Automatic Constraint Policy Optimization based on Continuous Constraint Interpolation Framework for Offline Reinforcement Learning

Xinchen Han*
SAMOVAR, Télécom SudParis
Institut Polytechnique de Paris
Paris, France

Qiuyang Fang
College of Artificial Intelligence
Nankai University
Tianjin, China

Hossam Afifi
SAMOVAR, Télécom SudParis
Institut Polytechnique de Paris
Paris, France

Michel Marot
SAMOVAR, Télécom SudParis
Institut Polytechnique de Paris
Paris, France

Abstract

Offline Reinforcement Learning (RL) relies on policy constraints to mitigate extrapolation error, where both the constraint form and constraint strength critically shape performance. However, most existing methods commit to a single constraint family: weighted behavior cloning, density regularization, or support constraints, without a unified principle that explains their connections or trade-offs. In this work, we propose Continuous Constraint Interpolation (CCI), a unified optimization framework in which these three constraint families arise as special cases along a common constraint spectrum. The CCI framework introduces a single interpolation parameter that enables smooth transitions and principled combinations across constraint types. Building on CCI, we develop Automatic Constraint Policy Optimization (ACPO), a practical primal–dual algorithm that adapts the interpolation parameter via a Lagrangian dual update. Moreover, we establish a maximum-entropy performance difference lemma and derive performance lower bounds for both the closed-form optimal policy and its parametric projection. Experiments on D4RL and NeoRL2 demonstrate robust gains across diverse domains, achieving state-of-the-art performance overall.

1 Introduction

RL tackles sequential decision-making through online interaction with an environment Han et al. [2024]. However, in many real-world applications, such interaction can be expensive, risky, or even infeasible. To address this limitation, offline RL, which aims to learn effective policies from pre-collected datasets without further interaction, has received growing attention in recent years. Yet, removing interaction also introduces fundamental challenges. In particular, extrapolation error Fujimoto et al. [2019] arises when Q-function estimates are queried on out-of-distribution (OOD) actions that are poorly supported by the dataset. Such errors can be amplified by bootstrapping, leading to severe overestimation and, ultimately, degraded performance and instability Levine et al. [2020], Kumar et al. [2020], Han et al. [2025d], Kim et al. [2025].

A dominant paradigm of offline RL methods mitigates extrapolation error by constraining the learned policy to stay close to the behavior policy that generated the dataset. In practice, however, the

*Corresponding author: xinchen.han@telecom-sudparis.eu.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

behavior policy is often unknown: offline data may come from random exploration, mixtures of multiple tasks, heterogeneous sources, or logged trajectories collected during online RL training Chen et al. [2022], Lyu et al. [2022]. Under such ambiguity, selecting an appropriate form and strength of policy constraint becomes crucial.

From a high-level perspective, mainstream policy-constrained approaches in offline RL can be organized into three categories: weighted behavior cloning (wBC), density regularization, and support constraints. wBC remains fundamentally imitative and typically reduces to weighted supervised learning on the dataset Chen et al. [2020], Florence et al. [2022], Xu et al. [2022], Fujimoto and Gu [2021]. Density regularization allows controlled distributional deviation from the behavior policy. In particular, KL-based density regularization is widely used to limit the magnitude of distributional shift from the behavior distribution Peters and Schaal [2007], Peng et al. [2019], Nair et al. [2020], Wang et al. [2020]. Support constraints, instead of matching the behavior distribution, encourage the learned policy to select actions within the support of the behavior policy Kumar et al. [2019], Han et al. [2025c], Wu et al. [2022], Mao et al. [2023]. However, overly conservative constraints can stabilize policy evaluation but undermine the core promise of offline RL—policy improvement beyond the dataset; while overly weak constraints fail to meaningfully restrict policy improvement, tending to select OOD actions and inducing large value misestimation that misguides optimization Wu et al. [2025], Han et al. [2025a].

To cope with this trade-off, several recent works Zhou et al. [2022], Luo et al. [2025], Han et al. [2025c] explore adaptive constraint schemes. However, existing approaches typically adapt within a single constraint paradigm, and a unified view that simultaneously covers and connects wBC, KL-based density regularization, and support constraints remains missing. Consequently, current methods still rely on task-dependent manual choices of constraint form and strength, rather than adapting them in a unified and principled way.

In this paper, we develop a unified optimization framework, the *Continuous Constraint Interpolation (CCI)* framework, and show that the above three constraint families arise as special cases. The CCI framework exposes a single interpolation parameter that enables smooth transitions and principled combinations among wBC, KL-based density regularization, and support constraints for policy improvement. Building on the CCI framework, we propose *Automatic Constraint Policy Optimization (ACPO)*, a practical offline RL algorithm that optimizes the interpolation parameter via a Lagrangian dual update, thereby eliminating the need to manually select fixed constraint forms and strengths. Intuitively, the interpolation parameter acts as a constraint "dial" that automatically adapts the constraints regime during training.

Our key contributions can be summarized as follows:

- **Unified CCI Framework.** We propose CCI, a unified optimization framework that bridges three major policy-constraint families: wBC, KL-based density regularization, and support constraints, and enables smooth transitions as well as principled combinations across constraint regimes.
- **Practical ACPO Algorithm.** Building on CCI, we develop ACPO, a practical primal–dual algorithm that automatically adapts the interpolation parameter through a Lagrangian dual update. We further establish a maximum-entropy performance difference lemma and derive performance lower bounds for both the optimal policy and its practical parametric projection.
- **Empirical Evaluation.** We perform extensive comparisons and ablations on the standard D4RL and the recent NeoRL2 benchmarks, showing that ACPO achieves consistent improvements across diverse domains and tasks.

2 Preliminaries

2.1 Maximum Entropy RL

Conventionally, an RL problem is defined as a Markov Decision Process (MDP) Sutton and Barto [2018], specified by a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $\mathcal{T}$ represents the transition function, r is the reward function, and $\gamma \in [0, 1)$ denotes the discount factor. The performance difference lemma Kakade and Langford [2002] in the standard RL setting

gives

$$\eta(\pi') - \eta(\pi) = \frac{1}{1-\gamma} \sum_s d_{\pi'}(s) \sum_a \pi'(a|s)\, A^{\pi}(s,a), \tag{1}$$

where $\eta(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$, and $\mathbb{E}_{\tau\sim\pi}[...]$ denotes the expectation under the trajectory τ induced by the interconnection of π and the environment. $d_{\pi'}(s)$ denotes the discounted state visitation distribution under π', and $A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$ is the advantage function.

In the maximum-entropy setting Haarnoja et al. [2017], Asadi and Littman [2017], Haarnoja et al. [2018], Xiao et al. [2019], the objective augments rewards with an entropy term:

$$\mathcal{J}(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t \Big(r(s_t, a_t) - \alpha \log \pi(a_t|s_t) \Big)\right], \tag{2}$$

where $\alpha > 0$ is the temperature parameter and $\mathcal{H}(s,\pi) = \mathbb{E}_{a\sim\pi(\cdot|s)}[-\log \pi(a|s)]$. The maximum-entropy value function is defined as $V^{\pi}(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^{\pi}(s,a) - \alpha \log \pi(a|s)\right]$. The soft Q-function is $Q^{\pi}(s,a) = r(s,a) + \gamma \mathbb{E}_{s'\sim\mathcal{T}(\cdot|s,a)}\left[V^{\pi}(s')\right]$.

2.2 wBC Methods

The generalized wBC objective can be defined as follows,

$$\max_{\pi} \mathcal{J}^{\text{wBC}}(\pi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[w^{\text{bc}}(s,a) \log \pi(a|s)\right], \tag{3}$$

where $w^{\text{bc}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ denotes a weight function. For example, $w^{\text{bc}}(s,a) = 1$ corresponds to the vanilla BC method, while more sophisticated choices are used in, e.g., Xu et al. [2022], Chen et al. [2020].

2.3 KL-based Density Regularization Methods

KL-based density regularization methods can be summarized by the following policy optimization problem Mao et al. [2023]:

$$\pi^*(a|s) = \arg\max_{\pi}\ \mathbb{E}_{a\sim\pi(\cdot|s)}\left[A^{\pi_{\text{pe}}}(s,a)\right] \quad \text{s.t. } D_{\text{KL}}[\pi(\cdot|s)\|\pi_{\text{base}}(\cdot|s)] \le \epsilon,\ \int_a \pi(a|s)\, da = 1. \tag{4}$$

where π_{base} denotes a reference (sampling) policy and π_{pe} denotes the evaluation policy. The optimization problem Eq. (4) admits a parametric solution, which can be written as

$$\max_{\pi} \mathcal{J}^{\text{KL}}(\pi) = \mathbb{E}_{s\sim\mathcal{D}, a\sim\pi_{\text{base}}}\left[w^{\text{kl}}(s,a) \log \pi(a|s)\right], \tag{5}$$

where $w^{\text{kl}}(s,a) = \exp\left(\frac{A^{\pi_{\text{pe}}}(s,a)}{\alpha}\right)$. Different choices of π_{base} and π_{pe} give rise to different algorithmic instantiations Nair et al. [2020], Wang et al. [2020], Siegel et al. [2020], while sharing the same KL-regularized principle.

2.4 Support Constraint Methods

The support constraint requires the learned policy to assign positive density only on actions supported by the behavior policy, e.g., $\pi_\beta(a|s) = 0 \ \Rightarrow\ \pi(a|s) = 0$.

InAC Xiao et al. [2023] introduces an in-sample softmax operator and implements support-constrained policy improvement by sampling actions from the dataset:

$$\max_{\pi} \mathcal{J}^{\text{Supp}}(\pi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[w^{\text{supp}}(s,a) \log \pi(a|s)\right], \tag{6}$$

where $w^{\text{supp}}(s,a) = \exp\left(\frac{A^{\pi}(s,a)}{\alpha} - \log \pi_\beta(a|s)\right)$.

Because $\exp(-\log \pi_\beta(a|s)) = \pi_\beta(a|s)^{-1}$, the above is equivalent to $\pi(a|s) = 0$ when $\pi_\beta(a|s) = 0$ and otherwise $\pi(a|s) \propto \exp(A^{\pi}(s,a)/\alpha)$ [2]. The learned policy π is not skewed by the action probabilities in π_β; it just has the same support.

[2] InAC defines $0 \cdot \infty = 0$.

3 Unified Policy Constraints for Offline RL

3.1 Continuous Constraint Interpolation Framework

The CCI framework provides a unified optimization view of three policy-constraint families: wBC, KL-based density regularization, and support constraints. It treats them as instances along a common constraint spectrum, and enables smooth switching and principled combinations via a single continuous interpolation parameter.

Consider the following constrained maximum-entropy policy optimization problem,

$$\begin{aligned} \max_{\pi} \ & \mathbb{E}_{s\sim\mathcal{D},\, a\sim\pi(\cdot|s)}\big[Q(s,a) - \alpha \log \pi(a|s)\big], \\ \text{s.t.} \ & \mathbb{E}_{s\sim\mathcal{D},\, a\sim\pi(\cdot|s)}\big[\log \pi_\beta(a|s)\big] \geq \epsilon, \\ & \int \pi(a|s)\, da = 1, \ \forall s. \end{aligned} \tag{7}$$

We do not impose per-state hard constraints, which would introduce an impractical (potentially infinite) number of constraints Wu et al. [2022]. Instead, we enforce the constraint on average over dataset states, yielding a single scalar constraint.

Solving the above problem Eq. (7) via the Lagrangian,

$$\begin{aligned} \mathcal{L}(\pi, \lambda, \nu) = \mathbb{E}_{s\sim\mathcal{D},\, a\sim\pi(\cdot|s)}\left[Q(s,a) - \alpha \log \pi(a|s)\right] + \lambda\Big(\mathbb{E}_{s\sim\mathcal{D},\, a\sim\pi(\cdot|s)}\left[\log \pi_\beta(a|s)\right] - \epsilon\Big) \\ + \mathbb{E}_{s\sim\mathcal{D}}\left[\nu(s)\Big(\int \pi(a|s)\, da - 1\Big)\right], \end{aligned} \tag{8}$$

we obtain the following closed-form optimal policy:

$$\pi^*_\lambda(a|s) = \exp\Big(\frac{Q(s,a) - Z_\lambda(s)}{\alpha}\Big)\pi_\beta(a|s)^{\lambda/\alpha}. \tag{9}$$

where $\lambda \geq 0$ denotes the dual variable, and $Z_\lambda(s) = \alpha \log \int \exp(Q(s,a)/\alpha)\, \pi_\beta(a|s)^{\lambda/\alpha}\, da$ is the log-normalizer.

Then, we project the non-parametric optimal policy π^*_λ into the parametric space π_θ by minimizing the reverse KL divergence,

$$\pi_\theta(a|s) = \arg\min_{\pi_\theta} \mathbb{E}_{s\sim\mathcal{D}}\big[\, D_{\mathrm{KL}}\left[\pi^*_\lambda(\cdot|s)||\pi_\theta(\cdot|s)\right]\big]. \tag{10}$$

This yields the following policy improvement objective under CCI:

$$\max_{\theta} \ \mathcal{J}^{\mathrm{CCI}}(\theta; \lambda) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[w_\lambda(s,a) \log \pi_\theta(a|s)\right], \tag{11}$$

where $w_\lambda(s,a) = \exp\Big(\frac{A(s,a)}{\alpha} + \frac{\lambda-\alpha}{\alpha}\log \pi_\beta(a|s)\Big)$.

Based on the policy improvement objective Eq. (11), varying λ enables continuous interpolation of the constraint spectrum, allowing smooth transitions and flexible combinations, shown in Tab. 1.

λ	Constraint	$w_\lambda(s,a)$
$\lambda = 0$	Support Constraint	$\exp\big(\frac{A(s,a)}{\alpha} - \log \pi_\beta(a\|s)\big)$.
$0 < \lambda < \alpha$	Support-to-Density Constraint	$\exp\big(\frac{A(s,a)}{\alpha} - \frac{\alpha-\lambda}{\alpha}\log \pi_\beta(a\|s)\big)$.
$\lambda = \alpha$	KL-based Density Constraint	$\exp\big(\frac{A(s,a)}{\alpha}\big)$.
$\lambda > \alpha$	Density-to-wBC Constraint	$\exp\big(\frac{A(s,a)}{\alpha} + \frac{\lambda-\alpha}{\alpha}\log \pi_\beta(a\|s)\big)$.
$\lambda \gg \alpha + A(s,a)(\log \pi_\beta(a\|s))^{-1}$	Practical wBC Constraint	$\approx \exp\big(\frac{\lambda-\alpha}{\alpha}\log \pi_\beta(a\|s)\big)$.

