# Learning Bellman Complete Representations for Offline Policy Evaluation

**Jonathan D. Chang** [*1] **Kaiwen Wang** [*1] **Nathan Kallus** [2] **Wen Sun** [1]

## Abstract

We study representation learning for Offline Reinforcement Learning (RL), focusing on the important task of Offline Policy Evaluation (OPE). Recent work shows that, in contrast to supervised learning, realizability of the Q-function is not enough for learning it. Two sufficient conditions for sample-efficient OPE are Bellman completeness and coverage. Prior work often assumes that representations satisfying these conditions are given, with results being mostly theoretical in nature. In this work, we propose BCRL, which directly learns from data an approximately linear Bellman complete representation with good coverage. With this learned representation, we perform OPE using Least Square Policy Evaluation (LSPE) with linear functions in our learned representation. We present an end-to-end theoretical analysis, showing that our two-stage algorithm enjoys polynomial sample complexity provided *some* representation in the rich class considered is linear Bellman complete. Empirically, we extensively evaluate our algorithm on challenging, image-based continuous control tasks from the Deepmind Control Suite. We show our representation enables better OPE compared to previous representation learning methods developed for off-policy RL (e.g., CURL, SPR). BCRL achieves competitive OPE error with the state-of-the-art method Fitted Q-Evaluation (FQE), and beats FQE when evaluating beyond the initial state distribution. Our ablations show that both linear Bellman complete and coverage components of our method are crucial.

[*]Equal contribution [1]Computer Science, Cornell University, Ithaca, NY, USA [2]Operations Research and Information Engineering, Cornell Tech, New York, NY, USA. Correspondence to: Jonathan D. Chang <https://jdchang1.github.io>, Kaiwen Wang <https://kaiwenw.github.io>.

*Proceedings of the $39^{th}$ International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022. 

## 1. Introduction

Deep Reinforcement Learning (RL) has developed agents that solve complex sequential decision making tasks, achieving new state-of-the-art results and surpassing expert human performance. Despite these impressive results, these algorithms often require a prohibitively large number of online interactions to scale to higher dimensional inputs.

To address these sample complexity demands, a recent line of work (van den Oord et al., 2018; Anand et al., 2019; Mazoure et al., 2020; Yang & Nachum, 2021) has incorporated advances in unsupervised representation learning from the supervised learning literature into developing RL agents. For example, CURL (Laskin et al., 2020) and SPR (Schwarzer et al., 2020) utilize contrastive representation objectives as auxiliary losses within an existing RL framework. These efforts are motivated by the tremendous success such self-supervised techniques have offered in computer vision, natural language processing, speech processing, and beyond. While these formulations have shown sample complexity improvements empirically, it remains an open question whether these approaches successfully address the unique challenges from RL that usually do not appear in supervised learning, such as exploration and exploitation, credit assignments, long horizon prediction, and distribution shift. In particular, recent work (Wang et al., 2021; Amortila et al., 2020; Foster et al., 2021) has shown that realizability of the learning target in RL (namely, the Q-function) is insufficient to avoid exponential dependence on problem parameters.

In this paper, we study representation learning for an important subtask of off-policy RL: offline policy evaluation (OPE). OPE is a critical component for any off-policy policy optimization approach (e.g., off-policy actor critic such as SAC, Haarnoja et al., 2018, and DDPG, Lillicrap et al., 2016). Moreover, OPE allows us to focus on issues arising from distribution shift and long horizon prediction. Specifically, we propose a new approach that leverages rich function approximation (e.g., deep neural networks) to learn a representation that is both Bellman complete and exploratory. A linear Bellman complete representation means that linear functions in the representation have zero inherent Bellman error (Antos et al., 2008), i.e., applying the Bellman operator on a linear function results in a new linear function. An exploratory representation means that

the resulting feature covariance matrix formed by the offline dataset is well-conditioned. These two representational properties ensure that, under linear function approximation (i.e., the linear evaluation protocol, Grill et al., 2020), classic least squares policy evaluation (LSPE) (Nedić & Bertsekas, 2003; Duan et al., 2020; Wang et al., 2020) can achieve accurate policy evaluation. We provide an end-to-end analysis showing that our representation learning approach together with LSPE ensures near-optimal policy evaluation with polynomial sample complexity.

Empirically, we extensively evaluate our method on image-based continuous control tasks from the Deepmind Control Suite. First, we compare against two representation learning approaches developed for off-policy RL: CURL (Laskin et al., 2020) and SPR (Schwarzer et al., 2020). These bear many similarities to contrastive learning techiniques for unsupervised representation learning: SimCLR (Chen et al., 2020) and Bootstrap your own latent (BYOL, Grill et al., 2020), respectively. Under the linear evaluation protocol (i.e., LSPE with a linear function on top of the learned representation), our approach consistently outperforms these baselines. We observe that representations learned by CURL and SPR sometimes even exhibit *instability when evaluated using LSPE* (prediction error blows up when more iterations of LSPE is applied), while our approach is more stable. This comparison demonstrates that representation learning in offline RL is more subtle and using representation techniques developed from supervised learning settings may not result in the best performance for offline RL. Our ablations show that both linear Bellman completeness and coverage are crucial, as our method also blows up if one ingredient is missing. Finally, `BCRL` achieves state-of-the-art OPE error when compared with other OPE methods, and improves the state-of-the-art when evaluating beyond the initial state distribution.[1]

### 1.1. Related Work

**Representation Learning in Offline RL:**

**From the theoretical side**, Hao et al. (2021) considers offline RL in sparse linear MDPs (Jin et al., 2020). Learning with sparsity can be understood as feature selection. Their work has a much stronger coverage condition than ours: namely, given a representation class , they assume that any feature $\phi$ 2 has global coverage under the offline data distribution, i.e., $\mathsf{E}_{s,a \ \nu}\phi(s,a)\phi(s,a)^{>}$ is well conditioned where $\nu$ is offline data distribution. In our work, we only assume that *there exists one* $\phi^{\star}$ that has global coverage, thus strictly generalizing their coverage condition. Uehara & Sun (2021) propose a general model-based offline RL approach that can perform representation learning for linear MDPs in the offline setting without global coverage. However, their algorithm is a version space approach and is not computationally efficient. Also, our linear Bellman completeness condition *strictly generalizes* linear MDPs. (Ni et al., 2021) consider learning state-action embeddings from a known RKHS. We use general function approximation that can be more powerful than RKHS. Finally, Parr et al. (2008) identifies bellman completeness as a desirable condition for feature selection in RL when analyzing an equivalence between linear value-function approximation and linear model approximation. In our work, we investigate how to learn bellman completeness representation, and also the role of coverage in our feature selection.

**From the empirical side**, Yang & Nachum (2021) evaluated a broad range of unsupervised objectives for learning pretrained representations from offline datasets for downstream Imitation Learning (IL), online RL, and offline RL. They found that the use of pretrained representations dramatically improved the downstream performance for policy learning algorithms. In this work, we aim to identify such an unsupervised representation objective and to extend the empirical evaluation to offline image-based continuous control tasks. Nachum & Yang (2021) presents a provable contrastive representation learning objective for IL (derived from maximum likelihood loss), learning state representations from expert data to do imitation with behavior cloning. Our approach instead focuses on learning state-action representations for OPE. Finally, both Song et al. (2016) and Chung et al. (2019) present algorithms to learn representations suitable for linear value function approximation. Song et al. (2016) is most relevant to our work where they aim to learn bellman complete features. Both works, however, work in the online setting and do not consider coverage induced from the representation which we identify is important for accurate OPE.

**Representation Learning in Online RL:**

**Theoretical works** on representation learning in online RL focus on learning representations to facilitate exploration from scratch. OLIVE (Jiang et al., 2017), BLin-UCB (Du et al., 2021), FLAMBE (Agarwal et al., 2020), Moffle (Modi et al., 2021), and Rep-UCB (Uehara et al., 2022) propose approaches for representation learning in linear MDPs where they assume that there exists a feature $\phi^{\star}$ 2 that admits the linear MDP structure for the ground truth transition. Zhang et al. (2021) posits an even stronger assumption where every feature $\phi$ 2 admits the linear MDP structure for the true transition. Note that we only assume that *there exists one* $\phi^{\star}$ that admits linear Bellman completeness, which strictly generalizes linear MDPs. Hence, our representation learning setting is more general than prior theoretical works. Note that we study the offline setting while prior works operate in the online setting which has additional challenges from online exploration.

[1] Code available at `https://github.com/CausalML/bcrl`.

**On the empirical side**, there is a large body of works that adapt existing self-supervised learning techniques developed from computer vision and NLP to RL. For learning representations from image-based RL tasks, CPC (van den Oord et al., 2018), ST-DIM (Anand et al., 2019), DRIML (Mazoure et al., 2020), CURL (Laskin et al., 2020), and SPR (Schwarzer et al., 2020) learn representations by optimizing various temporal contrastive losses. In particular, Laskin et al. (2020) proposes to interleave minimizing an InfoNCE objective with similarities to MoCo (He et al., 2019) and SimCLR (Chen et al., 2020) while learning a policy with SAC (Haarnoja et al., 2018). Moreover, Schwarzer et al. (2020) proposes learning a representation similar to BYOL (Grill et al., 2020) alongside a Q-function for Deep Q-Learning (Mnih et al., 2013). We compare our representational objective against CURL and SPR in Section 6, and demonstrate that under linear evaluation protocol, ours outperform both CURL and SPR. Note, we refer the readers to Schwarzer et al. (2020) for comparisons between SPR and CPC, ST-DIM, and DRIML.

## 2. Preliminaries

In this paper we consider the infinite horizon discounted MDP $\langle S, A, \gamma, P, r, d_0 \rangle$. where $S, A$ are state and action spaces which could contain a large number of states and actions or could even be infinite, $\gamma \in (0,1)$ is a discount factor, $P$ is the transition kernel, $r : S \times A \to \mathbb{R}$ is the reward function, and $d_0 \in \Delta(S)$ is the initial state distribution. We assume rewards are bounded by 1, i.e. $|r(s,a)| \leq 1$. Given a policy $\pi : S \to \Delta(A)$, we denote $V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r_h | \pi, s_0 := s\right]$ as the expected discounted total reward of policy $\pi$ starting at state $s$. We denote $V^\pi_{d_0} = \mathbb{E}_{s \sim d_0} V^\pi(s)$ as the expected discounted total reward of the policy $\pi$ starting at the initial state distribution $d_0$. We also denote average state-action distribution $d^\pi_{d_0}(s,a) = (1-\gamma)\sum_{h=0}^{\infty} \gamma^h d^\pi_h(s,a)$ where $d^\pi_h(s,a)$ is the probability of $\pi$ visiting the $(s,a)$ pair at time step $h$, starting from $d_0$.

In OPE, we seek to evaluate the expected discounted total reward $V^{\pi_e}$ of a target policy $\pi_e : S \to \Delta(A)$ given data drawn from an offline state-action distribution $\nu \in \Delta(S \times A)$. The latter can, for example, be the state-action distribution under a behavior policy $\pi_b$. Namely, the offline dataset $\mathcal{D} = \{s_i, a_i, r_i, s_i'\}_{i=1}^N$ consists of $N$ i.i.d tuples generated as $(s,a) \sim \nu, r = r(s,a), s' \sim P(\cdot|s,a)$.

We define Bellman operator $\mathcal{T}^\pi$ associated with $\pi$ as follows:

$$\mathcal{T}^\pi f(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a), a' \sim \pi(s')}[f(s',a')]$$

We may drop the superscript $\pi$ when it is clear from context, in particular when $\pi = \pi_e$ is the fixed target policy.

A representation, or feature, $\phi : S \times A \to \mathbb{R}^d$ is an embedding of state-action pairs into $d$-dimensional space. We let $\Sigma(\phi) = \mathbb{E}_\nu[\phi(s,a)\phi(s,a)^\top]$, $\hat{\Sigma}(\phi) = \frac{1}{N}\sum_{i=1}^N \phi(s_i,a_i)\phi(s_i,a_i)^\top$. We consider learning a representation from a feature hypothesis class $\Phi \subseteq [S \times A \to \mathbb{R}^d]$. We assume features have bounded norm: $\|\phi(s,a)\|_2 \leq 1, \forall s, a, \forall \phi \in \Phi$. In our experiments, $\Phi$ is convolutional neural nets with $d$ outputs.

**Notation** We denote $B_W := \{x \in \mathbb{R}^d : \|x\|_2 \leq W\}$ as the Euclidean Ball in $\mathbb{R}^d$ with radius $W$. Given a distribution $\nu \in \Delta(S \times A)$ and a function $f : S \times A \to \mathbb{R}$, we denote $L_2(\nu)$ norm as $\|f\|_\nu^2 = \mathbb{E}_{s,a \sim \nu} f^2(s,a)$. Given a positive definite matrix $\Sigma$, let $\|x\|_\Sigma = \sqrt{x^\top \Sigma x}$ and let $\lambda_{\min}(\Sigma)$ denote the minimum eigenvalue. When $\nu \ll \mu$ we let $\frac{d\nu}{d\mu}$ denote the Radon-Nikodym derivative. We use $\circ$ to denote composition, so $s' \sim P \circ (s,a)$, $a' \sim \pi \circ (s')$ is short-form for $s' \sim P(s,a), a' \sim \pi(s')$. For any function $f(s,a)$ and a policy $\pi$, we denote $f(s,\pi) = \mathbb{E}_{a \sim \pi(s)}[f(s,a)]$.

## 3. Linear Bellman Completeness

Before we introduce our representation learning approach, in this section, we first consider OPE with linear function approximation with a given representation $\phi$. Lessons from supervised learning or linear bandit may suggest that OPE should be possible with polynomially many offline samples, as long as (1) ground truth target $Q^{\pi_e}$ is linear in $\phi$ (i.e. $\exists w \in \mathbb{R}^d$, such that for all $s, a$, $Q^{\pi_e}(s,a) = w^\top \phi(s,a)$), and (2) the offline data provides sufficient coverage over $\pi_e$ (i.e., $\lambda_{\min}(\Sigma(\phi)) > 0$). Unfortunately, under these two assumptions, there are lower bounds indicating that for *any* OPE algorithm, there exists an MDP where one will need at least exponentially (exponential in horizon or $d$) many offline samples to provide an accurate estimate of $V^{\pi_e}$ (Wang et al., 2021; Foster et al., 2021; Amortila et al., 2020). This lower bound indicates that one needs additional structural conditions on the representation.

The additional condition the prior work has consider is Bellman completeness (BC). Since we seek to learn a representation rather than assume theoretical conditions on a given one, we will focus on an *approximate* version of BC.

**Definition 3.1** (Approximate Linear BC)**.** A representation $\phi$ is $\varepsilon_\nu$-approximately linear Bellman complete if,

$$\max_{w_1 \in B_W} \min_{w_2 \in B_W} \|w_2^\top \phi - \mathcal{T}^{\pi_e}(w_1^\top \phi)\|_\nu \leq \varepsilon_\nu.$$

Note the dependence on $\nu$, $\pi_e$, and $W$.

Intuitively the above condition is saying that for any linear function $w_1^\top \phi(s,a)$, $\mathcal{T}^\pi(w_1^\top \phi(s,a))$ itself can be approximated by another linear function under the distribution $\nu$.

*Remark* 3.2. Low-rank MDPs (Jiang et al., 2017; Agarwal et al., 2020) are subsumed in the exact linear BC model, i.e., $\varepsilon_\nu = 0$ under any distribution $\nu$.

**Algorithm 1** Least Squares Policy Evaluation (LSPE)

**Input:** Target policy $\pi_e$, features $\phi$, dataset $D$
Initialize $\widehat{\theta}_0 = \mathbf{0} \in B_W$.
**for** $k = 1, 2, ..., K$ **do**
  Set $\widehat{f}_{k-1}(s,a) = \widehat{\theta}_{k-1}^\top \phi(s,a)$,
    $\widehat{V}_{k-1}(s) = \mathbb{E}_{a \sim \pi_e(s)}[\widehat{f}_{k-1}(s,a)]$
  Perform linear regression:

$$\widehat{\theta}_k \in \arg\min_{\theta \in B_W} \frac{1}{N} \sum_{i=1}^{N} \left(\theta^\top \phi(s_i, a_i) - r_i - \gamma \widehat{V}_{k-1}(s_i')\right)^2$$

**end for**
Return $\widehat{f}_K$.

Note that the Bellman completeness condition is more subtle than the common realizability condition: for any fixed $\phi$, increasing its expressiveness (e.g., add more features) does not imply a monotonic decrease in $\varepsilon_\nu$. Thus common tricks such as lifting features to higher order polynomial kernel space or more general reproducing kernel Hilbert space does not imply the linear Bellman complete condition, nor does it improve the coverage condition. We next show approximate Linear BC together with coverage imply sample-efficient OPE via Least Square Policy Evaluation (LSPE) (Algorithm 1). We present our result using the relative condition number, defined for any initial state distribution $p_0$ as

$$\kappa(p_0) := \sup_{x \in \mathbb{R}^d} \frac{x^\top \mathbb{E}_{s,a \sim d_{p_0}^{\pi_e}} \phi(s,a)\phi(s,a)^\top x}{x^\top \Sigma(\phi) x}. \quad (1)$$

**Theorem 3.3** (Sample Complexity of LSPE)**.** *Assume feature $\phi$ satisfies approximate Linear BC with parameter $\varepsilon_\nu$. For any $\delta \in (0, 0.1)$, w.p. at least $1 - \delta$, we have for any initial state distribution $p_0$*

$$\left|V_{p_0}^{\pi_e} - \mathbb{E}_{s \sim p_0}\left[\widehat{f}_K(s, \pi_e)\right]\right| \le \frac{\gamma^{K/2}}{1-\gamma} + \frac{4\sqrt{\left\|\frac{\mathrm{d}d_{p_0}^{\pi_e}}{\mathrm{d}\nu}\right\|_1}}{(1-\gamma)^2}\varepsilon_\nu + \frac{480\sqrt{\kappa(p_0)}(1+W)d\log(N)\sqrt{\log(10/\delta)}}{(1-\gamma)^2\sqrt{N}},$$

*where $\widehat{f}_K$ is the output of Algorithm 1.*

Please see Appendix C.1 for proof. The above result holds simultaneously over all initial state distributions $p_0$ covered by the data distribution $\nu$. If $\nu$ has full coverage, i.e. if $\Sigma(\phi) \succeq \beta I$, as is commonly assumed in the literature, one can show that $\kappa(p_0) \le \beta^{-1}$ for any initial state distribution $p_0$. Also note that the concentrability coefficient $\left\|\frac{\mathrm{d}d_{p_0}^{\pi_e}}{\mathrm{d}\nu}\right\|_1$ shows up as $T^{\pi_e}\widehat{f}_{k-1}$ can be *nonlinear*. In the exact Linear BC case, where $\varepsilon_\nu = 0$ (i.e., there is a linear function that perfectly approximates $T^{\pi_e}\widehat{f}_{k-1}$ under $\nu$), the term involving the concentrability coefficient will be 0 and we can even avoid its finiteness.

# 4. Representation Learning

The previous section indicates sufficient conditions on the representation for efficient and accurate OPE with linear function approximation. However, a representation that is Bellman complete and also provides coverage is not available a priori, especially for high-dimensional settings (e.g., image-based control tasks as we consider in our experiments). Existing theoretically work often assume such representation is given. We propose to learn a representation $\phi$ that is approximately Bellman complete and also provides good coverage, via rich function approximation (e.g., a deep neural network). More formally, we want to learn a representation $\phi$ such that (1) it ensures approximate Bellman complete, i.e., $\varepsilon_\nu$ is small, and (2) has a good coverage, i.e., $\lambda_{\min}(\Sigma(\phi))$ is as large as possible, which are the two key properties to guarantee accurate and efficient OPE indicated by Theorem 3.3. To formulate the representation learning question, our key assumption is that our representation hypothesis class $\Phi$ is rich enough such that it contains at least one representation $\phi^\star$ that is linear Bellman complete (i.e., $\varepsilon_\nu = 0$) and has a good coverage.

