# Robust Transfer Learning With Side Information

**Akram Awad**[1] **Shihab Ahmed**[1] **Yue Wang**[1,2] **George Atia**[1,2]

[1]Department of Electrical Engineering, University of Central Florida, Orlando, Florida, USA
[2]Department of Computer Science, University of Central Florida, Orlando, Florida, USA

## Abstract

Robust Markov Decision Processes (MDPs) address environmental shift through distributionally robust optimization (DRO) by finding an optimal worst-case policy within an uncertainty set of transition kernels. However, standard DRO approaches require enlarging the uncertainty set under large shifts, which leads to overly conservative and pessimistic policies. In this paper, we propose a framework for transfer under environment shift that derives a robust target-domain policy via *estimate-centered* uncertainty sets, constructed through constrained estimation that integrates limited target samples with side information about the source-target dynamics. The side information includes bounds on feature moments, distributional distances, and density ratios, yielding improved kernel estimates and tighter uncertainty sets. Error bounds and convergence results are established for both robust and non-robust value functions. Moreover, we provide a finite-sample guarantee on the learned robust policy and analyze the robust sub-optimality gap. Under mild low-dimensional structure on the transition model, the side information reduces this gap and improves sample efficiency. We assess the performance of our approach across OpenAI Gym environments and classic control problems, consistently demonstrating superior target-domain performance over state-of-the-art robust and non-robust baselines.

## 1 INTRODUCTION

Transfer reinforcement learning (RL) seeks to leverage knowledge from a source environment to accelerate and stabilize learning in a related target environment. Such a scheme is essential in real-world applications where collecting sufficient data in the target environment is costly, dangerous, or otherwise infeasible, and where policies must instead be transferred from simulation or a related operational setting.

However, the difference between the two environments, often referred to as the *sim-to-real gap* or *environmental mismatch*, arise from modeling errors in simulation, unmodeled disturbances, adversarial perturbations, or non-stationary conditions, and can result in severe performance degradation when directly deploying trained policies. This underscores the need for principled transfer mechanisms that reduce adaptation time and minimize unduly data collection in the target domain.

A popular approach for transfer under model mismatch is the framework of *robust MDPs* and *robust RL* Nilim and Ghaoui [2003], Iyengar [2005], Wiesemann et al. [2013], which constructs an uncertainty set of transition kernels centered on the source environment. By optimizing the worst-case performance, robust RL yields policies that guarantee a lower bound on return when the target kernel lies within the set, enhancing robustness when deployed to out-of-distribution environments. This enhancement makes robust RL an attractive approach to transfer. However, when the two environments are substantially different, the uncertainty set must be expanded to cover the target one, often leading to overly conservative solutions. Such pessimism can cause robust policies to underperform in the target domain.

To mitigate over-conservatism, alternative transfer methods such as multi-task learning, domain randomization, and model-free domain adaptation have been explored (see Sec. 5). Multi-task RL seeks to learn representations that generalize across related tasks, while domain randomization exposes agents to diverse simulated conditions to promote robustness. Model-free adaptation methods reweight or adjust source samples to approximate the target distribution. Yet, these methods often fail when the target domain diverges sharply from the training conditions, as they do not explicitly account for the structure of uncertainty in the

*Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence* (UAI 2026), PMLR 337:255–286.

transition dynamics.

In this work, we develop a framework for robust transfer that explicitly leverages *side information*, i.e., prior knowledge about the relationship between the source and target environments. Such information may take the form of distance constraints, density ratios, or moment conditions that capture statistical or structural similarities between domains. By integrating side information with limited target samples, we construct improved estimates of the target transition kernel, yielding policies that are closer to optimal for the target and less conservative than standard robust RL. In effect, side information reduces the sim-to-real gap by anchoring the uncertainty set around an estimated target model rather than the source, thereby decreasing the adaptation time or the data required to achieve reliable performance.

We consider both the *non-robust setting*, which learns the optimal policy under the estimated target model, and the *robust setting*, which guards against residual model uncertainty around this estimate. Our main contributions are:

(1) We develop a side-information–based framework for estimating target transition kernels and learning robust policies, integrating structural constraints into the estimation process.
(2) We derive error bounds and establish convergence results for robust and non-robust value functions, providing asymptotic guarantees in terms of total variation distance and quality of side information.
(3) We provide finite-sample guarantees on the robust optimality gap under mild assumptions on the transition kernel subspace, demonstrating that side information reduces suboptimality and improves sample efficiency.
(4) Our approach is validated through experiments on OpenAI Gym and classic control tasks, showing consistent improvements over state-of-the-art baselines in both robust and non-robust settings.

## 2 PRELIMINARIES AND PROBLEM SETUP

**Markov Decision Process (MDP).** An MDP is denoted by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathsf{P}, r, \gamma)$, whose main components are the state space $\mathcal{S}$, the Action space $\mathcal{A}$, the discount factor $\gamma \in (0,1)$, the reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the transition kernels $\mathsf{P} = \{\mathsf{P}^{s,a} \in \Delta(\mathcal{S}), a \in \mathcal{A}, s \in \mathcal{S}\}$[1], where $\mathsf{P}^{s,a}$ is the distribution over the state space $\mathcal{S}$ of the next state conditioned on state $s$ and action $a$. Hence, $\mathsf{P}^{s,a}(s')$ denotes the probability of transitioning to state $s'$ if the agent is in state $s$ and takes action $a$. At each time step $t \geq 0$, the environment transitions to a state $s_{t+1}$ according to the transition probability $\mathsf{P}^{s_t,a_t}(s_{t+1})$ and yields a reward $r(s_t, a_t)$ for the agent.

[1] $\Delta(\mathcal{S})$ denotes the probability simplex over $\mathcal{S}$.

A stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is a distribution over the set of actions $\mathcal{A}$ for a given state $s$, which determines the probability of selecting a given action at a certain state. The value function of a stationary policy $\pi$ at state $s$ is defined as the expected discounted cumulative reward if the agent starts from state $s$ and takes actions according to policy $\pi$, i.e.,

$$V_{\mathsf{P}}^{\pi}(s) = \mathbb{E}_{\pi,\mathsf{P}}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \big| S_0 = s\right]. \quad (1)$$

The goal of the agent is to learn a stationary policy to maximize the expected cumulative reward, i.e., $\pi^* = \arg\max_\pi V_{\mathsf{P}}^{\pi}$.

**Robust MDP.** A robust MDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where the transition kernel is not fixed but comes from an uncertainty set $\mathcal{P}$. The environment can transit to the next state according to an arbitrary transition kernel belonging to the uncertainty set. In this work, we consider the $(s,a)$-rectangular uncertainty set Nilim and Ghaoui [2003], Iyengar [2005], i.e., $\mathcal{P} = \bigotimes_{s,a} \mathcal{P}_s^a$, where $\mathcal{P}_s^a \subseteq \Delta(\mathcal{S})$ and $\bigotimes_{s,a}$ is the Cartesian product over all state–action pairs. The robust discounted value function of policy $\pi$ at state $s$ is defined as

$$V_{\mathcal{P}}^{\pi}(s) \triangleq \min_{\eta \in \bigotimes_{t\geq 0} \mathcal{P}} \mathbb{E}_{\pi,\eta}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | S_0 = s\right], \quad (2)$$

where $\eta = (\mathsf{P}_0, \mathsf{P}_1, \dots)$. This accounts for the worst-case performance over the uncertainty set of transition kernels. The goal is to optimize the worst-case performance by maximizing the robust value function, i.e.,

$$\pi^* = \arg\max_{\pi} V_{\mathcal{P}}^{\pi}. \quad (3)$$

Equation (3) finds the optimal policy that maximizes the worst-case value function over the uncertainty set of distributions $\mathcal{P}$.

**Problem setup.** We consider an environment shift scenario, where an agent is trained in a source environment and subsequently deployed in a related, but distinct, target environment. This setting is formalized using two MDPs: the source domain MDP $\mathcal{M}_{\mathrm{s}} = (\mathcal{S}, \mathcal{A}, \mathsf{P}_{\mathrm{s}}, r, \gamma)$ and the target domain MDP $\mathcal{M}_{\mathrm{t}} = (\mathcal{S}, \mathcal{A}, \mathsf{P}_{\mathrm{t}}, r, \gamma)$, which only differ in their transition dynamics, i.e., $\mathsf{P}_{\mathrm{s}} \neq \mathsf{P}_{\mathrm{t}}$. Both $\mathcal{S}$ and $\mathcal{A}$ are assumed to be finite sets of sizes $S$ and $A$, respectively. The agent is provided with a fixed offline dataset $\mathcal{D} = \{(s_i, a_i, s_i')\}_{i=1}^N$, consisting of $N$ transitions. Each tuple is generated according to some distribution $\mu$ over state-action pairs, i.e., $(s_i, a_i) \sim \mu$, and the next state $s_i'$ is sampled from the nominal target transition kernel, $s_i' \sim \mathsf{P}_{\mathrm{t}}^{s_i,a_i}$. In addition, we assume the availability of side information $\Phi(\mathsf{P}_{\mathrm{s}}, \mathsf{P}_{\mathrm{t}})$, which captures structural or statistical relationships between the source and target transition kernels, defined in Section 3.

Our objective is to learn policies for the target domain using limited offline samples from the target MDP together with side information, under (i) a non-robust setting and (ii) a robust setting that accounts for transition uncertainty.

## 3 MAIN APPROACH

A key challenge is that the target environment is observable only through a *limited offline sample*. A natural baseline is offline RL Uehara and Sun [2022], Wang et al. [2024c], but reliable performance typically requires abundant, high-quality data; with scarce samples, offline RL often yields suboptimal policies due to coverage gaps and extrapolation error Levine et al. [2020]. Robust MDPs address distribution shift by optimizing over an uncertainty set *centered at the source* transition kernel Wang et al. [2023a], Wiesemann et al. [2013], Tamar et al. [2014], Lim and Autef [2019], Xu and Mannor [2010]. However, when the source–target shift is large, the uncertainty set must be enlarged to include the target dynamics, inducing excessive pessimism and degraded performance in the target domain.

Our key idea is to transfer knowledge from source to target by estimating the target transition kernel using limited offline target data together with side information that encodes relationships between the domains (e.g., bounds on moments, distributional distances, or density ratios). We then optimize policies *around this estimated target kernel* rather than around the source. Intuitively, if the estimate is closer to the true target dynamics than the source is, uncertainty sets centered at the estimate require smaller radii to cover the target, reducing conservatism while retaining robustness. We study both (i) a *non-robust* regime that learns the optimal policy for the estimated target model, and (ii) a *robust* regime that guards against residual model uncertainty around that estimate.

**Uncertainty Set Construction**. For radius $R \geq 0$, we define the $(s,a)$-rectangular set centered at $\mathsf{P}$ as

$$\mathcal{P}(\mathsf{P}, R) \triangleq \bigotimes_{s,a} \mathbb{B}_{\mathrm{TV}}(\mathsf{P}^{s,a}, R)\,, \tag{4}$$

where $\mathbb{B}_{\mathrm{TV}}(p,R) = \{q \in \Delta(\mathcal{S}) : \|q-p\|_{\mathrm{TV}} \leq R\}$ denotes the TV ball of radius $R$ centered at $p$. Hence, $\mathcal{P}(\mathsf{P}, R)$ is the Cartesian product of per-$(s,a)$ TV-balls centered at $\mathsf{P}^{s,a}$. The non-robust regime is the special case $R = 0$.

**Model-based Pipeline.** Our approach is model-based and proceeds in three steps:

(1) **Estimate target dynamics.** Using limited target samples and side information, compute a constrained estimator $\widehat{\mathsf{P}} = \{\widehat{\mathsf{P}}^{s,a} \in \Delta(\mathcal{S}), a \in \mathcal{A}, s \in \mathcal{S}\}$ of the target kernel $\mathsf{P}_{\mathrm{t}}$ (Sec. 3.1).

(2) **Optimize a policy.**

$$\text{(Non-robust)} \quad \pi^\star \in \arg\max_\pi V^\pi_{\widehat{\mathsf{P}}},$$
$$\text{(Robust)} \quad \pi^\star \in \arg\max_\pi \min_{Q \in \mathcal{P}(\widehat{\mathsf{P}}, R')} V^\pi_Q.$$

where $\mathcal{P}(\widehat{\mathsf{P}}, R')$ is defined w.r.t $\widehat{\mathsf{P}}$.

(3) **Evaluate on target.** Assess $V^{\pi^\star}$ in the target domain (non-robust) or its worst-case value over a target-domain uncertainty set (robust).

When $\widehat{\mathsf{P}}$ is closer (in TV) to $\mathsf{P}_{\mathrm{t}}$ than the source $\mathsf{P}_{\mathrm{s}}$ is, the radius needed to cover the target is smaller, yielding tighter worst-case evaluations and less pessimistic policies while preserving robustness guarantees. Our theory (Sec. 4) quantifies this via bounds that scale with $\max_{(s,a)} \|\widehat{\mathsf{P}}^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a}\|_{\mathrm{TV}}$.

Throughout this paper, we denote the target- and source-centered uncertainty sets by $\mathcal{P}(\mathsf{P}_{\mathrm{t}}, R)$ and $\mathcal{P}(\mathsf{P}_{\mathrm{s}}, R)$, respectively, and omit the dependence of the uncertainty set on the center when clear, e.g., $\mathcal{P}_{\mathrm{t}}(R) = \mathcal{P}(\mathsf{P}_{\mathrm{t}}, R)$, $\mathcal{P}_{\mathrm{s}}(R) = \mathcal{P}(\mathsf{P}_{\mathrm{s}}, R)$, and $\widehat{\mathcal{P}}(R) = \mathcal{P}(\widehat{\mathsf{P}}, R)$.

Figure 1 overviews the setting. Panel 1a shows a large environment shift between $\mathsf{P}_{\mathrm{s}}$ and $\mathsf{P}_{\mathrm{t}}$. Panel 1b illustrates the over-conservative strategy that centers an uncertainty set $\mathcal{P}_{\mathrm{s}}(R_{\mathrm{s}})$ of radius $R_{\mathrm{s}} \geq R$ at $\mathsf{P}_{\mathrm{s}}$. Panel 1c depicts our method: we center $\widehat{\mathcal{P}}(R')$ at the estimate $\widehat{\mathsf{P}}$, requiring a smaller $R'$ and improving target performance. In the non-robust case, both training and evaluation use singleton sets ($R = R' = 0$).

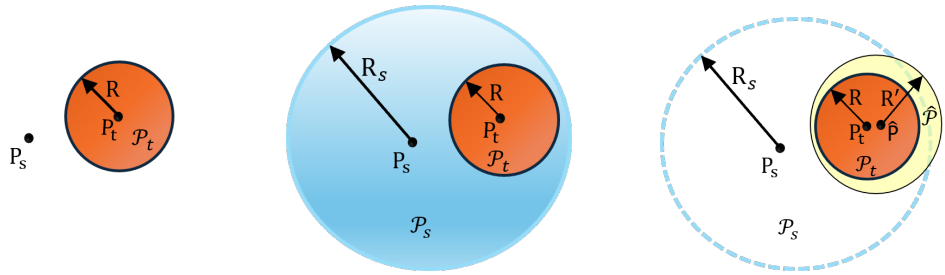
### 3.1 INFORMATION-BASED ESTIMATION

We propose an estimator for the target transition kernel that integrates limited offline target data with *side information* about the relationship between source and target dynamics. Given the dataset $\mathcal{D} = \{(s_i, a_i, s_i')\}_{i=1}^N$ from the target MDP, let $N(s,a) = \sum_{i=1}^N \mathbf{1}\{(s_i,a_i) = (s,a)\}$ and $N_{s,a}(s') = \sum_{i=1}^N \mathbf{1}\{(s_i,a_i,s_i') = (s,a,s')\}$, where $\mathbf{1}\{.\}$ denotes the indicator function. For each $(s,a)$, we estimate the next-state distribution $\widehat{\mathsf{P}}^{s,a} \in \Delta(\mathcal{S})$ as

$$\widehat{\mathsf{P}}^{s,a} \triangleq \begin{cases} \arg\max_{q \in \Delta(\mathcal{S})} \sum_{s' \in \mathcal{S}} N_{s,a}(s') \log q(s') & N(s,a) > 0, \\ \text{subject to} \quad \Phi\big(q, \mathsf{P}_{\mathrm{s}}^{s,a}\big) & \\ q_0^{s,a}, & N(s,a) = 0. \end{cases} \tag{5}$$

where $\Phi(\cdot, \mathsf{P}_{\mathrm{s}}^{s,a})$ encodes side information tying the target to the source, and $q_0^{s,a}$ is a prior-informed default when $(s,a)$ is unseen. In our experiments we set $q_0^{s,a} = \mathsf{P}_{\mathrm{s}}^{s,a}$ or uniform $q_0^{s,a} = \mathbf{1}/S$. We refer to (5) as the *Information-Based Estimator (IBE)*.

**Forms of Side Information.** We assume that, for each $(s,a)$, the source–target relationship satisfies $\Phi\big(\mathsf{P}_{\mathrm{t}}^{s,a}, \mathsf{P}_{\mathrm{s}}^{s,a}\big)$.




(a) (b) (c)

Figure 1: **(a) Environment shift: The source and target domain environments are relatively distant. (b) Over-conservative case: the source uncertainty set's radius is enlarged to include the target domain, which leads to an overly conservative policy. (c) Our approach**: We construct the uncertainty set around the estimated target dynamics, which are closer to the true target dynamics and therefore the set requires a smaller radius.

In addition to a no-side-information baseline (**Vanilla IBE**: unconstrained MLE on offline target transition counts), we instantiate $\Phi$ in four ways, yielding the variants summarized in Table 1. The constants $(d_{s,a}, \beta_{s,a}, B_{s,a})$ below may be per-$(s,a)$ or global.

(1) **Distance IBE.** Bound the discrepancy to the source $\mathrm{dist}\big(q, \mathsf{P}_{\mathrm{s}}^{s,a}\big) \le d_{s,a}$, where dist is either total variation $\|q - \mathsf{P}_{\mathrm{s}}^{s,a}\|_{\mathrm{TV}}$ or Wasserstein-1 $W_1(q, \mathsf{P}_{\mathrm{s}}^{s,a})$ with a specified ground cost (Appendix A), reflecting the assumption $\mathrm{dist}\big(\mathsf{P}_{\mathrm{t}}^{s,a}, \mathsf{P}_{\mathrm{s}}^{s,a}\big) \le d_{s,a}$. In robotics and control, such bounds arise naturally from known limits on physical parameter variations (e.g., friction coefficients, actuator gains, or payload mass estimated via calibration) and can be derived formally from Lipschitz continuity of the dynamics in the parameters (Appendix D). In finite state spaces, the radius $d_{s,a}$ can also be estimated from source and target samples via minimax-optimal divergence estimators Jiao et al. [2018], Nguyen et al. [2010].

(2) **Moment IBE.** Constrain feature moments. Let $\phi : \mathcal{S} \to \mathcal{H}$, mapping the state space to a Hilbert space, and $\mu(\mathsf{P}) = \mathbb{E}_{s' \sim \mathsf{P}}[\phi(s')]$. Impose $|\mu(q) - \mu(\mathsf{P}_{\mathrm{s}}^{s,a})| \le \beta_{s,a}$, matching the assumption $|\mu(\mathsf{P}_{\mathrm{t}}^{s,a}) - \mu(\mathsf{P}_{\mathrm{s}}^{s,a})| \le \beta_{s,a}$. This captures coarse aggregate information when full transition distributions are unavailable— for instance, known bounds on average velocity or energy dissipation in control systems—and is standard in empirical likelihood and conditional moment restriction frameworks Kremer et al. [2022].

(3) **Density IBE.** Assume absolute continuity and a bounded density ratio on the relevant support: $0 \le \frac{\mathsf{P}_{\mathrm{t}}^{s,a}(s')}{\mathsf{P}_{\mathrm{s}}^{s,a}(s')} \le B_{s,a}$ (whenever $\mathsf{P}_{\mathrm{s}}^{s,a}(s') > 0$), and enforce the element-wise constraint $0 \le q(s') \le B_{s,a}\, \mathsf{P}_{\mathrm{s}}^{s,a}(s')$, $\forall s' \in \mathcal{S}$. The cap $B_{s,a}$ encodes practical limits on importance reweighting under distribution shift, preventing excessive variance from extreme weights, and can be estimated from paired source–target samples via ratio estimation methods Kanamori et al. [2009], Nguyen et al. [2010]. If support mismatch is possible, we pre-smooth $\mathsf{P}_{\mathrm{s}}^{s,a}$ with small pseudocounts.