Table 1: Constraint spectrum induced by the interpolation parameter λ in the CCI framework.

The derivations for this subsection are provided in Appendix B.

3.2 Automatic Constraint Policy Optimization

Building on the CCI framework, we propose ACPO, a practical primal–dual method that automatically tunes the constraint form and strength parameter λ.

For a fixed Q-function, the objective in Eq. (7) is maximizing a concave objective in π, and the inequality constraint and normalization constraints are affine. Therefore, under standard regularity assumptions (ensuring $\mathcal{L}$ is well-defined) and Slater's condition, strong duality holds, and the primal problem is equivalent to the dual form,

$$\begin{aligned} &\min_{\lambda} \max_{\pi_\theta \in \Delta} \mathcal{L}(\pi_\theta, \lambda), \\ &\text{s.t. } \lambda \geq 0, \end{aligned} \tag{12}$$

where Δ denotes the set of valid policies.

ACPO alternates between (i) updating π_θ to maximize the CCI objective Eq. (11) for a fixed λ, and (ii) performing a projected gradient descent step on λ to enforce the constraint:

$$\lambda \leftarrow \Big[\lambda - \eta_\lambda \big(\mathbb{E}_{s\sim\mathcal{D},\, a\sim\pi_\theta}[\log \pi_\beta(a|s)] - \epsilon\big)\Big]_+. \tag{13}$$

This update increases λ when the constraint is violated, and decreases it when the constraint is overly conservative. By updating λ online via Eq. (13), ACPO automatically adapts the effective constraint regime to the dataset.

For policy evaluation, we adopt standard maximum-entropy critic updates with Clipped Double Q-Learning Fujimoto et al. [2018]:

$$\mathcal{L}_V(\psi) = \mathbb{E}_{s\sim\mathcal{D}, a\sim\pi_\theta(\cdot|s)}\left[\frac{1}{2}\big(V_\psi(s) - \min_{i\in\{1,2\}} (Q_{\phi_i}(s,a) - \alpha \log \pi_\theta(a|s))\big)^2\right], \tag{14}$$

$$\mathcal{L}_Q(\phi_i) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\frac{1}{2}\left(r + \gamma V_\psi(s') - Q_{\phi_i}(s,a)\right)^2\right]. \tag{15}$$

Following common practice in offline RL Xiao et al. [2023], Han et al. [2025c], we fit the behavior policy π_β on the dataset via maximum likelihood:

$$\mathcal{L}(\pi_\beta) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[-\log \pi_\beta(a|s)\right]. \tag{16}$$

Integrating policy improvement, policy evaluation, and dual optimization, the complete ACPO algorithm is summarized in Algorithm. 1.

Algorithm 1 ACPO

Input: Dataset $\mathcal{D} = \{(s, a, r, s')\}$.
Parameters: Q-networks $Q_{\phi_{1,2}}$, target Q-networks $Q_{\phi'_{1,2}}$, value network V_ψ, policy network π_θ, behavior policy network π_β.

// Pre-train behavior policy for N_β steps.
for $t = 1$ to N_β **do**
    Optimize π_β by Eq. (16).
end for
// Policy evaluation, improvement, and dual update.
for each gradient step **do**
    Sample $(s, a, r, s') \sim \mathcal{D}$.
    Optimize V_ψ by Eq. (14).
    Optimize Q_{ϕ_i} by Eq. (15).
    Update λ by Eq. (13).
    Optimize π_θ by Eq. (11).
    Soft-update target networks $\phi' \leftarrow (1-\tau)\phi' + \tau\phi$.
end for
return Policy π_θ.

3.3 Theoretical Analysis

The CCI framework admits a closed-form optimizer π^*_λ parameterized by the dual variable λ. We first formalize how λ controls conservatism by studying the constraint $\mathbb{E}_{a\sim\pi^*_\lambda(\cdot|s)}[\log \pi_\beta(a|s)]$. We then derive a maximum-entropy performance difference lemma and establish performance lower bounds for both the optimal policy π^*_λ and its parametric projection π_θ.

Proposition 3.1. *Fix any state s and temperature $\alpha > 0$. Let $\pi^*_\lambda(\cdot|s)$ denote the closed-form optimizer in Eq. (9) for a given $\lambda \geq 0$. Define*

$$g_s(\lambda) := \mathbb{E}_{a\sim\pi^*_\lambda(\cdot|s)}\big[\log \pi_\beta(a|s)\big]. \tag{17}$$

Then $g_s(\lambda)$ is monotone non-decreasing in λ, and

$$\frac{d}{d\lambda} g_s(\lambda) = \frac{1}{\alpha}\,\mathrm{Var}_{a\sim\pi^*_\lambda(\cdot|s)}\big(\log \pi_\beta(a|s)\big) \;\geq\; 0. \tag{18}$$

Consequently, the constraint $g(\lambda) := \mathbb{E}_{s\sim\mathcal{D}}[g_s(\lambda)]$ also satisfies $g'(\lambda) \geq 0$.

Proposition 3.1 shows the following intuition. As λ increases, the closed-form optimal policy π^*_λ assigns more probability mass to actions with larger π_β and yields a more conservative update. Conversely, a smaller λ allows π^*_λ to retain greater flexibility within the support of π_β and allows the update to be more strongly guided by advantage function. This monotonicity is consistent with the overall behavior of λ in the CCI framework, as shown in Tab. 1.

The detailed proof is provided in Appendix C.

Lemma 3.2 (Maximum Entropy Performance Difference)**.** *Let π and π_β be two stationary policies. Consider the maximum-entropy performance $\mathcal{J}(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\big(r(s_t,a_t) - \alpha\log\pi(a_t|s_t)\big)\right]$, then the maximum entropy performance difference satisfies*

$$\mathcal{J}(\pi) - \mathcal{J}(\pi_\beta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\pi}\Big[\mathbb{E}_{a\sim\pi(\cdot|s)}\big[\tilde{A}^{\pi_\beta}(s,a)\big] - \alpha\mathrm{KL}\big(\pi(\cdot|s)\|\pi_\beta(\cdot|s)\big)\Big], \tag{19}$$

where $\tilde{A}^{\pi_\beta}(s,a)$ is the advantage function under the shaped reward $\tilde{r}(s,a) = r(s,a) - \alpha\log\pi_\beta(a|s)$.

Theorem 3.3 (Performance Lower Bound for the Optimal Policy)**.** *Define $\epsilon_\beta(\pi) := \max_s\left|\mathbb{E}_{a\sim\pi(\cdot|s)}\big[\tilde{A}^{\pi_\beta}(s,a)\big]\right|$, $\kappa_\beta(\pi) := \max_s \mathrm{KL}\big(\pi(\cdot|s)\|\pi_\beta(\cdot|s)\big)$ and the average total variation distance under ν by $\Delta_{\mathrm{TV}}(\pi,\pi_\beta;\nu) := \mathbb{E}_{s\sim\nu}\big[D_{\mathrm{TV}}(\pi(\cdot|s),\pi_\beta(\cdot|s))\big]$. Assume $|\log\pi_\beta(a|s)| \leq B$ for all (s,a). For a given $\lambda \geq 0$ and $\alpha > 0$, π^*_λ satisfies the following maximum-entropy performance lower bound:*

$$\mathcal{J}(\pi^*_\lambda) \geq \mathcal{J}(\pi_\beta) - \frac{2B|\lambda-\alpha|}{1-\gamma}\Delta_{\mathrm{TV}}(\pi^*_\lambda,\pi_\beta;d_{\pi_\beta}) - \frac{4\alpha B + 2\gamma\Big(\epsilon_\beta(\pi^*_\lambda) + \alpha\kappa_\beta(\pi^*_\lambda)\Big)}{(1-\gamma)^2}\Delta_{\mathrm{TV}}(\pi^*_\lambda,\pi_\beta;d_{\pi_\beta}). \tag{20}$$

Theorem 3.4 (Performance Lower Bound for a Parametric Policy)**.** *Define the suboptimality gap of a parametric policy π_θ by $\delta^{\mathrm{sub}}_\theta(\lambda) := F_\lambda(\pi^*_\lambda) - F_\lambda(\pi_\theta) \;\geq\; 0$, where*

$$F_\lambda(\pi) = \mathbb{E}_{s\sim d_{\pi_\beta}}\Big[\mathbb{E}_{a\sim\pi(\cdot|s)}\tilde{A}^{\pi_\beta}(s,a) - \alpha\mathrm{KL}\big(\pi(\cdot|s)\|\pi_\beta(\cdot|s)\big) + (\lambda-\alpha)\mathbb{E}_{a\sim\pi(\cdot|s)}\log\pi_\beta(a|s)\Big].$$

Then π_θ satisfies

$$\begin{aligned}\mathcal{J}(\pi_\theta) \geq\; & \mathcal{J}(\pi_\beta) - \frac{2B|\lambda-\alpha|}{1-\gamma}\,\Delta_{\mathrm{TV}}\big(\pi_\theta,\pi_\beta;d_{\pi_\beta}\big) - \frac{\delta^{\mathrm{sub}}_\theta(\lambda)}{1-\gamma} \\ & - \frac{4\alpha B + 2\gamma\Big(\epsilon_\beta(\pi_\theta) + \alpha\kappa_\beta(\pi_\theta)\Big)}{(1-\gamma)^2}\Delta_{\mathrm{TV}}(\pi_\theta,\pi_\beta;d_{\pi_\beta}).\end{aligned} \tag{21}$$

The proofs of Theorems 3.3 and 3.4 suggest a decomposition of the lower bound into two components. The first-order term $\mathcal{O}\big(\frac{1}{1-\gamma}\big)$ evaluates policy improvement under the fixed state distribution d_{π_β}. The second-order term $\mathcal{O}\big(\frac{1}{(1-\gamma)^2}\big)$ captures distribution shift, arising from the mismatch between d_π

and d_{π_β}. Compared to the KL-density constraint in standard RL, whose performance lower bound is typically $\mathcal{J}^{\mathrm{KL}}(\pi^*) \geq \mathcal{J}(\pi_\beta) - \mathcal{O}\big(\frac{1}{(1-\gamma)^2}\big)$ Lyu et al. [2022], our bound additionally exposes an explicit first-order term and makes the role of λ transparent: the factor $|\lambda - \alpha|$ quantifies how the λ modulates the worst-case performance.

The detailed proof is provided in Appendix D.

4 Experiments

4.1 Comparisons on D4RL Benchmarks

Tasks. We conduct experiments in Gym-MuJoCo locomotion domains and more challenging AntMaze and Kitchen tasks Fu et al. [2021]. The main challenge of the latter is to learn policies for long-horizon planning from datasets that do not contain optimal trajectories, which require "stitching" ability.

Baselines. We select competitive offline RL baselines, including BCQ Fujimoto et al. [2019], IQL Kostrikov et al. [2021b], CQL Kumar et al. [2020], TD3+BC Fujimoto and Gu [2021], SPOT Wu et al. [2022], ACTIVE Chen et al. [2025], EQL Xu et al. [2023], OTR Luo et al. [2023].

Results. The whole results are shown in Tab. 2. Overall, ACPO achieves the best performance across all three D4RL domains, obtaining the highest total scores on Gym-MuJoCo, AntMaze, and Kitchen. These results indicate that ACPO provides a robust policy-constraint method across diverse offline RL domains rather than relying on a single fixed constraint regime.

Table 2: Averaged normalized scores on Gym-MuJoCo-v2, Antmaze-v2, and Kitchen-v0 datasets over five seeds. Scores are reported as mean ± standard deviation. **Bold** indicates the best performance in each task.

Dataset	BCQ	IQL	CQL	TD3+BC	SPOT	ACTIVE	EQL	OTR	ACPO
halfcheetah-med	47.0	47.4	44.0	48.3	**58.4**	50.6	47.2	43.3	52.5 ± 0.76
hopper-med	56.7	66.3	58.5	59.3	86.0	80.1	74.6	78.7	**97.9** ± 2.14
walker2d-med	72.6	78.3	72.5	83.7	86.4	84.8	83.2	79.4	**88.0** ± 1.1
halfcheetah-med-rep	40.4	44.2	45.5	44.6	**52.2**	51.1	44.5	41.3	47.5 ± 0.54
hopper-med-rep	53.3	94.7	95.0	60.9	100.2	**101.1**	98.1	84.8	100.6 ± 0.54
walker2d-med-rep	52.1	73.9	77.2	81.8	91.6	85.4	76.6	66.0	**93.6** ± 1.86
halfcheetah-med-exp	89.1	86.7	90.7	90.7	86.9	93.2	90.6	89.6	**97.2** ± 1.24
hopper-med-exp	81.8	91.5	105.4	98.0	99.3	89.2	105.5	93.2	**112.3** ± 0.72
walker2d-med-exp	109.0	109.6	109.6	110.1	112.0	112.0	110.2	109.3	**113.1** ± 0.22
Locomotion Total	602	692.6	698.4	677.4	773	747.5	730.5	685.6	**802.7** ± 9.12
antmaze-u	78.9	85.5	84.8	78.6	**93.5**	91.6	93.2	83.4	90.0 ± 3.0
antmaze-u-d	55.0	66.7	43.4	71.4	40.7	78.4	65.4	68.9	**86.0** ± 3.4
antmaze-m-p	0	72.2	65.2	10.6	74.7	74.6	**77.5**	70.5	76.0 ± 4.24
antmaze-m-d	0	71.0	54.0	3.0	**79.1**	75.2	70.0	70.4	76.0 ± 4.24
antmaze-l-p	6.7	39.6	38.4	0.2	35.3	51.2	45.6	45.3	**56.0** ± 4.94
antmaze-l-d	2.2	47.5	31.6	0.0	36.3	**48.2**	42.5	45.5	48.0 ± 4.98
Antmaze Total	142.8	382.5	317.4	163.8	359.6	419.2	394.2	384.0	**432.0** ± 24.8
kitchen-m	10.6	52.8	51.0	0.8	0.0	52.1	55.6	-	**72.5** ± 5.5
kitchen-p	18.9	46.1	49.8	0.0	0.0	69.6	**74.5**	-	64.0 ± 15.52
Kitchen Total	29.5	98.9	100.8	0.8	0.0	121.7	130.1	-	**136.5** ± 21.02

4.2 Comparisons on NeoRL2 Benchmarks

Tasks. NeoRL2 (Near Real-World Offline RL Benchmark) Gao et al. [2025] is a recent benchmark designed to reflect practical constraints in offline RL. It includes domains motivated by industrial control, nuclear fusion, and healthcare, featuring pronounced time-delay effects and external disturbances. Moreover, practical data acquisition limits the dataset size and coverage, which places higher demands on the robustness and adaptiveness of offline RL algorithms.

