]. Our treatment of data injection differs in two ways from the contextual bandit warm-starting literature. First, we focus on expanded settings where $\mathcal{Z}$ need not be a simplex,

but instead can be any finite action set including those that are combinatorial. A consequence of this expansion is that our pseudo-data injections need not correspond to feasible actions. Second, we draw from work on synthetic and catalytic priors [10, 17] to argue that pseudo-data injections and imposing prior distributions serve analogous roles and in some cases are equivalent in CLO with bandit feedback problems.

2 Inducing coverage through regularization

Inducing coverage in a finite dataset with many rare or missing actions is important to utilizing key theoretical results in the bandit CLO literature. This task can be reinterpreted as ensuring that $\hat{\Sigma}(X_i) = \sum_{z \in \mathcal{Z}} zz^\top e(z|X_i)$ is invertible and preferably well-conditioned on the span of $\mathcal{Z}$ for all contexts X_i in the data $\mathcal{D}$. In particular, this requires that the estimated propensity scores assign sufficient mass across all directions of $\mathcal{Z}$. In the case where $\hat{\Sigma}$ is independent of the context one might consider calculating propensity as an empirical frequency. However, this approach will assign 0 propensity to unobserved directions in the action space. On the other hand, when $\hat{\Sigma}(x)$ depends on the context, one might consider training a separate regression model $m : \mathbb{R}^c \to \Delta(|\mathcal{Z}|)$, where $\Delta(|\mathcal{Z}|)$ is the $|\mathcal{Z}|$-dimensional simplex. However, this approach may also result in ill-conditioned $\hat{\Sigma}(x)$ if $|\mathcal{Z}|$ is large, since many actions might receive very small propensity scores.

As a result, the simplest method to produce a well-conditioned $\hat{\Sigma}(x)$ that is invertible on the span of $\mathcal{Z}$ is to use lambda regularization. Specifically, Hu et. al [9] propose replacing $\hat{\Sigma}(x)$ with $\tilde{\Sigma}(x) = \hat{\Sigma}(x) + \lambda I$, where $\lambda = 1$. Despite being a simple fix, the lambda regularization technique admits a deeper interpretation as augmenting the available dataset $\mathcal{D}$ with additional pseudo-data. We denote the concatenation of two datasets $\mathcal{D}_1 = \{\alpha_i\}_{i=1}^{n_1}$ and $\mathcal{D}_2 = \{\beta_i\}_{i=1}^{n_2}$ as $\mathcal{D}_1 + \mathcal{D}_2 = \{\alpha_1, \ldots, \alpha_{n_1}, \beta_1, \ldots, \beta_{n_2}\}$.

Proposition 2.1 (Lambda Regularization Pseudo-Data)**.** *Solving problem* (5) *with* $\theta_{ISM}(x, z, c; \hat{\Sigma} + \lambda I, \hat{f})$ *over data* $\mathcal{D}$ *with* $|\mathcal{D}| = n$ *is equivalent to solving problem* (5) *with* $\theta_{DM}(x, z, c; \hat{f}, \hat{\Sigma})$ *over data* $\tilde{\mathcal{D}} = \mathcal{D} + \sum_{i=1}^{n} \{(X_i, \sqrt{\lambda} e_k, 0)\}_{k=1}^{d}$ *where* e_k *denotes the* k^{th} *standard basis vector.*

In other words, using lambda regularization corresponds to augmenting the dataset with 0-cost decisions. Practically, this implies that implementing lambda regularization can be equivalently implemented by adding pseudo-data to a dataset. For example, for a decision maker who must pick a path on a road network, lambda regularization with $\lambda = 1$ corresponds to adding single road actions with 0 cost for each observed context. From a decision making perspective, λ regularization creates an incentive to visit roads not previously covered by the dataset since it assumes 0 cost along unvisited segments of the road network. In this way, λ regularization can be viewed as imposing a certain type of prior information on the decision problem: unvisited roads should be explored since they offer an opportunity to find cheaper paths then those previously taken.

A natural question is whether coverage can be repaired while imposing arbitrary priors on the expected cost vector. For example consider a case where the practitioner holds the prior that actions not taken were likely avoided by the logging policy due to their high cost. In that case, lambda regularization might not be the preferred method of inducing coverage. To that end, we consider how arbitrary pseudo-data injections correspond to imposing different priors on the cost function. This relationship has been studied previously [10, 17], however we provide specialized results to the contextual bandit setting.

Proposition 2.2 (General pseudo-data)**.** *Let* $|\mathcal{D}| = n$*,* $\mu : \mathbb{R}^c \to \mathbb{R}^d$*, and* $\Lambda : \mathbb{R}^c \to \mathbb{R}^{d \times d}$ *with* $\Lambda(X_i) \succ 0$*. Assume a point wise prior* $f(X_i) \sim \mathcal{N}(\mu(X_i), \Lambda(X_i)^{-1})$ *and likelihood* $C_i \mid X_i, Z_i, f \sim \mathcal{N}(Z_i^\top f(X_i), \sigma^2)$ *for all* $(X_i, Z_i, C_i) \in \mathcal{D}$*. Then, for fixed* Σ *and any choice of* θ*, solving* (5) *over* $\mathcal{D}$ *using maximum-posterior estimation of* $\hat{f}$ *is equivalent to solving* (5) *over*

$$\tilde{D} = D + \sum_{i=1}^{n} \sum_{k=1}^{m_i} \{(X_i, \sigma L_k(X_i), \sigma L_k(X_i)\mu(X_i)\})$$

using the MSE estimation of $\hat{f}$ *described in problem* (3)*, where* $L(X_i) \in \mathbb{R}^{m_i \times n}$ *satisfies* $L(X_i)^\top L(X_i) = \Lambda(X_i)$ *and* $L_k(X_i)$ *denotes the* k^{th} *row.*

Proposition 2.2 informs the practical strategy of imposing general priors by injecting prior-aligned pseudo-data. A theoretical concern is whether such data injections create datasets that satisfy the coverage and ignorability assumptions 1.1. In the case where general pseudo-data $\mathcal{D}_{pd}$ is injected, ignorability is satisfied over the data if all data points in the augmented dataset $\tilde{\mathcal{D}} = \mathcal{D} + \mathcal{D}_{pd}$ satisfy $C_i = Z_i^\top f_0(X_i)$. Interestingly, in the case of lambda regularization where $f_0(x) \not\equiv 0$, injecting data $\sum_{i=1}^n \{(X_i, \sqrt{\lambda} e_k, 0)\}_{k=1}^d$ will create an augmented dataset $\tilde{D}$ that violates ignorability, since applying the $C_i = Z_i^\top f_0(x)$ condition to all data points in the augmenting dataset necessarily implies that $f_0(x) \equiv 0$. On the other hand, coverage can be guaranteed in the limit if the augmenting data is drawn i.i.d. from a prior distribution whose mixture with P satisfies coverage. This can be interpreted as selecting actions that explore previously unobserved directions in $\mathrm{span}(\mathcal{Z})$. We conclude this section by observing that the problem of inducing coverage should be viewed as a data injection problem where the data injected reflects some type of prior belief. This begs the question; what priors might be available in real-world settings and what type of pseudo-data injections should be applied to reflect those priors?

3 Direct pseudo-data injection

In this section we discuss two different priors that may be available to a practitioner, how those priors can be operationalized as pseudo-data injections to induce coverage, and we provide theoretical regret analysis for each type of pseudo-data injection introduced.

3.1 Pseudo-data injection given prior belief on data

First we discuss the general case where a practitioner holds a prior belief on the data generating distribution P. We refer to the prior distribution as Q. Consider a driver who must make a decision regarding a certain route that is not properly covered in the decision maker's experiential dataset $\mathcal{D}$. That driver might draw on a prior belief that the route in question has high cost given some context X. For instance, the driver might believe that the route under consideration is expensive during rush hour. In this way the decision maker is implicitly drawing from some prior distribution for information not covered in the dataset, injecting that pseudo-data into their dataset, and as a result inducing coverage on the augmented dataset. The same process can be replicated algorithmically by drawing data from a prior distribution Q to create an augmented dataset which satisfies coverage and produces new estimated nuisance functions. Note that the prior distribution can have $Q(z \in \mathcal{Z}|X, C) < 1$ almost surely on X. For, example one might wish to impose a prior on one road within a route, which need not correspond to a feasible action. A natural question arises in this setting: how is decision quality effected by the choice of prior distribution? To answer this question we first introduce a standard assumption in the CLO bandit literature.

Assumption 3.1 (Margin)**.** *Let π_0 be the optimal policy induced by $f_0(x)$ and define $\gamma(X) = \min_{z \neq \pi_0(X)} (f_0(X)^\top z - f_0(X)^\top \pi_0(X))$. Assume that there exists a $C > 0$ and $\alpha \geq 0$ such that*

$$\mathrm{Prob}_P(\gamma(X) \leq t) \leq C t^\alpha.$$

The margin assumption can be interpreted as controlling the gap between the first and second best solutions, with large α meaning that the sub-optimality gap tends to be large across contexts. Hu et al. show that assumption 3.1 holds with $\alpha = 1$ for sufficiently well-behaved f_0 and continuous X. Moreover, any CLO instance trivially satisfies this assumption with $\alpha = 0$ [8, Lemma 4,5]

Theorem 3.2. *Let C and α be the constants from assumption 3.1. Assume $\mathcal{D} \sim P^{\otimes n}$, where P satisfies ignorability and is the data generating distribution under logging policy $\pi_{\log}$. Assume that $\mathcal{D}_{pd} \sim Q^{\otimes m}$ such that $\tilde{P} = (1-\eta)P + \eta Q$ satisfies coverage for $\eta = m/(n+m)$ and that marginals $P_X = Q_X$. Let $\tilde{f}$ be a solution of* (3) *over $\tilde{D} = \mathcal{D} + \mathcal{D}_{pd}$ and $\tilde{\pi}$ be the induced policy. Assume that $\mathbb{E}_P[\|\tilde{f}^*(X) - \tilde{f}(X)\|_2^2] \leq \mathrm{Rate}_{\mathcal{F}}(n, m, \delta)$ with probability at least $1-\delta$ where $\tilde{f}^* \in \arg\min_{f \in \mathcal{F}^N} \mathbb{E}_{\tilde{P}}[(C - Z^\top f(X))^2]$. Then with probability at least $1-\delta$ we have*

$$\mathrm{Reg}(\tilde{\pi}) \leq 2BK^{\frac{1}{2}} \left(\underbrace{\mathrm{Rate}_{\mathcal{F}^N}(n, m, \delta)^{\frac{1}{2}}}_{\textit{Estimation Error}} + \underbrace{\frac{\eta}{\lambda_{\min}^+(\tilde{\Sigma})} \mathbb{E}_Q\left[\|Z(C - Z^\top f_0(X))\|_2^2\right]^{\frac{1}{2}}}_{\textit{Pseudo-Data Bias}} \right)^{\frac{3\alpha+2}{4(\alpha+1)}} \tag{6}$$

where $K = (1+\alpha)(\frac{C(2B)^\alpha}{\alpha^\alpha})^{\frac{1}{\alpha+1}}$, $\tilde{\Sigma}(X) = \mathbb{E}_{\tilde{P}}[ZZ^\top|X]$, $\lambda^+_{\min}$ *is the function returning the smallest positive eigen value and* $\lambda^+_{\min}(\tilde{\Sigma}) = \operatorname{ess\,inf}_x \lambda^+_{\min}(\tilde{\Sigma}(x))$.

