SUNDER: A GENERAL FRAMEWORK FOR DECOUPLED STATE–POLICY OPTIMIZATION IN OFFLINE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Distribution Correction Estimation (DICE) methods directly optimize policy-induced state-action occupancies, enabling distributional objectives beyond standard behavior-regularized reinforcement learning. However, joint occupancy regularization couples state visitation and action selection through a single regularization strength, even though these two forms of distribution shift may require different control. Naively separating them breaks the explicit affine structure underlying the standard DICE strong-duality argument. We introduce SUNDER (State–policy UNcoupling for Distribution Estimation via Ratio factorization), a general framework that separately controls state visitation and policy deviation. Although the resulting state–policy formulation is non-affine, we show that it admits a hidden convex structure that preserves strong duality and enables tractable elimination of the explicit optimization variables. We instantiate SUNDER for reward maximization, state-entropy maximization, and learning from observations. Under mismatches between state coverage and action quality, Opt-SUNDER outperforms OptiDICE across six maze domains, while achieving the best average performance among competitive recent baselines on 27 Minari tasks. Ent-SUNDER consistently outperforms SEMDICE across seven maze and four DMC domains with skewed dataset state distributions, and LfO-SUNDER achieves competitive performance on MuJoCo and Adroit tasks.

1 INTRODUCTION

Distribution correction estimation (DICE) was originally developed to estimate the state-action occupancy of a given target policy from off-policy data, prominently for off-policy evaluation(Nachum et al., 2019). Joint state-action correction is natural in this setting, since evaluating a target policy requires accounting for both its action selection and the resulting shift in state visitation. The role of DICE has since expanded to off-policy optimization, targeting objectives such as reward maximization(Lee et al., 2021), state-entropy maximization (SEM)(Lee et al., 2025), and learning from observations (LfO)(Kim et al., 2022a). This occupancy-based view provides a more flexible perspective than standard behavior-regularized reinforcement learning (RL), as it directly optimizes policy-induced distributions rather than only constraining policy deviation.

However, this broader role of DICE makes its inherited joint regularization potentially restrictive. In reward maximization, partial or truncated data collection(Ubellacker et al., 2024; Spencer et al., 2020) can misalign state coverage and action quality, making a single joint regularizer impose an unnecessary trade-off between the two. More fundamentally, the tension is structural for state-distribution objectives. SEM seeks visitation beyond the dataset state distribution, while LfO seeks expert-like states rather than behavior-data states. Thus, these objectives need not favor state regularization toward the offline data, even when policy regularization remains useful for limiting Out-of-Distribution (OOD) actions.

Why, then, not simply regularize state visitation and action selection separately? The difficulty is that this separation breaks the explicit affine structure used by standard DICE formulations to establish strong duality and eliminate otherwise intractable occupancy variables. To address this difficulty, we introduce SUNDER (State–policy UNcoupling for Distribution Estimation via Ratio factorization),

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the first general distribution-correction framework to separately control state and policy. Although SUNDER uses a non-affine state–policy formulation, its hidden convex structure establishes strong duality and allows the state–policy variables to be eliminated. Moreover, the framework is general enough to cover a broad range of existing DICE formulations.

We evaluate SUNDER from two perspectives. We first test whether SUNDER realizes the intended separation of state and policy regularization. For reward maximization, we instantiate SUNDER as Opt-SUNDER, the decoupled counterpart of OptiDICE, a reward-maximizing DICE method (Lee et al., 2021). Across six maze domains with controlled state–action quality mismatch, Opt-SUNDER outperforms OptiDICE. We then test whether decoupling sacrifices performance in standard settings. Across 27 Minari tasks spanning MuJoCo, Adroit, and AntMaze, Opt-SUNDER achieves the highest average performance in each domain. Second, we test whether SUNDER extends beyond reward maximization to state-distribution objectives. We consider state-entropy maximization (SEM) and learning from observations (LfO). In SEM, state regularization toward the dataset can hinder entropy maximization. In LfO, matching expert states does not require matching the behavior-data state distribution. For SEM, Ent-SUNDER, the decoupled counterpart of SEMDICE, a DICE method for state-entropy maximization (Lee et al., 2025), consistently outperforms SEMDICE across seven maze and four DMC domains with skewed dataset state distributions. For LfO, LfO-SUNDER achieves the best average performance among the evaluated baselines on Adroit while remaining competitive on MuJoCo.

2 PRELIMINARIES

Problem Setup and Notation. We consider a discounted Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, d_0, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $P(s' \mid s, a)$ is the transition kernel, $r(s, a)$ is the reward function, d_0 is the initial-state distribution, and $\gamma \in [0, 1)$ is the discount factor. For a policy π, we denote its normalized discounted state-action occupancy by

$$d^\pi(s, a) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \Pr_\pi(s_t = s, a_t = a). \tag{1}$$

Its state marginal is $\hat{d}^\pi(s) := \sum_a d^\pi(s, a)$, so that $d^\pi(s, a) = \hat{d}^\pi(s)\pi(a \mid s)$. More generally, for any state-action occupancy $d(s, a)$, we denote its state marginal by $\hat{d}(s) := \sum_a d(s, a)$ and its state-transition marginal by $d'(s, s') := \sum_a d(s, a)P(s' \mid s, a)$. We write $q_{\mathcal{B}}(s, a)$ for the state-action distribution induced by the offline buffer $\mathcal{B}$, and $\pi_{\mathcal{B}}(a \mid s) = q_{\mathcal{B}}(s, a)/\hat{q}_{\mathcal{B}}(s)$ for the corresponding behavior policy. Subscripts E and I similarly denote expert and imperfect data, respectively.

Strong Duality. Consider a maximization problem

$$\max_x J(x) \quad \text{s.t.} \quad h(x) = 0, \qquad \mathcal{C} := \{(u, t) : \exists\, x,\ h(x) = u,\ t \le J(x)\}\,.$$

If $\mathcal{C}$ is convex, then under standard regularity conditions a supporting-hyperplane argument establishes strong duality. A standard sufficient case is when J is concave and h is affine, for which the convexity of $\mathcal{C}$ follows. In this paper, we refer to this case as the *standard affine conditions*.

DICE-Based Optimization. Policy optimization can be equivalently expressed as optimization over discounted state-action occupancy measures $\max_{\pi \in \Pi} F(d^\pi) = \max_{d \in \mathcal{C}_{\text{flow}}} F(d)$. DICE methods optimize the latter occupancy formulation, where $\mathcal{C}_{\text{flow}}$ is the set of nonnegative occupancies satisfying Bellman-flow constraints

$$\sum_a d(s', a) = (1 - \gamma)d_0(s') + \gamma \sum_{s,a} d(s, a)P(s' \mid s, a), \quad \forall s'. \tag{2}$$

For nonnegative d, Eq. 2 characterizes exactly the policy-induced occupancies defined in Eq. 1.

Across DICE methods, the state-action occupancy d lies in $\mathcal{C}_{\text{flow}}$, while the task objective may additionally depend on induced occupancy variables such as $\hat{d}$ and d'. A broad class of DICE

objectives can be written as

$$\max_{d,\hat{d},d' \geq 0} \underbrace{G(\hat{d}, d') + \textstyle\sum_{s,a} d(s,a)c(s,a)}_{\text{Task objective}} - \underbrace{\alpha \textstyle\sum_{s,a} q_{\mathcal{B}}(s,a) f\left(\frac{d(s,a)}{q_{\mathcal{B}}(s,a)}\right)}_{\text{Joint occupancy regularization}} \tag{3}$$
$$\text{s.t.} \quad d \in \mathcal{C}_{\text{flow}}, \qquad \hat{d}(s) = \sum_a d(s,a), \qquad d'(s,s') = \sum_a d(s,a) P(s' \mid s,a).$$

Here, $\hat{d}$ and d' denote the state marginal and state-transition occupancy induced by d, respectively. These auxiliary variables and their corresponding constraints are introduced only when required by G. The function $c(s,a)$ specifies a local state-action objective, while q denotes the method-specific offline state-action distribution.

This formulation separates two roles. The task terms G and c specify the desired policy-induced distribution, while $D_f(d\|q)$ controls deviation of the joint state-action occupancy from offline data. Concretely, OptiDICE (Lee et al., 2021) uses $c(s,a) = r(s,a)$ and $G = 0$. SEMDICE (Lee et al., 2025), designed for state-entropy maximization (SEM), uses $G(\hat{d}) = H(\hat{d})$ and $c = 0$, and therefore introduces the state marginal $\hat{d}$. LobsDICE (Kim et al., 2022a), designed for learning from observations (LfO), uses $G(d') = -D_{\text{KL}}(d'\|d'^E)$ and $c = 0$, and therefore introduces the state-transition occupancy d'. In all three methods, the state-action occupancy d must lie in $\mathcal{C}_{\text{flow}}$.

However, direct optimization over $\mathcal{C}_{\text{flow}}$ generally requires access to the transition dynamics. DICE methods therefore typically use strong duality to eliminate the explicit occupancy variables. We illustrate this standard procedure with OptiDICE.

OptiDICE formulation. OptiDICE (Lee et al., 2021) considers

$$\max_{d \geq 0} \underbrace{\sum_{s,a} d(s,a) r(s,a)}_{\text{Return, Affine in } d} - \alpha \underbrace{\sum_{s,a} q_{\mathcal{B}}(s,a) f\left(\frac{d(s,a)}{q_{\mathcal{B}}(s,a)}\right)}_{\text{State-action regularization, Convex in } d}$$
$$\text{s.t.} \quad \underbrace{\sum_a d(s',a) - \gamma \sum_{s,a} d(s,a) P(s' \mid s,a)}_{\text{Bellman flow constraints, Affine in } d} = (1-\gamma) d_0(s'), \qquad \forall s'.$$

The first term maximizes the expected return, while the second regularizes the target occupancy toward the offline data distribution. The constraint ensures that d lies in the valid set $\mathcal{C}_{\text{flow}}$.

As discussed in Strong Duality, the objective is concave in d, while the Bellman-flow constraints are affine. Thus, OptiDICE satisfies the *standard affine conditions* for strong duality. We therefore form the Lagrangian and exchange the max–min order. The inner maximization over d can then be eliminated analytically using the convex conjugate of f, yielding

$$\min_V (1-\gamma)\mathbb{E}_{s \sim d_0}[V(s)] + \alpha \mathbb{E}_{(s,a) \sim q_{\mathcal{B}}} \left[f^*\left(\frac{r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')] - V(s)}{\alpha} \right) \right], \tag{4}$$

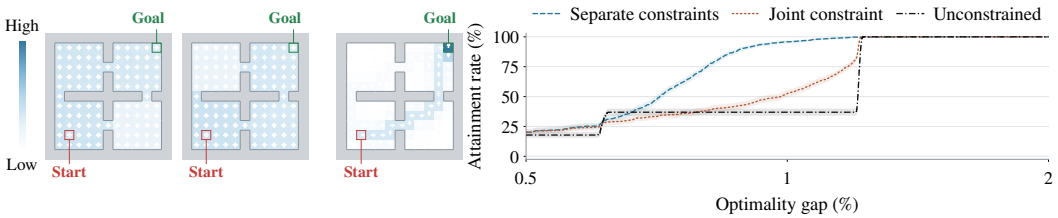
where $f^*(y) = \sup_{x \geq 0} \{xy - f(x)\}$. This is the standard strategy used by DICE methods to eliminate otherwise intractable occupancy variables. Standard DICE formulations therefore favor concave occupancy objectives with affine constraints, for which strong duality follows directly from the *standard affine conditions*.

3 UNCOUPLING FOR DISTRIBUTION ESTIMATION CORRECTION

In this section, we present the general SUNDER framework and show how it separately controls the state distribution and policy while preserving strong duality for eliminating intractable variables.

3.1 WHY SEPARATE STATE AND POLICY REGULARIZATION?

We first examine how separate and joint regularization behave under a plausible mismatch between state coverage and action quality. We construct such a mismatch in Fourrooms by collecting an



(a) Offline buffers (b) Optimal policy (c) Optimality attainment

Figure 1: Visualization of Fourrooms.

offline buffer with a uniform policy and randomly truncating data from a spatial region. This truncation skews the dataset state distribution, which differs visibly from the visitation distribution of the optimal policy in Fig. 1(a,b).

We then optimize reward-maximizing policies under three regularization schemes: joint KL regularization, separate state and policy KL regularization, and no regularization. Since our Fourrooms environment includes stochastic transitions, we estimate its transition matrix from the offline buffer, making unconstrained optimization susceptible to OOD exploitation. For each of 1,000 independently generated buffers, we select the best regularization coefficients over 10^{-7} to 1.

Fig. 1(c) reports, for each optimality-gap threshold, how many of the 1,000 instances achieve a policy within that gap of the optimum. At a 1% optimality-gap threshold, separate regularization succeeds in 958 of 1,000 instances, compared with 526 for joint regularization and 369 without regularization. The unconstrained variant is vulnerable to OOD exploitation, whereas joint regularization avoids part of this failure at the cost of unnecessarily constraining state visitation, motivating separate control of the two shifts.

3.2 Apparent Nonconvexity of Separate Regularization

A natural question is why we cannot simply regularize the state distribution $\hat{d}$ and the policy π separately. Doing so no longer preserves the explicit affine structure underlying the standard DICE strong-duality argument. Let π_{ref} denote the reference policy in the following formulation:

$$\max_{\hat{d},d',\pi\geq 0} \underbrace{G(\hat{d},d') + \sum_{s,a}\hat{d}(s)\pi(a\mid s)c(s,a)}_{\text{Task objective, nonlinear in } (\hat{d},\,d',\,\pi)} \underbrace{- \alpha \sum_{s,a}\hat{d}(s)\pi_{\mathrm{ref}}(a\mid s) f\left(\frac{\pi(a\mid s)}{\pi_{\mathrm{ref}}(a\mid s)}\right)}_{\text{Policy regularization, nonlinear in } (\hat{d},\,\pi)} \tag{5a}$$

$$\text{s.t.}\quad \hat{d}(s)\left(\sum_a \pi(a\mid s) - 1\right) = 0, \qquad \forall s, \quad \text{– Policy normalization; bilinear in } (\hat{d},\,\pi) \tag{5b}$$

$$d'(s,s') = \hat{d}(s)\sum_a \pi(a\mid s)P(s'\mid s,a),\ \forall s,s', \quad \text{–State-transition consistency; bilinear in } (\hat{d},\,\pi) \tag{5c}$$

$$\underbrace{\hat{d}(s')\sum_{a'}\pi(a'\mid s') - \gamma\sum_{s,a}\hat{d}(s)\pi(a\mid s)P(s'\mid s,a)}_{\text{Bellman flow, bilinear in } (\hat{d},\,\pi)} = (1-\gamma)d_0(s'), \qquad \forall s'. \tag{5d}$$

Under the policy-normalization and Bellman-flow constraints, $\hat{d}(s)\pi(a\mid s)$ forms a valid state-action occupancy induced by π. Equation 5c further defines d' as its corresponding state-transition occupancy. Thus, $\hat{d}$ and d' describe the distributions induced by the policy, while π controls action selection.

We use $G(\hat{d},d')$ to specify a task objective over the required induced distributions and $c(s,a)$ to define a state-action objective. As in the general DICE formulation, auxiliary occupancy variables such as d' need only be introduced when required by the task objective. The f-divergence term, in contrast, regularizes only the policy toward π_{ref}. By separating the task objective over induced distributions from policy regularization, the formulation independently controls state visitation and action selection, achieving the intended state-policy decoupling.

However, this separation introduces the bilinear term $\hat{d}(s)\pi(a \mid s)$, which appears in both the objective and the constraints. The resulting formulation is therefore no longer explicitly affine in $(\hat{d}, d', \pi)$, so the standard DICE strong-duality argument cannot be applied directly. We next show that, despite this apparent nonconvexity, the formulation retains a hidden convex structure that restores the strong-duality argument.

3.3 Hidden Convexity

Despite the apparent nonlinearity in $(\hat{d}, d', \pi)$, the problem has a hidden convex structure. Let $J(\hat{d}, d', \pi)$ denote the objective in Eq. 5a, and let u_π, $u_{d'}$, and u_d denote the residuals of the constraints in Eqs. 5b–5d, respectively. Let $u = (u_\pi, u_{d'}, u_d)$, and define

$$\mathcal{C} = \left\{ (u, t) : \exists\, \hat{d}, d', \pi \geq 0 \text{ such that } u = u(\hat{d}, d', \pi),\ t \leq J(\hat{d}, d', \pi) \right\}. \tag{5}$$

Proposition 1 (Hidden Convexity)**.** *For G jointly concave in $(\hat{d}, d')$, $\mathcal{C}$ defined in Eq. 5 is convex.*

Proof. Consider any $(u_1, t_1), (u_2, t_2) \in \mathcal{C}$, corresponding to $(\hat{d}_1, d'_1, \pi_1)$ and $(\hat{d}_2, d'_2, \pi_2)$. For $\theta \in [0, 1]$, define $\hat{d}_\theta = \theta\hat{d}_1 + (1-\theta)\hat{d}_2$, $d'_\theta = \theta d'_1 + (1-\theta)d'_2$, and $\pi_\theta = \big(\theta\hat{d}_1\pi_1 + (1-\theta)\hat{d}_2\pi_2\big)/\hat{d}_\theta$. with $\pi_\theta(\cdot \mid s)$ chosen arbitrarily when $\hat{d}_\theta(s) = 0$.