Baselines. We compare against representative baselines commonly reported on NeoRL2, including BC, CQL Kumar et al. [2020], EDAC An et al. [2021], MCQ Lyu et al. [2022], TD3+BC Fujimoto and Gu [2021], RAMBO Rigter et al. [2022], and MOBILE Sun et al. [2023] [3].

Results. As shown in Tab. 3, ACPO achieves clear improvements on NeoRL2. It attains the best performance on 4 out of 7 datasets and yields the highest total score across tasks, substantially outperforming the strongest baseline TD3+BC.

Table 3: Averaged normalized scores on NeoRL2 benchmark over three seeds. Scores are reported as mean $\pm$ standard deviation. **Bold** indicates the best performance in each task.

Dataset	Data	BC	CQL	EDAC	MCQ	TD3+BC	RAMBO	MOBILE	ACPO
Pipeline	69.3	68.6	81.1	72.9	49.7	82.0	24.1	65.5	**95.2** $\pm$ 4.62
Simglucose	73.9	75.1	11.0	8.1	29.6	74.2	10.8	9.3	**94.9** $\pm$ 12.73
RocketRecovery	75.3	72.8	74.3	65.7	76.5	79.7	-44.2	43.7	**84.6** $\pm$ 10.23
RandomFrictionHopper	28.7	28.0	33.0	34.7	31.7	29.5	29.6	**35.1**	33.7 $\pm$ 0.37
DMSD	56.6	65.1	70.2	**78.7**	77.8	60.0	76.2	64.4	74.9 $\pm$ 5.67
Fusion	48.8	55.2	55.9	58.0	49.7	54.6	59.6	5.0	**61.8** $\pm$ 2.07
SafetyHalfCheetah	**73.6**	70.2	71.2	53.1	54.7	68.6	-422.4	8.7	67.1 $\pm$ 6.90
NeoRL2 Total	426.2	435.0	396.7	371.2	369.7	448.6	-266.3	231.7	**512.2** $\pm$ 42.59

4.3 Constraint-Regime Comparisons and Adaptive λ Dynamics

Under the CCI framework, several representative policy constraints can be recovered by fixing the interpolation parameter λ. To isolate the effect of λ, we perform special-case comparisons by disabling the dual update in ACPO and training with a fixed λ throughout. These special cases match wBC, AWAC Nair et al. [2020] and InAC Xiao et al. [2023] at the level of the policy improvement. In contrast, the policy evaluation components may differ across methods; for a controlled comparison, we keep the policy evaluation fixed and use the same maximum-entropy critic updates Eq. (14) and (15) as ACPO in all variants.

Settings. For Gym-MuJoCo tasks, we set $\alpha = 0.1$. For NeoRL2 tasks, we set $\alpha = 0.5$. We instantiate three special cases by choosing: (i) practical wBC: $\lambda = 100$ to approximate the wBC regime, (ii) AWAC: $\lambda = \alpha$, and (iii) InAC: $\lambda = 0$.

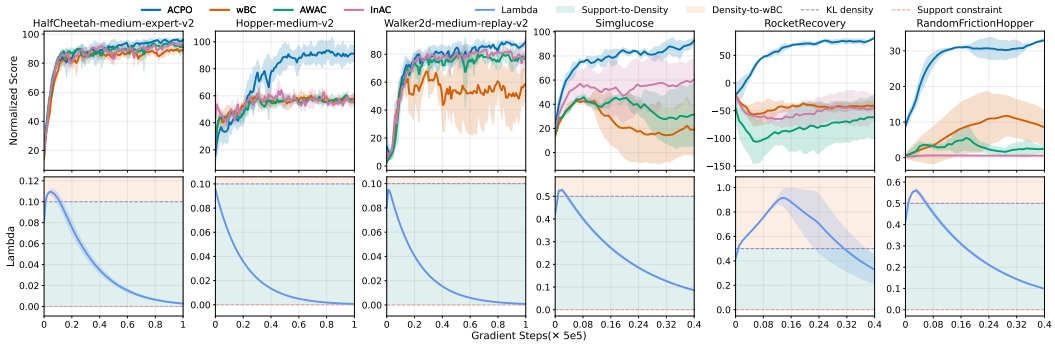


Figure 1: Constraint-regime comparisons and the evolution of λ learned by ACPO during training.

Fig. 1 reports training curves of ACPO against these three baselines across six datasets, together with the adaptive evolution of λ during training. From Fig. 1, we draw the following observations:

Overall Superiority. ACPO consistently achieves the strongest performance across the six tasks, with particularly clear gains on the more challenging NeoRL2 datasets where coverage is limited. This highlights the benefit of automatically adapting the effective constraint regime to the dataset.

λ Dynamics. We observe that λ is often non-monotonic, typically rising in early training and gradually decaying afterwards. Early in training, critic estimates are less calibrated and actor updates

[3]We focus on the strongest baselines used in NeoRL2 and omit MOPO Yu et al. [2020] and COMBO Yu et al. [2021] for brevity.

are more prone to exploiting OOD actions. The dual update reacts by increasing λ, strengthening the constraint and biasing updates toward higher-density behavior actions to stabilize learning. As training progresses and value estimates become better calibrated, the OOD risk diminishes; λ decreases accordingly, relaxing the constraint and enabling more effective policy improvement.

4.4 Adaptive λ versus Manual Schedules

The intermediate region $0 < \lambda < \alpha$ is an integral part of the CCI spectrum and continuously interpolates between support-based and density-based constraints. To assess whether simple schedules can match the adaptive dual update, we compare ACPO with three manually designed λ schedules on nine D4RL Gym-MuJoCo datasets.

Settings. We consider: (i) *linear decay*, which decreases λ linearly from 100 to 0 over the entire training process; (ii) a *two-stage schedule*, which sets $\lambda = \alpha$ for the first 50% of training and $\lambda = 0$ for the remaining 50%; and (iii) an *increase-then-decrease schedule*, which increases λ linearly from α to 2α during the first 33% of training, then decreases it linearly to 0 during the remaining 67%.

Table 4: Normalized scores for ACPO and manual λ schedules.

Dataset	ACPO	Linear Decay	Two-Stage	Inc.-then-Dec.
HalfCheetah-med	**52.5**	45.1	45.3	45.4
Hopper-med	**97.9**	61.1	62.3	67.2
Walker2d-med	**88.0**	69.2	69.5	70.2
HalfCheetah-med-rep	**47.5**	43.0	29.6	45.8
Hopper-med-rep	**100.6**	83.4	100.2	97.6
Walker2d-med-rep	**93.6**	50.9	71.7	66.2
HalfCheetah-med-exp	**97.2**	84.1	93.8	94.6
Hopper-med-exp	**112.3**	105.7	100.8	100.9
Walker2d-med-exp	**113.1**	108.1	108.5	108.1
Locomotion Total	**802.7**	650.6	681.7	696.0

Results. As shown in Tab. 4, ACPO outperforms all three manual schedules on each of the nine datasets, which supports the benefit of adapting λ through the dual update over the tested predefined schedules.

4.5 Behavior Density Estimator Ablation Study

Since ACPO explicitly depends on $\log \pi_\beta(a|s)$, the behavior density estimator directly affects both the constraint signal and the policy improvement weights. Besides the Gaussian estimator, we also consider a conditional variational auto-encoder (CVAE) Fujimoto et al. [2019], Han et al. [2025b] to model π_β, and adapt the IS-corrected method of Wu et al. [2022] to reduce bias in expectations involving π_β. We compare Gaussian and CVAE behavior models on 9 Gym-MuJoCo datasets, with results shown in Fig. 2.

As shown in Fig. 2, the two estimators perform similarly on most datasets, but differ noticeably on `walker2d-medium-v2` and `walker2d-medium-replay-v2`. To understand this gap, we visualize the distributions of $\log \pi_\beta(a|s)$ for (i) in-dataset actions a_{in} and (ii) perturbed OOD actions $a_{\text{ood}} = a_{\text{in}} + \xi z$, where $z \sim \mathcal{N}(0, I)$ and ξ controls the perturbation scale; the results are shown in Fig. 3.

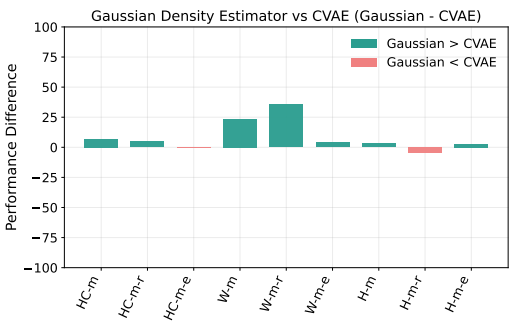


Figure 2: Normalized score differences on Gym-MuJoCo datasets.

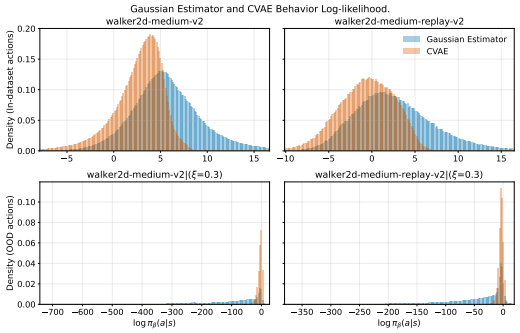


Figure 3: Distributions of $\log \pi_\beta(a|s)$ under Gaussian and CVAE behavior estimators.

Fig. 3 suggests why the CVAE underperforms on these two datasets. On in-dataset actions, it produces a sharper $\log \pi_\beta(a|s)$ distribution than the Gaussian estimator, indicating a more concentrated density estimate. On perturbed OOD actions, however, it does not consistently assign sufficiently low likelihood: the distribution of $\log \pi_\beta(a_{\text{ood}}|s)$ is not shifted left as much as expected. This tendency to overestimate OOD likelihood weakens the constraint in ACPO and ultimately degrades performance.

4.6 Sensitivity Analysis of Constraint Threshold ϵ

We also study the sensitivity of ACPO to the constraint threshold ϵ in Eq. (7). Tab. 5 reports results on the 9 Gym-MuJoCo datasets using $\epsilon \in \{-0.2, -0.7, -5.0\}$. Overall, ACPO is reasonably robust to ϵ: the average scores under $\epsilon = -0.2, -0.7, -5.0$ are close.

Table 5: Sensitivity to the ϵ on Gym-MuJoCo.

Dataset	$\epsilon = -0.2$	$\epsilon = -0.7$	$\epsilon = -5.0$
HalfCheetah-m	52.5 ± 0.76	48.2 ± 1.02	48.4 ± 0.72
HalfCheetah-m-e	97.2 ± 1.24	95.1 ± 0.40	95.7 ± 0.49
HalfCheetah-m-r	47.5 ± 0.54	47.2 ± 0.23	44.6 ± 0.67
Hopper-m	89.9 ± 10.88	95.9 ± 4.29	97.9 ± 2.14
Hopper-m-e	109.1 ± 8.36	110.3 ± 0.51	112.3 ± 0.72
Hopper-m-r	100.2 ± 1.02	99.6 ± 1.10	100.6 ± 0.54
Walker2d-m	84.3 ± 0.99	88.0 ± 1.1	85.1 ± 1.12
Walker2d-m-e	109.5 ± 0.22	113.1 ± 0.22	109.6 ± 0.34
Walker2d-m-r	82.3 ± 13.01	93.6 ± 1.86	85.8 ± 4.02
Average Score	85.8	87.9	86.7

4.7 Sensitivity Analysis of Dual Learning Rate η_λ

We analyze the sensitivity of ACPO to the dual learning rate η_λ in Eq. (13). We evaluate $\eta_\lambda \in \{1\text{e}{-5}, 1\text{e}{-4}, 3\text{e}{-4}\}$. The default value $\eta_\lambda = 1\text{e}{-5}$ achieves the highest average score. ACPO remains effective at the other tested values, although the lower aggregate scores indicate some sensitivity to more aggressive dual learning rates.

Table 6: Sensitivity to the η_λ on Gym-MuJoCo.

Dataset	$\eta_\lambda = 1\text{e}{-5}$	$1\text{e}{-4}$	$3\text{e}{-4}$
HalfCheetah-m	52.5	49.4	48.2
Hopper-m	97.9	78.8	92.7
Walker2d-m	88.0	85.5	79.3
HalfCheetah-m-r	47.5	49.4	44.5
Hopper-m-r	100.6	100.5	100.1
Walker2d-m-r	93.6	87.9	78.7
HalfCheetah-m-e	97.2	93.3	95.4
Hopper-m-e	112.3	109.1	110.4
Walker2d-m-e	113.1	110.3	108.6
Average Score	89.2	84.9	84.2

Training details for all experiments are provided in Appendix F.

5 Conclusion

We proposed CCI, a unified optimization framework that bridges weighted BC, KL-based density regularization, and support constraints as special cases along a common constraint spectrum. CCI exposes an explicit interpolation parameter that enables smooth transitions and principled combinations across constraint regimes. Building on CCI, we developed ACPO, a practical primal–dual algorithm that automatically adapts this constraint dial via a Lagrangian dual update. We further established a maximum-entropy performance difference lemma and derived performance lower bounds for both the closed-form optimal policy and its parametric projection. Experiments on D4RL and NeoRL2 show that ACPO achieves robust gains across diverse domains, often matching or exceeding strong baselines overall. As future work, extending ACPO to trajectory-level settings for long-horizon credit assignment and sparse rewards tasks is critical. Another promising direction is to incorporate dataset filtering to improve the reliability of behavior policy estimation.

