SUPPORTDICE: REPLAY-SUPPORTED OCCUPANCY OPTIMIZATION FOR OFFLINE-TO-ONLINE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline-to-online reinforcement learning aims to reuse an offline dataset while adapting as new interaction data arrives. We study this regime from the DICE (distribution-correction estimation) occupancy-optimization perspective. The central difficulty is that the replay buffer keeps changing: it tells us which state-action pairs have been observed and provides samples for training, but raw replay counts need not themselves form a valid discounted occupancy for policy optimization. We propose SupportDICE, which separates these roles by constructing an augmented replay-supported MDP whose exits from the currently covered region are redirected to an absorbing state. The resulting update optimizes valid occupancies on the current replay support, while a full-support anchor lets newly observed, flow-feasible state-action pairs enter future updates. In finite MDPs, the exact population updates have zero duality gap under finite-divergence feasibility; the fixed-support offline phase recovers DICE improvement in the augmented MDP, while the support-expanding online phase satisfies a per-step comparator guarantee on the current replay-supported problem. We also extend the population analysis to dominated standard Borel spaces. A tabular square-grid example shows that support expansion can expose a better feasible occupancy, while replay-centered DICE remains biased toward the old replay distribution. On D4RL benchmarks, SupportDICE achieves the highest average final scores on Adroit and Kitchen among the evaluated methods, while remaining competitive on AntMaze.

1 INTRODUCTION

Many reinforcement learning (RL) problems require both offline data and online adaptation. For example, in robotics, learning from scratch through online interaction can be expensive, slow, or unsafe. Pure offline RL avoids this cost by training from a fixed dataset, but its performance is limited by the dataset's coverage. *Offline-to-online (O2O) RL* offers a natural way to bridge this gap: an agent first uses offline data to obtain a strong initial policy and then uses limited online interaction to expand support and improve further. Existing O2O methods stabilize this transition through behavior regularization (Nair et al., 2020; Kostrikov et al., 2021), conservative or calibrated critics (Kumar et al., 2020; Nakamoto et al., 2023), and replay-mixture design (Ball et al., 2023).

This paper studies the O2O transition from a distribution correction estimation (DICE) perspective, which builds on the classical occupancy-measure view of MDPs (Manne, 1960; Puterman, 2014; Altman, 2021; Nachum & Dai, 2020; Lee et al., 2021). DICE methods use duality and density ratios to estimate occupancy-space updates from off-policy data, giving regularized policy improvement a convex formulation in occupancy space. However, standard DICE formulations are usually applied to settings with a fixed data source, whereas O2O replay keeps growing as new online interaction is added. Replay supplies samples and observed support, but its empirical frequencies need not themselves satisfy Bellman flow.

To address this issue, we propose *SupportDICE*, which makes two changes to the DICE update. First, it closes the current replay-supported region by sending unsupported exits to an absorbing state. This defines a valid occupancy problem on the currently available support. Second, it keeps a trust-region reference from the previous update while giving positive mass to all currently supported state-action

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.

pairs. As new online data enter replay, the next update can therefore use the new support wherever Bellman flow permits, without resetting its reference to the raw replay distribution.

On a fixed offline support, the exact update retains the usual DICE improvement property in the augmented MDP (Theorem 2). When online interaction expands replay support, SupportDICE obtains a per-step comparator guarantee against occupancies feasible on the current support (Theorem 3). The augmented flow constraints also change, so successive feasible sets need not be nested. An extension to dominated standard Borel spaces gives a corresponding continuous-space population formulation.

In summary, our contributions are listed as follows.

- We formulate SupportDICE, a replay-supported DICE update that separates replay as a changing sample source from the occupancy being optimized.
- We construct an augmented replay-supported MDP and an anchored update, establish finite-MDP duality and recovery conditions, and derive fixed-support improvement and current-support comparator guarantees, together with a dominated continuous-space extension.
- We validate the reference-update mechanism with an exact tabular square-grid example and evaluate SupportDICE on D4RL offline-to-online benchmarks.

2 Related work

Occupancy-measure RL and convex formulations. The occupancy-measure view of RL goes back to the linear-program formulations of MDPs and constrained MDPs (Manne, 1960; Puterman, 2014; Altman, 2021). In this view, policy optimization is expressed through visitation measures subject to Bellman-flow constraints; convex duality connects the resulting program to value-function objectives (Nachum & Dai, 2020). Our work is closest in spirit to this line, but studies a changing occupancy feasible set induced by O2O replay support. Continuous-MDP occupancy characterizations and disintegration results provide the geometric foundation for our dominated extension (Laroche & Des Combes, 2023; Jhaveri et al., 2026).

The DICE family and occupancy-space dual methods. DICE recasts distribution correction and policy optimization as convex dual problems over occupancies (Nachum et al., 2019a; Zhang et al., 2020a;b; Nachum et al., 2019b; Nachum & Dai, 2020). The line has expanded to direct policy optimization and constrained control through OptiDICE and COptiDICE (Lee et al., 2021; 2022), and to stronger practical optimization and generative policy extraction through ODICE and Diffusion-DICE (Mao et al., 2024b;a). COptiDICE's reduced-CMDP view under limited dataset support is the closest DICE precedent for restricting the feasible problem to observed support. Our augmented construction extends this support-restricted view to O2O: both the allowed coordinates and the augmented flow constraints change as replay support arrives. ODICE and recent semi-gradient analyses address the practical value optimization used in this setting (Mao et al., 2024b; Kim et al., 2025).

Offline-to-online and hybrid RL. Widely used O2O recipes rely on conservative critics, behavior regularization, or replay balancing (Nakamoto et al., 2023; Nair et al., 2020; Kostrikov et al., 2021; Kumar et al., 2020; Ball et al., 2023). Other methods make the transition mechanism explicit through support constraints, policy expansion, or unified objectives (Wu et al., 2022; Zhang et al., 2023; Lei et al., 2023; Xu et al., 2026). In parallel, hybrid-RL theory establishes statistical benefits from combining offline and online data under data-mixture and function-class assumptions (Song et al., 2022; Tan et al., 2024). Our method addresses a complementary question: how to define the occupancy-space update when replay support changes. In this sense, the framework complements value regularization, policy regularization, and replay engineering.

3 Preliminaries

We use the standard finite-MDP occupancy notation for the main theoretical development. Appendix A gives the finite derivations, and Appendix B gives a conditional continuous-space population extension under explicit domination and finite-divergence assumptions.

Discounted MDP and occupancy measures. Let $M = \langle \mathcal{S}, \mathcal{A}, P, r, \mu_0, \gamma \rangle$ be a discounted MDP with finite nonempty $\mathcal{S}$ and $\mathcal{A}$, bounded reward $|r| \le R_{\max}$, initial distribution μ_0, and $\gamma \in (0,1)$. For a Markov policy π, define the normalized discounted occupancy

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

so that $J(\pi) = \mathbb{E}_\pi[\sum_{t\ge 0} \gamma^t r(s_t,a_t)] = (1-\gamma)^{-1}\langle r, d^\pi \rangle$. The occupancy satisfies the Bellman-flow equations

$$\forall s \in \mathcal{S}, \quad \sum_a d(s,a) = (1-\gamma)\,\mu_0(s) + \gamma \sum_{\bar{s},\bar{a}} P(s \mid \bar{s},\bar{a})\, d(\bar{s},\bar{a}), \tag{2}$$

and, conversely, any nonnegative d satisfying (2) is the occupancy of a Markov policy obtained by disintegration $\pi_d(a|s) = d(s,a)/\sum_{a'} d(s,a')$ on states with positive marginal mass (Manne, 1960; Puterman, 2014; Altman, 2021). Summing (2) over states gives $\sum_{s,a} d(s,a) = 1$, so the feasible occupancies form the convex polytope $\mathcal{D} \doteq \{d \ge 0 : d \text{ satisfies (2)}\}$.

f-divergences and DICE duality. Let $f : [0,\infty) \to (-\infty, \infty]$ be proper, closed, convex, and normalized by $f(1) = 0$. We extend $f(u) = +\infty$ for $u < 0$, so its conjugate is the restricted Fenchel conjugate $f^*(y) = \sup_{u \ge 0}\{uy - f(u)\}$. For nonnegative measures d, q, write $D_f(d\|q) = \sum_x q(x) f(d(x)/q(x))$. We use the support-constrained convention $0f(0/0) = 0$ and set $D_f(d\|q) = +\infty$ if $q(x) = 0 < d(x)$ for some x. For the shifted-KL and Pearson generators used below, whose recession slopes are infinite, this agrees with the usual extended Csiszár convention.

Given a reference probability measure q that dominates the candidate occupancies, a DICE-style population update has the form

$$\max_{d \in \mathcal{D}} \langle r, d \rangle - \frac{1}{\eta} D_f(d \,\|\, q), \tag{3}$$

where $\eta > 0$ controls the regularization strength. For the finite discounted programs studied below, the existence of one finite-divergence feasible occupancy is enough for zero duality gap; the proof uses discounted-flow compactness (Appendix A.4). Applying Lagrangian duality to the Bellman-flow constraints then yields a convex minimization over a potential $v : \mathcal{S} \to \mathbb{R}$:

$$\inf_v\ (1-\gamma)\, \mathbb{E}_{s\sim\mu_0}[v(s)] + \frac{1}{\eta}\, \mathbb{E}_{(s,a)\sim q}\left[f^*\left(\eta \bar{A}_v(s,a)\right)\right], \tag{4}$$

where $\bar{A}_v(s,a) = r(s,a) + \gamma\, \mathbb{E}_{s'\sim P(\cdot|s,a)}[v(s')] - v(s)$ is the Bellman residual. Section 4.4 specializes this template to the replay-supported augmented MDP. The optimized object in (3) is an occupancy satisfying Bellman flow; the reference measure supplies domination and expectations. The O2O issue is that replay changes both the sampling support and the set of plausible occupancies over time, so the occupancy feasible set must be rebuilt as the buffer grows.

4 SUPPORTDICE

SupportDICE turns a changing replay buffer into a sequence of finite occupancy optimization problems. The construction has three steps. First, the replay buffer determines the currently supported environment state-action pairs (Section 4.1). Second, we close this partial support under Bellman flow by building an augmented MDP with an absorbing state (Section 4.2). Third, we run a DICE-style f-divergence-regularized occupancy update on the resulting replay-supported occupancy feasible set, using a full-support base measure so that newly observed coordinates can enter the finite-divergence domain whenever Bellman flow permits them to receive mass (Sections 4.3 and 4.4).

4.1 REPLAY BUFFER AND SUPPORT

At iteration k, the replay buffer induces the environment state-action support $\Omega_k^{\text{env}} \subseteq \mathcal{S} \times \mathcal{A}$. The buffer supplies samples for Monte Carlo estimation; Ω_k^{env} supplies the coordinates on which the next occupancy problem is allowed to place mass. In the population analysis, support is cumulative: $\Omega_{k+1}^{\text{env}} \supseteq \Omega_k^{\text{env}}$.

4.2 CLOSING REPLAY SUPPORT UNDER BELLMAN FLOW

The support Ω_k^{env} need not be closed under the environment dynamics: a supported action may transition to a state that has not yet appeared in replay. To obtain a Bellman-flow feasible control problem on the current support, define the replay-covered state set

$$\mathcal{S}_k^{\text{env}} \doteq \{s : \exists a,\ (s,a) \in \Omega_k^{\text{env}}\},$$

and the augmented state set $\mathcal{S}_k^{\text{aug}} \doteq \mathcal{S}_k^{\text{env}} \cup \{\perp\}$, where $\perp$ is absorbing with zero reward. The augmented initial distribution is $\mu_{0,k}^{\text{aug}}(s) = \mu_0(s)$ for $s \in \mathcal{S}_k^{\text{env}}$ and $\mu_{0,k}^{\text{aug}}(\perp) = \mu_0(\mathcal{S} \setminus \mathcal{S}_k^{\text{env}})$. Attempts to leave the current replay-covered state set are redirected to $\perp$:

$$P_k^{\text{aug}}(s' \mid s,a) = P(s' \mid s,a), \quad s' \in \mathcal{S}_k^{\text{env}}, \qquad P_k^{\text{aug}}(\perp \mid s,a) = P(\mathcal{S} \setminus \mathcal{S}_k^{\text{env}} \mid s,a), \tag{5}$$

for $s \in \mathcal{S}_k^{\text{env}}$, with $P_k^{\text{aug}}(\perp|\perp,a) = 1$. Rewards are $r_k^{\text{aug}}(s,a) = r(s,a)$ on $\mathcal{S}_k^{\text{env}}$ and $r_k^{\text{aug}}(\perp,a) = 0$.

The allowed augmented state-action support is

$$\Omega_k^{\text{aug}} \doteq \Omega_k^{\text{env}} \cup (\{\perp\} \times \mathcal{A}).$$

The replay-supported occupancy feasible set is

$$\mathcal{D}_k^{\text{aug}} \doteq \left\{ d \geq 0 : \text{supp}(d) \subseteq \Omega_k^{\text{aug}},\ d \text{ satisfies Bellman flow with } P_k^{\text{aug}} \text{ and } \mu_{0,k}^{\text{aug}} \right\}. \tag{6}$$

Thus $\mathcal{D}_k^{\text{aug}}$ contains exactly the Bellman-flow feasible occupancies that use replay-supported environment actions and synthetic absorbing-state actions. Since $\mathcal{S}_k^{\text{env}}$ is the state projection of Ω_k^{env}, every covered state has at least one allowed action, and $\mathcal{D}_k^{\text{aug}}$ is nonempty in finite spaces. The augmented kernel can be evaluated for all actions, but unsupported environment actions are outside Ω_k^{aug} and every feasible occupancy assigns them zero mass. For every $d \in \mathcal{D}_k^{\text{aug}}$, let π_d be its Markov-policy disintegration. The finite-MDP occupancy converse applied to M_k^{aug} gives

$$\langle r_k^{\text{aug}}, d \rangle = (1-\gamma) J_k^{\text{aug}}(\pi_d). \tag{7}$$

The augmented return differs from the original return only through trajectories that leave the replay-covered state set. For an arbitrary policy π, $J_k^{\text{aug}}(\pi)$ denotes its return under these augmented dynamics, using π on $\mathcal{S}_k^{\text{env}}$ and any fixed extension at $\perp$:

Lemma 1 (Absolute augmentation mismatch)**.** *Assume $|r(s,a)| \leq R_{\max}$. For any policy π, let $\tau_k = \inf\{t : s_t \notin \mathcal{S}_k^{\text{env}}\}$. Then*

$$\left|J(\pi) - J_k^{\text{aug}}(\pi)\right| \leq \frac{R_{\max}}{1-\gamma} \mathbb{E}_\pi[\gamma^{\tau_k} \mathbf{1}\{\tau_k < \infty\}] \leq \frac{R_{\max}}{1-\gamma} \sum_{t \geq 0} \gamma^t \Pr_\pi(s_t \notin \mathcal{S}_k^{\text{env}}). \tag{8}$$

Define $\Delta_k(\pi) \doteq \frac{R_{\max}}{1-\gamma} \mathbb{E}_\pi[\gamma^{\tau_k} \mathbf{1}\{\tau_k < \infty\}]$. Thus, for a fixed policy, the mismatch is at most $\Delta_k(\pi)$. Since replay support is cumulative, $\tau_{k+1} \geq \tau_k$ pathwise and hence $\Delta_{k+1}(\pi) \leq \Delta_k(\pi)$. This is monotonicity of the certified upper bound; the signed augmented-return error itself need not be monotone.