By construction, $\hat{d}_\theta\pi_\theta = \theta\hat{d}_1\pi_1 + (1-\theta)\hat{d}_2\pi_2$. Therefore, all constraint residuals satisfy $u(\hat{d}_\theta, d'_\theta, \pi_\theta) = \theta u(\hat{d}_1, d'_1, \pi_1) + (1-\theta)u(\hat{d}_2, d'_2, \pi_2)$.

The state-action term is linear in the induced occupancy under this construction:

$$\sum_{s,a} \hat{d}_\theta(s)\pi_\theta(a \mid s)c(s, a) = \theta \sum_{s,a} \hat{d}_1(s)\pi_1(a \mid s)c(s, a) + (1-\theta) \sum_{s,a} \hat{d}_2(s)\pi_2(a \mid s)c(s, a).$$

Moreover, by joint concavity of G, $G(\hat{d}_\theta, d'_\theta) \geq \theta G(\hat{d}_1, d'_1) + (1-\theta)G(\hat{d}_2, d'_2)$. By convexity of f,

$$\hat{d}_\theta(s)\pi_{\text{ref}}(a \mid s) f\left(\frac{\pi_\theta}{\pi_{\text{ref}}}\right) \leq \theta\hat{d}_1(s)\pi_{\text{ref}}(a \mid s) f\left(\frac{\pi_1}{\pi_{\text{ref}}}\right) + (1-\theta)\hat{d}_2(s)\pi_{\text{ref}}(a \mid s) f\left(\frac{\pi_2}{\pi_{\text{ref}}}\right).$$

Hence, $J(\hat{d}_\theta, d'_\theta, \pi_\theta) \geq \theta J(\hat{d}_1, d'_1, \pi_1) + (1-\theta)J(\hat{d}_2, d'_2, \pi_2)$. Let $u_\theta = \theta u_1 + (1-\theta)u_2$ and $t_\theta = \theta t_1 + (1-\theta)t_2$. Since $t_i \leq J(\hat{d}_i, d'_i, \pi_i)$, $t_\theta \leq J(\hat{d}_\theta, d'_\theta, \pi_\theta)$. Therefore $(u_\theta, t_\theta) \in \mathcal{C}$. Thus, for any jointly concave G, $\mathcal{C}$ is convex, establishing strong duality and enabling the elimination of otherwise intractable optimization variables, as in DICE.

3.4 SUNDER-Based Optimization

Having established strong duality, we now derive a dual objective from the general SUNDER primal in Eqs. 5a–5d. Whereas the standard DICE occupancy formulation eliminates the joint state-action occupancy d directly, SUNDER separately represents state visitation and action selection through $\hat{d}$ and π, while allowing additional induced occupancies such as d' when required by the task objective. We first eliminate π, and then jointly eliminate the remaining occupancy variables $(\hat{d}, d')$. We refer to this procedure as *SUNDER-based optimization*.

First, we form the Lagrangian of the general SUNDER primal. Let $V(s)$, $\lambda(s)$, and $Q(s, s')$ be the Lagrange multipliers for the Bellman-flow, policy-normalization, and state-transition consistency constraints in Eqs. 5d, 5b, and 5c, respectively. By Proposition 1, strong duality allows us to exchange the optimization order: $\max_{\hat{d}, d', \pi \geq 0} \min_{V, \lambda, Q} \mathcal{L} = \min_{V, \lambda, Q} \max_{\hat{d}, d', \pi \geq 0} \mathcal{L}$.

To simplify the elimination of π, we collect the terms multiplying $\pi(a \mid s)$ and define

$$e_\pi(s, a; V, \lambda, Q) := c(s, a) + \gamma\mathbb{E}_{s' \sim P(\cdot \mid s, a)}[V(s')] - V(s) - \lambda(s) - \mathbb{E}_{s' \sim P(\cdot \mid s, a)}[Q(s, s')].$$

The Lagrangian can then be written as

$$\mathcal{L} = (1-\gamma)\mathbb{E}_{s \sim d_0}[V(s)] + G(\hat{d}, d') + \sum_{s,s'} d'(s, s')Q(s, s')$$
$$+ \sum_s \hat{d}(s) \left[\lambda(s) + \sum_a \pi(a \mid s) e_\pi(s, a; V, \lambda, Q) - \alpha \sum_a \pi_{\text{ref}}(a \mid s) f\left(\frac{\pi(a \mid s)}{\pi_{\text{ref}}(a \mid s)}\right) \right].$$

Table 1: SUNDER instantiations defined by the task objective G, one-step local term $c(s, a, s')$, policy reference, and relation to prior formulations. Detailed derivations are provided in Appendix A.

Method	G	$c(s, a, s')$	Policy ref.	DICE Relation
Opt-SUNDER	$-\beta D_f\left(\hat{d} \,\|\, \hat{q}_{\mathcal{B}}\right)$	$r(s, a)$	$\pi_{\mathcal{B}}$	OptiDICE
Ent-SUNDER	$-D_f\left(\hat{d} \,\|\, u\right)$	0	$\pi_{\mathcal{B}}$	SEMDICE
SAFE-SUNDER	$-D_f\left(\hat{d} \,\middle\|\, \frac{\hat{q}_{\mathcal{B}} - \omega \hat{q}_I}{1-\omega}\right)$	0	$\pi_{\mathcal{B}}$	SafeDICE
SMO-SUNDER	$-\beta D_f\left(\hat{d} \,\|\, \hat{q}_I\right)$	$\log \frac{\hat{q}_E(s)}{\hat{q}_I(s)}$	π_I	SMODICE
Demo-SUNDER	$-\beta D_{\mathrm{KL}}\left(\hat{d} \,\|\, \hat{q}_{\mathcal{B}}\right)$	$\log \frac{q_E(s,a)}{q_{\mathcal{B}}(s,a)}$	$\pi_{\mathcal{B}}$	DemoDICE
Lobs-SUNDER	$-D_{\mathrm{KL}}(d' \,\|\, q'_I) - \beta D_{\mathrm{KL}}\left(\hat{d} \,\|\, \hat{q}_I\right)$	$\log \frac{q'_E(s,s')}{q'_I(s,s')}$	π_I	LobsDICE
LfO-SUNDER	$-D_f(\omega d' + (1-\omega) q'_I \,\|\, \omega q'_E + (1-\omega) q'_I)$	0	π_I	Proposed in this work

For fixed $\hat{d}$, d', V, λ, and Q, the maximization over π separates across states. By Fenchel duality,

$$\max_{\pi(\cdot|s)\geq 0} \left\{ \sum_a \pi(a \mid s) e_\pi(s,a) - \alpha \sum_a \pi_{ref}(a \mid s) f\left(\frac{\pi}{\pi_{ref}}\right) \right\} = \alpha\, \mathbb{E}_{a\sim\pi_{ref}(\cdot|s)} \left[f^*\left(\frac{e_\pi(s,a)}{\alpha}\right)\right].$$

After eliminating π, define $e_{\hat{d}}(s; V, \lambda, Q) := \lambda(s) + \alpha\, \mathbb{E}_{a\sim\pi_D(\cdot|s)} \left[f^*\left(\frac{e_\pi(s,a;V,\lambda,Q)}{\alpha}\right)\right]$. The remaining maximization is then $\max_{\hat{d},d'\geq 0} \left\{ G(\hat{d}, d') + \langle \hat{d}, e_{\hat{d}} \rangle + \langle d', Q \rangle \right\}$.

Since G is jointly concave in $(\hat{d}, d')$, $-G$ is jointly convex, with convex conjugate $(-G)^*(e_{\hat{d}}, Q) := \sup_{\hat{d},d'\geq 0} \left\{ \langle \hat{d}, e_{\hat{d}} \rangle + \langle d', Q \rangle + G(\hat{d}, d') \right\}$. Hence, the general SUNDER dual is

$$\min_{V,\lambda,Q} (1-\gamma)\mathbb{E}_{s\sim d_0}[V(s)] + (-G)^*(e_{\hat{d}}, Q), \tag{6}$$

where $e_{\hat{d}}$ is as defined above. Thus, SUNDER eliminates π through the conjugate of the policy regularizer and the induced occupancy variables through the conjugate of the task objective. Different task objectives correspond to different choices of G, c, and the induced occupancy variables, which we instantiate next.

3.5 SUNDER Instantiations

Table 1 summarizes choices for the SUNDER variants considered in this paper. Substituting each choice into the general SUNDER dual in Eq. 6 yields the corresponding DICE Relation (Lee et al., 2021; 2025; Jang et al., 2023; Kim et al., 2022b;a). The resulting objectives, their correspondence to prior DICE formulations, and complete derivations are provided in Appendix A.

Here, u denotes the uniform state distribution and $\omega \in [0, 1)$ denotes the mixture coefficient. The data-source subscripts follow the notation introduced in Section 2. For SAFE-SUNDER, $\hat{q}_{\mathcal{B}}$ denotes the state marginal of the unlabeled offline data, and $(\hat{q}_{\mathcal{B}} - \omega \hat{q}_I)/(1-\omega)$ is the corrected preferred-state distribution. The corresponding behavior policy is denoted by $\pi_{\mathcal{B}}$.

Unlike standard DICE formulations, SUNDER allows the state-distribution objective and policy regularization to be controlled independently. This decoupled structure extends naturally from state-level to transition-level objectives. We next examine how this decoupling affects optimization and performance across representative settings.

3.6 POLICY EXTRACTION

The SUNDER objectives above recover the optimal policy density ratio as

$$w_\pi^*(s,a) = (f^*)'(e_\pi(s,a;V^*,\lambda^*,Q^*)/\alpha) = \frac{\pi^*(a \mid s)}{\pi_{\mathrm{ref}}(a \mid s)},$$

where $e_\pi(s,a;V,\lambda,Q) = c(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[V(s')\right] - V(s) - \lambda(s) - \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[Q(s,s')\right].$

Although w_π^* specifies the target policy relative to π_{ref}, execution requires an explicit parametric policy. The simplest choice is weighted behavioral cloning, using w_π^* as the sample weight. For any fixed state s and policy-fitting loss $\ell_\theta(s,a)$,

$$\mathbb{E}_{a \sim \pi_{\mathrm{ref}}(\cdot|s)}\left[w_\pi^*(s,a)\ell_\theta(s,a)\right] = \mathbb{E}_{a \sim \pi^*(\cdot|s)}\left[\ell_\theta(s,a)\right].$$

A standard choice is weighted behavioral cloning with a Gaussian policy. This is convenient, but restricts the extracted conditional action distribution to a simple parametric family. Recent work therefore explores more expressive policies for extracting the DICE target policy.

Diffusion-DICE learns a diffusion-based behavior policy and uses the estimated distribution ratio to guide generation toward the target policy. In practice, it introduces an additional guidance scale beyond the theoretically derived guidance term, requiring environment-specific tuning that can materially affect performance. We instead introduce a conditional flow matching (CFM)(Tong et al., 2024) policy extractor that uses the density ratio directly as a sample weight. Specifically, we use the recovered density ratio to weight the CFM objective. Let $x_t = (1-t)\epsilon + ta,\ t \sim \mathcal{U}[0,1],\ \epsilon \sim \mathcal{N}(0,I)$, with target velocity $u_t = a - \epsilon$. We optimize

$$\min_\theta \mathbb{E}_{\substack{s \sim \mu,\, a \sim \pi_{\mathrm{ref}}(\cdot|s),\\ t \sim \mathcal{U}[0,1],\, \epsilon \sim \mathcal{N}(0,I)}} \left[\mathrm{sg}[w_\pi^*(s,a)]\, \|v_\theta(x_t,t,s) - u_t\|_2^2\right].$$

By the density-ratio identity above, this is equivalent to the standard CFM objective under π^*. With direct weighting, weighted CFM introduces no additional guidance-scale hyperparameter. Appendix D shows a substantial performance gap between flow-matching and Gaussian policy extractors.

4 EXPERIMENTS

We evaluate SUNDER from two perspectives. First, we test whether SUNDER realizes the intended separation of state and policy regularization. For reward maximization, we construct controlled settings in which state coverage and action quality favor different amounts of regularization. This lets us isolate whether Opt-SUNDER can adjust the two forms of distribution shift independently. We then evaluate Opt-SUNDER on standard offline RL benchmarks to test whether this added flexibility sacrifices performance in ordinary settings. Second, we test whether the same framework extends beyond reward maximization to state-distribution objectives. We consider state-entropy maximization and learning from observations, where the desired state distribution need not match the offline-data distribution. Appendix C formalizes this advantage, showing that joint DICE regularization cannot in general recover optima attainable by SUNDER.

4.1 STATE–POLICY DECOUPLING IN REWARD MAXIMIZATION

To test whether Opt-SUNDER realizes the intended separation of state and policy regularization, we construct two complementary offline datasets in which state and action quality are controlled independently.

We define state quality by proximity to the reward-optimal path in the maze, and action quality by the performance of the behavior policy. In the state-good, action-bad setting, the dataset contains high-quality states but actions from a poor policy; in the state-bad, action-good setting, we reverse these properties. These two settings induce opposite regularization preferences. Accordingly, we fix the unfavorable regularizer at 0.1 and vary the favorable one.

Table 2(a) shows the intended effect of decoupling. When states are poor but actions are good, Opt-SUNDER weakly constrains the state distribution and more strongly regularizes the policy, improving from 53.94 to 71.67. When states are good but actions are poor, it does the opposite,

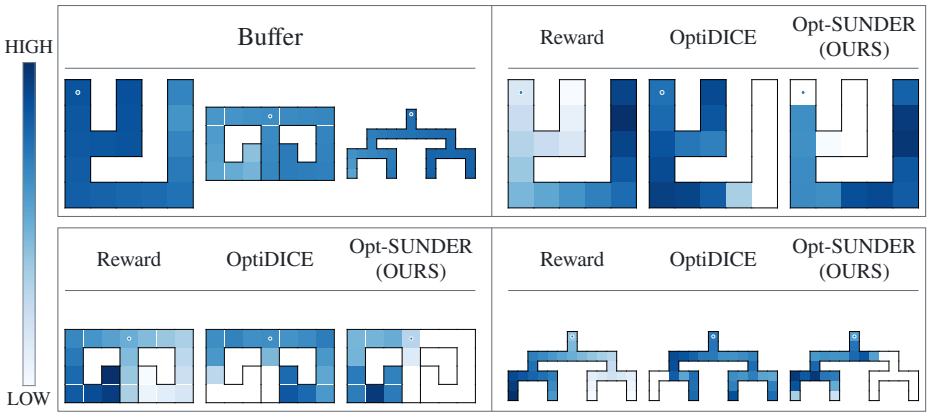


Figure 2: Maze visualization of the controlled state–action quality mismatch setting.

Table 2: Reward-maximization results in the controlled settings; standard offline RL benchmarks.

(a) Average return over six mazes (mean ± SE over 8 seeds; overall best results in bold). Regularization strengths: OptiDICE, $\alpha \to \rho(s,a)$; Opt-SUNDER, $\alpha \to \pi(a \mid s)$ and $\beta \to \hat{d}(s)$.

State Bad, Action Good			State Good, Action Bad		
α	**OptiDICE** (α)	**Opt-SUNDER** (α; $\beta = 0.1$)	α/β	**OptiDICE** (α)	**Opt-SUNDER** (β; $\alpha = 0.1$)
0.1	50.74 ± 1.62	53.94 ± 1.66	0.1	45.59 ± 0.95	42.40 ± 1.00
0.3	51.78 ± 1.04	63.86 ± 0.98	0.3	41.35 ± 2.29	47.98 ± 1.62
0.5	53.89 ± 1.65	67.62 ± 1.31	0.5	41.39 ± 1.67	52.61 ± 2.76
0.75	54.32 ± 1.18	70.13 ± 2.19	0.75	41.35 ± 1.79	52.67 ± 2.80
1.0	51.48 ± 0.94	$\mathbf{71.67 \pm 1.49}$	1.0	40.84 ± 1.32	$\mathbf{53.66 \pm 2.66}$

(b) Normalized scores over 27 Farama Minari tasks (mean ± SE over 8 seeds), averaged within each domain: **MuJoCo:** locomotion (9); **AntMaze:** sparse-reward navigation (6); **Adroit:** manipulation (12).