Acknowledgments and Disclosure of Funding

Funding. This work received no dedicated project funding.

Competing interests. The authors declare no competing interests related to this work.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 22–31. PMLR, 06–11 Aug 2017.

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble. *Advances in neural information processing systems*, 34:7436–7447, 2021.

Kavosh Asadi and Michael L. Littman. An alternative softmax operator for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 243–252. PMLR, 06–11 Aug 2017.

Arip Asadulaev, Rostislav Korst, Alexander Korotin, Vage Egiazarian, Andrey Filchenkov, and Evgeny Burnaev. Rethinking optimal transport in offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 123592–123607. Curran Associates, Inc., 2024. doi: 10.52202/079017-3929.

Chenjia Bai, Lingxiao Wang, Zhuoran Yang, Zhi-Hong Deng, Animesh Garg, Peng Liu, and Zhaoran Wang. Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning. In *International Conference on Learning Representations*, 2022.

Tianyuan Chen, Ronglong Cai, Faguo Wu, and Xiao Zhang. ACTIVE: Offline reinforcement learning via adaptive imitation and in-sample v-ensemble. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=qiluFujVc8`.

Xi Chen, Ali Ghadirzadeh, Tianhe Yu, Jianhao Wang, Yuan Gao, Wenzhe Li, Liang Bin, Chelsea Finn, and Chongjie Zhang. LAPO: Latent-variable advantage-weighted policy optimization for offline reinforcement learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Xinyue Chen, Zijian Zhou, Zheng Wang, Che Wang, Yanqiu Wu, and Keith Ross. Bail: Best-action imitation learning for batch deep reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18353–18363. Curran Associates, Inc., 2020.

Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 158–168. PMLR, 08–11 Nov 2022.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2021. URL `https://arxiv.org/abs/2004.07219`.

Scott Fujimoto and Shixiang (Shane) Gu. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 20132–20145. Curran Associates, Inc., 2021.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR, 10–15 Jul 2018.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2052–2062. PMLR, 09–15 Jun 2019.

Songyi Gao, Zuolin Tu, Rong-Jun Qin, Yi-Hao Sun, Xiong-Hui Chen, and Yang Yu. Neorl-2: Near real-world benchmarks for offline reinforcement learning with extended realistic scenarios, 2025.

Seyed Kamyar Seyed Ghasemipour, Dale Schuurmans, and Shixiang Shane Gu. Emaq: Expected-max q-learning operator for simple yet effective offline and online rl. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3682–3691. PMLR, 18–24 Jul 2021.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1352–1361. PMLR, 06–11 Aug 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 10–15 Jul 2018.

Xinchen Han, Hossam Afifi, Hassine Moungla, and Michel Marot. Leaky ppo: A simple and efficient rl algorithm for autonomous vehicles. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2024. doi: 10.1109/IJCNN60899.2024.10650450.

Xinchen Han, Hossam Afifi, and Michel Marot. Cosine similarity based adaptive implicit q-learning for offline reinforcement learning. In *2025 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, 2025a. doi: 10.1109/WCNC61545.2025.10978817.

Xinchen Han, Hossam Afifi, and Michel Marot. Elapse: Expand latent action projection space for policy optimization in offline reinforcement learning. *Neurocomputing*, 631:129665, 2025b. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2025.129665.

Xinchen Han, Hossam Afifi, and Michel Marot. PIQL: Projective implicit q-learning with support constraint for offline reinforcement learning. In *The 25th International Conference on Autonomous Agents and Multi-Agent Systems*, 2025c.

Xinchen Han, Hossam Afifi, Hassine Moungla, and Michel Marot. Attention ensemble mixture: a novel offline reinforcement learning algorithm for autonomous vehicles. *Applied Intelligence*, 55 (6):1–14, 2025d.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, page 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608737.

Jeonghye Kim, Yongjae Shin, Whiyoung Jung, Sunghoon Hong, Deunsol Yoon, Youngchul Sung, Kanghoon Lee, and Woohyung Lim. Penalizing infeasible actions and reward scaling in reinforcement learning with offline data. In *Forty-second International Conference on Machine Learning*, 2025.

Ilya Kostrikov, Rob Fergus, Jonathan Tompson, and Ofir Nachum. Offline reinforcement learning with fisher divergence critic regularization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5774–5783. PMLR, 18–24 Jul 2021a.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *Deep RL Workshop NeurIPS 2021*, 2021b. URL `https://openreview.net/forum?id=EblVBDNalKu`.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in neural information processing systems*, 32, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191. Curran Associates, Inc., 2020.

Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. Optidice: Offline policy optimization via stationary distribution correction estimation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6120–6130. PMLR, 18–24 Jul 2021.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Jie Liu, Yinmin Zhang, Chuming Li, Yaodong Yang, Yu Liu, and Wanli Ouyang. Adaptive pessimism via target q-value for offline reinforcement learning. *Neural Networks*, 180:106588, 2024. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2024.106588.

Qin-Wen Luo, Ming-Kun Xie, Ye-Wen Wang, and Sheng-Jun Huang. Learning to trust bellman updates: Selective state-adaptive regularization for offline RL. In *Forty-second International Conference on Machine Learning*, 2025.

Yicheng Luo, zhengyao jiang, Samuel Cohen, Edward Grefenstette, and Marc Peter Deisenroth. Optimal transport for offline imitation learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=MhuFzFsrfvH`.

Jiafei Lyu, Xiaoteng Ma, Xiu Li, and Zongqing Lu. Mildly conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 1711–1724. Curran Associates, Inc., 2022.

Liyuan Mao, Haoran Xu, Xianyuan Zhan, Weinan Zhang, and Amy Zhang. Diffusion-dice: In-sample diffusion guidance for offline reinforcement learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 98806–98834. Curran Associates, Inc., 2024. doi: 10.52202/079017-3136.

Yixiu Mao, Hongchang Zhang, Chen Chen, Yi Xu, and Xiangyang Ji. Supported trust region optimization for offline reinforcement learning. In *International Conference on Machine Learning*, pages 23829–23851. PMLR, 2023.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

Motoki Omura, Yusuke Mukuta, Kazuki Ota, Takayuki Osa, and Tatsuya Harada. Offline reinforcement learning with wasserstein regularization via optimal transport maps. In *Reinforcement Learning Conference*, 2025. URL `https://openreview.net/forum?id=RxdQBYKjtE`.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Zhiyong Peng, Yadong Liu, Haoqiang Chen, and Zongtan Zhou. Conservative network for offline reinforcement learning. *Knowledge-Based Systems*, 282:111101, 2023. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2023.111101.

Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th international conference on Machine learning*, pages 745–750, 2007.

Marc Rigter, Bruno Lacerda, and Nick Hawes. Rambo-rl: Robust adversarial model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 16082–16097. Curran Associates, Inc., 2022.

Noah Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, Nicolas Heess, and Martin Riedmiller. Keep doing what worked: Behavior modelling priors for offline reinforcement learning. In *International Conference on Learning Representations*, 2020.

Yihao Sun, Jiaji Zhang, Chengxing Jia, Haoxin Lin, Junyin Ye, and Yang Yu. Model-Bellman inconsistency for model-based offline reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 33177–33194. PMLR, 23–29 Jul 2023.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Ziyu Wang, Alexander Novikov, Konrad Zolna, Josh S Merel, Jost Tobias Springenberg, Scott E Reed, Bobak Shahriari, Noah Siegel, Caglar Gulcehre, Nicolas Heess, et al. Critic regularized regression. *Advances in Neural Information Processing Systems*, 33:7768–7778, 2020.

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.

Kun Wu, Yinuo Zhao, Zhiyuan Xu, Zhengping Che, Chengxiang Yin, Chi Harold Liu, Feifei Feng, and Jian Tang. Acl-ql: Adaptive conservative level in q-learning for offline reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):11399–11413, 2025. doi: 10.1109/TNNLS.2024.3497667.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning, 2020. URL `https://openreview.net/forum?id=BJg9hTNKPH`.

Chenjun Xiao, Ruitong Huang, Jincheng Mei, Dale Schuurmans, and Martin Müller. Maximum entropy monte-carlo planning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Chenjun Xiao, Han Wang, Yangchen Pan, Adam White, and Martha White. The in-sample softmax for offline reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Haoran Xu, Xianyuan Zhan, Honglei Yin, and Huiling Qin. Discriminator-weighted offline imitation learning from suboptimal demonstrations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 24725–24742. PMLR, 17–23 Jul 2022.

Haoran Xu, Li Jiang, Jianxiong Li, Zhuoran Yang, Zhaoran Wang, Victor Wai Kin Chan, and Xianyuan Zhan. Offline RL with no OOD actions: In-sample learning via implicit value regularization. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=ueYYgo2pSSU`.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 14129–14142. Curran Associates, Inc., 2020.

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. Combo: Conservative offline model-based policy optimization. In *Advances in Neural Information Processing Systems*, volume 34, pages 28954–28967. Curran Associates, Inc., 2021.

Yunfan Zhou, Xijun Li, and Qingyu Qu. Offline reinforcement learning with adaptive behavior regularization, 2022.

Appendix

A Related Works

In offline RL, policy constraints are not the only mechanism for mitigating extrapolation error induced by OOD actions. Beyond policy constraints, prior work has explored conservative value estimation Kumar et al. [2020], Peng et al. [2023], Liu et al. [2024], ensemble-based uncertainty-aware methods An et al. [2021], Chen et al. [2025], Bai et al. [2022], and DIstribution Correction Estimation (DICE) algorithms Nachum et al. [2019], Lee et al. [2021], Mao et al. [2024]. To highlight the advantages of our framework and algorithm, we focus the remainder of this section on policy-constrained offline RL and organize representative methods into three categories: wBC, density regularization, and support constraints.

wBC methods either cast policy learning as (weighted) supervised imitation on the dataset, or incorporate an explicit BC regularizer into the policy improvement objective. BAIL Chen et al. [2020] learns a value function and imitates only actions selected as high-quality under the learned value. IBC Florence et al. [2022] improves expressiveness via implicit energy-based policies, avoiding the limitations of explicit regression-based models. DWBC Xu et al. [2022] trains a discriminator to estimate expert-likeness and uses it to reweight the BC loss. In contrast, TD3+BC Fujimoto and Gu [2021] augments value-based policy improvement with a BC term, mitigating distributional drift during policy improvement.

Density regularization methods allow controlled deviation from the behavior policy by penalizing distributional shift at the policy level. The most common choice is KL-based regularization Peters and Schaal [2007], Peng et al. [2019], Nair et al. [2020], Wang et al. [2020], Kostrikov et al. [2021b], due to its elegant and tractable policy improvement objective, which keeps policy updates via stable in-sample learning. For completeness, we also review density regularization choices beyond KL divergence, including MMD-based constraints for sample-based distribution matching Wu et al. [2020], Fisher divergence Kostrikov et al. [2021a], and Wasserstein distance regularization Omura et al. [2025], Asadulaev et al. [2024].

Support constraint methods perform policy improvement with the policy constrained within the support of the behavior policy. EMaQ Ghasemipour et al. [2021] samples multiple candidate actions from behavior policy and performs backups over these in-support proposals. STR Mao et al. [2023] combines trust-region policy optimization with an explicit support constraint via Importance Sampling (IS), ensuring that each update remains within the behavior support. SPOT Wu et al. [2022] trains a VAE-based estimator to model the behavior support and introduces a simple plug-in regularization term that keeps policy improvement within the estimated support.

Overall, most existing approaches operate within a single policy-constraint paradigm. A unified optimization view that bridges all three families under one framework—enabling smooth transitions, flexible combinations, and automatic adaptation across constraint regimes—remains missing. We therefore propose the CCI framework and derive a practical algorithm, ACPO.

B Continuous Constraint Interpolation Framework

We start from the following constrained maximum-entropy policy optimization problem,

$$\begin{aligned} \max_{\pi} \mathbb{E}_{s\sim\mathcal{D},a\sim\pi(\cdot|s)}\big[Q(s,a) - \alpha\log\pi(a|s)\big], \\ \text{s.t. } \mathbb{E}_{s\sim\mathcal{D},a\sim\pi(\cdot|s)}\big[\log\pi_\beta(a|s)\big] \geq \epsilon, \\ \int_a \pi(a|s)\,da = 1,\ \forall s. \end{aligned} \tag{22}$$

Solving the above constrained problem Eq. (22) via the Lagrangian,

$$\begin{aligned} \mathcal{L}(\pi,\lambda,\nu) = \mathbb{E}_{s\sim\mathcal{D},a\sim\pi(\cdot|s)}\left[Q(s,a) - \alpha\log\pi(a|s)\right] + \lambda\left[\mathbb{E}_{s\sim\mathcal{D},a\sim\pi(\cdot|s)}\left[\log\pi_\beta(a|s) - \epsilon\right]\right] \\ + \mathbb{E}_{s\sim\mathcal{D}}\left[\nu(s)\big[\int_a \pi(a|s)\,da - 1\big]\right], \end{aligned} \tag{23}$$

To obtain the optimal non-parametric policy, we take the functional derivative of $\mathcal{L}(\pi, \lambda, \nu)$ w.r.t. $\pi(\cdot|s)$, and setting it to zero yields:

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}(\pi, \lambda, \nu)}{\partial \pi} \\ &= Q(s,a) - \alpha\big(1 + \log \pi(a|s)\big) + \lambda \log \pi_\beta(a|s) + \nu(s), \end{aligned} \tag{24}$$

Rearranging gives

$$\log \pi(a|s) = \frac{1}{\alpha}\Big(Q(s,a) + \lambda \log \pi_\beta(a|s) + \nu(s) - \alpha\Big). \tag{25}$$

Exponentiating both sides yields

$$\pi^*_\lambda(a|s) = \exp\left(\frac{\nu(s) - \alpha}{\alpha}\right) \exp\left(\frac{Q(s,a)}{\alpha}\right) \pi_\beta(a|s)^{\lambda/\alpha}. \tag{26}$$

Since the prefactor $\exp((\nu(s) - \alpha)/\alpha)$ does not depend on a, it can be absorbed into the normalizer, giving

$$\pi^*_\lambda(a|s) \;\propto\; \exp\left(\frac{Q(s,a)}{\alpha}\right) \pi_\beta(a|s)^{\lambda/\alpha}. \tag{27}$$

Enforcing $\int \pi^*_\lambda(a|s)\, da = 1$ yields the normalized closed-form solution

$$\pi^*_\lambda(a|s) = \exp\Big(\frac{Q(s,a) - Z_\lambda(s)}{\alpha}\Big) \pi_\beta(a|s)^{\lambda/\alpha}, \tag{28}$$

where $\lambda \geq 0$ and $\nu(s)$ are dual variables, and $Z_\lambda(s) = \alpha \log \int_a \exp\Big(\frac{Q(s,a)}{\alpha}\Big)\pi_\beta(a|s)^{\lambda/\alpha}\, da$.

