
Convergence Theorems for Entropy-Regularized and Distributional Reinforcement Learning

Yash Jhaveri*
Rutgers University–Newark

Harley Wiltzer*
Mila–Québec AI Institute
McGill University

Patrick Shafto
Rutgers University–Newark

Marc G. Bellemare†
Mila–Québec AI Institute
McGill University

David Meger
Mila–Québec AI Institute
McGill University

Abstract

In the pursuit of finding an optimal policy, reinforcement learning (RL) methods generally ignore the properties of learned policies apart from their expected return. Thus, even when successful, it is difficult to characterize which policies will be learned and what they will do. In this work, we present a theoretical framework for policy optimization that guarantees convergence to a particular optimal policy, via vanishing entropy regularization and a *temperature decoupling gambit*. Our approach realizes an interpretable, diversity-preserving optimal policy as the regularization temperature vanishes and ensures the convergence of policy derived objects–value functions and return distributions. In a particular instance of our method, for example, the realized policy samples all optimal actions uniformly. Leveraging our temperature decoupling gambit, we present an algorithm that estimates, to arbitrary accuracy, the return distribution associated to its interpretable, diversity-preserving optimal policy.

1 Introduction

In generic Markov Decision Processes (MDPs), many optimal policies exist. Thus, while certain policy optimization approaches can ensure convergent approximation to an optimal policy, they do not have control over which states these policies will visit, which actions they will play, or which long-term returns they can achieve. Indeed, the non-uniqueness of optimal policies renders any discussion of the properties of an optimal policy ambiguous, beyond its expected value.

A partial remedy to this problem is to regularize the RL objective in order to induce uniqueness. One popular approach to regularization is to penalize the value of a policy according to its KL divergence to a reference policy π^{ref} . This branch of RL is known as entropy-regularized RL (ERL). In ERL, for any positive regularization weight τ (also known as *temperature*), one and only one policy is optimal. Moreover, in a tabular MDP, τ -optimal policies and their derived objects (value functions, occupancy measures, and return distributions) converge to classically optimal policies and their derived objects. However, beyond tabular MDPs, the evolution of τ -optimal quantities, as a function of the temperature, is not well understood. Thus, as we decay the temperature to zero, we are, in some sense, back to where we started: living in ambiguity.

In this work, we introduce a *temperature decoupling gambit*, through which we can guarantee the convergence of resulting policies and their derived objects in the vanishing temperature limit.

*Equal contribution. Correspondence to yash.jhaveri@rutgers.edu, wiltzerh@mila.quebec.

†CIFAR AI Chair.

Much like how a gambit in chess sacrifices an immediate and shallow proxy of the objective (e.g., material count) for a long term positional advantage, the temperature decoupling gambit plays notably *suboptimal* policies for the τ -ERL objective to ensure convergence to RL optimality as $\tau \rightarrow 0$. This scheme entails estimating action-values under a target regularization temperature while playing policies with an amplified temperature. Furthermore, we characterize this limiting policy as a modification of the reference policy which “filters out” suboptimal actions. Even when τ -optimal policies converge in the vanishing temperature limit (such as in tabular MDPs), the limiting policy produced by the temperature decoupling gambit is distinct from the limiting policy found otherwise. The limiting policy found via our gambit preserves, quantifiably, more state-wise action diversity. Moreover, we show that this limiting policy achieves a notion of *reference-optimality* for RL, characterized by a new Bellman-like equation, whose unique fixed point upper bounds the (RL) performance of τ -optimal policies in general.

Our analysis additionally sheds light on the convergence of return distributions—the central objects of study in distributional RL (DRL) [6]. While optimal policies achieve the same return in expectation, they may vary drastically in other statistics, such as variance. In safety-critical applications, for example, understanding the distribution over returns is crucial. DRL provides techniques for estimating return distributions, primarily based on distributional dynamic programming methods which generalize dynamic programming approaches for estimating expected returns. However, it is well-known that existing distributional methods do not produce convergent iterates in the control setting [5]. Leveraging our convergence results for policies in ERL, we define the first algorithm for accurately estimating a reference-optimal return distribution, the return distribution associated to the interpretable, diverse policy realized by the temperature decoupling gambit.