**Assumption 4.1** (Representational power of $\Phi$)**.** There exists $\phi^\star \in \Phi$, such that (1) $\phi^\star$ achieves exact Linear BC (definition 3.1 with $\varepsilon_\nu = 0$), and (2) $\phi^\star$ induces coverage, i.e., $\lambda_{\min}(\Sigma(\phi^\star)) \ge \beta > 0$.

Our goal is to learn such a representation from the hypothesis class $\Phi$. Note that unlike prior RL representation learning works (Hao et al., 2021; Zhang et al., 2021), here we only assume that there exists a $\phi^\star$ has linear BC and induces good coverage, other candidate $\phi \in \Phi$ could have terrible coverage and does not necessarily have linear BC.

Before proceeding to the learning algorithm, we first present an *equivalent condition* for linear BC, which does not rely on the complicated min-max expression of Definition 3.1.

**Proposition 4.2.** *Consider a feature $\phi$ with full rank covariance $\Sigma(\phi)$. Given any $W > 0$, the feature $\phi$ being linear BC (under $B_W$) implies that there exist $(\rho, M) \in B_W \times \mathbb{R}^{d \times d}$ with $\|M\|_2 \le \sqrt{1 - \frac{k\rho k^2}{W^2}}$, and*

$$\mathbb{E}_\nu \left\| \begin{bmatrix} M \\ \rho^\top \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma \mathbb{E}_{s' \sim P(s,a)} \phi(s', \pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 = 0. \quad (2)$$

*On the other hand, if there exists $(\rho, M) \in B_W \times \mathbb{R}^{d \times d}$ with $\|M\|_2 < 1$ such that the above equality holds, then $\phi$ must satisfy exact linear BC with $W \le \frac{\|\rho\|_2}{1 - \|M\|_2}$.*

Please see Appendix D for proof. The above shows that linear BC is equivalent to a simple linear relationship between the feature and the expected next step's feature and reward. This motivates our feature learning algorithm: if we are capable of learning a representation $\phi$ such that the transition

from the current feature $\phi(s,a)$ to the expected next feature $\mathsf{E}_{s' \sim P(s,a)}[\phi(s', \pi_e)]$ and reward $r(s,a)$ is linear, then we've found a feature $\phi$ that is linear BC.

### 4.1. Algorithm

To learn the representation that achieves linear BC, we use Proposition 4.2 to design a bilevel optimization program. We start with deterministic transition as a warm up.

**Deterministic transition** Due to determinism in the transition, we do not have an expectation with respect to $s'$ anymore. So, we design the bilevel optimization as follows:

$$\min_{\phi \in \Phi} \Big[ \min_{(\rho,M) \in \Theta} \mathsf{E}_{\mathcal{D}} \left\| \begin{bmatrix} M \\ \rho^\intercal \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma \phi(s', \pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 \quad (3)$$
$$= \left\{ (\rho, M) \in B_{k\rho^\star k} \times \mathbb{R}^{d \times d} : \|M\|_2 \le \|M^\star\|_2 \right\},$$

where $\rho^\star, M^\star$ are the optimal $\rho, M$ for the linear BC $\phi^\star$ (in Assumption 4.1). Namely, we aim to search for a representation $\phi \in \Phi$, such that the relationship between $\phi(s,a)$ and the combination of the next time step's feature $\phi(s', \pi_e)$ and the reward, is linear. The spectral norm constraint in $\Theta$ is justified by Proposition 4.2. Solving the above bilevel moment condition finds a representation that achieves approximate linear BC.

However, there is no guarantee that such representation can provide a good coverage over $\pi_e$'s traces. We introduce regularizations for $\phi$ using the ideas from optimal designs, particularly the $E$-optimal design. Define the minimum eigenvalue regularization

$$R_E(\phi) := \lambda_{\min}\left(\mathsf{E}_D\left[\phi(s,a)\phi(s,a)^\intercal\right]\right),$$

as the smallest eigenvalue of the empirical feature covariance matrix under the representation $\phi$. Maximizing this quantity ensures that our feature provides good coverage, i.e. $\lambda_{min}(\Gamma(\phi))$ is as large as possible.

Thus, to learn a representation that is approximately linear Bellman complete and also provides sufficient coverage, we formulate the following constrained bilevel optimization:

$$\min_{\phi \in \Phi} \Big[ \min_{(\rho,M) \in \Theta} \mathsf{E}_{\mathcal{D}} \left\| \begin{bmatrix} M \\ \rho^\intercal \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma \phi(s', \pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2$$
$$s.t., R_E(\phi) \ge \beta/2.$$

To extend this to stochastic transitions, there is an additional double sampling issue, which we discuss and address now.

**Stochastic transition** Ideally, we would solve the following bilevel optimization problem,

$$\min_{\phi \in \Phi} \Big[ \min_{(\rho,M) \in \Theta} \mathsf{E}_{\mathcal{D}} \left\| \begin{bmatrix} M \\ \rho^\intercal \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma \mathsf{E}_{s' \sim P(s,a)}[\phi(s', \pi_e)] \\ r(s,a) \end{bmatrix} \right\|_2^2 \Big]. \quad (4)$$

However, we cannot directly optimize the above objective since we do not have access to $P(s,a)$ to compute the expected next step feature $\mathsf{E}_{s' \sim P(s,a)}\phi(s', \pi_e)$. Also note that the expectation $\mathsf{E}_{s'}$ is inside the square which means that we cannot even get an unbiased estimate of the gradient of $(\phi, M)$ with one sample $s' \sim P(s,a)$. This phenomenon is related to the *double sampling* issue on offline policy evaluation literature which forbids one to directly optimize Bellman residuals. Algorithms such as TD are designed to overcome the double sampling issue. Here, we use a different technique to tackle this issue (Chen & Jiang, 2019). We introduce a function class $\mathcal{G} \subseteq \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ which is rich enough to contain the expected next feature.

**Assumption 4.3.** For any $\phi \in \Phi$, we have that the mapping $(s,a) \to \gamma \mathsf{E}_{s' \sim P(s,a)}[\phi(s', \pi_e)]$ is in $\mathcal{G}$.

We form the optimization problem as follows:

$$\min_{\phi \in \Phi} \Big[ \min_{(\rho,M) \in \Theta} \mathsf{E}_{\mathcal{D}} \left\| \begin{bmatrix} M \\ \rho^\intercal \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma \phi(s', \pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2$$
$$\min_{g \in \mathcal{G}} \mathsf{E}_{\mathcal{D}} \|g(s,a) - \gamma \phi(s', \pi_e)\|_2^2 \Big]. \quad (5)$$

Note that under Assumption 4.3, the min over $\mathcal{G}$ will approximate $\gamma^2 \mathsf{E}_{\mathcal{D}} \|\mathsf{E}_{s' \sim P(s,a)}\phi(s', \pi_e) - \phi(s', \pi_e)\|_2^2$, i.e., the average variance induced by the stochastic transition. Thus, for a fixed $\phi$ and $M$, we can see that the following

$$\mathsf{E}_{\mathcal{D}} \|M\phi(s,a) - \gamma \phi(s', \pi_e)\|_2^2$$
$$- \gamma^2 \mathsf{E}_{\mathcal{D}} \|\mathsf{E}_{s' \sim P(s,a)}\phi(s', \pi_e) - \phi(s', \pi_e)\|_2^2$$

is indeed an unbiased estimate of:

$$\mathsf{E}_{s,a \sim \nu} \left\| M\phi(s,a) - \gamma \mathsf{E}_{s' \sim P(s,a)}\phi(s', \pi_e) \right\|_2^2$$

which matches to the ideal objective in Eq. 4. Thus solving for $\phi$ based Eq. 5 allows us to optimize Eq. 4, which allows us to learn an approximate linear Bellman complete representation. Similarly, we incorporate the $E$-optimal design here by adding a constraint that forcing the smallest eigenvalue of the empirical feature covariance matrix, i.e., $R_E(\phi)$, to be lower bounded.

Once we learn a representation that is approximately linear BC, and also induces sufficient coverage, we simply call the LSPE to estimate $V^{\pi_e}$. The whole procedure is summarized in Algorithm 2. Note in Alg 2 we put constraints to the objective function using Lagrangian multiplier.

There are other design choices that also encourage coverage. One particular choice we study empirically is motivated by the idea of D-optimal design. Here we aim to find a representation that maximizes the following log-determinant

$$R_D(\phi) := \ln\det\left(\mathsf{E}_D\left[\phi(s,a)\phi(s,a)^\intercal\right]\right).$$

**Algorithm 2** OPE with **B**ellman **C**omplete and exploratory **R**epresentation **L**earning (BCRL)

1: **Input:** Representation class , dataset $D$ of size $2N$, design regularization $R$, function class $G$, policy $\pi_e$.
2: Randomly split $D$ into two sets $D_1, D_2$ of size $N$.
3: If the system is stochastic, learn representation $\widehat{\phi}$ as,

$$\arg\min_{\phi \in} \left[ \min_{(\rho,M)\in} \mathbb{E}_{D_1} \left\| \begin{bmatrix} M \\ \rho^\top \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma\phi(s',\pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 \right.$$
$$\left. \min_{g\in G} \mathbb{E}_{D_1} \left[ g(s,a) - \gamma\phi(s',\pi_e)\right]_2^2 \right] + \lambda R(\phi)$$

4: Otherwise, for deterministic system, learn $\widehat{\phi}$ as,

$$\arg\min_{\phi \in} \left[ \min_{(\rho,M)\in} \mathbb{E}_{D_1} \left\| \begin{bmatrix} M \\ \rho^\top \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma\phi(s',\pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 \right] + \lambda R(\phi)$$

5: Return $\widehat{V} :=$ LSPE$(\pi_e, \widehat{\phi}, D_2)$.

When $D$ is large, then the regularization $R_D(\phi)$ approximates $\sum_i \ln(\sigma_i(\mathbb{E}_\nu[\phi(s,a)\phi(s,a)^\top]))$. Maximizing $R_D(\phi)$ then intuitively maximizes coverage over all directions. D-optimal design is widely used for bandits (Lattimore & Szepesvári, 2020) and RL (Wang et al., 2021; Agarwal et al., 2019) to design exploration distributions with global coverage. Note that, in contrast to these contexts where the feature is fixed and the distribution is optimized, we optimize the feature, given the data distribution $\nu$.

### 4.2. Sample Complexity Analysis

We now prove a finite sample utility guarantee for the empirical constrained bilevel optimization problem,

$$\widehat{\phi} \in \arg\min_{\phi \in} \left[ \min_{(\rho,M)\in} \mathbb{E}_D \left\| \begin{bmatrix} M \\ \rho^\top \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma\phi(s',\pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 \right.$$
$$\left. \min_{g\in G} \mathbb{E}_D \left[ g(s,a) - \gamma\phi(s',\pi_e)\right]_2^2 \right]. \quad (6)$$
$$s.t., R_E(\phi) \quad \beta/2.$$

For simplicity, we state our results for discrete function class and $G$. Note that our sample complexity only scales with respect $\ln(|\,|)$ and $\ln(|G|)$, which are the standard complexity measures for discrete function classes. We extend our analysis to infinite function classes under metric entropy assumptions (Wainwright, 2019; van der Vaart & Wellner, 1996) in the Appendix; see Assumption E.5.

The following theorem shows that Algorithm 2 learns a representation $\widehat{\phi}$ that is $O(N^{-1/2})$ approximately Linear BC and has coverage.

**Theorem 4.4.** *Assume Assumption 4.1 (and Assumption 4.3 if the system is stochastic). Let* $C_2 := \frac{96\ln^{1/2}(|\,|)+4\sqrt{d}+4\ln^{1/2}(1/\delta)}{\beta/4}$. *If* $N \quad C_2^2$, *then for any* $\delta \in (0,1)$, *w.p. at least* $1 \quad \delta$, *we have*

1. $\widehat{\phi}$ *satisfies* $\widehat{\varepsilon}$*-approximate Linear BC with*

$$\widehat{\varepsilon} \quad \frac{13d(1+W)^2\ln^{1/2}(4W|\,|N/\delta)}{\sqrt{N}} + \frac{7\gamma(1+W)\ln^{1/2}(2|G||/\delta)}{\sqrt{N}},$$

2. $\lambda_{min}(\,(\widehat{\phi})) \quad \beta/4$.

*If transitions are deterministic, treat* $\ln(|G|) = 0$.

*Proof Sketch.* First, we use Weyl's Perturbation Theorem and chaining to show that the eigenvalues of $(\phi)$ are close to $\widehat{}(\phi)$, uniformly over $\phi$. This implies that (a) $\lambda_{min}(\widehat{}(\phi^\star)) \quad \beta/2$, and hence is feasible in the empirical bilevel optimization Equation (6), and (b) $\lambda_{min}(\,(\widehat{\phi}))$ $\beta/4$. Since $\phi^\star$ is feasible, we apply uniform concentration arguments to argue that $\widehat{\phi}$ has low population loss (Equation (5)), which implies approximate Linear BC. □

The error term in $\widehat{\varepsilon}$ is comprised of the statistical errors of fitting and of fitting $G$ for the double sampling correction. In the contextual bandit setting, i.e. $\gamma = 0$, there is no transition, so the second term becomes 0. Chaining together with Theorem 3.3 gives the following end-to-end $\widetilde{O}(N^{-1/2})$ evaluation error guarantee for LSPE with the learned features:

**Theorem 4.5.** *Under Assumption 4.1 (and Assumption 4.3 if* $P(s,a)$ *is stochastic). Let* $C_2, \widehat{\varepsilon}$ *be as defined in Theorem 4.4. If* $N \quad C_2^2$, *then we have for any* $\delta \in (0, 1/2)$, *w.p. at least* $1 \quad 2\delta$, *for all distributions* $p_0$,

$$\left| V_{p_0}^{\pi_e} \quad \mathbb{E}_{s \sim p_0}\left[\widehat{f}_K(s,\pi_e)\right] \right| \quad \frac{\gamma^{K/2}}{1-\gamma} + \frac{4\sqrt{\left\| \frac{d d_{p_0}^{\pi_e}}{d\nu} \right\|_1}}{(1-\gamma)^2}\widehat{\varepsilon}$$
$$+ \frac{960\beta^{-1/2}(1+W)d\ln(N)\sqrt{\ln(10/\delta)}}{(1-\gamma)^2\sqrt{N}}.$$

**Comparison to FQE** What if one ignores the representation learning and just runs the Fitted Q-Evaluation (FQE) which directly performs least square fitting with the nonlinear function class $F := \{w^\top \phi(s,a) : \phi \in , \|w\|_2 \quad W\}$? As $N \to \infty$, FQE will suffer the following worst-case Bellman error (also called inherent Bellman error):

$$\varepsilon_{\text{ibe}} := \max_{f\in F} \min_{g\in F} \|g - \mathcal{T}^{\pi_e} f\|_\nu^2.$$

**Algorithm 3** Practical Instantiation of BCRL

**Input:** Offline dataset $D = \{s, a, s'\}$, target policy $\pi_e$
Initialize parameters for $\phi$, $M$, $\rho$, and $\bar{\phi}$
**for** t = 0, 1, . . . , T **do**
  $M_{t+1} \leftarrow M_t - \eta \nabla_M J(\phi_t, M_t, \rho_t, \bar{\phi}_t)$
  $\rho_{t+1} \leftarrow \rho_t - \eta \nabla_\rho J(\phi_t, M_t, \rho_t, \bar{\phi}_t)$
  $\phi_{t+1} \leftarrow \phi_t - \eta \nabla_\phi J(\phi_t, M_{t+1}, \rho_{t+1}, \bar{\phi}_t)$
  $\bar{\phi}_{t+1} \leftarrow \tau\phi_{t+1} + (1 - \tau)\bar{\phi}_t$
**end for**
Linear evaluation: $\widehat{V} = \text{LSPE}(\pi_e, \phi_T, D)$.

Note that our assumption that there exists a linear BC representation $\phi^\star$ does not imply that the worst-case Bellman error $\varepsilon_{ibe}$ is small. In contrast, when $N \to \infty$, our approach will accurately estimate $V^{\pi_e}_{p_0}$.

## 5. A Practical Implementation

In this section we instantiate a practical implementation to learn our representation using deep neural networks for our representation function class $\Phi$. Based on Equation (3), we first formalize our bilevel optimization objective:

$$J(\phi, M, \rho, \bar{\phi}) = \mathbb{E}_D \left\| \begin{bmatrix} M \\ \rho^\top \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma\bar{\phi}(s', \pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 - \lambda \log\det \mathbb{E}_D \left[\phi(s,a)\phi(s,a)^\top\right].$$

In our implementation, we replace the hard constraint presented in Section 4.1 with a Lagrange multiplier, i.e. we use the optimal design constraint as a regularization term when learning $\phi$. Specifically, we maximize the $\log\det$ of the covariance matrix induced by the feature, which maximizes all eigenvalues since $\log\det$ is the sum of the log eigenvalues. Our experiment results in Section 6.4 demonstrate that it indeed improves the condition number of $\Phi(\phi)$. In some of our experiments, we use a target network in our implementation. Namely, the updates make use of a target network $\bar{\phi}$ where the weights can be an exponential moving average of the representation network's weights. This idea of using a slow moving target network has been shown to stabilize training in both the RL (Mnih et al., 2013) and the representation learning literature (Grill et al., 2020).

As summarized in Algorithm 3, given our offline behavior dataset $D$ and target policy $\pi_e$, we iteratively update $M$ and $\phi$ and then use the resulting learned representation to perform OPE. Please see Appendix F for implementation and hyperparameter details. As we will show in our experiments, our update procedures for $\phi$ significantly minimizes the Bellman Completeness loss and also improve the condition number of $\Phi(\phi)$, which are the two key quantities to ensure good performance of LSPE with linear regression as shown in Theorem 3.3.

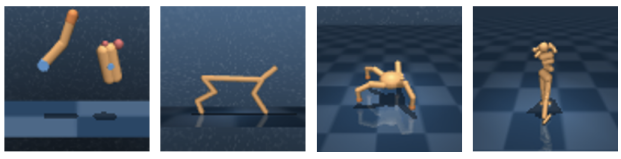

*Figure 1.* Representative frames from DeepMind Control Suite tasks. From left to right, `Finger Turn Hard`, `Cheetah Run`, `Quadruped Walk`, and `Humanoid Stand`.

## 6. Experiments

Our goal is to answer the following questions. (1) How do our representations perform on downstream LSPE compared to other popular unsupervised representation learning techniques? (2) How important are both the linear bellman completeness and optimal design components for learning representations? (3) How competitive is BCRL with other OPE methods, especially for evaluating beyond the initial state distribution?

Following the standards in evaluating representations in supervised learning, we used a **linear evaluation protocol**, i.e., linear regression in LSPE on top of a given representation. This allows us to focus on evaluating the quality of the representation. We compared our method to prior techniques on a range of challenging, image-based continuous control tasks from the DeepMind Control Suite benchmark (Tassa et al., 2018): `Finger Turn Hard`, `Cheetah Run`, `Quadruped Walk`, and `Humanoid Stand`. Frames from the tasks are shown in Figure 1. To investigate our learned representation, we benchmark our representation against two state-of-the-art self-supervised representation learning objectives adopted for RL: (1) CURL uses the InfoNCE objective to contrastively learn state-representations; and (2) SPR adopts the BYOL contrastive learning framework to learn representations with latent dynamics. Note, we modified CURL for OPE by optimizing the contrastive loss between state-action pairs rather than just states. For SPR, we did not include the Q prediction head and used their state representation plus latent dynamics as the state-action representation for downstream linear evaluation. We used the same architecture for the respective representations across all evaluated algorithms. To compare against other OPE methods, we additionally compared against Fitted Q-Evaluation (Munos & Szepesvári, 2008; Kostrikov & Nachum, 2020) (FQE), weighted doubly robust policy evaluation (Jiang & Li, 2016; Thomas & Brunskill, 2016) (DR), Dreamer-v2 (Hafner et al., 2021) model based evaluation (MB), and DICE (Yang et al., 2020). We modified implementations from the benchmark library released by Fu et al. (2021) for FQE and DR and used the authors' released codebases for Dreamer-v2 and BestDICE.[2]

[2] https://github.com/google-research/dice_rl.

