Balance Equation-based Distributionally Robust Offline Imitation Learning

Rishabh Agrawal
University of Southern California
rishabha.edu

Yusuf Alvi
University of Southern California
yalvi@usc.edu

Rahul Jain
University of Southern California
rahul.jain@usc.edu

Ashutosh Nayyar
University of Southern California
ashutosn@usc.edu

Abstract

Imitation Learning (IL) has proven highly effective for robotic and control tasks where manually designing reward functions or explicit controllers is infeasible. However, standard IL methods implicitly assume that the environment dynamics remain fixed between training and deployment. In practice, this assumption rarely holds where modeling inaccuracies, real-world parameter variations, and adversarial perturbations can all induce shifts in transition dynamics, leading to severe performance degradation. We address this challenge through Balance Equation-based Distributionally Robust Offline Imitation Learning, a framework that learns robust policies solely from expert demonstrations collected under nominal dynamics, without requiring further environment interaction. We formulate the problem as a distributionally robust optimization over an uncertainty set of transition models, seeking a policy that minimizes the imitation loss under the worst-case transition distribution. Importantly, we show that this robust objective can be reformulated entirely in terms of the nominal data distribution, enabling tractable offline learning. Empirical evaluations on continuous-control benchmarks demonstrate that our approach achieves superior robustness and generalization compared to state-of-the-art offline IL baselines, particularly under perturbed or shifted environments.

1 Introduction

Imitation Learning (IL) has become a cornerstone of modern AI, enabling agents to acquire complex skills in domains where manually engineering a reward function is difficult or impractical [51]. In applications ranging from robotic manipulation to autonomous driving, it is often far simpler to provide expert demonstrations than to design a dense, well-shaped reward signal that captures the desired behavior [37]. IL frameworks leverage these demonstrations to learn control policies directly, bypassing the challenges of reward engineering [5]. However, this success is predicated on the fragile assumption that the transition dynamics at test time match those under which the demonstrations were collected [43, 8]. In real-world systems, this rarely holds [32]. Physical parameters such as mass, friction, or motor response can drift over time [26]; unmodeled effects like actuator delay or unexpected contacts may arise [44, 18]; and the environment itself can undergo adversarial perturbations [14]. This creates a *robustness gap*, as policies effective in training environments degrade under perturbed dynamics [13].

A natural way to mitigate this robustness gap is to introduce interaction across multiple source environments, the target environment, or both during training. Under multi-source interaction, RIME

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Embodied and Safe-Assured Robotic Systems (E-SARS).

[8] learns a single policy that generalizes over varying dynamics by minimizing a Jensen–Shannon risk across expert policies, while ADAIL [25] leverages dynamics embeddings and domain-adversarial objectives to enforce invariance. In source-only interaction, AIRL [11] recovers reward functions that disentangle expert intent from environment-specific dynamics. For target-only adaptation, DYNAIL [24] fine-tunes policies by distinguishing source and target dynamics using limited target rollouts. Cross-domain alignment methods such as GAMA [20] and Cycle-Consistent IL [36] jointly interact with source and target domains to align MDPs through latent or state–action correspondences. Another line of work performs few-shot domain adaptation, where a source-trained policy is fine-tuned on a handful of target demonstrations to bridge domain gaps [47, 7]. While effective, these approaches still rely on interaction or target-domain data, which is impractical in settings like autonomous driving and surgical robotics, where rollouts are unsafe or costly [23], or when no expert exists under new dynamics.

Motivated by the discussed limits, we study *Distributionally Robust Offline Imitation Learning (DROIL)*. The goal of DROIL is to learn a *robust* policy, using only expert demonstrations from a single nominal environment *without any further interaction*, that remains effective under perturbations in transition dynamics. This enables IL agents to generalize across real-world dynamics variations while training entirely offline from one source of expert data, avoiding unsafe or costly rollouts. The setting is uniquely challenging: (i) standard IL methods such as Behavioral Cloning and Distribution Matching [33, 21] implicitly assume identical train–test dynamics, so robustness must be induced purely through the learning objective, as no further interactions are available to infer or adapt to perturbed transitions; (ii) Robustification under transition uncertainty yields a minimax problem whose inner maximization is challenging to solve, as it depends on unknown perturbed dynamics.

To address these challenges, we formulate robustness as a minimax optimization over an ambiguity set centered on the nominal transition dynamics and reformulate it in the f-divergence–based occupancy measure space. To ensure optimization only over stationary distributions, we impose a *Balance Equation* derived from Bellman flow consistency. By introducing a *triplet occupancy representation* and invoking strong duality, we transform the intractable adversarial maximization into a closed-form importance-weighting objective defined entirely under nominal expert data.

Our main contributions include: (i) *BE-DROIL*, the first distributionally robust imitation learning framework in the strict offline regime under Balance Equation constraint, using demonstrations from a single nominal environment; (ii) a novel theoretically grounded triplet-occupancy representation that eliminates explicit dependence on unknown dynamics, enabling robustness purely in the occupancy space; (iii) a scalable and practical alternating optimization algorithm based on converting the robust optimization into a tractable closed-form importance-weighting scheme under data collected from expert in nominal environment; and (iv) strong empirical robustness across continuous-control benchmarks under diverse transition perturbations.

2 Related Work

Offline Imitation Learning. Behavioral Cloning (BC) is a classical baseline in offline imitation learning [33], formulating imitation as supervised learning from expert trajectories that map states to actions. However, by disregarding environment dynamics and often operating with limited expert data, BC is prone to covariate shift and compounding error, which limit its generalization ability [38]. To incorporate transition structure, the *D̲istribution C̲orrection E̲stimation (DICE)* family [21, 19, 27] estimates stationary occupancy ratios without requiring explicit access to the behavior or learner policy, enabling off-policy evaluation and control under the stationary distribution.

Our formulation builds on this occupancy-based view but differs fundamentally: its importance weights stem from a robust inner maximization over transition uncertainty rather than policy mismatch. Recent studies [1–3] show that offline IL must satisfy a Markov balance relation linking the expert policy and transition dynamics. Still, BC, DICE, and Markov-balance methods all assume identical train–test dynamics and therefore fail under transition shifts.

Robust Reinforcement Learning. This line of work tackles transition uncertainty by optimizing worst-case returns over an ambiguity set of transition kernels. Classical approaches use rectangular uncertainty sets and min–max dynamic programming [28, 16], while later works extend to distributionally robust formulations based on total variation, f-divergence, or Wasserstein metrics [9, 48, 31]. These methods require reward feedback and often online rollouts to improve robustness. Offline variants such as robust fitted Q-iteration estimate worst-case value functions from fixed datasets using nominal-measure reformulations [29]. However, they still rely on known rewards, making them unsuitable for imitation learning, where only expert demonstrations are available. Our work builds on these robust foundations but reinterprets them for IL, without rewards or interaction, via a fully offline nominal-data formulation robust to transition uncertainty.

Robust Imitation Learning. Much of IL research focuses on robustness to *demonstration imperfections* (e.g., suboptimal or noisy labels) and stability in supervised imitation. Several methods learn from imperfect or negative demonstrations or denoise labels to mitigate covariate shift without altering dynamics [46, 41], while others improve stability through noise injection or corrective querying [22, 39]. Meta-IL and multi-task IL/IRL frameworks enhance adaptability across related tasks via shared experience but still assume consistent transition dynamics between training and deployment [10, 17, 50]. To address *dynamics mismatch*, domain randomization (DR) perturbs physical parameters such as mass or friction during training to improve sim-to-real transfer [32, 15], yet it requires a high-fidelity simulator, impractical in strictly offline IL where only nominal expert data exist. Recent methods enhance BC robustness under different failure modes: Wu et al. [45] enforce Lipschitz regularization to improve stability against input perturbations, while DRIL–DICE [40] mitigates covariate shift by introducing f-divergence regularization to align state–action distributions under nominal dynamics. However, both assume fixed train–test transition models. Among works explicitly considering transition uncertainty, RIME learns policies that generalize across families of MDPs via online rollouts for robustness estimation [8]. The most closely related method to ours is *Distributionally Robust Behavior Cloning (DRBC)*, which formulates behavior cloning under total-variation ambiguity sets around the transition model and optimizes for worst-case imitation loss [30]. However, DRBC maximizes over arbitrary distributions in the uncertainty set, many of which do not correspond to stationary occupancy measures induced by the expert policy under any admissible transition kernel. In contrast, our framework constrains the adversary to operate only over *valid stationary distributions* consistent with perturbed dynamics and generalizes beyond total variation to a broader class of f-divergence ambiguity sets, yielding a less pessimistic and more data-aligned robust objective.