2 Preliminaries

Given a Borel set $S \subset \mathbb{R}^n$, for some $n \in \mathbb{N}$, we let $M(S)$ and $M_b(S)$ denote the space of Borel measurable and bounded Borel measurable functions on S respectively. We let $\mathcal{P}(S)$ denote the space of Borel probability measures on S . From now on, measurability will always be with respect to Borel sets. Moreover, for any $\rho \in \mathcal{P}(Y)$ with $Y \subset \mathbb{R}^m$ and any measurable function $f : Y \rightarrow S$, the *push-forward* of ρ by f is $f_{\#}\rho := \rho \circ f^{-1} \in \mathcal{P}(S)$. Here f^{-1} is the preimage of f .

We single out two particular functions. The function $\text{proj}^{Y_k} : Y_1 \times \dots \times Y_n \rightarrow Y_k$ defined by $\text{proj}^{Y_k}(y_1, \dots, y_k, \dots, y_n) := y_k$ is the *projection function* of Y^n onto Y_k . We note that the push-forward of the projection map is marginalization: $\nu^\mu := \text{proj}_{\#}^Y \mu$ is the Y -marginal of $\mu \in \mathcal{P}(Y \times Z)$. The *bootstrap function* $b_{a,b} : \mathbb{R} \rightarrow \mathbb{R}$ is defined by $b_{a,b}(z) := a + bz$ from [6].

Our analysis works with conditional distributions, which we formalize as probability kernels, as well as a tensor-product notation constructing product measures and for disintegrating product measures. For any $Y \subset \mathbb{R}^m$ and $Z \subset \mathbb{R}^n$, the space of (Borel) *probability kernels* from Y to Z , denoted $K(Y, \mathcal{P}(Z))$, is the set of all indexed measures λ for which $y \mapsto \lambda_y(S)$ is measurable for each $S \in \mathcal{B}(Z)$, the Borel subsets of Z . Given $\lambda \in K(Y, \mathcal{P}(Z))$ and $\rho \in \mathcal{P}(Y)$, the *generalized product measure* $\lambda_{-} \otimes \rho \in \mathcal{P}(Y \times Z)$ is defined as follows:

$$\int \phi d(\lambda_{-} \otimes \rho) := \int \left[\int \phi(y, z) d\lambda_y(z) \right] d\rho(y) \quad \forall \phi \in M(Y \times Z).$$

Additionally, we can disintegrate any $\mu \in \mathcal{P}(Y \times Z)$ as a generalized product between either of its marginals and the induced conditional probabilities:

$$\mu = \pi_{-}^{\mu} \otimes \nu^{\mu} \quad \text{where} \quad \nu^{\mu} := \text{proj}_{\#}^Y \mu \quad \text{and} \quad \pi^{\mu} \in K(Y, \mathcal{P}(Z)).$$

An important subset of $K(Y, \mathcal{P}(Z))$ consists of those kernels with bounded p th moments,

$$\bar{K}^p(Y, \mathcal{P}(Z)) := \left\{ \lambda \in K(Y, \mathcal{P}(Z)) : \sup_{y \in Y} \int |z|^p d\lambda_y(z) < \infty \right\} \quad \text{for } p \in [1, \infty),$$

which can be metrized as complete metric spaces. In this work, we consider their metrization via the following metrics based on the Wasserstein metrics [40] d_p ,

$$\bar{d}_p(\lambda, \lambda') := \sup_y d_p(\lambda_y, \lambda'_y) \quad \text{and} \quad d_{p;q,\omega}(\lambda, \lambda') := \left(\int d_p(\lambda_y, \lambda'_y)^q d\omega(y) \right)^{1/q}, \quad (2.1)$$

where $p, q \in [1, \infty)$ and $\omega \in \mathcal{P}(\mathcal{Y})$. These metrize topologies on $\bar{K}^p(\mathcal{Y}, \mathcal{P}(\mathcal{Z}))$ akin to the weak topology on probability measures with finite p th moments.