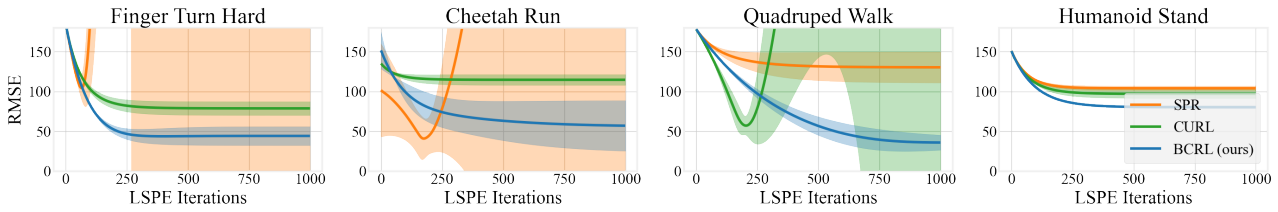


*Figure 2.* OPE curves across five seeds using representations trained with BCRL, SPR, and CURL on the offline datasets (Table 1).

Our target policies were trained using the authors' implementation of DRQ-v2 (Yarats et al., 2021), a state-of-the-art, off-policy actor critic algorithm for vision-based continuous control. With high-quality target policies, we collected 200 rollouts and did a linear evaluation protocol to predict the discounted return. For our offline behavior datasets, we collected 10K samples from a trained policy with mean performance roughly a quarter of that of the target policy (Table 1). All results are aggregated over five random seeds. See Appendix F for details on hyperparameters, environments, and dataset composition.

### 6.1. OPE via LSPE with Learned Representations

Figure 2 compares the OPE performance of BCRL against SPR and CURL. Representations learned by BCRL outperform those learned by SPR and CURL. On some tasks, SPR and CURL both exhibited an exponential error amplification with respect to the number of iterations of LSPE, while BCRL did not suffer from any blowup.

### 6.2. OPE Performance

Figure 3 compares the OPE performance of BCRL against multiple benchmarks from the OPE literature. BCRL is competitive with FQE and evaluates better than other benchmarks across the tasks that we tested on.

We also evaluated how well estimated values from BCRL rank policies. Following (Fu et al., 2021), we use the spearman ranking correlation metric to compute the correlation between ordinal rankings according to the OPE estimates and the ground truth values. For ranking, we evaluated three additional target policies with mean performances roughly 75%, 50%, and 10% of the target policy. Figure 3 presents the mean correlation of each evaluation algorithm across all tasks. BCRL is competitive with FQE and consistently better than other benchmarks at ranking policies.

### 6.3. Further Investigation of Different Settings

In this section, we consider two additional settings: (1) the offline dataset contains some on-policy data, which ensures that the offline data provides coverage over the evaluation policy's state action distribution; (2) we evaluate all estimators beyond the original initial state distribution $d_0$. The first setting ensures that baselines and our algorithm all satisfy the coverage condition, thus demonstrating the unique benefit of learning a Bellman complete representation. The second setting evaluates the robustness of our algorithm, i.e., the ability to estimate beyond the original $d_0$.

**Evaluation on On-Policy + Off-Policy Data:** To further investigate the importance of learning linear BC features, we experiment with learning representations from an offline dataset that also contains state-action pairs from the target policy. More specifically, we train our representations on a dataset containing 10K behavior policy and 10K target policy samples. Note, only the experiment in this paragraph uses this mixture dataset. With the addition of on-policy data from the target policy, we can focus on just the role of linear BC for OPE performance because the density ratio $\frac{\mathrm{d}d^{\pi_e}_{po}}{\mathrm{d}\nu}$ and the relative condition number (Eq. 1) is at most 2, i.e. we omit the design regularization and focused on minimizing the Bellman completeness loss. Figure 4 (Left) shows that BCRL outperforms baselines in this setting, even

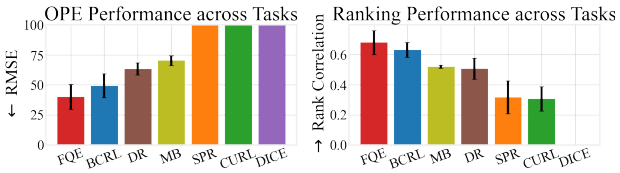


*Figure 3.* (Left) Root mean squared evaluation error across all tasks. (Right) Mean spearman ranking correlation across all tasks.

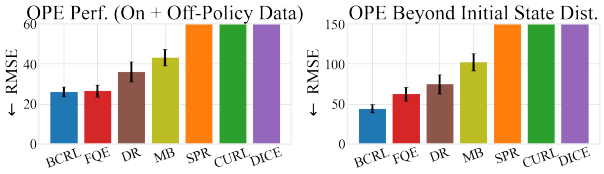


*Figure 4.* (Left) Evaluation on a mixture dataset with on-policy and off-policy data. Note the addition of target policy data bounds the relative condition number (Eq. 1). (Right) Evaluation beyond the initial state distribution (given just the offline data).

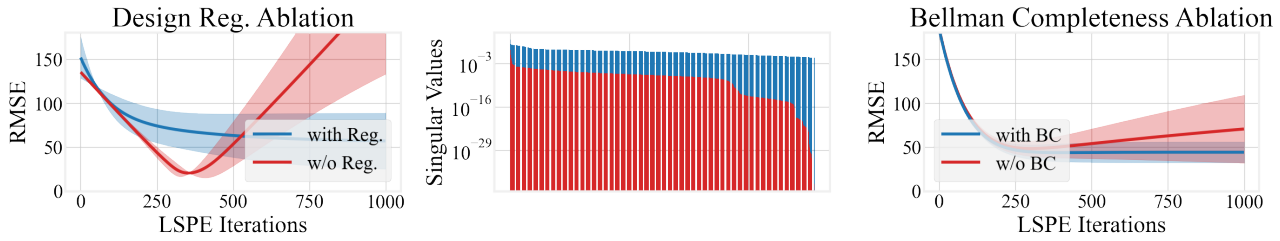


*Figure 5.* (Left) BCRL's OPE curves for Cheetah Run with (blue) and without (red) the $D$-optimal design based regularization. (Center) Bar plot of the singular values of the feature covariance matrices for the left plot. (Right) OPE curves for Finger Turn Hard with (blue) and without (red) optimizing for linear BC.

matching FQE performance across tasks. Our experiments corroborate our analysis that explicitly enforcing our learned representations to be linear BC improves OPE. Note that the fact that LSPE with SPR and CURL features still blows up under this mixture data means that the other representations' failures are not just due to coverage.

**Evaluation Beyond Initial State Distribution:** To further investigate the benefits of optimizing the $D$-optimal design to improve coverage, we investigate doing OPE *beyond* the initial state distribution $d_0$, which is supported by Theorem 4.5. Note that if our representation is exactly Bellman complete and also the corresponding feature covariance matrix is well-conditioned, we should be able to evaluate well on any states. Specifically, we experiment on evaluating at all timesteps in a target policy rollout, not just at the initial state distribution. Figure 4 (Right) shows that BCRL is able to more robustly evaluate out-of-distribution than all other benchmarks.

### 6.4. Ablation Studies

**Impact of Optimal Design Regularization:** To investigate the impact of maximizing the $D$-optimal design, we ablate the design regularization term from our objective and analyze the downstream evaluation performance and the respective feature covariance matrices on the offline dataset. Figure 5 (Center) presents a bar plot of the singular values of the feature covariance matrix $\Sigma(\phi) := \mathbb{E}_{s,a \sim \nu}[\phi(s,a)\phi(s,a)^\top]$. Figure 5 (Left) shows the downstream OPE performance for features trained with and without the design regularization on the Cheetah Run task. Note that without the regularization, we find that that the feature covariance matrix has much worse condition number, i.e. feature is less exploratory. As our analysis suggests, we also observe a deterioration in evaluation performance without the design regularization to explicitly learn exploratory features.

**Impact of Linear Bellman Completeness:** Figure 5 (Right) presents an ablation study where we only optimize for the design term in our objective. We find that downstream OPE performance degrades without directly optimizing for linear BC, suggesting that a feature with good coverage alone is not enough to avoid error amplification.

## 7. Conclusion

We present BCRL which leverages rich function approximation to learn Bellman complete and exploratory representations for stable and accurate offline policy evaluation. We provide a mathematical framework of representation learning in offline RL, which generalizes all existing representation learning frameworks from the RL theory literature. We provide an end-to-end theoretical analysis of our approach for OPE and demonstrate that BCRL can accurately estimate policy values with polynomial sample complexity. Notably, the complexity has no explicit dependence on the size of the state and action space, instead, it only depends on the statistical complexity of the representation hypothesis class. Experimentally, we extensively evaluate our approach on the DeepMind Control Suite, a set of image-based, continuous control, robotic tasks. First, we show that under the linear evaluation protocol – using linear regression on top of the representations inside the classic LSPE framework – our approach outperforms prior RL representation techniques CURL and SPR which leverage contrastive representation learning techniques SimCLR and BYOL respectively. We also show that BCRL achieves competitive OPE performance with the state-of-the-art FQE, and noticeably improves upon it when evaluating beyond the initial state distribution. Finally, our ablations show that approximate Linear Bellman Completeness and coverage are crucial ingredients to the success of our algorithm. Future work includes extending BCRL to offline policy optimization.

### Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1846210 and by a Cornell University Fellowship. We thank Rahul Kidambi, Ban Kawas, and the anonymous reviewers for useful discussions and feedback.

## References


Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *NeurIPS*, 33:20095–20107, 2020.

Amortila, P., Jiang, N., and Xie, T. A variant of the wang-foster-kakade lower bound for the discounted setting. *arXiv preprint arXiv:2011.01075*, 2020.

Anand, A., Racah, E., Ozair, S., Bengio, Y., Côté, M., and Hjelm, R. D. Unsupervised state representation learning in atari. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *NeurIPS*, pp. 8766–8779, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/6fb52e71b837628ac16539c1ff911667-Abstract.html.

Antos, A., Szepesvári, C., and Munos, R. Fitted q-iteration in continuous action-space mdps. In Platt, J., Koller, D., Singer, Y., and Roweis, S. (eds.), *NIPS*, volume 20. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2007/file/da0d1111d2dc5d489242e60ebcbaf988-Paper.pdf.

Bhatia, R. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *ICML*, pp. 1042–1051. PMLR, 2019.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020. URL https://arxiv.org/abs/2002.05709.

Chung, W., Nath, S., Joseph, A., and White, M. Two-timescale networks for nonlinear value function approximation. In *ICLR*, 2019. URL https://openreview.net/forum?id=rJleN20qK7.

Du, S. S., Kakade, S. M., Lee, J. D., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. *arXiv preprint arXiv:2103.10897*, 2021.

Duan, Y., Jia, Z., and Wang, M. Minimax-optimal off-policy evaluation with linear function approximation. In *ICML*, pp. 2701–2709. PMLR, 2020.

Foster, D. J., Krishnamurthy, A., Simchi-Levi, D., and Xu, Y. Offline reinforcement learning: Fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*, 2021.

Fu, J., Norouzi, M., Nachum, O., Tucker, G., Wang, Z., Novikov, A., Yang, M., Zhang, M. R., Chen, Y., Kumar, A., Paduraru, C., Levine, S., and Paine, T. L. Benchmarks for deep off-policy evaluation. *CoRR*, abs/2103.16596, 2021. URL https://arxiv.org/abs/2103.16596.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., and Valko, M. Bootstrap your own latent - a new approach to self-supervised learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *NeurIPS*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *ICML*, pp. 1861–1870. PMLR, 2018.

Hafner, D., Lillicrap, T. P., Norouzi, M., and Ba, J. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0oabwyZbOu.

Hao, B., Duan, Y., Lattimore, T., Szepesvári, C., and Wang, M. Sparse feature selection makes batch reinforcement learning more sample efficient. In *ICML*, pp. 4063–4073. PMLR, 2021.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. Momentum contrast for unsupervised visual representation learning. *CoRR*, abs/1911.05722, 2019. URL http://arxiv.org/abs/1911.05722.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *ICML*, pp. 652–661. PMLR, 2016.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *ICML*, pp. 1704–1713. PMLR, 2017.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *COLT*, pp. 2137–2143. PMLR, 2020.

Kostrikov, I. and Nachum, O. Statistical bootstrapping for uncertainty estimation in off-policy evaluation, 2020. URL https://arxiv.org/abs/2007.13609.

Laskin, M., Srinivas, A., and Abbeel, P. Curl: Contrastive unsupervised representations for reinforcement learning. In *ICML*, pp. 5639–5650. PMLR, 2020.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. In *ICLR*, 2016.

Mazoure, B., Tachet des Combes, R., Doan, T. L., Bachman, P., and Hjelm, R. D. Deep reinforcement and infomax learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *NeurIPS*, volume 33, pp. 3686–3698. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/26588e932c7ccfa1df309280702fe1b5-Paper.pdf.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. A. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013. URL http://arxiv.org/abs/1312.5602.

Modi, A., Chen, J., Krishnamurthy, A., Jiang, N., and Agarwal, A. Model-free representation learning and exploration in low-rank mdps. *arXiv preprint arXiv:2102.07035*, 2021.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *JMLR*, 9(5), 2008.

Nachum, O. and Yang, M. Provable representation learning for imitation with contrastive fourier features. *CoRR*, abs/2105.12272, 2021. URL https://arxiv.org/abs/2105.12272.

Nedić, A. and Bertsekas, D. P. Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dynamic Systems*, 13(1):79–110, 2003.

Ni, C., Zhang, A. R., Duan, Y., and Wang, M. Learning good state and action representations via tensor decomposition. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 1682–1687. IEEE, 2021.

Parr, R., Li, L., Taylor, G., Painter-Wakefield, C., and Littman, M. L. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In *ICML*, pp. 752–759, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390251. URL https://doi.org/10.1145/1390156.1390251.

Pollard, D. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pp. i–86. JSTOR, 1990.

Schwarzer, M., Anand, A., Goel, R., Hjelm, R. D., Courville, A. C., and Bachman, P. Data-efficient reinforcement learning with momentum predictive representations. *CoRR*, abs/2007.05929, 2020. URL https://arxiv.org/abs/2007.05929.

Song, Z., Parr, R. E., Liao, X., and Carin, L. Linear feature encoding for reinforcement learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *NIPS*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/8232e119d8f59aa83050a741631803a6-Paper.pdf.

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., Casas, D. d. L., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Thomas, P. S. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning, 2016. URL https://arxiv.org/abs/1604.00923.

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline RL in low-rank MDPs. In *ICLR*, 2022. URL https://openreview.net/forum?id=J4iSIR9fhY0.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

van der Vaart, A. W. and Wellner, J. A. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer New York, 1996. ISBN 9781475725476. doi: 10.1007/978-1-4757-2545-2. URL http://link.springer.com/10.1007/978-1-4757-2545-2.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, R., Foster, D. P., and Kakade, S. M. What are the statistical limits of offline RL with linear function approximation? *CoRR*, abs/2010.11895, 2020. URL https://arxiv.org/abs/2010.11895.

Wang, Y., Wang, R., and Kakade, S. M. An exponential lower bound for linearly-realizable mdps with constant suboptimality gap. *arXiv preprint arXiv:2103.12690*, 2021.

Yang, M. and Nachum, O. Representation matters: Offline pretraining for sequential decision making. In Meila, M. and Zhang, T. (eds.), *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11784–11794. PMLR, 2021. URL http://proceedings.mlr.press/v139/yang21h.html.

Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. Off-policy evaluation via the regularized lagrangian. *NeurIPS*, 33:6551–6561, 2020.

Yarats, D., Zhang, A., Kostrikov, I., Amos, B., Pineau, J., and Fergus, R. Improving sample efficiency in model-free reinforcement learning from images. *CoRR*, abs/1910.01741, 2019. URL http://arxiv.org/abs/1910.01741.

Yarats, D., Fergus, R., Lazaric, A., and Pinto, L. Mastering visual continuous control: Improved data-augmented reinforcement learning. *arXiv preprint arXiv:2107.09645*, 2021.

Zhang, W., He, J., Zhou, D., Zhang, A., and Gu, Q. Provably efficient representation learning in low-rank markov decision processes. *arXiv preprint arXiv:2106.11935*, 2021.

# Appendices

## A. Metric Entropy and Entropy Integral

We recall the standard notions of entropy integrals here, based on the following distance on ,

$$d\ (\phi, \widetilde{\phi}) = \mathsf{E}_D\left[\left\|\phi(s,a) \quad \widetilde{\phi}(s,a)\right\|_2\right]$$

Let $N(t, \ )$ denote the $t$-covering number under $d$ .

**Definition A.1.** Define the entropy integral, which we assume to be finite as,

$$\kappa(\ ) = \int_0^4 \log^{1/2} N(t, \ )\mathrm{d}t$$

When is finite, $N(t)$ $j\ j$, so $\kappa(\ )$ $O(\log^{1/2}(j\ j))$.

## B. Technical Lemmas

**Lemma B.1.** *Let $X_i$ be i.i.d. random variables s.t. $jX_ij$ $c$ and $\mathsf{E}\left[X_i^2\right]$ $\nu$, then for any $\delta$ 2 $(0,1)$, we have w.p. at least 1 $\delta$,*

$$\left|\frac{1}{N}\sum_{i=1}^N X_i \quad \mathsf{E}[X]\right| \quad \inf_{a>0}\frac{\nu}{2a} + \frac{(c+a)\log(2/\delta)}{N},$$

*and if $\nu$ $L\mathsf{E}[X]$ for some positive $L$, then in particular,*

$$\left|\frac{1}{N}\sum_{i=1}^N X_i \quad \mathsf{E}[X]\right| \quad \frac{1}{2}\mathsf{E}[X] + \frac{(c+L)\log(2/\delta)}{N}.$$

*Proof.* First, by Bernstein's inequality (Boucheron et al., 2013, Theorem 2.10), we have w.p. 1 $\delta$,

$$\left|\frac{1}{N}\sum_{i=1}^N X_i \quad \mathsf{E}[X]\right| \quad \sqrt{\frac{2\nu\log(2/\delta)}{N}} + \frac{c\log(2/\delta)}{N}$$

Using the fact that $2xy$ $x^2/a + ay^2$ for any $a > 0$, split the square root term,

$$\inf_{a>0}\frac{\nu}{2a} + \frac{(c+a)\log(2/\delta)}{N}$$

which yields the first part. If $\nu$ $L\mathsf{E}[X]$, picking $a = L$ concludes the proof. □

We now state several results of Orlicz norms (mostly from Pollard, 1990) for completeness. For an increasing, convex, positive function $: \mathsf{R}^+ \nearrow \mathsf{R}^+$, such that $(x)$ 2 $[0,1)$, define the Orlicz norm as

$$kZk \ := \inf\{C > 0\ j\ \mathsf{E}[\ (jZj/C)] \quad 1\}.$$

It is indeed a norm on the -Orlicz space of random variables $L$ $(\nu)$, since it can be interpreted as the Minkowski functional of the convex set $K = \{X : \mathsf{E}[\ (jXj)] \quad 1\}$.

Let $x_1, x_2, \ldots, x_N$ be i.i.d. datapoints drawn from some underlying distribution. Let $\omega$ denote the randomness of the $N$ sampled datapoints, and let $F_\omega = \left\{(f(x_i(\omega)))_{i=1}^N : f\ 2\ F\right\}$ $\mathsf{R}^N$ denote the (random) set of vectors from the data corresponding to $\omega$. Let $\boldsymbol{\sigma}$ denote a vector of $N$ i.i.d. Rademacher random variables ( 1 equi-probably), independent of all else.