Benchmark	FQL	GFP	IFQL	IQL	ReBRAC	OptiDICE (Flow)	OptiDICE (Gaussian)	Opt-SUNDER (Ours)
MuJoCo (9 tasks)	103.44 ± 0.56	107.20 ± 1.00	69.02 ± 2.20	72.08 ± 3.19	100.81 ± 0.82	$\mathbf{113.31 \pm 0.42}$	89.43 ± 2.92	$\mathbf{113.53 \pm 0.39}$
AntMaze (6 tasks)	103.91 ± 2.00	107.40 ± 1.49	$82.30 \pm 1.18^{\dagger}$	40.88 ± 1.35	71.57 ± 1.43	103.04 ± 1.72	18.76 ± 1.65	$\mathbf{117.26 \pm 1.92}$
Adroit (12 tasks)	45.59 ± 0.44	46.81 ± 1.00	45.11 ± 0.48	43.45 ± 0.31	38.39 ± 0.84	44.49 ± 1.24	34.85 ± 0.54	$\mathbf{48.15 \pm 0.27}$

improving from 42.40 to 53.66 as state regularization increases while policy regularization remains weak. OptiDICE cannot accommodate these asymmetric regularization preferences, as its joint occupancy regularizer ties the control of state and policy shifts to a single coefficient.

Figure 2 shows the corresponding state visitation in the state-bad, action-good setting. The buffer is concentrated in low-reward regions, and OptiDICE largely follows this distribution. Opt-SUNDER instead shifts visitation toward high-reward regions by weakening state regularization while retaining policy regularization, as intended by state-policy decoupling.

We next ask whether this flexibility remains useful in standard offline RL benchmarks. Table 2(b) reports normalized scores on 27 Minari tasks spanning MuJoCo, AntMaze, and Adroit. We compare against OptiDICE as well as strong offline RL baselines, including FQL(Park et al., 2025), GFP(Nguimatsia Tiofack et al., 2026), and ReBRAC(Tarasov et al., 2023). To isolate the effect of decoupling from policy extraction, we also evaluate OptiDICE with the same flow-based policy. Opt-SUNDER achieves the highest average score in each domain.

Table 3: Results on state-entropy maximization and learning from observations.

(a) State entropy $\hat{H}(s)$ on maze and continuous MDP tasks. (mean $\pm$ SE over 8 seeds)

Task / Domain	Buffer	APT	PROTO	SEMDICE	Ent-SUNDER (Ours)
Maze Avg.	0.646	0.299 ± 0.083	0.288 ± 0.061	0.837 ± 0.161	$\mathbf{1.460 \pm 0.089}$
Reacher	3.502	2.525 ± 0.075	3.008 ± 0.068	5.560 ± 0.022	$\mathbf{5.620 \pm 0.024}$
Acrobot	2.178	1.197 ± 0.269	1.476 ± 0.311	2.226 ± 0.138	$\mathbf{2.915 \pm 0.288}$
Point mass	3.602	4.442 ± 0.019	4.293 ± 0.044	7.654 ± 0.059	$\mathbf{7.795 \pm 0.059}$
Pendulum	3.093	3.043 ± 0.031	2.972 ± 0.020	$\mathbf{8.946 \pm 0.039}$	$\mathbf{9.002 \pm 0.021}$

(b) Normalized returns over 15 D4RL v1.1 datasets (mean $\pm$ SE over 5 seeds), averaged within each domain

Benchmark	ReCOIL	DILO	LobsDICE	SMODICE	OTR	IOSTOM	LfO-SUNDER (Ours)
MuJoCo (9 datasets)	79.5 ± 3.69	72.6 ± 2.42	61.4 ± 2.85	62.2 ± 1.48	80.7 ± 2.14	$\mathbf{97.0 \pm 0.88}$	67.50 ± 0.95
Adroit (6 datasets)	54.73 ± 5.28	21.78 ± 6.89	29.74 ± 2.32	25.30 ± 2.23	55.83 ± 7.43	62.23 ± 2.27	$\mathbf{83.21 \pm 2.28}$

4.2 STATE-DISTRIBUTION OBJECTIVES BEYOND REWARD MAXIMIZATION

We next test whether SUNDER extends beyond reward maximization to state-distribution objectives.

State-entropy maximization. SEM aims to increase the entropy of the policy-induced state distribution. When the offline state distribution has low entropy, regularizing state visitation toward it can directly conflict with this objective. Ent-SUNDER therefore performs SEM without dataset-state regularization, while retaining policy regularization. We evaluate it against SEMDICE and two additional SEM baselines, APT(Liu & Abbeel, 2021) and PROTO(Yarats et al., 2021), on datasets with skewed state distributions. Table 3a shows that Ent-SUNDER outperforms all baselines on every maze domain and on all four continuous-control tasks. This confirms that removing dataset-state regularization is beneficial when it conflicts with the entropy objective.

Learning from observations. LfO seeks expert-like transitions from imperfect offline data. While policy regularization toward the offline data remains useful for limiting OOD actions, constraining state visitation can hinder expert matching. LfO-SUNDER therefore separates expert-transition matching from policy regularization. Table 3b shows that LfO-SUNDER achieves the highest average return on Adroit (83.21) among a set of strong LfO baselinesPham et al. (2025); Sikchi et al. (2025; 2024); Luo et al. (2023) and outperforms LobsDICE(Kim et al., 2022a) and SMODICE(Ma et al., 2022) on MuJoCo. These results support separating expert matching from policy regularization.

5 CONCLUSION

We introduced SUNDER, a general distribution-correction framework that decouples state visitation from policy deviation. Although this separation makes the state–policy formulation non-affine, we showed that its hidden convex structure preserves strong duality and allows the explicit optimization variables to be eliminated. Thus, separate control of state and policy shift does not sacrifice the tractability underlying DICE. Our experiments show when this distinction matters. When state and action quality are mismatched, Opt-SUNDER can control the two shifts independently. At the same time, the Minari results show that this flexibility does not sacrifice performance in standard offline RL settings. These results show that decoupling state and policy regularization can be beneficial, motivating further study of when it is useful.

Limitations and Future Work. SUNDER exhibits the intended decoupling behavior and achieves the best average performance on the 27 reward-maximization tasks and on LfO manipulation. However, LfO-SUNDER does not match the strongest baselines on locomotion tasks. Understanding and reducing this gap is left for future work.

AI USE STATEMENT

We used generative AI tools for manuscript editing, code development, and occasional checks of mathematical derivations written by the authors. The theoretical formulation and interpretation of experimental results were developed by the authors. All AI-assisted work was reviewed and verified by the authors. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies offline reinforcement learning in simulated environments using benchmark datasets. Our experiments do not involve human subjects or personally identifiable information. We do not identify specific ethical concerns arising from the experiments conducted in this work. However, applying reinforcement learning methods to real-world systems may introduce safety considerations that are not captured by the benchmarks studied here.

REPRODUCIBILITY STATEMENT

We provide the details needed to reproduce both the theoretical and empirical results of this work. Section A presents the SUNDER framework, and Appendix A.1 provides detailed derivations of its instantiations. Section 4 describes the experimental settings and baselines, while Appendix D reports the full experimental results and selected hyperparameters. Code for reproducing the experiments is available at `https://anonymous.4open.science/r/SUNDER_Anonymous-652B/README.md`.

REFERENCES

Youngsoo Jang, Geon-Hyeong Kim, Jongmin Lee, Sungryull Sohn, Byoungjip Kim, Honglak Lee, and Moontae Lee. SafeDICE: Offline safe imitation learning with non-preferred demonstrations. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74921–74951, 2023. doi: 10.52202/075280-3274.

Geon-Hyeong Kim, Jongmin Lee, Youngsoo Jang, Hongseok Yang, and Kee-Eung Kim. Lobsdice: Offline learning from observation via stationary distribution correction estimation. In *Advances in Neural Information Processing Systems*, 2022a.

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*, 2022b.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022.

Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. Optidice: Offline policy optimization via stationary distribution correction estimation. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. URL `https://proceedings.mlr.press/v139/lee21f.html`.

Jongmin Lee, Meiqi Sun, and Pieter Abbeel. Semdice: Off-policy state entropy maximization via stationary distribution correction estimation. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://openreview.net/forum?id=rJ5g8ueQaI`.

Hao Liu and Pieter Abbeel. Behavior from the void: Unsupervised active pre-training. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18459–18473, 2021.

Yicheng Luo, Zhengyao Jiang, Samuel Cohen, Edward Grefenstette, and Marc Peter Deisenroth. Optimal transport for offline imitation learning. In *International Conference on Learning Representations*, 2023.

Yecheng Ma, Andrew Shen, Dinesh Jayaraman, and Osbert Bastani. Versatile offline imitation from observations and examples via regularized state-occupancy matching. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14639–14663. PMLR, 2022.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2315–2325, 2019.

Franki Nguimatsia Tiofack, Théotime Le Hellard, Fabian Schramm, Nicolas Perrin-Gilbert, and Justin Carpentier. Guided flow policy: Learning from high-value actions in offline reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EBjy1rmpv0.

Seohong Park, Qiyang Li, and Sergey Levine. Flow q-learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48104–48127. PMLR, 2025.

Quang Anh Pham, Janaka Brahmanage, Tien Mai, and Akshat Kumar. IOSTOM: Offline imitation learning from observations via state transition occupancy matching. In *Advances in Neural Information Processing Systems*, volume 38, pp. 65725–65760, 2025. doi: 10.52202/085713-1975.

Harshit Sikchi, Qinqing Zheng, Amy Zhang, and Scott Niekum. Dual rl: Unification and new methods for reinforcement and imitation learning. In *International Conference on Learning Representations*, 2024.

Harshit Sikchi, Caleb Chuck, Amy Zhang, and Scott Niekum. A dual approach to imitation learning from observations with offline datasets. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 1125–1147. PMLR, 2025.

Jonathan Spencer, Sanjiban Choudhury, Matt Barnes, Matthew Schmittle, Mung Chiang, Peter Ramadge, and Siddhartha Srinivasa. Learning from interventions: Human-robot interaction as both explicit and implicit feedback. In *Proceedings of Robotics: Science and Systems*, July 2020. doi: 10.15607/RSS.2020.XVI.055.

Denis Tarasov, Vladislav Kurenkov, Alexander Nikulin, and Sergey Kolesnikov. Revisiting the minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11592–11620, 2023. doi: 10.52202/075280-0511.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=CD9Snc73AW.

Samuel Ubellacker, Aaron Ray, James M. Bern, Jared Strader, and Luca Carlone. High-speed aerial grasping using a soft drone with onboard perception. *npj Robotics*, 2(1):5, 2024. doi: 10.1038/s44182-024-00012-1.

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Reinforcement learning with prototypical representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11920–11931. PMLR, 2021.

A Detailed Derivations of SUNDER Instantiations

The main text presents SUNDER as a general optimization framework in which different methods are obtained by specifying the task objective G, the state-action term c, the reference policy, and the induced occupancy variables required by the task. This appendix provides the detailed derivations of the instantiations summarized in Table 1.

We first summarize the resulting dual objectives for reference. We then derive each objective by specializing the general SUNDER dual in Eq. 6, making explicit the method-specific conjugate calculations and the elimination of the corresponding policy and occupancy variables. Since the hidden-convexity and strong-duality arguments have already been established for the general formulation, we do not repeat them for each instantiation. Instead, we focus on how each choice of G, c, and reference policy leads to its final optimization objective.

Table 4: Population-level dual objectives of the SUNDER instantiations.

Method	Population-level dual objective
Opt-SUNDER	$\min_{V,\lambda} (1-\gamma)\mathbb{E}_{s\sim d_0}[V(s)] + \beta\mathbb{E}_{s\sim\hat{q}_{\mathcal{B}}}\left[f^*\left(\frac{e_{\hat{d}(s)}}{\beta}\right)\right]$
Ent-SUNDER	$\min_{V,\lambda} (1-\gamma)\mathbb{E}_{s\sim d_0}[V(s)] + \mathbb{E}_{s\sim u}\left[f^*\left(e_{\hat{d}}(s)\right)\right]$
SAFE-SUNDER	$\min_{V,\lambda} (1-\gamma)\mathbb{E}_{s\sim d_0}[V(s)] + \mathbb{E}_{s\sim\frac{\hat{q}_{\mathcal{B}}-\omega\hat{q}_I}{1-\omega}}\left[f^*\left(e_{\hat{d}}(s)\right)\right]$
SMO-SUNDER	$\min_{V,\lambda} (1-\gamma)\mathbb{E}_{s\sim d_0}[V(s)] + \beta\mathbb{E}_{s\sim\hat{q}_I}\left[f^*\left(\frac{e_{\hat{d}}(s)}{\beta}\right)\right]$
Demo-SUNDER	$\min_{V,\lambda} (1-\gamma)\mathbb{E}_{s\sim d_0}[V(s)] + \beta\mathbb{E}_{s\sim\hat{q}_{\mathcal{B}}}\left[f^*_{\mathrm{KL}}\left(\frac{e_{\hat{d}}(s)}{\beta}\right)\right]$
Lobs-SUNDER	$\min_{V,\lambda,Q} (1-\gamma)\mathbb{E}_{s\sim d_0}[V(s)] + \beta\mathbb{E}_{s\sim\hat{q}_I}\left[f^*_{\mathrm{KL}}\left(\frac{e_{\hat{d}}(s)}{\beta}\right)\right] + \mathbb{E}_{(s,s')\sim q'_I}\left[f^*_{\mathrm{KL}}(Q(s,s'))\right]$
LfO-SUNDER	$\min_{V,Q} \omega(1-\gamma)\mathbb{E}_{s\sim d_0}[V(s)] + \mathbb{E}_{(s,s')\sim\omega q'_E+(1-\omega)q'_I}\left[f^*\left(\gamma V(s') - V(s) + Q(s,s') + e_{\hat{d}}(s;Q)\right)\right]$ $- (1-\omega)\mathbb{E}_{(s,s')\sim q'_I}\left[\gamma V(s') - V(s) + Q(s,s') + e_{\hat{d}}(s;Q)\right]$

A.1 Summary of Final Objectives

For reference, Table 4 summarizes the population-level dual objectives obtained from the SUNDER instantiations in Table 1. We report only the resulting optimization problems here; their complete derivations and practical sample-based forms are provided in the subsequent subsections.

We use the same notation as in the main text. In particular,

$$e_\pi(s,a;V,\lambda,Q) := c(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')] - V(s) - \lambda(s) - \mathbb{E}_{s'\sim P(\cdot|s,a)}[Q(s,s')], \quad (7)$$

and

$$e_{\hat{d}}(s;V,\lambda,Q) := \lambda(s) + \alpha\mathbb{E}_{a\sim\pi_{\mathrm{ref}}(\cdot|s)}\left[f^*\left(\frac{e_\pi(s,a;V,\lambda,Q)}{\alpha}\right)\right]. \quad (8)$$

When the task does not require a transition occupancy d', we set $Q \equiv 0$ and omit it from the notation.

A.2 Opt-SUNDER

Opt-SUNDER instantiates SUNDER for offline reward maximization while independently controlling state-distribution and policy shifts from the offline data. As summarized in Table 1, we set

$$G(\hat{d}) = -\beta D_f\left(\hat{d}\,\|\,\hat{q}_{\mathcal{B}}\right), \qquad c(s,a) = r(s,a), \qquad \pi_{\mathrm{ref}} = \pi_{\mathcal{B}}. \quad (9)$$

Since the task objective depends only on the state occupancy $\hat{d}$, no transition occupancy d' is required.

Substituting these choices into the general SUNDER primal gives

$$\max_{\hat{d},\pi\geq 0} \quad \sum_{s,a}\hat{d}(s)\pi(a\mid s)r(s,a) - \alpha\sum_{s,a}\hat{d}(s)\pi_{\mathcal{B}}(a\mid s)f\left(\frac{\pi(a\mid s)}{\pi_{\mathcal{B}}(a\mid s)}\right) - \beta\sum_s\hat{q}_{\mathcal{B}}(s)f\left(\frac{\hat{d}(s)}{\hat{q}_{\mathcal{B}}(s)}\right) \quad (10)$$

$$\text{s.t.} \quad \hat{d}(s)\left(\sum_a\pi(a\mid s) - 1\right) = 0, \qquad \forall s, \quad (11)$$

$$\hat{d}(s')\sum_{a'}\pi(a'\mid s') - \gamma\sum_{s,a}\hat{d}(s)\pi(a\mid s)P(s'\mid s,a) = (1-\gamma)d_0(s'), \qquad \forall s'. \quad (12)$$

The coefficient β therefore controls state-distribution deviation from $\hat{q}_{\mathcal{B}}$, whereas α independently controls policy deviation from $\pi_{\mathcal{B}}$.