2.1 Markov Decision Processes and Reinforcement Learning

A discounted MDP is a five-tuple (X, A, P, r, γ) . Here $X \subset \mathbb{R}^m$ is the *state space*, $A \subset \mathbb{R}^n$ is the *action space*, $r \in M_b(X \times A)$ is the *reward function*, and $\gamma \in (0, 1)$ is the *discount factor*.¹

Central to RL are policies. A *policy* is a probability kernel $\pi \in K(X, \mathcal{P}(A))$. Policies induce state transition kernels $\hat{P}^\pi$ as well as a state-action transition kernels $\check{P}^\pi$, given by

$$\hat{P}_x^\pi := \text{proj}_\#^X(P_{x,-} \otimes \pi_x) \in \mathcal{P}(X) \quad \text{and} \quad \check{P}_{x,a}^\pi := \pi_- \otimes P_{x,a} \in \mathcal{P}(X \times A),$$

respectively. Therefore, policies yield sequences of states as well as state-action pairs, labeled $(S_t^\pi)_{t \geq 0}$ and $(X_t^\pi, A_t^\pi)_{t \geq 0}$ respectively, whose sequences of laws $(\nu_t^\pi)_{t \geq 0}$ and $(\mu_t^\pi)_{t \geq 0}$ are given by

$$\nu_{t+1}^\pi := \hat{P}^\pi \nu_t^\pi \quad \text{with} \quad \nu_0^\pi := \nu_0 \quad \text{and} \quad \mu_{t+1}^\pi := \check{P}^\pi \mu_t^\pi \quad \text{with} \quad \mu_0^\pi := \pi_- \otimes \nu_0$$

for some $\nu_0 \in \mathcal{P}(X)$. Given $\nu_0 \in \mathcal{P}(X)$, the long-term behavior of any policy π can be encoded via its (discounted, state-action) *occupancy measure* μ^π , the set of which we denote by $\mathcal{O}(\nu_0)$,

$$\mathcal{O}(\nu_0) := \left\{ \mu^\pi \in \mathcal{P}(X \times A) : \mu^\pi := (1 - \gamma) \sum_{t \geq 0} \gamma^t \mu_t^\pi \text{ for some } \pi \in K(X, \mathcal{P}(A)) \right\}.$$

Policies also induce *return distribution functions* $\zeta^\pi \in K(X \times A, \mathcal{P}(\mathbb{R}))$ and $\eta^\pi \in K(X, \mathcal{P}(\mathbb{R}))$,

$$\zeta_{x,a}^\pi := \text{law} \left(\sum_{t \geq 0} \gamma^t r(X_t^\pi, A_t^\pi) \middle| X_0^\pi = x, A_0^\pi = a \right) \quad \text{and} \quad \eta_x^\pi := \text{proj}_\#^\mathbb{R}(\zeta_{x,-}^\pi \otimes \pi_x)$$

whose means, the *action-value function* $q^\pi \in M_b(X \times A)$ and the *value function* $v^\pi \in M_b(X)$,

$$q^\pi(x, a) := \mathbf{E}_{Z \sim \zeta_{x,a}^\pi} [Z] \quad \text{and} \quad v^\pi(x) := \mathbf{E}_{G \sim \eta_x^\pi} [G],$$

lead to the RL objective: find a $\pi^* \in K(X, \mathcal{P}(A))$ such that $q^{\pi^*} \geq q^\pi$ for all π . Such a policy is called *optimal*. Generally, many policies are optimal. However, their associated action-value functions are identical (see [32]). We denote this optimal action-value function by q^* .

2.2 Entropy-Regularized Reinforcement Learning

In ERL, the value of a policy is penalized by how far it diverges from a fixed *reference policy* $\pi^{\text{ref}} \in K(X, \mathcal{P}(A))$. In particular, the τ -ERL problem with *temperature* $\tau > 0$ is

$$\sup_{\mu^\pi \in \mathcal{O}(\nu_0)} \mathcal{J}_\tau(\mu) \quad \text{where} \quad \mathcal{J}_\tau(\mu) := \int r \, d\mu - \tau \mathcal{R}(\mu) \quad \text{and} \quad \mathcal{R}(\mu) := \int \text{KL}(\pi_x^\mu \parallel \pi_x^{\text{ref}}) \, d\nu^\mu(x).$$