Theorem 3.2 decomposes regret into an estimation term and a pseudo-data bias term. The estimation error captures the finite sample error of the regression estimator. These rates are generally well studied, vary depending on the hypothesis class $\mathcal{F}^N$, and typically goes to 0 as $n, m \to \infty$ [9]. We instead focus our attention on the second term which quantifies the bias of the pseudo-data. Specifically, the bias of the pseudo-data injection is measured as the extent to which the pseudo-data distribution Q violates the ignorability condition $\mathbb{E}[C|Z,X] = Z^\top f_0(X)$ which is satisfied under the original bandit data-generating distribution P from which $\mathcal{D}$ is drawn. The bound answers the motivating question for theorem 3.2 - how the quality of data injection impacts decision quality. Theorem 3.2 tells us that quality, as measured through ignorability violations, of the injected data directly impacts decision quality. In fact, when P satisfies ignorability and $Q = P$, then the bias term vanishes. This extreme case is intuitive since if we can draw an unlimited amount of data from the true distribution then regret should shrink. The intuition that regret shrinks when Q is similar to P is captured by corollary A.1, where the pseudo-data bias term is written in terms of the 1-Wasserstein distance which furnishes an intuitive but weaker bound. The bound also incorporates the degree to which coverage has been induced. For example, if there are still decision-relevant events for certain contexts with low probability after the pseudo-data injection then $(\lambda^+_{\min}(\tilde{\Sigma}))^{-1}$ can become very large causing the second term to dominate the rate.

Remark 3.3 (Lambda Regularization Specialization)**.** *Interestingly, lambda regularization can be viewed as a special case of theorem 3.2 where* Q *is chosen such that* $Q_X = P_X$, $Q(C = 0|X, Z) = 1$ *almost surely, and* $Q(Z|X, C) \sim \text{Uniform}(\sqrt{\lambda}e_1, \ldots, \sqrt{\lambda}e_d)$ *where* e_i *are standard basis vectors of* $\mathbb{R}^d$*. In this setting the psuedo-data bias reduces to the constant* $\sqrt{d}\,\mathbb{E}_P[\|f_0(X)\|_2^2]^{1/2}$*; see appendix A.6.*

3.2 Pseudo-data injection given prior belief on logging policy

It may be unreasonable to assume that a practitioner holds a prior belief on the entire data generation process. Suppose instead that they hold a prior belief only on the quality of the historical logging policy. This belief can itself be used to construct a prior Q, yielding a special case of section 3.1. To describe the quality of a historical logging policy, we introduce the concept of *rationality*.

Definition 3.4 (Rationality)**.** *Let* $\Delta(X) = \max_{z\in\mathcal{Z}} f_0(X)^\top z - \min_{z\in\mathcal{Z}} f_0(X)^\top z$*. A policy* π *is considered* R_π*- rational if*

$$R_\pi = 1 - \frac{\mathbb{E}[f_0(X)^\top \pi(X) - \min_{z\in\mathcal{Z}} f_0(X)^\top z]}{\mathbb{E}[\Delta(X)]}.$$

A prior belief on logging policy rationality $R_{\pi_{\log}}$ can help a decision maker choose how to induce coverage through data injection. For instance, if $R_{\pi_{\log}} = 0$ then the logging policy selects the worst action for any context almost surely. In this case, the decision maker may want to draw from a prior Q that assigns low cost to to unexplored directions in the action space. This choice incentivizes the exploration of unexplored actions. On the other hand, if $R_{\pi_{\log}} = 1$ then the logging policy is the optimal policy for any context almost surely. In this case, the decision maker may want to draw from a prior Q that assigns high cost to to unexplored directions in the action space, to discourage exploration. Two natural questions arise; given a prior belief on the rationality of the logging policy, what specific data should be injected and what is the decision quality of the post-injection policy?

Theorem 3.5. *Given all assumptions in theorem 3.2 and terms defined in definition 3.4, suppose* Q *is chosen such that for all* $(x, z, c) \sim Q$ *we have that* $c = M$ *and* $\|z\|_2 \le B$*. Then with probability at least* $1 - \delta$ *we have that*

$$\text{Reg}(\tilde{\pi}) \le 2BK^{1/2}\left(\text{Rate}_{\mathcal{F}^N}(n, m, \delta)^{1/2} + \frac{\eta B}{\lambda^+_{\min}(\tilde{\Sigma})}\mathbb{E}_Q[(M - Z^\top f_0(X))^2]^{1/2}\right)^{\frac{3\alpha+2}{4(\alpha+1)}}.$$

Moreover there exists a selection of Q *with support almost entirely composed of directions of* $\mathcal{Z}$ *not covered in* P *where* $\eta Q + (1-\eta)P$ *satisfies coverage, such that the regret bound is minimized by choosing*

$$M^* = \mathbb{E}_P[R_{\pi_{\log}} \max_{z\in\mathcal{Z}} f_0(X)^\top z + (1 - R_{\pi_{\log}}) \min_{z\in\mathcal{Z}} f_0(X)^\top z]$$

Theorem 3.5 reinforces the intuition described earlier. Specifically, if given the choice of augmenting dataset where costs are constant, the constant should be chosen as a function of the rationality of the logging policy. For example if $R_{\pi_{\log}} = 0$ then the constant cost should be chosen as the expected minimum cost to incentivize exploration. Alternatively, if $R_{\pi_{\log}} = 1$ then the constant cost should be chosen as the maximum expected cost to incentivize exploitation of actions already "found" by the logging policy. We note that this intuition holds formally under specific selections of Q. The proof of 3.5 constructs this Q by first placing uniform probability mass on uncovered directions of $\mathrm{span}(\mathcal{Z})$, then adding a correction term.

4 Experiments

We've discussed strategies to induce coverage in settings where the practitioner has some prior belief on the available data or about the rationality of the logging policy. We now test the performance of our proposed coverage induction methods with a focus on (1) how pseudo-data injection impacts performance when prior beliefs vary in quality and (2) in what settings our proposed coverage repair methods outperform lambda regularization. The overall experimental setup will dictate *where* we need to inject data to repair coverage, and the specific experiments will dictate *what* data we inject.

4.1 Experimental setup

We run out experiments in a simulated stochastic shortest path problem in accordance with [7, 9]. The task is to travel from a source node s to a sink node t on a 5 by 5 grid, resulting in a setup with $d = 40$ edges and 70 feasible paths, so $\mathcal{Z} \subset \{0,1\}^{40}$ and has cardinality 70. We consider 3-dimensional contexts X and a linear underlying cost function $f_0(x) = \mathbb{E}[Y|X = x]$; further implementation details are in Appendix B.1. Differing from previous work, expected arc costs are in part dependent on their position in the grid, such that arcs with lower index are generally cheaper to traverse than arcs with higher index - this ensures that historical policies that adhere to different rationalities will systematically avoid different regions of the grid. Historical logging policies $\pi_{\log}$ are generated according to a rationality parameter $r \in [0,1]$. For each historical context X_i, the policy selects the path whose total cost is the closest to $r \min_{z\in\mathcal{Z}} Y_i^\top z + (1-r)\max_{z\in\mathcal{Z}} Y_i^\top z$. By doing this, we obtain policies where $R_{\pi_{\log}} \approx r$. This results in each historical dataset $\mathcal{D}_r = \{(X_i, Z_i, C_i)\}_{i=1}^n$ having a lack of coverage; the details of how coverage is repaired depend on the experiment in question.

Once coverage is repaired, we proceed as in [9] to learn nuisance functions, compute the score function $\theta(x, z, c; f, \Sigma)$ according to the Doubly Robust Method, and optimize the resulting CLO problem. Throughout, we use the K-fold cross-fitting procedure in [5] with K=2. The components of the pipeline not directly related to bandit feedback are implemented using the PyEPO package [23], and we use the Perturbation Gradient Central (PGC) surrogate loss derived by [11]. Doubly robust θ and PGC are chosen due to their high performance in [9]. In each case we compute the regret as defined in equation 2, normalized by the expected cost of the optimal policy (normalized regret). For all experiments, reported values are averaged over 50 independent replications of the underlying data generation process. Specifically, the 50 datasets used are the same across the different experiments described below.

Experiment 1 - injecting rationality-based constant values The first experiment investigates how different historical policies are affected by injecting constant-valued pseudo-data, in accordance with section 3.2. While the proof of theorem 3.5 adds correction terms, injecting data using only the uniform component of Q is a heuristic that is easily implementable by practitioners, and would only require an assumption on historical rationality. Recall that for each instance in the experimental setup, we are provided a historical dataset $\mathcal{D}_r$ that lacks coverage according to a specified rationality r. Under bandit feedback, we cannot determine if we have access to the true minimum and maximum path and arc costs in the historical data. Thus, for each dataset we suggest the following bounds: the lower bound $l = 0$, and the upper bound $u = \max_{i\in[1,\ldots,n]} C_i$.[2] Given a parameter $v \in [0,1]$, we repair coverage as follows: for each *arc* that does not appear in any path taken in the historical data, 50 new observations are injected into the dataset. For each observation, x is sampled randomly from the

[2]Even larger upper bounds could be imposed if necessary, but for our purposes we predict that no single arc will ever have a cost higher than the highest *total path cost* observed.

observed contexts in $\mathcal{D}_r$, z is the one-hot vector encoding the relevant arc, and $c = (1-v)l + vu = v \max_{i \in [1,\dots,n]} C_i$. Injecting these observations suffices to restore coverage, after which we can continue the bandit CLO pipeline.

Experiment 2 - injecting from prior distribution The second experiment investigates the effect of injecting data from a believed prior distribution Q as a function of the "quality" of Q, in accordance with section 3.1. Given a dataset $\mathcal{D}_r$, we additionally create a synthetic prior Q with a quality parameter $\gamma \in [-1, 1]$. Q behaves as follows: whenever Q is queried with some x and z, Q returns the following cost:

$$q(x, z; \gamma) = \xi + \gamma(f_0(x)^\top z - \xi)$$

where $\xi = \mathbb{E}[Y^\top Z]$ is the expected path cost over all contexts and paths in the ground truth distribution. γ can thus be interpreted as follows: $\gamma = 1$ corresponds to the case where Q is an expert oracle for the underlying distribution, $\gamma = 0$ corresponds to the case where Q is a distribution with constant cost everywhere, and $\gamma = -1$ is an "adversarial" expert that mirrors all costs across ξ. Given $\mathcal{D}$ and Q_γ, we repair coverage as follows: for each *feasible path* that does not appear in the historical data, 50 new observations are injected into the dataset. For each observation corresponding to path z, x is sampled randomly from the observed contexts in $\mathcal{D}_r$, and $c = q(x, z; \gamma)$. Injecting these observations suffices to restore coverage, after which we can continue the bandit CLO pipeline.