Having established strong duality for the general SUNDER formulation, we can directly apply the elimination procedure in Eq. 6. Since d' is absent, we omit Q. Using $c(s,a) = r(s,a)$, the policy residual becomes

$$e_\pi(s,a;V,\lambda) = r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')] - V(s) - \lambda(s). \tag{13}$$

Eliminating the policy through the conjugate of its regularizer gives the corresponding state residual

$$e_{\hat{d}}(s;V,\lambda) = \lambda(s) + \alpha \mathbb{E}_{a\sim\pi_{\mathcal{B}}(\cdot|s)}\left[f^*\left(\frac{e_\pi(s,a;V,\lambda)}{\alpha}\right)\right]. \tag{14}$$

The remaining method-specific step is therefore to eliminate the state occupancy under the Opt-SUNDER state objective.

For $-G(\hat{d}) = \beta \sum_s \hat{q}_{\mathcal{B}}(s) f\left(\frac{\hat{d}(s)}{\hat{q}_{\mathcal{B}}(s)}\right)$, its convex conjugate is

$$(-G)^*(e_{\hat{d}}) = \sup_{\hat{d}\geq 0}\left\{\sum_s \hat{d}(s) e_{\hat{d}}(s) - \beta \sum_s \hat{q}_{\mathcal{B}}(s) f\left(\frac{\hat{d}(s)}{\hat{q}_{\mathcal{B}}(s)}\right)\right\}. \tag{15}$$

To evaluate this term, define the state density ratio $w_d(s) := \frac{\hat{d}(s)}{\hat{q}_{\mathcal{B}}(s)}$. Then

$$(-G)^*(e_{\hat{d}}) = \beta \sum_s \hat{q}_{\mathcal{B}}(s) \sup_{w_d(s)\geq 0}\left\{w_d(s)\frac{e_{\hat{d}}(s)}{\beta} - f(w_d(s))\right\} = \beta \mathbb{E}_{s\sim\hat{q}_{\mathcal{B}}}\left[f^*\left(\frac{e_{\hat{d}}(s)}{\beta}\right)\right]. \tag{16}$$

Thus, substituting Eq. 16 into the general SUNDER dual yields the Opt-SUNDER objective

$$\min_{V,\lambda} \quad (1-\gamma)\mathbb{E}_{s\sim d_0}[V(s)] + \beta \mathbb{E}_{s\sim\hat{q}_{\mathcal{B}}}\left[f^*\left(\frac{e_{\hat{d}}(s;V,\lambda)}{\beta}\right)\right] \tag{17}$$

with e_π and $e_{\hat{d}}$ defined in Eqs. 13–14.

When f^* is differentiable, the eliminated policy and state density ratios can be recovered as

$$w^*_\pi(s,a) = (f^*)'\left(\frac{e_\pi(s,a;V,\lambda)}{\alpha}\right), \qquad w^*_d(s) = (f^*)'\left(\frac{e_{\hat{d}}(s;V,\lambda)}{\beta}\right). \tag{18}$$

Hence, α and β control the policy and state-distribution corrections through separate density ratios. The resulting dual depends only on expectations under the initial-state and offline-data distributions and can therefore be optimized from offline transitions without explicitly optimizing over π or $\hat{d}$.

A.3 ENT-SUNDER

Ent-SUNDER instantiates SUNDER for state-entropy maximization while independently controlling policy deviation from the offline data. As summarized in Table 1, we set

$$G(\hat{d}) = -D_f\left(\hat{d} \,\|\, u\right), \qquad c(s,a) = 0, \qquad \pi_{\text{ref}} = \pi_{\mathcal{B}}. \tag{19}$$

Here, u denotes the uniform state distribution. For the KL divergence, maximizing $G(\hat{d})$ is equivalent to maximizing the entropy of $\hat{d}$ up to a constant. Since the task objective depends only on the state occupancy $\hat{d}$, no transition occupancy d' is required.

Substituting these choices into the general SUNDER primal gives

$$\max_{\hat{d},\pi\geq 0} \quad -\sum_s u(s) f\left(\frac{\hat{d}(s)}{u(s)}\right) - \alpha \sum_{s,a} \hat{d}(s)\pi_{\mathcal{B}}(a \mid s) f\left(\frac{\pi(a \mid s)}{\pi_{\mathcal{B}}(a \mid s)}\right) \tag{20}$$

$$\text{s.t.} \quad \hat{d}(s)\left(\sum_a \pi(a \mid s) - 1\right) = 0, \qquad \forall s, \tag{21}$$

$$\hat{d}(s')\sum_{a'} \pi(a' \mid s') - \gamma \sum_{s,a} \hat{d}(s)\pi(a \mid s) P(s' \mid s,a) = (1-\gamma) d_0(s'), \qquad \forall s'. \tag{22}$$

Unlike Opt-SUNDER, Ent-SUNDER does not regularize the state distribution toward the offline state distribution. Instead, the state objective directly encourages high-entropy visitation, while α controls policy deviation from $\pi_{\mathcal{B}}$.

Having established strong duality for the general SUNDER formulation, we can directly apply the elimination procedure in Eq. 6. Since d' is absent, we omit Q. Using $c(s,a)=0$, the policy residual becomes

$$e_\pi(s,a;V,\lambda) = \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')] - V(s) - \lambda(s). \tag{23}$$

Eliminating the policy through the conjugate of its regularizer gives the corresponding state residual

$$e_{\hat{d}}(s;V,\lambda) = \lambda(s) + \alpha \mathbb{E}_{a\sim\pi_{\mathcal{B}}(\cdot|s)}\left[f^*\left(\frac{e_\pi(s,a;V,\lambda)}{\alpha}\right)\right]. \tag{24}$$

The remaining method-specific step is therefore to eliminate the state occupancy under the Ent-SUNDER state objective.

For $-G(\hat{d}) = \sum_s u(s) f\left(\frac{\hat{d}(s)}{u(s)}\right)$, its convex conjugate is

$$(-G)^*(e_{\hat{d}}) = \sup_{\hat{d}\geq 0}\left\{\sum_s \hat{d}(s)e_{\hat{d}}(s) - \sum_s u(s) f\left(\frac{\hat{d}(s)}{u(s)}\right)\right\}. \tag{25}$$

To evaluate this term, define the state density ratio $w_d(s) := \frac{\hat{d}(s)}{u(s)}$. Then

$$(-G)^*(e_{\hat{d}}) = \sum_s u(s) \sup_{w_d(s)\geq 0}\left\{w_d(s)e_{\hat{d}}(s) - f(w_d(s))\right\} = \mathbb{E}_{s\sim u}\left[f^*\big(e_{\hat{d}}(s)\big)\right]. \tag{26}$$

Thus, substituting Eq. 26 into the general SUNDER dual yields the Ent-SUNDER objective

$$\min_{V,\lambda} \quad (1-\gamma)\mathbb{E}_{s\sim d_0}[V(s)] + \mathbb{E}_{s\sim u}\left[f^*\big(e_{\hat{d}}(s;V,\lambda)\big)\right] \tag{27}$$

with e_π and $e_{\hat{d}}$ defined in Eqs. 23–24.

When f^* is differentiable, the eliminated policy and state density ratios can be recovered as

$$w^*_\pi(s,a) = (f^*)'\left(\frac{e_\pi(s,a;V,\lambda)}{\alpha}\right), \qquad w^*_d(s) = (f^*)'\big(e_{\hat{d}}(s;V,\lambda)\big). \tag{28}$$

Hence, Ent-SUNDER optimizes the state-distribution objective toward the uniform distribution while separately controlling action selection through the policy density ratio.

Since offline samples are drawn from $\hat{q}_{\mathcal{B}}$ rather than u, we finally rewrite the state expectation under the offline state distribution. For a uniform distribution $u(s)=c$,

$$\mathbb{E}_{s\sim u}\left[f^*\big(e_{\hat{d}}(s;V,\lambda)\big)\right] = c\,\mathbb{E}_{s\sim\hat{q}_{\mathcal{B}}}\left[\frac{1}{\hat{q}_{\mathcal{B}}(s)} f^*\big(e_{\hat{d}}(s;V,\lambda)\big)\right]. \tag{29}$$

Therefore, the Ent-SUNDER dual can be evaluated from offline states as

$$\min_{V,\lambda} \quad (1-\gamma)\mathbb{E}_{s\sim d_0}[V(s)] + c\,\mathbb{E}_{s\sim\hat{q}_{\mathcal{B}}}\left[\frac{1}{\hat{q}_{\mathcal{B}}(s)} f^*\big(e_{\hat{d}}(s;V,\lambda)\big)\right]. \tag{30}$$

A.4 SAFE-SUNDER

SAFE-SUNDER instantiates SUNDER for recovering a preferred state distribution from an offline dataset containing an imperfect component. Assuming that an ω fraction of the offline state distribution is associated with the imperfect-data distribution $\hat{q}_I$, the corrected preferred-state reference is

$$\frac{\hat{q}_{\mathcal{B}} - \omega\hat{q}_I}{1-\omega}, \qquad \omega\in[0,1).$$

As summarized in Table 1, we therefore set

$$G(\hat{d}) = -D_f\left(\hat{d}\,\middle\|\,\frac{\hat{q}_{\mathcal{B}} - \omega\hat{q}_I}{1-\omega}\right), \qquad c(s,a)=0, \qquad \pi_{\text{ref}} = \pi_{\mathcal{B}}. \tag{31}$$

The task objective therefore matches the induced state occupancy to the corrected preferred-state distribution, while policy regularization is applied separately toward the behavior policy of the available offline data. Since the task objective depends only on the state occupancy $\hat{d}$, no transition occupancy d' is required.

Substituting these choices into the general SUNDER primal gives

$$\max_{\hat{d},\pi\geq 0} \quad -\sum_s \frac{\hat{q}_{\mathcal{B}}(s)-\omega\hat{q}_I(s)}{1-\omega} f\left(\frac{(1-\omega)\hat{d}(s)}{\hat{q}_{\mathcal{B}}(s)-\omega\hat{q}_I(s)}\right) - \alpha\sum_{s,a}\hat{d}(s)\pi_{\mathcal{B}}(a\mid s) f\left(\frac{\pi(a\mid s)}{\pi_{\mathcal{B}}(a\mid s)}\right) \tag{32}$$

$$\text{s.t.} \quad \hat{d}(s)\left(\sum_a \pi(a\mid s)-1\right)=0, \qquad \forall s, \tag{33}$$

$$\hat{d}(s')\sum_{a'}\pi(a'\mid s') - \gamma\sum_{s,a}\hat{d}(s)\pi(a\mid s)P(s'\mid s,a) = (1-\gamma)d_0(s'), \qquad \forall s'. \tag{34}$$

Thus, the state objective determines the preferred visitation distribution, whereas α independently controls policy deviation from $\pi_{\mathcal{B}}$.

Having established strong duality for the general SUNDER formulation, we can directly apply the elimination procedure in Eq. 6. Since d' is absent, we omit Q. Using $c(s,a)=0$, the policy residual becomes

$$e_\pi(s,a;V,\lambda) = \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')] - V(s) - \lambda(s). \tag{35}$$

Eliminating the policy through the conjugate of its regularizer gives the corresponding state residual

$$e_{\hat{d}}(s;V,\lambda) = \lambda(s) + \alpha\mathbb{E}_{a\sim\pi_{\mathcal{B}}(\cdot|s)}\left[f^*\left(\frac{e_\pi(s,a;V,\lambda)}{\alpha}\right)\right]. \tag{36}$$

The remaining method-specific step is therefore to eliminate the state occupancy under the SAFE-SUNDER state objective.

For $-G(\hat{d}) = \sum_s \frac{\hat{q}_{\mathcal{B}}(s)-\omega\hat{q}_I(s)}{1-\omega} f\left(\frac{(1-\omega)\hat{d}(s)}{\hat{q}_{\mathcal{B}}(s)-\omega\hat{q}_I(s)}\right)$, its convex conjugate is

$$(-G)^*(e_{\hat{d}}) = \sup_{\hat{d}\geq 0}\left\{\sum_s \hat{d}(s)e_{\hat{d}}(s) - \sum_s \frac{\hat{q}_{\mathcal{B}}(s)-\omega\hat{q}_I(s)}{1-\omega} f\left(\frac{(1-\omega)\hat{d}(s)}{\hat{q}_{\mathcal{B}}(s)-\omega\hat{q}_I(s)}\right)\right\}. \tag{37}$$

To evaluate this term, define the state density ratio $w_d(s) := \frac{(1-\omega)\hat{d}(s)}{\hat{q}_{\mathcal{B}}(s)-\omega\hat{q}_I(s)}$. Then

$$(-G)^*(e_{\hat{d}}) = \sum_s \frac{\hat{q}_{\mathcal{B}}(s)-\omega\hat{q}_I(s)}{1-\omega}\sup_{w_d(s)\geq 0}\left\{w_d(s)e_{\hat{d}}(s) - f(w_d(s))\right\} = \mathbb{E}_{s\sim\frac{\hat{q}_{\mathcal{B}}-\omega\hat{q}_I}{1-\omega}}\left[f^*\left(e_{\hat{d}}(s)\right)\right]. \tag{38}$$

Thus, substituting Eq. 38 into the general SUNDER dual yields the SAFE-SUNDER objective

$$\min_{V,\lambda} \quad (1-\gamma)\mathbb{E}_{s\sim d_0}[V(s)] + \mathbb{E}_{s\sim\frac{\hat{q}_{\mathcal{B}}-\omega\hat{q}_I}{1-\omega}}\left[f^*\left(e_{\hat{d}}(s;V,\lambda)\right)\right]. \tag{39}$$

with e_π and $e_{\hat{d}}$ defined in Eqs. 35–36.

When f^* is differentiable, the eliminated policy and state density ratios can be recovered as

$$w^*_\pi(s,a) = (f^*)'\left(\frac{e_\pi(s,a;V,\lambda)}{\alpha}\right), \qquad w^*_d(s) = (f^*)'\left(e_{\hat{d}}(s;V,\lambda)\right). \tag{40}$$

Hence, SAFE-SUNDER uses the corrected preferred-state distribution to control state visitation while separately controlling action selection through the policy density ratio.

To evaluate the state term directly from the available distributions, we use

$$\mathbb{E}_{s\sim\frac{\hat{q}_{\mathcal{B}}-\omega\hat{q}_I}{1-\omega}}\left[f^*\left(e_{\hat{d}}(s;V,\lambda)\right)\right] = \frac{1}{1-\omega}\left(\mathbb{E}_{s\sim\hat{q}_{\mathcal{B}}}\left[f^*\left(e_{\hat{d}}(s;V,\lambda)\right)\right] - \omega\mathbb{E}_{s\sim\hat{q}_I}\left[f^*\left(e_{\hat{d}}(s;V,\lambda)\right)\right]\right). \tag{41}$$

Therefore, the SAFE-SUNDER dual can be written entirely in terms of the available offline and imperfect-data state distributions as

$$\min_{V,\lambda} \quad (1-\gamma)\mathbb{E}_{s\sim d_0}[V(s)] + \frac{1}{1-\omega}\left(\mathbb{E}_{s\sim\hat{q}_{\mathcal{B}}}\left[f^*\left(e_{\hat{d}}(s;V,\lambda)\right)\right] - \omega\mathbb{E}_{s\sim\hat{q}_I}\left[f^*\left(e_{\hat{d}}(s;V,\lambda)\right)\right]\right). \tag{42}$$

A.5 SMO-SUNDER

SMO-SUNDER instantiates SUNDER for state-distribution matching in learning from observations. Let $\hat{q}_E$ and $\hat{q}_I$ denote the expert and imperfect-data state distributions, respectively. Following SMODICE, we define the state-based reward

$$R(s) := \log \frac{\hat{q}_E(s)}{\hat{q}_I(s)}.$$

The task objective encourages expert-like state visitation through this density-ratio reward, while state-distribution and policy deviations from the imperfect data are controlled independently. As summarized in Table 1, we set

$$G(\hat{d}) = -\beta D_f\left(\hat{d} \,\|\, \hat{q}_I\right), \qquad c(s,a) = R(s), \qquad \pi_{\text{ref}} = \pi_I. \tag{43}$$

Since the task objective depends only on the state occupancy $\hat{d}$, no transition occupancy d' is required. Substituting these choices into the general SUNDER primal gives

$$\max_{\hat{d},\pi\geq 0} \quad \sum_{s,a} \hat{d}(s)\pi(a \mid s)R(s) - \alpha \sum_{s,a} \hat{d}(s)\pi_I(a \mid s) f\left(\frac{\pi(a \mid s)}{\pi_I(a \mid s)}\right) - \beta \sum_s \hat{q}_I(s) f\left(\frac{\hat{d}(s)}{\hat{q}_I(s)}\right) \tag{44}$$

$$\text{s.t.} \quad \hat{d}(s)\left(\sum_a \pi(a \mid s) - 1\right) = 0, \qquad \forall s, \tag{45}$$

$$\hat{d}(s') \sum_{a'} \pi(a' \mid s') - \gamma \sum_{s,a} \hat{d}(s)\pi(a \mid s)P(s' \mid s,a) = (1-\gamma)d_0(s'), \qquad \forall s'. \tag{46}$$

Under the policy-normalization constraint, the first term is $\mathbb{E}_{s\sim\hat{d}}[R(s)]$. Thus, β controls state-distribution deviation from $\hat{q}_I$, whereas α independently controls policy deviation from π_I.