Similarly within InAC Xiao et al. [2023], we also replace $Z_\lambda(s)$ with value function $V(s)$. Therefore, the closed-form solution Eq. (28) can be transformed into the following form,

$$\pi^*_\lambda(a|s) \propto \exp\Big(\frac{A(s,a)}{\alpha}\Big) \pi_\beta(a|s)^{\lambda/\alpha}, \tag{29}$$

Then, we project the non-parametric optimal policy π^*_λ into the parametric space π_θ by minimizing the reverse KL divergence under the state distribution in dataset,

$$\begin{aligned} \pi_\theta(a|s) &= \arg\min_{\pi_\theta} \mathbb{E}_{s\sim\mathcal{D}}\big[\, D_{\text{KL}}\left[\pi^*_\lambda(\cdot|s)||\pi_\theta(\cdot|s)\right]\big] \\ &= \arg\min_{\theta} \mathbb{E}_{s\sim\mathcal{D}}\Big[\mathbb{E}_{a\sim\pi^*_\lambda(\cdot|s)}\big[\log \pi^*_\lambda(a|s) - \log \pi_\theta(a|s)\big]\Big] \\ &= \arg\min_{\theta} \mathbb{E}_{s\sim\mathcal{D}}\Big[\mathbb{E}_{a\sim\pi_\beta(\cdot|s)}\Big[-\exp\Big(\frac{A(s,a)}{\alpha}\Big)\pi_\beta(a|s)^{(\lambda-\alpha)/\alpha} \log \pi_\theta(a|s)\Big]\Big] \\ &= \arg\max_{\theta} \mathbb{E}_{s\sim\mathcal{D}}\Big[\mathbb{E}_{a\sim\pi_\beta(\cdot|s)}\Big[\exp\Big(\frac{A(s,a)}{\alpha} + \frac{\lambda-\alpha}{\alpha}\log \pi_\beta(a|s)\Big) \log \pi_\theta(a|s)\Big]\Big]. \end{aligned} \tag{30}$$

From Eq. (30), the policy improvement objective can be written in the dataset form

$$\max_\theta\; \mathcal{J}^{\text{CCI}}(\theta; \lambda) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\Big[w_\lambda(s,a)\, \log \pi_\theta(a|s)\Big], \tag{31}$$

where $w_\lambda(s,a) = \exp\Big(\frac{A(s,a)}{\alpha} + \frac{\lambda-\alpha}{\alpha}\log \pi_\beta(a|s)\Big)$. This yields the CCI policy improvement objective Eq. (11).

Varying λ continuously interpolates the influence of the weight term $w_\lambda(s,a)$, yielding several representative regimes:

① **Support Constraint** $\lambda = 0$**.** When $\lambda = 0$, the objective becomes

$$\max_\theta\; \mathcal{J}^{\text{CCI}}(\theta; 0) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\Big[\exp\Big(\frac{A(s,a)}{\alpha} - \log \pi_\beta(a|s)\Big) \log \pi_\theta(a|s)\Big], \tag{32}$$

which matches policy improvement objective used by InAC Xiao et al. [2023].

② **Support-to-Density Constraint** $0 < \lambda < \alpha$**.** When $0 < \lambda < \alpha$, the objective becomes

$$\max_{\theta} \mathcal{J}^{\text{CCI}}(\theta;\alpha) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\Big[\exp\Big(\frac{A(s,a)}{\alpha} + \frac{\alpha-\lambda}{\alpha}\log\pi_\beta(a|s)\Big)\log\pi_\theta(a|s)\Big], \tag{33}$$

As λ increases toward α, the update gradually shifts from support constraint toward KL-based density constraint.

③ **KL-based Density Constraint** $\lambda = \alpha$**.** When $\lambda = \alpha$, the objective becomes

$$\max_{\theta} \mathcal{J}^{\text{CCI}}(\theta;\alpha) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\Big[\exp\Big(\frac{A(s,a)}{\alpha}\Big)\log\pi_\theta(a|s)\Big], \tag{34}$$

which matches the policy improvement objective in AWAC Nair et al. [2020].

④ **Density-to-wBC Constraint** $\lambda > \alpha$**.** When $\lambda > \alpha$, the objective becomes

$$\max_{\theta} \mathcal{J}^{\text{CCI}}(\theta;\lambda) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\Big[\exp\Big(\underbrace{\tfrac{A(s,a)}{\alpha}}_{\text{KL-density term}} + \underbrace{\tfrac{\lambda-\alpha}{\alpha}\log\pi_\beta(a|s)}_{\text{wBC term}}\Big)\log\pi_\theta(a|s)\Big]. \tag{35}$$

⑤ **Practical wBC** $\lambda \gg \alpha + A(s,a)(\log\pi_\beta(a|s))^{-1}$**.** When $\lambda \gg \alpha + A(s,a)(\log\pi_\beta(a|s))^{-1}$, then $\frac{\lambda-\alpha}{\alpha}\log\pi_\beta(a|s) \gg \frac{A(s,a)}{\alpha}$, the objective becomes

$$\max_{\theta} \mathcal{J}^{\text{CCI}}(\theta;\lambda) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\Big[\exp\big(\frac{\lambda-\alpha}{\alpha}\log\pi_\beta(a|s)\big)\log\pi_\theta(a|s)\Big], \tag{36}$$

We note that the practical wBC regime corresponds to choosing $\lambda \gg \alpha + A(s,a)\big(\log\pi_\beta(a|s)\big)^{-1}$. Next, we provide an explicit upper bound on $\big|A(s,a)\,(\log\pi_\beta(a|s))^{-1}\big|$ under the maximum-entropy setting.

In practice, we stabilize training by clipping log-probabilities for any stochastic policy: $\log\pi(a|s) \in [\ell_{\min}, \ell_{\max}]$. Define

$$C_{\max} \triangleq \max\{|\ell_{\min}|, |\ell_{\max}|\} \Rightarrow |-\alpha\log\pi(a|s)| \le \alpha C_{\max},$$

In common RL reward setting, the reward function is bounded by $|r(s,a)| \le R_{\max}$. Then the per-step soft reward satisfies $\big|r(s,a) - \alpha\log\pi(a|s)\big| \le R_{\max} + \alpha C_{\max}$. By the discounted sum, we have $|Q^\pi(s,a)| \le \frac{R_{\max}+\alpha C_{\max}}{1-\gamma}$, $|V^\pi(s)| \le \frac{R_{\max}+\alpha C_{\max}}{1-\gamma}$. Therefore,

$$|A^\pi(s,a)| \le |Q^\pi(s,a)| + |V^\pi(s)| \le \frac{2\big(R_{\max}+\alpha C_{\max}\big)}{1-\gamma}. \tag{37}$$

It follows that

$$\left|\frac{A(s,a)}{\log\pi_\beta(a|s)}\right| \le \frac{|A(s,a)|}{C^\beta_{\min}+\delta} \le \frac{2\big(R_{\max}+\alpha C_{\max}\big)}{(1-\gamma)(C^\beta_{\min}+\delta)}. \tag{38}$$

Therefore, a sufficient condition for the practical wBC constraint is to choose $\lambda \gg \alpha + \frac{2\big(R_{\max}+\alpha C_{\max}\big)}{(1-\gamma)(C^\beta_{\min}+\delta)}$, where $C^\beta_{\min} \triangleq \min|\log\pi_\beta(a|s)|$, and δ is a small constant introduced to avoid division by zero.

Summarizing the above, we can obtain Tab. 1.

In addition, SAC Haarnoja et al. [2018] is another extreme initiation of the CCI framework, minimizing the forward KL divergence $D_{\text{KL}}\left[\pi_\theta(\cdot|s)||\pi^*_\lambda(\cdot|s)\right]$, and setting $\lambda = 0$, then the policy improvement objective is

$$\max_{\theta} \mathcal{J}^{\text{SAC}}(\theta;0) = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\big[\exp\big(\frac{A(s,a)}{\alpha} - \log\pi_\theta(a|s)\big)\big], \tag{39}$$

which matches the policy improvement objective in SAC.

C Proof for Proposition 3.1

Proposition C.1. *Fix any state s and temperature $\alpha > 0$. Let $\pi^*_\lambda(\cdot|s)$ denote the closed-form optimizer in Eq. (9) for a given $\lambda \geq 0$. Define*

$$g_s(\lambda) := \mathbb{E}_{a\sim\pi^*_\lambda(\cdot|s)}\big[\log \pi_\beta(a|s)\big]. \tag{40}$$

Then $g_s(\lambda)$ is monotone non-decreasing in λ, and

$$\frac{d}{d\lambda} g_s(\lambda) = \frac{1}{\alpha}\,\mathrm{Var}_{a\sim\pi^*_\lambda(\cdot|s)}\big(\log \pi_\beta(a|s)\big) \;\geq\; 0. \tag{41}$$

Consequently, $g(\lambda) := \mathbb{E}_{s\sim\mathcal{D}}[g_s(\lambda)]$ also satisfies $g'(\lambda) \geq 0$.

Proof. Fix s and abbreviate $Q(a) := Q(s,a)$ and $\pi_\beta(a) := \pi_\beta(a|s)$. Define the unnormalized density and the partition function

$$w_\lambda(a) := \exp\!\left(\frac{Q(a)}{\alpha}\right)\,\pi_\beta(a)^{\lambda/\alpha}, \qquad \mathcal{Z}(\lambda) := \int w_\lambda(a)\,da, \tag{42}$$

so that $\pi^*_\lambda(a|s) = w_\lambda(a)/\mathcal{Z}(\lambda)$. By definition,

$$g_s(\lambda) = \int \pi^*_\lambda(a|s)\;\log \pi_\beta(a)\,da. \tag{43}$$

Differentiating Eq. (43) w.r.t. λ and exchanging derivative and integral,

$$\frac{d}{d\lambda} g_s(\lambda) = \int \frac{\partial}{\partial\lambda}\pi^*_\lambda(a|s)\;\log \pi_\beta(a)\,da. \tag{44}$$

Since $\pi^*_\lambda = w_\lambda/\mathcal{Z}(\lambda)$, we have

$$\frac{\partial}{\partial\lambda}\pi^*_\lambda(a|s) = \frac{w'_\lambda(a)}{\mathcal{Z}(\lambda)} - \frac{w_\lambda(a)\mathcal{Z}'(\lambda)}{\mathcal{Z}(\lambda)^2} = \pi^*_\lambda(a|s)\Big(\frac{w'_\lambda(a)}{w_\lambda(a)} - \frac{\mathcal{Z}'(\lambda)}{\mathcal{Z}(\lambda)}\Big). \tag{45}$$

Using $w_\lambda(a) = \exp(Q(a)/\alpha)\exp((\lambda/\alpha)\log\pi_\beta(a))$, we obtain

$$\frac{w'_\lambda(a)}{w_\lambda(a)} = \frac{1}{\alpha}\log \pi_\beta(a). \tag{46}$$

Moreover,

$$\mathcal{Z}'(\lambda) = \int w'_\lambda(a)\,da = \int w_\lambda(a)\cdot\frac{1}{\alpha}\log\pi_\beta(a)\,da \Rightarrow \frac{\mathcal{Z}'(\lambda)}{\mathcal{Z}(\lambda)} = \frac{1}{\alpha}\int \pi^*_\lambda(a|s)\log\pi_\beta(a)\,da = \frac{1}{\alpha}g_s(\lambda). \tag{47}$$

Substituting Eq. (46), (47) into Eq. (45) yields

$$\frac{\partial}{\partial\lambda}\pi^*_\lambda(a|s) = \frac{1}{\alpha}\pi^*_\lambda(a|s)\Big(\log\pi_\beta(a) - g_s(\lambda)\Big). \tag{48}$$

Plugging Eq. (48) into Eq. (44) gives

$$\begin{aligned}
\frac{d}{d\lambda} g_s(\lambda) &= \frac{1}{\alpha}\int \pi^*_\lambda(a|s)\Big(\log\pi_\beta(a) - g_s(\lambda)\Big)\log\pi_\beta(a)\,da \\
&= \frac{1}{\alpha}\Big(\mathbb{E}_{\pi^*_\lambda}\big[(\log\pi_\beta)^2\big] - \mathbb{E}_{\pi^*_\lambda}[\log\pi_\beta]^2\Big) \\
&= \frac{1}{\alpha}\mathrm{Var}_{a\sim\pi^*_\lambda(\cdot|s)}\big(\log\pi_\beta(a)\big) \;\geq\; 0,
\end{aligned}$$

which proves Eq. (41). The statement for $g(\lambda) = \mathbb{E}_{s\sim\mathcal{D}}[g_s(\lambda)]$ follows by linearity of expectation. □

D Proof for Performance Lower Bound

Lemma D.1 (Maximum-Entropy Performance Difference Lemma)**.** *Let π and π_β be two stationary policies. Consider the maximum-entropy performance $\mathcal{J}(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\big(r(s_t,a_t) - \alpha\log\pi(a_t|s_t)\big)\right]$. Then the maximum-entropy performance difference satisfies*

$$\mathcal{J}(\pi) - \mathcal{J}(\pi_\beta) = \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d_\pi}\Big[\mathbb{E}_{a\sim\pi(\cdot|s)}\big[\tilde{A}^{\pi_\beta}(s,a)\big] - \alpha\, D_{\mathrm{KL}}\big(\pi(\cdot|s)\,\|\,\pi_\beta(\cdot|s)\big)\Big], \tag{49}$$

where $\tilde{A}^{\pi_\beta}(s,a)$ is the advantage function under the shaped reward $\tilde{r}(s,a) = r(s,a) - \alpha\log\pi_\beta(a|s)$.