When $\tau = 0$, we recover the linear programming formulation of the (expected-value) RL objective. In ERL, the regularizer $\mathcal{R}$ is strictly convex. Thus, $\mathcal{J}_\tau$ is strictly concave and its maximizer unique.²

Lemma 2.1. *The functional $\mathcal{R} : \mathcal{P}(X \times A) \rightarrow \mathbb{R}$ is strictly convex.*

[Proof]

Given Lemma 2.1, one might hope that the well-posedness of τ -ERL could be realized through simple, yet power methods like the direct method in the calculus of variations. However, outside the tabular case, this is unclear, for many reasons, the first of which is that $M_b(X \times A)$ is not separable.

The well-posedness of τ -ERL, however, can be established through other means. In particular, in τ -ERL, only one optimal policy exists, and it is characterized as a Boltzmann–Gibbs (BG) policy.

Definition 2.2. Let $q \in M(X \times A)$ and $\tau > 0$. We denote the *Boltzmann–Gibbs policy associated to q and τ* by $\mathcal{G}_\tau q$, and it is characterized by

$$d(\mathcal{G}_\tau q)_x(a) := e^{(q(x,a) - (\mathcal{V}_\tau q)(x))/\tau} d\pi_x^{\text{ref}}(a) \quad \text{with} \quad (\mathcal{V}_\tau q)(x) := \tau \log \int e^{q(x,a)/\tau} d\pi_x^{\text{ref}}(a).$$

We note that $(\mathcal{G}_\tau q)_x$ is well-defined if and only if $(\mathcal{V}_\tau q)(x) \in \mathbb{R}$.

¹ We expect many of our results can be extended to Polish spaces.

² The only work we are aware of that establishes a comparable result is [28]. However, their result is on tabular MDPs and establishes convexity on $\mathcal{O}(\nu_0)$, not on all of $\mathcal{P}(X \times A)$.

More specifically, it is well-known that the optimal policy of τ -ERL is the BG policy associated to the unique fixed point q_τ^* of the *soft Bellman optimality operator* $\mathcal{B}_\tau^* : M(\mathcal{X} \times \mathcal{A}) \rightarrow M(\mathcal{X} \times \mathcal{A})$,

$$(\mathcal{B}_\tau^* q)(x, a) := r(x, a) + \gamma \int (\mathcal{V}_\tau q)(x') dP_{x,a}(x').$$

(See Lemma A.7.) The following theorem summarizes the well-posedness of τ -ERL.

Theorem 2.3. *Let $\tau > 0$. The policy $\pi^{\tau,*} := \mathcal{G}_\tau q_\tau^*$ is optimal, and uniquely so. More precisely, for all $\nu_0, \nu'_0 \in \mathcal{P}(\mathcal{X})$, we have that $\arg \max_{\mathcal{O}(\nu_0)} \mathcal{J}_\tau = \pi^{\tau,*} = \arg \max_{\mathcal{O}(\nu'_0)} \mathcal{J}_\tau$.* [Proof]

In Appendix A, we prove Theorem 2.3 as well as a collection of supporting and related results that generalize well-known results in tabular MDPs. We include them for completeness.

In the remainder of this work, we study the evolution of τ -optimal objects as τ vanishes. In the tabular regime, where $M_b(\mathcal{X} \times \mathcal{A})$ is separable, one can establish the existence and uniqueness of a τ -optimal occupancy measure: μ_τ^* . Furthermore, under a compatibility assumption, one can prove that the limit of the sequence $(\mu_\tau^*)_{\tau>0}$ as τ vanishes exists and is unique as well.

Assumption 2.4. The intersection of $\{\arg \sup_{\mathcal{O}(\nu_0)} \mathcal{J}_0\}$ and $\{\mathcal{R} < \infty\}$ is nonempty.

Assumption 2.4 asks that our regularizer isn't identically $+\infty$ on the set of optimal policies. Without such an assumption, τ -ERL and RL have no meaningful relationship, as we shall see in Section 3.