4.3 ANCHORED UPDATES ON THE SUPPORTED FEASIBLE SET

The augmented MDP identifies which occupancies are feasible on the current support. The remaining issue is support trapping: a DICE update regularized against a carried reference can assign finite-divergence mass only where that reference has positive mass. After support expansion, a zero-extended previous reference assigns zero to newly observed coordinates, so those coordinates would remain unavailable without a new dominating measure. Let m_k denote the full-support base measure used for expectations in the dual. In the finite population theory we use

$$m_k = \text{Unif}(\Omega_k^{\text{aug}}), \tag{9}$$

viewed as a probability measure on Ω_k^{aug}. Any full-support choice is valid, with the ratio ρ_k adjusted accordingly. The base measure m_k supplies expectations and domination. Let $\bar{q}_k \in \Delta(\Omega_k^{\text{aug}})$ be any carried-over probability reference on the current augmented support; under support expansion, $\bar{q}_k$

need not satisfy the current Bellman-flow constraints. Operationally, when the support expands from Ω_k^{aug} to $\Omega_{k+1}^{\mathrm{aug}}$, the carried reference is extended by assigning zero mass to the new coordinates; the next base measure then supplies positive mass on the enlarged augmented support through m_{k+1}. If the recovered object is approximate, this zero-extension is normalized before being carried as the next reference. For the O2O update, define the anchored reference

$$q_k = (1-\varepsilon)\,\bar{q}_k + \varepsilon\, m_k, \qquad \varepsilon \in (0,1]. \tag{10}$$

Since m_k has full support on Ω_k^{aug} and $q_k \geq \varepsilon m_k$, the finite ratio

$$\rho_k(x) \doteq \frac{q_k(x)}{m_k(x)}, \qquad x \in \Omega_k^{\mathrm{aug}}, \tag{11}$$

is well-defined. The exact population update is

$$d_{k+1} \in \arg\max_{d \in \mathcal{D}_k^{\mathrm{aug}}} \Big\{ \langle r_k^{\mathrm{aug}}, d\rangle - \tfrac{1}{\eta}\, D_f(d \,\|\, q_k) \Big\}, \tag{P_k}$$

where $\eta > 0$. This is an f-divergence-regularized occupancy step on the current replay-supported feasible set; the anchor removes the divergence barrier on newly covered coordinates, while Bellman flow still determines whether they can receive positive occupancy.

4.4 Base-measure DICE dual

With the feasible set and anchored reference in place, the DICE dual is the standard Bellman-flow Lagrangian dual, written with expectations under the base measure m_k. For any potential $v : \mathcal{S}_k^{\mathrm{aug}} \to \mathbb{R}$, define

$$\bar{A}_v(s,a) = r_k^{\mathrm{aug}}(s,a) + \gamma\, \mathbb{E}_{s' \sim P_k^{\mathrm{aug}}(\cdot|s,a)}[v(s')] - v(s). \tag{12}$$

Theorem 1 (Strong duality; base-measure dual form)**.** *Assume that some $d^0 \in \mathcal{D}_k^{\mathrm{aug}}$ satisfies $D_f(d^0 \| q_k) < \infty$ (Assumption F_k'). Then (P_k) has a finite attained optimum and zero duality gap with the following dual infimum:*

$$\inf_v\ (1-\gamma)\, \mathbb{E}_{s \sim \mu_{0,k}^{\mathrm{aug}}}[v(s)] + \frac{1}{\eta}\, \mathbb{E}_{(s,a)\sim m_k}\left[\rho_k(s,a)\, f^*\big(\eta\, \bar{A}_v(s,a)\big)\right], \tag{V_k}$$

For shifted KL and Pearson, F_k' holds automatically because f is finite on $[0,\infty)$ and $\mathcal{D}_k^{\mathrm{aug}}$ is nonempty. Generators with a smaller effective domain require a separate finite-feasibility check. Dualizing Bellman flow and applying the conjugate of f gives the objective in (V_k), with $\mathbb{E}_{q_k}[\cdot] = \mathbb{E}_{m_k}[\rho_k\, \cdot]$. Appendix A.4 establishes zero duality gap through discounted-flow closedness.

Continuous-space extension. Appendix B extends the population program to standard Borel spaces under explicit domination, superlinearity, and finite-divergence feasibility. It establishes primal attainment and zero duality gap; recovery holds when a bounded dual minimizer exists.

4.5 Fenchel recovery and policy extraction

Under Assumption F_k' and Attain-k in Appendix A.4 (the latter is automatic for either Pearson convention), every primal–dual optimal pair obeys the pointwise Fenchel/KKT characterization below. Let v_k^* be a dual minimizer and write $z_{k+1}(x) = d_{k+1}(x)/q_k(x)$ on the support of q_k. Pointwise Fenchel optimality gives

$$z_{k+1}(s,a) \in \arg\max_{z \geq 0} \{ z\, \eta \bar{A}_{v_k^*}(s,a) - f(z) \}, \qquad d_{k+1} = q_k \odot z_{k+1}, \tag{13}$$

Equivalently, $\eta \bar{A}_{v_k^*}(s,a) \in \partial f(z_{k+1}(s,a))$. If every scalar argmax in (13) is a singleton, this characterization determines d_{k+1} from v_k^* alone; otherwise an arbitrary pointwise selection need not satisfy Bellman flow. For differentiable, strictly convex f, an interior maximizer satisfies $z = (f')^{-1}(\eta \bar{A}_{v_k^*}(s,a))$. For example, shifted KL $f(z) = z\log z - z + 1$ gives $z = \exp(\eta \bar{A}_{v_k^*}(s,a))$, while Pearson $f(z) = \frac{1}{2}(z-1)^2$ gives $z = \max\{0, 1 + \eta \bar{A}_{v_k^*}(s,a)\}$. Both generators are strictly convex, so their scalar maximizers and finite-divergence primal optimizers are unique. Appendix A.4

gives the boundary cases and proves finite dual attainment for both Pearson conventions in the finite program.

Given a feasible occupancy d_{k+1}, the exact induced policy is

$$\pi_{k+1}(a \mid s) = \frac{d_{k+1}(s,a)}{\sum_{a'} d_{k+1}(s,a')}, \tag{14}$$

on states with positive marginal mass, with arbitrary actions elsewhere. The corresponding population extraction weight under m_k is $d_{k+1}/m_k = \rho_k z_{k+1}$. This identifies the occupancy-weighted fitting target independently of the actor parameterization. The practical realization is specified in Appendix C.3.

4.6 Theoretical guarantees

The following guarantees are finite-dimensional primal optimality statements; full derivations are in the appendix.

Theorem 2 (Fixed-support augmented-return DICE improvement)**.** *Consider the fixed-support offline phase, where the replay-supported augmented MDP is fixed, the reference is centered at the current feasible occupancy $q_k = d_k$, and $d_k \in \mathcal{D}_k^{\mathrm{aug}}$. Let d_{k+1} solve the support-conservative update*

$$d_{k+1} \in \arg\max_{d \in \mathcal{D}_k^{\mathrm{aug}},\, d \ll d_k} \left\{ \langle r^{\mathrm{aug}}, d \rangle - \frac{1}{\eta} D_f(d \| d_k) \right\}.$$

Then

$$\langle r^{\mathrm{aug}}, d_{k+1} \rangle \geq \langle r^{\mathrm{aug}}, d_k \rangle + \tfrac{1}{\eta}\, D_f(d_{k+1} \,\|\, d_k). \tag{15}$$

Because $D_f(d_{k+1}\|d_k) \geq 0$, this establishes monotone improvement of the augmented return through (7). The online phase uses the anchored reference to incorporate newly observed flow-feasible coordinates, with the comparator guarantee below.

Theorem 3 (Per-step comparator bound)**.** *Let d_{k+1} solve* (P_k)*. For any comparator $d \in \mathcal{D}_k^{\mathrm{aug}}$ with $D_f(d\|q_k) < \infty$,*

$$\langle r_k^{\mathrm{aug}}, d_{k+1} \rangle \geq \langle r_k^{\mathrm{aug}}, d \rangle - \tfrac{1}{\eta}\, D_f(d \,\|\, q_k). \tag{16}$$

Summing this inequality bounds the cumulative comparison error against feasible finite-divergence comparators on the current supports. The bound depends on their divergences to the anchored references. For example, with shifted KL and $m_k = \mathrm{Unif}(\Omega_k^{\mathrm{aug}})$, the anchor implies $D_{\mathrm{KL}}(d\|q_k) \leq \log(|\Omega_k^{\mathrm{aug}}|/\varepsilon)$ for every feasible comparator $d \in \mathcal{D}_k^{\mathrm{aug}}$.

Original-MDP consequences. Let π_k, π_{k+1} be the policies induced by d_k, d_{k+1} in the fixed-support theorem, and let π_d be induced by a comparator d. Choose arbitrary but fixed extensions of these policies outside $\mathcal{S}_k^{\mathrm{env}}$, and use the same extensions in J and Δ_k. Combining the two theorems with Lemma 1 and (7) gives

$$J(\pi_{k+1}) \geq J(\pi_k) + \frac{D_f(d_{k+1}\|d_k)}{\eta(1-\gamma)} - \Delta_k(\pi_{k+1}) - \Delta_k(\pi_k), \tag{17}$$

$$J(\pi_{k+1}) \geq J(\pi_d) - \frac{D_f(d\|q_k)}{\eta(1-\gamma)} - \Delta_k(\pi_{k+1}) - \Delta_k(\pi_d). \tag{18}$$

4.7 Practical realization

The practical implementation parameterizes the DICE value function and trains it with replay minibatches. For policy learning, we use the FQL architecture (Park et al., 2025). The replay buffer retains all offline and online transitions and is sampled uniformly. Online fine-tuning uses the anchored reference with $\varepsilon = 0.3$: carried probability masses are zero-extended when new transitions arrive, and a periodic refresh updates the normalized weights. Appendix C.3 gives the policy losses, weight processing, and update schedule.

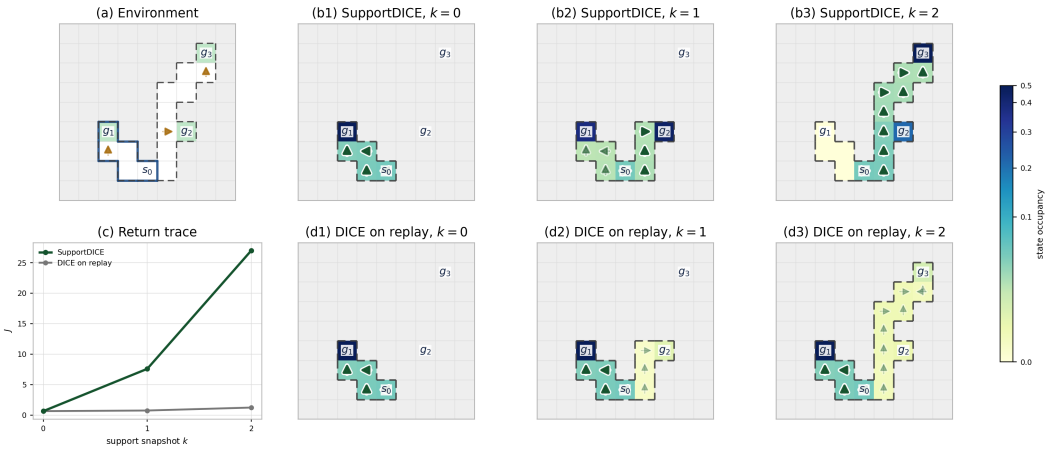


Figure 1: Square-grid finite-MDP trace. Panel (a) shows the environment and cumulative support sequence. Panels (b1)–(b3) show SupportDICE over successive support snapshots: it carries the previous occupancy as its reference and uses newly supported branches. Panel (c) summarizes the exact returns. Panels (d1)–(d3) show replay-centered DICE, which remains pulled toward the old low-reward branch. Darker shading denotes greater state occupancy; dark paths are greedy policy traces.

5 Experiments

We use the experiments for two complementary checks. First, a finite tabular example isolates the replay-support mechanism in the same exact population setting as the theory. We then evaluate SupportDICE on D4RL benchmarks.

5.1 Square-grid toy example

We use a deterministic 9×9 GridWorld to isolate how a replay-supported occupancy update uses newly observed support in an offline-heavy buffer. The MDP has discount 0.95 and three terminal goals with increasing rewards. We condition on a cumulative support path that initially covers the low-reward branch and later reveals the two better branches. At each snapshot, both methods solve the same augmented occupancy problem exactly. Replay-centered DICE resets its reference to the offline-heavy empirical replay distribution; SupportDICE carries its previous occupancy and mixes in the uniform base with $\varepsilon = 0.3$.

Figure 1 compares the two exact occupancy traces. SupportDICE moves from the low-reward branch to the middle branch and then to the high-reward branch as those branches enter support. Replay-centered DICE, even when rerun on the same expanded augmented problem, remains tied to the low-reward replay branch. Thus support expansion by itself is not enough: the reference must also let newly covered coordinates influence the next solution. Appendix C.1 gives the environment, solver, and exact returns.

5.2 D4RL benchmarks

We evaluate SupportDICE on continuous-control O2O benchmarks from D4RL, using the CORL evaluation pipeline (Tarasov et al., 2023). The suite covers AntMaze navigation, where an ant must reach sparse goals in mazes of increasing size; Adroit cloned manipulation, where a dexterous hand learns from cloned demonstrations; and Kitchen manipulation, where a policy must compose multiple subtasks in one rollout.

Baselines are Cal-QL, AWAC, IQL, CQL, and SPOT from CORL, together with RLPD (Ball et al., 2023) and PEX (Zhang et al., 2023). AntMaze and Adroit use one million online transitions, while Kitchen uses 1.25 million. Appendix C.6 gives the full mean and sample standard deviation;

Table 1: Final online normalized scores (four-seed means). Bold marks the highest and underline the second distinct reported mean in each row; ties share formatting. Sample standard deviations appear in Appendix C.6; implementation details appear in Appendix C.

Dataset	Support DICE	Cal-QL	AWAC	IQL	CQL	SPOT	RLPD	PEX
AntMaze								
umaze	**100.00**	99.75	98.75	96.50	99.50	99.50	98.00	95.50
umaze-diverse	**100.00**	98.50	0.00	63.75	99.00	95.00	97.00	63.00
medium-play	96.00	**98.75**	0.00	89.75	97.75	97.25	96.50	85.00
medium-diverse	96.00	98.25	0.00	92.25	97.25	94.50	**98.50**	87.00
large-play	96.50	**97.25**	0.00	64.50	88.25	87.00	85.50	61.00
large-diverse	**94.50**	91.50	0.00	64.25	91.75	81.00	84.00	73.00
Mean	97.17	**97.33**	16.46	78.50	95.58	92.38	93.25	77.42
Adroit								
pen	**147.32**	−2.68	86.82	102.02	−1.28	43.63	119.28	94.81
door	**99.50**	−0.33	0.01	20.34	−0.33	0.02	81.12	65.06
hammer	136.45	0.17	0.24	57.27	2.85	3.73	**137.87**	64.07
relocate	**73.79**	−0.31	−0.04	0.32	−0.33	−0.15	−0.04	−0.01
Mean	**114.27**	−0.79	21.76	44.99	0.23	11.81	84.56	55.98
Kitchen								
partial	**91.88**	8.33	0.00	75.00	0.00	24.38	65.00	54.38
mixed	66.25	8.33	**75.00**	65.00	0.83	15.00	73.75	59.38
complete	83.75	0.00	28.33	23.33	0.00	7.50	**90.63**	21.88
Mean	**80.63**	5.55	34.44	54.44	0.28	15.63	76.46	45.21

Appendix C.2 describes the benchmark protocols, including the shared Kitchen reward and evaluation protocol.

5.3 Results

Table 1 reports the final online evaluation averaged over four seeds on thirteen tasks.

On AntMaze, SupportDICE achieves a mean score of 97.17 across the six datasets, closely matching the best average performance of Cal-QL (97.33). SupportDICE obtains the highest score on three datasets, including perfect scores of 100.00 on both umaze and umaze-diverse, and the best score of 94.50 on `large-diverse`. It also consistently achieves scores above 96.00 on the remaining medium and large maze tasks.