(4) **Low-Dimensional Structure (LDS) IBE.** For each $(s,a)$, let $\{\mathsf{P}_\theta^{s,a}\}_{\theta \in \Theta \subset \mathbb{R}^{\mathsf{d}}}$ be a parametric family (e.g., softmax). Write $\mathsf{P}_{\mathrm{s}}^{s,a} = \mathsf{P}_{\theta_{\mathrm{s}}}^{s,a}$ and $\mathsf{P}_{\mathrm{t}}^{s,a} = \mathsf{P}_{\theta_{\mathrm{t}}}^{s,a}$. Let $\mathcal{I}_{\mathrm{shared}} \subseteq \{1, \dots, \mathsf{d}\}$ be the coordinates shared by source and target, and define the affine subspace $\Theta_0 \triangleq \big\{\theta \in \Theta : \theta(\mathcal{I}_{\mathrm{shared}}) = \theta_{\mathrm{s}}(\mathcal{I}_{\mathrm{shared}})\big\}$, whose intrinsic dimension is $\mathsf{d}_0 = \mathsf{d} - |\mathcal{I}_{\mathrm{shared}}| \ll \mathsf{d}$. Such shared coordinates arise naturally when source and target differ in only a subset of physical parameters (for instance, when kinematics are preserved across domains but actuator gains or payload vary) so that the intrinsic degrees of freedom governing the shift are low-dimensional. Estimate the remaining $\mathsf{d}_0$ free coordinates by constrained MLE (CMLE) on offline target counts:

$$\widehat{\theta}_{\mathrm{t}} \in \arg\max_{\theta \in \Theta_0} \sum_{s' \in \mathcal{S}} N_{s,a}(s') \log \mathsf{P}_\theta^{s,a}(s'), \quad \widehat{\mathsf{P}}^{s,a} = \mathsf{P}_{\widehat{\theta}_{\mathrm{t}}}^{s,a}.$$

If $N(s,a) = 0$, use the uniform-initialized default $1/S$.

LDS-IBE explicitly adopts the softmax parameterization and enforces a low-dimensional structure on the transition kernels. As discussed later in Sec. 4.2, this structural prior facilitates finite-sample guarantees and a suboptimality-gap analysis.

When multiple forms of side information are available simultaneously, they can be combined as joint constraints on the feasible set, i.e., the intersection of the individual feasible sets induced by each $\Phi$, allowing all available domain knowledge to be incorporated at once.

We note that stronger priors (tighter $\Phi$) shrink the feasible set in (5), yielding more concentrated estimators and improved lower bounds on variance, which we formalize using constrained Cramér–Rao bounds in Appendix B.

Moreover, the side information $d_{s,a}$, $B_{s,a}$, $\beta_{s,a}$ need only be valid upper bounds on the true target-source discrepancy, rather than exact values. Tighter upper bounds yield a smaller feasible set and potentially a better estimate,

Table 1: Information-Based Estimation (IBE) under different side information $\Phi$.

| Method | Optimization Problem |
|---|---|
| **Distance IBE** | $\max_{q\in\Delta(\mathcal{S})} \sum_{s'\in\mathcal{S}} N_{s,a}(s')\log q(s')$ s.t. $\mathrm{dist}(q, \mathsf{P}_{\mathrm{s}}^{s,a}) \le d_{s,a}$ |
| **Density IBE** | $\max_{q\in\Delta(\mathcal{S})} \sum_{s'\in\mathcal{S}} N_{s,a}(s')\log q(s')$ s.t. $q/\mathsf{P}_{\mathrm{s}}^{s,a} \le B_{s,a}$ |
| **Moment IBE** | $\max_{q\in\Delta(\mathcal{S})} \sum_{s'} N_{s,a}(s')\log q(s')$ s.t. $\lvert\mu(q) - \mu(\mathsf{P}_{\mathrm{s}}^{s,a})\rvert \le \beta_{s,a}$. |
| **LDS IBE** | $\max_{\theta\in\mathbb{R}^{d_0}} \sum_{s'\in\mathcal{S}} N_{s,a}(s')\log \mathsf{P}_{\theta}^{s,a}(s')$ |

whereas looser bounds enlarge the feasible set and reduce the amount of side information incorporated. In the limiting case of uninformative bounds, the constrained estimator reduces to Vanilla IBE, so the method degrades gracefully while preserving consistency.

**Remark 1.** *In Appendix C, we also consider a* value-aware *side-information constraint based on a Wasserstein radius with the task-relevant pseudometric* $d_V(s,s') \triangleq |V_{\mathcal{P}}^{\pi}(s) - V_{\mathcal{P}}^{\pi}(s')|$. *Intuitively, it measures shift in the geometry that matters for control: moving probability between states of similar value is cheap, while moves across large value gaps are expensive. This is relevant in sim-to-real settings where small perturbations in robot position or sensor readings leave task value nearly unchanged.*

**Policy Estimation and Evaluation.** After obtaining $\widehat{\mathsf{P}}$, we compute policies by Value Iteration (VI) (see Algorithm 1) in both regimes using standard (robust) Bellman updates. In the non-robust setting, we apply VI with $\widehat{\mathsf{P}}$. In the robust setting, we use $(s,a)$-rectangular TV balls centered at the estimate, $\widehat{\mathcal{P}}(R)$, and perform robust VI with support-function evaluations over $\widehat{\mathcal{P}}(R)$. For evaluation, policies are tested in the target domain via Algorithm 2: we report target-domain value (non-robust) and worst-case value over target-centered uncertainty sets $\mathcal{P}_{\mathrm{t}}(R)$ (robust); in the special case $R=0$ both reduce to the non-robust evaluation.

# 4 THEORETICAL ANALYSIS

We provide finite-sample and asymptotic guarantees for the side-information–guided estimator and the induced (robust and non-robust) policies. We consider a discounted MDP with $\gamma \in (0,1)$ and rewards in $[0,1]$. For exposition, we assume balanced coverage: for each $(s,a)$, we observe $n$ i.i.d. target transitions $x_1^{s,a},\dots,x_n^{s,a} \sim \mathsf{P}_{\mathrm{t}}^{s,a}$ and denote by $\widehat{\mathsf{P}}_n^{s,a}$ its IBE estimate. Collecting these estimates over all (s,a) pairs yields $\widehat{\mathsf{P}}_n = \{\widehat{\mathsf{P}}_n^{s,a} \in \Delta(\mathcal{S}) : s\in\mathcal{S}, a\in\mathcal{A}\}$. Define the estimate-centered uncertainty set $\widehat{\mathcal{P}}_n(R) \triangleq \bigotimes_{(s,a)} \mathbb{B}_{\mathrm{TV}}(\widehat{\mathsf{P}}_n^{s,a}, R)$, making the dependence on $n$ explicit. The implementation does not require balanced coverage and instead uses empirical counts $N(s,a)$ from the offline target data.

For any policy $\pi$ and uncertainty set $\mathcal{P}(R)$, let $V_{\mathcal{P}}^{\pi} = \inf_{Q\in\mathcal{P}(R)} V_Q^{\pi}$ denote the robust value. Define the target-centered and estimate-centered robust-optimal policies by $\pi^\star \in \arg\max_\pi V_{\mathcal{P}_{\mathrm{t}}}^{\pi}$, $\pi_n \in \arg\max_\pi V_{\widehat{\mathcal{P}}_n}^{\pi}$, respectively. We also write $V_{\widehat{\mathcal{P}}_n}^{\pi_n}$ and $V_{\mathcal{P}_{\mathrm{t}}}^{\pi_n}$ for the worst-case values of $\pi_n$ evaluated on $\widehat{\mathcal{P}}_n(R)$ and $\mathcal{P}_{\mathrm{t}}(R)$, respectively. The non-robust setting corresponds to $R=0$, where $\mathcal{P}_{\mathrm{t}}(0)$ and $\widehat{\mathcal{P}}_n(0)$ reduce to singletons. Proofs of all results are deferred to Appendix F.

## 4.1 IBE FOR TRANSFER

We show that IBE-based policies are consistent for the target domain. Let $\delta_n \triangleq \max_{(s,a)} \|\widehat{\mathsf{P}}_n^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a}\|_{\mathrm{TV}}$.

**Training and Evaluation Errors.** We distinguish (i) the *training error*, which compares the estimated-policy value on the estimate-centered set to the target-optimal value; and (ii) the *evaluation error*, which evaluates the estimated policy on the *target-centered* set.

**Theorem 1** (Training error)**.** *For rewards in* $[0,1]$ *and any* $\gamma\in(0,1)$*,*

$$\left\|V_{\widehat{\mathcal{P}}_n}^{\pi_n} - V_{\mathcal{P}_t}^{\pi^\star}\right\|_\infty \le \frac{2\,\delta_n}{(1-\gamma)^2} \tag{6}$$

The training error scales linearly with the uniform TV error $\delta_n$. Hence, the closer $\widehat{\mathsf{P}}_n^{s,a}$ is to $\mathsf{P}_{\mathrm{t}}^{s,a}$ uniformly over $(s,a)$, the closer the training value of $\pi_n$ is to the target-optimal robust value. As the estimator improves (e.g., via side information), $\delta_n \downarrow$ and $V_{\widehat{\mathcal{P}}_n}^{\pi_n} \to V_{\mathcal{P}_{\mathrm{t}}}^{\pi^\star}$.

**Theorem 2** (Evaluation error)**.** *Under the same conditions,*

$$\left\|V_{\mathcal{P}_{\mathrm{t}}}^{\pi_n} - V_{\mathcal{P}_{\mathrm{t}}}^{\pi^\star}\right\|_\infty \le \frac{4\,\delta_n}{(1-\gamma)^2}\,. \tag{7}$$

This bounds the *deployment-time* gap when evaluating $\pi_n$ on the *target-centered* set. It is at most twice the training

bound: moving from the estimate-centered set (used to learn $\pi_n$) to the target-centered set costs a factor of at most 2. Consequently, as $\delta_n \to 0$ the evaluation error vanishes, guaranteeing asymptotic optimality on the target-centered robust problem. The dependence of $\delta_n$ on the form of side information $\Phi$ is discussed in Appendix E.

**Corollary 3** (Consistency)**.** *If $\delta_n \to 0$ (e.g., $\widehat{\mathsf{P}}_n^{s,a} \to \mathsf{P}_{\mathrm{t}}^{s,a}$ in TV), then $\|V_{\mathcal{P}_{\mathrm{t}}}^{\pi_n} - V_{\mathcal{P}_{\mathrm{t}}}^{\pi^\star}\|_\infty \to 0$, so the IBE-learned policy is asymptotically optimal for the target-centered robust problem.*

Thus, convergence of the robust (and non-robust) value functions follows if the transition-kernel estimates converge in total variation. The next result shows that this holds for IBE.

**Proposition 4** (TV-consistency of IBE)**.** *For fixed $(s,a)$, let $\widehat{\mathsf{P}}_n^{s,a}$ be the IBE solution of* (5) *with a side-information constraint set $\Phi(\cdot, \mathsf{P}_{\mathrm{s}}^{s,a})$ from Table 1. Assume $x_1^{s,a}, \ldots, x_n^{s,a} \overset{i.i.d.}{\sim} \mathsf{P}_{\mathrm{t}}^{s,a}$ and that the true target kernel lies in the* relative interior *of the feasible set. Then, $\left\|\widehat{\mathsf{P}}_n^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a}\right\|_{\mathrm{TV}} \overset{n\to\infty}{\longrightarrow} 0$. If this holds for every $(s,a)$ with per-pair sample sizes $n \to \infty$, then $\delta_n \to 0$.*

Proposition 4 shows that IBE is *TV-consistent*; the estimates converge to the true target kernel in total variation. TV-consistency is a sufficient condition for value-function consistency via Theorems 1–2: both the training and evaluation gaps scale as $O(\delta_n)$, with $\delta_n \to 0$. Thus, the IBE-learned policy becomes asymptotically optimal for the target-centered robust problem, and the non-robust case follows as the special instance $R = 0$. Proposition 4 extends classical MLE consistency to the *constrained* setting induced by side information (Table 1): TV/$\mathrm{W}_1$ balls, moment sets, density-ratio boxes, and LDS subspaces all yield closed, convex feasible regions on which the constrained MLE remains consistent. This validates our transfer pipeline of estimating target dynamics with side information and then optimizing around that estimate, and explains the empirical improvements we observe in the target domain. We corroborate TV-consistency and the predicted error decay in Appendix I.1, and we empirically verify the bound from Theorem 2 in Appendix I.4.

### 4.2 SUBOPTIMALITY GAP AND FINITE-SAMPLE GUARANTEES

We now leverage IBE's consistency to obtain *finite-sample* guarantees for the learned *robust* policy, focusing on the LDS–IBE case. Under a mild structural assumption on the transition family, we choose a radius $R_n$ so that the *estimate-centered* uncertainty set $\widehat{\mathcal{P}}_n(R_n)$ (as defined earlier) contains the true target kernel with high probability. This yields a high-confidence *out-of-sample* lower bound on target performance and, consequently, a bound on the robust suboptimality gap.

**Assumption 5** (Parametric family and TV–Lipschitzness)**.** There exists $\Theta \subset \mathbb{R}^{\mathsf{d}}$ and a mapping $\theta \mapsto \mathsf{P}_\theta$ such that for some $L > 0$,

$$\left\|\mathsf{P}_\theta^{s,a} - \mathsf{P}_{\theta'}^{s,a}\right\|_{\mathrm{TV}} \le L\,\|\theta - \theta'\|_1 \quad \forall\, \theta, \theta' \in \Theta,\ \forall (s,a),$$

and the target kernel satisfies $\mathsf{P}_{\mathrm{t}}^{s,a} = \mathsf{P}_{\theta_{\mathrm{t}}}$ for some $\theta_{\mathrm{t}}$ in the LDS subspace $\Theta_0$ of intrinsic dimension $\mathsf{d}_0 \ll \mathsf{d}$.

**Lemma 6** (Finite-sample radius for LDS–IBE)**.** *Under Assumption 5 (TV–Lipschitz in $\ell_1$), for any $\delta_1 \in (0,1)$ there exists $C_0 > 0$ such that choosing*

$$R_n \;=\; LC_0\sqrt{\frac{\mathsf{d}_0}{n}} \tag{8}$$

*ensures $\mathsf{P}_{\mathrm{t}} \in \widehat{\mathcal{P}}_n(R_n)$ with probability at least $1 - \delta_1$.*

By construction, this gives a high-probability *out-of-sample* lower bound: for any policy $\pi$, $V_{\mathsf{P}_{\mathrm{t}}}^{\pi} \ge V_{\widehat{\mathcal{P}}_n(R_n)}^{\pi}$ with probability $\ge 1 - \delta_1$. LDS shrinks the effective dimension from $\mathsf{d}$ to $\mathsf{d}_0$, tightening $R_n$.

**Robust Suboptimality Gap.** Let $\pi_n \in \arg\max_\pi V_{\widehat{\mathcal{P}}_n}^{\pi}$ and $\pi^\star \in \arg\max_\pi V_{\mathcal{P}_{\mathrm{t}}}^{\pi}$. Define the (nonnegative) gap $\mathrm{Gap}(\pi) \triangleq V_{\mathcal{P}_{\mathrm{t}}}^{\pi^\star} - V_{\mathcal{P}_{\mathrm{t}}}^{\pi}$, where both $\mathcal{P}_{\mathrm{t}}$ and $\widehat{\mathcal{P}}_n$ are of radius $R_n$

**Theorem 7** (Suboptimality gap)**.** *Under Assumption 5 and the radius choice above, for rewards in $[0,1]$ and any $\gamma \in (0,1)$,*

$$\mathrm{Gap}(\pi_n) \;\le\; \frac{R_n}{(1-\gamma)^2} \;=\; \tilde{O}\left(\frac{L}{(1-\gamma)^2}\sqrt{\frac{\mathsf{d}_0}{n}}\right) \tag{9}$$

*with probability at least $1 - \delta_1$, where $\tilde{O}(.)$ is up to logarithmic factors.*

Hence, combining the high-probability coverage $\mathsf{P}_{\mathrm{t}} \in \widehat{\mathcal{P}}_n(R_n)$ with the evaluation bound of Theorem 2 yields a target-domain suboptimality gap that decays as $O(\sqrt{\mathsf{d}_0/n})$ (up to logs), rather than the $O(\sqrt{\mathsf{d}/n})$ rate obtained without side information. Theorem 7 makes this improvement explicit: by incorporating low-dimensional structure through LDS-IBE, the uncertainty radius $R_n$ required to cover the target dynamics is smaller, which directly translates into a tighter suboptimality gap.

**Remark 2** (Finite-sample rates for other IBE variants)**.** *The rate above is specific to LDS–IBE, where the low-dimensional structure reduces estimation to a finite-dimensional parametric problem, so standard concentration arguments yield a rate in $n$. For Moment–IBE and Density–IBE, $\delta_n$ can be controlled through dual-norm and integral-probability-metric arguments, as highlighted in Appendix E, but deriving an explicit rate is more involved and left as an open direction.*

# 5 RELATED WORK

**Robust RL.** Robust RL casts robustness as distributionally robust optimization (DRO), maximizing worst-case return over an uncertainty set of transition kernels [Iyengar, 2005, Nilim and Ghaoui, 2003, Wiesemann et al., 2013, Lim et al., 2013, Tamar et al., 2014]. Methods are either *model-based*, i.e., fit a model then apply robust dynamic programming [Lim and Autef, 2019, Yu and Xu, 2015, Yang et al., 2022, Shi et al., 2023, Wang et al., 2023a, Ho et al., 2021], or *model-free*, that is, optimize a robust objective without explicit modeling [Roy et al., 2017, Si et al., 2020, Zhou et al., 2021, Wang and Zou, 2021, Badrinath and Kalathil, 2021, Wang et al., 2024a, Liu et al., 2022, Wang et al., 2024b, 2023b]. Both typically rely on data from a single source environment. While worst-case guarantees yield lower bounds when the target kernel lies in the set, large source–target shifts force wide uncertainty sets, producing conservative policies that underperform in the target domain.

**Multi-Task RL.** Multi-task RL learns shared representations across tasks via joint training or transferable features to improve generalization across environments [Cheng et al., 2022, Agarwal et al., 2023, Zhang and Wang, 2021, Du et al., 2021]. Several works use this for *transfer*, aiming to boost a designated target task: Cheng et al. [2022] performs reward-free exploration on source tasks to learn representations, then adapts *online* on the target; Agarwal et al. [2023] strengthens this with a general state-dependent linear model and a cross-sampling scheme for in-distribution generalization. Related frameworks provide provable multi-task representation learning guarantees [Zhang and Wang, 2021, Du et al., 2021]. In contrast, our setting assumes only *limited offline target data* (no online interaction at deployment) and leverages side information to estimate target dynamics before (robustly) optimizing a policy.

**Domain Randomization and Meta Learning.** Domain randomization improves robustness by training over a wide distribution of simulated environments prior to deployment [Peng et al., 2018, Tobin et al., 2017, Chebotar et al., 2019]. Meta-RL trains across related tasks to learn fast adaptation from a small amount of *online* target experience [Finn et al., 2017, Nagabandi et al., 2019]. These approaches typically rely on broad simulator coverage or interaction at deployment; performance can degrade when the target lies outside the randomized/meta-training distribution or when online adaptation is limited. In contrast, we assume only *limited offline target data* and no online interaction, and we use side information to estimate target dynamics and (robustly) optimize the policy.

**Domain Adaptation (DA).** Prior DA methods mitigate source–target mismatch by transferring samples or reusing source data, often assuming shared dynamics [Taylor et al., 2008, Laroche and Barlier, 2017]. We instead consider *transition mismatch* and leverage side information to estimate target kernels, mitigating negative transfer under limited target data. Tirinzoni et al. [2018] extend FQI via importance weighting across multiple sources, fitting weights with Gaussian Processes , which limits applicability in discrete or non-Gaussian settings; our approach is agnostic to the transition model class. Wen et al. [2024] rank and filter source transitions using a mutual-information–based domain gap but do not address robustness to uncertain target dynamics; we explicitly model uncertainty sets around the estimated target kernel.

# 6 NUMERICAL EXPERIMENTS

We evaluate our approach against state-of-the-art baselines on six benchmarks. Due to space, we report *CartPole* here and defer the remaining tasks to Appendix I, which display similar trends.