Baseline - regularization As a comparative baseline, for each $\mathcal{D}_r$ we additionally learn a regularized policy where no pseudo-data is injected, and $\tilde{\Sigma}(x) = \hat{\Sigma}(x) + I$.11

Remark 4.1 (Feasibility of injected data)**.** *Note that in the first experiment we inject pseudo-data corresponding to specific arcs, while in the second experiment we inject pseudo-data corresponding to specific paths. In principle, either experiment could be run with either type of data injection - depending on the problem domain, practitioners may only have priors on arcs or paths but not both. Additionally, in domains such as the shortest path problem the combinatorial nature of the search space may lead to the number of unexplored paths far exceeding the number of unexplored arcs, leading to an explosion in pseudo-dataset size. We run our experiments as described to demonstrate that both approaches are possible.*

4.2 Results

Results are shown in figure 1. When injecting constant values, we find that choosing v closer to 0 yields lower regret when rationality is low and vice versa, which aligns with theorem (3.5). When injecting from a prior distribution, we find that distributions that are more aligned with the ground truth yield lower regret, which aligns with (3.2). However, a strictly optimal historical policy leads to the lowest regret regardless of the quality of the prior distribution. We suspect that injecting data from "poor" distributions in this case may actually be widening the gap between the observed optimal routes and the second best routes which makes it easier to learn a near optimal policy and corresponds with decreasing the constant K and α in equation (6).

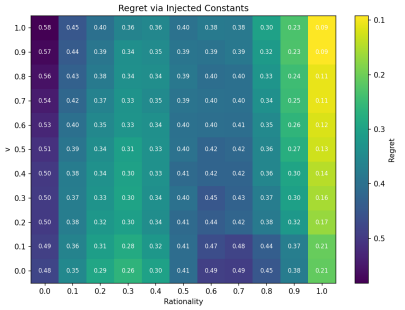
To compare against the lambda regularization baseline with $\lambda = 1$, we use the instances from experiment 1 where $v = r$ (i.e the optimal choice recommended by theorem 3.5) and the instances from experiment 2 where $\gamma = 0.6$. These correspond to instances where a practitioner is either injecting a reasonable constant given historical rationality, or injecting from a reasonably aligned prior distribution. In both cases, we find that the regularization baseline yields lower regret when rationality is low, and higher regret when rationality is high. This aligns with our intuitions that regularization encourages exploration, but that regimes of high rationality have useful structure in their historical decisions that is not exploited by regularization. Additionally we observe that for rationality approximately greater than 0.5, both injections from experiment 1 and 2 outperform regularization. This indicates that our proposed methods are attractive coverage repair schemes, when a practitioner believes their historical logging policy was at least better than a random policy. Indeed, we expect this to be the case in settings where historical decision makers have incentives to make "good" decisions (i.e., historical drivers are likely motivated to take fast routes). Full tables of results and some tangential discussions can be found in appendix B.2-B.3.

4.3 Real data experiments

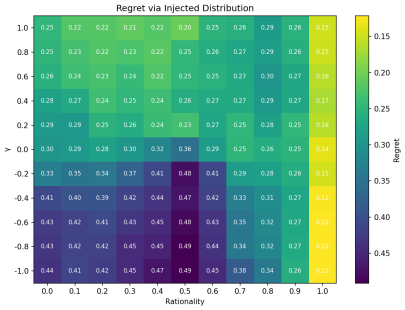
We also evaluate our methods on the real-world dataset from Uber Movement used in [9]. The dataset focuses on census tracts in downtown Los Angeles, and has 197-dimensional contexts. The road network has 93 arcs, and a total of 5902 feasible paths. Details regarding these experiments can be found in B.4 - in particular, when injecting from a prior distribution, we sample from the set of unexplored paths rather than injecting observations for all unexplored paths, to avoid the combinatorial blowup discussed in remark 4.1. When computing the normalized regrets of the learned policies, we see similar trends to the synthetic experiments. In particular, figure 2 indicates that our methods still outperform the regularization baseline when historical rationality is greater than 0.5, even when we can only inject constant values. In general, this indicates that our method produces better policies than those from regularization as long as historical decision makers were reasonably rational, *even when no other information about the system is known*. Full tables of results can be found in appendix B.5.

5 Discussion & limitations

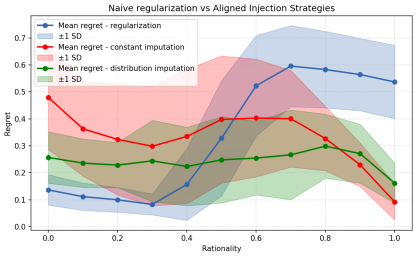
This paper studies the problem of inducing coverage in the CLO with bandit feedback setting. We propose two forms of pseudo-data injections aimed at repairing coverage using priors on the bandit data distribution or historical logging policy, and provide theoretical analysis for the regret of the post-injection policies. We acknowledge that our proposed methods have a few limitations. First, the methods are designed for practitioners with access to informative priors, which may not always be the case. Our work is directed toward cases where priors can be retrieved that are distinct from the historical data. Second, we acknowledge that constructing a dataset that injects feasible actions only on the uncovered regions of $\mathcal{Z}$ may be computationally difficult when $\mathcal{Z}$ is combinatorial. Our framework addresses this difficulty by allowing the injection of pseudo-actions outside of $\mathcal{Z}$, which can make coverage repair computationally easier at cost of reduced interpretability of the injected pseudo-actions in some cases. Third, prior-based pseudo-data injections may raise fairness concerns in sensitive domains such as healthcare or finance. If a practitioner's prior beliefs are biased, then pseudo-data injections based on those priors may reinforce historical inequalities. Fourth, pseudo-data injection based on poorly chosen priors may result in datasets that violate ignorability. While this does not pose any computational issues it does create systemic bias in the dataset that degrades regret. Note that we quantify and bound the bias within the presented regret bounds. Some interesting directions for future work *not covered* in this text include: more complex or domain-informed priors, active learning to inform the collection of coverage repair data, and investigating the bias-variance tradeoff as the number of injected data points increases.



(a) Injecting constant values.



(b) Injecting from prior distribution.



(c) Regret of policies obtained via regularization, constant imputation ($v = r$) and distribution imputation ($\gamma = 0.6$).

Figure 1: Regret of policies obtained via the described data injection and regularization procedures.

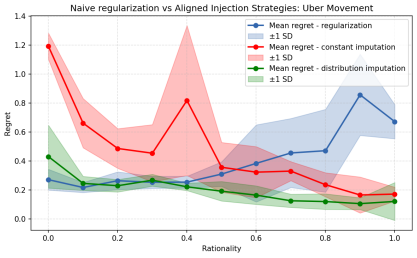


Figure 2: (Uber Movement) Regret of policies obtained via regularization, constant imputation ($v = r$) and distribution imputation ($\gamma = 0.6$).

References

[1] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs, 2012.

[2] Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

[3] Giuliano Basso. A hitchhiker ' s guide to wasserstein distances. 2015.

[4] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'11, page 2249–2257, Red Hook, NY, USA, 2011. Curran Associates Inc.

[5] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and causal parameters, 2024.

[6] Othman El Balghiti, Adam N. Elmachtoub, Paul Grigas, and Ambuj Tewari. Generalization bounds in the predict-then-optimize framework. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[7] Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1):9–26, January 2022.

[8] Yichun Hu, Nathan Kallus, and Xiaojie Mao. Fast rates for contextual linear optimization. *Management Science*, 68(6):4236–4245, 2022.

[9] Yichun Hu, Nathan Kallus, Xiaojie Mao, and Yanchen Wu. Contextual linear optimization with partial feedback, 2024.

[10] Dongming Huang, Feicheng Wang, Donald B. Rubin, and S. C. Kou. Catalytic priors: Using synthetic data to specify prior distributions in bayesian analysis, 2022.

[11] Michael Huang and Vishal Gupta. Decision-focused learning with directional gradients, 2024.

[12] Nathan Kallus. Recursive partitioning for personalization using observational data. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1789–1798. PMLR, 06–11 Aug 2017.

[13] Nathan Kallus and Xiaojie Mao. Stochastic optimization forests. *Management Science*, 69(4):1975–1994, April 2023.

[14] Kelsey Maass, Arun V. Sathanur, Arif Khan, and Robert Rallo. *Street-level Travel-time Estimation via Aggregated Uber Data*, pages 76–84.

[15] Bastian Oetomo, R. Malinga Perera, Renata Borovica-Gajic, and Benjamin I. P. Rubinstein. Cutting to the chase with warm-start contextual bandits. *Knowledge and Information Systems*, 65(9):3533–3565, April 2023.

[16] Maksim Pershin, Ivan Golovanov, Pavel Baltabaev, and Natalia Trankova. Calibration-gated llm pseudo-observations for online contextual bandits, 2026.

[17] Nick Polson and Vadim Sokolov. Synthetic priors, 2026.

[18] Utsav Sadana, Abhilash Chenreddy, Erick Delage, Alexandre Forel, Emma Frejinger, and Thibaut Vidal. A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research*, 320(2):271–289, 2025.

[19] Nian Si, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. Distributionally robust batch contextual bandits. *Management Science*, 69(10):5772–5793, October 2023.

[20] Alexander L. Strehl, John Langford, Lihong Li, and Sham M. Kakade. Learning from logged implicit exploration data. In *Proceedings of the 24th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'10, page 2217–2225, Red Hook, NY, USA, 2010. Curran Associates Inc.

[21] Y. Sun, Y. Ren, and X. Sun. Uber movement data: A proxy for average one-way commuting times by car. *ISPRS International Journal of Geo-Information*, 9(3):184, 2020.

[22] Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16(52):1731–1755, 2015.

[23] Bo Tang and Elias B Khalil. Pyepo: a pytorch-based end-to-end predict-then-optimize library for linear and integer programming. *Mathematical Programming Computation*, July 2024.

[24] Chicheng Zhang, Alekh Agarwal, Hal Daumé Iii, John Langford, and Sahand Negahban. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7335–7344. PMLR, 09–15 Jun 2019.

[25] Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183, January 2023.

Appendix

A Proofs

A.1 Proof of proposition 2.2

Proof. We proceed by first showing that the maximum-posterior estimator of f given the specified point wise prior and likelihood is equivariant to data injection then solving (3). The equivalence of solutions to (5) follows.