On Adroit cloned tasks, SupportDICE achieves the highest average score of 114.27, substantially outperforming the second-best RLPD score of 84.56. It obtains the best performance on three of the four tasks, achieving 147.32 on `pen`, 99.50 on `door`, and 73.79 on `relocate`. On hammer, SupportDICE also remains close to the best-performing RLPD (136.45 vs. 137.87).

On Kitchen, SupportDICE achieves the highest average score of 80.63, compared with 76.46 for the second-best RLPD. In particular, SupportDICE substantially outperforms the other methods on `partial`, achieving 91.88 compared with 75.00 for the next-best method. It also achieves 66.25 and 83.75 on `mixed` and `complete`, respectively.

Figure 2 complements Table 1 by showing the learning curves during online fine-tuning. On `relocate`, all methods start near zero, while SupportDICE steadily improves to around 70, whereas the compared baselines remain near zero throughout training. On `pen` and `door`, SupportDICE also improves substantially faster and maintains a clear advantage during most of fine-tuning. On Kitchen `partial` and `complete`, SupportDICE also reaches high scores earlier than RLPD.

Overall, these learning curves show that the improvements of SupportDICE are not limited to final performance, but are also reflected in the faster online fine-tuning on several tasks. Together with the tabular example, which isolates the role of the replay-supported occupancy update, the D4RL

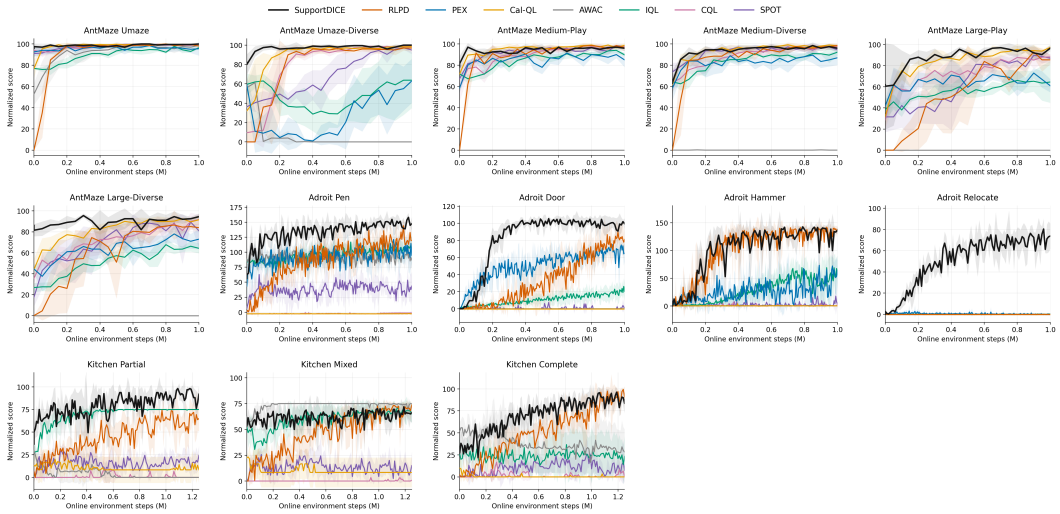


Figure 2: Online fine-tuning curves on the thirteen D4RL tasks in Table 1. Lines show the mean over four seeds; shaded regions show $\pm$ one sample standard deviation. SupportDICE makes substantial progress on `relocate`, where the compared baselines remain near zero.

results show that SupportDICE can effectively incorporate newly collected support without resetting the optimization reference to raw replay frequencies.

6 Conclusion

We introduced SupportDICE, an occupancy-space formulation of offline-to-online fine-tuning that keeps replay as a sampling and support mechanism while optimizing a Bellman-flow feasible occupancy. The key construction is an augmented replay-supported MDP: transitions that leave the current replay-covered state set are redirected to an absorbing state, and the occupancy problem is rebuilt as online data arrive. This yields a base-measure DICE dual, a constrained Fenchel recovery map, fixed-support improvement in the augmented MDP, and per-step guarantees over the current replay-supported feasible set. A dominated continuous-space extension gives corresponding population results.

The square-grid example isolates the reference mechanism in an exact finite MDP: the carried occupancy and full-support anchor let the update use newly observed branches, while replay-centered DICE remains tied to the old replay distribution. On D4RL benchmarks, SupportDICE is competitive on AntMaze and achieves the highest average final scores on Adroit and Kitchen among the evaluated methods. Future work should combine the replay-supported occupancy update with directed exploration, so that the method not only uses new replay support but also helps decide which support should be collected next.

Reproducibility Statement

Appendices A and B contain the assumptions and proofs. Appendix C provides implementation details, evaluation protocols, and full results.

AI Use Statement

We used generative AI tools to assist with checking theoretical statements and proofs, interpreting experimental results, and revising the manuscript. The authors retain full responsibility for the final content.

REFERENCES

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

Philip J Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning*, pp. 1577–1594. PMLR, 2023.

Yash Jhaveri, Harley Wiltzer, Patrick Shafto, Marc Bellemare, and David Meger. Convergence theorems for entropy-regularized and distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 38:30202–30246, 2026.

Woosung Kim, JunHo Seo, Jongmin Lee, and Byung-Jun Lee. Semi-gradient dice for offline constrained reinforcement learning. *arXiv preprint arXiv:2506.08644*, 2025.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.

Romain Laroche and Remi Tachet Des Combes. On the occupancy measure of non-markovian policies in continuous mdps. In *International Conference on Machine Learning*, pp. 18548–18562. PMLR, 2023.

Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pp. 6120–6130. PMLR, 2021.

Jongmin Lee, Cosmin Paduraru, Daniel J Mankowitz, Nicolas Heess, Doina Precup, Kee-Eung Kim, and Arthur Guez. Coptidice: Offline constrained reinforcement learning via stationary distribution correction estimation. *arXiv preprint arXiv:2204.08957*, 2022.

Kun Lei, Zhengmao He, Chenhao Lu, Kaizhe Hu, Yang Gao, and Huazhe Xu. Uni-o4: Unifying online and offline deep reinforcement learning with multi-step on-policy optimization. *arXiv preprint arXiv:2311.03351*, 2023.

Alan S Manne. Linear programming and sequential decisions. *Management Science*, 6(3):259–267, 1960.

Liyuan Mao, Haoran Xu, Xianyuan Zhan, Weinan Zhang, and Amy Zhang. Diffusion-dice: In-sample diffusion guidance for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 37:98806–98834, 2024a.

Liyuan Mao, Haoran Xu, Weinan Zhang, and Xianyuan Zhan. Odice: Revealing the mystery of distribution correction estimation via orthogonal-gradient update. *arXiv preprint arXiv:2402.00348*, 2024b.

Ofir Nachum and Bo Dai. Reinforcement learning via fenchel-rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in neural information processing systems*, 32, 2019a.

Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019b.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

Mitsuhiko Nakamoto, Simon Zhai, Anikait Singh, Max Sobol Mark, Yi Ma, Chelsea Finn, Aviral Kumar, and Sergey Levine. Cal-ql: Calibrated offline rl pre-training for efficient online fine-tuning. *Advances in Neural Information Processing Systems*, 36:62244–62269, 2023.

Seohong Park, Qiyang Li, and Sergey Levine. Flow q-learning. *arXiv preprint arXiv:2502.02538*, 2025.

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

R Tyrrell Rockafellar. Integral functionals, normal integrands and measurable selections. In *Nonlinear Operators and the Calculus of Variations: Summer School Held in Bruxelles 8–19 September 1975*, pp. 157–207. Springer, 2006.

Ralph Rockafellar. Integrals which are convex functionals. *Pacific journal of mathematics*, 24(3): 525–539, 1968.

Yuda Song, Yifei Zhou, Ayush Sekhari, J Andrew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid rl: Using both offline and online data can make rl efficient. *arXiv preprint arXiv:2210.06718*, 2022.

Kevin Tan, Wei Fan, and Yuting Wei. Hybrid reinforcement learning breaks sample size barriers in linear mdps. *Advances in Neural Information Processing Systems*, 37:120038–120077, 2024.

Denis Tarasov, Alexander Nikulin, Dmitry Akimov, Vladislav Kurenkov, and Sergey Kolesnikov. Corl: Research-oriented deep offline reinforcement learning library. *Advances in Neural Information Processing Systems*, 36:30997–31020, 2023.

Jialong Wu, Haixu Wu, Zihan Qiu, Jianmin Wang, and Mingsheng Long. Supported policy optimization for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 31278–31291, 2022.

Haoran Xu, Liyuan Mao, Hui Jin, Weinan Zhang, Xianyuan Zhan, and Amy Zhang. Uni-rl: Unifying online and offline rl via implicit value regularization. *Advances in Neural Information Processing Systems*, 38:133162–133187, 2026.

Haichao Zhang, We Xu, and Haonan Yu. Policy expansion for bridging offline-to-online reinforcement learning. *arXiv preprint arXiv:2302.00935*, 2023.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020a.

Shangtong Zhang, Bo Liu, and Shimon Whiteson. Gradientdice: Rethinking generalized offline estimation of stationary values. In *International Conference on Machine Learning*, pp. 11194–11203. PMLR, 2020b.

Zhiyuan Zhou, Andy Peng, Qiyang Li, Sergey Levine, and Aviral Kumar. Efficient online reinforcement learning fine-tuning need not retain offline data. In *International Conference on Learning Representations*, volume 2025, pp. 32343–32368, 2025.

A PROOFS AND TECHNICAL DETAILS

This appendix proves the finite-MDP population results using occupancy polytopes, discounted-flow closedness, and Fenchel duality. Appendix B gives the conditional continuous-space population extension, and Appendix C records practical implementation details, normalization, and actor projection.

A.1 FINITE MDP AND OCCUPANCY FACTS

Let $\mathcal{S}$ and $\mathcal{A}$ be finite nonempty sets, and let μ_0 be the initial state distribution. A Markov policy π induces the normalized discounted occupancy

$$d^\pi(s,a) = (1-\gamma)\sum_{t\geq 0}\gamma^t \Pr_\pi(s_t = s, a_t = a). \tag{19}$$

The unnormalized return satisfies

$$J(\pi) = \mathbb{E}_\pi\left[\sum_{t\geq 0}\gamma^t r(s_t,a_t)\right] = \frac{1}{1-\gamma}\langle r, d^\pi\rangle. \tag{20}$$

The occupancy satisfies Bellman flow:

$$\sum_a d^\pi(s,a) = (1-\gamma)\mu_0(s) + \gamma\sum_{\bar{s},\bar{a}} P(s \mid \bar{s},\bar{a})d^\pi(\bar{s},\bar{a}), \qquad \forall s \in \mathcal{S}. \tag{21}$$

Conversely, any nonnegative vector satisfying (21) is the discounted occupancy of a Markov policy. On states with positive marginal mass, the policy is

$$\pi_d(a \mid s) = \frac{d(s,a)}{\sum_{\bar{a}} d(s,\bar{a})}, \tag{22}$$

with arbitrary actions on zero-marginal states. This is the standard finite-MDP occupancy-LP converse used by (Manne, 1960; Puterman, 2014; Altman, 2021). Summing (21) over s gives

$$\sum_{s,a} d(s,a) = (1-\gamma) + \gamma\sum_{s,a} d(s,a), \tag{23}$$

so every feasible discounted occupancy has unit mass.

A.2 REPLAY-SUPPORTED AUGMENTED MDP

At O2O iteration k, let $\Omega_k^{\mathrm{env}} \subseteq \mathcal{S}\times\mathcal{A}$ be the finite replay-supported environment state-action support and let

$$\mathcal{S}_k^{\mathrm{env}} = \{s \in \mathcal{S} : \exists a,\ (s,a) \in \Omega_k^{\mathrm{env}}\}. \tag{24}$$

Define $\mathcal{S}_k^{\mathrm{aug}} = \mathcal{S}_k^{\mathrm{env}} \cup \{\perp\}$, where $\perp$ is absorbing with zero reward. The augmented initial distribution is

$$\mu_{0,k}^{\mathrm{aug}}(s) = \mu_0(s), \quad s \in \mathcal{S}_k^{\mathrm{env}}, \qquad \mu_{0,k}^{\mathrm{aug}}(\perp) = \mu_0(\mathcal{S}\setminus\mathcal{S}_k^{\mathrm{env}}). \tag{25}$$

The augmented kernel keeps transitions inside $\mathcal{S}_k^{\mathrm{env}}$ and redirects all mass leaving $\mathcal{S}_k^{\mathrm{env}}$ to $\perp$:

$$P_k^{\mathrm{aug}}(s' \mid s,a) = P(s' \mid s,a), \quad s' \in \mathcal{S}_k^{\mathrm{env}}, \qquad P_k^{\mathrm{aug}}(\perp \mid s,a) = P(\mathcal{S}\setminus\mathcal{S}_k^{\mathrm{env}} \mid s,a), \tag{26}$$

with $P_k^{\mathrm{aug}}(\perp \mid \perp, a) = 1$. The augmented reward is $r_k^{\mathrm{aug}}(s,a) = r(s,a)$ for $s \in \mathcal{S}_k^{\mathrm{env}}$ and $r_k^{\mathrm{aug}}(\perp,a) = 0$. The zero reward is the terminal-state convention used in the analysis; replacing it by another fixed terminal penalty changes the augmented reward vector but not the support-closure or Bellman-flow construction. This defines

$$M_k^{\mathrm{aug}} = \langle \mathcal{S}_k^{\mathrm{aug}}, \mathcal{A}, P_k^{\mathrm{aug}}, r_k^{\mathrm{aug}}, \mu_{0,k}^{\mathrm{aug}}, \gamma\rangle.$$

The kernel is row-stochastic: for $s \in \mathcal{S}_k^{\mathrm{env}}$,

$$\sum_{s'\in\mathcal{S}_k^{\mathrm{aug}}} P_k^{\mathrm{aug}}(s' \mid s,a) = P(\mathcal{S}_k^{\mathrm{env}} \mid s,a) + P(\mathcal{S}\setminus\mathcal{S}_k^{\mathrm{env}} \mid s,a) = 1,$$

and the absorbing row at $\perp$ also sums to one.