**Lemma B.2** (Symmetrization). *For any increasing, convex, positive function , we have*

$$\mathsf{E}_{\omega}\left[\left(\sup_{f\,2\,F_{\omega}}\left|\sum_{i=1}^{N} f_i \quad \mathsf{E}_{\omega} f_i\right|\right)\right] \quad \mathsf{E}_{\omega,}\left[\left(2\sup_{f\,2\,F_{\omega}} j\sigma \quad fj\right)\right].$$

*Proof.* See Theorem 2.2 of (Pollard, 1990). □

**Lemma B.3** (Contraction). *Let F R^N, and suppose λ : R^N 7! R^N s.t. each component λ_i : R 7! R is L-Lipschitz. Then, for any increasing, convex, positive function , we have*

$$\mathsf{E}\left[\left(\sup_{f\,2\,F} j\sigma \quad \lambda(f)j\right)\right] \quad \frac{3}{2}\mathsf{E}\left[\left(2L\sup_{f\,2\,F} j\sigma \quad fj\right)\right]$$

*Proof.* Apply Theorem 5.7 of (Pollard, 1990) to the functions $\lambda_i/L$, which are contractions. □

We now focus on the specific Orlicz space of sub-Gaussian random variables, with the function $(x) = \frac{1}{5}\exp(x^2)$. The -Orlicz norm of the maximum of random variables can be bounded by the maximum of the -Orlicz norms.

**Lemma B.4.** *For any random variables $Z_1, ..., Z_m$, we have*

$$k\max_{i\ m} jZ_i jk \qquad \sqrt{2 + \log(m)}\max_{i\ m} kZ_i k$$

*Proof.* See Lemma 3.2 of (Pollard, 1990). □

**Lemma B.5.** *For each f 2 R^N, we have kσ fk 2kfk_2.*

*Proof.* See Lemma 3.1 of (Pollard, 1990). □

The following is a truncated chaining result for Orlicz norms. This result is not new, but many sources state and prove it in terms of covering and Rademacher complexity, rather than for Orlicz norms. In particular, it generalizes Theorem 3.5 of (Pollard, 1990) – consider a sequence of $\alpha$'s converging to zero.

**Lemma B.6** (Chaining). *Let F R^N such that 0 2 F. Then,*

$$\left\|\sup_{f\,2\,F} j\sigma \quad fj\right\| \qquad \inf_{\alpha\ 0}\left\{3\alpha^{p}\overline{N} + 9\int_{\alpha}^{b}\sqrt{\log(D(\delta/2, F))}\mathrm{d}\delta\right\},$$

*where $b = \sup_{f\,2\,F} kfk_2$ and $D(\delta, F)$ is the Euclidean $\delta$-packing number for F.*

*Proof.* Suppose $b$ and all the packing numbers are finite, otherwise the right hand side is infinite and there is nothing to show. For an arbitrary $K > 1$, construct a sequence of $K$ finer and finer approximations to $F$,

$$f0g = F_0 \quad F_1 \qquad\qquad F_K \quad F_{K+1} = F$$

where for any $k$ 2 $[K]$, $F_k$ satisfies the property that for any $f$ 2 $F$, there exist $n_k(f)$ 2 $F_k$ s.t. $kn_k(f) \quad fk_2 \quad b2^{\ k}$. Indeed this can be done iteratively: for any $F_k$, we can construct $F_{k+1}$ by adding elements to construct a maximal $b2^{\ k}$ packing (maximality ensures the distance requirement, since the existence of any vector which has larger distance can be added to the packing). By definition of $D(\ , F)$, we have $jF_k j \quad D(b2^{\ k}, F)$.

For any $k \in [K]$, we have by triangle inequality,

$$\sup_{f \in F_{k+1}} |\sigma^\top (f - f)| \leq \sup_{f \in F_{k+1}} |\sigma^\top (f - n_k(f))| + \sup_{f \in F_{k+1}} |\sigma^\top (n_k(f) - f)|$$
$$= \sup_{f \in F_k} |\sigma^\top (f - f)| + \sup_{f \in F_{k+1}} |\sigma^\top (n_k(f) - f)|$$

If $k = K$, we can loosely bound the right-most term by Cauchy-Schwarz, since for any $f \in F$, we have $|\sigma^\top (n_k(f) - f)| \leq \sqrt{N} \|n_k(f) - f\|_2 \leq \sqrt{N} b 2^{-K}$. So, we have

$$\left\| \sup_{f \in F} |\sigma^\top (f - f)| \right\|_{\psi_2} \leq \left\| \max_{f \in F_K} |\sigma^\top (f - f)| \right\|_{\psi_2} + \frac{\sqrt{N} b 2^{-K}}{\sqrt{\log 5}},$$

since for any non-negative constant $c$, $\|c\|_{\psi_2} = \inf\{C > 0 : \frac{1}{5} \exp((c/C)^2) \leq 1\} = \frac{c}{\sqrt{\log 5}}$. If $k < K$, the suprema are taken over finite sets, so the maximum is attained. Hence, we can apply a special property of the $\psi_2$-Orlicz norm (Lemma B.4), to get,

$$\left\| \max_{f \in F_{k+1}} |\sigma^\top (f - f)| \right\|_{\psi_2} \leq \left\| \max_{f \in F_k} |\sigma^\top (f - f)| \right\|_{\psi_2} + \left\| \max_{f \in F_{k+1}} |\sigma^\top (n_k(f) - f)| \right\|_{\psi_2}$$
$$\left\| \max_{f \in F_k} |\sigma^\top (f - f)| \right\|_{\psi_2} + \sqrt{2 + \log(|F_{k+1}|)} \max_{f \in F_{k+1}} \|\sigma\|_2 \|n_k(f) - f\|_2$$

By Lemma B.5,

$$\left\| \max_{f \in F_k} |\sigma^\top (f - f)| \right\|_{\psi_2} + 2\sqrt{2 + \log(|F_{k+1}|)} \max_{f \in F_{k+1}} \|n_k(f) - f\|_2$$
$$\left\| \max_{f \in F_k} |\sigma^\top (f - f)| \right\|_{\psi_2} + 2\sqrt{2 + \log(|F_{k+1}|)} \; b 2^{-k}.$$

Also, note that since $F_0 = \{0\}$ by construction, we have $\|\max_{f \in F_0} |\sigma^\top (f - f)|\|_{\psi_2} = 0$. Unrolling this, we have

$$\|\sup_{f \in F} |\sigma^\top (f - f)|\|_{\psi_2} \leq \frac{\sqrt{N} b 2^{-K}}{\sqrt{\log 5}} + \sum_{k=0}^{K-1} 2^{-k} 2b \sqrt{2 + \log(D(b2^{-(k+1)}, F))}$$

Since for any $D \geq 2$, we have $\sqrt{2 + \log(1 + D)} / \sqrt{D} \leq 2.2$,

$$\frac{\sqrt{N} b 2^{-K}}{\sqrt{\log 5}} + 4.4b \sum_{k=0}^{K-1} 2^{-k} \sqrt{\log(D(b2^{-(k+1)}, F))}$$
$$= \frac{\sqrt{N} b 2^{-K}}{\sqrt{\log 5}} + 17.6b \sum_{k=0}^{K-1} (2^{-(k+1)} - 2^{-(k+2)}) \sqrt{\log(D(b2^{-(k+1)}, F))}$$

Since $D(\cdot, F)$ is a monotone decreasing,

$$\frac{\sqrt{N} b 2^{-K}}{\sqrt{\log 5}} + 17.6 \int_{b2^{-(K+1)}}^{b/2} \sqrt{\log(D(\delta, F))} d\delta$$
$$= \frac{\sqrt{N} b 2^{-K}}{\sqrt{\log 5}} + 8.8 \int_{b2^{-K}}^{b} \sqrt{\log(D(\delta/2, F))} d\delta.$$

Now consider any $\alpha > 0$. Pick $K$ such that $\frac{b}{2^{K+1}} \leq \alpha \leq \frac{b}{2^K}$, then we have

$$\|\sup_{f \in F} |\sigma^\top (f - f)|\|_{\psi_2} \leq \frac{2\alpha \sqrt{N}}{\sqrt{\log 5}} + 8.8 \int_{\alpha}^{b} \sqrt{\log(D(\delta/2, F))} d\delta.$$

Since $\alpha$ was arbitrary, the above bound holds when we take an infimum over $\alpha$. □

## C. Proofs for LSPE

We first show a generalization of the performance difference lemma (PDL).

**Lemma C.1** (Generalized PDL)**.** *For any policies $\pi, \pi'$, any function $f : S \times A \to \mathbb{R}$, and any initial state distribution $\mu$, we have*

$$V_\mu^\pi - \mathbb{E}_{s \sim \mu}[f(s, \pi')] = \frac{1}{1-\gamma}\mathbb{E}_{s,a \sim d_\mu^\pi}\left[\mathcal{T}^{\pi'} f(s,a) - f(s,a)\right] \tag{7}$$

*Proof.* Let $T^\pi$ be the distribution of trajectories $\tau = (s_0, a_0, s_1, a_1, s_2, a_2, ...)$ from rolling out $\pi$. So, we have,

$$\begin{aligned}
V_\mu^\pi - \mathbb{E}_{s \sim \mu}[f(s, \pi')] &= \mathbb{E}_{\tau \sim T^\pi}\left[\sum_{t=0}^{1} \gamma^t r(s_t, a_t)\right] - \mathbb{E}_{s \sim \mu}[f(s, \pi')] \\
&= \mathbb{E}_{\tau \sim T^\pi}\left[\sum_{t=0}^{1} \gamma^t \left(r(s_t, a_t) + f(s_t, \pi') - f(s_t, \pi')\right) - f(s_0, \pi')\right] \\
&= \mathbb{E}_{\tau \sim T^\pi}\left[\sum_{t=0}^{1} \gamma^t \left(r(s_t, a_t) + \gamma f(s_{t+1}, \pi') - f(s_t, \pi')\right)\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s,a \sim d_\mu^\pi}\left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)}[f(s', \pi')] - f(s, \pi')\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s,a \sim d_\mu^\pi}\left[\mathcal{T}^{\pi'} f(s,a) - f(s,a)\right].
\end{aligned}$$

□

This generalizes the PDL, which we can get by setting $f(s,a) = Q^{\pi'}(s,a)$:

$$\begin{aligned}
V_\mu^\pi - V_\mu^{\pi'} &= \frac{1}{1-\gamma}\mathbb{E}_{s,a \sim d_\mu^\pi}\left[\mathcal{T}^{\pi'} Q^{\pi'}(s,a) - Q^{\pi'}(s, \pi')\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s,a \sim d_\mu^\pi}\left[Q^{\pi'}(s,a) - V^{\pi'}(s)\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s,a \sim d_\mu^\pi}\left[A^{\pi'}(s,a)\right].
\end{aligned}$$

To prove our LSPE guarantee, we'll instantiate $f(s,a)$ to be the estimated $\widehat{f}(s,a)$ from LSPE, and set $\pi = \pi'$. This gives us an expression for the prediction error of LSPE,

$$\left|V_\mu^\pi - \mathbb{E}_{s \sim \mu}\left[\widehat{f}_k(s, \pi)\right]\right| = \frac{1}{1-\gamma}\left|\mathbb{E}_{d_\mu^\pi}\left[\mathcal{T}^\pi \widehat{f}_k(s,a) - f_k(s,a)\right]\right|.$$

We then upper bound the right hand side by its $L_2(d_\mu^\pi)$ norm, which is the Bellman error. Next, we bound the Bellman error of running LSPE by the regression losses at each step.

**Lemma C.2.** *Consider any policy $\pi$ and functions $f_1, \ldots, f_K : S \times A \to \mathbb{R}$ that satisfy $\max_{k=1,\ldots,K} \|f_k - \mathcal{T}^\pi f_{k-1}\|_{L_2(d_{p_0}^\pi)} \le \eta$, and $f_0(s,a) = 0$. Then, for all $k = 1, \ldots, K$, we have $\|f_k - \mathcal{T}^\pi f_k\|_{L_2(d_{p_0}^\pi)} \le \frac{4}{1-\gamma}\eta + \gamma^{k/2}$.*

*Proof.* For any $k = 1, \ldots, K$,

$$\begin{aligned}
\|f_k - \mathcal{T}^\pi f_k\|_{L_2(d_{p_0}^\pi)} &\le \|f_k - \mathcal{T}^\pi f_{k-1}\|_{L_2(d_{p_0}^\pi)} + \|\mathcal{T}^\pi f_{k-1} - \mathcal{T}^\pi f_k\|_{L_2(d_{p_0}^\pi)} \\
&\le \eta + \gamma(\mathbb{E}_{s,a \sim d_{p_0}^\pi}[(\mathbb{E}_{s',a' \sim P(s,a) \times \pi}[f_{k-1}(s',a') - f_k(s',a')])^2])^{1/2} \\
&\le \eta + \gamma(\mathbb{E}_{s,a,s',a' \sim d_{p_0}^\pi \times P \times \pi}[(f_{k-1}(s',a') - f_k(s',a'))^2])^{1/2}
\end{aligned}$$

Since $d^\pi_{p_0}(s,a) = \gamma \mathsf{E}_{s,a\ \ d^\pi_{p_0}} P(s|s,a)\pi(a|s) + (1\ \ \gamma)p_0(s)\pi(a|s)$, and the quantity inside the expectation is a square, and thus non-negative,

$$\begin{aligned} &\eta + \gamma(\gamma\ \ ^1\mathsf{E}_{s,a\ \ d^\pi_{p_0}}[(f_{k\ \ 1}(s,a)\ \ f_k(s,a))^2])^{1/2} \\ =\ &\eta + \sqrt{\gamma}\|f_{k\ \ 1}\ \ f_k\|_{L_2(d^\pi_{p_0})} \\ &\eta + \sqrt{\gamma}\left(\eta + \|f_{k\ \ 1}\ \ \mathcal{T}^\pi f_{k\ \ 1}\|_{L_2(d^\pi_{p_0})}\right). \end{aligned}$$

Unrolling the recursion and using $\|f_0\ \ \mathcal{T}^\pi f_0\|_{L_2(d^\pi_{p_0})}\ \ 1$, we have

$$\begin{aligned} \|f_K\ \ \mathcal{T}^\pi f_K\|_{L_2(d^\pi_{p_0})}\ \ \eta + \sqrt{\gamma}(\eta + \sqrt{\gamma}(\eta + \ldots\sqrt{\gamma}(\eta + 1)\ldots)) \\ = \frac{1\ \ \sqrt{\gamma}^K}{1\ \ \sqrt{\gamma}}\eta + \gamma^{K/2}, \end{aligned}$$

which gives the claim since $\frac{1}{1\ \ \sqrt{\gamma}}\ \ 2(1\ \ \gamma)\ \ ^1$ and $1\ \ \sqrt{\gamma}^K\ \ 1$. □

The following lemma is a "fast rates"-like result for norms. It shows that the norm induced by the empirical covariance matrix $\widehat{\ }$ can be bounded by the norm induced by two times the population covariance matrix $\ $, up to some $\widetilde{O}(N^{\ 1/2})$ terms.

**Lemma C.3** (Fast Rates for $\ $-norm)**.** *For any $\delta\ 2\ (0,1)$, we have with probability at least $1\ \ \delta$, for any $x\ 2\ B_W$, we have*

$$\|x\|_{\widehat{\ }}\ \ 2\|x\|_{\ }\ + 5W\sqrt{\frac{d\log(N/\delta)}{N}},$$

*and*

$$\|x\|_{\ }\ \ 2\|x\|_{\widehat{\ }} + 5W\sqrt{\frac{d\log(N/\delta)}{N}}.$$

*Proof.* First, fix any $x\ 2\ B_W$. Since $(x^\top\phi(s,a))^2\ \ W^2$ almost surely, we have $\mathsf{E}\left[(x^\top\phi(s,a))^4\right]\ \ W^2\mathsf{E}\left[(x^\top\phi(s,a))^2\right]$. So by Lemma B.1, w.p. at least $1\ \ \delta$,

$$\left|\frac{1}{N}\sum_{i=1}^N (x^\top\phi(s_i,a_i))^2\ \ \mathsf{E}\left[(x^\top\phi(s,a))^2\right]\right|\ \ \frac{1}{2}\mathsf{E}\left[(x^\top\phi(s,a))^2\right] + \frac{2W^2\log(2/\delta)}{N}.$$

In particular, this means $\|x\|^2_{\widehat{\ }}\ \ \frac{3}{2}\|x\|^2_{\ } + \frac{2W^2\log(2/\delta)}{N}$ and $\|x\|^2_{\ }\ \ 2\|x\|^2_{\widehat{\ }} + \frac{4W^2\log(2/\delta)}{N}$. Now union bound over a $W/\sqrt{N}$-net of $B_W$, which can be done with $(1 + 2\sqrt{N})^d$ elements. The approximation error from this cover is

$$\begin{aligned} \|x\|_{\widehat{\ }}\ \ &\|n(x)\|_{\widehat{\ }} + \|n(x)\ \ x\|_{\widehat{\ }} \\ &\sqrt{\frac{3}{2}}\|n(x)\|_{\ } + \sqrt{\frac{2W^2\log(2(1+2\sqrt{N})^d/\delta)}{N}} + \frac{W}{\sqrt{N}} \\ &\sqrt{\frac{3}{2}}\|x\|_{\ } + \sqrt{\frac{3}{2}}\|n(x)\ \ x\|_{\ } + \sqrt{\frac{2W^2\log(2(1+2\sqrt{N})^d/\delta)}{N}} + \frac{W}{\sqrt{N}} \\ &\sqrt{\frac{3}{2}}\|x\|_{\ } + 4W\sqrt{\frac{d\log(N/\delta)}{N}}, \end{aligned}$$

where $n(x)$ is the closest element in the net to $x$. Similarly,

$$\begin{aligned}
kxk \quad & kn(x)k \quad + kn(x) \quad xk \\
& 2kn(x)k\widehat{\ } + \sqrt{\frac{4W^2 \log(2(1+2^{p\overline{N})^d/\delta)}{N}} + p\frac{W}{\overline{N}} \\
& 2kxk\widehat{\ } + 2kn(x) \quad xk\widehat{\ } + + \sqrt{\frac{4W^2 \log(2(1+2^{p\overline{N})^d/\delta)}{N}} + p\frac{W}{\overline{N}} \\
& 2kxk\widehat{\ } + 5W\sqrt{\frac{d\log(N/\delta)}{N}}.
\end{aligned}$$

□

Now define the following notation for analyzing Least Squares Policy Evaluation (LSPE). For every target vector $k\vartheta k_2 \quad W$, define

$$\begin{aligned}
y^{\vartheta} &= r + \gamma\vartheta^{|}\phi(s^{0},\pi), \\
\theta_{\vartheta} &= \operatorname*{arg\,min}_{k\theta k \quad W} \mathsf{E}_{s,a \quad \nu,s^{0} \quad P(s,a)}[(y^{\vartheta} \quad \theta^{|}\phi(s,a))^2] := \ell(\theta,\vartheta), \\
y_i^{\vartheta} &= r_i + \gamma\vartheta^{|}\phi(s_i^{0},\pi), \\
\widehat{\theta}_{\vartheta} &= \operatorname*{arg\,min}_{k\theta k \quad W} \frac{1}{N}\sum_{i=1}^{N}(y_i^{\vartheta} \quad \theta^{|}\phi(s_i,a_i))^2 := \widehat{\ell}(\theta,\vartheta).
\end{aligned}$$

The following lemma are useful facts about the optimal $\theta_{\vartheta}$ and $\widehat{\theta}_{\vartheta}$.