Having established strong duality for the general SUNDER formulation, we can directly apply the elimination procedure in Eq. 6. Since d' is absent, we omit Q. Using $c(s,a) = R(s)$, the policy residual becomes

$$e_\pi(s,a;V,\lambda) = R(s) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')] - V(s) - \lambda(s). \tag{47}$$

Eliminating the policy through the conjugate of its regularizer gives the corresponding state residual

$$e_{\hat{d}}(s;V,\lambda) = \lambda(s) + \alpha \mathbb{E}_{a\sim\pi_I(\cdot|s)}\left[f^*\left(\frac{e_\pi(s,a;V,\lambda)}{\alpha}\right)\right]. \tag{48}$$

The remaining method-specific step is therefore to eliminate the state occupancy under the SMO-SUNDER state objective.

For $-G(\hat{d}) = \beta \sum_s \hat{q}_I(s) f\left(\frac{\hat{d}(s)}{\hat{q}_I(s)}\right)$, its convex conjugate is

$$(-G)^*(e_{\hat{d}}) = \sup_{\hat{d}\geq 0}\left\{\sum_s \hat{d}(s)e_{\hat{d}}(s) - \beta \sum_s \hat{q}_I(s) f\left(\frac{\hat{d}(s)}{\hat{q}_I(s)}\right)\right\}. \tag{49}$$

To evaluate this term, define the state density ratio $w_d(s) := \frac{\hat{d}(s)}{\hat{q}_I(s)}$. Then

$$(-G)^*(e_{\hat{d}}) = \beta \sum_s \hat{q}_I(s) \sup_{w_d(s)\geq 0}\left\{w_d(s)\frac{e_{\hat{d}}(s)}{\beta} - f(w_d(s))\right\} = \beta \mathbb{E}_{s\sim\hat{q}_I}\left[f^*\left(\frac{e_{\hat{d}}(s)}{\beta}\right)\right]. \tag{50}$$

Thus, substituting Eq. 50 into the general SUNDER dual yields the SMO-SUNDER objective

$$\min_{V,\lambda} \quad (1-\gamma)\mathbb{E}_{s\sim d_0}[V(s)] + \beta \mathbb{E}_{s\sim\hat{q}_I}\left[f^*\left(\frac{e_{\hat{d}}(s;V,\lambda)}{\beta}\right)\right]. \tag{51}$$

with e_π and $e_{\hat{d}}$ defined in Eqs. 47–48.

When f^* is differentiable, the eliminated policy and state density ratios can be recovered as

$$w^*_\pi(s,a) = (f^*)'\left(\frac{e_\pi(s,a;V,\lambda)}{\alpha}\right), \qquad w^*_d(s) = (f^*)'\left(\frac{e_{\hat{d}}(s;V,\lambda)}{\beta}\right). \tag{52}$$

Hence, SMO-SUNDER preserves the state-based expert-matching objective of SMODICE while allowing the state and policy corrections to be controlled independently through β and α, respectively.

For the KL divergence, setting $\alpha = \beta = 1$ recovers the coupled SMODICE objective. In particular,

$$\mathbb{E}_{s\sim\hat{d}}\left[\log\frac{\hat{q}_E(s)}{\hat{q}_I(s)}\right] - D_{\mathrm{KL}}\left(\hat{d}\,\|\,\hat{q}_I\right) = -D_{\mathrm{KL}}\left(\hat{d}\,\|\,\hat{q}_E\right), \tag{53}$$

so that the factorized objective becomes

$$\max_{\hat{d},\pi\geq 0} \quad -D_{\mathrm{KL}}\left(\hat{d}\,\|\,\hat{q}_E\right) - \mathbb{E}_{s\sim\hat{d}}\left[D_{\mathrm{KL}}(\pi(\cdot\mid s)\,\|\,\pi_I(\cdot\mid s))\right], \tag{54}$$

subject to Eqs. 45–46. Thus, SMODICE is recovered when the state and policy components are coupled with equal unit weight, whereas SMO-SUNDER allows these two components to be adjusted separately.

A.6 DEMO-SUNDER

Demo-SUNDER instantiates SUNDER for offline imitation learning from expert and supplementary imperfect demonstrations. Let q_E denote the expert state-action distribution and let $q_{\mathcal{B}}$ denote the state-action distribution of the available offline demonstrations, including both expert and imperfect data. Following DemoDICE, we define the state-action density-ratio reward

$$R(s,a) := \log\frac{q_E(s,a)}{q_{\mathcal{B}}(s,a)}.$$

The task objective encourages expert-like state-action visitation through this density-ratio reward, while state-distribution and policy deviations from the offline data are controlled independently. As summarized in Table 1, we set

$$G(\hat{d}) = -\beta D_{\mathrm{KL}}\left(\hat{d}\,\|\,\hat{q}_{\mathcal{B}}\right), \qquad c(s,a) = R(s,a), \qquad \pi_{\mathrm{ref}} = \pi_{\mathcal{B}}. \tag{55}$$

Since the task objective depends only on the state occupancy $\hat{d}$ and the state-action reward $R(s,a)$, no transition occupancy d' is required.

Substituting these choices into the general SUNDER primal gives

$$\max_{\hat{d},\pi\geq 0} \quad \sum_{s,a}\hat{d}(s)\pi(a\mid s)R(s,a) - \alpha\sum_{s,a}\hat{d}(s)\pi_{\mathcal{B}}(a\mid s)f_{\mathrm{KL}}\left(\frac{\pi(a\mid s)}{\pi_{\mathcal{B}}(a\mid s)}\right) - \beta\sum_s \hat{q}_{\mathcal{B}}(s)f_{\mathrm{KL}}\left(\frac{\hat{d}(s)}{\hat{q}_{\mathcal{B}}(s)}\right) \tag{56}$$

$$\text{s.t.} \quad \hat{d}(s)\left(\sum_a \pi(a\mid s) - 1\right) = 0, \qquad \forall s, \tag{57}$$

$$\hat{d}(s')\sum_{a'}\pi(a'\mid s') - \gamma\sum_{s,a}\hat{d}(s)\pi(a\mid s)P(s'\mid s,a) = (1-\gamma)d_0(s'), \qquad \forall s'. \tag{58}$$

Thus, β controls state-distribution deviation from $\hat{q}_{\mathcal{B}}$, whereas α independently controls policy deviation from $\pi_{\mathcal{B}}$.

Having established strong duality for the general SUNDER formulation, we can directly apply the elimination procedure in Eq. 6. Since d' is absent, we omit Q. Using $c(s,a) = R(s,a)$, the policy residual becomes

$$e_\pi(s,a;V,\lambda) = R(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot\mid s,a)}[V(s')] - V(s) - \lambda(s). \tag{59}$$

Eliminating the policy through the conjugate of its regularizer gives the corresponding state residual

$$e_{\hat{d}}(s;V,\lambda) = \lambda(s) + \alpha\mathbb{E}_{a\sim\pi_{\mathcal{B}}(\cdot\mid s)}\left[f^*_{\mathrm{KL}}\left(\frac{e_\pi(s,a;V,\lambda)}{\alpha}\right)\right]. \tag{60}$$

The remaining method-specific step is therefore to eliminate the state occupancy under the Demo-SUNDER state objective.

For $-G(\hat{d}) = \beta \sum_s \hat{q}_{\mathcal{B}}(s) f_{\mathrm{KL}}\left(\frac{\hat{d}(s)}{\hat{q}_{\mathcal{B}}(s)}\right)$, its convex conjugate is

$$(-G)^*(e_{\hat{d}}) = \sup_{\hat{d} \geq 0} \left\{ \sum_s \hat{d}(s) e_{\hat{d}}(s) - \beta \sum_s \hat{q}_{\mathcal{B}}(s) f_{\mathrm{KL}}\left(\frac{\hat{d}(s)}{\hat{q}_{\mathcal{B}}(s)}\right) \right\}. \tag{61}$$

To evaluate this term, define the state density ratio $w_d(s) := \frac{\hat{d}(s)}{\hat{q}_{\mathcal{B}}(s)}$. Then

$$(-G)^*(e_{\hat{d}}) = \beta \sum_s \hat{q}_{\mathcal{B}}(s) \sup_{w_d(s) \geq 0} \left\{ w_d(s) \frac{e_{\hat{d}}(s)}{\beta} - f_{\mathrm{KL}}(w_d(s)) \right\} = \beta \mathbb{E}_{s \sim \hat{q}_{\mathcal{B}}} \left[f^*_{\mathrm{KL}}\left(\frac{e_{\hat{d}}(s)}{\beta}\right) \right]. \tag{62}$$

Thus, substituting Eq. 62 into the general SUNDER dual yields the Demo-SUNDER objective

$$\min_{V,\lambda} \quad (1-\gamma)\mathbb{E}_{s \sim d_0}[V(s)] + \beta \mathbb{E}_{s \sim \hat{q}_{\mathcal{B}}} \left[f^*_{\mathrm{KL}}\left(\frac{e_{\hat{d}}(s; V, \lambda)}{\beta}\right) \right]. \tag{63}$$

with e_π and $e_{\hat{d}}$ defined in Eqs. 59–60.

When f^*_{KL} is differentiable, the eliminated policy and state density ratios can be recovered as

$$w^*_\pi(s,a) = (f^*_{\mathrm{KL}})'\left(\frac{e_\pi(s,a;V,\lambda)}{\alpha}\right), \qquad w^*_d(s) = (f^*_{\mathrm{KL}})'\left(\frac{e_{\hat{d}}(s;V,\lambda)}{\beta}\right). \tag{64}$$

Hence, Demo-SUNDER preserves the expert-matching objective of DemoDICE while allowing state-distribution and policy deviations from the offline data to be controlled independently through β and α, respectively.

To recover DemoDICE as a coupled special case, recall that its population objective can be written as $\max_{d \in \mathcal{C}_{\mathrm{flow}}} -D_{\mathrm{KL}}(d \,\|\, q_E) - \eta D_{\mathrm{KL}}(d \,\|\, q_{\mathcal{B}})$. Using $R(s,a) = \log(q_E(s,a)/q_{\mathcal{B}}(s,a))$, this objective is equivalently

$$\max_{d \in \mathcal{C}_{\mathrm{flow}}} \quad \mathbb{E}_{(s,a) \sim d}[R(s,a)] - (1+\eta) D_{\mathrm{KL}}(d \,\|\, q_{\mathcal{B}}). \tag{65}$$

Factorizing $d(s,a) = \hat{d}(s)\pi(a \mid s)$ and $q_{\mathcal{B}}(s,a) = \hat{q}_{\mathcal{B}}(s)\pi_{\mathcal{B}}(a \mid s)$ gives

$$D_{\mathrm{KL}}(d \,\|\, q_{\mathcal{B}}) = D_{\mathrm{KL}}\left(\hat{d} \,\|\, \hat{q}_{\mathcal{B}}\right) + \mathbb{E}_{s \sim \hat{d}}\left[D_{\mathrm{KL}}(\pi(\cdot \mid s) \,\|\, \pi_{\mathcal{B}}(\cdot \mid s))\right]. \tag{66}$$

Therefore, setting $\alpha = \beta = 1 + \eta$ in Demo-SUNDER recovers the coupled DemoDICE objective. Demo-SUNDER relaxes this coupling by allowing the state and policy components to use independent regularization strengths.

A.7 LOBS-SUNDER

Lobs-SUNDER instantiates SUNDER for learning from observations by matching the state-transition occupancy induced by the policy to the expert state-transition distribution. Let q'_E and q'_I denote the expert and imperfect-data state-transition distributions, respectively. Following the transition-matching objective of LobsDICE, we define the transition-based reward

$$R(s,s') := \log \frac{q'_E(s,s')}{q'_I(s,s')}.$$

Then

$$-D_{\mathrm{KL}}(d' \,\|\, q'_E) = \mathbb{E}_{(s,s') \sim d'}\left[R(s,s')\right] - D_{\mathrm{KL}}(d' \,\|\, q'_I). \tag{67}$$

Thus, expert transition matching can be expressed as a transition-based intrinsic reward together with a transition-distribution correction toward the imperfect data.

Lobs-SUNDER retains this transition objective while independently controlling state-distribution and policy deviations from the imperfect data. As summarized in Table 1, we set

$$G(\hat{d}, d') = \mathbb{E}_{(s,s')\sim d'}\left[R(s,s')\right] - D_{\mathrm{KL}}(d' \,\|\, q'_I) - \beta D_{\mathrm{KL}}\Big(\hat{d} \,\|\, \hat{q}_I\Big), \quad c(s,a) = 0, \quad \pi_{\mathrm{ref}} = \pi_I. \tag{68}$$

Unlike the preceding state-level instantiations, the task objective depends explicitly on the induced state-transition occupancy d'. Therefore, both $\hat{d}$ and d' are retained in the SUNDER formulation.

Substituting these choices into the general SUNDER primal gives

$$\begin{aligned}
\max_{\hat{d},d',\pi\geq 0} \quad & \sum_{s,s'} d'(s,s')R(s,s') - \sum_{s,s'} q'_I(s,s') f_{\mathrm{KL}}\left(\frac{d'(s,s')}{q'_I(s,s')}\right) - \beta \sum_s \hat{q}_I(s) f_{\mathrm{KL}}\left(\frac{\hat{d}(s)}{\hat{q}_I(s)}\right) \\
& - \alpha \sum_{s,a} \hat{d}(s)\pi_I(a \mid s) f_{\mathrm{KL}}\left(\frac{\pi(a \mid s)}{\pi_I(a \mid s)}\right) && (69) \\
\text{s.t.} \quad & \hat{d}(s)\left(\sum_a \pi(a \mid s) - 1\right) = 0, \qquad \forall s, && (70) \\
& d'(s,s') = \hat{d}(s)\sum_a \pi(a \mid s) P(s' \mid s,a), \qquad \forall s,s', && (71) \\
& \hat{d}(s')\sum_{a'} \pi(a' \mid s') - \gamma \sum_{s,a} \hat{d}(s)\pi(a \mid s) P(s' \mid s,a) = (1-\gamma) d_0(s'), \qquad \forall s'. && (72)
\end{aligned}$$

The transition objective preserves expert transition matching, whereas β and α independently control state-distribution and policy deviations from the imperfect data.

Having established strong duality for the general SUNDER formulation, we can directly apply the elimination procedure in Eq. 6. Since the task objective depends on d', we retain the transition-consistency multiplier Q. Using $c(s,a) = 0$, the policy residual becomes

$$e_\pi(s,a;V,\lambda,Q) = \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')] - V(s) - \lambda(s) - \mathbb{E}_{s'\sim P(\cdot|s,a)}[Q(s,s')]. \tag{73}$$

Eliminating the policy through the conjugate of its regularizer gives the corresponding state residual

$$e_{\hat{d}}(s;V,\lambda,Q) = \lambda(s) + \alpha \mathbb{E}_{a\sim\pi_I(\cdot|s)}\left[f^*_{\mathrm{KL}}\left(\frac{e_\pi(s,a;V,\lambda,Q)}{\alpha}\right)\right]. \tag{74}$$

The remaining method-specific step is therefore to jointly eliminate the state and transition occupancies under the Lobs-SUNDER objective.

For $-G(\hat{d}, d') = D_{\mathrm{KL}}(d' \,\|\, q'_I) + \beta D_{\mathrm{KL}}\Big(\hat{d} \,\|\, \hat{q}_I\Big) - \mathbb{E}_{(s,s')\sim d'}[R(s,s')]$, its convex conjugate is

$$\begin{aligned}
(-G)^*(e_{\hat{d}}, Q) = \sup_{\hat{d},d'\geq 0} \Bigg\{ & \sum_s \hat{d}(s) e_{\hat{d}}(s) + \sum_{s,s'} d'(s,s')\left(Q(s,s') + R(s,s')\right) \\
& -\beta \sum_s \hat{q}_I(s) f_{\mathrm{KL}}\left(\frac{\hat{d}(s)}{\hat{q}_I(s)}\right) - \sum_{s,s'} q'_I(s,s') f_{\mathrm{KL}}\left(\frac{d'(s,s')}{q'_I(s,s')}\right) \Bigg\}.
\end{aligned} \tag{75}$$

Since the state and transition terms are separable, the maximization can be evaluated independently. Define the density ratios $w_d(s) := \frac{\hat{d}(s)}{\hat{q}_I(s)}$ and $w_{d'}(s,s') := \frac{d'(s,s')}{q'_I(s,s')}$. Then

$$\begin{aligned}
(-G)^*(e_{\hat{d}}, Q) &= \beta \sum_s \hat{q}_I(s) \sup_{w_d(s)\geq 0} \left\{ w_d(s)\frac{e_{\hat{d}}(s)}{\beta} - f_{\mathrm{KL}}(w_d(s)) \right\} \\
&\quad + \sum_{s,s'} q'_I(s,s') \sup_{w_{d'}(s,s')\geq 0} \left\{ w_{d'}(s,s')\left(Q(s,s') + R(s,s')\right) - f_{\mathrm{KL}}(w_{d'}(s,s')) \right\} \\
&= \beta \mathbb{E}_{s\sim\hat{q}_I}\left[f^*_{\mathrm{KL}}\left(\frac{e_{\hat{d}}(s)}{\beta}\right)\right] + \mathbb{E}_{(s,s')\sim q'_I}\left[f^*_{\mathrm{KL}}(Q(s,s') + R(s,s'))\right].
\end{aligned} \tag{76}$$

Thus, substituting Eq. 76 into the general SUNDER dual yields the Lobs-SUNDER objective

$$\min_{V,\lambda,Q} \quad (1-\gamma)\mathbb{E}_{s\sim d_0}[V(s)] + \beta\mathbb{E}_{s\sim\hat{q}_I}\left[f^*_{\mathrm{KL}}\left(\frac{e_{\hat{d}}(s;V,\lambda,Q)}{\beta}\right)\right] + \mathbb{E}_{(s,s')\sim q'_I}\left[f^*_{\mathrm{KL}}(Q(s,s')+R(s,s'))\right]. \tag{77}$$

with e_π and $e_{\hat{d}}$ defined in Eqs. 73–74.