3 Preliminaries

The Imitation Learning Problem. We consider an infinite-horizon discounted Markov decision process (MDP) $M = (\mathcal{S}, \mathcal{A}, T^o, r, \gamma, \mu)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $T^o(s' \mid s, a)$ represents the transition dynamics, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\gamma \in (0, 1)$ is the discount factor, and μ is the initial state distribution such that $s_0 \sim \mu$. A policy π defines a distribution over actions conditioned on the state, $\pi(a_t|s_t) = \mathbb{P}_\pi(A_t = a_t \mid S_t = s_t)$ for each time step t. The discounted occupancy measure induced by a policy π on the $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ under the nominal transition kernel T^o is defined as $d^\pi_{T^o}(s, a, s') = (1 - \gamma)\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \, \mathbf{1}_{s_t=s, a_t=a, s_{t+1}=s'}\right]$, where the expectation is taken over trajectories generated by $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim T^o(\cdot|s_t, a_t)$, with s_0 drawn from μ. The marginal state–action and state-only occupancy measures are then given by $d^\pi_{T^o}(s, a) = \sum_{s'} d^\pi_{T^o}(s, a, s')$ and $d^\pi_{T^o}(s) = \sum_{a,s'} d^\pi_{T^o}(s, a, s')$ respectively.

In the *offline imitation learning (OIL)* setting, an agent has access only to demonstration trajectories collected from an expert policy π_D, denoted as $D = \{(s_0, a_0), (s_1, a_1), \ldots\}$, without any additional interaction with the environment. No reward information is provided in D, and the goal is to learn a *policy* $\pi^\star_{oil}$ that mimics the expert's behavior as closely as possible using this static dataset. Formally, the general learning objective is expressed as [49]

$$\pi^\star_{oil} \in \arg\min_{\pi \in \Pi} \, \mathbb{E}_{s \sim d^{\pi_D}_{T^o}} \left[\mathcal{L}(\pi(\cdot|s), \pi_D(\cdot|s))\right], \tag{1}$$

where Π denotes the set of stationary stochastic policies, and $\mathcal{L}$ is a divergence or distance metric (e.g., mean squared error or KL divergence) measuring the gap between learner and expert policies.

Distributionally Robust Setup. We adopt the robust Markov decision process (RMDP) formulation [28], defined by the tuple $M_{\text{rob}} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma, \mu)$, which extends the standard MDP by introducing an uncertainty set $\mathcal{T}(\rho')$ over transition models. The uncertainty set is factorized across all state–action pairs as

$$\mathcal{T}(\rho') = \bigotimes_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{T}_{s,a}(\rho')$$
$$\text{with} \quad \mathcal{T}_{s,a}(\rho') = \left\{ T_{s,a} \in \Delta(\mathcal{S}) \ : \ D_{TV}(T_{s,a}, T^o_{s,a}) \le \rho' \right\}. \tag{2}$$

Here, $T^o = \left(T^o_{s,a}, (s,a) \in \mathcal{S}\times\mathcal{A}\right)$ denotes the nominal transition kernel, $D_{TV}(\cdot,\cdot)$ measures the Total Variation distance between two distributions, $\Delta(\mathcal{S})$ is the probability distribution over $\mathcal{S}$, and $\rho' > 0$ specifies the radius of the ambiguity set that determines the degree of robustness. The *distributionally robust offline imitation learning (DROIL)* problem is then defined as

$$\pi^\star_{droil} \in \arg\min_{\pi\in\Pi} \max_{T\in\mathcal{T}} \mathbb{E}_{s\sim d_T^{\pi_D}} \left[\mathcal{L}(\pi(\cdot|s), \pi_D(\cdot|s))\right]. \tag{3}$$

where $d_T^{\pi_D}$ is the occupancy measure induced by expert policy π_D in transition kernel T. The objective of this problem is to learn a *robust policy* $\pi^\star_{droil}$ that minimizes the worst-case imitation loss evaluated over the state distributions induced by the expert policy π_D under all admissible transition kernels $T \in \mathcal{T}(\rho')$. The problem can equivalently be viewed as a game in which the adversary selects a transition kernel $T \in \mathcal{T}(\rho')$ that maximizes the imitation loss for a given learner's policy π, while the learner seeks a policy $\pi^\star_{droil}$ that minimizes this adversarial objective. In the offline setting, however, the key challenge is that expert demonstrations are collected only under the nominal transition kernel T^o, and no additional on-policy interactions are permitted in any $T \in \mathcal{T}(\rho')$.

Definition 1 (f-divergence). *Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a continuous and convex function, and let $p, q \in \Delta(\mathcal{X})$ denote two probability distributions over a domain $\mathcal{X}$. The f-divergence between p and q is defined as*

$$D_f(p \,\|\, q) = \mathbb{E}_{x\sim q}\left[f\left(\frac{p(x)}{q(x)}\right)\right]. \tag{4}$$

A widely used example of an f-divergence is the Kullback–Leibler (KL) divergence, which arises when $f(u) = u\log u$.

4 Methodology

We begin by deriving a theoretical result that quantifies how uncertainty in the transition model influences the induced occupancy measure for a fixed policy.

Lemma 1. *For any policy π and transition kernel $T \in \mathcal{T}(\rho')$, the following inequality holds:*

$$D_{\mathrm{TV}}(d_T^\pi, \, d_{T^o}^\pi) \le \frac{\rho'}{1-\gamma},$$

where $d_T^\pi, d_{T^0}^\pi$ are the (s,a,s') occupancy measures induced by policy π under transition kernels T and T^0 respectively.

Lemma 1 shows that, for any admissible transition kernel $T \in \mathcal{T}(\rho')$, the occupancy measure over triplets (s,a,s') induced by a policy π deviates from that under the nominal dynamics T^o by at most $\frac{\rho'}{1-\gamma}$ in Total Variation distance. The complete proof is provided in Appendix A. Having established this bound, we can equivalently express the inner maximization in the *DROIL* objective (3) as a maximization over occupancy measures rather than transition kernels, which upper bounds the maximization over transition kernels, where the occupancy uncertainty set is defined as

$$\mathcal{D}^{\pi_D}(\rho) = \left\{ d_T^{\pi_D} : D_{\mathrm{TV}}(d_T^{\pi_D}, \, d_{T^o}^{\pi_D}) \le \tfrac{\rho'}{1-\gamma} = \rho \right\}. \tag{5}$$

As we will show next, optimizing directly over $\mathcal{D}^{\pi_D}$ in (5) enables us to formulate our constrained problem entirely in the space of occupancy measures, thereby eliminating explicit dependence on $T \in \mathcal{T}(\rho')$, which is unknown and cannot be sampled from in the purely offline setting.

Robust Policy Learning. We now formulate the *DROIL* problem entirely in the space of occupancy measures. Specifically, we consider the following constrained minimax optimization problem:

$$\text{BE-DROIL} := \min_{\pi} \max_{d_T^{\pi_D}\ge 0} \mathbb{E}_{s\sim d_T^{\pi_D}}[\mathcal{L}(\pi(\cdot|s), \pi_D(\cdot|s))] \tag{6}$$

$$\text{s.t.} \quad \sum_{s'} d_T^{\pi_D}(s,a,s') = (1-\gamma)\mu(s)\pi_D(a|s) + \gamma\pi_D(a|s)\sum_{\tilde{s},\tilde{a}} d_T^{\pi_D}(\tilde{s},\tilde{a},s), \quad \forall (s,a) \tag{7}$$

$$D_f(d_T^{\pi_D}(s,a,s') \,\|\, d_{T^o}^{\pi_D}(s,a,s')) \le \rho. \tag{8}$$

Equation (6) corresponds to the *DROIL* problem defined in (3). The first constraint (7) enforces a *Bellman flow conservation (**balance equation**)* condition, ensuring that $d_T^{\pi_D}(s, a, s')$ represents a valid stationary occupancy measure realizable under the expert policy π_D and the transition kernel T, thereby preserving the temporal dependencies that characterize the underlying MDP. For a detailed treatment of this consistency property, refer to [34, 4]. Enforcing this condition prevents the assignment of arbitrary probability mass to transitions that cannot occur under the expert's policy in the underlying MDP. Based on the occupancy uncertainty set defined in (5), the second constraint (8) bounds the deviation of the occupancy measure of the expert's policy π_D under perturbed dynamics from that under nominal dynamics within an f-divergence ball of radius ρ. For $\rho = 0$, the formulation reduces to non-robust imitation learning in (1), but with Bellman flow consistency enforced in the nominal environment. Although we employ a general f-divergence formulation here, we later relate it to the Total Variation bound established in Lemma 1.

Why the Triplet (s, a, s') Occupancy Measure. A key design choice in our formulation is to define the occupancy measure over triplets (s, a, s'), rather than the conventional state–action pair (s, a). If the occupancy were defined only on (s, a), the Bellman flow constraint would explicitly depend on the transition kernel $T(s'|s, a)$ which is not known for any arbitrary $T \in \mathcal{T}(\rho')$ and cannot be estimated in offline settings where data is available only from a single nominal environment and no further interaction with the environment is permitted. By incorporating the next state s' directly into $d_T^{\pi_D}(s, a, s')$, the flow constraint can be expressed entirely in terms of occupancy measures, eliminating the need to deal with T explicitly. As we will show next, this formulation therefore enables robust policy learning in the fully offline regime.