Proof. Introduce π_β into the entropy-regularized return:

$$\begin{aligned}\mathcal{J}(\pi) &= \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t\big(r(s_t,a_t) - \alpha\log\pi(a_t|s_t)\big)\Big] \\ &= \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t\big(r(s_t,a_t) - \alpha\log\pi_\beta(a_t|s_t)\big)\Big] - \alpha\,\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t\log\frac{\pi(a_t|s_t)}{\pi_\beta(a_t|s_t)}\Big].\end{aligned} \tag{50}$$

Define the shaped return $\widetilde{\mathcal{J}}(\pi) := \mathbb{E}_{\tau\sim\pi}\big[\sum_{t=0}^{\infty}\gamma^t\,\tilde{r}(s_t,a_t)\big]$ with $\tilde{r}(s,a) = r(s,a) - \alpha\log\pi_\beta(a|s)$. Since $\log\frac{\pi_\beta(a|s)}{\pi_\beta(a|s)} = 0$, we have $\mathcal{J}(\pi_\beta) = \widetilde{\mathcal{J}}(\pi_\beta)$ and thus

$$\mathcal{J}(\pi) - \mathcal{J}(\pi_\beta) = \big(\widetilde{\mathcal{J}}(\pi) - \widetilde{\mathcal{J}}(\pi_\beta)\big) - \alpha\,\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t\log\frac{\pi(a_t|s_t)}{\pi_\beta(a_t|s_t)}\Big]. \tag{51}$$

For the first term, $\widetilde{\mathcal{J}}$ is a standard discounted return with reward $\tilde{r}$. By the performance difference theorem (e.g., Lemma 6.1 in Kakade and Langford [2002]),

$$\widetilde{\mathcal{J}}(\pi) - \widetilde{\mathcal{J}}(\pi_\beta) = \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d_\pi}\,\mathbb{E}_{a\sim\pi(\cdot|s)}\big[\tilde{A}^{\pi_\beta}(s,a)\big]. \tag{52}$$

For the second term, using the discounted occupancy identity $d_\pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\,\Pr(s_t = s \mid \pi)$, we obtain

$$\begin{aligned}\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t\log\frac{\pi(a_t|s_t)}{\pi_\beta(a_t|s_t)}\Big] &= \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d_\pi}\,\mathbb{E}_{a\sim\pi(\cdot|s)}\Big[\log\frac{\pi(a|s)}{\pi_\beta(a|s)}\Big] \\ &= \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d_\pi}\Big[D_{\mathrm{KL}}\big(\pi(\cdot|s)\,\|\,\pi_\beta(\cdot|s)\big)\Big].\end{aligned} \tag{53}$$

Substituting Eq. (52) and Eq. (53) into Eq. (51) yields Eq. (49). □

Theorem D.2 (Performance Lower Bound for the Optimal Policy)**.** *Define $\epsilon_\beta(\pi) := \max_s\Big|\mathbb{E}_{a\sim\pi(\cdot|s)}\big[\tilde{A}^{\pi_\beta}(s,a)\big]\Big|$, $\kappa_\beta(\pi) := \max_s \mathrm{KL}\big(\pi(\cdot|s)\|\pi_\beta(\cdot|s)\big)$ and the average total variation distance under ν by $\Delta_{\mathrm{TV}}(\pi,\pi_\beta;\nu) := \mathbb{E}_{s\sim\nu}\big[D_{\mathrm{TV}}(\pi(\cdot|s),\pi_\beta(\cdot|s))\big]$. Assume $|\log\pi_\beta(a|s)| \le B$ for all (s,a). For a given $\lambda \ge 0$ and $\alpha > 0$, π^*_λ satisfies the following maximum-entropy performance lower bound:*

$$\mathcal{J}(\pi^*_\lambda) \ge \mathcal{J}(\pi_\beta) - \frac{2B|\lambda-\alpha|}{1-\gamma}\Delta_{\mathrm{TV}}(\pi^*_\lambda,\pi_\beta;d_{\pi_\beta}) - \frac{4\alpha B + 2\gamma\Big(\epsilon_\beta(\pi^*_\lambda) + \alpha\kappa_\beta(\pi^*_\lambda)\Big)}{(1-\gamma)^2}\Delta_{\mathrm{TV}}(\pi^*_\lambda,\pi_\beta;d_{\pi_\beta}). \tag{54}$$

Proof. By the maximum-entropy performance difference lemma (Lemma D.1),

$$\mathcal{J}(\pi) - \mathcal{J}(\pi_\beta) = \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d_\pi}\Big[\mathbb{E}_{a\sim\pi(\cdot|s)}\big[\tilde{A}^{\pi_\beta}(s,a)\big] - \alpha\, D_{\mathrm{KL}}\big(\pi(\cdot|s)\,\|\,\pi_\beta(\cdot|s)\big)\Big]. \tag{55}$$

Let

$$f_\pi(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}\big[\tilde{A}^{\pi_\beta}(s,a)\big] - \alpha\, D_{\mathrm{KL}}\big(\pi(\cdot|s)\,\|\,\pi_\beta(\cdot|s)\big),$$

so that the RHS of (55) equals $\frac{1}{1-\gamma}\mathbb{E}_{d_\pi}[f_\pi(s)]$. Decomposing around d_{π_β} yields

$$\mathbb{E}_{d_\pi}[f_\pi] = \mathbb{E}_{d_{\pi_\beta}}[f_\pi] + \Big(\mathbb{E}_{d_\pi}[f_\pi] - \mathbb{E}_{d_{\pi_\beta}}[f_\pi]\Big). \tag{56}$$

By definition of $\epsilon_\beta(\pi)$ and $\kappa_\beta(\pi)$, for all s we have

$$|f_\pi(s)| \le \Big|\mathbb{E}_{a\sim\pi(\cdot|s)}\big[\tilde{A}^{\pi_\beta}(s,a)\big]\Big| + \alpha\, D_{\mathrm{KL}}\big(\pi(\cdot|s)\|\pi_\beta(\cdot|s)\big) \le \epsilon_\beta(\pi) + \alpha\kappa_\beta(\pi),$$

hence

$$\Big|\mathbb{E}_{d_\pi}[f_\pi] - \mathbb{E}_{d_{\pi_\beta}}[f_\pi]\Big| \le \|d_\pi - d_{\pi_\beta}\|_1\Big(\epsilon_\beta(\pi) + \alpha\kappa_\beta(\pi)\Big). \tag{57}$$

Moreover, the divergence between discounted state visitation distributions (Lemma 3 in Achiam et al. [2017]) gives

$$\|d_\pi - d_{\pi_\beta}\|_1 \le \frac{2\gamma}{1-\gamma}\,\Delta_{\mathrm{TV}}(\pi,\pi_\beta;d_{\pi_\beta}). \tag{58}$$

Combining (57) and (58) yields

$$\mathbb{E}_{d_\pi}[f_\pi] \ge \mathbb{E}_{d_{\pi_\beta}}[f_\pi] - \frac{2\gamma}{1-\gamma}\Big(\epsilon_\beta(\pi) + \alpha\kappa_\beta(\pi)\Big)\Delta_{\mathrm{TV}}(\pi,\pi_\beta;d_{\pi_\beta}). \tag{59}$$

Substituting (59) into (55) and setting $\pi = \pi^*_\lambda$ yields

$$\mathcal{J}(\pi^*_\lambda) - \mathcal{J}(\pi_\beta) \ge \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\pi_\beta}} f_{\pi^*_\lambda}(s) - \frac{2\gamma}{(1-\gamma)^2}\Big(\epsilon_\beta(\pi^*_\lambda) + \alpha\kappa_\beta(\pi^*_\lambda)\Big)\Delta_{\mathrm{TV}}(\pi^*_\lambda,\pi_\beta;d_{\pi_\beta}). \tag{60}$$

For the first term,

$$\mathbb{E}_{s\sim d_{\pi_\beta}} f_{\pi^*_\lambda}(s) = \mathbb{E}_{s\sim d_{\pi_\beta}}\Big[\mathbb{E}_{a\sim\pi^*_\lambda(\cdot|s)}\tilde{A}^{\pi_\beta}(s,a) - \alpha\, D_{\mathrm{KL}}\big(\pi^*_\lambda(\cdot|s)\|\pi_\beta(\cdot|s)\big)\Big].$$

Next, define the auxiliary objective

$$F(\pi) := \mathbb{E}_{s\sim d_{\pi_\beta}}\Big[\mathbb{E}_{a\sim\pi(\cdot|s)}A^{\pi_\beta}(s,a) - \alpha\, D_{\mathrm{KL}}\big(\pi(\cdot|s)\|\pi_\beta(\cdot|s)\big) + (\lambda-\alpha)\,\mathbb{E}_{a\sim\pi(\cdot|s)}\log\pi_\beta(a|s)\Big],$$

where $A^{\pi_\beta}(s,a)$ denotes the advantage under the original reward $r(s,a)$. Since π^*_λ maximizes F in the CCI optimization problem, we have $F(\pi^*_\lambda) \ge F(\pi_\beta)$. For π_β, note that $\mathbb{E}_{a\sim\pi_\beta(\cdot|s)}[A^{\pi_\beta}(s,a)] = 0$ and $D_{\mathrm{KL}}(\pi_\beta(\cdot|s)\|\pi_\beta(\cdot|s)) = 0$, hence

$$F(\pi_\beta) = (\lambda-\alpha)\, g_{d_{\pi_\beta}}(\pi_\beta), \qquad g_{d_{\pi_\beta}}(\pi) := \mathbb{E}_{s\sim d_{\pi_\beta},\, a\sim\pi(\cdot|s)}[\log\pi_\beta(a|s)].$$

Rearranging gives

$$\mathbb{E}_{s\sim d_{\pi_\beta}}\Big[\mathbb{E}_{a\sim\pi^*_\lambda(\cdot|s)}A^{\pi_\beta}(s,a) - \alpha\, D_{\mathrm{KL}}\big(\pi^*_\lambda(\cdot|s)\,\|\,\pi_\beta(\cdot|s)\big)\Big] \ge -(\lambda-\alpha)\big(g_{d_{\pi_\beta}}(\pi^*_\lambda) - g_{d_{\pi_\beta}}(\pi_\beta)\big). \tag{61}$$

We now relate $\tilde{A}^{\pi_\beta}$ to A^{π_β}. Let $c(s,a) := -\alpha\log\pi_\beta(a|s)$ and $\tilde{r}(s,a) = r(s,a) + c(s,a)$. By linearity of value functions in reward,

$$\tilde{A}^{\pi_\beta}(s,a) = A^{\pi_\beta}(s,a) + A_c^{\pi_\beta}(s,a),$$

where $A_c^{\pi_\beta}$ is the advantage under the reward c alone. Since $|\log\pi_\beta(a|s)| \le B$, we have $|c(s,a)| \le \alpha B$, which implies $\|A_c^{\pi_\beta}\|_\infty \le \frac{2\alpha B}{1-\gamma}$. Moreover, $\mathbb{E}_{a\sim\pi_\beta(\cdot|s)}[A_c^{\pi_\beta}(s,a)] = 0$ for all s, hence for any π and s,

$$\begin{aligned}
\Big|\mathbb{E}_{a\sim\pi(\cdot|s)}[A_c^{\pi_\beta}(s,a)]\Big| &= \Big|\int\big(\pi(a|s) - \pi_\beta(a|s)\big)\,A_c^{\pi_\beta}(s,a)\,da\Big| \\
&\le 2\|A_c^{\pi_\beta}\|_\infty\, D_{\mathrm{TV}}\big(\pi(\cdot|s),\pi_\beta(\cdot|s)\big) \\
&\le \frac{4\alpha B}{1-\gamma}\, D_{\mathrm{TV}}\big(\pi(\cdot|s),\pi_\beta(\cdot|s)\big).
\end{aligned} \tag{62}$$

Therefore, for $\pi = \pi^*_\lambda$,

$$\mathbb{E}_{s\sim d_{\pi_\beta}}\mathbb{E}_{a\sim\pi^*_\lambda(\cdot|s)}\tilde{A}^{\pi_\beta}(s,a) \geq \mathbb{E}_{s\sim d_{\pi_\beta}}\mathbb{E}_{a\sim\pi^*_\lambda(\cdot|s)}A^{\pi_\beta}(s,a) - \frac{4\alpha B}{1-\gamma}\,\Delta_{\mathrm{TV}}(\pi^*_\lambda,\pi_\beta;d_{\pi_\beta}). \tag{63}$$

Combining (61) and (63) yields

$$\mathbb{E}_{s\sim d_{\pi_\beta}} f_{\pi^*_\lambda}(s) \geq -(\lambda-\alpha)\big(g_{d_{\pi_\beta}}(\pi^*_\lambda) - g_{d_{\pi_\beta}}(\pi_\beta)\big) - \frac{4\alpha B}{1-\gamma}\,\Delta_{\mathrm{TV}}(\pi^*_\lambda,\pi_\beta;d_{\pi_\beta}). \tag{64}$$

Finally, for each state s, define

$$\Delta g_s := \mathbb{E}_{a\sim\pi^*_\lambda(\cdot|s)}[\log\pi_\beta(a|s)] - \mathbb{E}_{a\sim\pi_\beta(\cdot|s)}[\log\pi_\beta(a|s)].$$

Then

$$|\Delta g_s| \leq \int \big|\pi^*_\lambda(a|s) - \pi_\beta(a|s)\big|\cdot|\log\pi_\beta(a|s)|\,da \leq 2B\,D_{\mathrm{TV}}\big(\pi^*_\lambda(\cdot|s),\pi_\beta(\cdot|s)\big), \tag{65}$$

and averaging over $s\sim d_{\pi_\beta}$ gives

$$\big|g_{d_{\pi_\beta}}(\pi^*_\lambda) - g_{d_{\pi_\beta}}(\pi_\beta)\big| \leq 2B\,\Delta_{\mathrm{TV}}(\pi^*_\lambda,\pi_\beta;d_{\pi_\beta}). \tag{66}$$

Substituting (66) into (64) yields

$$\mathbb{E}_{s\sim d_{\pi_\beta}} f_{\pi^*_\lambda}(s) \geq -2B|\lambda-\alpha|\,\Delta_{\mathrm{TV}}(\pi^*_\lambda,\pi_\beta;d_{\pi_\beta}) - \frac{4\alpha B}{1-\gamma}\,\Delta_{\mathrm{TV}}(\pi^*_\lambda,\pi_\beta;d_{\pi_\beta}). \tag{67}$$

Plugging (67) into (60) completes the proof and yields Eq. (54). □

Theorem D.3 (Performance Lower Bound for a Parametric Policy)**.** *Define the suboptimality gap of a parametric policy π_θ by*

$$\delta^{\mathrm{sub}}_\theta(\lambda) := F_\lambda(\pi^*_\lambda) - F_\lambda(\pi_\theta)\ \geq\ 0, \tag{68}$$

where $F_\lambda(\pi) = \mathbb{E}_{s\sim d_{\pi_\beta}}\Big[\mathbb{E}_{a\sim\pi(\cdot|s)}\tilde{A}^{\pi_\beta}(s,a) - \alpha\,D_{\mathrm{KL}}\big(\pi(\cdot|s)\,\|\,\pi_\beta(\cdot|s)\big) + (\lambda-\alpha)\,\mathbb{E}_{a\sim\pi(\cdot|s)}\log\pi_\beta(a|s)\Big]$.