When f^*_{KL} is differentiable, the eliminated policy, state, and transition density ratios can be recovered as

$$w^*_\pi(s,a) = (f^*_{\mathrm{KL}})'\left(\frac{e_\pi(s,a;V,\lambda,Q)}{\alpha}\right), \qquad w^*_d(s) = (f^*_{\mathrm{KL}})'\left(\frac{e_{\hat{d}}(s;V,\lambda,Q)}{\beta}\right),$$
$$w^*_{d'}(s,s') = (f^*_{\mathrm{KL}})'(Q(s,s')+R(s,s')). \tag{78}$$

Hence, Lobs-SUNDER preserves the transition-level expert-matching objective of LobsDICE while independently controlling state-distribution and policy deviations from the imperfect data through β and α, respectively.

To recover LobsDICE as a coupled special case, recall that Eq. 67 gives

$$\mathbb{E}_{(s,s')\sim d'}\left[R(s,s')\right] - D_{\mathrm{KL}}(d' \,\|\, q'_I) = -D_{\mathrm{KL}}(d' \,\|\, q'_E). \tag{79}$$

Moreover, the KL chain rule gives

$$D_{\mathrm{KL}}(d \,\|\, q_I) = D_{\mathrm{KL}}\left(\hat{d} \,\|\, \hat{q}_I\right) + \mathbb{E}_{s\sim\hat{d}}\left[D_{\mathrm{KL}}(\pi(\cdot \mid s) \,\|\, \pi_I(\cdot \mid s))\right]. \tag{80}$$

Therefore, setting $\alpha = \beta = \eta$ reduces the Lobs-SUNDER primal to

$$\max_{d,d'\geq 0} \quad -D_{\mathrm{KL}}(d' \,\|\, q'_E) - \eta D_{\mathrm{KL}}(d \,\|\, q_I), \tag{81}$$

subject to the corresponding Bellman-flow and transition-consistency constraints. Thus, LobsDICE is recovered when the state and policy components of the imperfect-data regularizer share the same coefficient, whereas Lobs-SUNDER allows them to be adjusted independently.

A.8 LfO-SUNDER

Unlike the preceding instantiations, LfO-SUNDER is not obtained by directly decoupling a single existing DICE formulation. Instead, we develop a new transition-level SUNDER formulation for learning from observations.

Unlike Lobs-SUNDER, which we expressed in the general $(\hat{d}, d', \pi)$ parameterization to make its connection to LobsDICE explicit, here we use d' as the primary occupancy variable. A state-transition occupancy already contains its corresponding state occupancy as the row marginal

$$\hat{d}(s) = \sum_{s'} d'(s,s').$$

Thus, $\hat{d}$ need not be introduced as a separate optimization variable.

Let q'_E and q'_I denote the expert and imperfect-data state-transition distributions, respectively. We define the transition-level objective

$$G(d') = -D_f(\omega d' + (1-\omega)q'_I \,\|\, \omega q'_E + (1-\omega)q'_I), \qquad \omega \in (0,1), \tag{82}$$

and independently regularize the policy toward the imperfect-data policy π_I with strength α.

Using $\hat{d}(s) = \sum_{s'} d'(s,s')$, the resulting LfO-SUNDER primal is

$$\max_{d',\pi\geq 0} \quad -D_f(\omega d' + (1-\omega)q'_I \,\|\, \omega q'_E + (1-\omega)q'_I) - \alpha \sum_{s,a}\left(\sum_{s'} d'(s,s')\right)\pi_I(a \mid s) f\left(\frac{\pi}{\pi_I}\right)$$
$$\text{s.t.} \quad \sum_{s'} d'(s,s') = (1-\gamma)d_0(s) + \gamma\sum_{\bar{s}} d'(\bar{s},s), \qquad \forall s, \tag{83}$$
$$d'(s,s') = \left(\sum_{\bar{s}} d'(s,\bar{s})\right)\sum_a \pi(a \mid s)P(s' \mid s,a), \qquad \forall s,s'. \tag{84}$$

The first term matches the policy-induced transition occupancy to the expert transition distribution through a mixture with the imperfect transition distribution. The shared q'_I component anchors the comparison to the imperfect-data distribution, while ω controls the degree of transition-level expert matching. The second term independently regularizes action selection toward π_I through α. Thus, LfO-SUNDER separately controls transition-level distribution matching and policy deviation.

This transition-level parameterization also removes the need for a separate policy-normalization constraint. Summing Eq. 84 over s' gives

$$\sum_{s'} d'(s,s') = \left(\sum_{\bar{s}} d'(s,\bar{s})\right) \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s,a) = \left(\sum_{\bar{s}} d'(s,\bar{s})\right) \sum_a \pi(a \mid s), \quad (85)$$

where we used $\sum_{s'} P(s' \mid s,a) = 1$. Since the two transition sums represent the same row marginal,

$$\left(\sum_{s'} d'(s,s')\right) \left(\sum_a \pi(a \mid s) - 1\right) = 0. \quad (86)$$

Hence, policy normalization is automatically satisfied on the support of the induced state occupancy and need not be imposed as a separate constraint. Therefore, Eqs. 83 and 84 provide a compact representation of the same policy-induced occupancies as the general SUNDER constraints, with $\hat{d}$ recovered directly from the row marginal of d'.

Having established this compact representation, we next eliminate π and d' to obtain the corresponding dual objective. We introduce $V(s)$ and $Q(s,s')$ as the Lagrange multipliers for Eqs. 83 and 84, respectively. For convenience, we scale both constraint terms by ω. Since $\omega > 0$, this only rescales the corresponding multipliers. Using the same sign convention for Q as in the general SUNDER formulation, the Lagrangian is

$$\begin{aligned}\mathcal{L} =& - D_f(\omega d' + (1-\omega)q'_I \,\|\, \omega q'_E + (1-\omega)q'_I) - \alpha \sum_{s,a} \left(\sum_{\bar{s}} d'(s,\bar{s})\right) \pi_I(a \mid s) f\left(\frac{\pi(a \mid s)}{\pi_I(a \mid s)}\right) \\ &+ \omega \sum_{s'} V(s') \left[(1-\gamma) d_0(s') + \gamma \sum_s d'(s,s') - \sum_{s''} d'(s',s'')\right] \\ &+ \omega \sum_{s,s'} Q(s,s') \left[d'(s,s') - \left(\sum_{\bar{s}} d'(s,\bar{s})\right) \sum_a \pi(a \mid s) P(s' \mid s,a)\right]. \end{aligned} \quad (87)$$

We first eliminate the policy. Define

$$e_\pi(s,a;Q) := -\mathbb{E}_{s' \sim P(\cdot \mid s,a)}[Q(s,s')]. \quad (88)$$

The policy-dependent terms can then be written as

$$\sum_s \left(\sum_{\bar{s}} d'(s,\bar{s})\right) \sum_a \pi_I(a \mid s) \left[\omega \frac{\pi(a \mid s)}{\pi_I(a \mid s)} e_\pi(s,a;Q) - \alpha f\left(\frac{\pi(a \mid s)}{\pi_I(a \mid s)}\right)\right]. \quad (89)$$

Applying Fenchel conjugacy pointwise gives

$$\sup_{\pi(a|s) \geq 0} \left\{\omega \frac{\pi(a \mid s)}{\pi_I(a \mid s)} e_\pi(s,a;Q) - \alpha f\left(\frac{\pi(a \mid s)}{\pi_I(a \mid s)}\right)\right\} = \alpha f^*\left(\frac{\omega e_\pi(s,a;Q)}{\alpha}\right). \quad (90)$$

We therefore define

$$e_{\hat{d}}(s;Q) := \frac{\alpha}{\omega} \mathbb{E}_{a \sim \pi_I(\cdot \mid s)} \left[f^*\left(\frac{\omega e_\pi(s,a;Q)}{\alpha}\right)\right]. \quad (91)$$

After eliminating π, the remaining terms involving d' become

$$\begin{aligned}\max_{\pi \geq 0} \mathcal{L} =& \omega(1-\gamma)\mathbb{E}_{s \sim d_0}[V(s)] - D_f(\omega d' + (1-\omega)q'_I \,\|\, \omega q'_E + (1-\omega)q'_I) \\ &+ \omega \sum_{s,s'} d'(s,s') \left[\gamma V(s') - V(s) + Q(s,s') + e_{\hat{d}}(s;Q)\right]. \end{aligned} \quad (92)$$

It remains to eliminate the transition occupancy. Using

$$\omega d'(s,s') = [\omega d'(s,s') + (1-\omega)q'_I(s,s')] - (1-\omega)q'_I(s,s'), \tag{93}$$

the linear term in d' can be expressed under the same mixture that appears in the f-divergence. Following its empirical success in prior work (Pham et al., 2025), we assume compatibility of the Fenchel maximizer with $d' \geq 0$ and apply Fenchel conjugacy pointwise. gives the LfO-SUNDER dual

$$\begin{aligned}\min_{V,Q} \quad & \omega(1-\gamma)\mathbb{E}_{s\sim d_0}[V(s)] \\ & + \mathbb{E}_{(s,s')\sim \omega q'_E + (1-\omega)q'_I}\left[f^*\big(\gamma V(s') - V(s) + Q(s,s') + e_{\hat{d}}(s;Q)\big)\right] \\ & - (1-\omega)\mathbb{E}_{(s,s')\sim q'_I}\left[\gamma V(s') - V(s) + Q(s,s') + e_{\hat{d}}(s;Q)\right].\end{aligned} \tag{94}$$

Thus, the compact transition representation eliminates both explicit optimization variables π and d', leaving an objective over only V and Q.

When f^* is differentiable, the eliminated policy density ratio is recovered from the policy Fenchel optimum as

$$w^*_\pi(s,a) := \frac{\pi^*(a \mid s)}{\pi_I(a \mid s)} = (f^*)'\left(\frac{\omega e_\pi(s,a;Q)}{\alpha}\right). \tag{95}$$

Similarly, define the transition-mixture density ratio

$$w^*_{d'}(s,s') := \frac{\omega d'^*(s,s') + (1-\omega)q'_I(s,s')}{\omega q'_E(s,s') + (1-\omega)q'_I(s,s')}. \tag{96}$$

The Fenchel optimality condition gives

$$w^*_{d'}(s,s') = (f^*)'\big(\gamma V(s') - V(s) + Q(s,s') + e_{\hat{d}}(s;Q)\big). \tag{97}$$

Equivalently, the optimized transition occupancy can be recovered as

$$d'^*(s,s') = \frac{1}{\omega}\left[\left(\omega q'_E(s,s') + (1-\omega)q'_I(s,s')\right) w^*_{d'}(s,s') - (1-\omega)q'_I(s,s')\right]. \tag{98}$$

Hence, LfO-SUNDER uses transition-level distribution correction to move the induced occupancy toward expert transitions, while α independently controls policy deviation from the imperfect-data policy. The compact d' parameterization further removes the redundant state occupancy and policy-normalization multiplier, leaving only the transition-level dual variables V and Q.

B A COMPLEMENTARY OCCUPANCY VIEW OF SUNDER

The main text represents SUNDER through the state occupancy $\hat{d}$ and policy π, making state visitation and action selection explicit as separate optimization components. Proposition 1 establishes strong duality directly in this apparently non-affine representation.

Here, we provide a complementary view of the same structure in occupancy coordinates. Under the standard reparameterization $\rho = \hat{d}\pi$, the constraints become affine and the policy regularizer takes a convex perspective form. This view makes the convex geometry behind Proposition 1 explicit while leaving the underlying SUNDER objective unchanged.

The two representations therefore emphasize different aspects of SUNDER: $(\hat{d}, \pi)$ exposes the state–policy decomposition, whereas $(\rho, \hat{d})$ exposes its convex structure.

B.1 Perspective Convexity in Occupancy Coordinates

To expose the convex structure, define the state-action occupancy

$$\rho(s,a) := \hat{d}(s)\pi(a \mid s).$$

Using $(\rho, \hat{d}, d')$ as the optimization variables, the general SUNDER primal in Eqs. 5a–5d can be equivalently written as

$$\begin{aligned}
\max_{\rho,\hat{d},d'\geq 0} \quad & G(\hat{d}, d') + \sum_{s,a} \rho(s,a)c(s,a) - \alpha \sum_{s,a} \hat{d}(s)\pi_{\mathrm{ref}}(a \mid s) f\left(\frac{\rho(s,a)}{\hat{d}(s)\pi_{\mathrm{ref}}(a \mid s)}\right) \\
\text{s.t.} \quad & \hat{d}(s) = \sum_a \rho(s,a), \qquad \forall s, \\
& d'(s,s') = \sum_a \rho(s,a) P(s' \mid s,a), \qquad \forall s, s', \\
& \sum_a \rho(s',a) - \gamma \sum_{s,a} \rho(s,a) P(s' \mid s,a) = (1-\gamma) d_0(s'), \qquad \forall s'.
\end{aligned}$$

This is a reparameterization of the state–policy formulation. For any feasible $(\hat{d}, d', \pi)$, setting $\rho = \hat{d}\pi$ gives a feasible point above with the same objective value. Conversely, given a feasible $(\rho, \hat{d}, d')$, the corresponding policy is recovered as $\pi(a \mid s) = \rho(s,a)/\hat{d}(s)$ whenever $\hat{d}(s) > 0$, with $\pi(\cdot \mid s)$ defined arbitrarily otherwise. Moreover, $\rho(s,a)/(\hat{d}(s)\pi_{\mathrm{ref}}(a \mid s)) = \pi(a \mid s)/\pi_{\mathrm{ref}}(a \mid s)$, so the policy-regularization term is unchanged.

The convex structure becomes explicit in these coordinates. For convex f, its perspective

$$(x, t) \mapsto t f(x/t)$$

is jointly convex. Therefore,

$$(\rho(s,a), \hat{d}(s)) \mapsto \hat{d}(s)\pi_{\mathrm{ref}}(a \mid s) f\left(\frac{\rho(s,a)}{\hat{d}(s)\pi_{\mathrm{ref}}(a \mid s)}\right)$$

is jointly convex in $(\rho(s,a), \hat{d}(s))$. The local objective is affine in ρ, and $G(\hat{d}, d')$ is jointly concave by assumption. All constraints are affine in $(\rho, \hat{d}, d')$. Hence, the occupancy-coordinate formulation satisfies the standard affine conditions for strong duality.

This view also explains the mixing construction used in Proposition 1. There, two state–policy pairs are combined through $\hat{d}_\theta \pi_\theta = \theta \hat{d}_1 \pi_1 + (1-\theta)\hat{d}_2 \pi_2$. In occupancy coordinates, this is simply

$$\rho_\theta = \theta \rho_1 + (1-\theta)\rho_2.$$

Thus, the hidden convexity established directly in the state–policy representation corresponds to ordinary convexity of the induced occupancy together with the perspective convexity of the policy regularizer.

B.2 Complementary Roles of the Two Representations

The state–policy representation makes the two components controlled by SUNDER explicit:

$$\underbrace{\hat{d}(s)}_{\text{state visitation}}, \qquad \underbrace{\pi(a \mid s)}_{\text{action selection}}.$$

Here, $\hat{d}$ describes where the policy visits, while π describes how it acts at each state. This separation is the conceptual basis of SUNDER: objectives over induced state distributions and regularization of action selection can be specified independently, while their product $\hat{d}\pi$ remains the induced state-action occupancy.

This representation also matches the modular optimization used in the main text. SUNDER first eliminates π through the policy regularizer and then eliminates the remaining occupancy variables $(\hat{d}, d')$ according to the task objective G. Thus, the same state–policy decomposition that defines the framework also organizes its derivation across different instantiations.