Lagrangian Formulation. For a fixed learner policy π, we begin by considering the dual problem of the inner maximization problem in (6)-(8). This dual problem can be written as:

$$\begin{aligned}\min_{Q,\tau\geq 0}\ \max_{d_T^{\pi_D}\geq 0}\ & \mathbb{E}_{s\sim d_T^{\pi_D}}\big[L_\pi(s)\big] - \tau\big(D_f\big(d_T^{\pi_D}(s,a,s') \,\big\|\, d_{T^o}^{\pi_D}(s,a,s')\big) - \rho\big) \\ & - \sum_{s,a} Q(s,a)\left[\sum_{s'} d_T^{\pi_D}(s,a,s') - (1-\gamma)\mu(s)\pi_D(a|s) - \gamma\pi_D(a|s)\sum_{\tilde{s},\tilde{a}} d_T^{\pi_D}(\tilde{s},\tilde{a},s)\right].\end{aligned} \tag{9}$$

where $L_\pi(s) = \mathcal{L}(\pi(\cdot|s), \pi_D(\cdot|s))$, $Q(s, a) \in \mathbb{R}$ denotes the Lagrange multiplier associated with the Bellman flow constraint, and $\tau \geq 0$ is the multiplier for the f-divergence constraint.

Because the inner maximization problem in (6)-(8) is convex and admits a strictly feasible point (i.e. $d_{T^o}^{\pi_D}$), Slater's condition [6] ensures that strong duality holds. Consequently, the dual problem in (9) attains the same optimal value as the original maximization over $d_T^{\pi_D}$.

To simplify (9), we expand and reorganize several terms as follows.

$$\sum_{s,a} Q(s,a)[(1-\gamma)\mu(s)\pi_D(a|s)] = (1-\gamma)\ \mathbb{E}_{s\sim\mu,\,a\sim\pi_D(\cdot|s)}[Q(s,a)],$$

$$\begin{aligned}\sum_{s,a} Q(s,a)\left[\gamma\pi_D(a|s)\sum_{\tilde{s},\tilde{a}} d_T^{\pi_D}(\tilde{s},\tilde{a},s)\right] &= \sum_{s',a'} Q(s',a')\left[\gamma\pi_D(a'|s')\sum_{s,a} d_T^{\pi_D}(s,a,s')\right] \\ &= \sum_{s,a,s'} d_T^{\pi_D}(s,a,s')\left[\gamma\sum_{a'} Q(s',a')\pi_D(a'|s')\right] \\ &= \gamma\ \mathbb{E}_{s,a,s'\sim d_T^{\pi_D}}\left[\mathbb{E}_{a'\sim\pi_D(\cdot|s')}[Q(s',a')]\right],\end{aligned}$$

$$\sum_{s,a} Q(s,a)\left[-\sum_{s'} d_T^{\pi_D}(s,a,s')\right] = -\ \mathbb{E}_{s,a,s'\sim d_T^{\pi_D}}[Q(s,a)].$$

Substituting these identities into (9) yields the simplified Lagrangian:

$$\begin{aligned}\min_{Q,\tau\geq 0}\ \max_{d_T^{\pi_D}\geq 0}\ & (1-\gamma)\,\mathbb{E}_{s\sim\mu,\,a\sim\pi_D(\cdot|s)}[Q(s,a)] - \tau\, D_f(d_T^{\pi_D}(s,a,s') \,\|\, d_{T^o}^{\pi_D}(s,a,s')) + \rho\tau \\ & + \mathbb{E}_{s,a,s'\sim d_T^{\pi_D}}\Big[L_\pi(s) + \gamma\,\mathbb{E}_{a'\sim\pi_D(\cdot|s')}[Q(s',a')] - Q(s,a)\Big].\end{aligned} \tag{10}$$

This Lagrangian cannot be directly optimized using offline data because it depends on the unknown $d_T^{\pi_D}(s, a, s')$ for arbitrary $T \in \mathcal{T}(\rho')$. Since only expert trajectories under the nominal kernel T^o are available, we define an importance ratio

$$w(s, a, s') = \frac{d_T^{\pi_D}(s, a, s')}{d_{T^o}^{\pi_D}(s, a, s')}.$$

Using the definition of f-divergence, we can rewrite (10) as

$$\begin{aligned} \min_{Q, \tau \geq 0} \max_{w \geq 0} \; & (1-\gamma)\, \mathbb{E}_{s \sim \mu,\, a \sim \pi_D(\cdot|s)}[Q(s, a)] + \rho\tau \\ & + \mathbb{E}_{s,a,s' \sim d_{T^o}^{\pi_D}}\left[-\tau\, f(w(s, a, s')) + w(s, a, s')\, e_{Q,\pi}(s, a, s')\right], \end{aligned} \tag{11}$$

where $e_{Q,\pi}(s, a, s') = L_\pi(s) + \gamma\, \mathbb{E}_{a' \sim \pi_D(\cdot|s')}[Q(s', a')] - Q(s, a)$.

At this point, it is important to note that the original optimization problem in (6)–(8) involved a maximization over unknown occupancy distributions $d_T^{\pi_D}$ for any $T \in \mathcal{T}(\rho')$. Through the above reformulation, we convert this intractable objective into one that depends only on samples drawn from the expert policy under the nominal dynamics T^o. This transformation is crucial, as it enables practical optimization in a strictly offline setting using expert demonstration data alone.

Proposition 1. *For $\tau > 0$, the inner maximization in* (11) *admits the solution*

$$w^\star_{Q,\tau,\pi}(s, a, s') = \max\left(0,\; (f')^{-1}\left(\frac{e_{Q,\pi}(s, a, s')}{\tau}\right)\right).$$

For $\tau = 0$, $w^\star_{Q,\tau,\pi}(s, a, s') = +\infty$ if $e_{Q,\pi}(s, a, s') > 0$ and 0 otherwise.

The proof of Proposition 1 is provided in the Appendix A. Substituting $w^\star_{Q,\tau,\pi}$ from Proposition 1 into (11) simplifies the problem to a single minimization over Q and τ. This results in a two-stage learning procedure for the original optimization problem in (6)–(8), where the first stage estimates (Q, τ) under the nominal data distribution, and the second stage updates the policy parameters to minimize the expected weighted loss:

$$\begin{aligned} \min_{Q, \tau \geq 0} \quad & (1-\gamma)\, \mathbb{E}_{s \sim \mu,\, a \sim \pi_D(\cdot|s)}[Q(s, a)] + \rho\tau \\ & + \mathbb{E}_{(s,a,s') \sim d_{T^o}^{\pi_D}}\left[-\rho f\left(w^\star_{Q,\tau,\pi}(s, a, s')\right) + w^\star_{Q,\tau,\pi}(s, a, s')\, e_{Q,\pi}(s, a, s')\right] \\ \min_{\pi} \quad & \mathbb{E}_{(s,a,s') \sim d_{T^o}^{\pi_D}}\left[w^\star_{Q,\tau,\pi}(s, a, s')\, L_\pi(s)\right]. \end{aligned} \tag{12}$$

Optimization alternates between updating (Q, τ) and the policy π until convergence, yielding a robust policy that minimizes expected loss under worst-case transition perturbations.

Discussion on f-divergence. From Lemma 1, we established a Total Variation (TV) bound between the occupancy measures induced by a policy under perturbed and nominal transition kernels. To extend this result beyond TV, we adopt an f-divergence formulation in our constrained optimization problem. To keep the constraint theoretically meaningful, Lemma 2 ensures that any f-divergence whose generator is upper bounded by that of TV inherits the same upper bound between the occupancy measures induced by the same policy, as in Lemma 1. The proof is provided in Appendix A. Proposition 1 further requires the generator $f(\cdot)$ to be differentiable with an invertible derivative $f'(\cdot)$ to compute $(f')^{-1}$. Since the TV generator $f_{\mathrm{TV}}(t) = \frac{1}{2}|t-1|$ is non-differentiable at $t = 1$, we employ a smooth approximation, the *SoftTV* generator [40], defined as

$$f_{\mathrm{SoftTV}}(x) = \tfrac{1}{2}\log\big(\cosh(x-1)\big), \qquad \big(f'_{\mathrm{SoftTV}}\big)^{-1}(y) = \tanh^{-1}(2y) + 1. \tag{13}$$

Lemma 6 in Appendix A shows that $f_{\mathrm{SoftTV}}(x) \leq f_{\mathrm{TV}}(x)$ for all x, and combining this with Lemma 2 for $\alpha = 1$ gives $D_{f_{\mathrm{SoftTV}}}(d_T^\pi(s, a, s') \| d_{T^o}^\pi(s, a, s')) \leq \frac{\rho'}{1-\gamma}$. In our experiments, we use $f_{\mathrm{SoftTV}}(\cdot)$ as the generator for the f-divergence; however, any differentiable generator with an invertible derivative satisfying Lemma 2 can be used.