**Baselines.** We compare to FQI [Ernst et al., 2005], Importance-Weighted FQI (IWFQI) [Tirinzoni et al., 2018], IGDF [Wen et al., 2024], and standard Q-learning. FQI and Q-learning use only *target* samples; IWFQI and IGDF use both source and target data. The original IWFQI estimates weights via Gaussian Processes (GPs), restricting it to continuous states. To isolate its best-case performance, we instead feed it oracle (true) importance weights. Our method first estimates the target kernel with IBE, then computes an optimal policy via non-robust or robust RL (Appendix G). We report performance in the target domain under both evaluations.

**Side Information and Environments.** We instantiate IBE using the constraints in Table 1: Distance IBE (TV and $W_1$ with bounds set to the true source–target distances, i.e., $d_{s,a} = \|\mathsf{P}_{\mathrm{s}}^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a}\|_{\mathrm{TV}}$ and $W_1(\mathsf{P}_{\mathrm{s}}^{s,a}, \mathsf{P}_{\mathrm{t}}^{s,a}) = d_{s,a}$), Moment IBE (moment gap $\beta_{s,a} = |\mu(\mathsf{P}_{\mathrm{t}}^{s,a}) - \mu(\mathsf{P}_{\mathrm{s}}^{s,a})|$), Density IBE "global" ($B_{s,a} = \max(\mathsf{P}_{\mathrm{t}}^{s,a}/\mathsf{P}_{\mathrm{s}}^{s,a})$), and Density IBE "local" (elementwise $B_{s,a} = \mathsf{P}_{\mathrm{t}}^{s,a}/\mathsf{P}_{\mathrm{s}}^{s,a} + 1$). We test on three toy-text tasks (Frozen Lake, Cliff Walking, Taxi) and three classic control problems (Acrobot, Cart Pole, Pendulum) from OpenAI Gym [Brockman et al., 2016]. In all cases, source and target differ only in their transition kernels. Environment details are in Appendix H.

**Sampling and Evaluation.** We draw $N$ state–action pairs uniformly and record counts $N(s, a)$. For each occurrence, the next state is sampled from the target kernel. We vary per-pair sample sizes over $\{1, 5, 10, 50, 100, 150, \ldots, 10^4\}$. Kernel estimates are initialized at the source kernel, except for Vanilla IBE, which uses a uniform initializer to assess learning purely from target data. Results are averaged over 20 runs with 95% confidence intervals.

**Performance Results.** We evaluate IBE in both non-robust and robust regimes. Transition kernels are estimated via (5). Policies are computed with Algorithm 1 (Appendix G). We set the robustness radius to $R = 0$ (non-robust) and $R = 0.1$

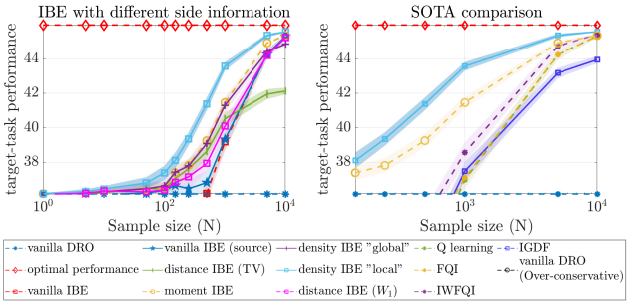



Figure 2: Target domain performance for the non-robust setting as a function of sample size for *CartPole*. The legend is shared between Figures 2 and 3.

(robust), and evaluate in the target domain with the same radius using Algorithm 2. Due to space, we show *CartPole* in the main text and defer the remaining environments to Appendix I.2. Performance is reported as the state-averaged value $\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} V^{\pi}(s)$ across varying target-sample budgets, averaged over 20 runs. Figures 2 and 3 plot non-robust and robust results, respectively: left panels compare IBE side-information variants; right panels compare our best IBE to SOTA baselines. Note that IGDF's assumptions fail for very small $N$ (e.g., $N \in \{1, 5\}$), so head-to-head comparisons with IGDF begin at $N \geq 150$, where all methods are well-defined.

As shown in the left panels of Figures 2 and 3, incorporating side information markedly improves target-domain performance over *Vanilla IBE* (samples only). We additionally initialize Vanilla IBE at the source kernel for unseen state-action pairs, matching our method's default, so as to isolate the effect of side information. The results indicate that the side information (not just initialization) drives the gains. Among our variants, *Density IBE (local)* performs best, plausibly because element-wise density-ratio bounds provide sharper guidance on the target transitions. In the right panels, *Density IBE (local)* also outperforms SOTA baselines in both non-robust and robust regimes, and the same is observed for other IBE variants, such as moment-IBE. We also compare to an over-conservative baseline that centers the uncertainty set at the *source* kernel and enlarges the radius to cover the target, $R + \|\mathsf{P}_{\mathrm{s}}^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a}\|_{\mathrm{TV}}$ (per $(s, a)$). As expected, this pessimistic construction yields poor target performance (see Figure 3).

**Dimension Effect.** We empirically test the dimension effect of LDS side information predicted by our theory. In *CartPole*, we parameterize the dynamics with a softmax model and follow the LDS-IBE procedure in Sec. 3.1 to estimate $\theta_{\mathrm{t}}$. We compare LDS-IBE (which constrains $\theta$ to the known low-dimensional subspace) to Vanilla IBE (no side information). Although the ambient parameter dimension is $\mathsf{d} = 4$ (Appendix H), the logits depend on $\theta$ through a $\mathsf{d}_0 = 2$ subspace, so the effective model dimension is two. Using the estimated dynamics, we learn a policy and

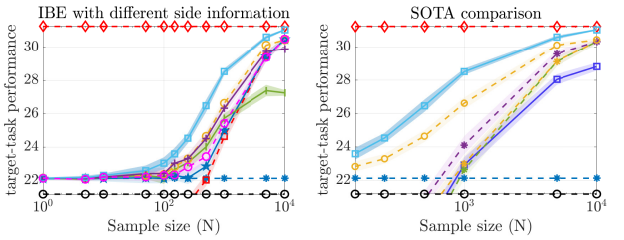



Figure 3: Target domain performance for the robust setting as a function of sample size for *CartPole*.

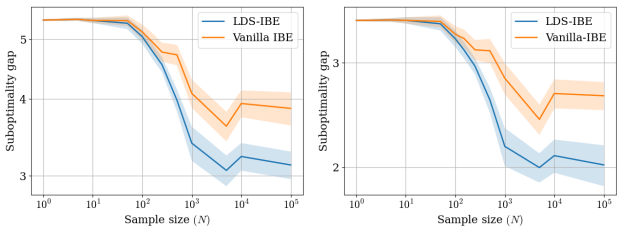



Figure 4: Suboptimality gap as a function of sample size $(N)$ (log-log scale) in the CartPole environment for LDS-IBE, for non-robust (left) and robust (right) scenarios.

evaluate it on the target environment in both non-robust and robust regimes. Figure 4 reports the suboptimality gap versus the number of target samples $N$ (left: non-robust, right: robust). As expected, LDS–IBE yields consistently smaller gaps than Vanilla IBE, reflecting the improved $\tilde{O}\big(\sqrt{\mathsf{d}_0/N}\big)$ dependence from exploiting the low-dimensional structure in the transition dynamics.

## 7 CONCLUSION

We proposed a model-based transfer RL framework leveraging limited offline target data and side information to estimate target dynamics and learn effective policies. Centering uncertainty sets at an information-based target estimate (IBE) rather than the source avoids the over-conservatism of source-centered DRO, yielding tighter radii and less pessimistic policies. Incorporating side-information constraints brings the estimate closer to the true target kernel, reducing the data needed for adaptation. Theoretically, we established value-function error bounds, scaling with the estimator's uniform TV error, and finite-sample guarantees. Under LDS, the robust suboptimality gap scales as $\tilde{O}(\sqrt{\mathsf{d}_0/n})$, explicitly quantifying the sample-efficiency gains from side information. Empirically, IBE—particularly its density-ratio and moment variants—consistently outperforms RL baselines in both robust and non-robust scenarios. Natural extensions include settings with reward shift between source and target, complementing the transition-kernel shift considered here, and extending the LDS-IBE finite-sample guarantees to continuous state and action spaces via function approximation, where guarantees would additionally depend on

approximation error.

## ACKNOWLEDGMENTS

This work was supported by DARPA under Agreement No. HR0011-24-9-0427 and NSF under Award CCF-2106339.

### References

Alekh Agarwal, Yuda Song, Wen Sun, Kaiwen Wang, Mengdi Wang, and Xuezhou Zhang. Provable benefits of representational transfer in reinforcement learning. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2114–2187. PMLR, 2023.

Kishan Panaganti Badrinath and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. pages 511–520. PMLR, 2021.

Andrew G Barto, Richard S Sutton, and Charles W Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, (5):834–846, 1983.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.

Yevgen Chebotar, Ankur Handa, Viktor Makoviychuk, Miles Macklin, Jan Issac, Nathan Ratliff, and Dieter Fox. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8973–8979. IEEE, 2019.

Yuan Cheng, Songtao Feng, Jing Yang, Hong Zhang, and Yingbin Liang. Provable benefit of multitask representation learning in reinforcement learning. *Advances in Neural Information Processing Systems*, 35:31741–31754, 2022.

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Partial policy iteration for L1-robust Markov decision processes. *Journal of Machine Learning Research*, 22 (275):1–46, 2021.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the $l\_\{1\}$ distance. *IEEE Transactions on Information Theory*, 64(10):6672–6706, 2018.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

Heiner Kremer, Jia-Jie Zhu, Krikamol Muandet, and Bernhard Schölkopf. Functional generalized empirical likelihood estimation for conditional moment restrictions. In *International Conference on Machine Learning*, pages 11665–11682. PMLR, 2022.

Romain Laroche and Merwan Barlier. Transfer reinforcement learning with shared dynamics. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2 edition, 2017.

Sergey Levine, Aviral Kumar, G. Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *ArXiv*, abs/2005.01643, 2020. URL `https://api.semanticscholar.org/CorpusID:218486979`.

Shiau Hong Lim and Arnaud Autef. Kernel-based reinforcement learning in robust Markov decision processes. In *International conference on machine learning*, pages 3973–3981. PMLR, 2019.

Shiau Hong Lim, Huan Xu, and Shie Mannor. Reinforcement learning in robust Markov decision processes. *Advances in neural information processing systems*, 26, 2013.

Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust $Q$-learning. pages 13623–13643. PMLR, 2022.

Thomas L Marzetta. A simple derivation of the constrained multiple parameter cramer-rao bound. *IEEE Transactions on Signal Processing*, 41(6):2247–2249, 1993.

Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S. Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. In

*International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyztsoC5Y7.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Arnab Nilim and Laurent Ghaoui. Robustness in Markov decision problems with uncertain transition matrices. *Advances in neural information processing systems*, 16, 2003.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018.

K. B. Petersen and M. S. Pedersen. The matrix cookbook, October 2008. URL http://www2.imm.dtu.dk/pubdb/p.php?3274. Version 20081110.

Aurko Roy, Huan Xu, and Sebastian Pokutta. Reinforcement learning under model mismatch. *Advances in neural information processing systems*, 30, 2017.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 36: 79903–79917, 2023.

Nian Si, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. Distributionally robust policy evaluation and learning in offline contextual bandits. pages 8884–8894. PMLR, 2020.

Richard S Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In D. Touretzky, M.C. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995.

Aviv Tamar, Shie Mannor, and Huan Xu. Scaling up robust mdps using function approximation. In *International conference on machine learning*, pages 181–189. PMLR, 2014.

Matthew E Taylor, Nicholas K Jong, and Peter Stone. Transferring instances for model-based reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II 19*, pages 488–505. Springer, 2008.

Andrea Tirinzoni, Andrea Sessa, Matteo Pirotta, and Marcello Restelli. Importance weighted transfer of samples in reinforcement learning. In *International Conference on Machine Learning*, pages 4936–4945. PMLR, 2018.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=tyrJsbKAe6.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. Sample complexity of variance-reduced distributionally robust q-learning. *Journal of Machine Learning Research*, 25(341):1–77, 2024a.

Yudan Wang, Shaofeng Zou, and Yue Wang. Model-free robust reinforcement learning with sample complexity analysis. In *Uncertainty in Artificial Intelligence*, pages 3470–3513. PMLR, 2024b.

Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

Yue Wang, Alvaro Velasquez, George Atia, Ashley Prater-Bennette, and Shaofeng Zou. Robust average-reward Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15215–15223, 2023a.

Yue Wang, Alvaro Velasquez, George K Atia, Ashley Prater-Bennette, and Shaofeng Zou. Model-free robust average-reward reinforcement learning. In *International Conference on Machine Learning*, pages 36431–36469. PMLR, 2023b.

Yue Wang, Jinjun Xiong, and Shaofeng Zou. Achieving the asymptotically optimal sample complexity of offline reinforcement learning: A dro-based approach. *Transactions on Machine Learning Research*, 2024c.

Xiaoyu Wen, Chenjia Bai, Kang Xu, Xudong Yu, Yang Zhang, Xuelong Li, and Zhen Wang. Contrastive representation for data filtering in cross-domain offline reinforcement learning. In *International Conference on Machine Learning*, pages 52720–52743. PMLR, 2024.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Huan Xu and Shie Mannor. Distributionally robust Markov decision processes. pages 2505–2513, 2010.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Toward theoretical understandings of robust Markov decision processes. *The Annals of Statistics*, 50(6):3223–3248, 2022.

Pengqian Yu and Huan Xu. Distributionally robust counterpart in Markov decision processes. *IEEE Transactions on Automatic Control*, 61(9):2538–2543, 2015.

Chicheng Zhang and Zhi Wang. Provably efficient multi-task reinforcement learning with model transfer. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19771–19783. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a440a3d316c5614c7a9310e902f4a43e-Paper.pdf.

Puning Zhao and Lifeng Lai. Minimax optimal estimation of kl divergence for continuous distributions. *IEEE Transactions on Information Theory*, 66(12):7787–7811, 2020.

Zhengqing Zhou, Qinxun Bai, Zhengyuan Zhou, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. pages 3331–3339. PMLR, 2021.

# Robust Transfer Learning With Side Information (Supplementary Material)

**Akram Awad**[1] **Shihab Ahmed**[1] **Yue Wang**[1,2] **George Atia**[1,2]

[1]Department of Electrical Engineering, University of Central Florida, Orlando, Florida, USA
[2]Department of Computer Science, University of Central Florida, Orlando, Florida, USA

## A DEFINITIONS

**Definition 8** (Total variation distance Levin et al. [2017])**.** Let $(\mathcal{X}, \mathcal{F})$ be a measurable space and let $\mathsf{P}, \mathsf{Q}$ be two probability distributions defined on it. The total variation (TV) distance is defined as

$$\|\mathsf{P} - \mathsf{Q}\|_{TV} \;:=\; \sup_{A\in\mathcal{F}} \big|\mathsf{P}(A) - \mathsf{Q}(A)\big|.$$

In the discrete case, $\|\mathsf{P} - \mathsf{Q}\|_{TV} = \frac{1}{2}\sum_{x\in\mathcal{X}} |\mathsf{P}(x) - \mathsf{Q}(x)|$.

**Definition 9** (Wasserstein–1 (Earth Mover's) distance Villani et al. [2008])**.** Let $(\mathcal{X}, \|\cdot\|)$ be a metric space and let $\mathsf{P}, \mathsf{Q}$ be probability measures on $\mathcal{X}$ with finite first moments. The Wasserstein–1 distance is

$$W_1(\mathsf{P}, \mathsf{Q}) \;:=\; \inf_{\rho\in\Pi(\mathsf{P},\mathsf{Q})} \int_{\mathcal{X}\times\mathcal{X}} \|x - y\|\, d\rho(x, y),$$

where $\Pi(\mathsf{P}, \mathsf{Q})$ is the set of couplings (joint distributions) with marginals $\mathsf{P}$ and $\mathsf{Q}$. For the cost $\|.\|$, we use the Euclidean distance, i.e., $\|x - y\|_2$, in this paper.

**Definition 10** (Value-Aware 1-Wasserstein distance )**.** Let $d_v(s, s') = |V^{\pi}_{\mathcal{P}}(s) - V^{\pi}_{\mathcal{P}}(s')|$ be a pseudometric, and $(\mathcal{S}, d_v)$ be a finite pseudometric space. The *Value-Aware 1-Wasserstein distance* between $\mathsf{P}$ and $\mathsf{Q}$ is defined as

$$W_{d_v}(\mathsf{P}, \mathsf{Q}) := \inf_{\rho\in\Pi(\mathsf{P},\mathsf{Q})} \sum_{s,s'\in\mathcal{S}} \rho(s, s') \cdot d_v(s, s') = \sup_{\|f\|_{\mathrm{Lip}(\tilde{d})}\le 1} \left|\mathbb{E}_{\mathsf{P}}[f] - \mathbb{E}_{\mathsf{Q}}[f]\right|,$$

## B CRAMÉR–RAO BOUNDS WITH SIDE INFORMATION

In this section, we examine the Cramér–Rao Bound (CRB) for the estimator $\widehat{\mathsf{P}}^{s,a}_n$ under the constrained estimation setting. The CRB provides a lower bound on the variance of a single parameter estimator or the covariance matrix of a vector with multiple parameters. In general, given an unbiased estimator $\hat{\zeta}$ of a $k$-element vector $\zeta$, the Fisher Information Matrix (FIM) is defined as $J(\zeta) = \mathbb{E}[\nabla \log(L(\zeta)) \nabla \log(L(\zeta))^{\top}]$, where $\nabla \log(L(\zeta))$ is the gradient of the log-likelihood function w.r.t. $\zeta$. The CRB then states that the covariance of the unbiased estimator is lower bounded by the inverse $J(\zeta)^{-1}$ of the FIM. Before presenting the main result of this subsection, we derive an expression for the FIM $J(\mathsf{P}^{s,a}_{\mathrm{t}})$ as stated in the next lemma.

**Lemma 11.** *Given $n$ i.i.d. samples $\{x_1, \ldots, x_n\} \sim \mathsf{P}^{s,a}_{\mathrm{t}}$, where $x \in \{s_1, \ldots, s_k\}$, with $\mathsf{P}^{s,a}_{\mathrm{t},j} := \Pr(x = s_j), \mathsf{P}^{s,a}_{\mathrm{t},j} \neq 0, \forall j \le k$, the FIM is given by*

$$J(\mathsf{P}^{s,a}_{\mathrm{t}}) = n\,\mathrm{diag}\left(\frac{1}{\mathsf{P}^{s,a}_{\mathrm{t},1}}, \ldots, \frac{1}{\mathsf{P}^{s,a}_{\mathrm{t},k}}\right). \tag{10}$$

We can readily state the main result of this subsection.