By Bayes' rule, $\mathrm{Prob}(f \mid \mathcal{D}) \propto \mathrm{Prob}(\mathcal{D} \mid f)\,\mathrm{Prob}(f)$. Therefore the maximum a posteriori estimator satisfies

$$\hat{f} \in \arg\max_{f \in \mathcal{F}^N} \mathrm{Prob}(f \mid \mathcal{D}) = \arg\min_{f \in \mathcal{F}^N} \Big\{ -\log \mathrm{Prob}(\mathcal{D} \mid f) - \log \mathrm{Prob}(f) \Big\}.$$

Under the likelihood assumption we have

$$\mathrm{Prob}(\mathcal{D} \mid f) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(C_i - Z_i^\top f(X_i))^2}{2\sigma^2} \right).$$

Therefore

$$\begin{aligned} -\log \mathrm{Prob}(\mathcal{D} \mid f) &= -\sum_{i=1}^{n} \log\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(C_i - Z_i^\top f(X_i))^2}{2\sigma^2} \right) \right] \\ &= \sum_{i=1}^{n} \frac{(C_i - Z_i^\top f(X_i))^2}{2\sigma^2} + \frac{n}{2} \log(2\pi\sigma^2). \end{aligned}$$

Similarly, under the prior assumption the negative log-prior is, up to an additive constant independent of f,

$$-\log \mathrm{Prob}(f) = \frac{1}{2} \sum_{i=1}^{n} (f(X_i) - \mu(X_i))^\top \Lambda(X_i)(f(X_i) - \mu(X_i)).$$

Combining the two terms, dropping additive constants, and multiplying through by $2\sigma^2$ gives

$$\hat{f} = \arg\min_{f \in \mathcal{F}^N} \left\{ \sum_{i=1}^{n} (C_i - Z_i^\top f(X_i))^2 + \sigma^2 \sum_{i=1}^{n} (f(X_i) - \mu(X_i))^\top \Lambda(X_i)(f(X_i) - \mu(X_i)) \right\}.$$

Now fix x and write $\beta = f(x)$, $\mu = \mu(x)$, and $\Lambda = \Lambda(x)$. Then the point wise objective becomes

$$\|Z\beta - C\|_2^2 + \sigma^2 (\beta - \mu)^\top \Lambda (\beta - \mu).$$

Since $\Lambda(X_i)$ is positive semi-definite it admits a decomposition $L(X_i)^\top L(X_i) = \Lambda(X_i)$ for all X_i, where $L_i \in \mathbb{R}^{m_i \times d}$; then

$$\sigma^2 (\beta - \mu)^\top \Lambda (\beta - \mu) = \sigma^2 (\beta - \mu)^\top L^\top L (\beta - \mu) = \|\sigma L \beta - \sigma L \mu\|_2^2.$$

Therefore $\hat{f}$ satisfies

$$\begin{aligned} \hat{f} &\in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \Big[(C_i - Z_i^\top f(X_i))^2 + \sigma^2 (f(X_i) - \mu(X_i))^\top \Lambda (f(X_i) - \mu(X_i)) \Big] \\ &\in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (C_i - Z_i^\top f(X_i))^2 + \sum_{i=1}^{n} \sum_{k=1}^{m_i} (\sigma L_k(X_i)^\top \mu(X_i) - \sigma L_k(X_i)^\top f(X_i))^2 \end{aligned}$$

This shows that the maximum posterior estimator is equivalent to solving problem (3) over data $\tilde{D} = D + \sum_{i=1}^{n} \sum_{k=1}^{m_i} \{(X_i, \sigma L_k(X_i), \sigma L_k(X_i)\mu(X_i)\})$, where $L_j(x)$ denotes the j^{th} row of $L(x)$.

We've established that if $\hat{f} \in \arg\max_{f \in \mathcal{F}^N} \mathrm{Prob}(f|\mathcal{D})$ then $\hat{f}$ is a solution to (3) over $\tilde{\mathcal{D}}$. We also assume that $\hat{\Sigma}$ is the same across both methods of solving (5); using MAP on $\mathcal{D}$ or MSE on $\tilde{D}$. Then $\theta(x, z, c; \hat{f}, \hat{\Sigma})$ are equivariant across both solution methods. Consequently both solution methods yield the same $\hat{\pi}$ of (5). □

A.2 Proof of proposition 2.1

Proof. Fix a constant x that appears in data $\mathcal{D}$. Let $\mathcal{I}_x = \{(X_i, Z_i, C_i) \in \mathcal{D} : x = X_i\}$ and $n_x = |\mathcal{I}_x|$. Let $\hat{f}_\lambda(x) = \beta_x$. Then the ridge regression problem is

$$\min_{\beta_x \in \mathbb{R}^d} \frac{1}{n_x} \sum_{(x, Z_i, C_i) \in \mathcal{I}_x} (C_i - Z_i^\top \beta_x)^2 + \lambda \|\beta_x\|.$$

Note that the problem furnishes some optimal β_x^* that satisfies

$$\begin{aligned}
0 &= -\frac{2}{n_x} \sum_{i \in \mathcal{I}_x} Z_i (C_i - Z_i^\top \beta_x^*) + 2\lambda \beta_x^* \\
0 &= \frac{1}{n_x} \sum_{i \in \mathcal{I}_x} Z_i (C_i - Z_i^\top \beta_x^*) - \lambda \beta_x^* \\
\left(\frac{1}{n_x} \sum_{i \in \mathcal{I}_x} Z_i Z_i^\top + \lambda I \right) \beta_x^* &= \frac{1}{n_x} \sum_{i \in \mathcal{I}_x} Z_i C_i \\
\beta_x^* &= \left(\frac{1}{n_x} \sum_{i \in \mathcal{I}_x} Z_i Z_i^\top + \lambda I \right)^{-1} \left(\frac{1}{n_x} \sum_{i \in \mathcal{I}_x} Z_i C_i \right) \\
\beta_x^* &= (\hat{\Sigma}(x) + \lambda I)^{-1} \hat{m}(x),
\end{aligned}$$

where $\hat{\Sigma}(x)$ is the nuisance function computed using empirical frequency based propensity scores and $\hat{m}(x)$ is the finite data approximation of $\mathbb{E}[ZC|X = x]$. Then for any fixed context x,

$$\begin{aligned}
\frac{1}{n_x} \sum_{i \in \mathcal{I}_x} \theta_{\mathrm{IS}}(x, Z_i, C_i; \hat{\Sigma} + \lambda I) &= (\hat{\Sigma}(x) + \lambda I)^{-1} \left(\frac{1}{n_x} \sum_{i \in \mathcal{I}_x} Z_i C_i \right) \\
&= (\hat{\Sigma}(x) + \lambda I)^{-1} \widehat{m}(x) \\
&= \hat{f}_\lambda(x) \\
&= \frac{1}{n_x} \sum_{i \in \mathcal{I}_x} \theta_{\mathrm{DM}}(x, Z_i, C_i; \hat{f}_\lambda).
\end{aligned}$$

The empirical average of the inverse-score-style estimator agrees exactly with the direct ridge estimator at each observed context x. Therefore solving problem (5) is equivalent for both choices of θ. Note that ridge regression is a special case of using point-wise priors defined in proposition 2.2 where $\Lambda(x) = I$, $\mu(x) = 0$, and $\sigma = \sqrt{\lambda}$. Therefore using ridge regression is equivalent to injecting data $\sum_{i=1}^n \{(X_i, \sqrt{\lambda} e_k, 0)\}_{k=1}^d$. □

A.3 Proof of theorem 3.2

Proof. Let π_0 be the policy induced by $f_0^{true}(x) = \mathbb{E}_{(X,Y)}[Y|X = x]$. Regret of policy $\tilde{\pi}$ is defined as

$$\mathrm{Reg}(\tilde{\pi}) = \mathbb{E}_P[f_0^{true}(X)^\top \tilde{\pi}(X) - f_0^{true}(X)^\top \pi_0(X)].$$

Recall Q is chosen such that $\tilde{P}$ satisfies coverage on the support S where $\mathrm{span}(\mathcal{Z}) \subseteq \mathrm{span}(S) \subseteq \mathbb{R}^d$. Consider the projection of $f_0(x)$ on to the $\mathrm{span}(\mathcal{Z})$. The resulting function yields the same policy $\pi_0(x) = \min_{z \in \mathcal{Z}} z^\top \Pi_{\mathcal{Z}} f_0^{true}(x)$, where Π_A is the projection operator onto a set A. Without loss of generality we assume that $f_0(x) = \Pi_{\mathcal{Z}} f_0^{true}(x)$. Regret can then be identically written as

$$\mathrm{Reg}(\tilde{\pi}) = \mathbb{E}_P[f_0(X)^\top \tilde{\pi}(X) - f_0(X)^\top \pi_0(X)].$$

Define the random variable $A(X) = \mathbb{I}\{\tilde{\pi}(X) \neq \pi_0(X)\}$ where $\mathbb{I}\{E\}$ is 1 if the event E takes place 0 otherwise. If $A(x) = 0$ then $\tilde{\pi}(x) = \pi_0(x)$ and $f_0^\top(X)\tilde{\pi}(X) - f_0(X)^\top \pi_0(X) = 0$ so we can equivalently write regret as

$$\mathrm{Reg}(\tilde{\pi}) = \mathbb{E}_P\Big[A(X) \cdot [f_0^\top(X)\tilde{\pi}(X) - f_0(X)^\top \pi_0(X)]\Big].$$

The function $\tilde{f}$ is found by solve (3) over data of size $n+m$ drawn from $\eta P+(1-\eta)Q$. Note that for any minimizer of (3), $\Pi_S \tilde{f}(x)$ is also a minimizer. We use the convention that the minimizer $\tilde{f}$ selected from the set of minimizers lies in S for all dataset sizes including the case where $n, m \to \infty$. Adding as subtracting $A(X)\tilde{f}(X)^\top \tilde{\pi}(X)$ yields

$$\mathrm{Reg}(\tilde{\pi}) = \mathbb{E}_P\Big[A(X)\cdot(f_0(X)-\tilde{f}(X))^\top\tilde{\pi}(X)\Big] + \mathbb{E}_P\Big[A(X)\cdot[\tilde{f}^\top(X)\tilde{\pi}(X) - f_0(X)^\top\pi_0(X)]\Big].$$

Adding and subtracting $A(X)\tilde{f}(X)\pi_0(X)$ yields

$$\begin{aligned}\mathrm{Reg}(\tilde{\pi}) = \mathbb{E}_P\Big[A(X)\cdot(f_0(X)-\tilde{f}(X))^\top\tilde{\pi}(X)\Big] + \\ \mathbb{E}_P\Big[A(X)\cdot[\tilde{f}^\top(X)\tilde{\pi}(X) - \tilde{f}(X)^\top\pi_0(X)]\Big] + \\ + \mathbb{E}_P\Big[A(X)\cdot(\tilde{f}(X)-f_0(X))^\top\pi_0(X)]\Big].\end{aligned}$$

Since $\tilde{\pi}$ is the policy induced by $\tilde{f}$, we have that $\tilde{f}(X)^\top z \geq \tilde{f}(X)\tilde{\pi}(X)$ almost surely so the middle term is ≤ 0 almost surely and we get

$$\mathrm{Reg}(\tilde{\pi}) \leq \mathbb{E}_P\Big[A(X)\cdot(f_0(X)-\tilde{f}(X))^\top\tilde{\pi}(X)\Big] + \mathbb{E}_P\Big[A(X)\cdot(\tilde{f}(X)-f_0(X))^\top\pi_0(X)]\Big].$$

Applying Cauchy Schwartz twice yields

$$\begin{aligned}\mathrm{Reg}(\tilde{\pi}) &\leq \mathbb{E}_P\Big[A(X)\cdot(f_0(X)-\tilde{f}(X))^\top\tilde{\pi}(X)\Big] + \mathbb{E}_P\Big[A(X)\cdot(\tilde{f}(X)-f_0(X))^\top\pi_0(X)]\Big] \\ &= \mathbb{E}_P\Big[A(X)\cdot(f_0(X)-\tilde{f}(X))^\top(\tilde{\pi}(X)-\pi_0(X))\Big] \\ &\leq \mathbb{E}_P\Big[|A(X)|\cdot|(f_0(X)-\tilde{f}(X))^\top(\tilde{\pi}(X)-\pi_0(X))|\Big] \\ &\leq \mathbb{E}_P\Big[|A(X)|\cdot\|(f_0(X)-\tilde{f}(X))\|_2\|(\tilde{\pi}(X)-\pi_0(X))\|_2\Big].\end{aligned}$$

Recall that $\sup_{z\in\mathcal{Z}}\|z\|_2 \leq B$ and applying Cauchy Schwartz over the expectation yields

$$\mathrm{Reg}(\tilde{\pi}) \leq 2B\mathbb{E}_P[A(X)]^{1/2}\cdot\mathbb{E}_P[\|f_0(X)-\tilde{f}(X)\|_2^2]^{1/2} \tag{7}$$