Lemma 2. *Let $f : [0, \infty) \to \mathbb{R}$ be an f-divergence generator with $f(1) = 0$ and $f(t) \leq \alpha\, f_{\mathrm{TV}}(t)$ for all $t \geq 0$, where $f_{\mathrm{TV}}(t) = \frac{1}{2}|t-1|$ is the generator of total variation. Then, for any policy π and transition kernel $T \in \mathcal{T}(\rho')$,*

$$D_f(d_T^\pi(s, a, s') \| d_{T^o}^\pi(s, a, s')) \leq \alpha\, D_{\mathrm{TV}}(d_T^\pi, d_{T^o}^\pi) \leq \alpha\, \frac{\rho'}{1-\gamma}.$$

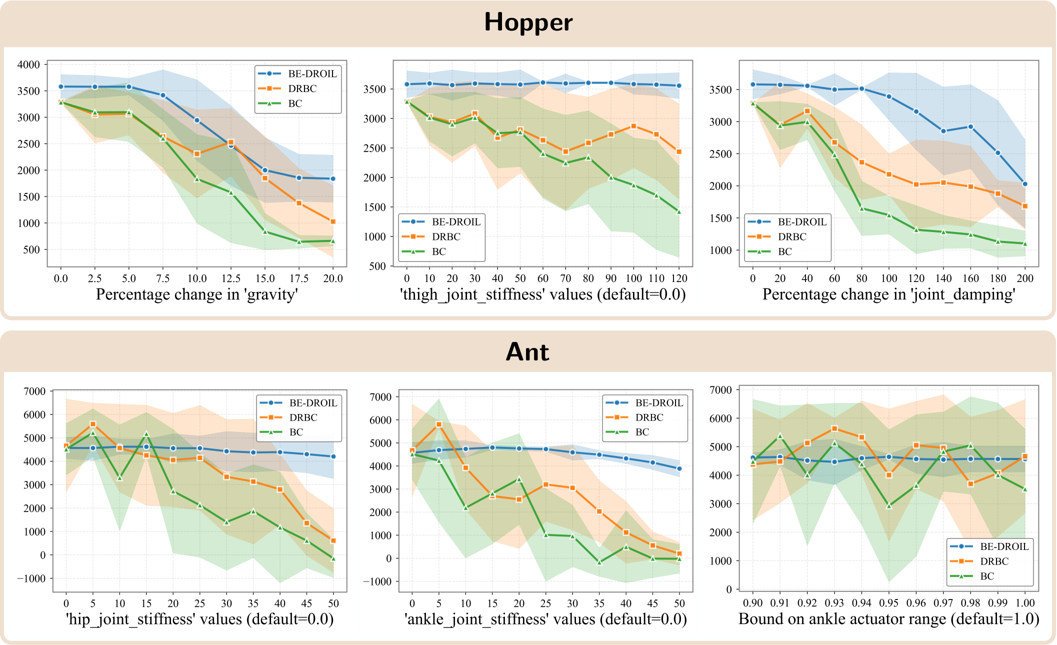

Figure 1: Perturbation results for Hopper and Ant, with Y-axis denoting average cumulative reward.

5 Experimental Results

Setup. We evaluate the empirical performance of our proposed BE-DROIL algorithm on the MuJoCo locomotion suite [42], a standard benchmark for imitation learning in continuous-control domains with varying levels of difficulty. To construct the expert demonstration dataset, we follow the protocol of Panaganti et al. [30], using expert trajectories generated by pre-trained TD3 [12] policies available in the RL Baselines3 Zoo repository [35]. The number of samples per environment is also aligned with Panaganti et al. [30] to ensure fair comparison. Additional training details are provided in the Appendix B.

Baseline Algorithms. We compare BE-DROIL against two baselines. Since fully offline robust imitation learning, where demonstrations originate from a single nominal environment, remains relatively unexplored, the only comparable prior method is Distributionally Robust Behavior Cloning (DRBC) [30]. In addition, we include standard Behavioral Cloning (BC) [33] as a non-robust baseline to highlight the importance of accounting for transition uncertainty when deployment dynamics differ from those seen during training.

Implementation. The term $(1-\gamma)\,\mathbb{E}_{s\sim\mu,,a\sim\pi_D(\cdot|s)}[Q(s,a)]$ in (12) requires sampling from the initial-state distribution μ. Since expert datasets contain few unique initial states (e.g., only six in Walker2d), following standard DICE-based practice [21], we treat every state within a trajectory as an effective initial state to ensure sufficient coverage for stable estimation. Network architectures and remaining optimization details are provided in the Appendix B.

Domains. To evaluate robustness under transition shifts, we assess performance in perturbed test environments where key physical parameters are modified to induce model mismatch. Following the perturbation protocol of Panaganti et al. [30], we vary parameters affecting torque generation, compliance, and energy dissipation, thereby capturing distinct modes of transition shift. Specifically, for Hopper, we perturb gravity, joint damping, and thigh-joint stiffness; for Ant, ankle- and hip-joint stiffness and actuator range; for Walker2d, gravity, actuator range, and foot-joint damping; and for HalfCheetah, back-joint stiffness, joint damping, and frictionloss.

Empirical Findings. Reported scores correspond to the mean and standard deviation of episodic returns over 100 independently seeded rollouts. Unlike DRBC, which tunes its hyperparameters (e.g., uncertainty radius and learning rate) separately for each environment, BE-DROIL uses a single set

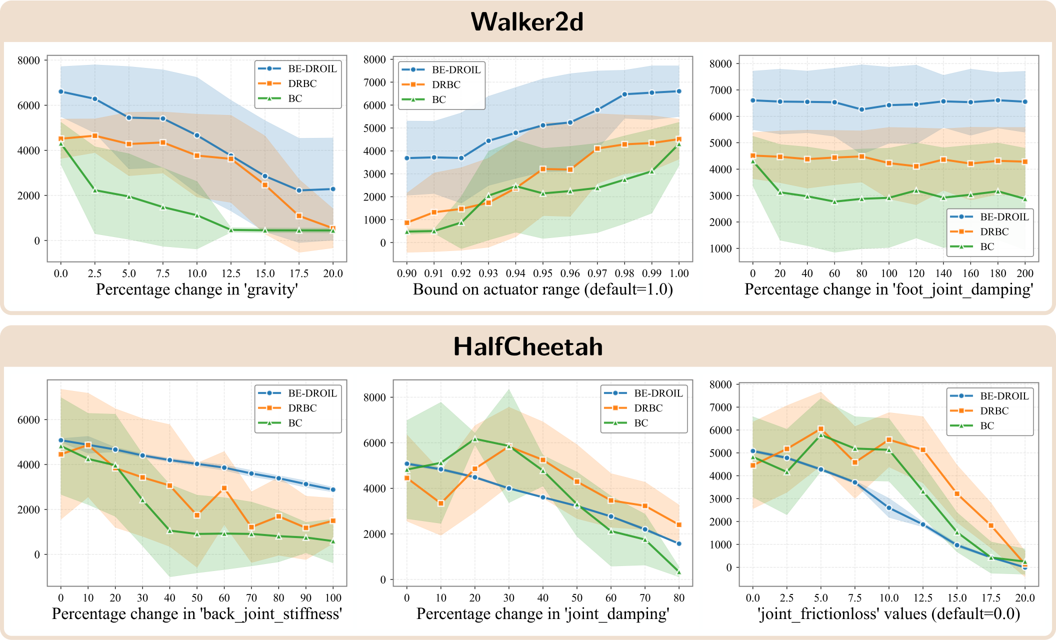

Figure 2: Perturbation results for Walker2d and HalfCheetah, with Y-axis denoting average cumulative reward.

optimized on Ant and applies it unchanged across all others. This design emphasizes generalization and fairness over environment-specific tuning.

Hopper. As gravity or damping increases, the agent encounters greater resistive forces and energy losses, making balance and propulsion harder to sustain. With higher stiffness, the leg becomes more rigid, reducing compliance and agility. Figure 1 shows that BC's performance drops sharply, failing to remain upright even under moderate perturbations. DRBC is more resilient but still declines steadily as shifts intensify. In contrast, BE-DROIL maintains high returns across all ranges, showing only gradual decay under extreme conditions while remaining consistently stable.

Ant. When the actuator range is reduced, the limbs lose torque authority, making it harder for the agent to recover from posture disturbances involving multiple legs. Figure 1 shows that BC collapses, unable to maintain a coherent gait as torque and flexibility decrease. DRBC declines moderately but retains basic stability across perturbations. BE-DROIL, however, sustains nearly constant returns, preserving coordination and balance even when both actuation strength and joint stiffness are severely limited.