**Theorem 12.** *Let $\widehat{\mathsf{P}}_n^{s,a}$ be an unbiased estimator of $\mathsf{P}_{\mathrm{t}}^{s,a}$ based on $n$ i.i.d. samples. For each of the following constraints, the diagonal elements of the CRB matrix, $C_{ii}(\mathsf{P}_{\mathrm{t}}^{s,a})$, are given by:*

- ***Regular Probability Constraint*** *( $\sum_i^k \mathsf{P}_{\mathrm{t},i}^{s,a} = 1$):* $\quad C_{ii}(\mathsf{P}_{\mathrm{t}}^{s,a}) = \frac{1}{n}\mathsf{P}_{\mathrm{t},i}^{s,a}(1-\mathsf{P}_{\mathrm{t},i}^{s,a})$
- ***Moment Constraints*** *( $\mathbb{E}_{\mathsf{P}_{\mathrm{t}}^{s,a}}[\phi(x)] = \mu$, where $x \in \mathcal{S}$ and $\phi : \mathcal{S} \to \mathbb{R}^m$, and $\sum_i^k \mathsf{P}_{\mathrm{t},i}^{s,a} = 1$) :*

$$C_{ii}(\mathsf{P}_{\mathrm{t}}^{s,a}) = C_{ii}^R(\mathsf{P}_{\mathrm{t}}^{s,a}) - \Delta_i \tag{11}$$

*where*

  - *$C_{ii}^R$ is the $i^{th}$ diagonal element of the CRB matrix with regular probability constraints.*
  - *$\Delta_i = \frac{(\mathsf{P}_{\mathrm{t},i}^{s,a})^2}{n^2}(a_i - \overline{a})\mathbb{S}^{-1}(a_i - \overline{a})$, where $a_i = \phi(x_i) - \phi(x_k) \in \mathbb{R}^m, i = 1, \ldots, k$, $\overline{a} = \sum_{i=1}^k a_i \mathsf{P}_{\mathrm{t},i}^{s,a}$, and $\mathbb{S} = \frac{1}{n}\operatorname{Cov}(a)$, where $\operatorname{Cov}(\mathrm{a})$ is the covariance matrix of a.*

Theorem 12 provides valuable insights into the quality of the estimator as additional constraints are imposed. Specifically, the diagonal elements of the CRB give lower bounds on the variances of $\widetilde{\mathsf{P}}_{n,i}^{s,a}, 1 \le i \le k$, where $\widehat{\mathsf{P}}_{i,n}^{s,a}$ is the $i^{th}$ element of $\widehat{\mathsf{P}}_n^{s,a}$. A smaller lower bound indicates the potential for a better minimum variance estimator. To quantify the CRB matrix, we can compute the trace of the CRB $C$, which provides a lower bound on the sum of the estimator variances. Formally, we have $\sum_{i=1}^k \operatorname{var}[\widehat{\mathsf{P}}_n^{s,a}] \ge \operatorname{trace}(C(\mathsf{P}_{\mathrm{t}}^{s,a})) = \sum_{i=1}^k C_{ii}(\mathsf{P}_{\mathrm{t}}^{s,a})$. According to Theorem 12, we conclude that:

$$\frac{1}{n}\sum_{i=1}^k \mathsf{P}_{\mathrm{t},i}^{s,a}(1-\mathsf{P}_{\mathrm{t},i}^{s,a}) \ge \sum_{i=1}^k \frac{\mathsf{P}_{\mathrm{t},i}^{s,a}(1-\mathsf{P}_{\mathrm{t},i}^{s,a})}{n} - \frac{(\mathsf{P}_{\mathrm{t},i}^{s,a})^2(a_i-\overline{a})^\top \mathbb{S}(a_i-\overline{a})}{n^2}. \tag{12}$$

Since $\mathbb{S}$ is positive definite (assuming $\operatorname{Cov}(a)$ is full rank), $\Delta_i \ge 0$.

Equation (12) demonstrates that having more prior information about $\mathsf{P}_{\mathrm{t}}^{s,a}$ potentially leads to a better estimate, given the smaller bound on the variance. We also provide empirical evidence of the information gain that the side information provides in Appendix I.5.

# C VALUE-AWARE SIDE INFORMATION.

In this section, we present an evaluation error bound under the value-aware Wasserstein distance side information, where the bound is presented in terms of the side information. Formally, the IBE under value-aware 1-Wasserstein distance constraint is given as

$$\widehat{\mathsf{P}}_n^{s,a} = \arg\max_{q \in \Delta(\mathcal{S})} \sum_{s' \in \mathcal{S}} n_j q(s') \quad \text{subject to} \quad W_{d_V}(q, \mathsf{P}_s^{s,a}) \le \beta_1, \tag{13}$$

where $W_{d_V}(\cdot,\cdot)$ is Value-Aware 1-Wasserstein distance (see Definition 10), and $n_j = N_{s,a}(j), \quad j \in \mathcal{S}$.

The following Theorem provides a bound on the evaluation error in terms of the side information $\beta_1$.

**Theorem 13** (Bound under Value-Aware Wasserstein Constraint)**.** *Let $W_{d_V}(\cdot,\cdot)$ be the Value-Aware 1-Wasserstein distance (see Definition 10), and $\widehat{\mathsf{P}}_n^{s,a}$ be IBE, given by* (13)*. Suppose $W_{d_V}(\mathsf{P}_{\mathrm{s}}^{s,a}, \mathsf{P}_{\mathrm{t}}^{s,a}) \le \beta_1$. Then, we have*

$$\left\| V_{\mathcal{P}_{\mathrm{t}}}^{\pi_n} - V_{\mathcal{P}_{\mathrm{t}}}^{*} \right\|_\infty \le \frac{4\gamma\beta_1}{1-\gamma}.$$

Theorem 13 presents a bound on the evaluation error in terms of the side information $\beta_1$. In practice, the target domain value function $V_{\mathcal{P}_{\mathrm{t}}}^{\pi}$ is unknown, so we consider using the value function under the source dynamics $\mathsf{P}_{\mathrm{s}}^{s,a}$ to define the ground metric. We define the source metric $\tilde{d}(s,s')$ as $\tilde{d}(s,s') := |V_{\mathsf{P}_s^{s,a}}^{\pi}(s) - V_{\mathsf{P}_s^{s,a}}^{\pi}(s')|$. Therefore, we solve

$$\widehat{\mathsf{P}}_n^{s,a} = \arg\max_{q \in \Delta(\mathcal{S})} \sum_{s' \in \mathcal{S}} n_j q(s') \quad \text{subject to} \quad W_{\tilde{d}}(q, \mathsf{P}_{\mathrm{s}}^{s,a}) \le \beta_1, \tag{14}$$

where $W_{\tilde{d}}(\cdot,\cdot)$ is a Value-Aware 1-Wasserstein distance with the source metric $\tilde{d}(s,s')$. We also define the mismatch factor w.r.t the uncertainty set $\mathcal{P} = \bigotimes_{(s,a)} \mathbb{B}_{\mathrm{TV}}\big(\mathsf{P}^{s,a},\, R\big)$ as,

$$L_{\mathcal{P}} := \sup_{s \neq s'} \frac{|V^{\pi}_{\mathcal{P}}(s) - V^{\pi}_{\mathcal{P}}(s')|}{|V^{\pi}_{\mathsf{P}^{s,a}_{\mathrm{s}}}(s) - V^{\pi}_{\mathsf{P}^{s,a}_{\mathrm{s}}}(s')|}. \tag{15}$$

**Theorem 14** (Deviation Bound with Source-Defined Metric)**.** *Let $W_{\tilde{d}}(\cdot,\cdot)$ be the Value-Aware 1-Wasserstein distance defined w.r.t the source metric $\tilde{d}$, and $\widehat{\mathsf{P}}^{s,a}_n$ be the IBE, given by* (14)*. Suppose $W_{\tilde{d}}(\mathsf{P}^{s,a}_{\mathrm{s}}, \mathsf{P}^{s,a}_{\mathrm{t}}) \leq \beta_1$. Then, the value error is bounded as,*

$$\Big\| V^{\pi_n}_{\mathcal{P}_{\mathrm{t}}} - V^{*}_{\mathcal{P}_{\mathrm{t}}} \Big\|_{\infty} \leq \frac{4\gamma L_{\mathcal{P}_n}}{1-\gamma} \cdot \beta_1.$$

*where $L_{\mathcal{P}_n}$ is the mismatch factor w.r.t the uncertainty set $\mathcal{P}_n$.*

Theorem 14 bounds the evaluation error in terms of the side information $\beta_1$, with more practical constraints $\Phi$ that do not involve the target domain value function $V^{\pi}_{\mathcal{P}_{\mathrm{t}}}$.

# D DERIVING SIDE INFORMATION FROM SYSTEM KNOWLEDGE

In many robotics/control settings, the next state is generated as $s' = f(s,a,\theta) + \xi$, where $\theta$ collects physical parameters (e.g., masses, friction, damping, actuator gains) and $\xi$ is additive noise with fixed law. Calibration, manufacturer tolerances, and system identification routinely yield *bounded parameter sets* $\|\theta_t - \theta_s\| \leq \varepsilon$. For any fixed $(s,a)$, consider the two next-state laws $\mathsf{P}^{s,a}_{\theta_1}, \mathsf{P}^{s,a}_{\theta_2}$ induced by $\theta_1, \theta_2$. We obtain

$$W_1\big(\mathsf{P}^{s,a}_{\theta_1}, \mathsf{P}^{s,a}_{\theta_2}\big) \;\leq\; \mathbb{E}\,\big\| f(s,a,\theta_1) - f(s,a,\theta_2) \big\|.$$

If $f$ is $L_\theta$-Lipschitz in $\theta$ (uniformly in $(s,a)$), then

$$W_1\big(\mathsf{P}^{s,a}_{\theta_1}, \mathsf{P}^{s,a}_{\theta_2}\big) \;\leq\; L_\theta\, \|\theta_1 - \theta_2\| \quad\Rightarrow\quad W_1\big(\mathsf{P}^{s,a}_{\theta_t}, \mathsf{P}^{s,a}_{\theta_s}\big) \;\leq\; L_\theta\, \varepsilon.$$

On discrete spaces with minimum separation $m > 0$ under the ground cost, Kantorovich–Rubinstein yields

$$\|\mathsf{P}^{s,a}_{\theta_t} - \mathsf{P}^{s,a}_{\theta_s}\|_{\mathrm{TV}} \;\leq\; \tfrac{1}{m}\, W_1\big(\mathsf{P}^{s,a}_{\theta_t}, \mathsf{P}^{s,a}_{\theta_s}\big) \;\leq\; \tfrac{L_\theta}{m}\, \varepsilon,$$

providing a non-vacuous TV and Wasserstein radius for our Distance–IBE constraints. When physical parameter bounds are unavailable, the Wasserstein radius $d_{s,a}$ can alternatively be obtained from data via KL estimation. Specifically, on a state space of bounded diameter $D$, Pinsker-type inequalities give

$$W_1\big(\mathsf{P}^{s,a}_{\theta_t}, \mathsf{P}^{s,a}_{\theta_s}\big) \;\leq\; D\sqrt{\tfrac{1}{2}\mathrm{KL}\big(\mathsf{P}^{s,a}_{\theta_t} \| \mathsf{P}^{s,a}_{\theta_s}\big)},$$

so that a minimax-optimal KL estimate Zhao and Lai [2020] directly yields a valid Wasserstein radius, connecting sample-based divergence estimation to our Distance–IBE framework.

# E EXPLICIT DEPENDENCE OF $\delta_n$ ON DIFFERENT FORMS OF THE SIDE INFORMATION $\Phi$

**Overview.** Our value guarantees are indexed by the deviation $\delta_n \triangleq \max_{s,a} \|\widehat{\mathsf{P}}^{s,a}_n - \mathsf{P}^{s,a}_{\mathrm{t}}\|_{\mathrm{TV}}$. This appendix explains how different forms of side information $\Phi$ enter the bounds by restricting the feasible set of the constrained estimator and thereby reducing the deviation term $\delta_n$.

**Distance IBE (TV / Wasserstein constraints).** Suppose the side information specifies a source-anchored radius $d$, e.g., $W_1(\mathsf{P}^{s,a}_{\mathrm{t}}, \mathsf{P}^{s,a}_{\mathrm{s}}) \leq d$ or $\|\mathsf{P}^{s,a}_{\mathrm{t}} - \mathsf{P}^{s,a}_{\mathrm{s}}\|_{\mathrm{TV}} \leq d$, and that the estimator $\widehat{\mathsf{P}}^{s,a}_n$ is constrained to the same ball around the source kernel $\mathsf{P}^{s,a}_{\mathrm{s}}$. Then, by the triangle inequality,

$$\delta_n = \max_{s,a} \|\widehat{\mathsf{P}}^{s,a}_n - \mathsf{P}^{s,a}_{\mathrm{t}}\|_{\mathrm{TV}} \leq \max_{s,a} \|\mathsf{P}^{s,a}_{\mathrm{t}} - \mathsf{P}^{s,a}_{\mathrm{s}}\|_{\mathrm{TV}} + \max_{s,a} \|\widehat{\mathsf{P}}^{s,a}_n - \mathsf{P}^{s,a}_{\mathrm{s}}\|_{\mathrm{TV}} \leq 2d.$$

This yields an explicit, non-vacuous pre-asymptotic bound on $\delta_n$ in terms of the side-information radius $d$. When side information is expressed in $W_1$ or KL, analogous bounds follow via standard inequalities such as Kantorovich–Rubinstein and Pinsker's inequality (e.g., on discrete $\mathcal{S}$, $\mathrm{TV} \leq \frac{1}{m} W_1$ and $\mathrm{TV} \leq \sqrt{\frac{1}{2}\mathrm{KL}}$).

**Moment IBE.** A constraint of the form $\|\mu(q) - \mu(\mathsf{P}_{\mathrm{s}}^{s,a})\| \le \beta_{s,a}$ for feature map $\phi$ induces an integral probability metric–type control: $\sup_{\|g\|\le 1} \left|\mathbb{E}_q[g(\phi)] - \mathbb{E}_{\mathsf{P}_{\mathrm{t}}^{s,a}}[g(\phi)]\right| \le \beta_{s,a} + \|\mu(\mathsf{P}_{\mathrm{t}}^{s,a}) - \mu(\mathsf{P}_{\mathrm{s}}^{s,a})\|$. When the value function $V$ is Lipschitz in $\phi$, this control translates into a TV or $W_1$ bound, thereby reducing $\delta_n$ relative to unconstrained maximum-likelihood estimation at the same sample size $n$.

**Density IBE.** Density-ratio constraints of the form $0 \le q \le B\,\mathsf{P}_{\mathrm{s}}$ enforce support overlap and prevent off-support mass. For bounded $B$, such constraints yield uniform bounds on deviations of the form $|q^\top v - (\mathsf{P}_{\mathrm{t}})^\top v|$ for bounded value functions $v$, which again translate into TV control via dual norms, tightening $\delta_n$.

**LDS IBE.** Under a $\mathsf{d}_0$-dimensional parameterization and a TV-Lipschitz mapping $\theta \mapsto \mathsf{P}_\theta^{s,a}$, the constrained maximum-likelihood estimator admits a finite-sample deviation $R_n = \tilde{O}(\sqrt{\mathsf{d}_0/n})$, implying $\delta_n = \tilde{O}(\sqrt{\mathsf{d}_0/n})$. This rate underlies the improved robust suboptimality gap established in Theorem 7.

Collectively, these instantiations show how different choices of side information $\Phi$ enter the guarantees explicitly through $\delta_n$, and hence directly affect both the training and evaluation errors as well as the robust suboptimality gap.

# F PROOFS

## F.1 PROOF OF THEOREM 1

*Proof.* Considering any state $s$, we have that

$$|V_{\widehat{\mathcal{P}}_n}^{\pi_n}(s) - V_{\mathcal{P}_{\mathrm{t}}}^{\pi^*}(s)| \le \max_{\pi} |V_{\widehat{\mathcal{P}}_n}^{\pi}(s) - V_{\mathcal{P}_{\mathrm{t}}}^{\pi}(s)|, \tag{16}$$

where the inequality is from the fact that $|\max f - \max g| \le \max|f - g|$.

We denote the $(s,a)$-uncertainty sets $\mathbb{B}_{\mathrm{TV}}\big(\mathsf{P}_{\mathrm{t}}^{s,a},\, R\big)$ and $\mathbb{B}_{\mathrm{TV}}\big(\widehat{\mathsf{P}}_n^{s,a},\, R\big)$ as $(\mathcal{P}_{\mathrm{t}})_s^a$ and $(\widehat{\mathcal{P}}_n)_s^a$, respectively.

For any fixed policy $\pi$, we have

$$\begin{aligned}
|V_{\widehat{\mathcal{P}}_n}^{\pi}(s) - V_{\mathcal{P}_{\mathrm{t}}}^{\pi}(s)| &\overset{(a)}{=} |r_\pi(s) + \gamma\sigma_{(\widehat{\mathcal{P}}_n)_s^{\pi(s)}}(V_{\widehat{\mathcal{P}}_n}^{\pi}) - r_\pi(s) - \gamma\sigma_{(\mathcal{P}_{\mathrm{t}})_s^a}(V_{\mathcal{P}_{\mathrm{t}}}^{\pi})| \\
&= |\gamma\sigma_{(\mathcal{P}_n)_s^{\pi(s)}}(V_{\widehat{\mathcal{P}}_n}^{\pi}) - \gamma\sigma_{(\mathcal{P}_{\mathrm{t}})_s^a}(V_{\mathcal{P}_{\mathrm{t}}}^{\pi})| \\
&= |\gamma\sigma_{(\mathcal{P}_n)_s^{\pi(s)}}(V_{\widehat{\mathcal{P}}_n}^{\pi}) - \gamma\sigma_{(\mathcal{P}_{\mathrm{t}})_s^a}(V_{\widehat{\mathcal{P}}_n}^{\pi}) + \gamma\sigma_{(\mathcal{P}_{\mathrm{t}})_s^a}(V_{\widehat{\mathcal{P}}_n}^{\pi}) - \gamma\sigma_{(\mathcal{P}_{\mathrm{t}})_s^a}(V_{\mathcal{P}_{\mathrm{t}}}^{\pi})| \\
&\overset{(b)}{\le} \gamma\|V_{\widehat{\mathcal{P}}_n}^{\pi} - V_{\mathcal{P}_{\mathrm{t}}}^{\pi}\|_\infty + \gamma|\sigma_{(\mathcal{P}_n)_s^{\pi(s)}}(V_{\widehat{\mathcal{P}}_n}^{\pi}) - \sigma_{(\mathcal{P}_{\mathrm{t}})_s^a}(V_{\widehat{\mathcal{P}}_n}^{\pi})|,
\end{aligned} \tag{17}$$

where $(a)$ is from the robust Bellman equation, and $(b)$ is from the Lipschitz smoothness of $\sigma_{\mathcal{P}}(\cdot)$ Wang and Zou [2021].

Consider the second term $|\sigma_{(\mathcal{P}_n)_s^{\pi(s)}}(V_{\widehat{\mathcal{P}}_n}^{\pi}) - \sigma_{(\mathcal{P}_{\mathrm{t}})_s^a}(V_{\widehat{\mathcal{P}}_n}^{\pi})|$. Note that the support function can be solved by its dual form:

$$\sigma_{\mathcal{P}}(V) = \min_{\mathsf{P}\in\mathcal{P}} \mathsf{P}^\top V = \max_{\alpha\ge 0}\{\mathsf{P}_0^\top(V-\alpha) - R\mathbf{Span}(V-\alpha)\}, \tag{18}$$

where $\mathsf{P}_0$ is the center of $\mathcal{P}$, $R$ is the radius, and $\mathbf{Span}(V) = \max_i V(i) - \min_i V(i)$, for the vector $V = [V(1), \ldots, V(S)]$. Hence, (16) can be further bounded as

$$|V_{\widehat{\mathcal{P}}_n}^{\pi_n}(s) - V_{\mathcal{P}_{\mathrm{t}}}^{\pi^*}(s)| \le \gamma\|V_{\widehat{\mathcal{P}}_n}^{\pi} - V_{\mathcal{P}_{\mathrm{t}}}^{\pi}\| + \gamma|(\widehat{\mathsf{P}}_n^{s,\pi(s)} - \mathsf{P}_{\mathrm{t}}^{s,\pi(s)})^\top V_{\widehat{\mathcal{P}}_n}^{\pi}|\,, \tag{19}$$

which is from the fact that $|\max_x f(x) - \max_x g(x)| \le \max_x |f(x) - g(x)|$. Note that (19) holds for any $s$ and any policy $\pi$, thus

$$\|V_{\widehat{\mathcal{P}}_n}^{\pi_n} - V_{\mathcal{P}_{\mathrm{t}}}^{\pi^*}\|_\infty \le \gamma\|V_{\widehat{\mathcal{P}}_n}^{\pi} - V_{\mathcal{P}_{\mathrm{t}}}^{\pi}\|_\infty + \gamma\max_{(s,a)}\|(\widehat{\mathsf{P}}_n^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a})^\top V_{\widehat{\mathcal{P}}_n}^{\pi}\|_1\,. \tag{20}$$

Then, recursively applying (20), and using the fact $V_{\widehat{\mathcal{P}}_n}^{\pi}(s) \le \frac{1}{1-\gamma}$, implies that

$$\|V_{\widehat{\mathcal{P}}_n}^{\pi_n} - V_{\mathcal{P}_{\mathrm{t}}}^{\pi^*}\|_\infty \le \sum_{t=1}^{\infty}\frac{\gamma^t}{1-\gamma}\max_{(s,a)}\|\widehat{\mathsf{P}}_n^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a}\|_1 = 2\max_{(s,a)}\frac{\|\widehat{\mathsf{P}}_n^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a}\|_{TV}}{(1-\gamma)^2}\,. \tag{21}$$

□

### F.2 PROOF OF THEOREM 2

*Proof.* Denote the optimal policy w.r.t. the uncertainty set $\mathcal{P}_n$ by $\pi_n$, i.e.,

$$\pi_n = \arg\max_{\pi} V^{\pi}_{\mathcal{P}_n}. \tag{22}$$

Then, we have that

$$\|V^{\pi_n}_{\mathcal{P}_\mathrm{t}} - V^{\pi^*}_{\mathcal{P}_\mathrm{t}}\|_\infty \leq \|V^{\pi_n}_{\mathcal{P}_\mathrm{t}} - V^{\pi_n}_{\mathcal{P}_n}\|_\infty + \|V^{\pi_n}_{\mathcal{P}_n} - V^{\pi^*}_{\mathcal{P}_\mathrm{t}}\|_\infty \tag{23}$$

.