Next we bound $\mathbb{E}_P[A(X)] = \mathrm{Prob}_P(\tilde{\pi}(X)\neq\pi_0(X))$. Note that $\tilde{\pi}(X)\neq\pi_0(X)$ then

$$\tilde{f}(X)\tilde{\pi}(X) - \tilde{f}(X)\pi_0(X) < 0$$

since the event occurs if an action appears better under $\tilde{f}$ then under f_0. Also note that

$$f_0(X)^\top\tilde{\pi}(X) - f_0(X)^\top\pi_0(X) \geq \gamma(X),$$

where

$$\gamma(X) = \min_{z\neq\pi_0(X)}(f_0(X)^\top z - f_0(X)^\top\pi_0(X)).$$

Subtracting the two previous inequalities yields

$$A(X)=1 \implies (f_0(X)-\tilde{f}(X))^\top(\tilde{\pi}(X)-\pi_0(X)) \geq \gamma(X).$$

Applying Cauchy Schwartz and the bound on $\|z\|_2 \leq B$ yields

$$\mathrm{Prob}_P(A(X)=1) \leq \mathrm{Prob}_P\left(\|\tilde{f}(X)-f_0(X)\|_2 \geq \frac{\gamma(X)}{2B}\right).$$

We can further bound this probability so that for any $t>0$,

$$\mathrm{Prob}_P(A(X)=1) \leq \mathrm{Prob}_P(\gamma(X)\leq t) + \mathrm{Prob}_P\left(\|\tilde{f}(X)-f_0(X)\|_2 \geq \frac{t}{2B}\right).$$

Applying the margin assumption 3.1 on the left term and Markov's inequality on the right term yields

$$\mathrm{Prob}_P(A(X)=1) \leq Ct^\alpha + \frac{2B}{t}\mathbb{E}_P[\|\tilde{f}(X)-f_0(X)\|_2].$$

Since the bound holds for all $t > 0$ we can minimize t to retrieve the tightest bound. To that end, define $\phi(t) = Ct^\alpha + \frac{2B}{t}E$ and $E = \mathbb{E}_P[\|\tilde{f}(X) - f_0(X)\|_2]$. Using the first order condition yields

$$\phi'(t) = \alpha C t^{\alpha-1} - \frac{2BE}{t^2} = 0 \implies t^* = \left(\frac{2BE}{\alpha C}\right)^{1/(\alpha+1)}$$

Plugging t^* back in to the probability bound yields

$$\mathrm{Prob}_P(A(X) = 1) \leq K\mathbb{E}_P[\|\tilde{f}(X) - f_0(X)\|_2]^{\alpha/(\alpha+1)}, \quad K = (1+\alpha)\Big(\frac{C(2B)^\alpha}{\alpha^\alpha}\Big)^{1/(\alpha+1)}$$

Plugging the probability bound back into equation (7) gives

$$\mathrm{Reg}(\tilde{\pi}) \leq K^{1/2}\mathbb{E}_P[\|\tilde{f}(X) - f_0(X)\|_2]^{\frac{\alpha}{2(\alpha+1)}}\mathbb{E}_P[\|f_0(X) - \tilde{f}(X)\|_2^2]^{1/2}.$$

Using Jensens inequality yields

$$\begin{aligned}\mathrm{Reg}(\tilde{\pi}) &\leq 2BK^{1/2}\mathbb{E}_P[\|\tilde{f}(X) - f_0(X)\|_2^2]^{\frac{\alpha}{4(\alpha+1)}}\mathbb{E}_P[\|f_0(X) - \tilde{f}(X)\|_2^2]^{1/2} \\ &= 2BK^{1/2}\mathbb{E}_P[\|f_0(X) - \tilde{f}(X)\|_2^2]^{\frac{3\alpha+2}{4(\alpha+1)}}.\end{aligned} \tag{8}$$

Next we bound $\mathbb{E}_P[\|\tilde{f}(X) - f_0(X)\|_2^2]$. First note that if the marginal distribution of X under P, Q, and $\tilde{P}$ are identical by assumption, then we can write $\mathbb{E}_P[\|\tilde{f}(X) - f_0(X)\|_2^2] = \mathbb{E}_{\tilde{P}}[\|\tilde{f}(X) - f_0(X)\|_2^2]$. Let $\tilde{f}^*$ be the population estimator, i.e. $\tilde{f}^* \in \arg\min_{f\in\mathcal{F}^N}\mathbb{E}_{\tilde{P}}[(C - Z^\top f(X))^2]$. Recall, we can assume that $\tilde{f}^*(x)$ lies in S without loss of generality. Now consider the event

$$\mathcal{E}_\delta = \left\{\tilde{\mathcal{D}} \;\middle|\; \mathbb{E}_{\tilde{P}}[\|\tilde{f}^*(X) - \tilde{f}(X)\|_2^2] \leq \mathrm{Rate}_{\mathcal{F}^N}(n, m, \delta)\right\},$$

where $\mathrm{Rate}_{\mathcal{F}^N}$ is defined such that $\mathrm{Prob}(\mathcal{E}_\delta) \geq 1 - \delta$. Adding and subtracting $\tilde{f}^*$ and applying the triangle inequality yields

$$\begin{aligned}\mathbb{E}_{\tilde{P}}[\|\tilde{f}(X) - f_0(X)\|_2^2]^{1/2} &\leq \mathbb{E}_{\tilde{P}}[\|\tilde{f}(X) - \tilde{f}^*(X)\|_2^2]^{1/2} + \mathbb{E}_{\tilde{P}}[\|\tilde{f}^*(X) - f_0(X)\|_2^2]^{1/2} \\ &\leq \mathrm{Rate}_{\mathcal{F}^N}(n, m, \delta)^{1/2} + \mathbb{E}_{\tilde{P}}[\|\tilde{f}^*(X) - f_0(X)\|_2^2]^{1/2}\end{aligned}$$

with probability at least $1 - \delta$. To bound $\mathbb{E}_{\tilde{P}}[\|\tilde{f}^*(X) - f_0(X)\|_2^2]$ define

$$\begin{aligned}\tilde{\mu}(x) &= (1-\eta)\mathbb{E}_P[ZC \mid X = x] + \eta\mathbb{E}_Q[ZC \mid X = x], \\ \tilde{\Sigma}(x) &= (1-\eta)\mathbb{E}_P[ZZ^\top \mid X = x] + \eta\mathbb{E}_Q[ZZ^\top \mid X = x].\end{aligned}$$

Since P satisfies ignorability we know that

$$\mu_0(x) = \mathbb{E}_P[ZC \mid X = x] = \mathbb{E}_P[ZZ^\top \mid X = x]f_0(x) = \Sigma_0(x)f_0(x).$$

Because $\tilde{f}^*$ is the population least squares minimizer under $\tilde{P}$, it satisfies

$$\begin{aligned}0 &= \nabla_f\mathbb{E}_{\tilde{P}}[(C - Z^\top f^*_{pd}(x))^2 \mid X = x] \\ 0 &= -2\mathbb{E}_{\tilde{P}}[Z(C - Z^\top f^*_{pd}(x)) \mid X = x] \\ \mathbb{E}_{\tilde{P}}[ZC \mid X = x] &= \mathbb{E}_{\tilde{P}}[ZZ^\top f^*_{pd}(x) \mid X = x] \\ \tilde{\mu}(x) &= \tilde{\Sigma}(x)f^*_{pd}(x).\end{aligned}$$

Combining this result with the definition of $\tilde{\Sigma}$ yields

$$(1-\eta)\Sigma_0(x)f_0(x) + \eta\mathbb{E}_Q[ZC \mid X = x] = \big((1-\eta)\Sigma_0(x) + \eta\mathbb{E}_Q[ZZ^\top \mid X = x]\big)\tilde{f}^*(x).$$

Now define $\epsilon(x) = \mathbb{E}_Q[Z(C - Z^\top f_0(x))|X = x]$. Then

$$\mathbb{E}_Q[ZC \mid X = x] = \mathbb{E}_Q[ZZ^\top \mid X = x]f_0(x) + \epsilon(x),$$

so

$$(1-\eta)\Sigma_0(x)f_0(x) + \eta\big(\mathbb{E}_Q[ZZ^\top \mid X=x]f_0(x) + \epsilon(x)\big)$$
$$= \big((1-\eta)\Sigma_0(x) + \eta\mathbb{E}_Q[ZZ^\top \mid X=x]\big)\tilde{f}^*(x).$$

Rearranging gives

$$\tilde{\Sigma}(x)\big(\tilde{f}^*(x) - f_0(x)\big) = \eta\,\epsilon(x).$$

Recall that $\tilde{f}^*(x), f_0(x) \in \operatorname{span}(S)$ and $\tilde{\Sigma}(x)$ is invertible on the span of S since $\tilde{\Sigma}(x) = \mathbb{E}_{\tilde{P}}[ZZ^\top|X=x]$. Consequently, $\tilde{f}^*(x) - f_0(x) \in \operatorname{span}(S)$ and

$$\tilde{f}^*(x) - f_0(x) = \eta\tilde{\Sigma}(x)^\dagger\epsilon(x),$$

where $\dagger$ denotes the Moore-Penrose inverse. Take the norm of both sides, applying Cauchy Schwartz and squaring yields

$$\|\tilde{f}^*(x) - f_0(x)\|_2^2 \le \eta^2\|\tilde{\Sigma}(x)^\dagger\|_2^2\|\epsilon(x)\|_2^2 \le \frac{\eta^2}{\lambda_{\min}^+(\tilde{\Sigma}(x))^2}\|\epsilon(x)\|_2^2.$$

Taking the expectation with respect to $\tilde{P}$ yields

$$\mathbb{E}_{\tilde{P}}\|\tilde{f}^*(X) - f_0(X)\|_2^2] \le \eta^2\mathbb{E}_Q\left[\frac{\|\epsilon(X)\|_2^2]}{\lambda_{\min}^+(\tilde{\Sigma}(X))^2}\right]$$

since $E_{\tilde{P}}[\mathbb{E}_Q[\cdot|X=x]] = \mathbb{E}_Q[\cdot]$ due to the equivalence of X marginals. Plugging this bound back into equation (8) gives

$$\operatorname{Reg}(\tilde{\pi}) \le 2BK^{1/2}\left(\operatorname{Rate}_{\mathcal{F}^N}(n,m,\delta)^{1/2} + \left(\eta^2\mathbb{E}_Q\left[\frac{\|\epsilon(X)\|_2^2]}{\lambda_{\min}^+(\tilde{\Sigma}(X))^2}\right]\right)^{1/2}\right)^{\frac{3\alpha+2}{4(\alpha+1)}}.$$

Let $\lambda_{\min}^+(\tilde{\Sigma}) = \operatorname{ess\,inf}_x \lambda_{\min}^+\tilde{\Sigma}(x)$ which is > 0 since $\tilde{\Sigma}(X)$ satisfies coverage and is as a result invertable over the span of $\mathcal{Z}$ and has at least one positive eigen value. The bound can then be rewritten as

$$\operatorname{Reg}(\tilde{\pi}) \le 2BK^{1/2}\left(\operatorname{Rate}_{\mathcal{F}^N}(n,m,\delta)^{1/2} + \frac{\eta}{\lambda_{\min}^+(\tilde{\Sigma})}\mathbb{E}_Q\Big[\|\epsilon(X)\|_2^2]\Big]^{1/2}\right)^{\frac{3\alpha+2}{4(\alpha+1)}}. \tag{9}$$