Walker2d. Figure 2 shows that BC quickly loses stability as its legs fall out of sync, leading to frequent collapses. DRBC preserves locomotion under moderate perturbations but gradually deteriorates as gravity and damping increase. BE-DROIL remains the most stable across all conditions, maintaining coordinated walking and high returns with only mild decline even under strong environmental and control shifts.

HalfCheetah. Figure 2 shows that BC's performance drops sharply under these perturbations, while DRBC maintains greater stability. BE-DROIL performs best under stiffness changes but exhibits steeper declines under increased damping and friction. This gap likely stems from DRBC using environment-specific uncertainty radii and learning rates optimized for HalfCheetah, whereas BE-DROIL employs a fixed configuration across all domains.

Overall, BE-DROIL achieves state-of-the-art robustness on Ant, Hopper, and Walker2d, showing low variance and smaller degradation than baselines under actuator, stiffness, and gravity perturbations. Using a fixed hyperparameter configuration across all tasks ensures fair cross-domain evaluation but results in weaker performance under high damping and friction loss in HalfCheetah. Preliminary re-tuning indicates that minor adjustments to the learning rate and uncertainty radius largely close this gap, confirming BE-DROIL's effectiveness when properly scaled. These results demonstrate principled distributional robustness under transition shifts, with strong generalization and consistent performance across domains without environment-specific tuning.

6 Conclusion

This work introduced BE-DROIL, a principled framework for *distributionally robust imitation learning* under dynamics mismatch in the strictly offline setting, where only expert demonstrations from a single nominal environment are available. We formulated robust policy learning as a constrained minimax optimization over an f-divergence ambiguity set around the nominal transition kernel and derived an equivalent offline objective via a triplet occupancy formulation that eliminates explicit dependence on unknown dynamics. Using convex duality, we reduced the adversarial problem to a tractable importance-weighting scheme under nominal data, yielding an efficient alternating optimization algorithm. Empirically, BE-DROIL demonstrates consistent robustness gains over existing baselines across multiple continuous-control benchmarks with perturbed dynamics. Future work will extend this framework to non-stationary dynamics and large-scale imitation foundation models, as well as deriving analogues of Lemma 1 for alternative divergence measures to broaden the class of admissible f-divergence generators in our optimization framework.

7 Impact Statement

This work introduces BE-DROIL, a framework for distributionally robust offline imitation learning that enhances policy reliability under dynamics shift. The method can reduce risky data collection in robotics or healthcare by learning entirely from offline demonstrations, but it may also propagate biases or unsafe behaviors present in expert data. Robustness guarantees are bounded by the chosen uncertainty set and may not hold under unmodeled or extreme shifts. The approach should therefore be deployed cautiously in real-world systems, with safety audits, drift monitoring, and human oversight. We will release code to ensure reproducibility while maintaining ethical and license compliance. Overall, BE-DROIL advances safer, more generalizable policy learning, provided users remain mindful of its assumptions and scope.

References

[1] Rishabh Agrawal, Nathan Dahlin, Rahul Jain, and Ashutosh Nayyar. Policy optimization for strictly batch imitation learning. In *OPT 2024: Optimization for Machine Learning*, 2024. URL `https://openreview.net/forum?id=5L3qmIOXPz`.

[2] Rishabh Agrawal, Nathan Dahlin, Rahul Jain, and Ashutosh Nayyar. Markov balance satisfaction improves performance in strictly batch offline imitation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 15311–15319, 2025.

[3] Rishabh Agrawal, Nathan Dahlin, Rahul Jain, and Ashutosh Nayyar. Conditional kernel imitation learning for continuous state environments. In Necmiye Ozay, Laura Balzano, Dimitra Panagou, and Alessandro Abate, editors, *Proceedings of the 7th Annual Learning for Dynamics &amp; Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pages 469–483. PMLR, 04–06 Jun 2025. URL `https://proceedings.mlr.press/v283/agrawal25a.html`.

[4] Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

[5] Shuanghao Bai, Wanqi Zhou, Pengxiang Ding, Wei Zhao, Donglin Wang, and Badong Chen. Rethinking latent redundancy in behavior cloning: An information bottleneck approach for robot manipulation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 2560–2580. PMLR, 13–19 Jul 2025. URL `https://proceedings.mlr.press/v267/bai25e.html`.

[6] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[8] Jongseong Chae, Seungyul Han, Whiyoung Jung, Myungsik Cho, Sungho Choi, and Youngchul Sung. Robust imitation learning against variations in environment dynamics. In *International Conference on Machine Learning*, pages 2828–2852. PMLR, 2022.

[9] Esther Derman and Shie Mannor. Distributional robustness and regularization in reinforcement learning. *arXiv preprint arXiv:2003.02894*, 2020.

[10] Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot visual imitation learning via meta-learning. In *Conference on robot learning*, pages 357–368. PMLR, 2017.

[11] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rkHywl-A-`.

[12] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

[13] Ryan Hoque, Ajay Mandlekar, Caelan Garrett, Ken Goldberg, and Dieter Fox. Intervengen: Interventional data generation for robust and data-efficient robot imitation learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2840–2846. IEEE, 2024.

[14] Hao-Lun Hsu, Haocheng Meng, Shaocheng Luo, Juncheng Dong, Vahid Tarokh, and Miroslav Pajic. Reforma: Robust reinforcement learning via adaptive adversary for drones flying under disturbances. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5169–5175. IEEE, 2024.

[15] Wenlong Huang, Igor Mordatch, Pieter Abbeel, and Deepak Pathak. Generalization in dexterous manipulation via geometry-aware multi-task learning. *arXiv preprint arXiv:2111.03062*, 2021.

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

[17] Stephen James, Michael Bloesch, and Andrew J Davison. Task-embedded control networks for few-shot imitation learning. In *Conference on robot learning*, pages 783–795. PMLR, 2018.

[18] Yongpeng Jiang, Mingrui Yu, Xinghao Zhu, Masayoshi Tomizuka, and Xiang Li. Contact-implicit model predictive control for dexterous in-hand manipulation: A long-horizon and robust approach. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5260–5266. IEEE, 2024.

[19] Geon-Hyeong Kim, Seokin Seo, Jongmin Lee, Wonseok Jeon, HyeongJoo Hwang, Hongseok Yang, and Kee-Eung Kim. DemoDICE: Offline imitation learning with supplementary imperfect demonstrations. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=BrPdX1bDZkQ`.

[20] Kuno Kim, Yihong Gu, Jiaming Song, Shengjia Zhao, and Stefano Ermon. Domain adaptive imitation learning. In *International Conference on Machine Learning*, pages 5286–5295. PMLR, 2020.

[21] Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=Hyg-JC4FDr`.

[22] Michael Laskey, Jonathan Lee, Roy Fox, Anca Dragan, and Ken Goldberg. Dart: Noise injection for robust imitation learning. In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg, editors, *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 143–156. PMLR, 13–15 Nov 2017. URL `https://proceedings.mlr.press/v78/laskey17a.html`.

[23] Luc Le Mero, Dewei Yi, Mehrdad Dianati, and Alexandros Mouzakitis. A survey on imitation learning techniques for end-to-end autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):14128–14147, 2022.

[24] Zixuan Liu, Liu Liu, Bingzhe Wu, Lanqing Li, Xueqian Wang, Bo Yuan, and Peilin Zhao. Dynamics adapted imitation learning. *Transactions on Machine Learning Research*, 2023.

[25] Yiren Lu and Jonathan Tompson. Adail: Adaptive adversarial imitation learning. *arXiv preprint arXiv:2008.12647*, 2020.

[26] Jiafei Lyu, Kang Xu, Jiacheng Xu, Jing-Wen Yang, Zongzhang Zhang, Chenjia Bai, Zongqing Lu, Xiu Li, et al. Odrl: A benchmark for off-dynamics reinforcement learning. *Advances in Neural Information Processing Systems*, 37:59859–59911, 2024.

[27] Liyuan Mao, Haoran Xu, Weinan Zhang, and Xianyuan Zhan. ODICE: Revealing the mystery of distribution correction estimation via orthogonal-gradient update. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=L8UNn7Llt4`.

[28] Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[29] Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. *Advances in neural information processing systems*, 35: 32211–32224, 2022.

[30] Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Distributionally robust behavioral cloning for robust imitation learning. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 1342–1347. IEEE, 2023.

[31] Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Bridging distributionally robust learning and offline rl: An approach to mitigate distribution shift and partial data coverage. In Necmiye Ozay, Laura Balzano, Dimitra Panagou, and Alessandro Abate, editors, *Proceedings of the 7th Annual Learning for Dynamics &amp; Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pages 619–634. PMLR, 04–06 Jun 2025. URL `https://proceedings.mlr.press/v283/panaganti25a.html`.

[32] 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.

[33] Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1:305–313, 1988.

[34] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[35] Antonin Raffin. Rl baselines3 zoo. `https://github.com/DLR-RM/rl-baselines3-zoo`, 2020.