Let

$$\Omega_k^{\mathrm{aug}} = \Omega_k^{\mathrm{env}} \cup (\{\perp\} \times \mathcal{A}) \tag{27}$$

be the allowed augmented state-action support. We write $\mathcal{D}_k^{\mathrm{aug}}$ for the set of nonnegative vectors supported on Ω_k^{aug} satisfying augmented flow:

$$\sum_{a\in\mathcal{A}} d(s,a) = (1-\gamma)\mu_{0,k}^{\mathrm{aug}}(s) + \gamma \sum_{\bar{s},\bar{a}} P_k^{\mathrm{aug}}(s \mid \bar{s},\bar{a}) d(\bar{s},\bar{a}), \quad \forall s \in \mathcal{S}_k^{\mathrm{aug}}. \tag{28}$$

This is the replay-supported occupancy feasible set. Since $\mathcal{S}_k^{\mathrm{env}}$ is the state projection of Ω_k^{env}, every $s \in \mathcal{S}_k^{\mathrm{env}}$ has at least one allowed action by construction. Therefore $\mathcal{D}_k^{\mathrm{aug}}$ is nonempty: choose any Markov policy supported on allowed actions in $\mathcal{S}_k^{\mathrm{env}}$ and arbitrary at $\perp$; its augmented discounted occupancy lies in $\mathcal{D}_k^{\mathrm{aug}}$, since it assigns positive mass only to allowed pairs in Ω_k^{env} and to synthetic pairs $(\perp, a) \in \Omega_k^{\mathrm{aug}}$. This support-restricted view is in the same spirit as COptiDICE's reduced-CMDP interpretation under limited dataset support; the absorbing state makes support leakage explicit and yields a sequence of current-support occupancy polytopes. These polytopes need not be nested because P_k^{aug} changes with coverage (Lee et al., 2022). For any $d \in \mathcal{D}_k^{\mathrm{aug}}$, the total discounted mass at $\perp$ is determined by initial mass outside $\mathcal{S}_k^{\mathrm{env}}$ and leakage from supported pairs:

$$\sum_a d(\perp, a) = \frac{(1-\gamma)\mu_{0,k}^{\mathrm{aug}}(\perp) + \gamma \sum_{(\bar{s},\bar{a})\in\Omega_k^{\mathrm{env}}} P_k^{\mathrm{aug}}(\perp \mid \bar{s},\bar{a}) d(\bar{s},\bar{a})}{1-\gamma}.$$

Throughout the exact population analysis, Ω_k^{env} denotes the cumulative support observed up to iteration k. For every $d \in \mathcal{D}_k^{\mathrm{aug}}$, the occupancy converse applied to M_k^{aug} shows that its disintegration π_d has augmented occupancy d, and hence

$$\langle r_k^{\mathrm{aug}}, d\rangle = (1-\gamma) J_k^{\mathrm{aug}}(\pi_d). \tag{29}$$

A.3 Base measure, anchor, and primal update

Let m_k be a population base probability measure with full support on Ω_k^{aug}. We use the simple convention

$$m_k = \mathrm{Unif}(\Omega_k^{\mathrm{aug}}), \tag{30}$$

viewed as a measure on Ω_k^{aug}. This is only a convention; any probability measure with positive mass on every element of Ω_k^{aug} can be used, with ρ_k changed accordingly. The replay buffer defines support and supplies samples. In the exact finite program, Ω_k^{aug} consists of aggregated state–action coordinates, not independently controlled duplicate replay indices. If mini-batches are drawn from a law h_k different from m_k, expectations under q_k require the weight q_k/h_k (assuming domination); equivalently, one may choose $m_k = h_k$. The absorbing-state coordinates are synthetic bookkeeping terms and can be enumerated or sampled separately. The base measure is not assumed to satisfy Bellman flow and is not assumed to be an occupancy. Let $f : [0,\infty) \to (-\infty, \infty]$ be proper, closed, convex, and normalized by $f(1) = 0$. We extend $f(u) = +\infty$ for $u < 0$, so

$$f^*(y) = \sup_{u\geq 0}\{uy - f(u)\} \tag{31}$$

is the restricted conjugate over nonnegative ratios.

For a nonnegative vector d and probability vector q, define

$$D_f(d\|q) = \sum_{x\in\Omega_k^{\mathrm{aug}}} q(x) f\left(\frac{d(x)}{q(x)}\right), \qquad x = (s,a). \tag{32}$$

We use the support-constrained convention: coordinates with $d(x) = q(x) = 0$ contribute $0f(0/0) = 0$, while $q(x) = 0 < d(x)$ makes $D_f(d\|q) = +\infty$. For shifted KL and Pearson, whose recession slopes are infinite, this agrees with the usual extended Csiszár convention. Let $\bar{q}_k \in \Delta(\Omega_k^{\mathrm{aug}})$ be any carried-over probability reference on the current augmented support. It need not satisfy Bellman flow in the current augmented MDP; after support expansion, a previous occupancy is generally only a reference measure, not a feasible point of the new flow constraints. Define

$$q_k = (1-\varepsilon)\bar{q}_k + \varepsilon m_k, \qquad \varepsilon \in (0,1]. \tag{33}$$

Hence $q_k(x) \geq \varepsilon m_k(x) > 0$ on Ω_k^{aug}, so q_k and m_k have the same support, $m_k(x)/q_k(x) \leq 1/\varepsilon$, and

$$\rho_k(x) = \frac{q_k(x)}{m_k(x)} \tag{34}$$

is well-defined. The exact O2O population update is

$$d_{k+1} \in \arg\max_{d \in \mathcal{D}_k^{\mathrm{aug}}} \left\{ \langle r_k^{\mathrm{aug}}, d \rangle - \frac{1}{\eta} D_f(d \| q_k) \right\}, \qquad \eta > 0. \tag{35}$$

A.4 Finite-divergence feasibility and dual attainment

Let $X_k = \Omega_k^{\mathrm{aug}}$ and write augmented flow as $B_k d = b_k$, where

$$(B_k d)(s) = \sum_a d(s,a) - \gamma \sum_{(\bar{s},\bar{a}) \in X_k} P_k^{\mathrm{aug}}(s \mid \bar{s}, \bar{a}) d(\bar{s}, \bar{a}), \qquad b_k = (1-\gamma)\mu_{0,k}^{\mathrm{aug}}. \tag{36}$$

Define the negated-primal integrand

$$\Phi_k(d) = \frac{1}{\eta} \sum_{x \in X_k} q_k(x) f\left(\frac{d(x)}{q_k(x)} \right) - \langle r_k^{\mathrm{aug}}, d \rangle, \tag{37}$$

with $f(u) = +\infty$ for $u < 0$ as above.

Assumption F'_k: finite-divergence feasibility. There exists $d^0 \in \mathcal{D}_k^{\mathrm{aug}}$ such that $D_f(d^0 \| q_k) < \infty$.

Lemma A.0 (discounted-flow closedness). Under F'_k, the primal value in (35) is finite and attained, and the primal optimum equals the infimum of its full-coordinate Lagrangian dual. No relative-interior condition is required.

Proof. Row-stochasticity gives, for every $d \geq 0$,

$$\mathbf{1}^\top B_k d = (1-\gamma)\mathbf{1}^\top d. \tag{38}$$

In particular, $B_k d = b_k$ implies $\mathbf{1}^\top d = 1$; hence $\mathcal{D}_k^{\mathrm{aug}}$ is compact. The function Φ_k is proper, closed, and convex under F'_k, so its minimum over this compact feasible set is finite and attained.

For the no-gap statement, define the convex perturbation value

$$p(c) = \inf\{\Phi_k(d) : d \geq 0,\ B_k d = c\}, \tag{39}$$

with the infimum of an empty fiber equal to $+\infty$. Every nonempty fiber is bounded because (38) fixes $\|d\|_1 = \mathbf{1}^\top c/(1-\gamma)$; since Φ_k is proper and lower semicontinuous, it is bounded below on each such compact fiber, so $p(c) > -\infty$.

To prove lower semicontinuity at b_k, let $c_n \to b_k$ and suppose $\liminf_n p(c_n) < \infty$. Pass to a subsequence along which $p(c_n)$ converges to $\liminf_j p(c_j)$, and relabel it by n. Choose $d_n \geq 0$ with $B_k d_n = c_n$ and $\Phi_k(d_n) \leq p(c_n) + 1/n$. Equation (38) places d_n in a common compact set. Since d_n belongs to the defining fiber, $p(c_n) \leq \Phi_k(d_n)$ as well; hence $0 \leq \Phi_k(d_n) - p(c_n) \leq 1/n$. Along a further subsequence, $d_n \to \bar{d}$; then $B_k \bar{d} = b_k$ and lower semicontinuity gives

$$p(b_k) \leq \Phi_k(\bar{d}) \leq \liminf_n \Phi_k(d_n) = \liminf_n p(c_n). \tag{40}$$

The value function p is convex; it is proper because F'_k gives $p(b_k) < \infty$ and the compact-fiber argument gives $p(c) > -\infty$ everywhere. For a proper convex function, the biconjugation theorem gives $p^{**} = \mathrm{cl}\, p$. Lower semicontinuity at b_k therefore implies $p(b_k) = (\mathrm{cl}\, p)(b_k) = p^{**}(b_k)$. For completeness, its conjugate is

$$\begin{aligned} p^*(\lambda) &= \sup_{d \geq 0} \{ \langle \lambda, B_k d \rangle - \Phi_k(d) \} \\ &= \frac{1}{\eta} \sum_{x \in X_k} q_k(x) f^*\big(\eta[\, r_k^{\mathrm{aug}} + B_k^\top \lambda\,](x)\big)\,. \end{aligned} \tag{41}$$

Since $p(b_k) = \sup_\lambda \{\langle \lambda, b_k \rangle - p^*(\lambda)\}$, negating and setting $\lambda = -v$ gives the full-coordinate Lagrangian dual: $b_k^\top v = (1-\gamma)\langle \mu_{0,k}^{\mathrm{aug}}, v \rangle$ and $r_k^{\mathrm{aug}} - B_k^\top v = \bar{A}_v$. Hence the primal optimum equals the dual infimum. This argument establishes an infimum identity only; it does not assert dual attainment.

Checking F'_k. For shifted KL and the halved or unhalved Pearson generators, f is finite on all of $[0, \infty)$ and $q_k > 0$ on X_k. Since $\mathcal{D}_k^{\mathrm{aug}} \neq \varnothing$, F'_k then holds automatically. For a generator with a smaller effective domain, the correct check is the existence of at least one feasible d satisfying $d(x)/q_k(x) \in \operatorname{dom} f$ for every x. A structural-zero coordinate makes this impossible when $f(0) = +\infty$; merely deleting that coordinate after m_k, q_k, and the primal have been defined would instead change the optimization problem.

Assumption Attain-k: dual attainment. When pointwise KKT recovery from a finite potential is claimed, the dual below has a minimizer v_k^*. Lemma A.0 and Theorem A.1 do not need this assumption. Attainment can fail when an optimal ratio lies at a boundary point admitting no finite subgradient, including $z = 0$ for shifted KL and finite endpoints of more general effective domains. For the Pearson generators, finite-dimensional polyhedral KKT conditions imply this assumption automatically, as shown below.

Active and inactive coordinates. Let

$$\Omega_k^{\mathrm{act}} = \{x \in X_k : \exists d \in \mathcal{D}_k^{\mathrm{aug}} \text{ with } d(x) > 0\}. \tag{42}$$

In the directed graph containing an edge $s \to s'$ whenever some allowed (s, a) has $P_k^{\mathrm{aug}}(s' \mid s, a) > 0$, a coordinate (s, a) is active if and only if s is reachable from $\operatorname{supp} \mu_{0,k}^{\mathrm{aug}}$. Let R_k and U_k denote the reachable and unreachable state sets and put $h = \mathbf{1}_{U_k}$. Then

$$b_k^\top h = 0, \qquad [B_k^\top h](s, a) = 0 \ \ (s \in R_k), \qquad [B_k^\top h](s, a) \geq 1 - \gamma \ \ (s \in U_k). \tag{43}$$

Consequently, along $v_t = v + th$, every active Bellman residual is unchanged and every inactive residual tends to $-\infty$. If $f(0) < \infty$, closed convex conjugacy gives

$$\lim_{y \to -\infty} f^*(y) = -f(0), \tag{44}$$

which recovers in the full dual the constant contribution of the inactive primal coordinates. In general all unreachable states must be shifted together; changing only $v(\perp)$ need not preserve the other residuals.

For shifted KL, a finite dual minimizer exists if and only if $\Omega_k^{\mathrm{act}} = X_k$. Necessity follows because finite v gives $q_k(x)e^{\eta \bar{A}_v(x)} > 0$ at every coordinate. Conversely, when every coordinate is active, averaging one feasible occupancy that is positive at each coordinate gives a strictly positive feasible occupancy. If a shifted-KL primal optimum had a zero coordinate, the directional derivative of Φ_k toward that positive feasible occupancy would be $-\infty$, contradicting optimality. Thus the optimum is strictly positive, and the affine first-order condition then supplies a finite flow multiplier.

Automatic Pearson attainment. Let $f_c(z) = c(z - 1)^2$ on $z \geq 0$, where $c = \frac{1}{2}$ gives the halved Pearson generator and $c = 1$ gives the unhalved convention. Because $q_k > 0$ on X_k, the negated primal objective is the restriction to the nonnegative orthant of the continuously differentiable convex quadratic

$$\Phi_{k,c}(d) = \frac{c}{\eta} \sum_{x \in X_k} \frac{(d(x) - q_k(x))^2}{q_k(x)} - \langle r_k^{\mathrm{aug}}, d \rangle. \tag{45}$$

Let d^* minimize (45) over the nonempty polyhedron $\mathcal{C}_k = \{d \geq 0 : B_k d = b_k\}$. First-order optimality and the normal-cone formula for a polyhedron give finite vectors v^* and μ^* such that

$$0 = \frac{2c}{\eta} \left(\frac{d^*}{q_k} - \mathbf{1} \right) - r_k^{\mathrm{aug}} + B_k^\top v^* - \mu^*, \qquad \mu^* \geq 0, \qquad \mu^* \odot d^* = 0. \tag{46}$$

Writing $z^* = d^*/q_k$ and $\bar{A}_{v^*} = r_k^{\mathrm{aug}} - B_k^\top v^*$, the positive coordinates of (46) satisfy $z^* = 1 + \eta \bar{A}_{v^*}/(2c)$, while a zero coordinate satisfies $\bar{A}_{v^*} \leq -2c/\eta$. Hence, componentwise,

$$z^*(x) = \max\left\{0, 1 + \frac{\eta \bar{A}_{v^*}(x)}{2c}\right\}. \tag{47}$$

This is the unique scalar maximizer of the Pearson Lagrangian at every coordinate. Therefore d^* maximizes $\mathcal{L}(\cdot, v^*)$ over the nonnegative cone, the dual objective at the finite potential v^* equals the

primal optimum, and v^* is a dual minimizer. Thus Attain-k holds automatically for both Pearson conventions, including in the presence of inactive coordinates.

As an optional sufficient condition, if replay support is constructed from complete trajectories started from μ_0 and predecessor state–action pairs are retained cumulatively, every replay-covered environment state is reachable. If, in addition, initial mass lies outside the covered set or some allowed environment pair leaks from it, then $\perp$ is reachable as well and $\Omega_k^{\mathrm{act}} = X_k$. This is a provenance condition on replay, not a consequence of an arbitrary collection of supported pairs.

A.5 Base-measure dual

For $v : \mathcal{S}_k^{\mathrm{aug}} \to \mathbb{R}$ define

$$\bar{A}_v(s,a) = r_k^{\mathrm{aug}}(s,a) + \gamma \sum_{s'} P_k^{\mathrm{aug}}(s' \mid s,a) v(s') - v(s). \tag{48}$$

The absorbing-state boundary case is discussed in Assumption Attain-k above.

Theorem A.1 (proof of Theorem 1). Under Assumption F'_k, (35) has a finite attained optimum and zero duality gap with the dual infimum

$$\inf_v (1-\gamma) \sum_{s \in \mathcal{S}_k^{\mathrm{aug}}} \mu_{0,k}^{\mathrm{aug}}(s) v(s) + \frac{1}{\eta} \sum_{(s,a) \in \Omega_k^{\mathrm{aug}}} q_k(s,a) f^*(\eta \bar{A}_v(s,a)). \tag{49}$$

Equivalently, using the base measure m_k,

$$\inf_v (1-\gamma) \mathbb{E}_{s \sim \mu_{0,k}^{\mathrm{aug}}}[v(s)] + \frac{1}{\eta} \mathbb{E}_{(s,a)\sim m_k} \left[\rho_k(s,a) f^*(\eta \bar{A}_v(s,a)) \right]. \tag{50}$$

No Bellman-flow property of m_k is used.