The term $\mathbb{E}_Q[\|\epsilon(X)\|_2^2]$ can be further simplifies using Jensens inequality we yields

$$\mathbb{E}_Q[\|\epsilon(X)\|_2^2] = \mathbb{E}_Q[\|\mathbb{E}_Q[Z(C - Z^\top f_0(x))|X=x]\|_2^2] \le \mathbb{E}_Q[\|Z(C - Z^\top f_0(X))\|_2^2].$$

Plugging this into the bound yields the result

$$\operatorname{Reg}(\tilde{\pi}) \le 2BK^{1/2}\left(\operatorname{Rate}_{\mathcal{F}^N}(n,m,\delta)^{1/2} + \frac{\eta}{\lambda_{\min}^+(\tilde{\Sigma})}\mathbb{E}_Q\Big[\|Z(C - Z^\top f_0(X))\|_2^2\Big]^{1/2}\right)^{\frac{3\alpha+2}{4(\alpha+1)}}.$$

□

A.4 Wasserstein corollary of theorem 3.2

Corollary A.1. *Given all assumptions in theorem 3.2 and if $h(z,c;x) = z(c - z^\top f_0(x))$ is Lipchitz on z, c with constant $\bar{L}$ almost surely then with probability at least $1-\delta$ we have*

$$\operatorname{Reg}(\tilde{\pi}) \le 2BK^{1/2}\left(\operatorname{Rate}_{\mathcal{F}^N}(n,m,\delta)^{1/2} + \frac{\eta\bar{L}}{\lambda_{\min}^+(\tilde{\Sigma})}\mathbb{E}_Q\Big[W_1\Big(Q(Z,C|X), P(Z,C|X)\Big)\Big]\right)^{\frac{3\alpha+2}{4(\alpha+1)}}$$

where W_1 is the 1-Wasserstein distance.

Proof. Since P satisfies ignorability, $u^\top \mathbb{E}_P[Z(C - Z^\top f_0(X))|X] = 0$. So for each fixed x we can write

$$u^\top \mathbb{E}_Q[Z(C - Z^\top f_0(X))|X = x] = \mathbb{E}_Q[u^\top h(z, c; x)|X = x] - \mathbb{E}_P[u^\top h(z, c; x)|X = x], \quad (10)$$

where u is fixed with $\|u\|_2 = 1$.

Recall that by assumption h is $\bar{L}$ lipschitz and so $u^\top h$ is still $\bar{L}$ Lipschitz. So by Kantorovich–Rubinstein duality [3], the 1-Wasserstein distance can be written as

$$W_1(P, Q) = \sup_{h:|h(x)-h(y)|\leq\|x-y\|_2} \Big(\mathbb{E}_Q[u^\top h(z, c; x)] - \mathbb{E}_P[u^\top h(z, c; x)]\Big) \quad \forall u : \|u\|_2 = 1.$$

Consequently, taking the supremum over h on the left hand side of equation (10) gives

$$u^\top \mathbb{E}_Q[Z(C - Z^\top f_0(X))|X = x] \leq \bar{L}\mathbb{E}_Q\Big[W_1\Big(Q(Z, C|X), P(Z, C|X)\Big)\Big].$$

Take the supremum of the right hand side over all u such that $\|u\|_2 = 1$, then taking the expectation with respect to Q yields the bound

$$\|\mathbb{E}_Q[Z(C - Z^\top f_0(X))|X = x]\|_2 \leq \bar{L}\mathbb{E}_Q\Big[W_1\Big(Q(Z, C|X), P(Z, C|X)\Big)\Big].$$

Squaring and taking the expectation with respect to Q yields

$$\mathbb{E}_Q\Big[\|\mathbb{E}_Q[Z(C - Z^\top f_0(X))|X = x]\|_2^2\Big] \leq \bar{L}^2\mathbb{E}_Q\Big[W_1\Big(Q(Z, C|X), P(Z, C|X)\Big)\Big]^2.$$

Plugging into equation (9) yields

$$\mathrm{Reg}(\tilde{\pi}) \leq 2BK^{1/2}\left(\mathrm{Rate}_{\mathcal{F}^N}(n, m, \delta)^{1/2} + \frac{\eta\bar{L}}{\lambda^+_{\min}(\tilde{\Sigma})}\mathbb{E}_Q\Big[W_1\Big(Q(Z, C|X), P(Z, C|X)\Big)\Big]\right)^{\frac{3\alpha+2}{4(\alpha+1)}}$$

□

A.5 Proof of theorem 3.5

Proof. Assume that any action z such that $Q(z|x, c) > 0$ satisfies $\|z\|_2 \leq B$. So $(X, Z, C) \sim Q$ necessarily implies that

$$\mathbb{E}_Q\Big[\|Z(M - Z^\top f_0(X))\|_2^2\Big] = \mathbb{E}_Q[\|Z\|_2^2(M - Z^\top f_0(X))^2] \leq B^2\mathbb{E}_Q[(M - Z^\top f_0(X))^2].$$

Minimizing the term with respect to M using first order conditions yields

$$2B^2\mathbb{E}_Q[(M^* - Z^\top f_0(X))] = 0 \implies M^* = \mathbb{E}_Q[Z^\top f_0(X)].$$

Plugging M^* back into equation (9) gives the presented regret bound.

Next we show that there exists a Q such that M^* minimizes the right hand side of the regret bound, which entails finding a Q such that $\mathbb{E}_Q[Z^\top f_0(X)] = M^*$ and $\eta Q + (1 - \eta)P$ satisfies coverage. Consider, the set of directions $v_1(x), \ldots, v_s(x) \in \mathrm{span}(\mathcal{Z})$ such that

$$\mathrm{span}(\mathcal{Z}) \subseteq \mathrm{span}\Big(\mathrm{supp}(P_{Z|X=x}) \cup \{v_1(x), \ldots, v_s(x)\}\Big), \quad P_X - a.s.$$

Define Q^{cov} such that Q^{cov} satisfies $Q^{cov}_X = P_X$ and $Q^{cov}(Z = v_k(X)|X) = 1/s$ for all $k = 1, \ldots, s$. If $\mathbb{E}_{Q^{cov}}[Z^\top f_0(X)] = M^*$ then set $Q = Q^{cov}$ and we're done, since by construction $\eta Q^{cov} + (1 - \eta)P$ satisfies coverage for $\eta > 0$. Otherwise, choose ρ such that

$$\rho = \sup\left\{r \in [0, 1] : \frac{M^* - r\mu_{cov}}{1 - r} \in \Big(\mathbb{E}_P[\min_{z\in\mathcal{Z}} f_0(X)^\top z], \mathbb{E}_P[\max_{z\in\mathcal{Z}} f_0(X)^\top z]\Big)\right\},$$

where $\mu_{cov} = \mathbb{E}_{Q^{cov}}[Z^\top f_0(X)]$. We define $T_\rho := (M^* - \rho\mu_{cov})/(1 - \rho)$. Also define $Q^{\min}$ and $Q^{\max}$ such that $Q^{\min}(z \in \min_{z\in\mathcal{Z}} f_0(X)^\top z|X) = 1$ and $Q^{\min}(z \in \max_{z\in\mathcal{Z}} f_0(X)^\top z|X) = 1$ almost surely on P_X. Choose

$$p_\rho := \frac{T_\rho - \mathbb{E}_P[\min_{z\in\mathcal{Z}} f_0(X)^\top z]}{\mathbb{E}_P[\max_{z\in\mathcal{Z}} f_0(X)^\top z - \min_{z\in\mathcal{Z}} f_0(X)^\top z]} \in (0, 1).$$

Define

$$Q = \rho Q^{cov} + (1-\rho)(p_\rho Q^{\max} + (1-p_\rho)Q^{\min}).$$

Then

$$E_Q[Z^\top f_0(X)] = p\mu_{cov} + (1-\rho)(p_\rho \mathbb{E}_P[\max_{z\in\mathcal{Z}} f_0(X)^\top z] + (1-p_\rho)$$
$$\mathbb{E}_P[\min_{z\in\mathcal{Z}} f_0(X)^\top z]) = \rho\mu_{cov} + (1-\rho) = M^*.$$

So M^* is the unique minimizer of the right hand side of the regret bound. Again we can conclude that by construction Q gives positive probability mass to each $v_k(x)$ so the span of the support of $\tilde{P}_{Z|X}$ is equal to the span of $\mathcal{Z}$. Thus, $(1-\eta)P + \eta Q$ satisfies coverage.

□

A.6 Specialization of regret bound (6) to lambda regularization

Let C and α satisfy assumption 3.1 and P be the distribution induced by logging policy $\pi_{\log}$ that does not satisfy coverage. Assume that we have a dataset $\mathcal{D} \sim P^{\otimes n}$. Recall by proposition 2.1 that ridge regression corresponds to using an augmenting data set $\sum_{i=1}^n \{(X_i, \sqrt{\lambda}e_k, 0)\}_{k=1}^d$. This corresponds to drawing pseudo-data from a distribution Q where $Q_X = P_X$, $Q(C=0|X,Z) = 1$ almost surely, and $Q(Z|X,C) \sim \text{Uniform}(\sqrt{\lambda}e_1, \ldots, \sqrt{\lambda}e_d)$ where e_i are standard basis vectors of $\mathbb{R}^d$. In this setting $\eta = nd/(n+nd) = d/(d+1)$ and $\tilde{\Sigma}(x) = (1-\eta)\Sigma_P + (\eta\lambda/d)I$ so $\lambda^+_{\min}(\tilde{\Sigma}) = \eta\lambda/d$. The pseudo-data bias term of (6) then reduces to

$$\frac{\eta}{\lambda^+_{\min}(\tilde{\Sigma})}\mathbb{E}_Q\left[\|Z(C - Z^\top f_0(X))\|_2^2\right]^{\frac{1}{2}} = \frac{d}{\lambda}\mathbb{E}_Q\left[\|ZZ^\top f_0(X)\|_2^2\right]^{\frac{1}{2}}$$
$$= \frac{d}{\lambda}\cdot\sqrt{\frac{\lambda^2}{d}}\mathbb{E}_Q[\|f_0(X)\|_2^2]^{\frac{1}{2}} = \sqrt{d}\,\mathbb{E}_Q[\|f_0(X)\|_2^2]^{\frac{1}{2}}$$

Note that because $P_X = Q_X$ the final form of the pseudo-data bias can be written as an expectation over P.

B Additional experimental details

The following section provides additional implementation details for the experiments provided in the paper. All experiments were implemented on a cloud computing platform with a mixture of the following compute nodes:

- AMD EPYC 7513 CPU @ 2.60 GHz, 248 GB RAM
- AMD EPYC 7542 CPU @ 2.90 GHz, 248 GB RAM
- INTEL XEON Silver 4116 CPU @ 2.10 GHz, 185 GB RAM
- INTEL XEON Silver 4116 CPU @ 2.10 GHz, 89 GB RAM

B.1 Implementation details

Covariates $X = (X_1, X_2, X_3)^\top \in \mathbb{R}^3$ are generated i.i.d. from independent standard normal distribution. The full feedback vector $Y \in \mathbb{R}^{40}$ is simulated according to $Y = f_0 + \epsilon$, where $f^*(X) = a + W_1X_1 + W_2X_2 + W_3X_3$ and ϵ is random noise drawn from the uniform distribution $U[-0.5, 0.5]$. Each coefficient vector $W_i \in \mathbb{R}^{40}$ is generated with elements drawn independently from $\text{Uniform}[0,1]$, and the bias vector $a \in \mathbb{R}^{40}$ has elements drawn from $\mathcal{N}(2 + i/10, 1)$. Notably, the bias vector a has dependence on the arc index to encode geographical information into the graph beyond the random draws of W: edge 0 has its bias drawn from $\mathcal{N}(2,1)$, whereas edge 39 has its bias drawn from $\mathcal{N}(5.9, 1)$. This is done so that, for example, an optimal historical logging policy will generally avoid edges with high indices when possible, and an "anti-optimal" historical policy

will generally avoid edges with low indices, giving us regions of coverage that meaningfully depend on historical rationality.