[36] Dripta S Raychaudhuri, Sujoy Paul, Jeroen Vanbaar, and Amit K Roy-Chowdhury. Cross-domain imitation from observations. In *International conference on machine learning*, pages 8902–8912. PMLR, 2021.

[37] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

[38] Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 661–668. JMLR Workshop and Conference Proceedings, 2010.

[39] Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 627–635, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL `https://proceedings.mlr.press/v15/ross11a.html`.

[40] Seokin Seo, Byung-Jun Lee, Jongmin Lee, HyeongJoo Hwang, Hongseok Yang, and Kee-Eung Kim. Mitigating covariate shift in behavioral cloning via robust stationary distribution correction. *Advances in Neural Information Processing Systems*, 37:109177–109201, 2024.

[41] Voot Tangkaratt, Nontawat Charoenphakdee, and Masashi Sugiyama. Robust imitation learning from noisy demonstrations. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 298–306. PMLR, 13–15 Apr 2021. URL `https://proceedings.mlr.press/v130/tangkaratt21a.html`.

[42] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

[43] Luca Viano, Yu-Ting Huang, Parameswaran Kamalaruban, Adrian Weller, and Volkan Cevher. Robust inverse reinforcement learning under transition dynamics mismatch. *Advances in Neural Information Processing Systems*, 34:25917–25931, 2021.

[44] Wei Wang, Dongqi Han, Xufang Luo, and Dongsheng Li. Addressing signal delay in deep reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2023.

[45] Shili Wu, Yizhao Jin, Puhua Niu, Aniruddha Datta, and Sean B Andersson. Robust behavior cloning via global lipschitz regularization. *arXiv preprint arXiv:2506.19250*, 2025.

[46] Yueh-Hua Wu, Nontawat Charoenphakdee, Han Bao, Voot Tangkaratt, and Masashi Sugiyama. Imitation learning from imperfect demonstration. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6818–6827. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/wu19a.html`.

[47] Tianhe Yu, Chelsea Finn, Annie Xie, Sudeep Dasari, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. *arXiv preprint arXiv:1802.01557*, 2018.

[48] Zhuodong Yu, Ling Dai, Shaohang Xu, Siyang Gao, and Chin Pang Ho. Fast bellman updates for wasserstein distributionally robust mdps. *Advances in Neural Information Processing Systems*, 36:30554–30578, 2023.

[49] Yisong Yue and Hoang M. Le. Imitation learning (tutorial). *International Conference on Machine Learning (ICML)*, 2018.

[50] Allan Zhou, Eric Jang, Daniel Kappler, Alex Herzog, Mohi Khansari, Paul Wohlhart, Yunfei Bai, Mrinal Kalakrishnan, Sergey Levine, and Chelsea Finn. Watch, try, learn: Meta-learning from demonstrations and rewards. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SJg5J6NtDr`.

[51] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

A Theoretical Derivations.

Lemma 3. *For any policy π and transition kernel $T \in \mathcal{T}(\rho')$, the following holds:*

$$D_{\mathrm{TV}}(d_T^\pi(s), d_{T^o}^\pi(s)) \leq \frac{\gamma\rho'}{1-\gamma}.$$

Proof. See Lemma 7 in Panaganti et al. [30] for a detailed analysis. □

Lemma 4. *For any policy π and transition kernel $T \in \mathcal{T}(\rho')$, the following holds:*

$$D_{\mathrm{TV}}(d_T^\pi(s,a), d_{T^o}^\pi(s,a)) \leq \frac{\gamma\rho'}{1-\gamma}.$$

Proof. We build upon the Lemma 3 that quantifies how uncertainty in the transition model influences the induced occupancy measure on the state-space for a fixed policy.

Since $d_T^\pi(s,a) = \pi(a|s)d_T^\pi(s)$, it follows that

$$\begin{aligned} D_{\mathrm{TV}}(d_T^\pi(s,a), d_{T^o}^\pi(s,a)) &= \frac{1}{2}\sum_{s,a} |d_T^\pi(s,a) - d_{T^o}^\pi(s,a)| \\ &= \frac{1}{2}\sum_{s,a} \pi(a|s)|d_T^\pi(s) - d_{T^o}^\pi(s)| \\ &= D_{\mathrm{TV}}(d_T^\pi(s), d_{T^o}^\pi(s)) \\ &\leq \frac{\gamma\rho'}{1-\gamma}. \end{aligned}$$

□

Having computed the occupancy measure total-variation distance over state–action pairs for a fixed policy under two distinct transition kernels, we now restate Lemma 1 and provide its proof.

Lemma 1. *Consider any policy π and $T \in \mathcal{T}(\rho')$. Then,*

$$D_{\mathrm{TV}}\big(d_T^\pi(s,a,s'),\, d_{T^o}^\pi(s,a,s')\big) \leq \frac{\rho'}{1-\gamma}.$$

Proof. We have

$$d_T^\pi(s,a,s') = d_T^\pi(s,a)T_{s,a}(s'), \quad d_{T^o}^\pi(s,a,s') = d_{T^o}^\pi(s,a)T_{s,a}^o(s').$$

Then,

$$\begin{aligned} &\|d_T^\pi(s,a,s') - d_{T^o}^\pi(s,a,s')\|_1 \\ &= \sum_{s,a,s'} \big|d_T^\pi(s,a)T_{s,a}(s') - d_{T^o}^\pi(s,a)T_{s,a}^o(s')\big| \\ &\leq \sum_{s,a,s'} d_T^\pi(s,a)\,\big|T_{s,a}(s') - T_{s,a}^o(s')\big| + \sum_{s,a,s'} |d_T^\pi(s,a) - d_{T^o}^\pi(s,a)|\, T_{s,a}^o(s'). \end{aligned}$$

For the first term, using $\|T_{s,a} - T_{s,a}^o\|_1 = 2D_{\mathrm{TV}}(T_{s,a}, T_{s,a}^o) \leq 2\rho'$,

$$\sum_{s,a,s'} d_T^\pi(s,a)\,\big|T_{s,a}(s') - T_{s,a}^o(s')\big| = \sum_{s,a} d_T^\pi(s,a)\|T_{s,a} - T_{s,a}^o\|_1 \leq 2\rho'.$$

For the second term, since $\sum_{s'} T_{s,a}^o(s') = 1$, we have

$$\begin{aligned} \sum_{s,a,s'} |d_T^\pi(s,a) - d_{T^o}^\pi(s,a)|\, T_{s,a}^o(s') &= \sum_{s,a} |d_T^\pi(s,a) - d_{T^o}^\pi(s,a)| \\ &= \|d_T^\pi(s,a) - d_{T^o}^\pi(s,a)\|_1. \end{aligned}$$

By lemma 4,

$$D_{\mathrm{TV}}(d_T^\pi(s,a), d_{T^o}^\pi(s,a)) \le \frac{\gamma\rho'}{1-\gamma},$$

which implies

$$\|d_T^\pi(s,a) - d_{T^o}^\pi(s,a)\|_1 \le \frac{2\gamma\rho'}{1-\gamma}.$$

Combining both terms,

$$\|d_T^\pi(s,a,s') - d_{T^o}^\pi(s,a,s')\|_1 \le 2\rho' + \frac{2\gamma\rho'}{1-\gamma} = \frac{2\rho'}{1-\gamma}.$$

Hence,

$$D_{\mathrm{TV}}\big(d_T^\pi(s,a,s'), d_{T^o}^\pi(s,a,s')\big) = \frac{1}{2}\|d_T^\pi(s,a,s') - d_{T^o}^\pi(s,a,s')\|_1 \le \frac{\rho'}{1-\gamma}.$$

□

Proposition 1. *When $\tau > 0$, for the inner maximization subproblem in* (11),

$$w^\star_{Q,\tau,\pi}(s,a,s') := \arg\max_{w\ge 0}\Big\{(1-\gamma)\ \mathbb{E}_{s\sim\mu,\,a\sim\pi_D(\cdot|s)}[Q(s,a)] + \rho\tau + \mathbb{E}_{s,a,s'\sim d_{T^o}^{\pi_D}}\big[-\tau f(w(s,a,s')) + w(s,a,s')\, e_{Q,\pi}(s,a,s')\big]\Big\},$$

the optimizer admits the closed-form expression

$$w^\star_{Q,\tau,\pi}(s,a,s') = \max\bigg(0,\ (f')^{-1}\bigg(\frac{e_{Q,\pi}(s,a,s')}{\tau}\bigg)\bigg), \quad \forall(s,a,s'),$$

where $(f')^{-1}$ denotes the inverse mapping of f', which exists and is strictly increasing because f is strictly convex. For $\tau = 0$, $w^\star_{Q,\tau,\pi}(s,a,s') = +\infty$ if $e_{Q,\pi}(s,a,s') > 0$ and 0 otherwise.