Following the same technique in the proof as of Theorem 1, we know that

$$\|V^{\pi_n}_{\mathcal{P}_n} - V^{\pi^*}_{\mathcal{P}_\mathrm{t}}\|_\infty \leq \max_{\pi} \|V^{\pi}_{\widehat{\mathcal{P}}_n} - V^{\pi}_{\mathcal{P}_\mathrm{t}}\|_\infty \leq 2 \max_{(s,a)} \frac{\|\widehat{\mathsf{P}}^{s,a}_n - \mathsf{P}^{s,a}_\mathrm{t}\|_{TV}}{(1-\gamma)^2} \tag{24}$$

and

$$\begin{aligned}
\|V^{\pi_n}_{\mathcal{P}_\mathrm{t}} - V^{\pi_n}_{\mathcal{P}_n}\|_\infty &\leq \max_{\pi} \|V^{\pi}_{\widehat{\mathcal{P}}_n} - V^{\pi}_{\mathcal{P}_\mathrm{t}}\|_\infty \\
&\leq \sum_{t=1}^{\infty} \frac{\gamma^t}{1-\gamma} \max_{(s,a)} \|\widehat{\mathsf{P}}^{s,a}_n - \mathsf{P}^{s,a}_\mathrm{t}\|_1 \\
&= 2 \max_{(s,a)} \frac{\|\widehat{\mathsf{P}}^{s,a}_n - \mathsf{P}^{s,a}_\mathrm{t}\|_{TV}}{(1-\gamma)^2}.
\end{aligned} \tag{25}$$

Combining RHS of (24) and (25) completes the proof.

□

### F.3 PROOF OF COROLLARY 3

*Proof.* The proof follows directly from the error bounds provided in Theorem 1 and 2. □

### F.4 PROOF OF PROPOSITION 4

We denote the optimal vanilla IBE by $\widetilde{\mathsf{P}}^{s,a}_n$, while the contrained IBE is denoted by $\widehat{\mathsf{P}}^{s,a}_n$ . We first establish the following lemma on the convergence of the vanilla IBE in total variation.

**Lemma 15.** *Let $\widetilde{\mathsf{P}}^{s,a}_n$ be the vanilla IBE of $\mathsf{P}^{s,a}_\mathrm{t}$, then*

$$\lim_{n\to\infty} \|\widetilde{\mathsf{P}}^{s,a}_n - \mathsf{P}^{s,a}_t\|_{TV} = 0. \tag{26}$$

*Proof.* By the consistency of the MLE, for a sequence $x^n = (x_1, \ldots, x_n)$ drawn i.i.d. from the discrete distribution $\mathsf{P}^{s,a}_\mathrm{t}$, the MLE estimate of $\mathsf{P}^{s,a}_\mathrm{t}$ is given by

$$\widetilde{\mathsf{P}}^{s,a}_{n,j} = \frac{n_j}{n}, \quad j = 1, \ldots, k \tag{27}$$

where $n_j := \sum_{i=1}^{n} \mathbf{1}(x_i = j)$ represents the number of times state $j$ is observed in the sample.

By the Strong Law of Large Numbers (SLLN), we have that

$$\frac{n_j}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i = j) \xrightarrow{\text{a.s.}} \mathsf{P}^{s,a}_{\mathrm{t},j}, \quad \forall j \in [k].$$

By the almost sure convergence, $\Pr(w \in \Omega : \lim_{n\to\infty} \frac{n_j}{n} = \mathsf{P}_{\mathrm{t},j}^{s,a}) = 1$, where $\Omega$ is the sample space. Hence, for any $\epsilon > 0$, $\exists N_j$ such that

$$\left| \frac{n_j}{n} - \mathsf{P}_{\mathrm{t},j}^{s,a} \right| \leq \frac{\epsilon}{k}, \quad \forall n \geq N_j. \tag{28}$$

Let $N = \max(N_1, \ldots, N_k)$. It follows that

$$\sum_{j=1}^{k} \left| \frac{n_j}{n} - \mathsf{P}_{\mathrm{t},j}^{s,a} \right| \leq \epsilon, \quad \forall n \geq N. \tag{29}$$

By (27), $\sum_{j=1}^{k} \left| \widetilde{\mathsf{P}}_{n,j}^{s,a} - \mathsf{P}_{\mathrm{t},j}^{s,a} \right| \leq \epsilon, \quad \forall n \geq N$. Hence, $\|\widetilde{\mathsf{P}}_{n}^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a}\|_{TV} \to 0$, which completes the proof. □

Now, we are ready to prove Proposition 4.

*Proof.* We prove the results for each IBE method listed in Table 1. Both the vanilla IBE and the IBE programs have a unique solution for each $n$, as the objective is strictly concave and the constraints are convex. We use the results of Lemma 15, which states that for any $\varepsilon > 0$, $\exists N_0(\varepsilon)$ such that $\|\widetilde{\mathsf{P}}_{n}^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a}\|_{TV} \leq \varepsilon, \forall n > N_0$, to prove each IBE case.

**Distance IBE**.

Let $d_{s,a}^0 = \|\mathsf{P}_{\mathrm{s}}^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a}\|_1 = 2\|\mathsf{P}_{\mathrm{s}}^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a}\|_{TV}$ . Since $\mathsf{P}_{\mathrm{t}}^{s,a}$ is an interior point of the constraint set, we can choose $\varepsilon \leq d_{s,a} - \frac{d_{s,a}^0}{2}$. Hence, for sufficiently large $n$,

$$\|\widetilde{\mathsf{P}}_{n}^{s,a} - \mathsf{P}_{\mathrm{s}}^{s,a}\|_{TV} \overset{(a)}{\leq} \|\widetilde{\mathsf{P}}_{n}^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a}\|_{TV} + \|\mathsf{P}_{\mathrm{s}}^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a}\|_{TV} \overset{(b)}{\leq} \varepsilon + \frac{d_{s,a}^0}{2} \overset{(c)}{\leq} d_{s,a} \tag{30}$$

where (a) follows from the triangular inequality, (b) by Theorem 15, and (c) from the choice of $\varepsilon$. Hence, $\widetilde{\mathsf{P}}_{n}^{s,a}$ satisfies the distance constraint. Therefore, by the uniqueness of $\widehat{\mathsf{P}}_{n}^{s,a}$ (the solution of Equation (5)), it follows that $\widetilde{\mathsf{P}}_{n}^{s,a}$ solves the distance IBE program. Consequently, $\widehat{\mathsf{P}}_{n}^{s,a} \to \mathsf{P}_{\mathrm{t}}^{s,a}$ in total variation.

**Density IBE**.
We first denote the density ratio between $\mathsf{P}_{\mathrm{t}}^{s,a}$ and $\mathsf{P}_{\mathrm{s}}^{s,a}$ by $B_{s,a}^0$, i.e.

$$B_{s,a}^0 = \frac{\mathsf{P}_{\mathrm{t}}^{s,a}}{\mathsf{P}_{\mathrm{s}}^{s,a}} \tag{31}$$

By Theorem 15, each entry of $\widetilde{\mathsf{P}}_{n}^{s,a}$ converges to the corresponding entry of $\mathsf{P}_{\mathrm{t}}^{s,a}$. For sufficiently large $n$, we have

$$\left| \frac{\widetilde{\mathsf{P}}_{n,i}^{s,a}}{\mathsf{P}_{\mathrm{t},i}^{s,a}} - 1 \right| \leq \varepsilon \implies \frac{\widetilde{\mathsf{P}}_{n,i}^{s,a}}{\mathsf{P}_{\mathrm{t},i}^{s,a}} \leq 1 + \varepsilon \tag{32}$$

where $\widetilde{\mathsf{P}}_{n,i}^{s,a}$ and $\mathsf{P}_{\mathrm{t},i}^{s,a}$ are the $i^{th}$ element of $\widetilde{\mathsf{P}}_{n}^{s,a}$ and $\mathsf{P}_{\mathrm{t}}^{s,a}$, respectively. Then, we get that

$$\frac{\widetilde{\mathsf{P}}_{n,i}^{s,a}}{\mathsf{P}_{\mathrm{s},i}^{s,a}} \overset{(a)}{\leq} B_{s,a}^{0,i}(1+\varepsilon) \overset{(b)}{\leq} B_{s,a}^{i} \tag{33}$$

where $B_{s,a}^{0,i}$, $B_{s,a}^{i}$ and $\mathsf{P}_{\mathrm{s},i}^{s,a}$ are the $i^{th}$ entry of $B_{s,a}^0$, $B_{s,a}$, and $\mathsf{P}_{\mathrm{s}}^{s,a}$, respectively. Here: (a) follows from Eqs. (31) and (32), while (b) by the choice of $\varepsilon$. We conclude that $\widetilde{\mathsf{P}}_{n}^{s,a}$ satisfies the Density IBE's constraint. Since $\widehat{\mathsf{P}}_{n}^{s,a}$ is a unique solution of Equation (5), $\widetilde{\mathsf{P}}_{n}^{s,a}$ solves the density IBE. Thus, the result follows.

**Moment IBE**.
We define $\beta_{s,a}^0$ as the absolute difference between the embedding means of $\mathsf{P}_{\mathrm{s}}^{s,a}$ and $\mathsf{P}_{\mathrm{t}}^{s,a}$, i.e.

$$\beta_{s,a}^0 = \left| \mu(\mathsf{P}_{\mathrm{t}}^{s,a}) - \mu(\mathsf{P}_{\mathrm{s}}^{s,a}) \right|. \tag{34}$$

Now, we prove that each entry of $|\mu(\widetilde{\mathsf{P}}_n^{s,a}) - \mu(\mathsf{P}_{\mathrm{t}}^{s,a})|$ goes to zero as $n \to \infty$. Suppose $\phi$ maps the state space (of $K$ states) to a high-dimensional feature space of dimension $M$, and consider the matrix $A$ of features of size $M \times K$, i.e.

$$A = \begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_M \end{bmatrix} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,K} \\ \vdots & \ddots & \vdots \\ a_{M,1} & \cdots & a_{M,K} \end{bmatrix} \tag{35}$$

where $\phi(x_i) = (a_{1,i}, \ldots, a_{M,i})^\top, x_i \in \mathcal{S}, i \leq K$.

$$\begin{aligned} |\mu(\widetilde{\mathsf{P}}_n^{s,a}) - \mu(\mathsf{P}_{\mathrm{t}}^{s,a})| &= |A(\widetilde{\mathsf{P}}_n^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a})| \\ &= \begin{bmatrix} |\sum_{i=1}^K a_{1,i}(\widetilde{\mathsf{P}}_{n,i}^{s,a} - \mathsf{P}_{\mathrm{t},i}^{s,a})| \\ \vdots \\ |\sum_{i=1}^K a_{M,i}(\widetilde{\mathsf{P}}_{n,i}^{s,a} - \mathsf{P}_{\mathrm{t},i}^{s,a})| \end{bmatrix} \\ &\leq \begin{bmatrix} \|\mathbf{a}_1^\top(\widetilde{\mathsf{P}}_n^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a})\|_1 \\ \vdots \\ \|\mathbf{a}_M^\top(\widetilde{\mathsf{P}}_n^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a})\|_1 \end{bmatrix} \\ &\overset{(*)}{\leq} \begin{bmatrix} \|A\|_\infty \|(\widetilde{\mathsf{P}}_n^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a})\|_1 \\ \vdots \\ \|A\|_\infty \|(\widetilde{\mathsf{P}}_n^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a})\|_1 \end{bmatrix} \\ &= 2\|A\|_\infty \|\widetilde{\mathsf{P}}_n^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a}\|_{TV} \mathbf{1}\,, \end{aligned} \tag{36}$$

where $\mathbf{1}$ is the vector of all ones and (*) follows from Hölder's inequality. Hence, by Theorem 15, we have:

$$\lim_{n\to\infty} |\mu(\widetilde{\mathsf{P}}_n^{s,a}) - \mu(\mathsf{P}_{\mathrm{t}}^{s,a})| = 0\,. \tag{37}$$

For sufficiently large $n$,

$$\begin{aligned} |\mu(\widetilde{\mathsf{P}}_n^{s,a}) - \mu(\mathsf{P}_{\mathrm{s}}^{s,a})| &= |A(\widetilde{\mathsf{P}}_n^{s,a} - \mathsf{P}_{\mathrm{s}}^{s,a})| \\ &= |A(\widetilde{\mathsf{P}}_n^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a} + \mathsf{P}_{\mathrm{t}}^{s,a} - \mathsf{P}_{\mathrm{s}}^{s,a})| \\ &\overset{(a)}{\leq} |A(\widetilde{\mathsf{P}}_n^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a})| + |A(\mathsf{P}_{\mathrm{t}}^{s,a} - \mathsf{P}_{\mathrm{s}}^{s,a})| \\ &\overset{(b)}{\leq} \varepsilon + \beta_{s,a}^0 \\ &\overset{(c)}{\leq} \beta_{s,a} \end{aligned} \tag{38}$$

where (a) follows by the triangular inequality, (b) by using Eqs. (34) and (37), and (c) by the choice $\varepsilon$. Then, $\widehat{\mathsf{P}}_n^{s,a}$ satisfies the moment constraints. The result follows by the uniqueness of $\widehat{\mathsf{P}}_n^{s,a}$.

□

### F.5 PROOF OF LEMMA 6

*Proof.* Define the constrained MLE

$$\widehat{\theta} = \arg\max_{\theta \in \Theta_0} \mathcal{L}(\theta)$$

where $\mathcal{L}$ is the log-likelihood function under $\mathsf{P}_\theta$. By the consistency of the MLE, we have with probability at least $1 - \delta_1$ that

$$\|\widehat{\theta} - \theta^*\| \leq C_0 \sqrt{\frac{\mathsf{d}_0}{n}}$$

for some constant $C_0$. By Assumption 5, we have

$$\|\mathsf{P}_{\widehat{\theta}} - \mathsf{P}_{\mathrm{t}}^{s,a}\|_{TV} \leq \|\widehat{\theta} - \theta^*\| \leq LC_0\sqrt{\frac{\mathsf{d}_0}{n}} =: R_n \,.$$

This ensures $\mathsf{P}_{\mathrm{t}} \in \widehat{\mathcal{P}}_n(R_n)$ with probability at least $1 - \delta_1$. □

### F.6 PROOF OF THEOREM 7

*Proof.* The proof follows from Lemma 6 and Theorem 2.

□

### F.7 PROOF OF THEOREM 13

*Proof.* Denote the optimal policy w.r.t the uncertainty set $\mathcal{P}_n$ by $\pi_n$, i.e.,

$$\pi_n = \arg\max_{\pi} V^{\pi}_{\mathcal{P}_n}. \tag{39}$$

Then, we have that

$$\|V^{\pi_n}_{\mathcal{P}_{\mathrm{t}}} - V^{\pi^*}_{\mathcal{P}_{\mathrm{t}}}\|_\infty \leq \|V^{\pi_n}_{\mathcal{P}_{\mathrm{t}}} - V^{\pi_n}_{\mathcal{P}_n}\|_\infty + \|V^{\pi_n}_{\mathcal{P}_n} - V^{\pi^*}_{\mathcal{P}_{\mathrm{t}}}\|_\infty. \tag{40}$$

For the first term on the RHS of (40), we follow the proof of Theorem 1. Therefore, we have for any $s \in \mathcal{S}$:

$$|V^{\pi_n}_{\widehat{\mathcal{P}}_n}(s) - V^{\pi^*}_{\mathcal{P}_{\mathrm{t}}}(s)| \leq \max_{\pi} |V^{\pi}_{\widehat{\mathcal{P}}_n}(s) - V^{\pi}_{\mathcal{P}_{\mathrm{t}}}(s)|. \tag{41}$$

For any policy $\pi$, we have

$$|V^{\pi}_{\widehat{\mathcal{P}}_n}(s) - V^{\pi}_{\mathcal{P}_{\mathrm{t}}}(s)| \leq \gamma\|V^{\pi}_{\widehat{\mathcal{P}}_n} - V^{\pi}_{\mathcal{P}_{\mathrm{t}}}\| + \gamma|(\widehat{\mathsf{P}}_n^{s,\pi(s)} - \mathsf{P}_{\mathrm{t}}^{s,\pi(s)})^\top V^{\pi}_{\widehat{\mathcal{P}}_n}| \,, \tag{42}$$

By Kantorovich duality, given two distributions $\mathsf{P}, \mathsf{Q} \in \Delta(\mathcal{S})$, the Wasserstein-1 distance induced by $d_V$ is defined as

$$W_{d_V}(\mathsf{P}, \mathsf{Q}) := \sup_{\|f\|_{\mathrm{Lip}(d_V)} \leq 1} |\mathbb{E}_{\mathsf{P}}[f] - \mathbb{E}_{\mathsf{Q}}[f]| \,,$$

where

$$\|f\|_{\mathrm{Lip}(d_V)} := \sup_{s \neq s'} \frac{|f(s) - f(s')|}{d_V(s, s')}.$$

Since $V^{\pi}_{\widehat{\mathcal{P}}_n}$ is 1-Lipschitz with respect to $d_V$, we have

$$|(\widehat{\mathsf{P}}_n^{s,\pi(s)} - \mathsf{P}_{\mathrm{t}}^{s,\pi(s)})^\top V^{\pi}_{\widehat{\mathcal{P}}_n}| \leq W_{d_V}(\widehat{\mathsf{P}}_n^{s,\pi(s)}, \mathsf{P}_{\mathrm{t}}^{s,\pi(s)}). \tag{43}$$

By the triangle inequality, we obtain,

$$W_{d_V}(\widehat{\mathsf{P}}_n^{s,\pi(s)}, \mathsf{P}_{\mathrm{t}}^{s,\pi(s)}) \leq W_{d_V}(\widehat{\mathsf{P}}_n^{s,\pi(s)}, \mathsf{P}_{\mathrm{s}}^{s,\pi(s)}) + W_{d_V}(\mathsf{P}_{\mathrm{s}}^{s,\pi(s)}, \mathsf{P}_{\mathrm{t}}^{s,\pi(s)}) \leq 2\beta_1.$$

Therefore, we have,

$$|V^{\pi}_{\widehat{\mathcal{P}}_n}(s) - V^{\pi}_{\mathcal{P}_{\mathrm{t}}}(s)| \leq \gamma\|V^{\pi}_{\widehat{\mathcal{P}}_n} - V^{\pi}_{\mathcal{P}_{\mathrm{t}}}\|_\infty + 2\gamma\beta_1.$$

Applying the above recursively, we get

$$\|V^{\pi_n}_{\widehat{\mathcal{P}}_n} - V^{\pi^*}_{\mathcal{P}_{\mathrm{t}}}\|_\infty \leq \frac{2\beta_1\gamma}{(1-\gamma)}.$$

Also, we have

$$\|V^{\pi_n}_{\mathcal{P}_{\mathrm{t}}} - V^{\pi_n}_{\mathcal{P}_n}\|_\infty \leq \max_{\pi} \|V^{\pi}_{\widehat{\mathcal{P}}_n} - V^{\pi}_{\mathcal{P}_{\mathrm{t}}}\|_\infty \leq \frac{2\beta_1\gamma}{(1-\gamma)}.$$

Combining the RHS of both terms of (40) completes the proof.