The occupancy representation emphasizes a different aspect of the same problem. By absorbing the state–policy product into $\rho = \hat{d}\pi$, it makes the affine constraints and the perspective-convex

structure of the policy regularizer explicit. The two representations therefore serve complementary roles: $(\hat{d}, \pi)$ exposes what SUNDER controls, whereas $(\rho, \hat{d})$ exposes the convex geometry that makes this control tractable.

C Why Decoupling Matters: A Theoretical Perspective

One may still wonder whether separate coefficients are actually necessary. Suppose a desired policy is obtained by regularizing state and policy shift with different strengths. Since a single joint coefficient does not force the resulting state and policy shifts to be equal, one might expect that a suitable joint coefficient could nevertheless recover the same policy. The following result shows that this is not possible in general.

To formalize separate regularization, consider an occupancy $\rho(s, a) = \hat{d}\pi$, and let $J(\rho)$ denote the task objective. Define $D_s(\rho) := D_f(\hat{d}\|\hat{q}_B)$ and $D_\pi(\rho) := \mathbb{E}[D_f(\pi(\cdot \mid s)\|\pi_B(\cdot \mid s))]$. Separate regularization then optimizes

$$F_{\alpha,\beta}(\rho) := J(\rho) - \alpha D_\pi(\rho) - \beta D_s(\rho),$$

where α and β control policy and state shift, respectively. Under KL regularization, the chain rule gives $D_{\mathrm{KL}}(\rho\|q_B) = D_s(\rho) + D_\pi(\rho)$, so joint regularization with coefficient η corresponds to $F_{\eta,\eta}$. To state the result, define the two-budget value function

$$\mathcal{V}(\epsilon_s, \epsilon_\pi) := \max_{\rho \in \mathcal{C}_{\mathrm{flow}}} \{J(\rho) : D_s(\rho) \le \epsilon_s, D_\pi(\rho) \le \epsilon_\pi\}.$$

Theorem 1 (Non-recoverability under joint KL regularization)**.** *Let $\rho^\star \in \arg\max_{\rho \in \mathcal{C}_{\mathrm{flow}}} F_{\alpha,\beta}(\rho)$ with $\alpha, \beta > 0$, and define $\epsilon^\star := (D_s(\rho^\star), D_\pi(\rho^\star))$. Suppose $\mathcal{V}$ is differentiable at $\epsilon^\star$. If $\alpha \neq \beta$, then*

$$\rho^\star \notin \arg\max_{\rho \in \mathcal{C}_{\mathrm{flow}}} F_{\eta,\eta}(\rho), \qquad \forall \eta > 0.$$

That is, no choice of a joint coefficient η can recover the separately regularized optimum.

Proof. Optimality of $\rho^\star$ under $F_{\alpha,\beta}$ gives, for any feasible ρ,

$$J(\rho) \le J(\rho^\star) + \beta(D_s(\rho) - D_s(\rho^\star)) + \alpha(D_\pi(\rho) - D_\pi(\rho^\star)).$$

Hence, (β, α) supports $\mathcal{V}$ at $\epsilon^\star$, and differentiability gives $\nabla\mathcal{V}(\epsilon^\star) = (\beta, \alpha)$. If the same $\rho^\star$ were optimal under joint KL regularization with some $\eta > 0$, the KL chain rule similarly gives $\nabla\mathcal{V}(\epsilon^\star) = (\eta, \eta)$. Therefore $\alpha = \beta = \eta$, contradicting $\alpha \neq \beta$. □

Thus, when the desired optimum requires unequal state and policy regularization, standard joint regularization cannot generally recover it. This motivates a framework that controls the two shifts separately. The KL result, however, relies on the exact chain rule. We next ask whether the same non-recoverability persists for smooth f-divergences, where such an exact decomposition is unavailable. To answer this, we compare the local state and policy curvature around q_B.

Let Z have orthonormal columns spanning the Bellman-feasible tangent space at q_B, so a nearby feasible occupancy can be written as $\rho = q_B + Zx$. Assume that f is twice continuously differentiable near 1 with $f''(1) > 0$, and define

$$H_s := \nabla_x^2 D_s(q_B + Zx)\big|_{x=0}, \qquad H_\pi := \nabla_x^2 D_\pi(q_B + Zx)\big|_{x=0}.$$

The standard quadratic expansion of f-divergence gives

$$H := \nabla_x^2 D_f(q_B + Zx\|q_B)\big|_{x=0} = H_s + H_\pi \succ 0.$$

Fix $\alpha, \beta > 0$ with $\alpha \neq \beta$, and define $A := \alpha H_\pi + \beta H_s$ and $g := Z^\top r$. For reward scale $t > 0$, let

$$\rho_t^\star = q_B + Zx_t \in \arg\max_{\rho \in \mathcal{C}_{\mathrm{flow}}} \{t\langle r, \rho\rangle - \alpha D_\pi(\rho) - \beta D_s(\rho)\}.$$

Theorem 2 (Generic local non-recoverability beyond KL)**.** *Suppose H_s and H_π are linearly independent. Then, for Lebesgue-almost every projected reward $g = Z^\top r$, there exists $t_0 > 0$ such that*

$$\rho_t^\star \notin \arg\max_{\rho \in \mathcal{C}_{\mathrm{flow}}} \{t\langle r, \rho\rangle - \eta D_f(\rho\|q_B)\}, \qquad \forall\, 0 < t < t_0,\ \forall\, \eta > 0.$$

Table 5: Normalized returns on the Farama Minari MuJoCo (v0), D4RL AntMaze (v1), and D4RL Adroit (v2) datasets, loaded through Minari. Results are reported as mean ± standard error (SE); entries whose intervals overlap with that of the highest-mean method are bolded.

Tasks	FQL	GFP	IFQL	IQL	ReBRAC	OptiDICE (Flow)	OptiDICE (Gaussian)	Opt-SUNDER (Flow)	Opt-SUNDER (One-step)
MuJoCo									
halfcheetah-simple	70.02 ± 0.06	**72.60 ± 0.19**	67.67 ± 0.08	64.86 ± 0.24	70.97 ± 0.19	60.88 ± 0.05	60.55 ± 0.13	60.71 ± 0.15	60.80 ± 0.08
halfcheetah-medium	110.29 ± 3.17	117.02 ± 4.22	59.13 ± 2.17	78.64 ± 5.61	109.63 ± 5.18	**122.83 ± 1.34**	2.14 ± 1.22	**124.28 ± 1.80**	**122.99 ± 2.55**
halfcheetah-expert	123.97 ± 3.62	133.01 ± 1.98	76.46 ± 7.30	83.85 ± 10.89	22.20 ± 1.57	139.17 ± 0.50	139.11 ± 0.56	**141.57 ± 0.35**	**141.86 ± 0.92**
hopper-simple	82.79 ± 2.03	90.94 ± 0.91	32.39 ± 3.75	41.79 ± 7.70	**97.18 ± 2.20**	97.83 ± 0.14	98.17 ± 0.19	**98.59 ± 0.08**	97.89 ± 0.13
hopper-medium	66.75 ± 6.85	71.66 ± 6.73	91.22 ± 3.66	78.67 ± 10.13	**110.52 ± 0.38**	110.08 ± 0.05	**108.95 ± 1.87**	**110.16 ± 0.05**	**110.17 ± 0.10**
hopper-expert	106.40 ± 2.78	108.55 ± 5.09	76.51 ± 5.34	48.85 ± 6.94	**124.39 ± 0.97**	117.80 ± 3.16	118.48 ± 3.53	115.98 ± 3.23	115.28 ± 4.36
walker2d-simple	**89.67 ± 0.44**	**90.04 ± 0.53**	34.64 ± 4.60	74.29 ± 3.94	**90.91 ± 0.99**	89.40 ± 0.30	**90.78 ± 0.21**	**90.18 ± 0.08**	**90.36 ± 0.22**
walker2d-medium	**132.98 ± 0.73**	**132.92 ± 0.50**	106.02 ± 3.72	101.18 ± 8.69	**132.30 ± 0.57**	**132.35 ± 0.67**	54.83 ± 17.71	**132.90 ± 0.54**	**131.49 ± 1.10**
walker2d-expert	148.11 ± 1.20	148.08 ± 1.50	77.18 ± 10.71	76.58 ± 16.55	149.25 ± 0.37	149.46 ± 0.32	131.83 ± 14.30	**150.21 ± 0.15**	**150.03 ± 0.53**
Average	103.44 ± 0.56	107.20 ± 1.00	69.02 ± 2.20	72.08 ± 3.19	100.81 ± 0.82	**113.31 ± 0.42**	89.43 ± 2.92	**113.53 ± 0.39**	**113.43 ± 0.59**
AntMaze									
umaze	**98.54 ± 2.14**	**96.44 ± 1.37**	73.93 ± 2.25	83.07 ± 2.32	**95.78 ± 0.81**	**96.52 ± 1.81**	59.76 ± 6.51	**98.11 ± 1.16**	**96.24 ± 1.22**
umaze-diverse	**118.62 ± 1.76**	**116.82 ± 1.73**	104.65 ± 2.69	96.37 ± 3.19	109.28 ± 2.79	104.41 ± 1.69	41.57 ± 5.61	**117.43 ± 2.03**	**117.12 ± 1.74**
medium-play	134.01 ± 6.17	123.63 ± 3.12	108.68 ± 3.51†	36.13 ± 4.19	67.24 ± 3.48	129.70 ± 2.99	6.24 ± 1.64	**149.33 ± 5.50**	**156.37 ± 5.45**
medium-diverse	127.42 ± 5.61	**162.34 ± 6.26**	102.65 ± 4.07	6.55 ± 2.01	47.50 ± 6.83	140.90 ± 4.76	1.48 ± 1.03	**152.52 ± 4.02**	147.87 ± 3.05
large-play	61.76 ± 3.66	56.27 ± 3.65	46.68 ± 4.33	11.38 ± 2.17	45.49 ± 6.15	62.92 ± 5.77	2.39 ± 1.56	**81.10 ± 3.72**	68.37 ± 3.21
large-diverse	83.09 ± 2.89	88.90 ± 3.43	59.55 ± 2.34	11.77 ± 0.78	64.14 ± 2.65	83.78 ± 4.13	1.13 ± 0.74	**105.05 ± 3.84**	93.09 ± 4.27
Average	103.91 ± 2.00	107.40 ± 1.49	82.30 ± 1.18†	40.88 ± 1.35	71.57 ± 1.43	103.04 ± 1.72	18.76 ± 1.65	**117.26 ± 1.92**	113.18 ± 1.41
Adroit									
door-cloned	**1.26 ± 0.21**	0.26 ± 0.10	**0.77 ± 0.29**	0.24 ± 0.01	0.08 ± 0.02	0.23 ± 0.01	0.11 ± 0.03	0.23 ± 0.01	0.223 ± 0.003
door-expert	101.91 ± 0.49	96.52 ± 6.06	96.57 ± 2.85	100.97 ± 0.89	21.89 ± 11.49	**105.43 ± 0.11**	104.21 ± 0.89	105.16 ± 0.12	104.60 ± 0.37
door-human	0.23 ± 0.13	0.26 ± 0.03	3.80 ± 0.70	0.13 ± 0.03	0.18 ± 0.02	0.10 ± 0.04	0.16 ± 0.02	**5.83 ± 0.65**	3.70 ± 0.58
hammer-cloned	**25.75 ± 3.76**	11.12 ± 5.15	1.30 ± 0.27	0.55 ± 0.08	7.17 ± 1.15	2.72 ± 0.24	2.63 ± 1.29	1.23 ± 0.48	0.83 ± 0.23
hammer-expert	130.84 ± 1.01	123.80 ± 4.26	123.46 ± 2.84	132.20 ± 0.20	131.29 ± 0.15	**133.89 ± 0.26**	132.96 ± 0.15	132.28 ± 0.07	132.81 ± 0.07
hammer-human	**4.03 ± 0.87**	2.15 ± 0.99	2.51 ± 0.40	0.94 ± 0.41	0.46 ± 0.30	**4.08 ± 0.77**	0.49 ± 0.16	**3.00 ± 0.67**	1.72 ± 0.35
pen-cloned	**54.22 ± 3.01**	**58.50 ± 2.86**	**58.17 ± 2.36**	43.73 ± 1.92	**55.56 ± 2.84**	**56.74 ± 6.18**	2.99 ± 1.10	**53.98 ± 1.88**	**54.07 ± 1.69**
pen-expert	112.48 ± 0.85	**117.04 ± 0.57**	108.94 ± 1.52	101.80 ± 2.49	112.39 ± 2.02	81.99 ± 2.74	80.04 ± 3.21	97.19 ± 1.55	99.06 ± 1.03
pen-human	11.10 ± 2.30	48.35 ± 2.57	53.86 ± 2.84	38.77 ± 2.23	28.08 ± 1.38	46.50 ± 7.42	5.29 ± 1.12	75.03 ± 1.23	**77.93 ± 1.12**
relocate-cloned	0.22 ± 0.09	**0.53 ± 0.17**	0.01 ± 0.02	0.01 ± 0.01	0.09 ± 0.04	0.00 ± 0.00	0.03 ± 0.02	0.00 ± 0.00	0.00 ± 0.00
relocate-expert	**105.08 ± 0.42**	**103.24 ± 2.07**	91.87 ± 3.71	102.05 ± 0.79	103.31 ± 1.22	102.20 ± 1.55	89.29 ± 5.24	**104.89 ± 0.48**	**104.71 ± 0.68**
relocate-human	0.01 ± 0.06	0.00 ± 0.00	**0.05 ± 0.05**	0.03 ± 0.01	**0.10 ± 0.02**	−0.05 ± 0.01	0.05 ± 0.01	0.00 ± 0.00	0.00 ± 0.00
Average	45.59 ± 0.44	46.81 ± 1.00	45.11 ± 0.48	43.45 ± 0.31	38.39 ± 0.84	44.49 ± 1.24	34.85 ± 0.54	**48.15 ± 0.27**	**48.30 ± 0.21**

Proof. We first identify how separate regularization moves the optimum away from q_B. For sufficiently small t, the optimum remains in a neighborhood of q_B, and the first-order optimality condition and the implicit function theorem give $x_t = tA^{-1}g + o(t)$. Thus, the relative state and policy weights determine the local direction of the optimum through the curvature A.

The remaining question is whether joint regularization can follow the same direction by choosing an appropriate coefficient. Suppose it recovers the same optimum along a sequence $t_n \downarrow 0$, with coefficients $\eta_n > 0$. The joint first-order condition gives $t_n g = \eta_n \nabla_x D_f(q_B + Zx_{t_n} \| q_B)$. Since the joint divergence has local curvature H, $\nabla_x D_f(q_B + Zx \| q_B) = Hx + o(\|x\|)$. Substituting the expansion of x_{t_n} therefore gives $g = \eta_n \left(HA^{-1}g + o(1)\right)$. For $g \neq 0$, this implies, along a subsequence, $\eta_n \to \eta_0 \in (0, \infty)$ and hence $g = \eta_0 HA^{-1}g$. Letting $v = A^{-1}g$, recovery therefore requires $Av = \eta_0 Hv$. Thus, the direction selected by separate regularization must be a generalized eigendirection of the separate and joint curvatures.

It remains to determine whether this alignment is generic. If $A = \lambda H$ for some λ, then

$$(\alpha - \lambda)H_\pi + (\beta - \lambda)H_s = 0.$$

Linear independence of H_s and H_π would then imply $\alpha = \beta = \lambda$, contradicting $\alpha \neq \beta$. Hence, A and H are not proportional. Since A and H are symmetric and $H \succ 0$, their generalized eigendirections form a finite union of proper linear subspaces. The corresponding reward directions $g = Av$ therefore also form a measure-zero set.

For almost every g, joint recovery thus cannot occur for arbitrarily small t. Otherwise, a sequence $t_n \downarrow 0$ satisfying the recovery condition above would exist. Therefore, some $t_0 > 0$ satisfies the claimed non-recoverability for every $0 < t < t_0$ and every $\eta > 0$. □

D Experimental Results

This appendix provides the full task-level results and selected hyperparameters for the experiments in Section 4. We also report a one-step policy distilled from flow-based Opt-SUNDER to test whether iterative flow inference can be removed with little performance loss.