Proof. We begin with the case when $\tau > 0$. Define the objective functional

$$\mathcal{L}(Q,\tau,w) := (1-\gamma)\,\mathbb{E}_{s\sim\mu,\,a\sim\pi_D(\cdot|s)}[Q(s,a)] + \rho\tau + \mathbb{E}_{s,a,s'\sim d_{T^o}^{\pi_D}}\big[-\tau f(w(s,a,s')) + w(s,a,s')\, e_{Q,\pi}(s,a,s')\big].$$

For fixed (Q,τ), the maximization $\max_{w\ge0}\mathcal{L}(Q,\tau,w)$ can be viewed as a constrained optimization problem over w. Since the first two terms in $\mathcal{L}(Q,\tau,w)$ do not depend on w, the optimization effectively reduces to maximizing only the expectation term involving w. That is,

$$\begin{aligned}&\max_{w\ge0} \quad \mathbb{E}_{s,a,s'\sim d_{T^o}^{\pi_D}}\big[-\tau f(w(s,a,s')) + w(s,a,s')\, e_{Q,\pi}(s,a,s')\big]\\ \Leftrightarrow\ &\max_{w\ge0} \quad \sum_{s,a,s'} d_{T^o}^{\pi_D}(s,a,s')\big[-\tau f(w(s,a,s')) + w(s,a,s')\, e_{Q,\pi}(s,a,s')\big].\end{aligned}$$

Each term in the summation depends only on its corresponding local variable $w(s,a,s')$, which implies that the overall maximization can be solved independently for each (s,a,s'). Accordingly, for every (s,a,s') with $d_{T^o}^{\pi_D}(s,a,s') > 0$, we consider the scalar subproblem:

$$\begin{aligned}\max_{w(s,a,s')} \quad & -\tau f(w(s,a,s')) + e_{Q,\pi}(s,a,s')\, w(s,a,s'),\\ \text{s.t.} \quad & -w(s,a,s') \le 0.\end{aligned}$$

Introducing the Lagrange multiplier $\kappa(s,a,s') \ge 0$ for the constraint $-w(s,a,s') \le 0$ yields the Lagrangian

$$\mathcal{J}(w,\kappa) = -\tau f(w(s,a,s')) + e_{Q,\pi}(s,a,s')\, w(s,a,s') - \kappa(s,a,s')\,(-w(s,a,s')).$$

Simplifying,

$$\mathcal{J}(w,\kappa) = -\tau f(w(s,a,s')) + \big(e_{Q,\pi}(s,a,s') + \kappa(s,a,s')\big)\, w(s,a,s').$$

Because f is strictly-convex and the feasible set $\{w \ge 0\}$ is convex, the objective is concave in w, and thus strong duality holds. Consequently, the Karush–Kuhn–Tucker (KKT) conditions are both necessary and sufficient for optimality. We now state the **KKT** conditions corresponding to the above Lagrangian formulation.

- **Primal feasibility:** $w^*(s,a,s') \geq 0$.
- **Dual feasibility:** $\kappa^*(s,a,s') \geq 0$.
- **Complementary slackness:** $\kappa^*(s,a,s')\, w^*(s,a,s') = 0$.
- **Stationarity:** The first-order condition is

$$-\tau f'(w^*(s,a,s')) + e_{Q,\pi}(s,a,s') + \kappa^*(s,a,s') = 0.$$

We now construct a candidate pair $(w^\star_{Q,\tau,\pi}, \kappa^\star)$ and show that it satisfies all KKT conditions:

$$\begin{aligned} w^\star_{Q,\tau,\pi}(s,a,s') &= \max\left(0,\, (f')^{-1}\left(\frac{e_{Q,\pi}(s,a,s')}{\tau}\right)\right), \\ \kappa^\star(s,a,s') &= \begin{cases} 0, & \text{if } (f')^{-1}\left(\frac{e_{Q,\pi}(s,a,s')}{\tau}\right) > 0, \\ \tau f'(0) - e_{Q,\pi}(s,a,s'), & \text{otherwise.} \end{cases} \end{aligned} \tag{14}$$

Verification of KKT conditions. By construction, $w^\star_{Q,\tau,\pi}(s,a,s') \geq 0$, ensuring primal feasibility. We now verify the remaining KKT conditions by considering two complementary cases.

- **Case (1):** $(f')^{-1}\left(\frac{e_{Q,\pi}(s,a,s')}{\tau}\right) > 0$. In this case, $w^\star_{Q,\tau,\pi}(s,a,s') = (f')^{-1}\left(\frac{e_{Q,\pi}(s,a,s')}{\tau}\right)$ and $\kappa^\star(s,a,s') = 0$. It is straightforward to verify that $w^\star_{Q,\tau,\pi}(s,a,s'), \kappa^\star(s,a,s')$ satisfy the KKT conditions in this case.
- **Case (2):** $(f')^{-1}\left(\frac{e_{Q,\pi}(s,a,s')}{\tau}\right) \leq 0$. In this case, $w^\star_{Q,\tau,\pi}(s,a,s') = 0$ and $\kappa^\star(s,a,s') = \tau f'(0) - e_{Q,\pi}(s,a,s')$. It is straightforward to verify that primal feasibility, complementary slackness and stationarity conditions are satisfied in this case. To verify dual feasibility, note that $(f')^{-1}\left(\frac{e_{Q,\pi}(s,a,s')}{\tau}\right) \leq 0$ implies that $\left(\frac{e_{Q,\pi}(s,a,s')}{\tau}\right) \leq f'(0)$ and hence $\kappa^\star(s,a,s') = \tau f'(0) - e_{Q,\pi}(s,a,s') \geq 0$.

Therefore, the pair $(w^\star_{Q,\tau,\pi}, \kappa^\star)$ defined in (14) satisfies all KKT conditions and is thus optimal. For those (s,a,s') where $d^{\pi_D}_{T^o}(s,a,s') = 0$, the corresponding $w(s,a,s')$ values do not affect the objective and can therefore be chosen arbitrarily. For consistency, we define them in the same manner as for the tuples (s,a,s') with $d^{\pi_D}_{T^o}(s,a,s') > 0$. Consequently, the closed-form optimal solution is

$$w^\star_{Q,\tau,\pi}(s,a,s') = \max\left(0,\, (f')^{-1}\left(\frac{e_{Q,\pi}(s,a,s')}{\tau}\right)\right), \quad \forall (s,a,s').$$

When $\tau = 0$,

$$\mathcal{L}(Q,\tau,w) := (1-\gamma)\,\mathbb{E}_{s\sim\mu,\, a\sim\pi_D(\cdot|s)}[Q(s,a)] + \mathbb{E}_{s,a,s'\sim d^{\pi_D}_{T^o}}\left[w(s,a,s')\, e_{Q,\pi}(s,a,s')\right].$$

It is straightforward to see that maximizing this with respect to $w \geq 0$ gives

$$w^\star_{Q,\tau,\pi}(s,a,s') = \begin{cases} +\infty, & \text{if } e_{Q,\pi}(s,a,s') > 0, \\ 0, & \text{if } e_{Q,\pi}(s,a,s') < 0, \\ \text{arbitrary}, & \text{if } e_{Q,\pi}(s,a,s') = 0. \end{cases}$$

We consider the arbitrary value as 0 to complete the proof. □

Lemma 5 (Monotonicity of f-divergence)**.** *Let P and Q be probability distributions on $\mathcal{X}$ with $P \ll Q$. If two generator functions f and g satisfy $f(x) \leq g(x)$ for all $x \geq 0$, then*

$$D_f(P\|Q) \leq D_g(P\|Q).$$

Proof. Since $f(x) \leq g(x)$ pointwise, it follows for Q-almost every x that $f\left(\frac{P(x)}{Q(x)}\right) \leq g\left(\frac{P(x)}{Q(x)}\right)$. Taking expectations with respect to Q preserves the inequality:

$$\mathbb{E}_{x\sim Q}\left[f\left(\tfrac{P(x)}{Q(x)}\right)\right] \leq \mathbb{E}_{x\sim Q}\left[g\left(\tfrac{P(x)}{Q(x)}\right)\right].$$

Hence $D_f(P\|Q) \leq D_g(P\|Q)$, with equality only when the two functions agree Q-almost everywhere. □

Lemma 2. *Let $f : [0,\infty)\to\mathbb{R}$ be any f-divergence generator with $f(1) = 0$. Assume there exists $\alpha\geq 0$ such that*

$$f(t) \;\leq\; \alpha\, f_{\mathrm{TV}}(t) \quad \textit{for all } t \geq 0, \qquad \textit{where } f_{\mathrm{TV}}(t) = \tfrac{1}{2}|t-1|.$$

Then for any policy π and any transition kernel $T \in \mathcal{T}$,

$$D_f\big(d_T^\pi(s,a,s') \,\big\|\, d_{T^o}^\pi(s,a,s')\big) \;\leq\; \alpha\, D_{\mathrm{TV}}(d_T^\pi(s,a,s'),\, d_{T^o}^\pi(s,a,s'))\,.$$