We perform two nuisance estimations to compute the score function $\theta(x, z, c; f, \Sigma)$. Firstly, we estimate $\hat{f}$ via least squares regression, leaving the hypothesis space correctly specified (i.e., $\hat{f}$ is linear). To learn the conditional Gram matrix $\hat{\Sigma}$, we use a logistic regression to estimate the propensity scores $\hat{e}(z|X) = \mathbb{P}(Z = z|X)$ for each feasible x, then estimate $\hat{\Sigma}(X) = \sum_{j=1}^{m} z_j z_j^\top \hat{e}(z_j|X)$. For all experiments, 2000 historical datapoints are generated, with a train-test split of $[0.75, 0.25]$. Sizes of injected datasets are specified in the relevant experiments.

B.2 Full synthetic results

Full tabular results from figure 1 are included in tables 1-3:

Table 1: Regret via injected constants. Mean regret (standard deviation) across runs, as a function of rationality (columns) and value ratio v (rows). Lower regret indicates better performance.

	Rationality										
v	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1.0	0.58 (0.22)	0.45 (0.21)	0.40 (0.23)	0.36 (0.24)	0.36 (0.23)	0.40 (0.23)	0.38 (0.18)	0.38 (0.18)	0.30 (0.11)	0.23 (0.08)	0.09 (0.07)
0.9	0.57 (0.22)	0.44 (0.21)	0.39 (0.21)	0.34 (0.21)	0.35 (0.23)	0.39 (0.22)	0.39 (0.20)	0.39 (0.17)	0.32 (0.12)	0.23 (0.08)	0.09 (0.07)
0.8	0.56 (0.22)	0.43 (0.20)	0.38 (0.23)	0.34 (0.22)	0.34 (0.24)	0.39 (0.24)	0.40 (0.21)	0.40 (0.17)	0.33 (0.12)	0.24 (0.08)	0.11 (0.07)
0.7	0.54 (0.22)	0.42 (0.20)	0.37 (0.21)	0.33 (0.24)	0.35 (0.25)	0.39 (0.24)	0.40 (0.20)	0.40 (0.18)	0.34 (0.13)	0.25 (0.07)	0.11 (0.07)
0.6	0.53 (0.21)	0.40 (0.20)	0.35 (0.22)	0.33 (0.23)	0.34 (0.24)	0.40 (0.24)	0.40 (0.22)	0.41 (0.18)	0.35 (0.13)	0.26 (0.09)	0.12 (0.07)
0.5	0.51 (0.20)	0.39 (0.19)	0.34 (0.20)	0.31 (0.22)	0.33 (0.25)	0.40 (0.23)	0.42 (0.23)	0.42 (0.18)	0.36 (0.12)	0.27 (0.08)	0.13 (0.07)
0.4	0.50 (0.19)	0.38 (0.17)	0.34 (0.21)	0.30 (0.22)	0.33 (0.25)	0.41 (0.24)	0.42 (0.24)	0.42 (0.19)	0.36 (0.12)	0.30 (0.09)	0.14 (0.07)
0.3	0.50 (0.19)	0.37 (0.17)	0.33 (0.22)	0.30 (0.22)	0.34 (0.25)	0.40 (0.25)	0.45 (0.24)	0.43 (0.18)	0.37 (0.13)	0.30 (0.10)	0.16 (0.07)
0.2	0.50 (0.19)	0.38 (0.20)	0.32 (0.20)	0.30 (0.22)	0.34 (0.25)	0.41 (0.24)	0.44 (0.23)	0.42 (0.18)	0.38 (0.12)	0.32 (0.10)	0.17 (0.07)
0.1	0.49 (0.19)	0.36 (0.17)	0.31 (0.19)	0.28 (0.20)	0.32 (0.23)	0.41 (0.26)	0.47 (0.25)	0.48 (0.19)	0.44 (0.14)	0.37 (0.11)	0.21 (0.09)
0.0	0.48 (0.19)	0.35 (0.17)	0.29 (0.19)	0.26 (0.19)	0.30 (0.23)	0.41 (0.26)	0.49 (0.25)	0.49 (0.18)	0.45 (0.14)	0.38 (0.11)	0.21 (0.09)

Table 2: Regret via injection from prior distribution. Mean regret (standard deviation) across runs, as a function of rationality (columns) and prior quality γ (rows). Lower regret indicates better performance.

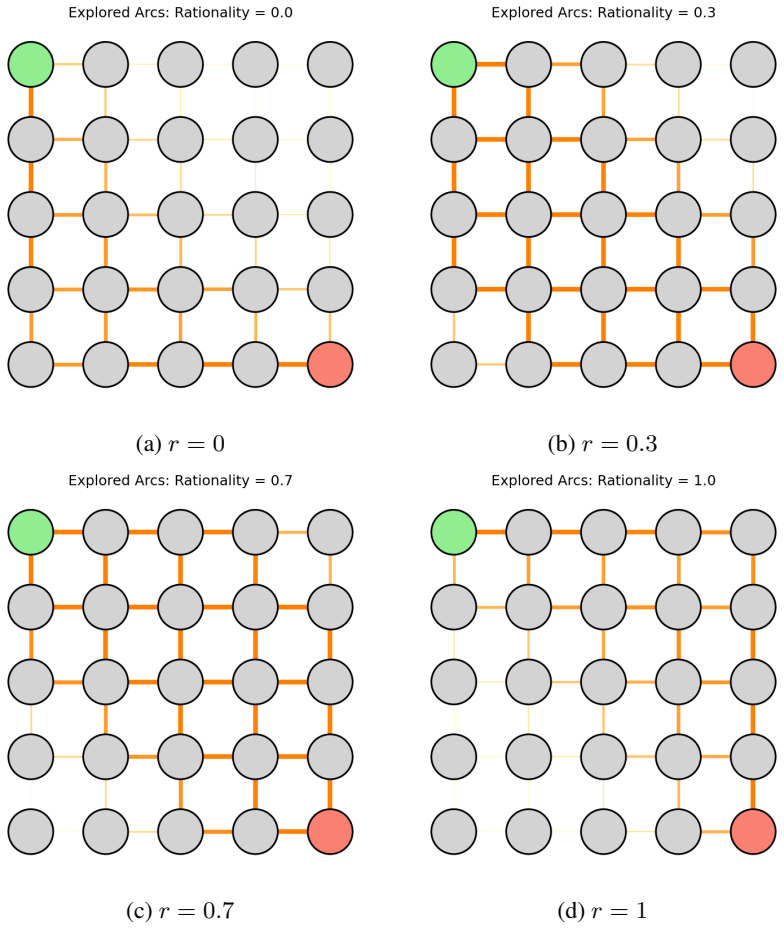
	Rationality										
γ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1.0	0.25 (0.09)	0.22 (0.08)	0.22 (0.08)	0.21 (0.11)	0.22 (0.11)	0.20 (0.09)	0.25 (0.13)	0.26 (0.13)	0.29 (0.11)	0.26 (0.11)	0.15 (0.08)
0.8	0.25 (0.12)	0.23 (0.08)	0.22 (0.10)	0.23 (0.13)	0.22 (0.10)	0.25 (0.14)	0.26 (0.15)	0.27 (0.14)	0.29 (0.13)	0.26 (0.10)	0.15 (0.09)
0.6	0.26 (0.09)	0.24 (0.09)	0.23 (0.08)	0.24 (0.15)	0.22 (0.14)	0.25 (0.16)	0.25 (0.14)	0.27 (0.16)	0.30 (0.12)	0.27 (0.11)	0.16 (0.07)
0.4	0.28 (0.10)	0.27 (0.10)	0.24 (0.12)	0.25 (0.13)	0.24 (0.14)	0.26 (0.14)	0.27 (0.14)	0.27 (0.13)	0.29 (0.13)	0.27 (0.12)	0.17 (0.08)
0.2	0.29 (0.11)	0.29 (0.14)	0.25 (0.10)	0.26 (0.14)	0.24 (0.12)	0.23 (0.14)	0.27 (0.15)	0.25 (0.15)	0.28 (0.13)	0.25 (0.10)	0.16 (0.08)
0.0	0.30 (0.14)	0.29 (0.11)	0.28 (0.12)	0.30 (0.14)	0.32 (0.18)	0.36 (0.12)	0.29 (0.16)	0.25 (0.15)	0.26 (0.13)	0.25 (0.15)	0.14 (0.06)
-0.2	0.33 (0.14)	0.35 (0.12)	0.34 (0.12)	0.37 (0.15)	0.41 (0.17)	0.48 (0.17)	0.41 (0.18)	0.29 (0.13)	0.28 (0.11)	0.26 (0.15)	0.15 (0.06)
-0.4	0.41 (0.14)	0.40 (0.12)	0.39 (0.13)	0.42 (0.16)	0.44 (0.18)	0.47 (0.17)	0.42 (0.17)	0.33 (0.14)	0.31 (0.12)	0.27 (0.15)	0.12 (0.04)
-0.6	0.43 (0.14)	0.42 (0.13)	0.41 (0.14)	0.43 (0.16)	0.45 (0.18)	0.48 (0.18)	0.43 (0.17)	0.35 (0.14)	0.32 (0.14)	0.27 (0.14)	0.12 (0.04)
-0.8	0.43 (0.15)	0.42 (0.13)	0.42 (0.14)	0.45 (0.18)	0.45 (0.19)	0.49 (0.18)	0.44 (0.17)	0.34 (0.15)	0.32 (0.13)	0.27 (0.13)	0.12 (0.04)
-1.0	0.44 (0.16)	0.41 (0.13)	0.42 (0.14)	0.45 (0.17)	0.47 (0.19)	0.49 (0.19)	0.45 (0.18)	0.38 (0.14)	0.34 (0.12)	0.26 (0.11)	0.12 (0.05)

Table 3: Regret via regularization. Mean regret (standard deviation) across runs, as a function of rationality (columns). Lower regret indicates better performance.

Rationality											
0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.14 (0.06)	0.11 (0.05)	0.10 (0.05)	0.08 (0.04)	0.16 (0.13)	0.33 (0.21)	0.52 (0.19)	0.56 (0.15)	0.58 (0.13)	0.56 (0.13)	0.54 (0.14)

B.3 Regarding topology

Our discussion of the experimental results shown in figure 1 primarily focuses on vertical trends: for a fixed r, how does regret vary across different settings of v or γ? We note here that when considering horizontal trends, regret does not necessarily correlate with historical rationality. Of particular interest is injecting from a misaligned distribution when $r \approx 0.5$ - since paths with cost close to the mean behave the same regardless of γ, the historical observations effectively provide no additional benefit, whereas even a policy with $r = 0$ could at least learn some edges to avoid. In general, however, this lack of correlation is likely due to the fact that the relation between historical rationality and historical coverage is dependent on the underlying topology of the shortest path problem. In our setup, figure



(a) $r = 0$ (b) $r = 0.3$

(c) $r = 0.7$ (d) $r = 1$

Figure 3: Historical exploration of policies according to different rationalities. Arc thickness corresponds to empirical frequency of the arc in the observable data. The green node indicates the source, and the red node indicates the sink.