Proof. For a multiplier v, subtract the flow residual from the primal objective:

$$\begin{aligned} \mathcal{L}(d,v) = \langle r_k^{\mathrm{aug}}, d \rangle &- \frac{1}{\eta} D_f(d \| q_k) \\ &- \sum_{s \in \mathcal{S}_k^{\mathrm{aug}}} v(s) \left[\sum_a d(s,a) - \gamma \sum_{\bar{s},\bar{a}} P_k^{\mathrm{aug}}(s \mid \bar{s},\bar{a}) d(\bar{s},\bar{a}) - (1-\gamma) \mu_{0,k}^{\mathrm{aug}}(s) \right]. \end{aligned} \tag{51}$$

For feasible d, the residual is zero. In the finite setting,

$$\sum_s v(s) \sum_{\bar{s},\bar{a}} P_k^{\mathrm{aug}}(s \mid \bar{s},\bar{a}) d(\bar{s},\bar{a}) = \sum_{\bar{s},\bar{a}} d(\bar{s},\bar{a}) \sum_s P_k^{\mathrm{aug}}(s \mid \bar{s},\bar{a}) v(s), \tag{52}$$

so

$$\mathcal{L}(d,v) = (1-\gamma) \langle \mu_{0,k}^{\mathrm{aug}}, v \rangle + \sum_{(s,a) \in \Omega_k^{\mathrm{aug}}} d(s,a) \bar{A}_v(s,a) - \frac{1}{\eta} D_f(d \| q_k). \tag{53}$$

Because $q_k(x) > 0$ on Ω_k^{aug}, write $d(x) = q_k(x) z(x)$ with $z(x) \geq 0$.

Unit-mass remark. No normalization constraint is lost in the next display. Summing the augmented flow equations (28) over every state in $\mathcal{S}_k^{\mathrm{aug}}$, including $\perp$, and using row-stochasticity of P_k^{aug} gives $\sum_x d(x) = 1$. After these flow constraints have been dualized, the Lagrangian supremum is therefore correctly taken over the nonnegative cone. For fixed v,

$$\begin{aligned} \sup_{d \geq 0} \left\{ \sum_x d(x) \bar{A}_v(x) - \frac{1}{\eta} D_f(d \| q_k) \right\} &= \sup_{z \geq 0} \sum_x q_k(x) \left[z(x) \bar{A}_v(x) - \frac{1}{\eta} f(z(x)) \right] \\ &= \frac{1}{\eta} \sum_x q_k(x) f^*(\eta \bar{A}_v(x)). \end{aligned} \tag{54}$$

Adding the initial-state term gives (49). Lemma A.0 licenses equality between the primal optimum and the dual infimum. Finally, the change of base $q_k = \rho_k m_k$ gives the base-measure form (50).

Exact offline support-restricted dual. The fixed-support offline update uses $q_k = d_k$, which can vanish on X_k. For an exact feasible occupancy d_k, define

$$X_k^{\mathrm{off}} = \mathrm{supp}(d_k), \qquad \mathcal{D}_k^{\mathrm{off}} = \{d \in \mathcal{D}_k^{\mathrm{aug}} : d \ll d_k\}. \tag{55}$$

On X_k^{off}, the reference d_k is strictly positive and $d_k \in \mathcal{D}_k^{\mathrm{off}}$ witnesses F'_k. Lemma A.0 and the preceding derivation therefore apply verbatim with all coordinate sums restricted to X_k^{off}. Equivalently, zero-reference coordinates contribute the conjugate of the constraint $d(x) = 0$, namely zero, and can be omitted. In the base-measure notation it is safest to sum only over X_k^{off}, rather than leave an ambiguous product $0 \cdot f^*(\eta \bar{A}_v)$ when f^* is infinite. Because d_k is strictly positive on X_k^{off}, every restricted coordinate is active; consequently the shifted-KL and Pearson offline duals are also attained on this restricted problem.

For an approximate carried object $\hat{d}_k$, the same restricted dual is valid only if

$$\{d \in \mathcal{D}_k^{\mathrm{aug}} : d \ll \hat{d}_k,\ D_f(d\|\hat{d}_k) < \infty\} \neq \varnothing. \tag{56}$$

The vector $\hat{d}_k$ itself witnesses this condition only when it satisfies Bellman flow. Otherwise a feasibility projection or a full-support anchored reference is required; the exact offline monotonicity theorem does not cover that approximate replacement.

A.6 Pointwise Fenchel/KKT recovery

Proposition A.2 (pointwise KKT characterization). Assume F'_k and Attain-k; for either Pearson generator, the latter follows automatically from (45)– (47). Let d_{k+1} solve (35) and let v_k^* solve (50). Define $z_{k+1}(x) = d_{k+1}(x)/q_k(x)$ on the support of q_k. Then, for every $x = (s, a)$ with $q_k(x) > 0$,

$$z_{k+1}(x) \in \partial f^*(\eta \bar{A}_{v_k^*}(x)), \qquad \eta \bar{A}_{v_k^*}(x) \in \partial f(z_{k+1}(x)). \tag{57}$$

Equivalently,

$$z_{k+1}(x) \in \arg\max_{z \geq 0} \{z\, \eta \bar{A}_{v_k^*}(x) - f(z)\}. \tag{58}$$

For differentiable, strictly convex f, an interior maximizer satisfies $z_{k+1}(x) = (f')^{-1}(\eta \bar{A}_{v_k^*}(x))$. If the maximizer lies at the boundary, the constrained argmax (58) is the correct statement. For shifted KL $f(z) = z \log z - z + 1$ with the lower-semicontinuous extension $f(0) = 1$, $f^*(y) = e^y - 1$ and $z = \exp(\eta \bar{A}_{v_k^*}(x))$ at finite potentials. Thus an inactive coordinate precludes a finite shifted-KL dual minimizer and is represented only by a limiting dual sequence. For Pearson $f(z) = \frac{1}{2}(z-1)^2$ on $z \geq 0$, $z = \max\{0, 1 + \eta \bar{A}_{v_k^*}(x)\}$. Under ODICE's unhalved Pearson convention $f(z) = (z-1)^2$, exact KKT recovery gives $z = \max\{0, 1 + \eta \bar{A}_{v_k^*}(x)/2\}$.

Converse (recovery from the dual alone). If, for every x, the scalar argmax in (58) is a singleton $z^\dagger(x)$, then $\mathcal{L}(\cdot, v_k^*)$ has a unique maximizer over the nonnegative cone. Since every primal optimum is one such maximizer,

$$d_{k+1} = q_k \odot z^\dagger, \tag{59}$$

and the pointwise vector automatically satisfies augmented Bellman flow. Shifted KL and the halved or unhalved Pearson generators are strictly convex, so their scalar maximizers are unique; moreover $D_f(\cdot\|q_k)$ is strictly convex when $q_k > 0$, and hence the finite-divergence primal optimizer is unique. For a non-strict generator, Proposition A.2 is only a characterization of an optimal pair: an arbitrary pointwise selection can violate flow.

Proof. Primal-dual optimality gives

$$P^* = \mathcal{L}(d_{k+1}, v_k^*) = \inf_v \sup_{d \geq 0} \mathcal{L}(d, v) = \sup_{d \geq 0} \mathcal{L}(d, v_k^*). \tag{60}$$

Thus d_{k+1} maximizes $\mathcal{L}(\cdot, v_k^*)$ over the nonnegative cone. After $d = q_k z$, the integrand and constraint $z(x) \geq 0$ decouple across x. Since $q_k(x) > 0$, each $z_{k+1}(x)$ maximizes $z \mapsto z\, \eta \bar{A}_{v_k^*}(x) - f(z)$ over $z \geq 0$.

Equivalently, define the Fenchel gap

$$G(x) = f(z_{k+1}(x)) + f^*(\eta \bar{A}_{v_k^*}(x)) - z_{k+1}(x) \eta \bar{A}_{v_k^*}(x). \tag{61}$$

Fenchel-Young gives $G(x) \geq 0$. Strong duality and attainment imply

$$\sum_{x \in \Omega_k^{\text{aug}}} q_k(x) G(x) = 0. \tag{62}$$

Since $q_k(x) > 0$ on the finite support, $G(x) = 0$ for every x. Equality in Fenchel-Young is equivalent to (57).

A.7 Policy extraction

For any $d \in \mathcal{D}_k^{\text{aug}}$, define $d_S(s) = \sum_a d(s,a)$. The exact induced policy is

$$\pi_d(a \mid s) = \frac{d(s,a)}{d_S(s)} \tag{63}$$

when $d_S(s) > 0$, with arbitrary actions on zero-marginal states. In the augmented MDP this policy has discounted occupancy d.

Realizability. For exact weighted-behavior-cloning recovery, the policy class must contain $\pi_{d_{k+1}}(\cdot \mid s)$ on every state with positive d_{k+1} marginal. Otherwise WBC is a projection onto the chosen policy class, as in practical DICE implementations.

The finite exact WBC objective is

$$\max_{\pi} \sum_{(s,a) \in \Omega_k^{\text{aug}}} d_{k+1}(s,a) \log \pi(a \mid s). \tag{64}$$

Changing the sampling base to m_k gives

$$\max_{\pi} \mathbb{E}_{(s,a) \sim m_k} \left[\omega_k(s,a) \log \pi(a \mid s) \right], \qquad \omega_k(s,a) = \rho_k(s,a) z_{k+1}(s,a). \tag{65}$$

Thus if samples are drawn from m_k, the exact population weight is $\rho_k z_{k+1} = d_{k+1}/m_k$. For environment actor fitting, terms with state $\perp$ may be ignored or treated as synthetic bookkeeping terms; the action distribution at $\perp$ is arbitrary and is never executed in the original environment.

A.8 Augmentation mismatch

Let $\tau_k = \inf\{t : s_t \notin \mathcal{S}_k^{\text{env}}\}$. Extend π arbitrarily at $\perp$. Couple the original and augmented trajectories by the same transition randomness so that they agree until τ_k; when the original transition lands outside $\mathcal{S}_k^{\text{env}}$, the augmented transition lands at $\perp$ and then stays there with zero reward. On the event $\{\tau_k = 0\}$, the augmented trajectory starts at $\perp$, and the same tail bound below applies from time zero. If $|r| \leq R_{\max}$, then the pre-exit rewards agree, the augmented post-exit rewards are zero, and the triangle inequality gives

$$\begin{aligned} |J(\pi) - J_k^{\text{aug}}(\pi)| &\leq \mathbb{E}_\pi \left[\sum_{t \geq \tau_k} \gamma^t R_{\max} \mathbf{1}\{\tau_k < \infty\} \right] \\ &= \frac{R_{\max}}{1-\gamma} \mathbb{E}_\pi[\gamma^{\tau_k} \mathbf{1}\{\tau_k < \infty\}]. \end{aligned} \tag{66}$$

Since $\{\tau_k = t\} \subseteq \{s_t \notin \mathcal{S}_k^{\text{env}}\}$,

$$\mathbb{E}_\pi[\gamma^{\tau_k} \mathbf{1}\{\tau_k < \infty\}] = \sum_{t \geq 0} \gamma^t \Pr_\pi(\tau_k = t) \leq \sum_{t \geq 0} \gamma^t \Pr_\pi(s_t \notin \mathcal{S}_k^{\text{env}}), \tag{67}$$

which proves Lemma 1.

Define

$$\Delta_k(\pi) = \frac{R_{\max}}{1-\gamma} \mathbb{E}_\pi[\gamma^{\tau_k} \mathbf{1}\{\tau_k < \infty\}]. \tag{68}$$

Because $\mathcal{S}_k^{\text{env}} \subseteq \mathcal{S}_{k+1}^{\text{env}}$, the same original-MDP trajectory satisfies $\tau_{k+1} \geq \tau_k$ and therefore $\Delta_{k+1}(\pi) \leq \Delta_k(\pi)$ for every fixed π. Only this certified upper bound is monotone; cancellation among signed rewards can make the actual absolute mismatch nonmonotone.

A.9 IMPROVEMENT AND COMPARATOR BOUNDS

Theorem A.3: fixed-support augmented-return DICE improvement (proof of Theorem 2). Assume the augmented MDP and feasible set are fixed across iterations. For the fixed-support offline-phase theorem, let $d_k \in \mathcal{D}_k^{\mathrm{aug}}$ and define d_{k+1} by the support-conservative update

$$d_{k+1} \in \arg\max_{d \in \mathcal{D}_k^{\mathrm{aug}},\, d \ll d_k} \left\{ \langle r^{\mathrm{aug}}, d \rangle - \frac{1}{\eta} D_f(d \| d_k) \right\}. \tag{69}$$

The restricted feasible set is compact, contains d_k, and the objective is upper semicontinuous and finite at d_k; hence this argmax is nonempty. Then

$$\langle r^{\mathrm{aug}}, d_{k+1} \rangle \geq \langle r^{\mathrm{aug}}, d_k \rangle + \frac{1}{\eta} D_f(d_{k+1} \| d_k). \tag{70}$$

Proof. By optimality of d_{k+1} and feasibility of d_k,

$$\langle r^{\mathrm{aug}}, d_{k+1} \rangle - \frac{1}{\eta} D_f(d_{k+1} \| d_k) \geq \langle r^{\mathrm{aug}}, d_k \rangle - \frac{1}{\eta} D_f(d_k \| d_k). \tag{71}$$

Since $f(1) = 0$, $D_f(d_k \| d_k) = 0$. Rearranging proves the claim. Jensen's inequality gives $D_f(d_{k+1} \| d_k) \geq 0$, so (29) identifies this as augmented-return monotonicity. This theorem is support-conservative: if a candidate is not absolutely continuous with respect to d_k, its divergence is infinite and it is excluded. In the growing-support O2O setting, the anchored choice $\varepsilon > 0$ removes the divergence barrier for newly collected coordinates that are compatible with flow and leads to the local comparator guarantee in Theorem 3.

Theorem A.4: per-step local comparator bound (proof of Theorem 3). Under F'_k, Lemma A.0 ensures that a finite optimizer d_{k+1} of (35) exists. For any comparator $d \in \mathcal{D}_k^{\mathrm{aug}}$ with $D_f(d \| q_k) < \infty$,

$$\langle r_k^{\mathrm{aug}}, d_{k+1} \rangle \geq \langle r_k^{\mathrm{aug}}, d \rangle - \frac{1}{\eta} D_f(d \| q_k). \tag{72}$$

Proof. By optimality of d_{k+1},

$$\langle r_k^{\mathrm{aug}}, d_{k+1} \rangle - \frac{1}{\eta} D_f(d_{k+1} \| q_k) \geq \langle r_k^{\mathrm{aug}}, d \rangle - \frac{1}{\eta} D_f(d \| q_k). \tag{73}$$

Rearranging gives

$$\langle r_k^{\mathrm{aug}}, d_{k+1} \rangle \geq \langle r_k^{\mathrm{aug}}, d \rangle - \frac{1}{\eta} D_f(d \| q_k) + \frac{1}{\eta} D_f(d_{k+1} \| q_k). \tag{74}$$

The final term is nonnegative. Indeed, d_{k+1} is a probability measure because augmented flow implies unit mass, and q_k is a probability measure because it is a convex combination of $\bar{q}_k$ and m_k. For probability measures p, q, Jensen and $f(1) = 0$ give

$$D_f(p \| q) = \mathbb{E}_q \left[f\left(\frac{dp}{dq} \right) \right] \geq f\left(\mathbb{E}_q \frac{dp}{dq} \right) = f(1) = 0. \tag{75}$$

Dropping the nonnegative term proves the bound. The result requires only local domination of the comparator by q_k; it does not require m_k itself to be a Bellman-flow occupancy.