Then π_θ satisfies

$$\begin{aligned}\mathcal{J}(\pi_\theta) \geq\ \mathcal{J}(\pi_\beta) &- \frac{2B|\lambda-\alpha|}{1-\gamma}\,\Delta_{\mathrm{TV}}\big(\pi_\theta,\pi_\beta;d_{\pi_\beta}\big) - \frac{\delta^{\mathrm{sub}}_\theta(\lambda)}{1-\gamma}\\ &- \frac{4\alpha B + 2\gamma\Big(\epsilon_\beta(\pi_\theta) + \alpha\kappa_\beta(\pi_\theta)\Big)}{(1-\gamma)^2}\,\Delta_{\mathrm{TV}}\big(\pi_\theta,\pi_\beta;d_{\pi_\beta}\big).\end{aligned} \tag{69}$$

Proof. The derivation follows the same decomposition as in the proof of Theorem D.2. Applying Eq. (55) with $\pi = \pi_\theta$ and using the same occupancy-shift bound yields

$$\mathcal{J}(\pi_\theta) - \mathcal{J}(\pi_\beta) \geq \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d_{\pi_\beta}} f_{\pi_\theta}(s) - \frac{2\gamma}{(1-\gamma)^2}\Big(\epsilon_\beta(\pi_\theta) + \alpha\kappa_\beta(\pi_\theta)\Big)\Delta_{\mathrm{TV}}(\pi_\theta,\pi_\beta;d_{\pi_\beta}), \tag{70}$$

where

$$f_\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\big[\tilde{A}^{\pi_\beta}(s,a)\big] - \alpha\,D_{\mathrm{KL}}\big(\pi(\cdot|s)\,\|\,\pi_\beta(\cdot|s)\big).$$

We first lower bound the shaped-advantage term by the original advantage term, as in Eq. (63):

$$\mathbb{E}_{s\sim d_{\pi_\beta},\,a\sim\pi(\cdot|s)}[\tilde{A}^{\pi_\beta}(s,a)] \geq \mathbb{E}_{s\sim d_{\pi_\beta},\,a\sim\pi(\cdot|s)}[A^{\pi_\beta}(s,a)] - \frac{4\alpha B}{1-\gamma}\,\Delta_{\mathrm{TV}}(\pi,\pi_\beta;d_{\pi_\beta}). \tag{71}$$

Next, we convert the suboptimality gap in Eq. (68) into a baseline-anchored inequality. Since π^*_λ maximizes F_λ, we have $F_\lambda(\pi^*_\lambda) \geq F_\lambda(\pi_\beta)$. Combining this with Eq. (68) gives

$$F_\lambda(\pi_\theta) = F_\lambda(\pi^*_\lambda) - \delta^{\mathrm{sub}}_\theta(\lambda) \geq F_\lambda(\pi_\beta) - \delta^{\mathrm{sub}}_\theta(\lambda). \tag{72}$$

For π_β, $\mathbb{E}_{a\sim\pi_\beta}[A^{\pi_\beta}(s,a)] = 0$ and $D_{\mathrm{KL}}(\pi_\beta\|\pi_\beta) = 0$, hence $F_\lambda(\pi_\beta) = (\lambda-\alpha)\, g_{d_{\pi_\beta}}(\pi_\beta)$ where

$$g_{d_{\pi_\beta}}(\pi) := \mathbb{E}_{s\sim d_{\pi_\beta},\, a\sim\pi(\cdot|s)}[\log \pi_\beta(a|s)].$$

Rearranging Eq. (72) yields

$$\begin{aligned}\mathbb{E}_{s\sim d_{\pi_\beta}, a\sim\pi_\theta(\cdot|s)}\big[A^{\pi_\beta}(s,a)\big] - \alpha\, \mathbb{E}_{s\sim d_{\pi_\beta}}\Big[D_{\mathrm{KL}}\big(\pi_\theta(\cdot|s)\,\|\,\pi_\beta(\cdot|s)\big)\Big] \\ \geq -(\lambda-\alpha)\Big(g_{d_{\pi_\beta}}(\pi_\theta) - g_{d_{\pi_\beta}}(\pi_\beta)\Big) - \delta_\theta^{\mathrm{sub}}(\lambda).\end{aligned} \tag{73}$$

Finally, we bound the density-shaping difference by total variation. For each state s, define

$$\Delta g_s := \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[\log \pi_\beta(a|s)] - \mathbb{E}_{a\sim\pi_\beta(\cdot|s)}[\log \pi_\beta(a|s)].$$

Then

$$|\Delta g_s| \leq 2B\, D_{\mathrm{TV}}\big(\pi_\theta(\cdot|s), \pi_\beta(\cdot|s)\big), \tag{74}$$

and averaging over $s\sim d_{\pi_\beta}$ yields

$$\big|g_{d_{\pi_\beta}}(\pi_\theta) - g_{d_{\pi_\beta}}(\pi_\beta)\big| \leq 2B\, \Delta_{\mathrm{TV}}(\pi_\theta, \pi_\beta; d_{\pi_\beta}). \tag{75}$$

Combining Eq. (71), Eq. (73), and Eq. (75) yields

$$\begin{aligned}\mathbb{E}_{s\sim d_{\pi_\beta}} f_{\pi_\theta}(s) &= \mathbb{E}_{s\sim d_{\pi_\beta}}\Big[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\tilde{A}^{\pi_\beta}(s,a) - \alpha\, D_{\mathrm{KL}}\big(\pi_\theta(\cdot|s)\,\|\,\pi_\beta(\cdot|s)\big)\Big] \\ &\geq -2B|\lambda-\alpha|\, \Delta_{\mathrm{TV}}(\pi_\theta, \pi_\beta; d_{\pi_\beta}) - \frac{4\alpha B}{1-\gamma}\, \Delta_{\mathrm{TV}}(\pi_\theta, \pi_\beta; d_{\pi_\beta}) - \delta_\theta^{\mathrm{sub}}(\lambda).\end{aligned} \tag{76}$$

Plugging Eq. (76) into Eq. (70) completes the proof and yields Eq. (69). □

E Additional Experiments

E.1 Linear-Decay Schedules with Different Initial Values

We evaluate linear-decay schedules initialized at $\lambda = 100$ and $\lambda = 1$, each decreasing to 0 over training. We compare both schedules with ACPO on the nine D4RL Gym-MuJoCo datasets, keeping the training budget and all other experimental settings unchanged. Tab. 7 reports the results.

Table 7: ACPO and linear-decay schedules with different initial values on D4RL Gym-MuJoCo.

Dataset	ACPO	Linear Decay ($100\to0$)	Linear Decay ($1\to0$)
HalfCheetah-med	**52.5**	45.1	45.3
Hopper-med	**97.9**	61.1	63.9
Walker2d-med	**88.0**	69.2	69.5
HalfCheetah-med-rep	**47.5**	43.0	42.5
Hopper-med-rep	**100.6**	83.4	100.2
Walker2d-med-rep	**93.6**	50.9	72.7
HalfCheetah-med-exp	**97.2**	84.1	94.0
Hopper-med-exp	**112.3**	105.7	104.1
Walker2d-med-exp	**113.1**	108.1	108.8
Locomotion Total	**802.7**	650.6	701.0

E.2 Comparison with Fixed Constraint Regimes

We compare ACPO with three variants that keep λ fixed throughout training: $\lambda = 100$ (practical wBC), $\lambda = \alpha$ (KL-based density regularization), and $\lambda = 0$ (support constraint). We use the same policy-evaluation updates as ACPO and keep the remaining settings consistent with the corresponding D4RL and NeoRL2 experiments. For Gym-MuJoCo and NeoRL2, $\alpha = 0.1$ and $\alpha = 0.5$, respectively. Tab. 8 and 9 report the full results.

Table 8: Comparison with fixed constraint regimes on D4RL.

Dataset	wBC ($\lambda = 100$)	KL ($\lambda = \alpha$)	Support ($\lambda = 0$)	ACPO
HalfCheetah-med	42.6	47.9	48.3	**52.5**
Hopper-med	52.9	59.8	60.3	**97.9**
Walker2d-med	75.3	83.1	82.7	**88.0**
HalfCheetah-med-rep	36.6	44.8	44.3	**47.5**
Hopper-med-rep	18.1	69.8	92.1	**100.6**
Walker2d-med-rep	26.0	78.1	69.8	**93.6**
HalfCheetah-med-exp	55.2	64.9	83.5	**97.2**
Hopper-med-exp	52.5	100.1	93.8	**112.3**
Walker2d-med-exp	107.5	110.0	109.0	**113.1**
Locomotion Total	466.7	658.5	683.8	**802.7**
Antmaze-u	54.6	80.0	54.0	**90.0**
Antmaze-u-d	45.6	52.0	60.0	**86.0**
Antmaze-m-p	0.0	0.0	52.0	**76.0**
Antmaze-m-d	0.0	0.2	50.0	**76.0**
Antmaze-l-p	0.0	0.0	40.0	**56.0**
Antmaze-l-d	0.0	0.0	20.0	**48.0**
Antmaze Total	100.2	132.2	276.0	**432.0**
Kitchen-mixed	47.5	57.5	60.0	**72.5**
Kitchen-partial	33.9	55.0	47.5	**64.0**
Kitchen Total	81.4	112.5	107.5	**136.5**

Table 9: Comparison with fixed constraint regimes on NeoRL2.

Dataset	wBC ($\lambda = 100$)	KL ($\lambda = \alpha$)	Support ($\lambda = 0$)	ACPO
Pipeline	29.9	36.7	57.5	**95.2**
Simglucose	25.2	32.8	70.5	**94.9**
RocketRecovery	-39.3	-61.6	-44.1	**84.6**
RandomFrictionHopper	9.9	2.7	0.5	**33.7**
DMSD	69.7	71.5	70.0	**74.9**
Fusion	54.7	54.0	55.2	**61.8**
SafetyHalfCheetah	8.7	53.5	65.4	**67.1**
NeoRL2 Total	158.8	189.6	275.0	**512.2**

F Experimental Details

F.1 Implementation Details for D4RL Experiments

The vast majority of hyperparameters used in our D4RL experiments are summarized in Tab. 10.

The dataset abbreviations used in Tab. 2 and their corresponding full names are listed in Tab. 11.

Training curves for all D4RL tasks are provided in Fig. 5.

F.2 Implementation Details for NeoRL2 Experiments

The vast majority of hyperparameters used in our NeoRL2 experiments are summarized in Tab. 12.

Training curves for all NeoRL2 tasks are provided in Fig. 6.

F.3 Additional Details for Constraint-Regime Comparisons and Adaptive λ Dynamics

In Subsection 4.3, we keep all hyperparameters consistent with Tab. 10 and Tab. 12. We additionally specify the key settings in the main text: *"For Gym-MuJoCo tasks, we set* $\alpha = 0.1$. *For NeoRL2 tasks, we set* $\alpha = 0.5$. *We instantiate three special cases by choosing: (i) practical wBC:* $\lambda = 100$ *to approximate the wBC regime, (ii) AWAC:* $\lambda = \alpha$*, and (iii) InAC:* $\lambda = 0$*."*

Table 10: D4RL hyperparameters, where $s_{\dim}$ and $a_{\dim}$ are the state and action dimensions, respectively.

Hyperparameter	Value	Task
Actor architecture	$s_{\dim}$ – 256 – 256 – $2a_{\dim}$	All
Q-network	$(s_{\dim}+a_{\dim})$ – 256 – 256 – 1	All
Value network	$s_{\dim}$ – 256 – 256 – 1	All
Behavior network	$s_{\dim}$ – 512 – 512 – $2a_{\dim}$	All
Activation	ReLU	All
Learning rate (actor & critic)	$5e-4 \rightarrow 1e-4$ (CosineAnnealingLR)	All
Learning rate (behavior policy)	$1e-4$	All
Learning rate (λ)	$1e-5$	All
Target Q soft update rate τ	$5e-3$	All
Evaluation epochs	10	Gym-MuJoCo & Kitchen
	50	AntMaze
Evaluation frequency	$5e3$	All
Batch size	256	All
Discount factor γ	0.99	All
Temperature α	0.1	All
Initial λ	0.1 ($\lambda = \alpha$)	All
ϵ in Eq. (7)	-5.0	Hopper
	-0.2	HalfCheetah & AntMaze
	-0.7	Walker2d
	-1.0	Kitchen

Table 11: Dataset abbreviations for D4RL.