**Lemma C.4.** *For any $\vartheta_1, \vartheta_2$, we have*

$$\begin{aligned}
k\theta_{\vartheta_1} \quad \theta_{\vartheta_2}k \quad & \sqrt{2\gamma W k\vartheta_1 \quad \vartheta_2 k_2}, \\
k\widehat{\theta}_{\vartheta_1} \quad \widehat{\theta}_{\vartheta_2}k\widehat{\ } \quad & \sqrt{2\gamma W k\vartheta_1 \quad \vartheta_2 k_2}.
\end{aligned}$$

*For any $\theta, \vartheta$, we have*

$$\ell(\theta,\vartheta) \quad \ell(\theta_{\vartheta},\vartheta) \quad k\theta \quad \theta_{\vartheta}k^2 .$$

*Proof.* First, recall that $\theta_{\vartheta}$ minimizes $f(\theta) = \mathsf{E}\left[(y^{\vartheta} \quad \theta^{|}\phi(s,a))^2\right]$, which has the Jacobian $r f(\theta) = 2\mathsf{E}\left[(\theta^{|}\phi(s,a) \quad y^{\vartheta})\phi(s,a)\right]$. Since $\theta_{\vartheta}$ is optimal over $B_W$, the necessary optimality condition is that $r f(\theta_{\vartheta}) \ 2 \ N_{B_W}(\theta_{\vartheta})$, the normal cone of $B_W$ at $\theta_{\vartheta}$, i.e. for any $\theta \ 2 \ B_W$, we have $h\mathsf{E}\left[\phi(s,a)(y^{\vartheta} \quad \theta_{\vartheta}^{|}\phi(s,a))\right], \theta \quad \theta_{\vartheta} i \quad 0$. In particular, we have

$$\begin{aligned}
h\mathsf{E}\left[\phi(s,a)(y^{\vartheta_1} \quad \theta_{\vartheta_1}^{|}\phi(s,a))\right], \theta_{\vartheta_2} \quad \theta_{\vartheta_1} i \quad 0 \\
h\mathsf{E}\left[\phi(s,a)(y^{\vartheta_2} \quad \theta_{\vartheta_2}^{|}\phi(s,a))\right], \theta_{\vartheta_1} \quad \theta_{\vartheta_2} i \quad 0
\end{aligned}$$

adding the two we get

$$\begin{aligned}
k\theta_{\vartheta_1} \quad \theta_{\vartheta_2}k^2 &= h\mathsf{E}\left[\phi(s,a)(\theta_{\vartheta_1} \quad \theta_{\vartheta_2})^{|}\phi(s,a))\right], \theta_{\vartheta_1} \quad \theta_{\vartheta_2} i \\
& \quad h\mathsf{E}\left[(y^{\vartheta_1} \quad y^{\vartheta_2})\phi(s,a)\right], \theta_{\vartheta_1} \quad \theta_{\vartheta_2} i \\
&= \gamma h\mathsf{E}\left[\phi(s,a)\phi(s^{0},\pi)^{|}\right](\vartheta_1 \quad \vartheta_2), \theta_{\vartheta_1} \quad \theta_{\vartheta_2} i \\
& \quad \gamma k\mathsf{E}\left[\phi(s,a)\phi(s^{0},\pi)^{|}\right]k_2 k\vartheta_1 \quad \vartheta_2 k_2 k\theta_{\vartheta_1} \quad \theta_{\vartheta_2}k_2 \\
& \quad 2\gamma W k\vartheta_1 \quad \vartheta_2 k_2.
\end{aligned}$$

The claim with $\widehat{\theta}_{\vartheta_1}, \widehat{\theta}_{\vartheta_2}$ follows by the same argument.

For the second claim, we first apply the Parallelogram law, followed by the first-order optimality of $\theta_\vartheta$,

$$\begin{aligned}\ell(\theta,\vartheta) &= \ell(\theta_\vartheta,\vartheta) + \mathsf{E}_\nu((\theta_\vartheta - \theta)^\top\phi(s,a))^2 + 2\mathsf{E}_\nu[(y^\vartheta - \theta^\top\phi(s,a))\phi(s,a)^\top(\theta_\vartheta - \theta)] \\ &\le \ell(\theta_\vartheta,\vartheta) + \mathsf{E}_\nu((\theta_\vartheta - \theta)^\top\phi(s,a))^2 + 0 \\ &= \ell(\theta_\vartheta,\vartheta) + \|\theta - \theta_\vartheta\|^2 .\end{aligned}$$

□

Now we show our key lemma about the concentration of least squares, uniformly over all targets generated by $\vartheta \in B_W$.

**Lemma C.5** (Concentration for Least Squares)**.** *Let $\mathcal{F} = \{(s,a) \mapsto \theta^\top\phi(s,a) : \|\theta\|_2 \le W\}$, with $\|\phi(s,a)\|_2 \le 1$. Then, for any $\delta \in (0,1)$ w.p. at least $1-\delta$, we have*

$$\sup_{\vartheta \in B_W} \|\widehat{\theta}_\vartheta - \theta_\vartheta\| \le 120 d(1+W)\frac{\log(N)\sqrt{\log(10/\delta)}}{\overline{N}}.$$

*Proof.* By Lemma C.3, we have that w.p. at least $1-\delta$, we can bound the random $\|\cdot\|$ $\|\cdot\|_{\widehat{\cdot}}$ norm by $\|\cdot\|$ $\|\cdot\|$, and vice versa, simultaneously over all vectors in a ball,

$$\mathcal{E} = \{\forall x \in B_{2W}, \|x\|_{\widehat{\cdot}} \le 2\|x\| + t \text{ and } \|x\| \le 2\|x\|_{\widehat{\cdot}} + t\},$$

provided $t \ge 5W\sqrt{\frac{d\log(N/\delta)}{N}}$. For the remainder of this proof, we *condition* on this high-probability event. That is, any probability and expectations will be *implicitly conditioned* on $\mathcal{E}$.

First, we'll show that for an arbitrary and fixed $\vartheta \in B_W$, w.p. at least $1-\delta$, we have

$$\|\widehat{\theta}_\vartheta - \theta_\vartheta\| \le t.$$

To do so, we will bound the probability of the complement, which in turn can be simplified by the following chain of arguments,

$$\|\widehat{\theta}_\vartheta - \theta_\vartheta\| \ge t$$

By $\ell(\theta,\vartheta) - \ell(\theta_\vartheta,\vartheta) \le \|\theta - \theta_\vartheta\|$ from Lemma C.4,

$$\begin{aligned}&\Rightarrow \ell(\widehat{\theta}_\vartheta,\vartheta) - \ell(\theta_\vartheta,\vartheta) \ge t^2 \\ &\Rightarrow \exists\|\theta - \theta_\vartheta\| \le W : \ell(\theta,\vartheta) - \ell(\theta_\vartheta,\vartheta) \ge t^2, \widehat{\ell}(\theta,\vartheta) - \widehat{\ell}(\theta_\vartheta,\vartheta) \le 0\end{aligned}$$

By convexity and continuity of $\ell(\cdot,\vartheta)$, we can make this strict equality. Indeed, given this, by Intermediate Value Theorem, there exists $\lambda \in [0,1]$ such that $\theta^0 = (1-\lambda)\theta + \lambda\theta_\vartheta$ has $\ell(\theta^0,\vartheta) - \ell(\theta_\vartheta,\vartheta) = \nu$. Then by convexity, $\widehat{\ell}(\theta^0,\vartheta) - \widehat{\ell}(\theta_\vartheta,\vartheta) \le (1-\lambda)\widehat{\ell}(\theta,\vartheta) - (1-\lambda)\widehat{\ell}(\theta_\vartheta,\vartheta) \le 0$.

$$\begin{aligned}&\Rightarrow \exists\|\theta - \theta_\vartheta\| \le W : \ell(\theta,\vartheta) - \ell(\theta_\vartheta,\vartheta) = t^2, \widehat{\ell}(\theta,\vartheta) - \widehat{\ell}(\theta_\vartheta,\vartheta) \le 0 \\ &\Rightarrow \exists\|\theta - \theta_\vartheta\| \le W : \|\theta - \theta_\vartheta\| \ge t, (\ell(\theta,\vartheta) - \widehat{\ell}(\theta,\vartheta)) - (\ell(\theta_\vartheta,\vartheta) - \widehat{\ell}(\theta_\vartheta,\vartheta)) \ge t^2\end{aligned}$$

By conditioning on $\mathcal{E}$,

$$\Rightarrow \exists\|\theta - \theta_\vartheta\| \le W : \|\theta - \theta_\vartheta\|_{\widehat{\cdot}} \le 3t, (\ell(\theta,\vartheta) - \widehat{\ell}(\theta,\vartheta)) - (\ell(\theta_\vartheta,\vartheta) - \widehat{\ell}(\theta_\vartheta,\vartheta)) \ge t^2$$

Hence, we now focus on bounding

$$\mathsf{P}\left(\sup_{\theta \in \mathcal{F}}\left|\sum_{i=1}^N \zeta_i(\theta,\vartheta) - \zeta_i(\theta_\vartheta,\vartheta) - \mathsf{E}_\nu[\zeta_i(\theta,\vartheta) - \zeta_i(\theta_\vartheta,\vartheta)]\right| \ge Nt^2\right) \le \delta,$$

where we define

$$\begin{aligned}\zeta_i(\theta,\vartheta) &= (y_i^{\vartheta} \quad \theta^{\intercal}\phi(s_i,a_i))^2,\\ &= f\theta \ : \ k\theta k \quad W, k\theta \quad \theta_{\vartheta} k_{\frown} \quad 3tg. \end{aligned} \tag{8}$$

Since $\zeta_i(\theta,\vartheta) \quad \zeta_i(\theta_{\vartheta},\theta) = \lambda_i((\theta \quad \theta_{\vartheta})^{\intercal}\phi(s_i,a_i))$ where $\lambda_i(x) = x(2y_i^{\vartheta} \quad (\theta + \theta_{\vartheta})^{\intercal}\phi(s_i,a_i))$ is $2(1+W)$-Lipschitz and $\lambda_i(0) = 0$, the contraction lemma (Lemma B.3) tells us that it suffices to consider a simpler class. Define

$$\omega_i(\theta,\vartheta) = (\theta \quad \theta_{\vartheta})^{\intercal}\phi(s_i,a_i)$$

and $\boldsymbol{\omega}(\theta,\vartheta) = (\omega_i(\theta,\vartheta))_{i=1}^N$. By Chaining (Lemma B.6), we have

$$\begin{aligned}&\mathsf{E}\left[\quad \left(\frac{1}{J}\sup j\boldsymbol{\sigma} \quad \boldsymbol{\omega}(\theta,\vartheta)j\right)\right] \quad 1\\ &\text{where } \ J = \inf_{\alpha \ 0}\left\{3\alpha\sqrt{N} + 9\int_{\alpha}^{b}\sqrt{\log(D(\delta/2,\boldsymbol{\omega}(\quad)))}\mathrm{d}\delta\right\},\\ &\boldsymbol{\omega}(\quad) = f\boldsymbol{\omega}(\theta,\vartheta) \ : \ \theta \ 2 \quad g,\end{aligned}$$

$D(\delta,F)$ is the Euclidean packing number of $F \quad \mathsf{R}^N$, and $b = \sup \quad k\boldsymbol{\omega}(\theta,\vartheta)k_2$ is the envelope.

Now we'll bound the truncated entropy integral, $J$. First notice that $b \quad 3t\sqrt{N}$, based on the definition of (Equation (8)), which localizes in $k \quad k_{\frown}$,

$$\frac{b}{\sqrt{N}} = \sup\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\theta \quad \theta_{\vartheta})^{\intercal}\phi(s_i,a_i)\phi(s_i,a_i)^{\intercal}(\theta \quad \theta_{\vartheta})} = \sup k\theta \quad \theta_{\vartheta}k_{\frown} \quad 3t.$$

Now, we bound the packing number $D(\ ,\boldsymbol{\omega}(\quad))$. Let $\theta_1,\theta_2 \ 2 \ B_W$ be arbitrary,

$$\frac{k\boldsymbol{\omega}(\theta_1,\vartheta) \quad \boldsymbol{\omega}(\theta_2,\vartheta)k_2}{\sqrt{N}} = k\theta_1 \quad \theta_2 k_{\frown} \quad \sqrt{\sigma_{max}(\widehat{\ })}k\theta_1 \quad \theta_2 k_2 \quad k\theta_1 \quad \theta_2 k_2.$$

So for any $\varepsilon$, we can construct an $\varepsilon$-cover by setting $k\theta_1 \quad \theta_2 k_2 \quad \varepsilon/\sqrt{N}$, which requires $(W(1+2\sqrt{N}/\varepsilon))^d$ points. Let $N(\varepsilon,F)$ denote the Euclidean covering number of $F \quad \mathsf{R}^N$. Then,

$$\log D(\varepsilon,\boldsymbol{\omega}(\quad)) \quad \log N(\varepsilon/2,\boldsymbol{\omega}(\quad)) \quad d\log W(1+4\sqrt{N}/\varepsilon)$$

So,

$$\begin{aligned}J \quad &\int_0^{3t\sqrt{N}}\sqrt{\log(D(\varepsilon/2,\boldsymbol{\omega}(\quad)))}\mathrm{d}\varepsilon\\ &\int_0^{3t\sqrt{N}}\sqrt{d\log(W(1+8\sqrt{N}/\varepsilon))}\mathrm{d}\varepsilon\\ &3t\sqrt{dN}(\sqrt{\log W} + \int_0^1\sqrt{\log(1+3/(\varepsilon t))})\mathrm{d}\varepsilon\\ &3t\sqrt{dN}(\sqrt{\log W} + \sqrt{\log(4/t)}),\end{aligned}$$

since $\int_0^1\sqrt{\log(1+c/\varepsilon)}\mathrm{d}\varepsilon \quad \sqrt{\log(1+c)}$ for any $c > 0$, and assuming $t \quad 1$.

Now we put everything together. Let $c$ denote a positive constant,

$$\mathsf{P}\left(\sup\left|\sum_{i=1}^{N}\zeta_i(\theta,\vartheta) \quad \zeta_i(\theta_{\vartheta},\vartheta) \quad \mathsf{E}_{\nu}[\zeta_i(\theta,\vartheta) \quad \zeta_i(\theta_{\vartheta},\vartheta)]\right| \quad Nt^2\right)$$

Since is increasing,

$$= \mathsf{P}\left( \left( c \sup \left| \sum_{i=1}^{N} \zeta_i(\theta, \vartheta) \quad \zeta_i(\theta_\vartheta, \vartheta) \quad \mathsf{E}_\nu [\zeta_i(\theta, \vartheta) \quad \zeta_i(\theta_\vartheta, \vartheta)] \right| \right) \qquad (cNt^2) \right)$$

By Markov's inequality,

$$\frac{\mathsf{E}\left[ \left( c \sup \left| \sum_{i=1}^{N} \zeta_i(\theta, \vartheta) \quad \zeta_i(\theta_\vartheta, \vartheta) \quad \mathsf{E}_\nu [\zeta_i(\theta, \vartheta) \quad \zeta_i(\theta_\vartheta, \vartheta)] \right| \right) \right]}{(cNt^2)}$$

By Symmetrization (Lemma B.2),

$$\frac{\mathsf{E}[\mathsf{E}[ \quad (2c \sup \quad j\sigma \quad (\zeta(\theta, \vartheta) \quad \zeta(\theta_\vartheta, \vartheta))j)]]}{(cNt^2)}$$

By Contraction (Lemma B.3) and that $\zeta_i(\theta, \vartheta) \quad \zeta_i(\theta_\vartheta, \vartheta) = \lambda_i(\omega_i(\theta, \vartheta))$ where $\lambda_i$ is $L = 2(1 + W)$ Lipschitz,

$$\frac{3}{2} \quad \frac{\mathsf{E}[\mathsf{E}[ \quad (4cL \sup \quad j\sigma \quad \omega(\theta, \vartheta)j)]]}{(cNt^2)}$$

Setting $c = \frac{1}{4LJ}$ and applying Chaining (Lemma B.6), the numerator is bounded by 1,

$$8 \exp\left( \left( \frac{Nt^2}{4LJ} \right)^2 \right)$$

Applying upper bound on $J$,

$$8 \exp\left( \left( \frac{Nt^2}{8(1 + W) \quad 3t \sqrt{dN}(\sqrt{\log W} + \sqrt{\log(4/t)})} \right)^2 \right)$$

$$8 \exp\left( \frac{Nt^2}{24^2(1 + W)^2 d(\log W + \log(4/t))} \right)$$

Now, we set $t = 24(1 + W)\sqrt{\frac{d \log(N) \log(1/\delta)}{N}}$,

$$8 \exp\left( \frac{\log(N) \log(1/\delta)}{\log(N/d \log(N))} \right)$$

Since $\log(N) \quad \log(N/d \log(N))$,

$$8\delta.$$

Hence, we have shown that for an arbitrary and fixed $\vartheta \in B_W$, w.p. $1 \quad 9\delta$, we have

$$k\widehat{\theta}_\vartheta \quad \theta_\vartheta k \quad < t = 24(1 + W)\sqrt{\frac{d \log(N) \log(1/\delta)}{N}}.$$

We finally apply a union bound over $\vartheta \in B_W$. Consider a $W/N$-cover of $\vartheta \in B_W$, which requires $(1 + 2N)^d$ points. Then, for any $\vartheta \in B_W$, we have

$$\begin{aligned}
k\widehat{\theta}_\vartheta \quad \theta_\vartheta k \qquad & k\widehat{\theta}_\vartheta \quad \widehat{\theta}_{n(\vartheta)} k \quad + k\widehat{\theta}_{n(\vartheta)} \quad \theta_{n(\vartheta)} k \quad + k\theta_{n(\vartheta)} \quad \theta_\vartheta k \\
& (2k\widehat{\theta}_\vartheta \quad \widehat{\theta}_{n(\vartheta)} k \frown + t) + t + \sqrt{2\gamma W k n(\vartheta) \quad \vartheta k} \\
& 2t + 3\sqrt{2\gamma W k n(\vartheta) \quad \vartheta k} \\
& 2t + 3\sqrt{2\gamma W^2/N} \\
& 5t.
\end{aligned}$$

Thus, we have shown that w.p. 1 $\delta$,

$$\sup_{\vartheta 2 B_W} k\widehat{\theta}_\vartheta \quad \theta_\vartheta k \quad < 120d(1+W)\frac{\log(N)\sqrt{\log(10/\delta)}}{\overline{N}}.$$

□

A nice corollary is that when provides full coverage, i.e. $\beta I$ for some positive $\beta$, then we can bound

$$\sup_{\vartheta 2 B_W} k\widehat{\theta}_\vartheta \quad \theta_\vartheta k_2 \quad \sigma_{min}(\ )^{1/2} \sup_{\vartheta 2 B_W} k\widehat{\theta}_\vartheta \quad \theta_\vartheta k \quad \beta^{1/2} \sup_{\vartheta 2 B_W} k\widehat{\theta}_\vartheta \quad \theta_\vartheta k \ .$$

### C.1. Main LSPE theorem

We now prove our LSPE sample complexity guarantee Theorem 3.3.