Cumulative local comparator bound. For any sequence $d_k^\star \in \mathcal{D}_k^{\mathrm{aug}}$ with $D_f(d_k^\star \| q_k) < \infty$, summing Theorem A.4 gives

$$\sum_{k=0}^{K-1} \left[\langle r_k^{\mathrm{aug}}, d_k^\star \rangle - \langle r_k^{\mathrm{aug}}, d_{k+1} \rangle \right] \leq \frac{1}{\eta} \sum_{k=0}^{K-1} D_f(d_k^\star \| q_k). \tag{76}$$

This bounds the cumulative comparison error against feasible finite-divergence comparators on the current supports. The bound depends on their divergences to the anchored references. For shifted KL and $m_k = \mathrm{Unif}(\Omega_k^{\mathrm{aug}})$, the anchor gives $q_k(x) \geq \varepsilon / |\Omega_k^{\mathrm{aug}}|$ for every $x \in \Omega_k^{\mathrm{aug}}$, so any feasible comparator $d \in \mathcal{D}_k^{\mathrm{aug}}$ satisfies

$$D_{\mathrm{KL}}(d \| q_k) \leq \log \frac{|\Omega_k^{\mathrm{aug}}|}{\varepsilon}. \tag{77}$$

Corollary A.5: original-MDP forms. Let π_k, π_{k+1} be induced by d_k, d_{k+1} in Theorem A.3 and let π_d be induced by a comparator d in Theorem A.4. Choose arbitrary but fixed extensions of these policies outside $\mathcal{S}_k^{\mathrm{env}}$, and use the same extensions in J and Δ_k. Combining (29), Lemma 1, and Theorems A.3–A.4 gives

$$J(\pi_{k+1}) \geq J(\pi_k) + \frac{D_f(d_{k+1}\|d_k)}{\eta(1-\gamma)} - \Delta_k(\pi_{k+1}) - \Delta_k(\pi_k), \tag{78}$$

$$J(\pi_{k+1}) \geq J(\pi_d) - \frac{D_f(d\|q_k)}{\eta(1-\gamma)} - \Delta_k(\pi_{k+1}) - \Delta_k(\pi_d). \tag{79}$$

Proof. Divide the corresponding augmented occupancy inequality by $1-\gamma$ using (29). For the updated policy use $J(\pi_{k+1}) \geq J_k^{\mathrm{aug}}(\pi_{k+1}) - \Delta_k(\pi_{k+1})$; for the current or comparator policy use $J_k^{\mathrm{aug}}(\pi) \geq J(\pi) - \Delta_k(\pi)$. Substitution gives the two displays.

B Continuous-space population theory under domination

Scope. This section extends the population theorem in Appendix A to standard Borel spaces. We consider occupancies that are absolutely continuous with respect to a population base measure. We first prove the update at a fixed iteration, then give a construction that keeps the anchored references well defined across iterations.

B.1 Measurable augmented MDP and dominated occupancies

Fix an iteration and suppress the subscript k. Let $\mathcal{S}$ and $\mathcal{A}$ be standard Borel spaces, let $E \subseteq \mathcal{S}$ be a Borel replay-covered state region, and let $\Gamma \subseteq E \times \mathcal{A}$ be a Borel graph of allowed actions. We assume that there exists at least one Borel Markov kernel supported on Γ; this is automatic under standard measurable-selection hypotheses, for example for a measurable nonempty compact-valued action correspondence. Introduce an absorbing state $\perp$ and one dummy absorbing action $a_\perp$. With the unions understood as disjoint, define

$$\mathcal{S}^+ = E \cup \{\perp\}, \qquad \mathcal{A}^+ = \mathcal{A} \cup \{a_\perp\}, \qquad X = \Gamma \cup \{(\perp, a_\perp)\}. \tag{80}$$

The singleton action at $\perp$ is without loss because the action taken after absorption has no effect on either dynamics or reward.

Let $P(\cdot \mid s,a)$ be a Borel stochastic kernel on $\mathcal{S}$, let $r : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ be bounded and Borel measurable, and let μ_0 be the initial distribution. The augmented objects are

$$\begin{aligned} P^+(C \mid s,a) &= P(C \mid s,a), && C \subseteq E,\ (s,a) \in \Gamma, \\ P^+(\{\perp\} \mid s,a) &= P(\mathcal{S} \setminus E \mid s,a), && (s,a) \in \Gamma, \\ P^+(\{\perp\} \mid \perp, a_\perp) &= 1, && \end{aligned} \tag{81}$$

$$\begin{aligned} \mu_0^+(C) &= \mu_0(C), && C \subseteq E, \\ \mu_0^+(\{\perp\}) &= \mu_0(\mathcal{S} \setminus E). && \end{aligned} \tag{82}$$

Define the augmented reward by $r^+(s,a) = r(s,a)$ on Γ and $r^+(\perp, a_\perp) = 0$. Thus $r^+ : X \to \mathbb{R}$ is bounded and Borel measurable, and the original and augmented trajectories receive identical rewards before the first exit from E.

An admissible policy is a Borel Markov kernel $\pi(\cdot \mid s)$ on $\mathcal{A}^+$ satisfying $\pi(\Gamma_s \mid s) = 1$ for $s \in E$, where $\Gamma_s = \{a : (s,a) \in \Gamma\}$, and $\pi(\{a_\perp\} \mid \perp) = 1$. Its normalized discounted occupancy and unnormalized augmented return are

$$d^\pi = (1-\gamma)\sum_{t\geq 0} \gamma^t \mathcal{L}_\pi(S_t, A_t), \tag{83}$$

$$J^+(\pi) = \mathbb{E}_\pi\left[\sum_{t\geq 0} \gamma^t r^+(S_t, A_t)\right]. \tag{84}$$

The return is absolutely convergent because r^+ is bounded and $0 < \gamma < 1$.

Let $m \in \mathcal{P}(X)$ be a population base measure and let $\bar{q} \in \mathcal{P}(X)$ satisfy $\bar{q} \ll m$. For $\varepsilon \in (0, 1]$, define

$$q = (1-\varepsilon)\bar{q} + \varepsilon m, \qquad \rho = \frac{dq}{dm} = (1-\varepsilon)\frac{d\bar{q}}{dm} + \varepsilon. \tag{85}$$

Thus $q \sim m$ and $\rho \geq \varepsilon$ m-almost everywhere. The continuous anchor therefore dominates every occupancy d satisfying $d \ll m$; unlike the finite anchor, it cannot dominate arbitrary probability measures on X.

For $z \in L^1(q)$ with $z \geq 0$, write $d_z = zq$. We say that z satisfies weak discounted Bellman flow if, for every bounded Borel $\varphi : \mathcal{S}^+ \to \mathbb{R}$,

$$\int_X \left[\varphi(s) - \gamma \int_{\mathcal{S}^+} \varphi(s') P^+(ds' \mid s, a)\right] z(s,a) q(ds, da) = (1-\gamma) \int_{\mathcal{S}^+} \varphi(s) \mu_0^+(ds). \tag{86}$$

Denote the set of such densities by $\mathcal{Z}(q)$ and the corresponding dominated occupancy set by

$$\mathcal{D}(q) = \{d_z : z \in \mathcal{Z}(q)\}. \tag{87}$$

Lemma B.1 (Weak-flow occupancy converse)**.** *Every $z \in \mathcal{Z}(q)$ has $\int z\, dq = 1$. Moreover, d_z admits a disintegration*

$$d_z(ds, da) = d_{z,S}(ds)\pi_z(da \mid s), \tag{88}$$

where an admissible Borel Markov policy π_z can be chosen so that its normalized discounted occupancy in the augmented MDP is exactly d_z. Consequently,

$$\int_X r^+ \, dd_z = (1-\gamma) J^+(\pi_z). \tag{89}$$

Proof. Taking $\varphi \equiv 1$ in (86) gives $(1-\gamma)\int z\, dq = 1 - \gamma$. Since $\mathcal{S}^+$ and the augmented action space are standard Borel, the disintegration theorem gives (88). It is supported on the allowed graph for $d_{z,S}$-almost every state. On the exceptional null set it can be replaced by the admissible kernel assumed above.

Equation (86) is the measure identity

$$d_{z,S} = (1-\gamma)\mu_0^+ + \gamma P^*_{\pi_z} d_{z,S}. \tag{90}$$

The right side defines a contraction with modulus γ on finite signed state measures under total variation. Hence (90) has a unique solution, namely $(1-\gamma)\sum_{t\geq 0} \gamma^t \mu_0^+ P^t_{\pi_z}$. Therefore d_z is the discounted occupancy of π_z. The bounded-reward identity follows by Fubini's theorem, since the discounted reward series is absolutely integrable. This is the discounted finite-measure case of the general continuous-space occupancy/disintegration result of Laroche & Des Combes (2023).

Lemma B.2 (An operational domination condition)**.** *Suppose the base measure is the occupancy $m = d^\beta$ of an admissible Markov coverage policy β. If another admissible Markov policy π satisfies*

$$\pi(\cdot \mid s) \ll \beta(\cdot \mid s) \quad \textit{for every } s \in \mathcal{S}^+, \tag{91}$$

then $d^\pi \ll m$, and hence $d^\pi \ll q$.

Proof. Induction on time shows that the state-action marginal generated by π at time t is absolutely continuous with respect to the corresponding marginal generated by β: action-kernel domination gives the state-action step and applying the same transition kernel preserves absolute continuity at the next state step. If $d^\beta(B) = 0$, every nonnegative discounted time marginal of β assigns zero mass to B, so every time marginal of π does as well. Thus $d^\pi(B) = 0$. Finally, $q \geq \varepsilon m$ implies $m \ll q$.

Lemma B.2 makes the density restriction concrete. For example, nondegenerate Gaussian action kernels are mutually absolutely continuous, even when the environment dynamics are deterministic. More generally, a mixture of coverage-policy occupancies can be used as m. If a carried reference $\bar{q}$ is not already dominated by a feasible coverage occupancy d^β, one may choose, for $\alpha \in (0, 1)$,

$$m = \alpha\bar{q} + (1-\alpha)d^\beta. \tag{92}$$

Then $\bar{q} \ll m$, while $d^\beta \ll m$ with $dd^\beta/dm \leq 1/(1-\alpha)$. Since $q \geq \varepsilon m$, the coverage occupancy is a q-dominated feasible witness with bounded density. Thus shifted KL and Pearson satisfy condition CF even though this mixture base m need not itself satisfy Bellman flow.

Lemma B.3 (A sufficient sequential anchor construction)**.** *At each iteration k, let $\bar{q}_k$ be a probability measure on the current allowed graph X_k and let d^{β_k} be one feasible normalized occupancy of the current augmented MDP. Fix $\alpha, \varepsilon \in (0,1)$ and define*

$$m_k = \alpha\bar{q}_k + (1-\alpha)d^{\beta_k}, \qquad q_k = (1-\varepsilon)\bar{q}_k + \varepsilon m_k. \tag{93}$$

Then $\bar{q}_k \ll m_k$, $q_k \sim m_k$, and

$$\frac{dd^{\beta_k}}{dq_k} \le \frac{1}{\varepsilon(1-\alpha)} \qquad q_k\text{-a.e.} \tag{94}$$

Consequently condition CF holds for shifted KL and Pearson, and more generally whenever f is finite on $[0, 1/\{\varepsilon(1-\alpha)\}]$. If a recovered optimizer is carried to the next allowed graph as $\bar{q}_{k+1}$—for example by zero extension when the graphs are cumulative—the same construction applies recursively. No inter-iteration condition such as $m_k \ll m_{k+1}$ is required.

Proof. The first two claims follow directly from the two mixtures. Moreover $q_k \ge \varepsilon m_k \ge \varepsilon(1-\alpha)d^{\beta_k}$, which proves (94). The displayed density is feasible, nonnegative, and bounded in the stated interval, so its f-integral is finite. Every recovered occupancy is a probability measure by Lemma B.1; inserting its transported copy as a component of m_{k+1} re-establishes the same-round domination directly.

B.2 The L^1 Bellman-flow operator

Let q_S be the state marginal of q and let

$$q_P(C) = \int_X P^+(C \mid s,a)q(ds,da). \tag{95}$$

Introduce the finite state reference measure

$$\lambda = \frac{1}{3}(\mu_0^+ + q_S + q_P). \tag{96}$$

For a signed $z \in L^1(q)$, both $(zq)_S$ and $P_*^+(zq)$ are absolutely continuous with respect to λ. Define the bounded linear operator

$$\mathsf{B} : L^1(q) \to L^1(\lambda), \qquad \mathsf{B}z = \frac{d\{(zq)_S - \gamma P_*^+(zq)\}}{d\lambda}. \tag{97}$$

It satisfies

$$\|\mathsf{B}z\|_{L^1(\lambda)} \le (1+\gamma)\|z\|_{L^1(q)}, \qquad \int \mathsf{B}z\, d\lambda = (1-\gamma)\int z\, dq. \tag{98}$$

Let

$$b = (1-\gamma)\frac{d\mu_0^+}{d\lambda} \in L^1(\lambda). \tag{99}$$

Then weak Bellman flow is equivalent to $\mathsf{B}z = b$ in $L^1(\lambda)$. For $v \in L^\infty(\lambda)$, the adjoint is the q-a.e. defined bounded function

$$(\mathsf{B}^*v)(s,a) = v(s) - \gamma\int v(s')P^+(ds' \mid s,a). \tag{100}$$

The expression is independent of the chosen λ-a.e. representative: $q_S \ll \lambda$ and $q_P \ll \lambda$ imply that changes on a λ-null set vanish in both terms q-almost everywhere.

B.3 Continuous-space strong duality

Assume throughout this subsection that

1. $0 < \gamma < 1$ and $r^+ \in L^\infty(q)$;
2. $f : [0,\infty) \to (-\infty, +\infty]$ is proper, lower semicontinuous, convex, $f(1) = 0$, and is extended by $+\infty$ on $(-\infty, 0)$;

3. f is superlinear:
$$\lim_{t\to\infty} \frac{f(t)}{t} = +\infty; \tag{101}$$
4. the dominated finite-divergence feasibility condition holds:
$$\exists z^0 \in \mathcal{Z}(q) \quad \text{such that} \quad \int_X f(z^0)\, dq < \infty. \tag{CF}$$

Shifted KL and both Pearson conventions satisfy (101). Condition CF is automatic, for these generators, if m itself is a feasible occupancy: since $q \geq \varepsilon m$, $dm/dq \leq 1/\varepsilon$ and is a finite-divergence feasible density.

This domination restriction is also exactly the finite-divergence domain of the usual extended Csiszár divergence for the generators used in the paper. Indeed, if $d = zq + d^\perp$ is the Lebesgue decomposition of a probability measure relative to q, then
$$D_f(d\|q) = \int f(z)dq + f'_\infty d^\perp(X), \qquad f'_\infty = \lim_{t\to\infty} \frac{f(t)}{t}. \tag{102}$$
Under (101), $f'_\infty = +\infty$, so finite divergence forces $d^\perp = 0$, equivalently $d \ll q$. Thus (87) does not discard any finite-divergence competitor for shifted KL or Pearson; it makes their implicit coverage condition explicit.

Equivalent occupancy formulation. Let $\mathcal{D}^+_{\text{all}} = \{d^\pi : \pi \text{ is admissible}\}$ denote all normalized occupancies of the augmented MDP, without imposing domination in the definition. Under (101) and condition CF,
$$\begin{aligned}\sup_{d\in\mathcal{D}^+_{\text{all}}} &\left\{ \int_X r^+\, dd - \frac{1}{\eta} D_f(d\|q) \right\} \\ &= \sup_{z\in\mathcal{Z}(q)} \left\{ \int_X r^+ z\, dq - \frac{1}{\eta} \int_X f(z)\, dq \right\}. \end{aligned} \tag{103}$$
Indeed, every singular occupancy has infinite divergence and objective value $-\infty$, while every q-dominated occupancy has a density in $\mathcal{Z}(q)$ by the weak-flow identity. Thus the right-hand problem contains all finite-divergence competitors of the full occupancy problem.

Building on existing characterizations of discounted occupancies through measurable Bellman flow and disintegration into Markov kernels (Laroche & Des Combes, 2023; Jhaveri et al., 2026), we establish the SupportDICE population update for a declared domination class and a superlinear f-divergence: the infinite-dimensional primal is attained and its bounded-potential DICE dual has the same value, including structural-zero contributions.