Dataset abbreviation	Dataset full name
halfcheetah-med	halfcheetah-medium-v2
hopper-med	hopper-medium-v2
walker2d-med	walker2d-medium-v2
halfcheetah-med-rep	halfcheetah-medium-replay-v2
hopper-med-rep	hopper-medium-replay-v2
walker2d-med-rep	walker2d-medium-replay-v2
halfcheetah-med-exp	halfcheetah-medium-expert-v2
hopper-med-exp	hopper-medium-expert-v2
walker2d-med-exp	walker2d-medium-expert-v2
antmaze-u	antmaze-umaze-v2
antmaze-u-d	antmaze-umaze-diverse-v2
antmaze-m-p	antmaze-medium-play-v2
antmaze-m-d	antmaze-medium-diverse-v2
antmaze-l-p	antmaze-large-play-v2
antmaze-l-d	antmaze-large-diverse-v2
kitchen-m	kitchen-mixed-v0
kitchen-p	kitchen-partial-v0

F.4 Additional Details for Adaptive λ versus Manual Schedules

In Subsection 4.4, we use the D4RL Gym-MuJoCo settings in Tab. 10 and compare ACPO with three manual schedules on all nine datasets: linear decay from $\lambda = 100$ to 0, a two-stage schedule with $\lambda = \alpha$ for the first half and $\lambda = 0$ for the second half, and an increase-then-decrease schedule from α to 2α in the first third and then to 0 over the remaining training.

Table 12: NeoRL2 hyperparameters. Architectures are denoted by MLP layer sizes, where s_{dim} and a_{dim} are the state and action dimensions, respectively.

Hyperparameter	Value	Task
Actor architecture	s_{dim} – 256 – 256 – $2a_{\mathrm{dim}}$	All
Q-network	$(s_{\mathrm{dim}}+a_{\mathrm{dim}})$ – 256 – 256 – 1	All
Value network	s_{dim} – 256 – 256 – 1	All
Behavior network	s_{dim} – 512 – 512 – $2a_{\mathrm{dim}}$	All
Activation	ReLU	All
Learning rate (actor & critic)	$5e-4 \rightarrow 1e-4$ (CosineAnnealingLR)	All
Learning rate (behavior policy)	$1e-4$	All
Learning rate (λ)	$1e-5$	All
Target Q soft update rate τ	$5e-3$	All
Evaluation epochs	10	All
Evaluation frequency	$5e3$	All
Batch size	256	All
Discount factor γ	0.99	All
Temperature α	0.5	All
Initial λ	0.5 ($\lambda = \alpha$)	All
ϵ in Eq. (7)	-5.0	Simglucose & Pipeline
	-10.0	SafetyHalfCheetah & RocketRecovery & Fusion
	-2.0	RandomFrictionHopper
	-0.5	DMSD

F.5 Additional Details for Behavior Density Estimator Ablation

In Subsection 4.5, we run inference with the trained Gaussian and CVAE behavior models on the full datasets of `walker2d-medium-v2` and `walker2d-medium-replay-v2`, producing the distribution curves shown in Fig. 3. In addition to $\xi = 0.3$, we also evaluate $\xi \in \{0.1, 0.5, 0.8\}$, with the corresponding curves reported in Fig. 4. The conclusions remain unchanged: compared with the Gaussian estimator, the CVAE likelihood on perturbed OOD actions is not consistently suppressed.

F.6 Additional Details for Sensitivity Analysis of ϵ

In Subsection 4.6, we keep the vast majority of hyperparameters consistent with Tab. 10, while varying ϵ over $\{-0.2, -0.7, -5.0\}$.

F.7 Additional Details for Sensitivity Analysis of η_λ

In Subsection 4.7, we use the D4RL Gym-MuJoCo settings in Tab. 10 and evaluate $\eta_\lambda \in \{1\mathrm{e}{-5}, 1\mathrm{e}{-4}, 3\mathrm{e}{-4}\}$ on all nine datasets, keeping the other hyperparameters fixed.

G Broader Impacts

This work may have positive societal impact by improving offline RL in domains where additional online interaction is expensive or difficult. However, the method is evaluated only on benchmark datasets, and its behavior in real-world settings may differ substantially under model error, or imperfect data coverage. Moreover, policies learned from offline data may inherit limitations or biases of the underlying datasets. Therefore, practical deployment should require careful domain-specific validation and additional safety assessment.

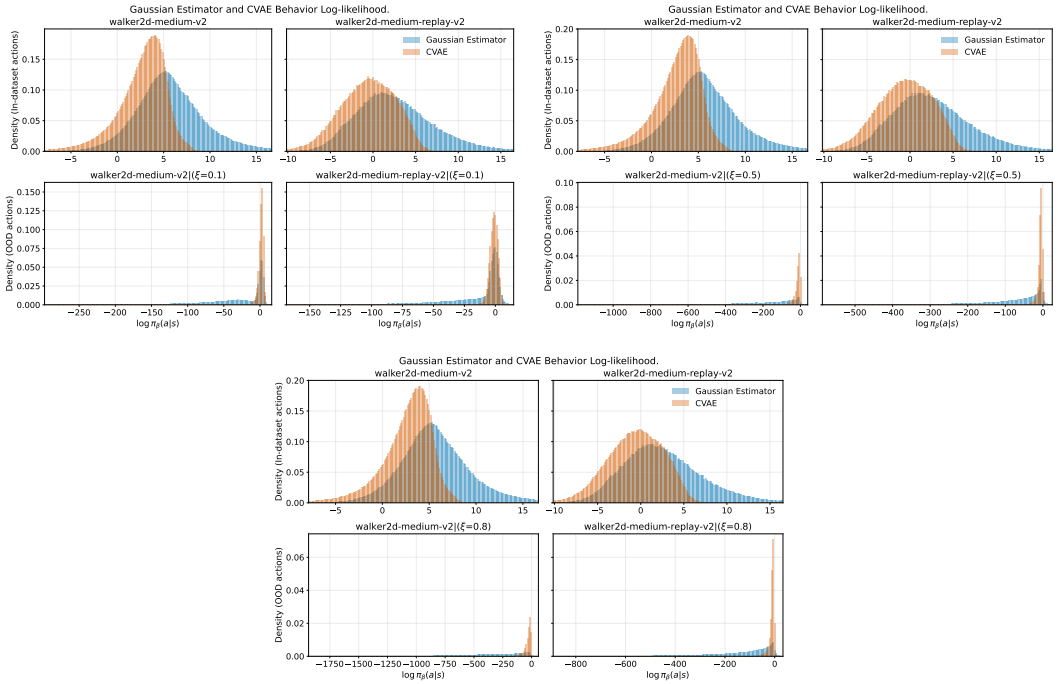


Figure 4: Gaussian and CVAE behavior models under $\xi \in \{0.1, 0.5, 0.8\}$.

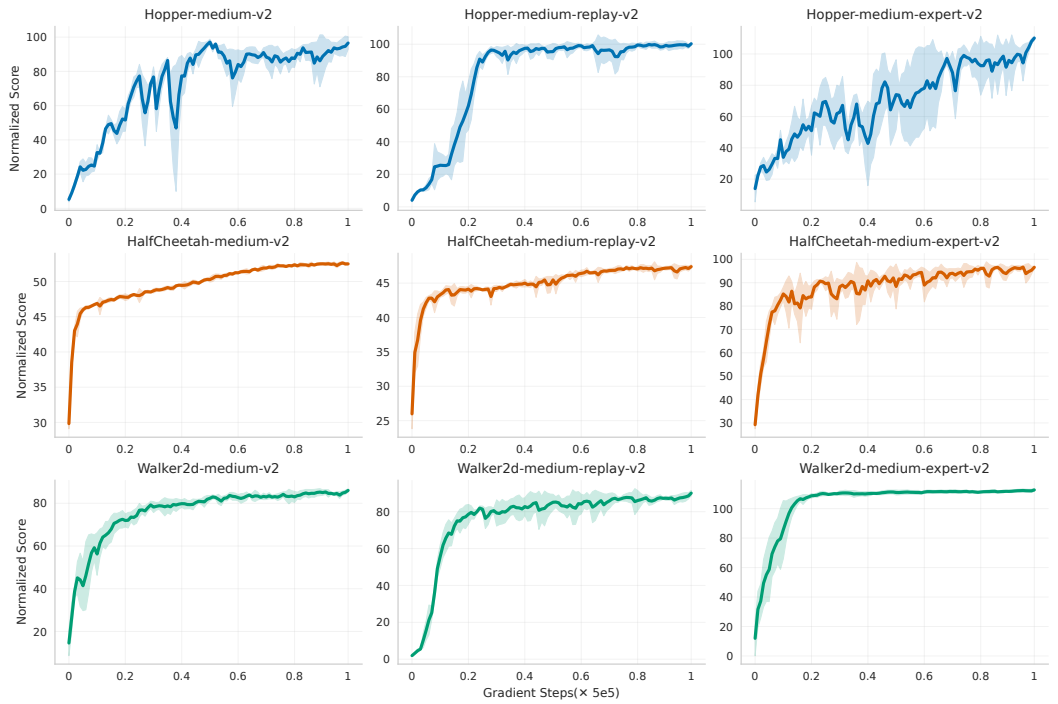


Figure 5: D4RL Learning Curves.

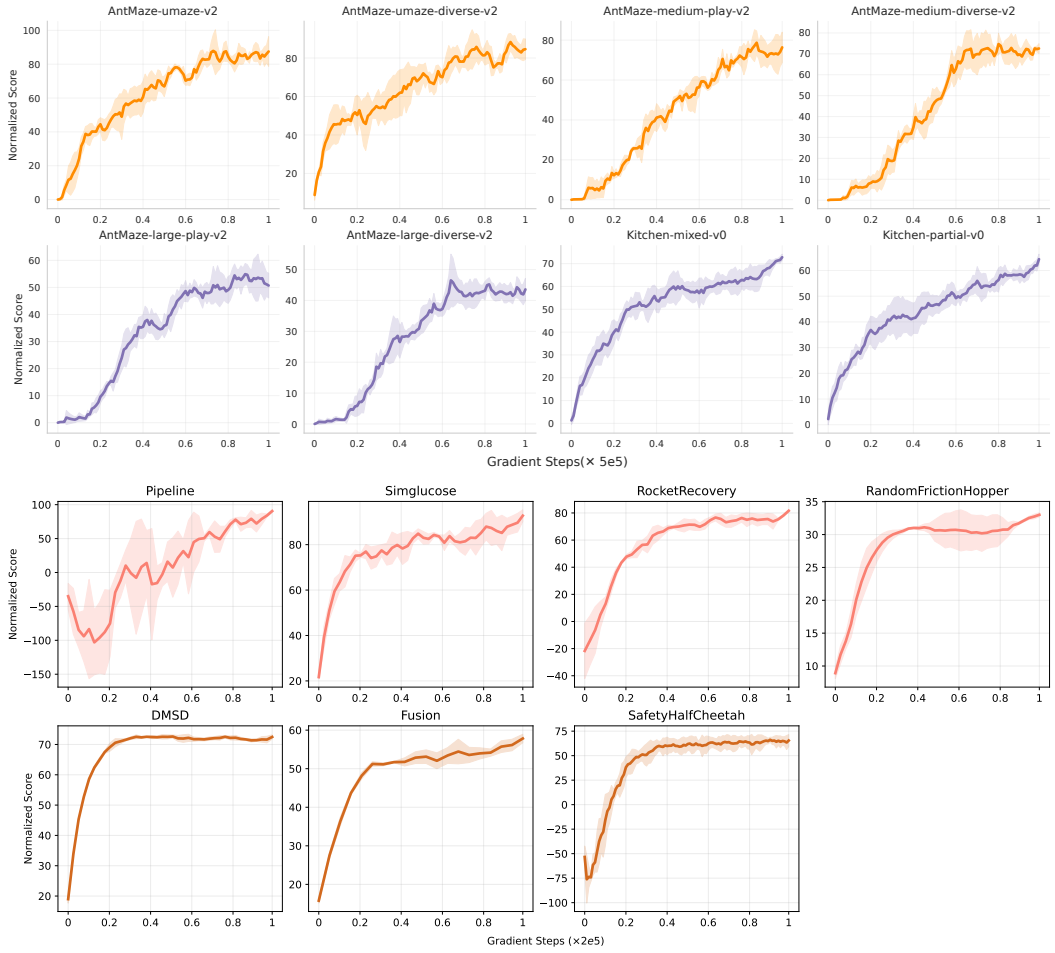


Figure 6: Antmaze, Kitchen and NeoRL2 Learning Curves.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the main contributions of the paper: a unified CCI framework, the ACPO algorithm with adaptive constraint tuning, theoretical analysis, and empirical evaluation on D4RL and NeoRL2.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in the conclusion through the identified future directions and in the broader-impact discussion (Appendix G), including that the current method is evaluated mainly on standard offline RL benchmarks and has not been validated in real-world deployment settings. In particular, practical performance may be affected by factors such as dataset quality and coverage, as well as the additional safety requirements of deployment in real decision-making environments.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The theoretical results are stated in Subsection 3.3 with explicit assumptions, propositions, lemmas, and theorems. Full proofs of the monotonicity result, the maximum-entropy performance difference lemma, and the two performance lower bounds are provided in Appendix C, D.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the information needed to reproduce the main experimental results, including detailed experimental settings and hyperparameters in Appendix F. In addition, the source code is provided in the supplementary material to support reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiments use publicly available offline RL benchmarks, including D4RL and NeoRL2. We also provide the code in the supplementary material, together with the implementation details in Appendix F, to support reproduction of the main results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All essential experimental details are provided in Appendix F, including hyperparameter settings, architectures, and evaluation protocols for the reported benchmarks and ablations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main experimental results are reported with mean and standard deviation over multiple random seeds, as shown in Tables 2, 3, 5.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources used for the experiments are described in Appendix F. In particular, we report the hardware platform used to run the experiments, including the CPU, system memory, and GPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Broader impacts are discussed in Appendix G, including both potential benefits and possible risks or limitations associated with deployment.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper does not release high-risk models or sensitive datasets that would require special safeguards for responsible release.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All existing assets used in this work, including benchmark datasets and baseline methods, are properly cited and described in the paper.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide the implementation code of the proposed algorithm in the supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used only for spelling and grammar checking.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.