3 shows how historical exploration is affected by rationality. Even though a historical policy with $r = 0.3$ may have had high historical regret, it may have explored enough of the action space to make good future decisions. More rigorously exploring the interplay between graph topology and policy exploration is an interesting line of future work.

B.4 Real data experiments

We additionally test our methods on a dataset from Uber Movement (`https://movement.uber.com`).[3] The dataset focuses on census tracts in downtown Los Angeles, and has 197-dimensional contexts. The road network has 93 arcs, and a total of 5902 feasible paths with span 37. The dataset consists of 3640 data points across the years 2018-2019; in our experiments we take 900 of these points as our training sample.

Historical datasets $\mathcal{D}_r$ are generated from rationalities r in the same fashion as the synthetic experiments. Constant value pseudo-data injection also is implemented identically to the synthetic case. When injecting data from prior distributions, it is now unreasonable to inject data for every unexplored path, due to the combinatorial blowup mentioned in remark 4.1. Thus, we instead generate one observation for each context in our historical data, where the path is sampled from the set of unobserved paths (while ensuring that all arcs are eventually explored). We note that due to the increased difficulty of learning policies on the real-world datasets compared to the synthetic datasets,

[3]The website is no longer actively maintained as of 11/2025. Data was acquired prior to shutdown.

and since each experiment learns 11^2 policies, we only report results for 5 independent replications: reported standard deviations are thus more variable than in the synthetic case.

B.5 Full Uber Movement results

Results for the real-world experiments are given in tables 4-6. Overall, we see similar trends to those observed in the synthetic experiments. The main difference is that regularization baseline is no longer the clear front-runner in low rationality regimes, as injection from prior distributions remains competitive across several choices of γ.

Table 4: Regret via injected constants (Uber Movement). Mean regret (standard deviation) across runs, as a function of rationality (columns) and value ratio v (rows). Lower regret indicates better performance.

	Rationality										
v	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1.0	1.45 (0.07)	1.46 (0.04)	1.21 (0.24)	1.15 (0.08)	0.89 (0.15)	0.72 (0.10)	0.69 (0.19)	0.49 (0.12)	0.29 (0.11)	0.34 (0.25)	0.17 (0.04)
0.9	1.41 (0.07)	1.45 (0.09)	1.18 (0.23)	0.91 (0.21)	0.97 (0.19)	0.57 (0.11)	0.48 (0.13)	0.36 (0.11)	0.20 (0.03)	0.16 (0.11)	0.18 (0.03)
0.8	1.40 (0.11)	1.31 (0.07)	1.12 (0.22)	1.06 (0.18)	0.80 (0.09)	0.60 (0.11)	0.45 (0.09)	0.23 (0.08)	0.24 (0.07)	0.22 (0.07)	0.11 (0.05)
0.7	1.44 (0.06)	1.28 (0.13)	1.03 (0.15)	0.96 (0.26)	0.77 (0.18)	0.59 (0.16)	0.45 (0.07)	0.33 (0.06)	0.21 (0.07)	0.29 (0.23)	0.19 (0.04)
0.6	1.35 (0.09)	1.03 (0.27)	0.89 (0.18)	0.93 (0.20)	0.64 (0.36)	0.37 (0.17)	0.32 (0.16)	0.29 (0.05)	0.16 (0.07)	0.33 (0.26)	0.19 (0.02)
0.5	1.28 (0.11)	0.73 (0.08)	0.71 (0.10)	0.64 (0.14)	0.58 (0.14)	0.36 (0.15)	0.12 (0.14)	0.24 (0.12)	0.36 (0.15)	0.53 (0.27)	0.22 (0.04)
0.4	1.10 (0.09)	0.70 (0.11)	0.52 (0.12)	0.49 (0.16)	0.82 (0.46)	0.63 (0.18)	0.26 (0.10)	0.19 (0.11)	0.41 (0.18)	0.99 (0.13)	0.21 (0.04)
0.3	1.20 (0.09)	0.67 (0.10)	0.55 (0.12)	0.45 (0.18)	0.45 (0.09)	0.53 (0.07)	0.41 (0.08)	0.72 (0.22)	0.67 (0.26)	1.00 (0.08)	0.21 (0.02)
0.2	1.14 (0.04)	0.69 (0.09)	0.49 (0.12)	0.61 (0.18)	0.38 (0.42)	0.55 (0.37)	0.52 (0.22)	0.38 (0.37)	0.94 (0.07)	1.05 (0.13)	0.23 (0.03)
0.1	1.18 (0.08)	0.66 (0.15)	0.53 (0.21)	0.47 (0.18)	0.76 (0.28)	0.53 (0.19)	0.42 (0.08)	0.32 (0.10)	0.94 (0.12)	1.10 (0.11)	0.20 (0.04)
0.0	1.19 (0.08)	0.78 (0.07)	0.39 (0.12)	0.32 (0.07)	0.58 (0.37)	0.67 (0.08)	0.45 (0.05)	0.25 (0.13)	0.94 (0.14)	0.98 (0.10)	0.24 (0.02)

Table 5: Regret via injection from prior distribution (Uber Movement). Mean regret (standard deviation) across runs, as a function of rationality (columns) and prior quality γ (rows). Lower regret indicates better performance.

	Rationality										
γ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1.0	0.39 (0.16)	0.18 (0.06)	0.25 (0.04)	0.29 (0.06)	0.24 (0.06)	0.20 (0.10)	0.11 (0.05)	0.16 (0.05)	0.11 (0.04)	0.12 (0.02)	0.08 (0.11)
0.8	0.38 (0.13)	0.28 (0.04)	0.16 (0.03)	0.24 (0.08)	0.22 (0.03)	0.17 (0.09)	0.21 (0.04)	0.18 (0.05)	0.13 (0.02)	0.15 (0.01)	0.08 (0.10)
0.6	0.43 (0.19)	0.24 (0.04)	0.23 (0.04)	0.27 (0.04)	0.22 (0.02)	0.19 (0.06)	0.16 (0.06)	0.12 (0.04)	0.12 (0.05)	0.10 (0.04)	0.12 (0.12)
0.4	0.42 (0.16)	0.25 (0.04)	0.25 (0.02)	0.24 (0.06)	0.25 (0.05)	0.19 (0.03)	0.16 (0.06)	0.16 (0.04)	0.07 (0.03)	0.08 (0.04)	0.11 (0.10)
0.2	0.28 (0.08)	0.34 (0.05)	0.26 (0.09)	0.24 (0.01)	0.25 (0.06)	0.25 (0.06)	0.16 (0.05)	0.15 (0.07)	0.13 (0.04)	0.18 (0.12)	0.10 (0.11)
0.0	0.48 (0.26)	0.33 (0.15)	0.28 (0.09)	0.26 (0.04)	0.23 (0.09)	0.23 (0.06)	0.21 (0.02)	0.15 (0.07)	0.16 (0.03)	0.14 (0.06)	0.02 (0.02)
-0.2	0.45 (0.17)	0.32 (0.09)	0.26 (0.04)	0.23 (0.06)	0.29 (0.08)	0.23 (0.06)	0.22 (0.08)	0.22 (0.09)	0.15 (0.04)	0.24 (0.09)	0.06 (0.07)
-0.4	0.43 (0.16)	0.23 (0.06)	0.25 (0.05)	0.26 (0.02)	0.31 (0.10)	0.29 (0.08)	0.16 (0.03)	0.21 (0.04)	0.27 (0.02)	0.24 (0.08)	0.12 (0.12)
-0.6	0.38 (0.17)	0.28 (0.07)	0.27 (0.04)	0.26 (0.04)	0.37 (0.06)	0.26 (0.08)	0.19 (0.05)	0.30 (0.13)	0.33 (0.03)	0.26 (0.06)	0.15 (0.12)
-0.8	0.49 (0.20)	0.29 (0.10)	0.22 (0.04)	0.28 (0.04)	0.24 (0.12)	0.27 (0.11)	0.26 (0.10)	0.27 (0.10)	0.39 (0.08)	0.28 (0.05)	0.06 (0.03)
-1.0	0.23 (0.04)	0.22 (0.08)	0.29 (0.10)	0.29 (0.07)	0.31 (0.12)	0.29 (0.14)	0.33 (0.20)	0.36 (0.15)	0.52 (0.12)	0.41 (0.11)	0.13 (0.11)

Table 6: Regret via regularization (Uber Movement). Mean regret (standard deviation) across runs, as a function of rationality (columns). Lower regret indicates better performance.

Rationality											
0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.27 (0.07)	0.22 (0.03)	0.26 (0.06)	0.25 (0.04)	0.25(0.04)	0.31 (0.08)	0.38 (0.27)	0.45 (0.24)	0.47 (0.29)	0.86 (0.28)	0.67 (0.12)

C Additional Project Requirements

C.1 Code Access

`https://github.com/ryanedmonds2000/End2End`

C.2 Author Contribution Statement

R.E proposed the original research direction of inducing coverage through data injection. Both authors jointly developed the proposed prior-based psuedo-data injection methods. J.S wrote section 1-3,5 and appendix A, and formalized/proved the analytical results contained in the paper. R.E wrote section 4 and appendix B and wrote the code for the experiments contained in the paper.

C.3 Presentation Feedback Incorporation

Vishal Feedback: To address the first comment on a broader lit review, we incorporated a greater range of contextual bandit sources and previous ideas from the warm-start literature. Additionally we explained where we are positioned with respect to the rest of the warm-start literature and how our approach differs. Second we clarified the difference between the finite estimation error and population level bias induced by pseudo-data both in the theorems' text and the surrounding prose. We emphasize that the implications associated with the population bias term is the main focus of our work. Third we clearly describe the the "knobs" available to us at the start of the experimental section, both to inform the experiments run and frame how the analysis can be operationalized.

Reviewer 1: To address the first comment of clarifying rationality, we simplified the definition in this text and added an explanation of what rationality means along with its implications. Second the reviewer requested that the experiments be discussed in greater detail. To that end we clearly explained the experimental setup in the text, added an appendix with additional experimental details, and added visualization to illustrate how we generate the synthetic data.

Reviewer 2: To address the first comment of clarifying the relationship between regularization and data injection we formalize the correspondence as proposition 2.1. We also describe the correspondence using the intuition of priors both in prose and formally with proposition 2.2. Second review 2 (along with reviewer 1) requested an experiment with real world data, which we incorporate by testing out method on a real-world stochastic shortest path dataset provided by "Uber Movement". As per the reviewers recommendation we clarify that the proposed methods are limited to and designed for settings where priors are available to the practitioner.

C.4 Code Review Feedback Incorporation

The feedback we received primarily focused on organizational structure of the codebase, rather than mechanics. Thus, no changes needed to be made to the code for the purposes of rerunning any experiments. The main point of feedback was that the various experimental files reuse a fair bit of code which could be either factored out or exposed to the CLI as a single script with more input parameters. These are changes that we plan to incorporate should the code become public in the future.