Theorem 2.5. *Suppose that $r \in M_b(\mathcal{X} \times \mathcal{A})$ and that $\mathcal{X} \times \mathcal{A}$ is finite. For every $\tau > 0$, let μ_τ^* be the maximizer of $\mathcal{J}_\tau$ over $\mathcal{O}(\nu_0)$. If Assumption 2.4 holds, the sequence $(\mu_\tau^*)_{\tau>0}$ has a unique setwise limit as τ tends to zero. This limit μ_0^* is the minimizer of $\mathcal{R}$ over $\arg \sup_{\mathcal{O}(\nu_0)} \mathcal{J}_0$.* [Proof]

Consequently, in the tabular setting, the sequence $(\pi^{\tau,*})_{\tau>0}$ has a unique limit.

Remark 2.6. Even if Theorem 2.5 could be extended to hold true in continuous MDPs, occupancy measure convergence *does not* guarantee policy convergence, outside of the tabular setting. Theorem 2.5 is a statement about a sequence of *joint distributions*. A policy convergence statement would be one about a sequence of conditional distributions (i.e., probability kernels). In general, the convergence of a sequence of joint distributions does not imply the convergence of the associated sequence of conditional distributions with respect to a fixed marginal (see, e.g., [7, Example 10.4.24]). While it is possible that the structure $\mathcal{O}(\nu_0)$ permits a type of policy convergence, we are unaware of any such result for continuous MDPs.

3 Convergence to Optimality: The Temperature Decoupling Gambit

While ERL has a unique solution, this identifiability comes at a cost with respect to RL: the resulting policy is suboptimal for RL. In this section, we analyze vanishing-temperature limits in τ -ERL. Our main results for this section—Theorems 3.9 and 3.10—show that policies and their return distributions converge under the scheme of Definition 3.7 to interpretable, optimal limits as $\tau \rightarrow 0$.

To understand the ways in which τ -ERL converges to RL, we define a (new) π^{ref} -sensitive variant of the Bellman optimality operator, the *Bellman reference-optimality operator*. We call its unique fixed point the *reference-optimal action-value function*.

Lemma 3.1. *Let $r \in M_b(\mathcal{X} \times \mathcal{A})$, $\gamma < 1$, and $\mathcal{B}_{\text{ref}}^* : M(\mathcal{X} \times \mathcal{A}) \rightarrow M(\mathcal{X} \times \mathcal{A})$ be defined by*

$$(\mathcal{B}_{\text{ref}}^* q)(x, a) := r(x, a) + \gamma \int \text{ess sup}_{\pi^{\text{ref}}} q(x', \cdot) dP_{x,a}(x').$$

Then $\mathcal{B}_{\text{ref}}^$ is a contraction on $M_b(\mathcal{X} \times \mathcal{A})$. Thus, it has a unique fixed point q_{ref}^* .* [Proof]

Generally, q_{ref}^* is distinct from q^* . Yet, ERL recovers q_{ref}^* in the vanishing temperature limit.

Theorem 3.2. *We have that $q_\tau^* \rightarrow q_{\text{ref}}^*$ monotonically as $\tau \rightarrow 0$.* [Proof]

Theorem 3.2 implies that optimal policies, in general, cannot be recovered by taking vanishing temperature limits in ERL. We formalize a notion of *reference-optimality* to highlight this distinction.

Definition 3.3. A policy $\pi \in \mathcal{K}(\mathcal{X}, \mathcal{P}(\mathcal{A}))$ is said to be *reference-optimal* (against π^{ref}) if $q^\pi \geq q_{\text{ref}}^*$. Moreover, π is said to be ϵ -reference optimal if $q^\pi \geq q_{\text{ref}}^* - \epsilon$.

Generally, $q_{\text{ref}}^* < q^*$. For instance, consider an MDP with one state $\perp$ (a bandit), $A = [0, 1]$, and $\pi_{\perp}^{\text{ref}} = \mathcal{U}(A)$. If $r(\perp, \cdot) = \delta_{1/2}$, then $\sup_A q^*(\perp, \cdot) = 1$, while $\text{ess sup}_{\pi_{\perp}^{\text{ref}}} q^*(\perp, \cdot) = 0$. However, in many interesting cases, reference-optimal policies *are* optimal in the classic sense. When A is discrete and π_x^{ref} is supported on all of A —a ubiquitous assumption in ERL—then indeed $q^* = q_{\text{ref}}^*$. Likewise, when A is continuous and (P, r) satisfy certain regularity conditions, then q^* is continuous [20]. In these case, a reference-optimal policy is optimal.