**Theorem 3.3** (Sample Complexity of LSPE)**.** *Assume feature $\phi$ satisfies approximate Linear BC with parameter $\varepsilon_\nu$. For any $\delta$ 2 (0, 0.1), w.p. at least* 1 $\delta$*, we have for any initial state distribution $p_0$*

$$\left|V_{p_0}^{\pi_e} \quad \mathsf{E}_{s\ p_0}\left[\widehat{f}_K(s,\pi_e)\right]\right| \quad \frac{\gamma^{K/2}}{1\quad\gamma} + \frac{4\sqrt{\left\|\frac{\mathrm{d}d_{p_0}^{\pi_e}}{\mathrm{d}\nu}\right\|_1}}{(1\quad\gamma)^2}\varepsilon_\nu + \frac{480\sqrt{\kappa(p_0)}(1+W)d\log(N)\sqrt{\log(10/\delta)}}{(1\quad\gamma)^2\overline{N}},$$

*where $\widehat{f}_K$ is the output of Algorithm 1.*

*Proof of Theorem 3.3.* By Lemmas C.1 and C.2,

$$\left|V_{p_0}^{\pi} \quad \mathsf{E}_{s\ p_0}\left[\widehat{f}_k(s,\pi)\right]\right| \quad \frac{4}{(1\quad\gamma)^2}\max_{k=1,2,\ldots} k\widehat{f}_k \quad T^{\pi}\widehat{f}_k\ {}_1 k_{L_2(d_{p_0}^{\pi})} + \frac{\gamma^{k/2}}{1\quad\gamma}.$$

Next, we bound the maximum regression error. Consider any initial state distribution $p_0$, then we have

$$\begin{aligned}
\max_{k=1,2,\ldots} k\widehat{f}_k \quad T^{\pi}\widehat{f}_k\ {}_1 k_{L_2(d_{p_0}^{\pi})} \quad & \sup_{k\vartheta k\ W} k\widehat{\theta}_\vartheta^{\mid}\phi \quad T^{\pi}(\vartheta^{\mid}\phi)k_{L_2(d_{p_0}^{\pi})} \\
& \sup_{k\vartheta k\ W} k\widehat{\theta}_\vartheta^{\mid}\phi \quad \theta_\vartheta^{\mid}\phi k_{L_2(d_{p_0}^{\pi})} + \sup_{k\vartheta k\ W} k\theta_\vartheta^{\mid}\phi \quad T^{\pi}(\vartheta^{\mid}\phi)k_{L_2(d_{p_0}^{\pi})} \\
& \sqrt{\kappa(p_0)} \sup_{k\vartheta k\ W} k\widehat{\theta}_\vartheta \quad \theta_\vartheta k \ + \sqrt{\left\|\frac{\mathrm{d}d_{p_0}^{\pi}}{\mathrm{d}\nu}\right\|_1} \sup_{k\vartheta k\ W} k\theta_\vartheta^{\mid}\phi \quad T^{\pi}(\vartheta^{\mid}\phi)k_{L_2(\nu)} \\
& \sqrt{\kappa(p_0)} \sup_{\vartheta 2 B_W} k\widehat{\theta}_\vartheta \quad \theta_\vartheta k \ + \sqrt{\left\|\frac{\mathrm{d}d_{p_0}^{\pi}}{\mathrm{d}\nu}\right\|_1}\varepsilon_\nu.
\end{aligned}$$

The quantity $\sup_{\vartheta 2 B_W} k\widehat{\theta}_\vartheta \quad \theta_\vartheta k$ can be directly bounded by Lemma C.5 w.p. at least 1 $\delta$. Thus, we have shown the desired result: for any initial state distribution $p_0$,

$$\left|V_{p_0}^{\pi} \quad \mathsf{E}_{s,a\ p_0\ \pi}\left[\widehat{f}_k(s,a)\right]\right| \quad \frac{4}{(1\quad\gamma)^2}\left(\sqrt{\kappa(p_0)}120d(1+W)\frac{\log(N)\sqrt{\log(10/\delta)}}{\overline{N}} + \sqrt{\left\|\frac{\mathrm{d}d_{p_0}^{\pi}}{\mathrm{d}\nu}\right\|_1}\right)\varepsilon_\nu + \frac{\gamma^{k/2}}{1\quad\gamma}.$$

□

## D. Proofs for Linear BC Equivalence

**Proposition 4.2.** *Consider a feature $\phi$ with full rank covariance $(\phi)$. Given any $W > 0$, the feature $\phi$ being linear BC (under $B_W$) implies that there exist $(\rho, M)$ 2 $B_W$ $\mathsf{R}^{d \ d}$ with $\|M\|_2 \ \sqrt{1 \ \frac{\|\rho\|^2}{W^2}}$, and*

$$\mathsf{E}_\nu \left\| \begin{bmatrix} M \\ \rho^\top \end{bmatrix} \phi(s,a) \ \begin{bmatrix} \gamma \mathsf{E}_{s' \ P(s,a)} \phi(s', \pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 = 0. \tag{2}$$

*On the other hand, if there exists $(\rho, M)$ 2 $B_W$ $\mathsf{R}^{d \ d}$ with $\|M\|_2 < 1$ such that the above equality holds, then $\phi$ must satisfy exact linear BC with $W$ $\frac{\|\rho\|_2}{1 \ \|M\|_2}$.*

*Proof.* ( ( =) Suppose that $\|M\|_2 < 1$ and $\rho$ 2 $B_W$ satisfy,

$$\mathsf{E}_\nu \left\| \begin{bmatrix} M \\ \rho^\top \end{bmatrix} \phi(s,a) \ \begin{bmatrix} \gamma \mathsf{E}_{s' \ P(s,a)} \phi(s', \pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 = 0.$$

Then, for any $w_1$ 2 $B_W$, setting $w_2 = \rho + M^\top w_1$ satisfies,

$$\begin{aligned} \|w_2^\top \phi \ T^{\pi_e}(w_1^\top \phi)\|_\nu^2 &= \mathsf{E}_\nu \left\| w_2^\top \phi(s,a) \ r(s,a) \ \gamma \mathsf{E}_{s' \ P(s,a)} [w_1^\top \phi(s', \pi_e)] \right\|_2^2 \\ &= \mathsf{E}_\nu \left\| \rho^\top \phi(s,a) \ r(s,a) + w_1^\top M \phi(s,a) \ \gamma \mathsf{E}_{s' \ P(s,a)} [w_1^\top \phi(s', \pi_e)] \right\|_2^2 \\ &= \mathsf{E}_\nu \left\| w_1^\top \left( M \phi(s,a) \ \gamma \mathsf{E}_{s' \ P(s,a)} [\phi(s', \pi_e)] \right) \right\|_2^2 \\ &= 0. \end{aligned}$$

Also, we have $\|w_2\|_2 \ \|\rho\|_2 + \|M\|_2 \|w_1\|_2 \ W$ since $W \ \frac{\|\rho\|_2}{1 \ \|M\|_2}$. Thus, $\phi$ satisfies exact Linear BC.

(= ) ) Suppose $\phi$ satisfies exact Linear BC, that is

$$\max_{w_1 2 B_W} \min_{w_2 2 B_W} \|w_2^\top \phi \ T^\pi(w_1^\top \phi)\|_\nu^2 = 0.$$

To see that there exists $\rho$ 2 $B_W$ that linearizes the reward w.r.t $\phi$ under $\nu$, set $w_1 = 0$, and we have:

$$\min_{w_2 2 B_W} \mathsf{E}_{s,a \ \nu} \left\| w_2^\top \phi(s,a) \ r(s,a) \right\|_2^2 = 0.$$

Let $\rho$ to be the minimizer of the above objective.

Now we need to show that there exists a $M$ 2 $\mathsf{R}^{d \ d}$ with $\|M\|_2 < \sqrt{1 \ \frac{\|\rho\|^2}{W^2}}$ that satisfies

$$\mathsf{E}_{s,a \ \nu} \left\| M \phi(s,a) \ \gamma \mathsf{E}_{s' \ P(s,a)} \phi(s', \pi_e) \right\|_2^2 = 0.$$

To extract the $i$-th row of $M$, plug in $w_i = We_i$ (note that $w_i$ 2 $B_W$). By exact Linear BC, we know that there exists a vector $v_i$ 2 $B_W$, such that:

$$\left\| v_i^\top \phi(s,a) \ \rho^\top \phi(s,a) \ \gamma W \mathsf{E}_{s' \ P(s,a)} e_i^\top \phi(s', \pi_e) \right\|_\nu = 0.$$

Repeating this for every $i$ 2 [$d$], we can construct $M$ as follows,

$$M = \frac{1}{W} \begin{bmatrix} (v_1 \ \rho)^\top \\ \vdots \\ (v_d \ \rho)^\top \end{bmatrix},$$

which satisfies,

$$\begin{aligned}
&\mathsf{E}_{s,a\ \ \nu}\left\|M\phi(s,a)\quad \gamma\mathsf{E}_{s'\ \ P(s,a)}\phi(s',\pi_e)\right\|_2^2\\
&=\sum_{i=1}^{d}\mathsf{E}_{s,a\ \ \nu}\left\|e_i^{\intercal}\left(M\phi(s,a)\quad \gamma\mathsf{E}_{s'\ \ P(s,a)}\phi(s',\pi_e)\right)\right\|_2^2\\
&=\sum_{i=1}^{d}\mathsf{E}_{s,a\ \ \nu}\left\|\frac{1}{W}(v_i\quad \rho)^{\intercal}\phi(s,a)\quad \gamma\mathsf{E}_{s'\ \ P(s,a)}e_i^{\intercal}\phi(s',\pi_e)\right\|_2^2=0.
\end{aligned}$$

Hence, we have $\left\|M\phi(s,a)\quad \gamma\mathsf{E}_{s'\ \ P(s,a)}\phi(s',\pi_e)\right\|_{\nu}^2=0$.

Finally, we must show that $\|M\|_2<\sqrt{1\quad\frac{k\rho k^2}{W^2}}$. First we show that $\|M\|_2\quad 1$. For any $w_1\ 2\ B_W$, by exact linear BC, there exists $w_2\ 2\ B_W$ s.t. $\left\|w_2^{\intercal}\phi(s,a)\quad r(s,a)\quad \gamma\mathsf{E}_{s'\ \ P(s,a)}\left[w_1^{\intercal}\phi(s',\pi_e)\right]\right\|_{\nu}^2=0$, and by the construction of $M$, satisfies $\|(w_2\quad\rho\quad M^{\intercal}w_1)^{\intercal}\phi(s,a)\|_{\nu}^2=\|w_2\quad\rho\quad M^{\intercal}w_1\|_{(\phi)}^2=0$. Since $(\phi)$ is positive definite, we have that $w_2=\rho+M^{\intercal}w_1$ is the unique choice of $w_2$, which by exact linear BC is in $B_W$. Hence, we've shown that for any $w_1\ 2\ B_W$, we also have that $\rho+M^{\intercal}w_1\ 2\ B_W$. Now take $w_1$ and $\quad w_1$, subtracting the two expressions yields that $2M^{\star\intercal}w_1\ 2\ B_{2W}$. Since this is true for arbitrary $w_1$, taking supremum over $w_1\ 2\ B_W$ shows that $\|M\|_2\quad 1$.

Now we show that the inequality must be strict. Consider the singular value decomposition: $M=\sum_{i=1}^{d}\sigma_i u_i v_i^{\intercal}$ where $\{u_i\}_{i2[d]}$ and $\{v_i\}_{i2[d]}$ are each an orthonormal basis of $\mathsf{R}^d$, and $\sigma_i$ is the $i$-th largest singular value. Without loss of generality, suppose $\rho^{\intercal}u_1\quad 0$, since we can always flip the sign of $u_1$. If we pick $x=Wv_1\ 2\ B_W$, we have $Mx=W\sigma_1u_1$. By the argument in the previous paragraph, since $x\ 2\ B_W$, we have $\rho+Mx\ 2\ B_W$, implying that

$$W^2\quad\|Mx+\rho\|_2^2=\|W\sigma_1u_1+(\rho^{\intercal}u_1)u_1+(\rho\quad(\rho^{\intercal}u_1)u_1)\|_2^2$$

Since $u_1$ and $(\rho\quad(\rho^{\intercal}u_1)u_1)$ are orthogonal, by Pythagoras, we have

$$\begin{aligned}
&=\|W\sigma_1+\rho^{\intercal}u_1\|^2+\|(\rho\quad(\rho^{\intercal}u_1)u_1)\|_2^2\\
&=(W\sigma_1)^2+2W\sigma_1\rho^{\intercal}u_1+(\rho^{\intercal}u_1)^2+\|(\rho\quad(\rho^{\intercal}u_1)u_1)\|_2^2\\
&=(W\sigma_1)^2+2W\sigma_1\rho^{\intercal}u_1+\|(\rho^{\intercal}u_1)u_1\|_2^2+\|(\rho\quad(\rho^{\intercal}u_1)u_1)\|_2^2
\end{aligned}$$

By Pythagoras,

$$=(W\sigma_1)^2+2W\sigma_1\rho^{\intercal}u_1+\|\rho\|_2^2.$$

Hence, we get the following inequality:

$$W^2\sigma_1^2+2W(\rho^{\intercal}u_1)\sigma_1+\|\rho\|^2\|_2\quad W^2\quad 0.$$

Solve for $\sigma_1$ and using the fact that $\rho^{\intercal}u_1\quad 0$, we have that,

$$\sigma_1\quad\frac{\rho^{\intercal}u_1+\sqrt{(\rho^{\intercal}u_1)^2+(W^2\quad\|\rho\|^2)}}{W}\quad\frac{\sqrt{W^2\quad\|\rho\|^2}}{W}\quad\sqrt{1\quad\frac{\|\rho\|^2}{W^2}}.$$

We finally show that $\|\rho\|_2<W$ unless $M=0$. We prove this by contradiction. Assume $\|\rho\|_2\quad W$. Following the above argument, take any $w_1\ 2\ B_W$, we must have $w_2:=\rho+M^{\intercal}w_1\ 2\ B_W$. We discuss two cases.

First if $\rho\ \not2$ range$(M^{\intercal})$. In this case, we must have $\|w_2\|_2^2=\|\rho\|_2^2+\|M^{\intercal}w_1\|_2^2=W^2+\|M^{\intercal}w_1\|_2^2>W^2$, as long as $M$ has non-zero entries. Thus, this case leads to contradiction .

Second, if $\rho\ 2$ range$(M^{\intercal})$. In this case, there must exist a vector $x\ \not6\ 0$ such that $M^{\intercal}x=\rho$. Consider $x:=W\frac{x}{\|x\|\|_2}\ 2\ B_W$. We have $w_2:=\rho+M^{\intercal}x=\rho\left(1+\frac{W}{\|x\|\|_2}\right)$, which means that $\|w_2\|_2>W$, which causes contradiction again.

So unless $M=0$, which only happens when $\gamma=0$ (i.e. horizon is 1), we have $\|\rho\|_2<W$. □

We now show an approximate version to the equivalence of Proposition 4.2. First, recall the bilevel loss from Equation (3). We use $L_{lbc}$ and $\widehat{L}_{lbc}$ to denote the population and empirical versions as follows,

$$L_{lbc}(\phi) = \min_{(\rho,M)\,2} \quad \mathsf{E}_{\nu} \left\| \begin{bmatrix} M \\ \rho^{\mid} \end{bmatrix} \phi(s,a) \quad \begin{bmatrix} \gamma \mathsf{E}_{s^0 \quad P(s,a)} [\phi(s^0, \pi_e)] \\ r(s,a) \end{bmatrix} \right\|_2^2 \tag{9}$$

$$\widehat{L}_{lbc}(\phi) = \min_{(\rho,M)\,2} \quad \mathsf{E}_{D} \left\| \begin{bmatrix} M \\ \rho^{\mid} \end{bmatrix} \phi(s,a) \quad \begin{bmatrix} \gamma \phi(s^0, \pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 \quad \min_{g 2 G} \mathsf{E}_{D} \, k g(s,a) \quad \gamma \phi(s^0, \pi_e) k_2^2 , \tag{10}$$

where $= \left\{ (\rho, M)\ 2\ B_W \quad \mathsf{R}^{d \quad d} : k\rho k \quad k\rho^{\star} k, k M k_2 \quad k M^{\star} k_2 \right\}$ and $(\rho^{\star}, M^{\star})$ are corresponding to the linear BC $\phi^{\star}$.

**Lemma D.1.** *Suppose a feature $\widehat{\phi}$ satisfies $L_{lbc}(\widehat{\phi}) \quad \varepsilon^2$. Then, $\widehat{\phi}$ is $\varepsilon(1 + W)$-approximately Linear BC, provided $W \quad \frac{k\rho^{\star} k_2}{1 \quad k M^{\star} k_2}$.*

*Proof.* Suppose $L_{lbc}(\widehat{\phi}) \quad \varepsilon^2$, so there exists $\widehat{M}$ (s.t. $k\widehat{M} k_2 \quad k M^{\star} k_2 < 1$) and $\widehat{\rho}\ 2\ B_W$ s.t.

$$\mathsf{E}_{s,a \quad \nu} \left\| \begin{bmatrix} \widehat{M} \\ \widehat{\rho}^{\mid} \end{bmatrix} \phi(s,a) \quad \begin{bmatrix} \gamma \mathsf{E}_{s^0 \quad P(s,a)} \phi(s^0, \pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 \quad \varepsilon^2$$

For any $w_1\ 2\ B_W$, we can take $w_2 = \widehat{\rho} + \widehat{M}^{\mid} w_1$. Then, $k w_2 k_2 \quad k\widehat{\rho} k_2 + k\widehat{M} k_2 W \quad k\rho^{\star} k_2 + k M^{\star} k_2 W \quad W$ by our assumption on $W$. Hence,

$$\begin{aligned}
&\max_{w_1 2 B_W} \min_{w_2 2 B_W} \left\| w_2^{\mid} \phi(s,a) \quad r(s,a) \quad \gamma \mathsf{E}_{s^0 \quad P(s,a)} [w_1^{\mid} \phi(s^0, \pi_e)] \right\|_{\nu} \\
&\quad \max_{w_1 2 B_W} \left\| (\widehat{\rho} + \widehat{M}^{\mid} w_1)^{\mid} \phi(s,a) \quad r(s,a) \quad \gamma \mathsf{E}_{s^0 \quad P(s,a)} [w_1^{\mid} \phi(s^0, \pi_e)] \right\|_{\nu} \\
&= \max_{w_1 2 B_W} \sqrt{\mathsf{E}_{s,a \quad \nu} \left[ \left( (\widehat{\rho} + \widehat{M}^{\mid} w_1)^{\mid} \phi(s,a) \quad r(s,a) \quad \gamma \mathsf{E}_{s^0 \quad P(s,a)} [w_1^{\mid} \phi(s^0, \pi_e)] \right)^2 \right]} \\
&\quad \max_{w_1 2 B_W} \sqrt{\mathsf{E}_{s,a \quad \nu} \left[ (\widehat{\rho}^{\mid} \phi(s,a) \quad r(s,a))^2 \right]} + \sqrt{\mathsf{E}_{s,a \quad \nu} \left[ \left( w_1^{\mid} (\widehat{M} \phi(s,a) \quad \gamma \mathsf{E}_{s^0 \quad P(s,a)} [\phi(s^0, \pi_e)]) \right)^2 \right]} \\
&\quad \varepsilon(1 + W),
\end{aligned}$$

as desired.

□

## E. Proofs for Representation Learning

To simplify analysis, assume that the functions in $G$ have bounded $\ell_2$ norm, i.e. $8g\ 2\ G, s, a\ 2\ S \quad A, k g(s,a) k_2 \quad \gamma$. This is reasonable, and can always be achieved by clipping without loss of accuracy, since the target for $g(s,a)$ is $\gamma \mathsf{E}_{s^0 \quad P(s,a)} [\phi(s^0, \pi_e)]$ and $k\phi(s,a) k_2 \quad 1$ for any $s, a\ 2\ S \quad A$.

### E.1. Lemmas

Recall that $\lambda_k(A)$ denotes the $k$-th largest eigenvalue of a matrix $A$, i.e. $\lambda_1(A), \lambda_n(A)$ give the largest and smallest eigenvalues respectively.

**Lemma E.1** (Weyl's Perturbation Theorem)**.** *Let $A, B\ 2\ \mathsf{C}^{n \quad n}$ be Hermitian matrices. Then*

$$\max_k j \lambda_k(A) \quad \lambda_k(B) j \quad k A \quad B k_2 .$$

*Proof.* Please see (Bhatia, 2013, Corollary III.2.6). □

We extend this to be uniform over all .