□

### F.8 PROOF OF THEOREM 14

*Proof.* Repeat the proof of Theorem 13. The key change is that $V^{\pi}_{\widehat{\mathcal{P}}_n}$ may not be 1-Lipschitz under $\tilde{d}$, but it is $L_{\mathcal{P}_n}$-Lipschitz, i.e.,

$$|V^{\pi}_{\widehat{\mathcal{P}}_n}(s) - V^{\pi}_{\widehat{\mathcal{P}}_n}(s')| \leq L_{\mathcal{P}_n} \cdot \tilde{d}(s, s'),$$

Therefore,

$$|(\widehat{\mathsf{P}}_n^{s,\pi(s)} - \mathsf{P}_{\mathrm{t}}^{s,\pi(s)})^{\top} V^{\pi}_{\widehat{\mathcal{P}}_n}| \leq L_{\mathcal{P}_n} W_{\tilde{d}}(\widehat{\mathsf{P}}_n^{s,\pi(s)}, \mathsf{P}_{\mathrm{t}}^{s,\pi(s)}).$$

The rest of the proof proceeds identically. □

### F.9 PROOF OF LEMMA 11

*Proof.* To simplify notation, we denote $\mathsf{P}_{\mathrm{t}}^{s,a}$ by $\mathsf{P}^{'}$. Let $S$ be a random vector that takes one of the values from $\{e_1, \ldots, e_k\}$, where $e_j, j = 1, \ldots, k$, is a vector with a 1 in the $j$-th position and 0 elsewhere (one-hot encoding). The entries are denoted $s_j$, where $\Pr(s_j = 1) = \Pr(S = e_j) = \mathsf{P}^{'}_j$, so the vector $\mathsf{P}^{'} = [\mathsf{P}^{'}_1, \ldots, \mathsf{P}^{'}_k]^T$, and $\sum_{j=1}^{k} \mathsf{P}^{'}_j = 1$. To compute the Fisher Information Matrix (FIM), we define the likelihood

$$\Pr(S|\mathsf{P}^{'}) = \prod_{j=1}^{k} {\mathsf{P}^{'}_j}^{s_j}.$$

Hence,

$$\log \Pr(s|\mathsf{P}^{'}) = \sum_{j=1}^{k} s_j \log \mathsf{P}^{'}_j.$$

Therefore, the FIM, $I_S(\mathsf{P}^{'}) = \mathbb{E}[\nabla \log \Pr(s|\mathsf{P}^{'}) \nabla \log \Pr(s|\mathsf{P}^{'})^T]$, is given by:

$$I_S(\mathsf{P}^{'}) = \mathbb{E}\left( \begin{bmatrix} s_1/\mathsf{P}^{'}_1 \\ \vdots \\ s_k/\mathsf{P}^{'}_k \end{bmatrix} \begin{bmatrix} s_1/\mathsf{P}^{'}_1 & \cdots & s_k/\mathsf{P}^{'}_k \end{bmatrix} \right).$$

This simplifies to

$$I_S(\mathsf{P}^{'}) = \mathbb{E}\begin{bmatrix} (s_1/\mathsf{P}^{'}_1)^2 & s_1 s_2/(\mathsf{P}^{'}_1 \mathsf{P}^{'}_2) & \cdots & s_1 s_k/(\mathsf{P}^{'}_1 \mathsf{P}^{'}_k) \\ \vdots & \ddots & \ddots & \vdots \\ s_k s_1/(\mathsf{P}^{'}_k \mathsf{P}^{'}_1) & \cdots & \cdots & (s_k/\mathsf{P}^{'}_k)^2 \end{bmatrix}.$$

Now, $\mathbb{E}[(s_i/\mathsf{P}^{'}_i)^2] = \frac{1}{\mathsf{P}^{'2}_i}\mathbb{E}[s_i^2] = 1/\mathsf{P}^{'2}_i(1 \cdot \mathsf{P}^{'}_i + 0) = 1/\mathsf{P}^{'}_i, i = 1, \ldots, k$ and $\mathbb{E}[(s_i s_j/\mathsf{P}^{'}_i \mathsf{P}^{'}_j)] = (1/\mathsf{P}^{'}_i \mathsf{P}^{'}_j)\mathbb{E}[s_i s_j] = 0$, for $i \neq j$. Thus,

$$I_S(\mathsf{P}^{'}) = \mathrm{diag}\left( \frac{1}{\mathsf{P}^{'}_1}, \ldots, \frac{1}{\mathsf{P}^{'}_k} \right).$$

If we observe $n$ i.i.d. vectors $S$, we get that $J(\mathsf{P}^{'}) = nI_S(\mathsf{P}^{'})$.

□

### F.10 PROOF OF THEOREM 12

We use the same notation simplifications as in the proof of Lemma 11. We start by bringing in the following lemma.

**Lemma 16** (Theroem 2, Marzetta [1993])**.** *Let the $k$-parameter vector $\widehat{\theta}$ be a constrained unbiased estimator of the parameter $\theta$, with $M$ constrains $f(\theta) = 0$ for some real valued function $f$. Then, the CRB is given by:*

$$C(\theta) = J^{-1} - J^{-1}F(F^T J^{-1} F)^{-1} F^T J^{-1}. \tag{44}$$

*where J is the FIM and F is $k \times M$ gradient matrix w.r.t $\theta$.*

*Proof.* **Regular probability constraint** ($\sum_i^k \mathsf{P}_i^{'} = 1$):

Due to the normalization constraint $\sum_i^k \mathsf{P}_i^{'} = 1$, the probabilities $\mathsf{P}_i^{'}$ are dependent. To handle this, we express $\mathsf{P}_k^{'} = 1 - \sum_i^{k-1} \mathsf{P}_i^{'}$, and use a reduced parameter vector $\theta = [\mathsf{P}_1^{'}, \ldots, \mathsf{P}_{k-1}^{'}]$. Computing the FIM, we get:

$$J = n \left( \operatorname{diag}\left( \left[ \frac{1}{\mathsf{P}_1^{'}}, \ldots, \frac{1}{\mathsf{P}_{k-1}^{'}} \right] \right) + \frac{1}{\mathsf{P}_k^{'}} \mathbf{1}_{k-1} \mathbf{1}_{k-1}^{\top} \right) \tag{45}$$

where $\mathbf{1}_{k-1}$ is a vector of all ones of size $k-1$. Using Sherman-Morrison-Woodbury formula Petersen and Pedersen [2008], we compute the the CRB matrix, i.e., $J^{-1}$:

$$C(\mathsf{P}^{'}) = J^{-1} = \frac{1}{n} \left( \operatorname{diag}(\mathsf{P}^{'}) - \mathsf{P}^{'} (\mathsf{P}^{'})^{\top} \right) . \tag{46}$$

**Moment constraint** ( $\mathbb{E}_{\mathsf{P}'}[\phi(x)] = \mu$, where $x \in \mathcal{S}$ and $\phi : \mathcal{S} \to \mathbb{R}^m$, and $\sum_i^k \mathsf{P}_i^{'} = 1$): By Lemma 16, the CRB matrix $C$ is given by:

$$C(\mathsf{P}^{'}) = {C_R}^{-1} - {C_R}^{-1} F (F^T {C_R}^{-1} F)^{-1} F^{\top} {C_R}^{-1} . \tag{47}$$

Here, we use $C_R$ to denote the CRB in Equation (46), incorporating only the regular probability constraint. The constraints, $f(\mathsf{P}^{'}) = \mathbb{E}_{\mathsf{P}'}[\phi(x)] - \mu = 0 \in \mathbb{R}^m$, can be rewritten as:

$$\begin{aligned} f(\mathsf{P}^{'}) &= \sum_{i=1}^{k-1} \phi(x_i) \mathsf{P}_i^{'} + \phi(x_k) \mathsf{P}_k^{'} - \mu = 0 \\ &\overset{(*)}{=} \sum_{i=1}^{k-1} (\phi(x_i) - \phi(x_k)) \mathsf{P}_i^{'} + \phi(x_k) - \mu = 0 \\ &= \sum_{i=1}^{k-1} a_i \mathsf{P}_i^{'} + b = 0 \quad (m \text{ constraints}) \end{aligned} \tag{48}$$

where (*) follows from the constraint $\mathsf{P}_k^{'} = 1 - \sum_{i=1}^{k-1} \mathsf{P}_i^{'}$, $a_i = \phi(x_i) - \phi(x_k) \in \mathbb{R}^m$, and $b = \phi(x_k) - \mu \in \mathbb{R}^m$.

Let the constraint set $f(\mathsf{P}^{'}) = [f_1, \ldots, f_m]$, then the gradient matrix F of size $m \times (k-1)$ is given by:

$$\begin{bmatrix} \frac{df_1}{d\mathsf{P}_1^{'}} & \cdots & \frac{df_1}{d\mathsf{P}_{k-1}^{'}} \\ \vdots & \ddots & \vdots \\ \frac{df_m}{d\mathsf{P}_1^{'}} & \cdots & \frac{df_m}{d\mathsf{P}_{k-1}^{'}} \end{bmatrix} = \begin{bmatrix} a_{1,1} & \ldots & a_{1,k-1} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \ldots & a_{m,k-1} \end{bmatrix} \tag{49}$$

where $a_i = [a_{1,i}, \ldots, a_{m,i}]$. Then,

$$\begin{aligned} \mathbb{S} \triangleq F^T {C_R}^{-1} F &= \frac{1}{n} \left( \sum_{i=1}^{k-1} a_i a_i^{\top} \mathsf{P}_i^{'} - \left( \sum_{i=1}^{k-1} a_i \mathsf{P}_i^{'} \right) \left( \sum_{j=1}^{k-1} a_j \mathsf{P}_j^{'} \right) \right) \\ &= \frac{1}{n} \operatorname{Cov}(a) \end{aligned} \tag{50}$$

where $a$ is a vector that takes the value $a_i$ with probability $\mathsf{P}_i^{'}$, and $\mathrm{Cov}(a)$ is its covariance matrix. Also, $C_R^{-1}F = \frac{1}{n}(\mathrm{diag}(\mathsf{P}_i^{'}) - \mathsf{P}^{'}(\mathsf{P}^{'})^{\top})F = \frac{1}{n}(\mathbb{P} - \mathsf{P}^{'}\overline{a}^{\top})$, where $\mathbb{P}$ is the $(k-1)\times m$ matrix with rows $\mathsf{P}_i^{'}a_i$ and $\overline{a} = \sum_{i=1}^{k} a_i \mathsf{P}_{\mathrm{t},i}^{s,a}$. Thus,

$$C(\mathsf{P}^{'}) = {C_R}^{-1} - ({C_R}^{-1}F)\mathbb{S}^{-1}({C_R}^{-1}F)^{\top}$$

and

$$\mathrm{Var}[\mathsf{P}_i^{'}] \geq ({C_R}^{-1})_{ii} - ({C_R}^{-1}F)_i^{\top}\mathbb{S}^{-1}(C_R^{-1}F)_i$$

where $({C_R}^{-1})_{ii} = \frac{\mathsf{P}_i^{'}(1-\mathsf{P}_i^{'})}{n}$ and $({C_R}^{-1}F)_i^{\top} = \frac{\mathsf{P}_i^{'}}{n}(a_i - \overline{a})$ is the $i^{th}$ row of $({C_R}^{-1}F)$. Then,

$$\begin{aligned}
\Delta_i &:= ({C_R}^{-1}F)_i^{\top}\mathbb{S}^{-1}(C_R^{-1}F)_i \\
&= \left(\frac{\mathsf{P}_i^{'}}{n}(a_i - \overline{a})\right)^{\top}\mathbb{S}^{-1}\left(\frac{\mathsf{P}_i^{'}}{n}(a_i - \overline{a})\right) \\
&= \frac{(\mathsf{P}_i^{'})^2}{n^2}(a_i - \overline{a})^{\top}\mathbb{S}^{-1}(a_i - \overline{a})\,. \qquad (51)
\end{aligned}$$

Since $\mathbb{S}$ is positive definite (assuming $\mathrm{Cov}(a)$ is full rank), $\Delta_i \geq 0$, indicating that the lower bound on the variance of the estimator is reduced with the incorporation of the moment constraint.

We remark that, if $\mathrm{Cov}(a)$ is not a full rank matrix, then $\mathbb{S}^{-1}$ can be replaced with the Moore-Penrose pseudoinverse $\mathbb{S}^{\dagger}$, i.e., $C(\mathsf{P}^{'}) = {C_R}^{-1} - ({C_R}^{-1}F)\mathbb{S}^{\dagger}({C^R}^{-1}F)^{\top}$. The matrix $\Delta := ({C_R}^{-1}F)\mathbb{S}^{\dagger}({C_R}^{-1}F)^{\top}$ is positive semi-definite (PSD), so

$$C(\mathsf{P}^{'}) = {C_R}^{-1} - \Delta \preceq {C_R}^{-1}\,,$$

so the constrained matrix is less than or equal to the unconstrained one in the PSD ordering, indicating that the reduction in the lower bound on the variance still holds.

□

# G ALGORITHMS

To estimate the optimal policy of the worst-case value function as in Equation (3), we utilize the robust value iteration algorithm presented in Algorithm 1, where $\sigma_{\mathcal{P}_s^a}(V) \triangleq \min_{\mathsf{P}\in\mathcal{P}_s^a} \mathsf{P}^{\top}V$ is the support function of $V$ on $\mathcal{P}_s^a$. Moreover, Algorithm 1 can also be used to solve for the optimal policy of the value function for the non-robust, if we set $\mathcal{P}_s^a = \{\mathsf{P}^{s,a}\}$, in which case $\sigma_{\mathcal{P}_s^a}(V) = {\mathsf{P}^{s,a}}^{\top}V$

**Algorithm 1** Robust Policy Estimation

**Input**: $\gamma, V_0(s) = 0, \forall s, T$

**for** $t = 0, 1, ..., T-1$ **do**
  **for** all $s \in \mathcal{S}$ **do**
    $V_{t+1}(s) \leftarrow \max_{a\in\mathcal{A}} \left\{r(s,a) + \gamma\sigma_{\mathcal{P}_s^a}(V_t)\right\}$
  **end for**
**end for**
**for** $s \in \mathcal{S}$ **do**
  $\pi_T(s) \leftarrow \arg\max_{a\in\mathcal{A}} \left\{r(s,a) + \gamma\sigma_{\mathcal{P}_s^a}(V_T)\right\}$
**end for**
**return** $V_T, \pi_T$

Based on a similar approach as in policy estimation, any policy $\pi$ can be evaluated on a general uncertainty set, as shown in Algorithm 2, where in the non-robust we set $R = 0$.

Table 2: A description of the source and target domain transition kernels for toy-text environments.

| Transition kernel | Description |
|---|---|
| $\mathsf{P}_{\mathrm{s}}^{s,a}(s_{\text{intended}})$ | The agent moves to the state in the intended direction with probability $$(1-\alpha)\left(r_s + 2\frac{1-r_s}{|\mathcal{S}|}\right).$$ |
| $\mathsf{P}_{\mathrm{s}}^{s,a}(s_{\text{opposite}})$ | The agent may move to the state opposite the intended direction with probability $$\alpha\left(r_s + 2\frac{1-r_s}{|\mathcal{S}|}\right).$$ |
| $\mathsf{P}_{\mathrm{s}}^{s,a}(s_{\text{other}})$ | The agent may move to all other states with probability $$\frac{1-r_s}{|\mathcal{S}|}.$$ |
| $\mathsf{P}_{\mathrm{t}}^{s,a}(s_{\text{intended}})$ | The agent moves to the state in the intended direction with probability $$\alpha\left(r_t + 2\frac{1-r_t}{|\mathcal{S}|}\right).$$ |
| $\mathsf{P}_{\mathrm{t}}^{s,a}(s_{\text{opposite}})$ | The agent may move to the state opposite the intended direction with probability $$(1-\alpha)\left(r_t + 2\frac{1-r_t}{|\mathcal{S}|}\right).$$ |
| $\mathsf{P}_{\mathrm{t}}^{s,a}(s_{\text{other}})$ | The agent may move to all other states with probability $$\frac{1-r_t}{|\mathcal{S}|}.$$ |

**Algorithm 2** Robust Policy Evaluation

**Input**: $\pi, \gamma, V_0(s) = 0, \forall s, T$

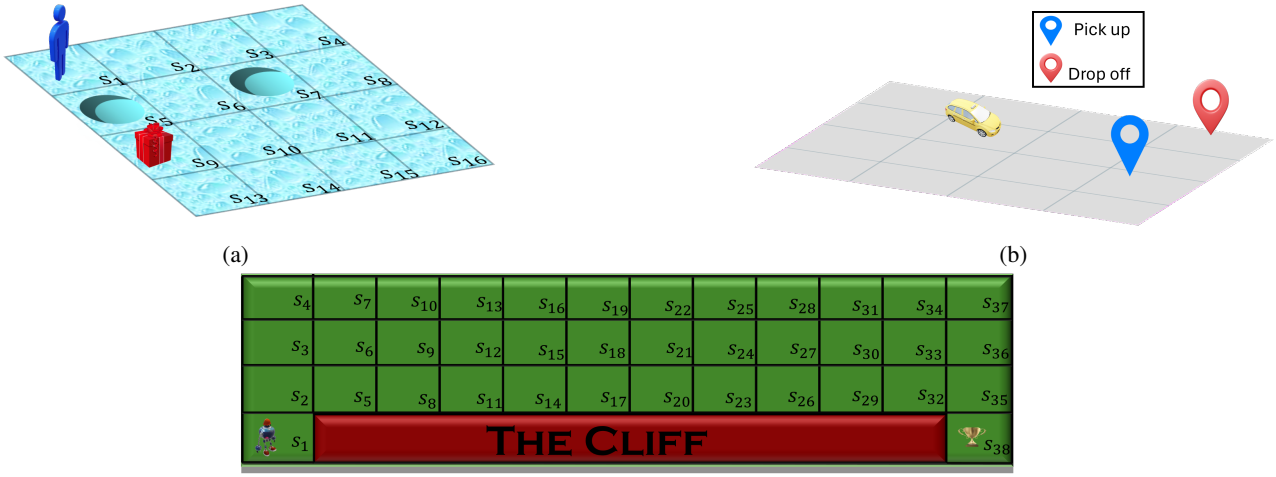
**for** $t = 0, 1, ..., T-1$ **do**
  **for** all $s \in \mathcal{S}$ **do**
    $V_{t+1}(s) \leftarrow \mathbb{E}_\pi[r(s, A) + \gamma \sigma_{\mathcal{P}_s^A}(V_t)]$
  **end for**
**end for**
**return** $V_T$

# H DESCRIPTION OF ENVIRONMENTS

## H.1 TOY TEXT ENVIRONMENTS

**Frozen Lake Environment**: This environment is depicted in Figure 5a. We consider a $4 \times 4$ frozen lake grid, consisting of a start state, an end state, two hole states, and one prize state, where each square represents a state. The main goal is to traverse the frozen lake from the start state to the end state without falling into holes. The agent can take four actions: right, left, up and down. The agent loses '-1' if it steps on a hole, receives '+5' for stepping on the prize state, and '-0.04' for any other frozen state. Transitions to the next state occur with a certain probability upon taking an action.

**Cliff Walking Environment**: The goal of the agent in the cliff walking environment is to reach the end (prize) state, starting




(a) (b)

(c)

Figure 5: A depiction OpenAI Gym toy text environments:**(a)** Frozen Lake environment **(b)** Taxi environment**(c)** Cliff Walking environment

from the first state while avoiding the cliff. The environment consists of 38 states: a start state, an end state, and 36 other states. The agent incurs a penalty of '-100' if it steps on the cliff and receives '+50' upon reaching the end state. For all other states, the agent receives a reward of '-1'. A schematic of this environment is depicted in Figure 5c.

**Taxi Environment**: In the taxi environment, the goal for the agent (taxi) is to navigate through the grid using movement actions (up, down, left, right), locate the passenger, pick up the passenger, and later drop off the passenger at the destination using pick-up and drop-off actions, respectively. In our experiment, we consider a $6 \times 6$ grid with fixed pick-up and drop-off locations, as shown in Figure 5b. The passenger can be located either at the pick-up or drop-off locations, or in the taxi, resulting in a total of 108 states and 6 possible actions. The agent is penalized '-1' for each movement action taken. Wrongful pick-up and drop-off actions incur a penalty of '-10'. The agent receives a reward of '+20' for successfully picking up the passenger from the designated location and '+50' for successfully dropping off the passenger, signifying task completion.

A detailed description of the transition kernels for both the source and target domains is provided in Table 2. We define $s_{intended}$ as the next state the agent is expected to move to when it takes action $a$ starting from state $s$. Conversely, $s_{opposite}$ refers to the state opposite to $s_{intended}$. For example, consider the frozen lake environment in the source domain. If the agent is at state $s_2$ and takes action $a$ = 'right,' it will move to $s_{intended} = s_3$ with probability $\mathsf{P}_{\mathrm{s}}^{s,a}(s_{intended})$, to $s_{opposite} = s_1$ with probability $\mathsf{P}_{\mathrm{s}}^{s,a}(s_{opposite})$, and to any other state $s_{other}$ with probability $\mathsf{P}_{\mathrm{s}}^{s,a}(s_{other})$. For the frozen lake environment, we use $r_s = 0.3$, $r_t = 0.8$, and $\alpha = 0.7$. For cliff walking, we set $r_s = 0.8$, $r_t = 0.2$, and $\alpha = 0.3$. For the taxi environment, we choose $r_s = 0.4$, $r_t = 0.8$, and $\alpha = 0.2$. To ensure validity, we normalize each $\mathsf{P}_{\mathrm{s}}^{s,a}$ and $\mathsf{P}_{\mathrm{t}}^{s,a}$ so that they sum up to 1.