Consider the continuous population update
$$\sup_{z\in\mathcal{Z}(q)} \left\{ \int_X r^+(x) z(x) q(dx) - \frac{1}{\eta} \int_X f(z(x)) q(dx) \right\}. \tag{CP}$$

Theorem B.1 (Dominated continuous-space SupportDICE duality)**.** *Under the preceding assumptions,* (CP) *has a finite attained optimum and has zero duality gap with*
$$\inf_{v\in L^\infty(\lambda)} \left\{ (1-\gamma) \int_{\mathcal{S}^+} v\, d\mu_0^+ + \frac{1}{\eta} \int_X f^*\big(\eta \bar{A}_v(x)\big)\, q(dx) \right\}, \tag{CV}$$
where
$$\bar{A}_v(s,a) = r^+(s,a) + \gamma \int v(s') P^+(ds' \mid s,a) - v(s) = r^+(s,a) - \mathsf{B}^* v(s,a). \tag{104}$$
Equivalently, using (85)*, the second term in* (CV) *is*
$$\frac{1}{\eta} \int_X \rho(x) f^*(\eta \bar{A}_v(x)) m(dx). \tag{105}$$
The theorem asserts equality with the dual infimum; it does not assert that a minimizer v^* exists.

Proof: primal existence. Define on $L^1(q)$

$$\Phi(z) = \frac{1}{\eta}\int f(z)\,dq - \int r^+ z\,dq. \tag{106}$$

The extension $f(t) = +\infty$ for $t < 0$ incorporates nonnegativity. A proper closed convex function on the real line has an affine minorant, so there are $a, c \in \mathbb{R}$ with $f(t) \geq at + c$ on $[0, \infty)$. Consequently Φ is bounded below on every set of densities having bounded L^1 mass. It is convex and norm lower semicontinuous on $L^1(q)$: from any norm-convergent sequence take an almost-everywhere convergent subsequence and apply Fatou's lemma to the nonnegative integrand $f(t) - at - c$. A norm-closed convex epigraph is weakly closed, so Φ is also weakly lower semicontinuous.

Every feasible density has unit mass by Lemma B.1. Let (z_n) be a minimizing sequence. Boundedness of $\Phi(z_n)$, bounded reward, and unit mass imply $\sup_n \int f(z_n)dq < \infty$. The nonnegative function $g(t) = f(t) - at - c$ is superlinear. The de la Vallée–Poussin criterion therefore makes (z_n) uniformly integrable; the Dunford–Pettis theorem yields a weakly convergent subsequence in $L^1(q)$. Its limit is nonnegative, and bounded linearity of B makes the equality $\mathsf{B}z = b$ weakly closed. Weak lower semicontinuity of Φ then proves primal attainment. Feasibility of z^0, the affine lower bound, and bounded reward show that the attained value is finite.

Proof: lower semicontinuity of the perturbation value. For $c \in L^1(\lambda)$ define

$$p(c) = \inf\{\Phi(z) : z \in L^1(q),\ \mathsf{B}z = c\}, \tag{107}$$

with empty fibers assigned $+\infty$. The affine lower bound and (98) imply $p(c) > -\infty$ on every nonempty fiber; condition CF makes $p(b) < \infty$. Hence p is proper and convex. It also has a continuous affine minorant. Indeed, with $R = \|r^+\|_\infty$,

$$\ell := \inf_{t\geq 0}\left\{\frac{1}{\eta}f(t) - Rt\right\} > -\infty \quad\Longrightarrow\quad p(c) \geq \ell \quad \text{for every } c \in L^1(\lambda). \tag{108}$$

The scalar infimum is finite by properness, lower semicontinuity, and superlinearity of f; the inequality follows pointwise for every finite $\Phi(z)$, while it is trivial on empty fibers.

Let $c_n \to b$ in $L^1(\lambda)$ and suppose $\liminf_n p(c_n) < \infty$. Pass to a subsequence along which $p(c_n)$ converges to $\liminf_j p(c_j)$, and relabel it by n. Choose z_n in the corresponding fibers such that

$$p(c_n) \leq \Phi(z_n) \leq p(c_n) + 1/n. \tag{109}$$

The mass identity in (98) gives

$$\int z_n\,dq = \frac{1}{1-\gamma}\int c_n\,d\lambda, \tag{110}$$

so the nonnegative z_n have uniformly bounded mass. Exactly as above, (109), bounded reward, and superlinearity give uniform integrability. Pass to $z_n \rightharpoonup \bar{z}$ in $L^1(q)$. Then $\mathsf{B}\bar{z} = b$ and

$$p(b) \leq \Phi(\bar{z}) \leq \liminf_n \Phi(z_n) = \liminf_n p(c_n). \tag{111}$$

Thus p is lower semicontinuous at b.

Proof: conjugate and dual. For $u \in L^\infty(\lambda)$,

$$\begin{aligned} p^*(u) &= \sup_{z\in L^1(q)}\{\langle u, \mathsf{B}z\rangle_\lambda - \Phi(z)\} \\ &= \sup_{z\in L^1(q)}\int_X \left[z(r^+ + \mathsf{B}^*u) - \frac{1}{\eta}f(z)\right] dq \\ &= \frac{1}{\eta}\int_X f^*\big(\eta[r^+ + \mathsf{B}^*u]\big)\,dq. \end{aligned} \tag{112}$$

The interchange in the last equality follows as below. The upper bound is pointwise Fenchel–Young. For the reverse bound, maximize the scalar integrand over $0 \leq z \leq n$. The smallest maximizer is a Borel function of the bounded measurable coefficient $r^+ + \mathsf{B}^*u$, and is itself bounded, hence belongs to $L^1(q)$. Letting $n \uparrow \infty$ and using monotone convergence after subtracting the integrable value at

$z = 1$ gives (112). Superlinearity makes f^* finite on $\mathbb{R}$, so the displayed integral is finite for bounded u. This is also the scalar special case of conjugacy for convex normal integrands (Rockafellar, 1968; 2006).

Because p is proper and has the continuous affine minorant (108), Fenchel–Moreau biconjugation on the dual pair $(L^1(\lambda), L^\infty(\lambda))$ gives $p^{**} = \operatorname{cl} p$. For convex epigraphs, norm and weak closure coincide. The pointwise lower semicontinuity (111) therefore yields

$$p(b) = p^{**}(b) = \sup_{u \in L^\infty(\lambda)} \{\langle u, b\rangle_\lambda - p^*(u)\}. \tag{113}$$

The value of (CP) is $-p(b)$. Set $u = -v$, use $\langle v, b\rangle = (1-\gamma)\int v\, d\mu_0^+$ and $r^+ - \mathsf{B}^* v = \bar{A}_v$ to obtain (CV). Finally $dq = \rho\, dm$ gives (105).

B.4 Attainment, structural zeros, and recovery

Proposition B.1 (Almost-everywhere Fenchel recovery)**.** *Let z^* solve* (CP)*. If $v^* \in L^\infty(\lambda)$ attains* (CV)*, then*

$$z^*(x) \in \arg\max_{z \geq 0} \{z\eta \bar{A}_{v^*}(x) - f(z)\} \quad q\text{-a.e.}, \tag{114}$$

or equivalently $\eta \bar{A}_{v^*}(x) \in \partial f(z^*(x))$ q-a.e. If the scalar maximizer is unique, it recovers the unique primal density q-almost everywhere and the recovered measure automatically satisfies weak Bellman flow.

Proof. Primal feasibility, equality of attained primal and dual values, and (104) give

$$\int_X \left[f(z^*) + f^*(\eta \bar{A}_{v^*}) - z^* \eta \bar{A}_{v^*}\right] dq = 0. \tag{115}$$

The integrand is nonnegative by Fenchel–Young, hence it vanishes q-almost everywhere. Equality in Fenchel–Young is exactly (114).

Proposition B.2 (Structural-zero boundary cases)**.** *Let $Z \subseteq X$ be Borel with $q(Z) > 0$ and suppose every $z \in \mathcal{Z}(q)$ satisfies $z = 0$ q-a.e. on Z.*

1. *If $f(0) < \infty$, the full primal contains the constant contribution $q(Z)f(0)/\eta$ in its penalty. The proof of Theorem B.1 already retains this term; deleting Z would define a different optimization problem unless the constant is carried explicitly.*
2. *If $f(0) = +\infty$, condition CF fails.*
3. *For shifted KL, $f(z) = z\log z - z + 1$, the dual cannot have a bounded minimizer: Proposition B.1 would require $z^* = \exp(\eta \bar{A}_{v^*}) > 0$ q-a.e., contradicting structural inactivity on Z.*
4. *Pearson recovery can represent $z^* = 0$ at a finite residual, but this pointwise fact alone does not guarantee existence of a global L^∞ Bellman-flow multiplier.*

B.5 Continuous return guarantees

The fixed-support improvement and local comparator arguments require only primal optimality and therefore extend without additional dual assumptions. For the fixed-support update, take $q = m = \bar{q} = d_k$ in Theorem B.1. If d_k is feasible, then $z \equiv 1$ satisfies weak flow and $\int f(1)\, dd_k = 0$, so condition CF holds and the theorem supplies an optimizer. Restricting the update to $d \ll d_k$, every such optimizer d_{k+1} satisfies

$$\int r^+\, dd_{k+1} \geq \int r^+\, dd_k + \frac{1}{\eta} D_f(d_{k+1} \| d_k). \tag{116}$$

For the anchored update, every comparator $d \in \mathcal{D}(q)$ with finite divergence satisfies

$$\int r^+\, dd_{k+1} \geq \int r^+\, dd - \frac{1}{\eta} D_f(d \| q). \tag{117}$$

The proofs are the same two-line optimality comparisons as in the finite theory; nonnegativity follows from Jensen because both measures are probabilities.

The finite uniform-KL estimate changes in continuous spaces. If $d \ll m$ and $D_{\mathrm{KL}}(d\|m) < \infty$, then $\rho = dq/dm \geq \varepsilon$ gives

$$D_{\mathrm{KL}}(d\|q) = \int \log\left(\frac{dd/dm}{dq/dm}\right) dd \leq D_{\mathrm{KL}}(d\|m) + \log\frac{1}{\varepsilon}. \tag{118}$$

For the augmentation-mismatch statement, fix any Borel extension $\widetilde{\pi}$ of an admissible augmented policy π to the original MDP outside E. On the canonical trajectory space under $\widetilde{\pi}$, map the first transition leaving E and the entire subsequent tail to $\perp$; the pushforward path has the augmented law and agrees with the original path before the exit. For $\tau_E = \inf\{t : S_t \notin E\}$ and $|r| \leq R_{\max}$,

$$|J(\widetilde{\pi}) - J^+(\pi)| \leq \frac{R_{\max}}{1-\gamma}\mathbb{E}_{\widetilde{\pi}}[\gamma^{\tau_E}\mathbf{1}\{\tau_E < \infty\}]. \tag{119}$$

Combining (89), (116)–(117), and (119) gives the same original-MDP inequalities as in the finite theorem, with integrals in place of sums.

B.6 Population sampling

The analysis uses a population coverage law m and exact Bellman flow on measurable state and action regions. Replay minibatches provide empirical estimates of expectations under this population law. When E and Γ are inferred from continuous samples, they are modeled as measurable regions and action correspondences. Appendix C.3 gives the numerical value and policy updates.

C Experimental and implementation details

C.1 Square-grid toy details

The toy in Figure 1 is a deterministic finite-MDP illustration of the augmented closure and carried-anchor mechanisms. The grid has size 9×9, discount $\gamma = 0.95$, start cell $s_0 = (7, 4)$, and actions $\mathcal{A} = \{\mathrm{up}, \mathrm{down}, \mathrm{left}, \mathrm{right}\}$. Boundary actions use the standard stay-at-wall convention. There are three terminal goal cells

$$g_1 = (5, 2), \qquad g_2 = (5, 6), \qquad g_3 = (1, 7),$$

with rewards $0.8, 16, 48$, respectively, on the source-action transitions whose next state is g_1, g_2, g_3. Once observed, goal cells are included in the replay-supported domain with zero-reward self-loops, following the usual discounted-MDP terminal convention.

The realized cumulative source-state support is

$$\begin{aligned}
\mathcal{S}_0 &= \{(7,4),(7,3),(6,3),(6,2),(5,2)\},\\
\mathcal{S}_1 &= \mathcal{S}_0 \cup \{(7,5),(6,5),(5,5),(5,6)\},\\
\mathcal{S}_2 &= \mathcal{S}_1 \cup \{(4,5),(3,5),(3,6),(2,6),(2,7),(1,7)\}.
\end{aligned}$$

At snapshot j, the environment replay atoms are $\Omega_j^{\mathrm{env}} = \mathcal{S}_j \times \mathcal{A}$, and the augmented MDP redirects transitions from $\mathcal{S}_j$ to states outside $\mathcal{S}_j$ to $\perp$.

The finite-theory base measure is $m_j = \mathrm{Unif}(\Omega_j^{\mathrm{aug}})$. The toy also defines an offline-heavy replay reference b_j on the same augmented domain. Every supported pair receives a small background weight 0.01; forward edges along the initial low-reward branch receive weight 280; forward edges along the middle branch receive weight 0.04; newly observed forward edges on the high branch receive weight 1.5; backward/noisy edges on the high branch receive weight 80; terminal self-loops receive weights 30 at g_1 and 0.10 at g_2, g_3; and synthetic $\perp$ actions receive weight 1. These weights are normalized to form b_j. This creates a finite offline-heavy replay distribution in which better online branches are present but rare.

Both occupancy rows in Figure 1 solve the same exact finite halved-Pearson occupancy program

$$\min_{d \geq 0} \frac{1}{2}\sum_{x \in \Omega_j^{\mathrm{aug}}} \frac{(d(x) - q(x))^2}{q(x)} - \eta \sum_x d(x) r_j^{\mathrm{aug}}(x), \qquad d \in \mathcal{D}_j^{\mathrm{aug}}.$$

We use $\eta = 18$ and $\varepsilon = 0.3$. SupportDICE sets $q_0 = b_0$ and, for $j > 0$, sets $q_j = (1-\varepsilon)\bar{q}_j + \varepsilon m_j$, where $\bar{q}_j$ is the previous exact SupportDICE occupancy zero-extended to the new support. The replay-centered DICE comparison instead sets $q_j = b_j$ at every snapshot.

All reported values are exact finite-MDP population returns. SupportDICE obtains

$$0.669420, \quad 7.610584, \quad 27.007902,$$

and its greedy path from s_0 reaches g_1, then g_2, then g_3 as support expands. Replay-centered DICE obtains

$$0.669420, \quad 0.769936, \quad 1.254903,$$

and its greedy path remains on the g_1 branch at all three snapshots. The Bellman-flow residual of the reported occupancy solutions is below 10^{-12} in ℓ_∞ norm.

C.2 Benchmark protocol and reporting

We use six AntMaze-v2 datasets, the four Adroit cloned-v1 datasets, and Kitchen partial, mixed, and complete-v0. Every numeric performance cell contains all four seeds $\{0, 1, 2, 3\}$. *Final* is each run's evaluation at the prescribed online endpoint. We report the mean and sample standard deviation across seeds (denominator $n-1$). Tables report two-decimal summary precision; ties at that precision share the same formatting.

For AntMaze and Adroit, the Cal-QL, AWAC, IQL, CQL, and SPOT entries use CORL fine-tuning runs with 1M offline updates and 1M online environment steps. Kitchen uses the shared protocol below, 500k offline updates, and 1.25M online steps.

RLPD uses its released implementation.

C.3 Policy implementation

The population construction determines an occupancy target independently of a policy parameterization. We realize the target using an occupancy-weighted flow teacher and a Q-guided one-step student, following the policy architecture of Flow Q-Learning (FQL) (Park et al., 2025). The DICE value solver is separate from the TD-Q critic and retains the ODICE forward/backward orthogonal-gradient surrogate (Mao et al., 2024b).