**Lemma E.2** (Uniform spectrum concentration)**.** *For any $\delta \in (0,1)$, w.p. $1-\delta$,*

$$\sup_{\phi \in \ ,k \in [d]} \left|\lambda_k( \ (\phi)) - \lambda_k(\widehat{\ }(\phi))\right| \quad \frac{1}{\overline{N}}\left(96\kappa( \ ) + 4d + 4\log^{1/2}(1/\delta)\right)$$

*Proof.* First observe that,

$$\begin{aligned}\sup_{\phi \in } \sup_{k} \left|\lambda_k( \ (\phi)) - \lambda_k(\widehat{\ }(\phi))\right| \quad & \sup_{\phi \in } \left\| \ (\phi) \quad \widehat{\ }(\phi)\right\|_2 \\ &= \sup_{\phi \in \ ,kxk_2 \ 1} x^\top( \ (\phi) \quad \widehat{\ }(\phi))x \\ &= \sup_{\phi \in \ ,kxk_2 \ 1} (\mathsf{E}_\nu \quad \mathsf{E}_D)(x^\top \phi(s,a))^2\end{aligned}$$

Now we want to bound the Rademacher complexity of the class

$$F = \left\{(s,a) \mapsto (x^\top \phi(s,a))^2, \phi \in \quad ,kxk_2 \quad 1\right\}$$

First, to bound the envelope, we have $(x^\top \phi(s,a))^2 \quad 1$. To cover, consider any $\phi \in$ and $x \in \mathsf{R}^d$ s.t. $kxk_2 \quad 1$. Pick $\widetilde{\phi}, \widetilde{x}$ close to $\phi, x$, so that

$$\begin{aligned}&\sqrt{\frac{1}{N}\sum_{i=1}^N \left((x^\top \phi(s_i,a_i))^2 \quad (\widetilde{x}^\top \widetilde{\phi}(s_i,a_i))^2\right)^2} \\ &2\sqrt{\frac{1}{N}\sum_{i=1}^N \left((x^\top \phi(s_i,a_i) \quad \widetilde{x}^\top \widetilde{\phi}(s_i,a_i))\right)^2} \\ &2\left(\sqrt{\frac{1}{N}\sum_{i=1}^N \left(x^\top(\phi(s_i,a_i) \quad \widetilde{\phi}(s_i,a_i))\right)^2} + \sqrt{\frac{1}{N}\sum_{i=1}^N \left((x \quad \widetilde{x})^\top \widetilde{\phi}(s_i,a_i)\right)^2}\right) \\ &2\left(d \ (\phi,\widetilde{\phi}) + kx \quad \widetilde{x}k_2\right)\end{aligned}$$

So it suffices to take $d \ (\phi,\widetilde{\phi}), kx \quad \widetilde{x}k_2 \quad t/4$ to $t$-cover $F$ in $L_2(D)$. Note the $t/4$-covering number for $x$ in the unit ball is $(1 + 8/t)^d$. Thus, by Dudley's entropy bound ((5.48) of (Wainwright, 2019)),

$$\begin{aligned}R_N(F) \quad & \frac{24}{\overline{N}} \int_0^1 \log^{1/2}(N(t/4, \ ) \ (1 + 8/t)^d)\mathrm{d}t \\ & \frac{96}{\overline{N}}\left(\kappa( \ ) + 4\sqrt{d}\right)\end{aligned}$$

Thus, by Theorem 4.10 of (Wainwright, 2019), w.p. $1 \quad \delta$,

$$\begin{aligned}\sup_{\phi \in \ ,kxk_2 \ 1} (\mathsf{E}_\nu \quad \mathsf{E}_D)(x^\top \phi(s,a))^2 \quad & 2R_N(F) + \frac{4\log^{1/2}(1/\delta)}{\overline{N}} \\ & \frac{1}{\overline{N}}\left(96\kappa( \ ) + 4\sqrt{d} + 4\log^{1/2}(1/\delta)\right)\end{aligned}$$

□

We now prove the double sampling lemma, i.e. modified Bellman Residual Minimization (Chen & Jiang, 2019). This will help deal with the double sampling issue when transitions are stochastic. Recall that $G$ is a function class of functions $g : X \mapsto \mathsf{R}^d$. Let $\nu$ be a distribution over $X \quad \mathsf{R}^d$ and, for any $x \in X$, let $P(x)$ be a distribution over $Y \quad \mathsf{R}^d$.

**Lemma E.3** (Double Sampling)**.** *Suppose $x \perp\!\!\!\perp \mathsf{E}_{y \sim P(x)}[y]$ 2 $G$. Then,*

$$\mathsf{E}_{x \sim \nu}\left[\left\|x - \mathsf{E}_{y \sim P(x)}[y]\right\|_2^2\right] = \mathsf{E}_{x \sim \nu, y \sim P(x)}\left[\|x - y\|_2^2\right] - \inf_{g \in G} \mathsf{E}_{x \sim \nu, y \sim P(x)}\left[\|g(x) - y\|_2^2\right]$$

*Proof.*

$$\begin{aligned}
&\mathsf{E}_{x \sim \nu, y \sim P(x)}\left[\|x - y\|_2^2\right] - \mathsf{E}_{x \sim \nu}\left[\left\|x - \mathsf{E}_{y \sim P(x)}[y]\right\|_2^2\right] \\
&= \mathsf{E}_{x \sim \nu}\left[\mathsf{E}_{y \sim P(x)}\left[\|x\|_2^2 - 2\langle x, y\rangle + \|y\|_2^2\right] - \left(\|x\|_2^2 - 2\left\langle x, \mathsf{E}_{y \sim P(x)}[y]\right\rangle + \left\|\mathsf{E}_{y \sim P(x)}[y]\right\|_2^2\right)\right] \\
&= \mathsf{E}_{x \sim \nu}\left[\mathsf{E}_{y \sim P(x)}\left[\|y\|_2^2\right] - \left\|\mathsf{E}_{y \sim P(x)}[y]\right\|_2^2\right] \\
&= \mathsf{E}_{x \sim \nu}\left[\mathsf{E}_{y \sim P(x)}\left[\left\|y - \mathsf{E}_{y \sim P(x)}[y]\right\|_2^2\right]\right]
\end{aligned}$$

where the last step uses the fact that $\left\|\mathsf{E}_{y \sim P(x)}[y]\right\|_2^2 = \mathsf{E}_{y \sim P(x)}\left[\left\langle y, \mathsf{E}_{y \sim P(x)}[y]\right\rangle\right]$, and completing the square. Now, observe that, assuming $g^\star(x) \perp\!\!\!\perp \mathsf{E}_{y \sim P(x)}[y]$ 2 $G$, we have that it is the minimizer of,

$$g^\star \in \arg\min_{g \in G} \mathsf{E}_{x \sim \nu, y \sim P(x)}\left[\|g(x) - y\|_2^2\right] \tag{11}$$

which completes the proof. □

### E.2. Concentration lemmas

For any $\phi$, define the optimal $\rho, M, g$ for our losses as follows:

$$\begin{aligned}
\rho_\phi &\in \arg\min_{(\rho, _) \in \mathcal{Z}} \mathsf{E}_\nu\left[(\rho^\top \phi(s,a) - r(s,a))^2\right] \\
M_\phi &\in \arg\min_{(_, M) \in \mathcal{Z}} \mathsf{E}_{\nu\ P}\left[\|M\phi(s,a) - \gamma\phi(s', \pi_e)\|_2^2\right] \\
g_\phi &\in \arg\min_{g \in G} \mathsf{E}_{\nu\ P}\left[\|g(s,a) - \gamma\phi(s', \pi)\|_2^2\right]
\end{aligned}$$

Similarly, define $\widehat{\rho}_\phi, \widehat{M}_\phi, \widehat{g}_\phi$ to be minimizers of the above losses when expectation is taken over the empirical distribution $D$, instead of the population distribution $\nu$. Observe that the unconstrained minimization yields a closed form solution for $g_\phi$ as $g_\phi(s,a) = \gamma \mathsf{E}_{s' \sim P(s,a)}[\phi(s', \pi)]$ – Assumption 4.3 posits that $G$ is rich enough to capture this.

The key property of our squared losses is that the second moment can be upper bounded by the expectation, which allows us to invoke the second part of the above Lemma B.1. We now combine this with covering to get uniform convergence results.

**Lemma E.4.** *For any $\delta \in (0,1)$, w.p. at least $1 - \delta$, for any $\phi \in \Phi, \rho \in B_W$ and $M \in \mathbb{R}^{d \times d}$ with $\|M\|_2 \le 1$, we have*

$$\begin{aligned}
&\left|\mathsf{E}_D\left[(\rho^\top \phi(s,a) - r(s,a))^2\right] - \mathsf{E}_\nu\left[(\rho^\top \phi(s,a) - r(s,a))^2\right]\right| \le \frac{1}{2}\mathsf{E}_\nu\left[(\rho^\top \phi(s,a) - r(s,a))^2\right] + \varepsilon_\rho, \\
&\left|\mathsf{E}_D\left[\|M\phi(s,a) - \gamma\phi(s', \pi)\|_2^2 - \|g_\phi(s,a) - \gamma\phi(s', \pi)\|_2^2\right] - \mathsf{E}_{\nu\ P}\left[\|M\phi(s,a) - \gamma\phi(s', \pi)\|_2^2 - \|g_\phi(s,a) - \gamma\phi(s', \pi)\|_2^2\right]\right| \\
&\quad \le \frac{1}{2}\mathsf{E}_{\nu\ P}\left[\|M\phi(s,a) - \gamma\phi(s', \pi)\|_2^2 - \|g_\phi(s,a) - \gamma\phi(s', \pi)\|_2^2\right] + \varepsilon_M,
\end{aligned}$$

*and, assuming realizability (Assumption 4.3), for every $g \in G$, we have*

$$\begin{aligned}
&\left|\mathsf{E}_D\left[\|g(s,a) - \gamma\phi^\star(s', \pi)\|_2^2 - \|g_{\phi^\star}(s,a) - \gamma\phi^\star(s', \pi)\|_2^2\right] - \mathsf{E}_\nu\left[\|g(s,a) - g_{\phi^\star}(s,a)\|_2^2\right]\right| \\
&\quad \le \frac{1}{2}\mathsf{E}_\nu\left[\|g(s,a) - g_{\phi^\star}(s,a)\|_2^2\right] + \varepsilon_g,
\end{aligned}$$

*where $\varepsilon_\rho, \varepsilon_M, \rho_g$ are defined below.*

**Finite function classes** Assuming and $G$ are finite, we have

$$\varepsilon_\rho \quad \frac{6d(1+W)^2 \log(4W j \quad j N/\delta)}{N}$$
$$\varepsilon_M \quad \frac{32d^2 \log(4 j \quad j N/\delta)}{N}$$
$$\varepsilon_g \quad \frac{20\gamma^2 \log(2 j G j /\delta)}{N}.$$

*Proof for* $\varepsilon_\rho$. For a fixed $\phi \; 2 \quad , \rho \; 2 \; B_W$, apply Lemma B.1 to $X_i = (\rho^| \phi(s_i, a_i) \quad r(s_i, a_i))^2$. The envelope is $j X_i j \quad (1+W)^2$ and the second moment is bounded $\mathsf{E}\left[X_i^2\right] \quad (1+W)^2 \mathsf{E}[X_i]$. So, the error from the lemma is $\frac{2(1+W)^2 \log(2/\delta)}{N}$. Now union bound over an $\varepsilon$-net of $B_W$. Since $\left|(\rho^| \phi(s,a) \quad r(s,a))^2 \quad (\widetilde{\rho}^| \phi(s,a) \quad r(s,a))^2\right| = j(\rho \quad \widetilde{\rho})^| \phi(s,a)((\rho+\widetilde{\rho})^| \phi(s,a) + 2r(s,a)) j \quad 2(1+W) k\widetilde{\rho} \quad \rho k_2$, we consider a $\frac{1}{N}$-net of $B_W$ which requires $(1+2WN)^d$ points. The error from this $\varepsilon$-net approximation is at most $\frac{4(1+W)}{N}$. Finally, union bound over . □

*Proof for* $\varepsilon_M$. For a fixed $\phi \; 2 \quad$ and $M \; 2 \; \mathsf{R}^{d \quad d}$ s.t. $k M k_2 < 1$, apply Lemma B.1 to $X_i = k a k_2^2 \quad k b k_2^2$ where $a = M\phi(s_i, a_i) \quad \gamma\phi(s_i^0, \pi), b = g_\phi(s_i, a_i) \quad \gamma\phi(s_i^0, \pi)$. The envelope is $j X_i j = j X_i j \quad k a k_2^2 + k b k_2^2 \quad (1+\gamma)^2 + (2\gamma)^2 \quad 8$. Further, observe that $k a k_2^2 \quad k b k_2^2 = \hbar a + b, a \quad b i \quad k a + b k_2 k a \quad b k_2$. So, the second moment is bounded $\mathsf{E}\left[X_i^2\right]$ $\mathsf{E}\left[k a + b k_2^2 k a \quad b k_2^2\right] \quad (1+3\gamma)^2 \mathsf{E}\left[k M\phi(s_i, a_i) \quad \gamma g_\phi(s_i, a_i) k_2^2\right] \quad 16\mathsf{E}[X_i]$, where we used Lemma E.3 to give us

$$\mathsf{E}\left[k M\phi(s_i, a_i) \quad \gamma g_\phi(s_i, a_i) k_2^2\right] = \mathsf{E}\left[k M\phi(s_i, a_i) \quad \gamma\phi(s_i^0, \pi) k_2^2 \quad k g_\phi(s_i, a_i) \quad \gamma\phi(s_i^0, \pi) k_2^2\right].$$

So, the error from the lemma is $\frac{24 \log(2/\delta)}{N}$. Now union bound over an $\varepsilon$-net of $\left\{M \; 2 \; \mathsf{R}^{d \quad d} : k M k_2 \quad 1\right\}$. Observe that

$$\left|\left(k M\phi(s,a) \quad \gamma\phi(s^0,\pi) k_2^2 \quad k g_\phi(s,a) \quad \gamma\phi(s^0,\pi) k_2^2\right) \quad \left(k\widetilde{M}\phi(s,a) \quad \gamma\phi(s^0,\pi) k_2^2 \quad k g_\phi(s,a) \quad \gamma\phi(s^0,\pi) k_2^2\right)\right|$$
$$\left\|(M \quad \widetilde{M})\phi(s_i,a_i)\right\|_2 \left\|M\phi(s_i,a_i) + \widetilde{M}\phi(s_i,a_i) \quad 2\gamma\phi(s_i^0,\pi)\right\|_2$$
$$\left\|(M \quad \widetilde{M})\right\|_2 \quad (2+2\gamma) \quad 4\left\|(M \quad \widetilde{M})\right\|_2.$$

Consider a $\frac{1}{N}$-net (under $k k_F$) for $f M \; 2 \; \mathsf{R}^{d \quad d} : k M k_F \quad \sqrt{d} g$, which requires $(1+2N\sqrt{d})^{d^2}$ points since it is like the $\ell_2$ for a $d^2$-dimensional vector. This is a $\frac{1}{N}$-net (under $k k_2$) for the subset $\left\{M \; 2 \; \mathsf{R}^{d \quad d} : k M k_2 \quad 1\right\}$ since $k M k_2 \quad k M k_F \quad \sqrt{d} k M k_2$. The error from this $\varepsilon$-net approximation is at most $\frac{8}{N}$. Finally, union bound over . □

*Proof for* $\varepsilon_g$. For a fixed $g \; 2 \; G$, apply Lemma B.1 to $X_i = k g(s_i, a_i) \quad \gamma\phi^\star(s_i^0, \pi) k_2^2 \quad k g_{\phi^\star}(s,a) \quad \gamma\phi^\star(s^0,\pi) k_2^2$, the excess regression loss. Under realizability Assumption 4.3, we have $\mathsf{E}[X_i] = k g \quad g_{\phi^\star} k_2^2$, since

$$\mathsf{E}[X_i] = \mathsf{E}\left[k g(s,a) \quad \gamma\phi^\star(s^0,\pi) k_2^2 \quad k g_{\phi^\star}(s,a) \quad \gamma\phi^\star(s^0,\pi) k_2^2\right]$$
$$= \mathsf{E}\left[k g(s,a) \quad g_{\phi^\star}(s,a) k_2^2 + 2\hbar g_{\phi^\star}(s,a) \quad \gamma\phi^\star(s^0,\pi), g(s,a) \quad g_{\phi^\star}(s,a) i\right]$$

By definition of $g_{\phi^\star}$, we have $\mathsf{E}_{s^0 \quad P(s,a)}[g_{\phi^\star}(s,a) \quad \gamma\phi^\star(s^0,\pi)] = 0$, so,

$$= \mathsf{E}\left[k g(s,a) \quad g_{\phi^\star}(s,a) k_2^2\right]$$

The envelope is $j X_i j \quad (2\gamma)^2$ and the second moment is bounded $\mathsf{E}\left[X_i^2\right]$ $\mathsf{E}\left[k g(s,a) + g_{\phi^\star}(s,a) \quad 2\gamma \mathsf{E}_{a^0 \quad \pi(s^0)}[\phi^\star(s^0,a^0)] k_2^2 k g(s,a) \quad g_{\phi^\star}(s,a) k_2^2\right] \quad (4\gamma)^2 \mathsf{E}[X_i]$. So the error term from the lemma is $\frac{20\gamma^2 \log(2/\delta)}{N}$. Now union bound over $G$. □

**Infinite function classes** When and $G$ are infinite, we need to assume some metric entropy conditions. Then in the final step of the finite-class proofs above, we union bound on a well-chosen $\varepsilon$-net and collect an additional approximation error which is on the order of $O(1/N)$.

**Assumption E.5.** For $F\ 2\ f\ ,Gg$, we assume there exists $p\ 2\ \mathsf{R}_{++}$ such that $N(t,F)\ .\ t^{-p}$, where the net is under the following distances,

$$
\begin{aligned}
d\ (\phi,\widetilde{\phi}) &= \mathsf{E}_D\left[\left\|\phi(s,a)\quad\widetilde{\phi}(s,a)\right\|_2\right] + \mathsf{E}_\nu\left[\left\|\phi(s,a)\quad\widetilde{\phi}(s,a)\right\|_2\right] \\
&\quad + \mathsf{E}_D\left[\left\|\phi(s',\pi)\quad\widetilde{\phi}(s',\pi)\right\|_2\right] + \mathsf{E}_{s,a\ \ D,s'\ \ P(s,a)}\left[\left\|\phi(s',\pi)\quad\widetilde{\phi}(s',\pi)\right\|_2\right] \\
&\quad + \mathsf{E}_{s,a\ \ \nu,s'\ \ P(s,a)}\left[\left\|\phi(s',\pi)\quad\widetilde{\phi}(s',\pi)\right\|_2\right] \\
d_G(g,\widetilde{g}) &= \sqrt{\mathsf{E}_D\left[kg(s_i,a_i)\quad\widetilde{g}(s_i,a_i)k_2^2\right]} + \sqrt{\mathsf{E}_\nu\left[kg(s,a)\quad\widetilde{g}(s,a)k_2^2\right]}
\end{aligned}
$$

Note that this assumption is automatically satisfied for any $p > 0$ by VC classes (van der Vaart & Wellner, 1996, Theorem 2.6.4). Under this assumption, we have

$$
\begin{aligned}
\varepsilon_\rho\ &.\ \frac{d(1+W)^2(1+p)\log(WN/\delta)}{N} \\
\varepsilon_M\ &.\ \frac{d^2(1+p)\log(N/\delta)}{N} \\
\varepsilon_g\ &.\ \frac{\gamma^2 p\log(N/\delta)}{N}.
\end{aligned}
$$

*Proof for* $\varepsilon_\rho$. From before, we showed for a fixed $\phi\ \ 2$ , we have $\varepsilon_\rho\ .\ \frac{d(1+W)^2\log(WN/\delta)}{N}$. Observe that $\left|(\rho^{\mathsf{I}}\phi(s,a)\quad r(s,a))^2\quad(\rho^{\mathsf{I}}\widetilde{\phi}(s,a)\quad r(s,a))^2\right| = \left|\rho^{\mathsf{I}}(\phi(s,a)\quad\widetilde{\phi}(s,a))(\rho^{\mathsf{I}}(\phi(s,a)+\widetilde{\phi}(s,a))+2r(s,a))\right|\quad 2W(1+W)k\phi(s,a)\quad\widetilde{\phi}(s,a)k_2$, and so the difference of the loss with $\phi$ and the loss with $\widetilde{\phi}$ is bounded by $2W(1+W)d\ (\phi,\widetilde{\phi})$. Now union bound over a $\frac{1}{N}$-net, which requires $O(N^p)$ points by Assumption E.5. The approximation error using the net is $\frac{2W(1+W)}{N}$. □