### H.2 CLASSIC CONTROL PROBLEMS

We consider three classical control problems from OpenAIGym: Cart pole, Acrobot, and Pendulum. We have made some modifications to the environments to suit our problem setup.

**Cart Pole Environment**: The goal is to balance a pole attached upright to the cart by applying forces to the left or right of the cart Barto et al. [1983]. The action space consists of two actions: right and left. The observation space is continuous and consists of a tuple of four quantities: Cart Position, Cart Velocity, Pole Angle, Pole Angular Velocity. Their corresponding ranges are $([-4.8, 4.8], (-\infty, \infty), [-24^o, 24^o], (-\infty, \infty))$. We discretize the observation space, with each combination of values representing a distinct state. The new ranges for each quantity are $([-4.8, 4.8], [-0.5, 0.5], [-24^o, 24^o], [-5, 5])$, and the number of discrete values is $(4, 4, 5, 3)$, respectively. A reward of '25' is received if the pole angle is 0 and the cart velocity is between -0.17 and +0.17. An additional reward of '25' is granted if the cart position is between -1.6 and 1.6. When the pole angle is between $(-12^o, 12^o)$ and the cart velocity is between -0.17 and +0.17, a reward of '10' is received. For all other cases, no rewards are received. The Cart Pole environment is depicted in Figure 6a.

Table 3: A description of the source and target domain transition kernels for classic control environments.

| Transition kernel | Description |
|---|---|
| $\mathsf{P}_{\mathsf{s}}^{s,a}(s_{\text{random}_1})$ | The agent moves to a random state $s_{\text{random}_1}$ with probability $\alpha r_s$. |
| $\mathsf{P}_{\mathsf{s}}^{s,a}(s_{\text{random}_2})$ | The agent moves to a random state $s_{\text{random}_2}$ with probability $(1-\alpha) r_s$. |
| $\mathsf{P}_{\mathsf{s}}^{s,a}(s_{\text{other}})$ | The agent may move to all other states with probability $\frac{r_s}{\lvert\mathcal{S}\rvert}$. |
| $\mathsf{P}_{\mathsf{t}}^{s,a}(s_{\text{random}_1})$ | The agent moves to a random state $s_{\text{random}_1}$ with probability $(1-\alpha) r_t$. |
| $\mathsf{P}_{\mathsf{t}}^{s,a}(s_{\text{random}_2})$ | The agent moves to a random state $s_{\text{random}_2}$ with probability $\alpha r_t$. |
| $\mathsf{P}_{\mathsf{t}}^{s,a}(s_{\text{other}})$ | The agent may move to all other states with probability $\frac{r_t}{\lvert\mathcal{S}\rvert}$. |

**Cart Pole Environment (LDS scenario)**: For each $(s, a)$, the source and target transition kernels are softmaxes over next–state features $\phi(s') \in \mathbb{R}^4$ (constructed from *Cart Position, Cart Velocity, Pole Angle, Pole Angular Velocity*), the source and target transition kernels for each $(s, a) \in \mathcal{S} \times \mathcal{A}$ are modeled as follows:

$$\mathsf{P}_s(s' \mid s, a) = \frac{\exp\{\theta_s^{(s,a)\top}\phi(s')\}}{\sum_{\bar{s}} \exp\{\theta_s^{(s,a)\top}\phi(\bar{s})\}}, \qquad \mathsf{P}_t(s' \mid s, a) = \frac{\exp\{\theta_t^{(s,a)\top}\phi(s')\}}{\sum_{\bar{s}} \exp\{\theta_t^{(s,a)\top}\phi(\bar{s})\}}.$$

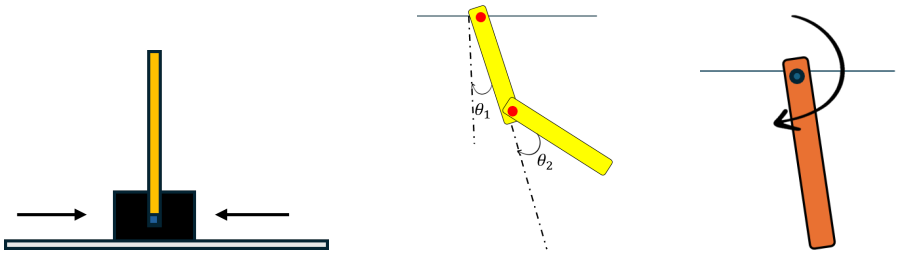
With $\theta_s^{(s,a)} = [\gamma_{sa};\, \alpha_{sa}^{(s)}]$ and $\theta_t^{(s,a)} = [\gamma_{sa};\, \alpha_{sa}^{(t)}]$ with the first $\mathsf{d} - \mathsf{d}_0$ coordinates shared ($\gamma_{sa} \in \mathbb{R}^{\mathsf{d}-\mathsf{d}_0}$) and the last $\mathsf{d}_0 = 2$ coordinates private ($\alpha_{sa}^{(s)}, \alpha_{sa}^{(t)} \in \mathbb{R}^2$); hence the source–target shift in the logits (and thus in the kernels) lies in a 2-dimensional subspace.

**Acrobot Environment**: The Acrobot problem, introduced in Sutton [1995], features a chain with two links, one fixed and one free end, as depicted in Figure 6b. The objective is to swing the free end of the chain above a certain height, starting from a downward hanging position. The joint between the two links is actuated, allowing for the application of torque to achieve the desired swing. The available actions involve applying a torque to the actuated joint, with options of $\{1, 0, -1\}$ Newton-meter (Nm). Observations in this environment consist of tuples of 6 quantities: $(\cos\theta_1, \sin\theta_1, \cos\theta_2, \sin\theta_2, V_{\theta_1}, V_{\theta_2})$. Here, $\theta_1$ and $\theta_2$ represent the angles of the first link w.r.t. the normal direction and the angle of the second link relative to the first link respectively. The terms $V_{\theta_1}$ and $V_{\theta_2}$ denote the angular velocities of $\theta_1$ $\theta_2$, respectively.

For the observation space, we use a tuple of four quantities $(\cos\theta_1, \cos\theta_2, V_{\theta_1}, V_{\theta_2})$, with a range of $([-1, 1], [-1, 1], [-4\pi, 4\pi], [-9\pi, 9\pi])$, and the number of discrete values $(6, 6, 2, 2)$, respectively. For the reward, we use the quantity $\phi = -\cos(\theta_1) - \cos(\theta_1 + \theta_2)$ to determine different reward levels. Specifically, for $\phi$ falling within predefined ranges ($\phi \geq 1$, $0.5 \leq \phi \leq 1$, $0.25 \leq \phi \leq 0.5$, $0 \leq \phi \leq 0.25$), rewards of 20, 15, 10, 5 are granted, respectively.

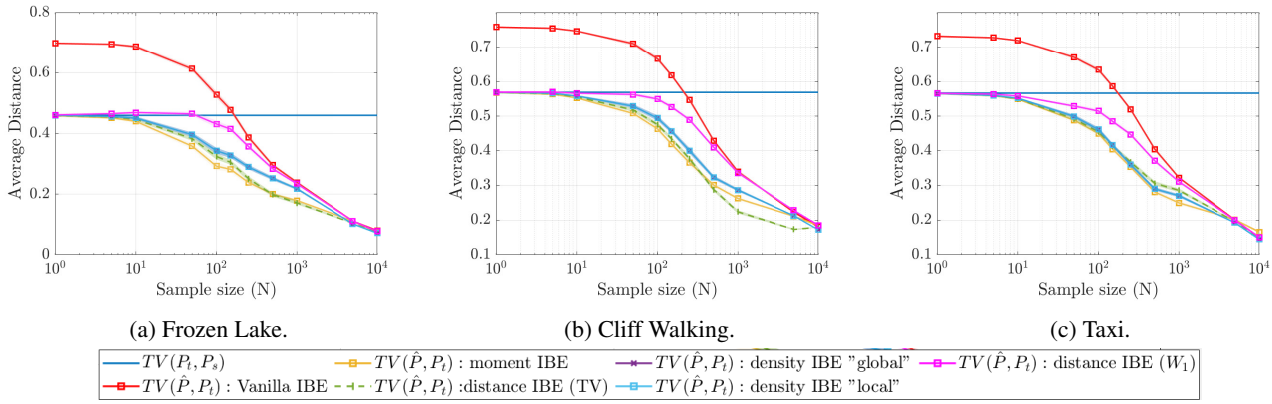
**Pendulum Environment**: The Pendulum problem is another classic control environment that focuses on the dynamic control of an inverted pendulum. Illustrated in Figure 6c, the system comprises a pendulum affixed at one end to a fixed point. The objective is to apply torque and stabilize the pendulum's center of gravity directly above the pivot, swinging it into an upright position. The action space is continuous, within the range of $[-2, 2]$ Nm. Observations consist of arrays of 3 quantities representing the $x$ and $y$ coordinates of the free end of the pendulum, and the angular velocity, with ranges of $[-1, 1]$, $[-1, 1]$, and $[-8, 8]$, respectively. In our experiment, we discretize the action space into five torque levels $\{-2, -1, 0, 1, 2\}$ Nm. The observation space is segmented into 240 unique states by discretizing the pendulum angle $[-\pi, \pi]$ rad into 12 segments and the angular velocity ranging $[-10, 10]$ rad/s into 20 segments. Rewards are structured to prioritize stabilizing the pendulum near the upright position. Maintaining the pendulum within $[-1, 1]$ of vertical and sustaining a low angular velocity of $[-1, 1]$ earns 100 points. A reward of 50 points is granted for keeping the angle within $[-0.5, 0.5]$, and 10 points are awarded for all other states.

Table 3 provides a description of the transition kernels for both the source and target domains in the classic control problems. For each $(s, a)$ pair, we generate two random states, $s_{random_1}$ and $s_{random_2}$. Taking the source domain as an example, the agent transitions to $s_{random_1}$ and $s_{random_2}$ with probabilities $\mathsf{P}_{\mathsf{s}}^{s,a}(s_{random_1})$ and $\mathsf{P}_{\mathsf{s}}^{s,a}(s_{random_2})$, respectively. For all other states, $s_{other}$, the agent transitions with probability $\mathsf{P}_{\mathsf{s}}^{s,a}(s_{other})$. We set $r_s = 0.6$, $r_t = 0.7$, and $\alpha = 0.2$ for all environments. To ensure validity, each $\mathsf{P}_{\mathsf{s}}^{s,a}$ and $\mathsf{P}_{\mathsf{t}}^{s,a}$ is normalized to sum to 1.




(a) (b) (c)

Figure 6: A depiction OpenAI Gym classic control problems: **(a)** Cart Pole. **(b)** Acrobot. **(c)** Pendulum.




(a) Frozen Lake. (b) Cliff Walking. (c) Taxi.

Figure 7: Average total variation distance between the estimated and the target domain transition kernels over 20 runs as a function of sample size for the toy text environments.

# I ADDITIONAL EXPERIMENTS

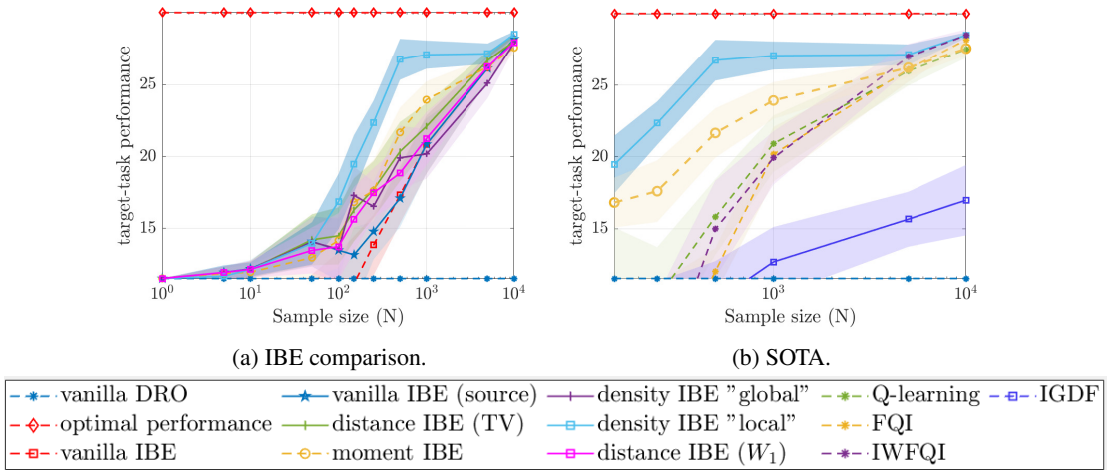
## I.1 CONVERGENCE OF THE ESTIMATORS

We experimentally verify the convergence of our estimators to the target domain transition kernels, as theoretically established in Proposition 4. Figure 7 shows the average distance over all $(s, a)$ pairs between our IBE estimates $\widehat{\mathsf{P}}^{s,a}$ and the true target transition kernels $\mathsf{P}_{\mathrm{t}}^{s,a}$, i.e., $\frac{1}{|\mathcal{S}||\mathcal{A}|}\sum_{(s,a)} \|\widehat{\mathsf{P}}^{s,a} - \mathsf{P}_{\mathrm{t}}^{s,a}\|_{TV}$, averaged over 20 runs, versus the sample size for different toy text environments, including Frozen Lake, Cliff Walking, and Taxi. We also report the average distance between the source and target domain transition kernels. The results indicate that the IBE converges asymptotically to the target domain kernels as $N \to \infty$.

The results demonstrate that our estimators can converge to the true target transition kernels under different types of side of information.

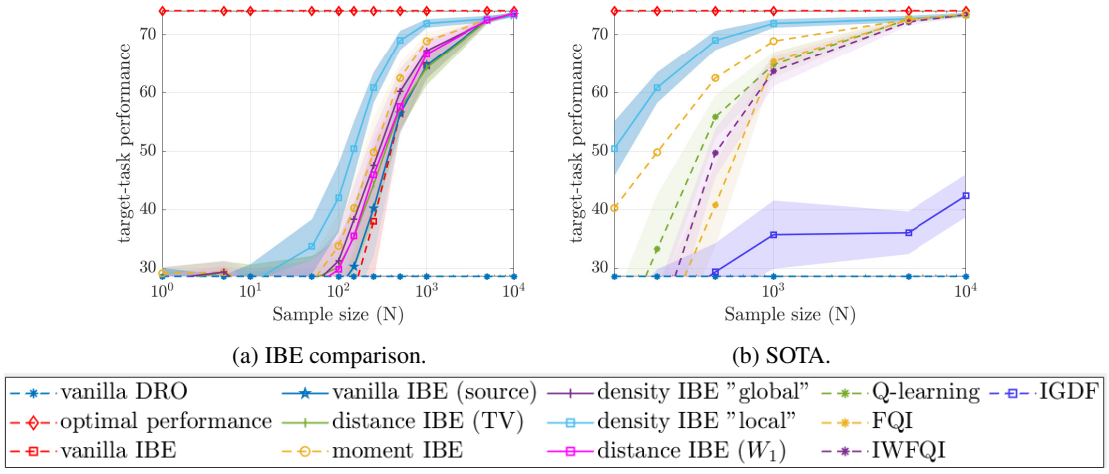
## I.2 NON-ROBUST SETTING

In this section, we provide the performance results for the omitted environments for the non-robust setting. Under this setting, we estimation the IBE policies using Algorithm 1, where $R = 0$. Then, the estimated policies, including the IBE policies and the state-of-the-art (SOTA) ones, are evaluated using Algorithm 2, using the same radius. Figures 8, 9, 11, 10, and 12 demonstrate the target task performance, the average value function, i.e., $\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} V^{\pi}(s)$, as a function of the sample size $N$ for the non-robust scenario. Each figure shows the IBE with different side information (left panel) and a comparison with SOTA methods (right panel).



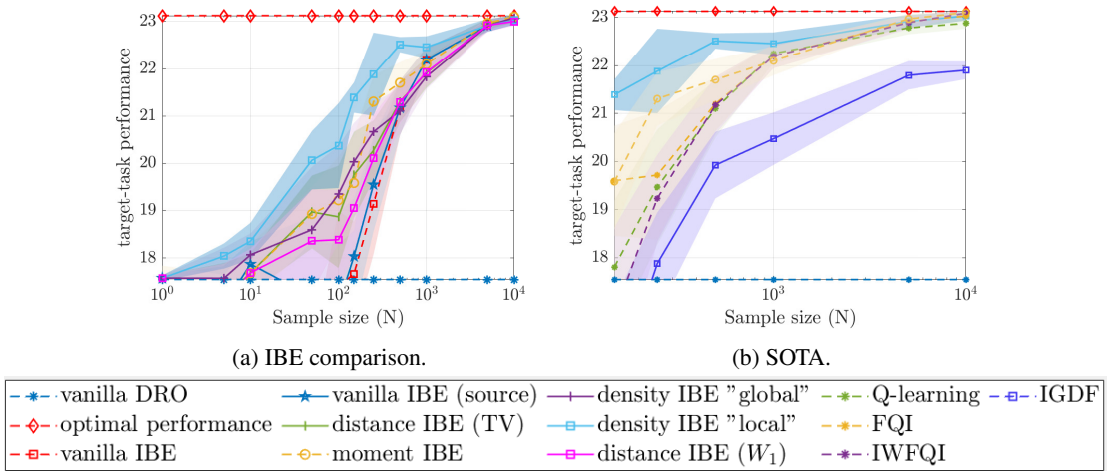

(a) IBE comparison. (b) SOTA.

Figure 8: Target domain performance for the non-robust setting averaged over 20 runs as a function of sample size for Cliff Walking environment, and the corresponding 95% confidence intervals.



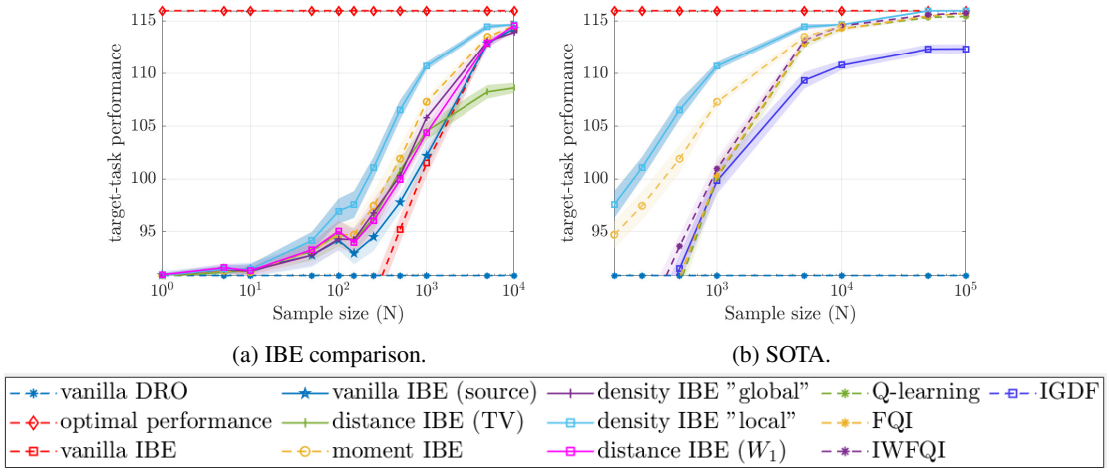

(a) IBE comparison. (b) SOTA.

Figure 9: Target domain performance for the non-robust setting averaged over 20 runs as a function of sample size for Taxi environment, and the corresponding 95% confidence intervals.