Teacher and student objectives. Let p be the action dimension, $\|y\|_{\mathrm{av}}^2 = p^{-1}\sum_{j=1}^p y_j^2$, $z \sim \mathcal{N}(0, I)$, $t \sim \mathrm{Unif}[0,1]$, and $x_t = (1-t)z + ta$. Write $\widehat{m}_k$ for uniform sampling over replay rows, and let h_k be the processed replay weights defined below. The teacher velocity field u_θ minimizes

$$\mathcal{L}_T = \mathbb{E}_{(s,a)\sim\widehat{m}_k, z, t}\left[\mathrm{sg}(h_k(s,a))\|u_\theta(s, x_t, t) - (a - z)\|_{\mathrm{av}}^2\right]. \tag{120}$$

Here sg stops gradients through the weights. Starting from the same noise z, ten Euler steps generate the teacher action $a_T(s, z)$. The student minimizes

$$\mathcal{L}_S = \alpha\,\mathbb{E}_{s\sim\widehat{m}_{k,S},z}\|a_\phi(s,z) - \mathrm{sg}(a_T(s,z))\|_{\mathrm{av}}^2 - \mathbb{E}_{s\sim\widehat{m}_{k,S},z} Q_{\mathrm{guide}}(s, \mathrm{clip}(a_\phi(s,z))). \tag{121}$$

Clipping uses the environment action bounds. Only the teacher regression is directly weighted by h_k; the distillation, Q-guidance, and critic TD objectives are not. The deployed policy generates one student action in one forward pass. If $h = d/m$ is an exact density ratio, the change-of-measure identity $\mathbb{E}_m[h\ell] = \mathbb{E}_d[\ell]$ explains the occupancy-weighted teacher objective. Fitting and integrating the flow and distilling the Q-guided student are the practical policy-realization steps.

Replay mass and refresh. The replay buffer retains all offline and online transitions without eviction. For N rows, the uniform sampling probability is $1/N$. Store a carried probability mass $\bar{d}_i$ for each row; appending a row leaves old masses unchanged and gives the new row mass zero. The reference-to-sampling ratio is $(1-\varepsilon)N\bar{d}_i + \varepsilon$. We use $\varepsilon = 0.3$, $\eta_{\mathrm{tr}} = 1$, and the *unhalved* Pearson generator $f(z) = (z-1)^2$, whose constrained recovery is $[1 + \eta A/2]_+$. Every 5,000 online transitions, a full-buffer refresh computes

$$u_i = \left[(1-\varepsilon)N\bar{d}_i + \varepsilon\right] \min_j [1 + \eta_{\mathrm{tr}}\widehat{A}_{V_j,i}/2]_+, \tag{122}$$

Table 2: Policy-learning recipes. TD target aggregation and student guidance are distinct. Kitchen guidance uses its bounded critic divided by 400.

Task group	α	TD target	Q_{guide}
AntMaze umaze / medium	10	twin-Q mean	twin-Q mean
AntMaze large	3	twin-Q mean	twin-Q mean
Adroit pen / door / hammer	1000	twin-Q minimum	twin-Q mean
Adroit relocate	10000	twin-Q mean	twin-Q mean
Kitchen	10	twin-Q minimum	twin-Q minimum / 400

where $\widehat{A}_{V_j,i}$ is the sampled residual from DICE value estimate j, using its reward convention and terminal mask. The resulting weights define an empirical occupancy estimate.

When the full-buffer mean of u exceeds 10^{-12}, first divide u by that mean. A global bisection then selects a scale c for

$$h_i = \min\{3, cu_i\}, \qquad \frac{1}{N}\sum_{i=1}^{N} h_i = 1, \qquad \widehat{d}_i = h_i/N. \tag{123}$$

If all raw weights are near zero, use $h_i = 1$. If the positive support is too small to meet both the cap and unit-mean requirement, floor every weight at 3×10^{-6}, normalize globally, and apply the same scale-and-cap procedure. A minibatch teacher update uses $B^{-1}\sum_{i\in B} h_i \ell_i$, with no minibatch weight renormalization. Between refreshes, the teacher reads $h_i = N\widehat{d}_i$ at the current buffer size. Buffer growth can therefore temporarily raise an old row's ratio above three; the cap applies at refresh, not at every update. New rows receive teacher mass at a subsequent refresh.

Initialization and networks. The flow teacher, one-step student, and each TD-Q network use four hidden layers of width 512, with LayerNorm in Q but not in the flow actor. Their Adam learning rate is 3×10^{-4}. DICE value networks use three hidden layers of width 256, ReLU activations, learning rate 10^{-4}, and target-update coefficient 0.005. Batch size is 256 and discount is 0.99. The policy learner is first pretrained on offline data with unweighted flow regression and its own Q/student objectives. Online fine-tuning restores its full agent and optimizer state together with a separately pretrained DICE solver. For AntMaze and Adroit, the DICE loader imports value, target-value, and available Adam state; replay mass starts uniformly and is refreshed from the loaded values before online learning. Kitchen instead restores a reward-compatible DICE snapshot including its mass and solver state, then switches the online anchor mixture to 0.3.

AntMaze uses rewards $-1/0$ for the policy learner and $-3/0$ for the DICE surrogate.

The DICE value solver uses $\lambda = 0.6$ on AntMaze umaze, 0.4 on AntMaze umaze-diverse, and 0.7 on all other tasks. The backward-gradient scale is 1.0 on AntMaze umaze and umaze-diverse, 0.2 on AntMaze medium-play, large-play, large-diverse, and Adroit hammer, 0.6 on Adroit door, and 0.4 on all other tasks.

C.4 Kitchen reward, termination, and value protocol

Kitchen uses native 60-dimensional observations, nine-dimensional actions, and a maximum horizon of 280 steps. Following the WSRL-style reward protocol (Zhou et al., 2025), offline rewards are the stored dataset reward minus four; online rewards are the number of subtasks completed so far minus four. Offline transitions retain the standard Q-learning dataset rows and source terminal flags. Online episodes terminate after completing all four subtasks; time-limit truncations retain bootstrapping.

Evaluation reports the final completed-subtask count multiplied by 25. All reported Kitchen baselines share this environment/reward/evaluation adapter and retain their own learners.

For SupportDICE's policy learner, $r \in [-4, 0]$ and $\gamma = 0.99$ motivate the discounted value range $[-400, 0]$. We implement a bounded sigmoid Q head, clip TD targets to this interval, use twin-Q minimum for both targets and actor guidance, and divide the guidance term by 400. These are SupportDICE-specific stabilization choices. DICE initialization uses compatible shifted rewards.

C.5 Relation to the population guarantees

Theorems 1–3 and Appendix B analyze exact population occupancy updates. Our implementation uses sampled DICE backups, processed replay weights, and parameterized policy learning. The absorbing state enters the population formulation; the numerical update uses the replay losses specified above. Appendix B.6 describes the population sampling setting.

C.6 Full benchmark results

Table 3 reports scores before and after online fine-tuning, as mean and sample standard deviation over four seeds.

Table 3: Scores before and after online fine-tuning on D4RL. Each cell reports the score before fine-tuning on the first line and the final score on the second line (preceded by $\rightarrow$), both as mean $\pm$ sample standard deviation over four seeds. Bold marks the highest final mean and underline the second distinct final mean in each task; ties share formatting. See Appendix C.2 for the evaluation protocols.

Dataset	SupportDICE	Cal-QL	AWAC	IQL	CQL	SPOT	RLPD	PEX
AntMaze								
	96.00 ± 1.63	76.75 ± 8.69	52.75 ± 10.01	77.00 ± 0.82	94.00 ± 1.83	91.00 ± 2.94	0.00 ± 0.00	92.50 ± 4.12
umaze	$\rightarrow \mathbf{100.00} \pm 0.00$	$\rightarrow \underline{99.75} \pm 0.50$	$\rightarrow 98.75 \pm 1.26$	$\rightarrow 96.50 \pm 1.29$	$\rightarrow 99.50 \pm 1.00$	$\rightarrow 99.50 \pm 0.58$	$\rightarrow 98.00 \pm 0.00$	$\rightarrow 95.50 \pm 6.40$
umaze-	82.00 ± 17.28	32.00 ± 32.09	56.00 ± 3.16	59.50 ± 11.03	9.50 ± 11.45	36.25 ± 2.50	0.00 ± 0.00	59.00 ± 9.31
diverse	$\rightarrow \mathbf{100.00} \pm 0.00$	$\rightarrow 98.50 \pm 1.29$	$\rightarrow 0.00 \pm 0.00$	$\rightarrow 63.75 \pm 28.89$	$\rightarrow \underline{99.00} \pm 1.41$	$\rightarrow 95.00 \pm 4.24$	$\rightarrow 97.00 \pm 3.46$	$\rightarrow 63.00 \pm 22.77$
medium-	78.50 ± 6.61	71.75 ± 3.77	0.00 ± 0.00	71.75 ± 3.40	59.00 ± 12.91	67.25 ± 12.09	0.00 ± 0.00	58.50 ± 4.43
play	$\rightarrow 96.00 \pm 2.83$	$\rightarrow \mathbf{98.75} \pm 1.89$	$\rightarrow 0.00 \pm 0.00$	$\rightarrow 89.75 \pm 1.26$	$\rightarrow \underline{97.75} \pm 1.50$	$\rightarrow 97.25 \pm 1.50$	$\rightarrow 96.50 \pm 3.00$	$\rightarrow 85.00 \pm 6.83$
medium-	63.00 ± 9.87	62.00 ± 4.97	0.00 ± 0.00	64.25 ± 2.22	63.50 ± 7.90	73.75 ± 8.42	0.00 ± 0.00	58.50 ± 11.47
diverse	$\rightarrow 96.00 \pm 3.27$	$\rightarrow \underline{98.25} \pm 1.71$	$\rightarrow 0.00 \pm 0.00$	$\rightarrow 92.25 \pm 3.30$	$\rightarrow 97.25 \pm 2.22$	$\rightarrow 94.50 \pm 1.91$	$\rightarrow \mathbf{98.50} \pm 1.91$	$\rightarrow 87.00 \pm 6.83$
large-	62.00 ± 41.54	31.75 ± 10.24	0.00 ± 0.00	38.50 ± 10.08	28.75 ± 8.96	31.50 ± 14.53	0.00 ± 0.00	42.00 ± 5.66
play	$\rightarrow \underline{96.50} \pm 3.00$	$\rightarrow \mathbf{97.25} \pm 2.06$	$\rightarrow 0.00 \pm 0.00$	$\rightarrow 64.50 \pm 19.67$	$\rightarrow 88.25 \pm 2.63$	$\rightarrow 87.00 \pm 3.74$	$\rightarrow 85.50 \pm 4.12$	$\rightarrow 61.00 \pm 6.22$
large-	86.00 ± 9.38	44.00 ± 10.03	0.00 ± 0.00	26.75 ± 4.35	35.50 ± 4.20	17.50 ± 8.39	0.00 ± 0.00	44.50 ± 4.12
diverse	$\rightarrow \mathbf{94.50} \pm 3.00$	$\rightarrow 91.50 \pm 4.51$	$\rightarrow 0.00 \pm 0.00$	$\rightarrow 64.25 \pm 4.79$	$\rightarrow \underline{91.75} \pm 4.57$	$\rightarrow 81.00 \pm 16.33$	$\rightarrow 84.00 \pm 3.65$	$\rightarrow 73.00 \pm 9.59$
Adroit cloned								
	56.47 ± 10.03	-2.66 ± 0.05	88.66 ± 17.44	84.19 ± 4.57	-2.76 ± 0.09	6.19 ± 6.01	0.63 ± 3.69	43.32 ± 39.54
pen	$\rightarrow \mathbf{147.32} \pm 10.96$	$\rightarrow -2.68 \pm 0.14$	$\rightarrow 86.82 \pm 12.84$	$\rightarrow 102.02 \pm 23.96$	$\rightarrow -1.28 \pm 2.49$	$\rightarrow 43.63 \pm 23.20$	$\rightarrow \underline{119.28} \pm 26.39$	$\rightarrow 94.81 \pm 21.51$
	0.08 ± 0.04	-0.33 ± 0.01	0.93 ± 1.92	1.19 ± 1.08	-0.33 ± 0.01	-0.21 ± 0.16	-0.05 ± 0.02	-0.10 ± 0.01
door	$\rightarrow \mathbf{99.50} \pm 6.19$	$\rightarrow -0.33 \pm 0.01$	$\rightarrow 0.01 \pm 0.01$	$\rightarrow 20.34 \pm 10.76$	$\rightarrow -0.33 \pm 0.01$	$\rightarrow 0.02 \pm 0.36$	$\rightarrow \underline{81.12} \pm 15.59$	$\rightarrow 65.06 \pm 9.85$
	0.17 ± 0.07	0.25 ± 0.04	1.80 ± 3.47	1.35 ± 0.37	0.56 ± 0.63	3.97 ± 7.38	0.20 ± 0.03	1.77 ± 1.41
hammer	$\rightarrow \underline{136.45} \pm 6.25$	$\rightarrow 0.17 \pm 0.19$	$\rightarrow 0.24 \pm 0.04$	$\rightarrow 57.27 \pm 32.89$	$\rightarrow 2.85 \pm 5.56$	$\rightarrow 3.73 \pm 5.76$	$\rightarrow \mathbf{137.87} \pm 6.80$	$\rightarrow 64.07 \pm 18.59$
	0.11 ± 0.12	-0.31 ± 0.06	-0.04 ± 0.04	0.04 ± 0.05	-0.33 ± 0.01	-0.24 ± 0.01	-0.01 ± 0.03	-0.07 ± 0.03
relocate	$\rightarrow \mathbf{73.79} \pm 11.87$	$\rightarrow -0.31 \pm 0.04$	$\rightarrow -0.04 \pm 0.02$	$\rightarrow \underline{0.32} \pm 0.44$	$\rightarrow -0.33 \pm 0.01$	$\rightarrow -0.15 \pm 0.06$	$\rightarrow -0.04 \pm 0.04$	$\rightarrow -0.01 \pm 0.21$
Kitchen								
	36.88 ± 5.15	13.33 ± 11.27	30.00 ± 6.12	27.50 ± 10.61	0.00 ± 0.00	6.88 ± 12.14	6.25 ± 12.50	62.50 ± 9.79
partial	$\rightarrow \mathbf{91.88} \pm 8.98$	$\rightarrow 8.33 \pm 14.43$	$\rightarrow 0.00 \pm 0.00$	$\rightarrow \underline{75.00} \pm 0.00$	$\rightarrow 0.00 \pm 0.00$	$\rightarrow 24.38 \pm 9.87$	$\rightarrow 65.00 \pm 13.99$	$\rightarrow 54.38 \pm 15.33$
	54.38 ± 5.91	22.50 ± 5.00	62.50 ± 5.40	52.50 ± 3.54	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	54.38 ± 5.54
mixed	$\rightarrow 66.25 \pm 9.46$	$\rightarrow 8.33 \pm 14.43$	$\rightarrow \mathbf{75.00} \pm 0.00$	$\rightarrow 65.00 \pm 8.90$	$\rightarrow 0.83 \pm 1.44$	$\rightarrow 15.00 \pm 12.91$	$\rightarrow \underline{73.75} \pm 1.44$	$\rightarrow 59.38 \pm 8.00$
	28.75 ± 14.51	10.00 ± 7.50	51.67 ± 10.41	26.67 ± 18.43	0.00 ± 0.00	0.00 ± 0.00	2.50 ± 5.00	50.63 ± 11.25
complete	$\rightarrow \underline{83.75} \pm 11.09$	$\rightarrow 0.00 \pm 0.00$	$\rightarrow 28.33 \pm 25.66$	$\rightarrow 23.33 \pm 22.55$	$\rightarrow 0.00 \pm 0.00$	$\rightarrow 7.50 \pm 12.99$	$\rightarrow \mathbf{90.63} \pm 12.31$	$\rightarrow 21.88 \pm 29.68$