*Proof for* $\varepsilon_M$. From before, we showed for a fixed $\phi\ \ 2$ , we have $\varepsilon_M\ .\ \frac{d^2\log(N/\delta)}{N}$. Observe that

$$
\begin{aligned}
&\left|\left(kM\phi(s,a)\quad\gamma\phi(s',\pi)k_2^2\quad kg_\phi(s,a)\quad\gamma\phi(s',\pi)k_2^2\right)\quad\left(kM\widetilde{\phi}(s,a)\quad\gamma\widetilde{\phi}(s',\pi)k_2^2\quad kg_{\widetilde{\phi}}(s,a)\quad\gamma\widetilde{\phi}(s',\pi)k_2^2\right)\right| \\
&\quad kM(\phi(s,a)\quad\widetilde{\phi}(s,a))\quad\gamma(\phi(s',\pi)\quad\widetilde{\phi}(s',\pi))kkM(\phi(s,a)+\widetilde{\phi}(s,a))\quad\gamma(\phi(s',\pi)\quad\widetilde{\phi}(s',\pi))k \\
&+ kg_\phi(s,a)\quad g_{\widetilde{\phi}}(s,a)\quad\gamma(\phi(s',\pi)\quad\widetilde{\phi}(s',\pi))kkg_\phi(s,a)+g_{\widetilde{\phi}}(s,a)\quad\gamma(\phi(s',\pi)+\widetilde{\phi}(s',\pi))k \\
&\quad\left(k\phi(s,a)\quad\widetilde{\phi}(s,a)k+\gamma k\phi(s',\pi)\quad\widetilde{\phi}(s',\pi)k\right)\quad(2+2\gamma) \\
&+\left(kg_\phi(s,a)\quad g_{\widetilde{\phi}}(s,a)k+\gamma k\phi(s',\pi)\quad\widetilde{\phi}(s',\pi)k\right)\quad(2+2\gamma)
\end{aligned}
$$

Using closed form solution for $g_\phi$,

$$16d\ (\phi,\widetilde{\phi}).$$

Now union bound over a $\frac{1}{N}$-net, which requires $O(N^p)$ points by Assumption E.5. The approximation error using the net is at most $\frac{16}{N}$. □

*Proof for $\varepsilon_g$.* From before, we showed for a fixed $g \in \mathcal{G}$, we have $\varepsilon_g \lesssim \frac{\gamma^2 \log(1/\delta)}{N}$. Observe that

$$\begin{aligned}&\left|\|g(s,a) - \gamma\phi^\star(s',\pi)\|^2 - \|\widetilde{g}(s,a) - \gamma\phi^\star(s',\pi)\|^2\right| \\ &\quad \|g(s,a) - \widetilde{g}(s,a)\|\,\|g(s,a) + \widetilde{g}(s,a) - 2\gamma\phi^\star(s',\pi)\| \\ &\quad (2+2\gamma)d_{\mathcal{G}}(g,\widetilde{g}).\end{aligned}$$

Now union bound over a $\frac{1}{N}$-net, which requires $O(N^p)$ points by Assumption E.5. The approximation error using the net is at most $\frac{4}{N}$. □

### E.3. Main Results

**Lemma E.6.** *Suppose $\phi^\star \in \Phi$ is Linear BC. Suppose Assumption 4.3 if transitions are stochastic. Moreover, suppose $\phi^\star$ is feasible in the bilevel optimization (and so $\widehat{L}_{lbc}(\widehat{\phi}) \le \widehat{L}_{lbc}(\phi^\star)$). Then, w.p. $1-5\delta$,*

$$L_{lbc}(\widehat{\phi}) \lesssim \frac{24d(1+W)^2\log(4W|\Phi|N/\delta) + 128d^2\log(4|\Phi|N/\delta) + 40\gamma^2\log(2|\mathcal{G}|/\delta)}{N}$$

*Proof.*

$$\begin{aligned}&L_{lbc}(\widehat{\phi}) \\ &= \mathbb{E}_\nu\left[(\rho_{\widehat{\phi}}^\top\widehat{\phi}(s,a) - r(s,a))^2\right] + \mathbb{E}_{\nu\times P}\left[\|M_{\widehat{\phi}}\widehat{\phi}(s,a) - \gamma\widehat{\phi}(s',\pi)\|_2^2\right] - \mathbb{E}_{\nu\times P}\left[\|g_{\widehat{\phi}}(s,a) - \gamma\widehat{\phi}(s',\pi)\|_2^2\right]\end{aligned}$$

Since $\rho_{\widehat{\phi}}, M_{\widehat{\phi}}$ are minimizers under $\nu$,

$$\mathbb{E}_\nu\left[(\widehat{\rho}_{\widehat{\phi}}^\top\widehat{\phi}(s,a) - r(s,a))^2\right] + \mathbb{E}_{\nu\times P}\left[\|\widehat{M}_{\widehat{\phi}}\widehat{\phi}(s,a) - \gamma\widehat{\phi}(s',\pi)\|_2^2\right] - \mathbb{E}_{\nu\times P}\left[\|g_{\widehat{\phi}}(s,a) - \gamma\widehat{\phi}(s',\pi)\|_2^2\right]$$

By the $\rho, M$ parts of Lemma E.4,

$$2\mathbb{E}_D\left[(\widehat{\rho}_{\widehat{\phi}}^\top\widehat{\phi}(s,a) - r(s,a))^2\right] + 2\mathbb{E}_D\left[\|\widehat{M}_{\widehat{\phi}}\widehat{\phi}(s,a) - \gamma\widehat{\phi}(s',\pi)\|_2^2\right] - 2\mathbb{E}_D\left[\|g_{\widehat{\phi}}(s,a) - \gamma\widehat{\phi}(s',\pi)\|_2^2\right] + 2\varepsilon_\rho + 2\varepsilon_M$$

Since $\widehat{g}_{\widehat{\phi}}$ minimizes under $D$,

$$2\mathbb{E}_D\left[(\widehat{\rho}_{\widehat{\phi}}^\top\widehat{\phi}(s,a) - r(s,a))^2\right] + 2\mathbb{E}_D\left[\|\widehat{M}_{\widehat{\phi}}\widehat{\phi}(s,a) - \gamma\widehat{\phi}(s',\pi)\|_2^2\right] - 2\mathbb{E}_D\left[\|\widehat{g}_{\widehat{\phi}}(s,a) - \gamma\widehat{\phi}(s',\pi)\|_2^2\right] + 2\varepsilon_\rho + 2\varepsilon_M$$

By optimality of $\widehat{\phi}$ under $\widehat{L}_{lbc}$,

$$\begin{aligned}&2\mathbb{E}_D\left[(\widehat{\rho}_{\phi^\star}^\top\phi^\star(s,a) - r(s,a))^2\right] + 2\mathbb{E}_D\left[\|\widehat{M}_{\phi^\star}\phi^\star(s,a) - \gamma\phi^\star(s',\pi)\|_2^2\right] - 2\mathbb{E}_D\left[\|\widehat{g}_{\phi^\star}(s,a) - \gamma\phi^\star(s',\pi)\|_2^2\right] \\ &+ 2\varepsilon_\rho + 2\varepsilon_M\end{aligned}$$

By the $\mathcal{G}$ part of Lemma E.4, we have $\mathbb{E}_D\left[\|g_{\phi^\star}(s,a) - \phi^\star(s',\pi)\|_2^2 - \|\widehat{g}_{\phi^\star}(s,a) - \phi^\star(s',\pi)\|_2^2\right] \le \frac{1}{2}\mathbb{E}_\nu\left[\|g_{\phi^\star} - \widehat{g}_{\phi^\star}\|_2^2\right] + \varepsilon_g \le \varepsilon_g$. Then, using the optimality of $\widehat{\rho}$ and $\widehat{M}$ under $D$,

$$\begin{aligned}&2\mathbb{E}_D\left[(\rho_{\phi^\star}^\top\phi^\star(s,a) - r(s,a))^2\right] + 2\mathbb{E}_D\left[\|M_{\phi^\star}\phi^\star(s,a) - \phi^\star(s',\pi)\|_2^2\right] - 2\mathbb{E}_D\left[\|g_{\phi^\star}(s,a) - \phi^\star(s',\pi)\|_2^2\right] \\ &+ 2\varepsilon_\rho + 2\varepsilon_M + 2\varepsilon_g\end{aligned}$$

By the $\rho, M$ parts of Lemma E.4,

$$\begin{aligned}&3\mathbb{E}_\nu\left[(\rho_{\phi^\star}^\top\phi^\star(s,a) - r(s,a))^2\right] + 3\mathbb{E}_{\nu\times P}\left[\|M_{\phi^\star}\phi^\star(s,a) - \phi^\star(s',\pi)\|_2^2\right] - 3\mathbb{E}_\nu\left[\|g_{\phi^\star}(s,a) - \phi^\star(s',\pi)\|_2^2\right] \\ &+ 4\varepsilon_\rho + 4\varepsilon_M + 2\varepsilon_g\end{aligned}$$

By Assumption 4.3 and Lemma E.3,

$= 3L_{lbc}(\phi^\star) + 4\varepsilon_\rho + 4\varepsilon_M + 2\varepsilon_g$

By assumption that $\phi^\star$ is Linear BC and Proposition 4.2,

$= 4\varepsilon_\rho + 4\varepsilon_M + 2\varepsilon_g.$

□

We now prove Theorem 4.4, in the general stochastic case. The deterministic transitions case is subsumed by ignoring the minimization over $G$, i.e. setting the complexity term of $G$ to zero.

**Theorem 4.4.** *Assume Assumption 4.1 (and Assumption 4.3 if the system is stochastic). Let* $C_2 := \frac{96\log^{1/2}(|\mathcal{I}||\mathcal{J}|)+4\sqrt{d}+4\log^{1/2}(1/\delta)}{\beta/4}$. *If* $N \geq C_2^2$, *then for any* $\delta \in (0,1)$, *w.p. at least* $1-\delta$, *we have*

*1.* $\widehat{\phi}$ *satisfies* $\widehat{\varepsilon}$*-approximate Linear BC with*

$$\widehat{\varepsilon} \leq \frac{13d(1+W)^2\log^{1/2}(4W|\mathcal{I}||\mathcal{J}|N/\delta)}{\sqrt{N}} + \frac{7\gamma(1+W)\log^{1/2}(2|\mathcal{G}|/\delta)}{\sqrt{N}},$$

*2.* $\lambda_{min}(\mathbb{E}_{\mathcal{D}}[\phi(\widehat{\phi})]) \geq \beta/4.$

*If transitions are deterministic, treat* $\log(|\mathcal{G}|) = 0$.

*Proof of Theorem 4.4.* First, by our assumption that $\sqrt{N} \geq 4(96\kappa(\mathcal{D}) + 4\sqrt{d} + 4\log^{1/2}(1/\delta))/\beta$, Lemma E.2 implies that w.p. at least $1-\delta$, we have $\sup_{\phi\in\Phi}\left|\lambda_{min}(\mathbb{E}_{\mathcal{D}}[\phi(\phi)]) - \lambda_{min}(\widehat{\mathbb{E}}_{\mathcal{D}}[\phi(\phi)])\right| \leq \beta/4$. Under this high probability event, we have two important consequences:

1. $\phi^\star$ is feasible in Equation (6), since $\lambda_{min}(\widehat{\mathbb{E}}_{\mathcal{D}}[\phi(\phi^\star)]) \geq \lambda_{min}(\mathbb{E}_{\mathcal{D}}[\phi(\phi^\star)]) - \beta/4 \geq \beta(1-1/4) \geq \beta/2$. In particular, this means $\widehat{L}_{lbc}(\widehat{\phi}) \leq \widehat{L}_{lbc}(\phi^\star)$.
2. The covariance of $\widehat{\phi}$ has lower-bounded eigenvalues, since $\lambda_{min}(\mathbb{E}_{\mathcal{D}}[\phi(\widehat{\phi})]) \geq \lambda_{min}(\widehat{\mathbb{E}}_{\mathcal{D}}[\phi(\widehat{\phi})]) - \beta/4 \geq \beta(1/2-1/4) \geq \beta/4$.

Now, apply Lemma E.6 to bound $L_{lbc}(\widehat{\phi})$, so w.p. $1-5\delta$,

$$L_{lbc}(\widehat{\phi}) \leq \frac{24d(1+W)^2\log(4W|\mathcal{I}||\mathcal{J}|N/\delta) + 128d^2\log(4|\mathcal{I}||\mathcal{J}|N/\delta) + 40\gamma^2\log(2|\mathcal{G}|/\delta)}{N}$$

By Lemma D.1, we have that $\widehat{\phi}$ is $\widehat{\varepsilon}$-approximately Linear BC, with parameter

$$\widehat{\varepsilon} \leq (1+W)\sqrt{\frac{24d(1+W)^2\log(4W|\mathcal{I}||\mathcal{J}|N/\delta) + 128d^2\log(4|\mathcal{I}||\mathcal{J}|N/\delta) + 40\gamma^2\log(2|\mathcal{G}|/\delta)}{N}}$$
$$\frac{13d(1+W)^2\log^{1/2}(4W|\mathcal{I}||\mathcal{J}|N/\delta)}{\sqrt{N}} + \frac{7\gamma(1+W)\log^{1/2}(2|\mathcal{G}|/\delta)}{\sqrt{N}}$$

Finally, we remark that the $\widehat{W}$ for $\widehat{\phi}$ (in the approximately Linear BC case) is upper bounded by a polynomial in $W^\star$ in the assumed exact Linear BC of $\phi^\star$. Consider our assumption that $\phi^\star$ is exactly Linear BC with $W^\star = W$ (use $\star$ to highlight that it is the $W$ in the assumption, which we now show matches the $W$ in the result). Then, by Proposition 4.2, $\|M^\star\|_2 \leq \sqrt{1 + \frac{\|\rho^\star\|_2^2}{W^{\star 2}}}$. Hence, it suffices to minimize over this smaller ball for $\widehat{M}$, so that $\|\widehat{M}\|_2 \leq \|M^\star\|_2$. Now, take the smallest possible $W$ in Lemma D.1, so that

$$\widehat{W} = \frac{\|\rho^\star\|_2}{1 - \|M^\star\|_2} \leq \frac{\|\rho^\star\|_2}{1 - \sqrt{1 + \frac{\|\rho^\star\|_2^2}{W^{\star 2}}}} = \frac{\|\rho^\star\|_2 \left(1 + \sqrt{1 + \frac{\|\rho^\star\|_2^2}{W^{\star 2}}}\right)}{\frac{\|\rho^\star\|_2^2}{W^{\star 2}}} \leq \frac{2W^{\star 2}}{\|\rho^\star\|_2},$$

which is a polynomial in $W^\star$. □

Our end-to-end result is deduced by chaining our LSPE theorem and the above theorem together.

**Theorem 4.5.** *Under Assumption 4.1 (and Assumption 4.3 if $P(s,a)$ is stochastic). Let $C_2, \widehat{\varepsilon}$ be as defined in Theorem 4.4. If $N \geq C_2^2$, then we have for any $\delta \in (0, 1/2)$, w.p. at least $1 - 2\delta$, for all distributions $p_0$,*

$$\left|V_{p_0}^{\pi_e} - \mathbb{E}_{s \sim p_0}\left[\widehat{f}_K(s, \pi_e)\right]\right| \leq \frac{\gamma^{K/2}}{1-\gamma} + \frac{4\sqrt{\left\|\frac{\mathrm{d}d_{p_0}^{\pi_e}}{\mathrm{d}\nu}\right\|_1}}{(1-\gamma)^2}\widehat{\varepsilon} + \frac{960\beta^{1/2}(1+W)d\log(N)\sqrt{\log(10/\delta)}}{(1-\gamma)^2 N}.$$

*Proof of Theorem 4.5.* We first apply Theorem 4.4 to see that $\widehat{\phi}$ satisfies the two properties needed for LSPE. It is indeed approximately Linear BC, with $\widehat{\varepsilon}$ specified in the theorem, and also has coverage (i.e. $\lambda_{min}(\mathbb{E}_{\nu}[\widehat{\phi}\widehat{\phi}^\top]) \geq \beta/4$). Using these two facts, and on a separate independent dataset $D_2$ (needs to be a separate dataset since $\widehat{\phi}$ is data-dependent), we run LSPE and directly apply Theorem 3.3 for the result. □

## F. Implementation Details

Here we detail all environment specifications and hyperparameters used in the main text.

### F.1. Dataset Details

Using the publicly released implementation for DrQ-v2, we trained high quality target policies and saved checkpoints for offline behavior datasets. We refer the readers to Yarats et al. (2021) for exact hyperparameters.

### F.2. Environment Details

Following the standards used by DrQ-v2 (Yarats et al., 2021), all environments have a maximum horizon length of 500 timesteps. This is achieved by the behavior/target policy having an action repeat of 2 frames. Furthermore, each state is 3 stacked frames that are each $84 \times 84$ dimensional RGB images (thus $9 \times 84 \times 84$).

| Task | Target Performance | Behavior Performance |
|---|---|---|
| `Finger Turn Hard` | 927 | 226 (24%) |
| `Cheetah Run` | 758 | 192 (25%) |
| `Quadruped Walk` | 873 | 236 (27%) |
| `Humanoid Stand` | 827 | 277 (33%) |

*Table 1.* Performance for target and behavior policies used to collect evaluation and offline datasets respectively.

| Task | Action Space Dimension | Task Traits | Reward Type |
|---|---|---|---|
| `Finger Turn Hard` | 2 | turn | sparse |
| `Cheetah Run` | 6 | locomotion | dense |
| `Quadruped Walk` | 12 | locomotion | dense |
| `Humanoid Stand` | 21 | stand | dense |

*Table 2.* Task descriptions, action space dimension, and reward type for each tested environment.

### F.3. Representation Architecture and Hyperparameter Details

We adopt the same network architecture as DrQ-v2's critic, first introduced in SAC-AE (Yarats et al., 2019). More specifically, to process pixel input, we have a 5 layer ConvNet with 3 3 kernels and 32 channels with `ReLU` activations. The first convolutional layer has a stride of 2 while the rest has stride 1. The output is fed through a single fully connected layer normalized by LayerNorm. Finally, there is a `tanh` nonlinearity on the outputted 50 dimensional state-representation. The action is then concatenated with this output and fed into a 4-layer MLP all with `ReLU` activations.

| Hyperparameter | Value |
|---|---|
| Feature Dimension | 512 |
| Weight Initialization | orthogonal init. |
| Optimizer | Adam |
| Learning Rate | 1 10 $^{5}$ |
| Batch Size | 2048 |
| Training Epochs | 200 |
| $\tau$ (target) | 0.005 |
| $\lambda_{\text{Design}}$ | 5 10 $^{6}$ |

*Table 3.* Hyperparameters used for `BCRL`

### F.4. Benchmarks and Metrics

**Modifications to CURL:** Originally, CURL only does contrastive learning between the image states with data augmentation. For OPE, apply the same CURL objective to the state-action feature detailed in the previous section. Note we also train CURL with the same random cropping image augmentations presented by the authors. Finally, since we are not interleaving the representation learning with SAC, we do not have a Q prediction head.

**Modifications to SPR:** We use the same image encoder as our features for SPR. The main difference is in the architecture of the projection layers where we implement as 3-layer mlp with `ReLU` activations. Note that these are additional parameters that neither CURL nor `BCRL` require. Finally, similarly to CURL, we do not have an additional Q-prediction head.

**Spearman Ranking Correlation Metric:** This rank correlation measures the correlation between the ordinal rankings of the value estimates and the ground truth returns. As defined in Fu et al. (2021), we have for policies $1, 2, \ldots, N$, true returns $V_{1:N}$, and estimated returns $\hat{V}_{1:N}$:

$$\text{Ranking Correlation} = \frac{\text{Cov}\left(V_{1:N}, \hat{V}_{1:N}\right)}{\sigma\left(V_{1:N}\right)\sigma(\hat{V}_{1:N})}$$