Table 6: State entropy $\hat{H}(s)$ on maze and continuous-control tasks, estimated by discretizing the listed state coordinates to approximate the state density. (mean $\pm$ SE over 8 seeds)

Task / Domain	Coordinates	Buffer	APT	PROTO	SEMDICE	Ent-SUNDER(Ours)
Maze square_a	(x, y)	0.572	0.592 ± 0.098	0.295 ± 0.111	0.722 ± 0.147	$\mathbf{1.326 \pm 0.086}$
Maze square_b	(x, y)	0.572	0.091 ± 0.027	0.147 ± 0.096	0.772 ± 0.113	$\mathbf{1.333 \pm 0.138}$
Maze square_c	(x, y)	0.572	0.356 ± 0.112	0.094 ± 0.021	0.755 ± 0.189	$\mathbf{1.403 \pm 0.062}$
Maze square_d	(x, y)	0.606	0.263 ± 0.085	0.213 ± 0.112	1.151 ± 0.243	$\mathbf{1.940 \pm 0.066}$
Maze tree	(x, y)	0.648	0.198 ± 0.083	0.199 ± 0.073	0.359 ± 0.094	$\mathbf{1.071 \pm 0.074}$
Maze bottleneck	(x, y)	0.776	0.377 ± 0.099	0.803 ± 0.000	1.118 ± 0.168	$\mathbf{1.749 \pm 0.071}$
Maze large	(x, y)	0.776	0.219 ± 0.077	0.263 ± 0.015	0.981 ± 0.171	$\mathbf{1.399 \pm 0.124}$
Reacher	$(\theta_1, \theta_2, \dot{\theta}_1, \dot{\theta}_2)$	3.502	2.525 ± 0.075	3.008 ± 0.068	5.560 ± 0.022	$\mathbf{5.620 \pm 0.024}$
Acrobot	(θ_1, θ_2)	2.178	1.197 ± 0.269	1.476 ± 0.311	2.226 ± 0.138	$\mathbf{2.915 \pm 0.288}$
Point mass	(x, y, v_x, v_y)	3.602	4.442 ± 0.019	4.293 ± 0.044	7.654 ± 0.059	$\mathbf{7.795 \pm 0.059}$
Pendulum	$(\theta, \dot{\theta})$	3.093	3.043 ± 0.031	2.972 ± 0.020	$\mathbf{8.946 \pm 0.039}$	$\mathbf{9.002 \pm 0.021}$

Table 7: Normalized returns over 15 D4RL datasets (mean $\pm$ SE over 5 seeds), averaged within each domain: **MuJoCo** locomotion (9) and **Adroit** manipulation (6). LobsDICE was implemented following the equations in the original paper and added to our anonymized GitHub repository. All other baselines use their official GitHub implementations and released configurations.

Offline Dataset	Expert Traj.	ReCOIL	DILO	LobsDICE	SMODICE	OTR	IOSTOM	LfO-SUNDER (OURS)
MuJoCo								
Hopper Medium	200	$\mathbf{101.2 \pm 10.4}$	$\mathbf{108.4 \pm 4.0}$	1.82 ± 0.0	48.19 ± 1.51	74.8 ± 8.2	$\mathbf{109.8 \pm 0.5}$	59.89 ± 6.46
Hopper Medium	30	49.7 ± 25.5	45.3 ± 17.8	59.10 ± 1.21	53.92 ± 2.16	81.5 ± 4.0	$\mathbf{106.2 \pm 1.9}$	53.66 ± 1.73
Hopper Medium-Replay	200	96.6 ± 2.5	31.1 ± 10.7	73.87 ± 16.22	45.42 ± 4.45	52.3 ± 29.9	$\mathbf{109.8 \pm 1.0}$	57.43 ± 7.47
Hopper Medium-Replay	30	55.7 ± 17.4	37.2 ± 19.7	24.82 ± 2.40	29.20 ± 4.07	85.0 ± 5.0	$\mathbf{109.2 \pm 1.3}$	37.78 ± 6.65
Hopper Medium-Expert	200	103.7 ± 2.9	$\mathbf{110.0 \pm 2.6}$	0.0 ± 0.0	69.70 ± 5.73	$\mathbf{100.7 \pm 12.1}$	$\mathbf{109.4 \pm 0.3}$	100.87 ± 0.19
Hopper Medium-Expert	30	$\mathbf{107.8 \pm 3.5}$	$\mathbf{108.0 \pm 3.5}$	0.68 ± 0.0	74.17 ± 4.01	$\mathbf{102.1 \pm 11.3}$	$\mathbf{110.4 \pm 0.5}$	$\mathbf{110.82 \pm 0.20}$
HalfCheetah Medium	200	60.5 ± 24.9	$\mathbf{90.0 \pm 0.7}$	75.34 ± 3.25	56.21 ± 3.45	67.7 ± 7.8	$\mathbf{90.5 \pm 0.8}$	52.72 ± 9.00
HalfCheetah Medium	30	41.8 ± 0.7	23.5 ± 15.5	$\mathbf{57.76 \pm 2.20}$	22.06 ± 5.48	42.8 ± 0.2	$\mathbf{52.7 \pm 11.8}$	42.74 ± 0.25
HalfCheetah Medium-Replay	200	38.2 ± 15.2	$\mathbf{92.0 \pm 0.9}$	80.98 ± 3.17	72.60 ± 5.16	84.2 ± 4.1	84.9 ± 5.7	51.63 ± 3.42
HalfCheetah Medium-Replay	30	$\mathbf{27.0 \pm 17.9}$	0.5 ± 1.1	$\mathbf{48.32 \pm 4.96}$	$\mathbf{42.11 \pm 3.05}$	37.9 ± 1.5	$\mathbf{36.3 \pm 7.4}$	34.69 ± 0.54
HalfCheetah Medium-Expert	200	$\mathbf{90.0 \pm 2.5}$	$\mathbf{92.2 \pm 0.5}$	$\mathbf{91.45 \pm 1.53}$	81.82 ± 1.20	$\mathbf{91.7 \pm 1.1}$	$\mathbf{89.2 \pm 3.4}$	84.77 ± 0.26
HalfCheetah Medium-Expert	30	91.4 ± 0.5	$\mathbf{77.9 \pm 20.5}$	$\mathbf{93.79 \pm 1.10}$	79.81 ± 2.05	$\mathbf{90.9 \pm 2.7}$	$\mathbf{91.0 \pm 1.9}$	74.17 ± 2.09
Walker2d Medium	200	$\mathbf{106.7 \pm 2.9}$	83.4 ± 11.2	90.91 ± 15.32	73.27 ± 8.62	83.8 ± 4.4	$\mathbf{108.2 \pm 0.2}$	88.18 ± 6.74
Walker2d Medium	30	57.5 ± 15.5	72.2 ± 4.1	91.13 ± 4.40	74.23 ± 2.96	79.9 ± 2.7	$\mathbf{108.3 \pm 0.3}$	79.53 ± 6.56
Walker2d Medium-Replay	200	$\mathbf{105.3 \pm 6.4}$	$\mathbf{98.2 \pm 13.8}$	105.54 ± 1.81	81.30 ± 11.97	89.1 ± 8.4	$\mathbf{107.9 \pm 0.0}$	35.17 ± 2.59
Walker2d Medium-Replay	30	$\mathbf{81.0 \pm 43.1}$	20.1 ± 7.0	91.70 ± 10.19	40.85 ± 10.51	70.1 ± 4.7	$\mathbf{107.9 \pm 1.0}$	34.42 ± 4.27
Walker2d Medium-Expert	200	$\mathbf{108.6 \pm 0.2}$	108.4 ± 0.3	72.13 ± 35.92	79.97 ± 13.21	$\mathbf{109.3 \pm 0.5}$	107.5 ± 0.5	108.33 ± 0.22
Walker2d Medium-Expert	30	108.1 ± 0.6	108.2 ± 0.1	45.82 ± 25.51	95.17 ± 4.98	$\mathbf{109.4 \pm 0.6}$	107.5 ± 0.5	108.14 ± 0.21
MuJoCo Average	–	79.5 ± 3.69	72.6 ± 2.42	61.4 ± 2.85	62.2 ± 1.48	80.7 ± 2.14	$\mathbf{97.0 \pm 0.88}$	67.50 ± 0.95
Adroit								
Pen Human	30	$\mathbf{88.6 \pm 18.7}$	7.8 ± 9.7	86.56 ± 6.55	61.76 ± 8.54	$\mathbf{90.3 \pm 19.4}$	$\mathbf{99.7 \pm 4.2}$	$\mathbf{105.86 \pm 5.57}$
Door Human	30	94.2 ± 5.7	0.3 ± 1.0	13.47 ± 2.77	9.31 ± 1.64	$\mathbf{88.9 \pm 17.4}$	$\mathbf{101.7 \pm 1.3}$	$\mathbf{99.60 \pm 1.46}$
Hammer Human	30	$\mathbf{106.8 \pm 14.2}$	0.2 ± 0.3	47.86 ± 8.13	68.77 ± 9.47	$\mathbf{106.2 \pm 22.4}$	73.5 ± 9.5	$\mathbf{108.24 \pm 5.36}$
Pen Cloned	30	$\mathbf{33.6 \pm 19.7}$	21.3 ± 17.7	30.54 ± 8.79	3.42 ± 2.35	$\mathbf{49.2 \pm 28.4}$	9.6 ± 2.8	$\mathbf{51.34 \pm 6.49}$
Door Cloned	30	1.7 ± 3.1	$\mathbf{55.2 \pm 29.4}$	-0.24 ± 0.12	5.58 ± 2.86	0.0 ± 0.0	33.3 ± 6.0	$\mathbf{85.15 \pm 6.83}$
Hammer Cloned	30	3.5 ± 4.8	$\mathbf{45.9 \pm 20.9}$	0.27 ± 0.02	2.96 ± 0.64	0.4 ± 0.0	$\mathbf{55.6 \pm 5.6}$	$\mathbf{49.06 \pm 8.38}$
Adroit Average	–	54.73 ± 5.28	21.78 ± 6.89	29.74 ± 2.32	25.30 ± 2.23	55.83 ± 7.43	62.23 ± 2.27	$\mathbf{83.21 \pm 2.28}$

Offline reward maximization. Table 5 reports the full task-level results underlying the domain averages in Table 2(b). We also compare OPT-SUNDER (FLOW) and OPT-SUNDER (ONE-STEP) to evaluate one-step policy distillation.

One-step policy distillation. OPT-SUNDER (FLOW) uses 30 integration steps at evaluation. To remove this iterative inference cost, we distill it into OPT-SUNDER (ONE-STEP). Specifically, we train an expressive network to directly match the action produced by the flow policy for each state–latent pair (s, z). We fix $z = 0$ at evaluation, which we found to work well empirically. As shown in Table 5, OPT-SUNDER (ONE-STEP) retains nearly all of the performance of OPT-SUNDER (FLOW): the average scores change from 113.53 to 113.43 on MuJoCo, 117.26 to 113.18 on AntMaze, and 48.15 to 48.30 on Adroit.

State-distribution objectives. We next report the full task-level results for the state-distribution objectives in Section 4. Ent-SUNDER achieves higher state entropy than SEMDICE across all maze and continuous-control tasks. For LfO, Table 7 reports results across datasets and expert-data

Table 8: Selected hyperparameters for SUNDER.

(a) Opt-SUNDER.

MuJoCo			AntMaze			Adroit		
Task	α	β	**Task**	α	β	**Task**	α	β
halfcheetah-simple	1	.25	umaze	.5	.1	door-cloned	1	.25
halfcheetah-medium	1	.5	umaze-diverse	.5	1	door-human	.5	.5
halfcheetah-expert	1	.5	medium-play	.5	.75	door-expert	.1	.005
hopper-simple	1	.25	medium-diverse	.5	.75	hammer-cloned	.1	.01
hopper-medium	1	.5	large-play	.5	.75	hammer-human	.1	.01
hopper-expert	1	1	large-diverse	.5	.1	hammer-expert	.1	.01
walker2d-simple	1	.5				pen-cloned	.1	.005
walker2d-medium	1	.75				pen-human	.1	.005
walker2d-expert	1	1				pen-expert	.1	.01
						relocate-cloned	.1	.005
						relocate-human	.1	.01
						relocate-expert	.1	.01

(b) LfO-SUNDER.

MuJoCo				Adroit		
Dataset	**Exp.**	ω	α	**Dataset**	ω	α
Hopper Medium	200	.25	.025	Pen Human	.25	.025
Hopper Medium	30	.5	.25	Door Human	.25	.25
Hopper Medium-Replay	200	.5	.25	Hammer Human	.1	.25
Hopper Medium-Replay	30	.1	.025	Pen Cloned	.25	.025
Hopper Medium-Expert	200	.01	.5	Door Cloned	.1	.025
Hopper Medium-Expert	30	.1	.01	Hammer Cloned	.1	.025
HalfCheetah Medium	200	.1	.025			
HalfCheetah Medium	30	.1	.1			
HalfCheetah Medium-Replay	200	.5	.025			
HalfCheetah Medium-Replay	30	.1	.25			
HalfCheetah Medium-Expert	200	.1	.01			
HalfCheetah Medium-Expert	30	.1	.5			
Walker2d Medium	200	.5	.1			
Walker2d Medium	30	.25	.1			
Walker2d Medium-Replay	200	.1	.1			
Walker2d Medium-Replay	30	.25	.25			
Walker2d Medium-Expert	200	.1	.01			
Walker2d Medium-Expert	30	.1	.01			

Flow policy on MuJoCo; Gaussian policy on Adroit.

budgets. LfO-SUNDER performs strongly on Adroit, while its performance on MuJoCo is more variable.

Flow versus Gaussian policy extraction. For OptiDICE, replacing the Gaussian policy with a flow policy improves the average score from 89.43 to 113.31 on MuJoCo, 18.76 to 103.04 on AntMaze, and 34.85 to 44.49 on Adroit. We therefore use flow policies in the main experiments.

Hyperparameters. Tables 8a and 8b report the selected regularization coefficients for Opt-SUNDER and LfO-SUNDER, respectively. State and policy coefficients are tuned independently.

OptiDICE hyperparameters. As with Opt-SUNDER, we choose domain-specific hyperparameter regions expected to perform well and sweep nearby values. For OptiDICE, we sweep

$$\alpha \in \{0.1,\, 0.25,\, 0.5,\, 0.75,\, 1.0\}$$

using a single tuning seed for each task, and report the best-performing setting in the main results.

D.1 Rule-based hyperparameter selection without per-task tuning

The main Minari results2 use per-task hyperparameters, as do the corresponding tuned baselines(Park et al., 2025; Nguimatsia Tiofack et al., 2026; Tarasov et al., 2023; Kostrikov et al., 2022). Since Opt-SUNDER separately controls policy and state regularization through α and β, respectively, a natural question is whether its performance depends critically on task-specific tuning. We therefore consider a rule-based selection scheme that determines α and β using only statistics of the offline dataset. The rules are fixed before evaluation and use no policy evaluation performance for hyperparameter selection.

We first determine β from a coarse measure of dataset quality. We use the same quality-to-β mapping across all three benchmark domains:

$$\beta = \begin{cases} 0.25, & \text{quality} < 50\%, \\ 0.50, & 50\% \le \text{quality} < 75\%, \\ 0.75, & 75\% \le \text{quality} < 90\%, \\ 1.00, & \text{quality} \ge 90\%. \end{cases}$$

The definition of dataset quality follows the information naturally available in each benchmark. For MuJoCo, we use the mean dataset return normalized by the expert return, and fix $\alpha = 1$. Thus, higher-quality datasets receive stronger state regularization through larger β.

For Adroit, we define dataset quality as the fraction of successful trajectories. We select the policy-regularization coefficient as

$$\alpha = \begin{cases} 0.25, & q < 50\% \text{ and } N_{\text{traj}} \ge 100, \\ 0.50, & q < 90\% \text{ otherwise}, \\ 1.00, & q \ge 90\%, \end{cases}$$

Table 9: Comparison of per-task tuning and rule-based hyperparameter selection for Opt-SUNDER.

Hyperparameter selection	MuJoCo	AntMaze	Adroit	Overall
Per-task tuning	113.53 ± 0.39	117.26 ± 1.92	48.15 ± 0.27	85.30
Dataset-statistics rule	113.14 ± 0.47	115.49 ± 1.64	46.42 ± 0.22	84.01

where q is the dataset success rate. We select β using the common quality rule above. When the dataset contains fewer than 100 trajectories, we additionally enforce $\beta \geq 0.5$ to avoid weak state regularization with limited data.

For AntMaze, we use the dataset success rate as the quality measure. Because these tasks are long-horizon and sparse-reward, we fix $\alpha = 0.5$ and select only β using the common quality rule.

We use $\gamma = 0.99$ for MuJoCo and Adroit and $\gamma = 0.995$ for AntMaze. All hyperparameters are therefore determined from the benchmark domain and offline-dataset statistics, without using per-task evaluation returns.

Rule-based selection. Table X compares rule-based hyperparameter selection with the per-task tuning used in the main results. Without using per-task evaluation feedback, the rule-based scheme retains most of the performance: the overall normalized return decreases only from 85.30 to 84.01 across 27 tasks. The differences are also small within each domain, with drops of 0.39 on MuJoCo, 1.77 on AntMaze, and 1.73 on Adroit.

These results suggest that Opt-SUNDER's performance does not primarily depend on per-task tuning of α and β. Most of the performance is retained when the coefficients are selected from coarse offline-dataset statistics, consistent with their distinct roles in controlling policy and state regularization.