In particular, combining with Lemma 1 yields

$$D_f\big(d_T^\pi(s,a,s') \,\big\|\, d_{T^o}^\pi(s,a,s')\big) \;\leq\; \alpha\,\frac{\rho'}{1-\gamma}.$$

As a special case, if $f(t) \leq f_{\mathrm{TV}}(t)$ for all $t \geq 0$ (i.e., $\alpha = 1$), then

$$D_f\big(d_T^\pi(s,a,s') \,\big\|\, d_{T^o}^\pi(s,a,s')\big) \;\leq\; \frac{\rho'}{1-\gamma}.$$

Proof. The assumption $f(t) \leq \alpha f_{\mathrm{TV}}(t)$ for all $t \geq 0$ implies, by Lemma 5, that

$$D_f(P\|Q) \;\leq\; \alpha\, D_{f_{\mathrm{TV}}}(P\|Q) \;=\; \alpha\, D_{\mathrm{TV}}(P,Q).$$

Applying this to $P = d_T^\pi(s,a,s')$ and $Q = d_{T^o}^\pi(s,a,s')$ gives

$$D_f(d_T^\pi\|d_{T^o}^\pi) \;\leq\; \alpha\, D_{\mathrm{TV}}(d_T^\pi, d_{T^o}^\pi)\,.$$

Finally, using Lemma 1, which bounds $D_{\mathrm{TV}}(d_T^\pi, d_{T^o}^\pi) \leq \rho'/(1-\gamma)$, we obtain the desired inequality:

$$D_f(d_T^\pi\|d_{T^o}^\pi) \leq \alpha\,\frac{\rho'}{1-\gamma}.$$

□

Lemma 6. *For all $x \in \mathbb{R}$, the soft total variation generator satisfies*

$$f_{\mathrm{SoftTV}}(x) = \tfrac{1}{2}\log\big(\cosh(x-1)\big) \leq \tfrac{1}{2}|x-1| = f_{\mathrm{TV}}(x).$$

Proof. Let $t = x-1$. Then $\cosh t = \frac{e^t+e^{-t}}{2}$. Using this, we have

$$\cosh t = \frac{e^t+e^{-t}}{2} \leq \frac{e^{|t|}+e^{|t|}}{2} = e^{|t|}.$$

Taking logarithms gives $\log(\cosh t) \leq |t|$. Multiplying both sides by $\frac{1}{2}$ yields

$$\tfrac{1}{2}\log(\cosh t) \leq \tfrac{1}{2}|t|,$$

and substituting back $t = x-1$ completes the proof. □

Environment	Expert data size N	Robustness param. ρ_r	Policy layers	Activation	Max steps	Learning rate	LR decay	Decay rate	Decay freq.
Hopper-v3	2000	0.2	(256, 256)	Tanh	2M	1×10^{-4}	True	0.9	10k
HalfCheetah-v3	3000	0.3	(256, 256)	Tanh	2M	1×10^{-5}	True	0.9	10k
Walker2d-v3	6000	0.5	(256, 256)	Tanh	2M	1×10^{-4}	True	0.95	10k
Ant-v3	8000	0.6	(256, 256)	Tanh	2M	1×10^{-4}	True	0.9	10k

Table 1: Hyperparameter configuration for DRBC across MuJoCo environments.

Component	Architecture	Hidden units	Activation	Batch size	Learning rate	Steps	γ	Update ratio (policy : Q, τ)
Policy π	TanhGaussian MLP	(256, 256)	ReLU	512	5×10^{-5}	1M	0.99	1:1
Q-function Q	MLP	(256, 256)	ReLU	512	5×10^{-5}	1M	0.99	1:1

Table 2: Hyperparameter configuration for BE-DROIL.

B Experimental Settings

Environments and Expert Data. We evaluate all algorithms on four continuous-control benchmarks from the MuJoCo suite [42]: Hopper-v3, HalfCheetah-v3, Walker2d-v3, and Ant-v3, following the setup of Panaganti et al. [30]. Expert demonstrations are generated using pre-trained TD3 [12] policies available in the RL Baselines3 Zoo repository [35]. The number of expert trajectories for each environment matches the DRBC configuration to ensure fair comparison and consistent dataset coverage across all methods. Details of the expert dataset sizes for each environment are provided in Table 1.

DRBC and BC Baselines. Both DRBC and BC are implemented in PyTorch using identical policy architectures composed of two hidden layers with 256 units and `tanh` activations. Each model is trained for 2×10^6 gradient steps with a batch size of 256, using the Adam optimizer. Learning rates and decay schedules follow Panaganti et al. [30]: the initial learning rate is 1×10^{-4} for all environments except HalfCheetah (1×10^{-5}), decayed exponentially by a factor of 0.9–0.95 every $10,000$ steps. The DRBC robustness parameter ρ_r is set to $\{0.2, 0.3, 0.5, 0.6\}$ for Hopper, HalfCheetah, Walker2d, and Ant respectively, as summarized in Table 1. BC uses the same architecture and optimizer settings but excludes the robust dual updates and uncertainty term.

BE-DROIL. Our proposed BE-DROIL algorithm is implemented in PyTorch. The policy is parameterized by a `TanhGaussianPolicy`, consisting of two fully connected layers with 256 hidden units , followed by a Gaussian distribution transformed through $\tanh$ to enforce action bounds. Both the policy and $Q-$function are trained with a learning rate of 5×10^{-5} using the Adam optimizer and a batch size of 512. The $Q-$function follows a two-layer MLP with 256 hidden units and `ReLU` activations. The discount factor is fixed at $\gamma = 0.99$.

The loss function $L_\pi(\cdot)$ minimizes the mean-squared error between the learner's and expert's action distributions,

$$L_\pi(s) = \mathrm{MSE}(a, a_D), \quad a \sim \pi(\cdot|s),\ a_D \sim \pi_D(\cdot|s),$$

where π and π_D denote the learner and expert policies, respectively. Training proceeds for 1 million gradient steps, alternating between one policy update and one joint update of the Q-function and τ network. Unlike DRBC, no environment-specific hyperparameter tuning is performed, i.e. identical configurations are used across all domains to ensure fair cross-environment evaluation. The complete BE-DROIL configuration is listed in Table 2.

Evaluation Protocol. Each trained policy is evaluated in perturbed test environments obtained by varying physical parameters such as gravity, actuator range, and joint stiffness to induce model mismatch. Performance is reported as the mean and standard deviation of episodic returns over 100 rollouts with independent random seeds.

C Details of f-divergence Options

We summarize several generator functions f commonly employed in DICE-style objectives, along with their inverse gradients and the corresponding mappings from a scaled score $z := \frac{e_{Q,\pi}(s,a,s')}{\tau}$ to the optimal nonnegative weight $w^\star_{Q,\tau,\pi}(s,a,s')$. The latter is obtained from the first-order condition

$f'(w) = z$ (projected to $w \geq 0$ when required). These definitions and mappings are provided in Table 3.

We define the following helper functions:

$$\mathrm{ReLU}(x) := \max\{0, x\}, \qquad \mathrm{ELU}(x) := \begin{cases} e^x - 1, & x < 0, \\ x, & x \geq 0, \end{cases}$$

$$f_{\text{soft-}\chi^2}(x) := \begin{cases} x \log x - x + 1, & 0 < x < 1, \\ (x-1)^2, & x \geq 1, \end{cases} \qquad \left(f'_{\text{soft-}\chi^2}\right)^{-1}(y) := \begin{cases} e^{\,y}, & y < 0, \\ y + 1, & y \geq 0. \end{cases}$$

Divergence	$\boldsymbol{f(x)}$	$\boldsymbol{\left(f'\right)^{-1}(y)}$	$\boldsymbol{w^\star_{Q,\tau,\pi}}$ **for** $\boldsymbol{z = \frac{e_{Q,\pi}}{\tau}}$
Soft TV	$\frac{1}{2}\log\big(\cosh(x-1)\big)$	$\tanh^{-1}(2y) + 1$	$\mathrm{ReLU}\big(\tanh^{-1}(2z) + 1\big)$
Total Variation	$\frac{1}{2}\lvert x - 1\rvert$	—	—
KL (forward)	$x \log x$	$e^{\,y-1}$	$\exp(z) - 1$
χ^2	$\frac{1}{2}(x-1)^2$	$y + 1$	$\mathrm{ReLU}(z + 1)$
Soft χ^2	$f_{\text{soft-}\chi^2}(x)$	$\left(f'_{\text{soft-}\chi^2}\right)^{-1}(y)$	$\mathrm{ELU}(z) + 1$

Table 3: Summary of common f-generators, their inverse gradients, and the closed-form mappings from the normalized score z to the optimal weight $w^\star_{Q,\tau,\pi}$. A dash indicates that no simple closed form exists for that column.