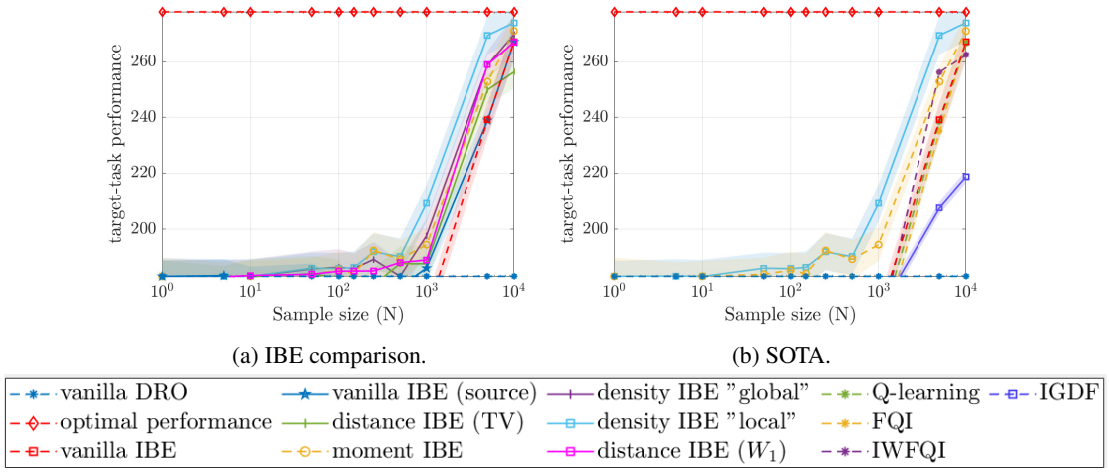

(a) IBE comparison. (b) SOTA.

Figure 10: Target domain performance for the non-robust setting averaged over 20 runs as a function of sample size for Frozen Lake environment, and the corresponding 95% confidence intervals.



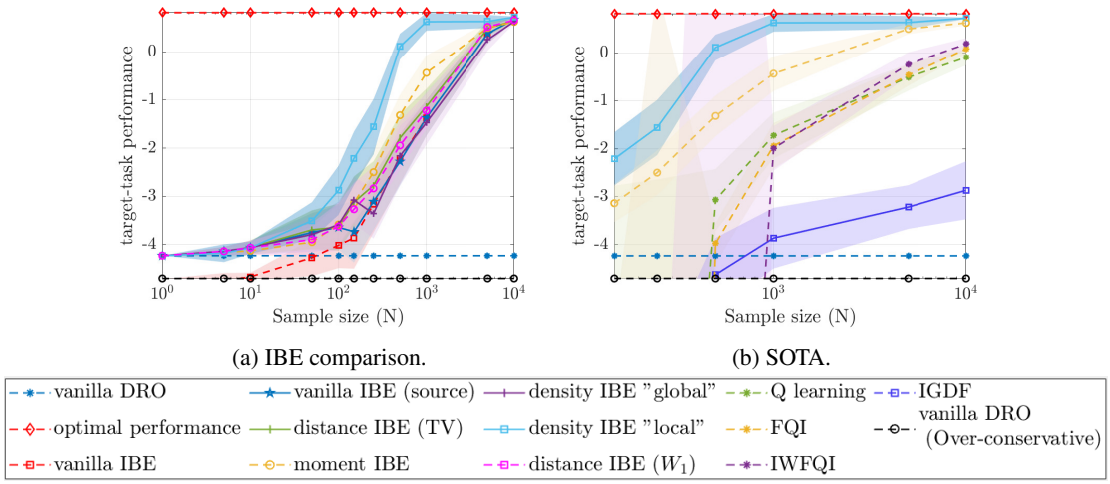

(a) IBE comparison. (b) SOTA.

Figure 11: Target domain performance for the non-robust setting averaged over 20 runs as a function of sample size for Acrobot environment, and the corresponding 95% confidence intervals.



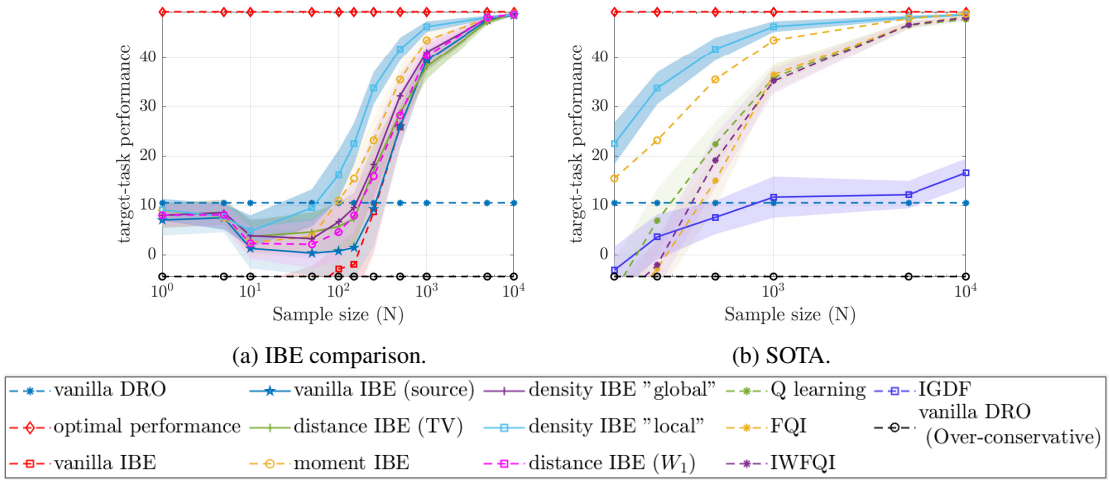

(a) IBE comparison. (b) SOTA.

Figure 12: Target domain performance for the non-robust setting averaged over 20 runs as a function of sample size for Pendulum environment, and the corresponding 95% confidence intervals.




(a) IBE comparison. (b) SOTA.

Figure 13: Target domain performance for the robust setting averaged over 20 runs as a function of sample size for Cliff Walking environment, and the corresponding 95% confidence intervals.




(a) IBE comparison. (b) SOTA.

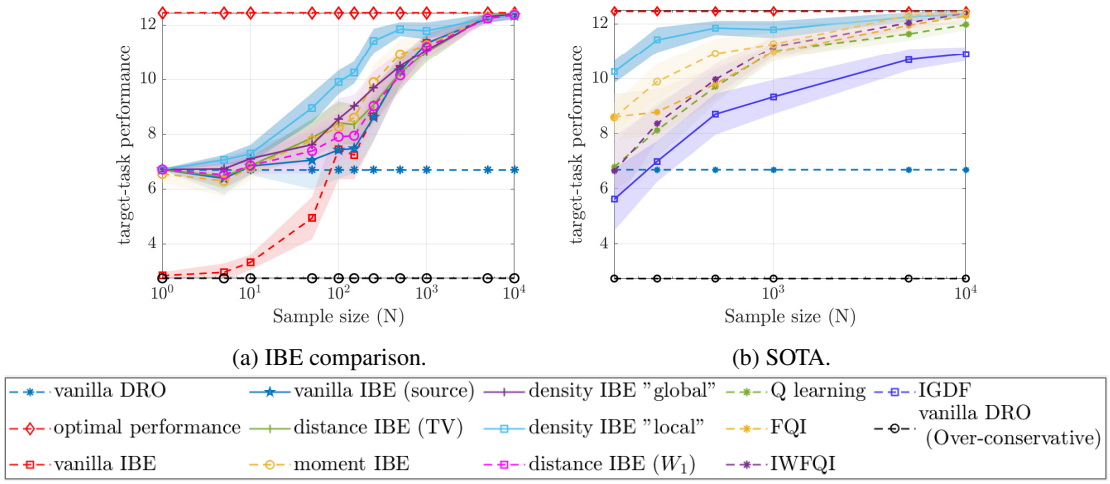
Figure 14: Target domain performance for the robust setting averaged over 20 runs as a function of sample size for Taxi environment, and the corresponding 95% confidence intervals.

### I.3 ROBUST SETTING

In this part, we present the IBE performance on the target domain for the robust scenario. Algorithm 1 is used to estimate the IBE polices, where we choose $R = 0.1$. All polices, including the IBE polices and the state-of-art (SOTA) polices are evaluated on the target domain under uncertainty using Algorithm 2, using the radius $R = 0.1$. Figures 13, 14, 15, 16, and 17 show the target task performance, namely, the average value function, i.e., $\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} V^{\pi}(s)$, as a function of the sample size $N$ for the robust scenario. Each figure shows the IBE with different side information (left panel) and a comparison with the SOTA methods (right panel).

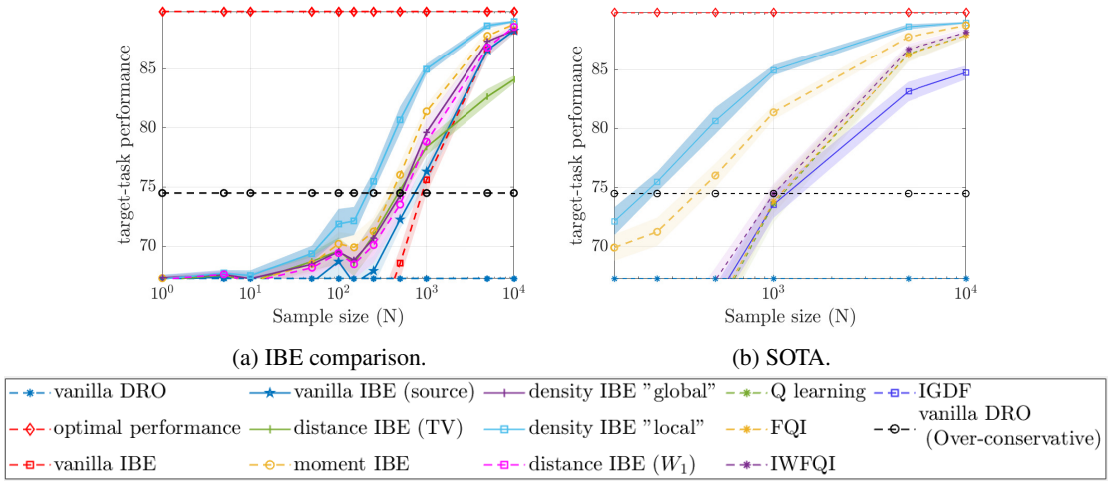
### I.4 EVALUATION ERROR: EXPERIMENTAL VERIFICATION

In this experiment, we verify the theoretical results of the evaluation error bound presented in Theorem 2 as well as the convergence results of Corollary 3. We consider the frozen lake environment, following the same settings and sampling procedure as in section 6. The policy is estimated by Algorithm 1 using the transition kernels estimated using the approaches outlined in Table 1. Then, each policy is evaluated on the target domain environment. We consider both non-robust and robust scenarios. Figure 18 and 19 show the evaluation (EV) error (LHS of Equation (7)) and the bound (RHS of Equation




(a) IBE comparison. (b) SOTA.

Figure 15: Target domain performance for the robust setting averaged over 20 runs as a function of sample size for Frozen Lake environment, and the corresponding 95% confidence intervals.




(a) IBE comparison. (b) SOTA.

Figure 16: Target domain performance for the robust setting averaged over 20 runs as a function of sample size for Acrobot environment, and the corresponding 95% confidence intervals.

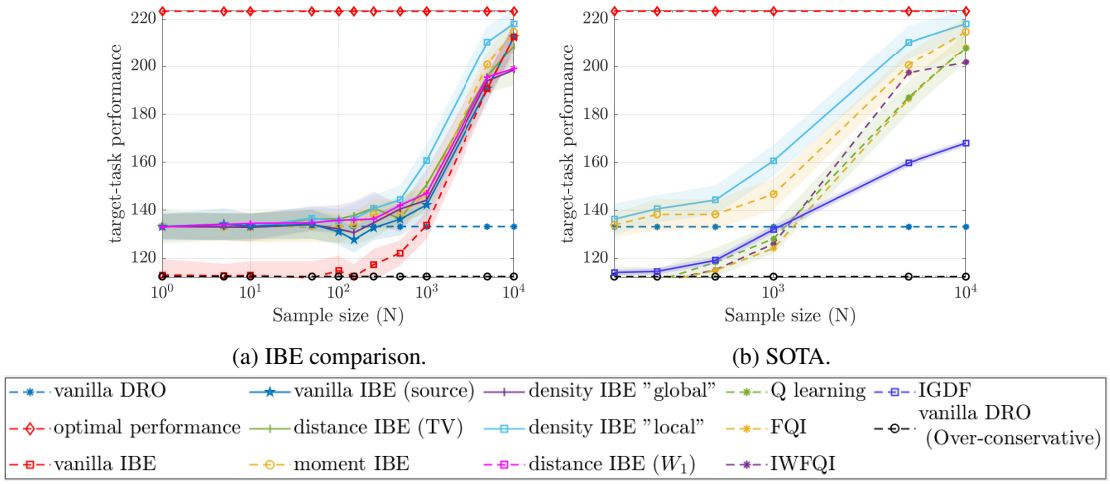



Figure 17: Target domain performance for the robust setting averaged over 20 runs as a function of sample size for Pendulum environment, and the corresponding 95% confidence intervals.

(7)) as a function of the sample size for each estimation approach for the non-robust and robust settings, respectively. As observed, the evaluation error is always upper bounded by the computed bound for all sample sizes. We also observe that the error and the bound decrease as the sample size increases which verifies the convergence results in Corollary 3.

### I.5 THE GAIN OF PRIOR INFORMATION

In this experiment, we investigate the information gain from prior information about the estimated kernels. In particular, given a $k \times k$ FIM $J(\mathsf{P}_{\mathrm{t}}^{s,a})$, we seek to measure the amount of information that $n$ i.i.d observations provide about the kernel $\mathsf{P}_{\mathrm{t}}^{s,a}$, given some prior information. Therefore, we consider the following optimization problem:

$$\min_{q \in \Delta(\mathcal{S})} \operatorname{trace}(J(q)) \quad \text{s.t.} \quad \mathbb{E}^q[x^j] \leq c_j, 1 \leq j \leq M. \tag{52}$$

Here, we assume knowledge of the bounds $c_j$ first $M$ moments. Note that we examine the worst-case scenario for the information the FIM carries by considering the minimum. We compare two scenarios: (i) 'Without Constraints', in which we omit the expectation constraints, and (ii) 'With Constraints', where the constraints are included. Figure 20 shows the optimal value of the $\operatorname{trace}(J(\widehat{\mathsf{P}}_n^{s,a}))$ (*left*) and its inverse (*right*) on a logarithmic scale. As shown, including the constraints significantly increases the amount of information about the estimator leading to a much better estimate.

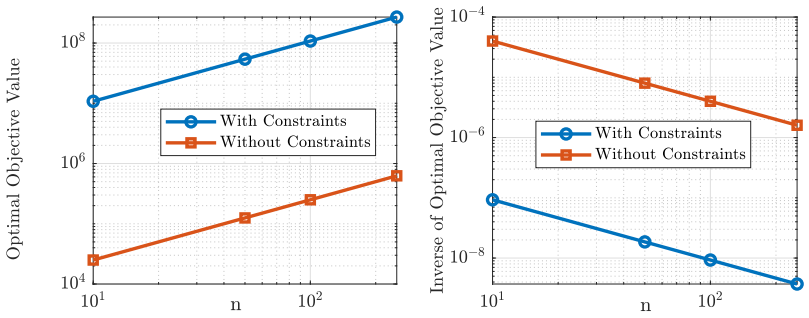



Figure 20: The trace of the FIM as a function of the sample size $n$ (left) and its inverse (right).

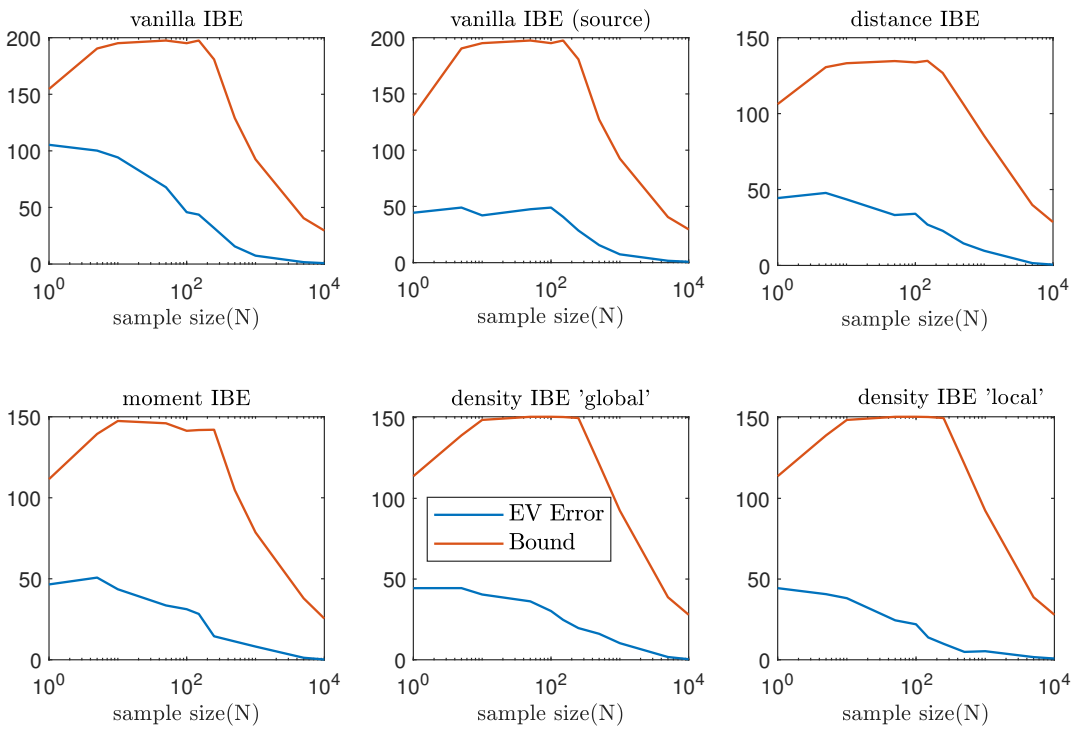



Figure 18: The Evaluation (EV) Error and the bound for each MLE method as a function of the sample size for the non-robust scenario

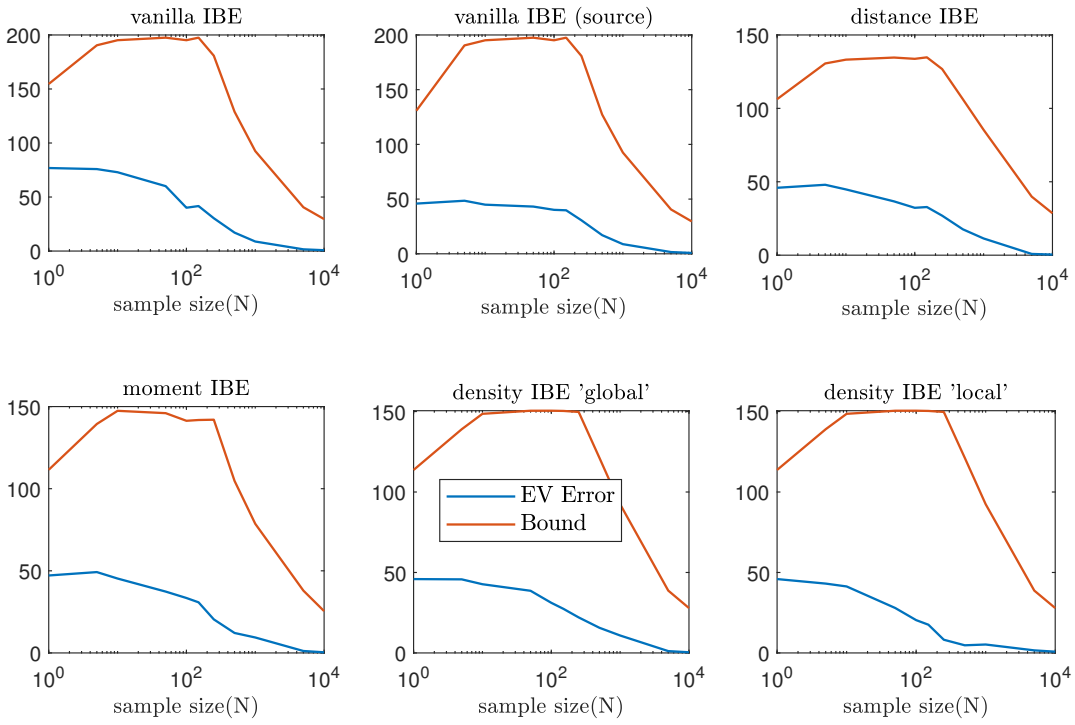



Figure 19: The Evaluation (EV) Error and the bound for each IBE method as a function of the sample size for the robust setting.