When $q_{\text{ref}}^* \neq q^*$, even state-of-the-art continuous-control methods, entropy-regularized or otherwise, can at best hope to achieve q_{ref}^* , and not q^* . This is because, when $q_{\text{ref}}^* \neq q^*$, optimal actions form a measure 0 set. And so, even rich policy classes, such as neural-network-parameterized Gaussian policies [19] or diffusion policies [9] will not sample these actions, with probability 1. Thus, moving forward, we establish q_{ref}^* as a “skyline” for optimal performance. In other words, we strive to achieve convergence to reference-optimal policies.

Under the next assumption, we can derive convergent policy optimization schemes as τ tends to zero.

Assumption 3.4. A constant $p_{\text{ref}} > 0$ exists for which

$$\inf_{\tau > 0} \inf_{x \in X} \pi_x^{\text{ref}} \left(\left\{ a \in A : q_{\tau}^*(x, a) = \text{ess sup}_{\pi_x^{\text{ref}}} q_{\tau}^*(x, \cdot) \right\} \right) \geq p_{\text{ref}}.$$

Remark 3.5. If A is discrete and π_x^{ref} is uniformly lower bounded, Assumption 3.4 holds. This is a standard assumption. When A is continuous, this assumption is more difficult to guarantee. Intuitively, it asks that there is enough mass surrounding the optima of the entropy-regularized optimal value functions q_{τ}^* for $\text{KL}((\mathcal{G}_{\tau} q_{\tau}^*)_x \parallel \pi_x^{\text{ref}})$ to remain bounded in the limit.

A result key to the remainder of our work is the following bound on the total variation distance between pairs of BG policies in terms of their temperature and the distance between their potentials.

Theorem 3.6. Let $q, q' \in M(X \times A)$. For any $\tau > 0$ and any $x \in X$,

$$\begin{aligned} & \|(\mathcal{G}_{\tau} q)_x - (\mathcal{G}_{\tau} q')_x\|_{\text{TV}} \\ & \leq \min \left\{ \sqrt{\tau^{-1} \|q(x, \cdot) - q'(x, \cdot)\|_{L^{\infty}(\pi_x^{\text{ref}})}}, \frac{1}{2} \sinh(4\tau^{-1} \|q(x, \cdot) - q'(x, \cdot)\|_{L^{\infty}(\pi_x^{\text{ref}})}) \right\}. \end{aligned}$$

In particular,

$$\|(\mathcal{G}_{\tau} q)_x - (\mathcal{G}_{\tau} q')_x\|_{\text{TV}} \leq \frac{2e - 3}{4} \tau^{-1} \|q(x, \cdot) - q'(x, \cdot)\|_{L^{\infty}(\pi_x^{\text{ref}})},$$

if $\|q(x, \cdot) - q'(x, \cdot)\|_{L^{\infty}(\pi_x^{\text{ref}})} < \tau/2$.

[Proof]

While q_{τ}^* and $\mathcal{V}_{\tau} q_{\tau}^*$ converge in the zero-temperature limit, whether or not τ -regularized optimal policies $\pi^{\tau, \star}$ converge is still unclear. Indeed, under Assumption 3.4, $\|q_{\text{ref}}^* - q_{\tau}^*\|_{\infty} \lesssim \tau$ (see Lemma B.10). However, the log-probabilities of an action a under $\pi^{\tau, \star}$ are amplified by τ^{-1} . Hence, the total variation difference between the BG policy at temperature τ and potential q_{ref}^* and $\pi^{\tau, \star}$ may not vanish as τ vanishes. Based on this insight, we introduce the *temperature decoupling gambit*.

Definition 3.7. Given $\tau > 0$, the *temperature decoupling gambit* specifies an alternate temperature $\sigma = \sigma(\tau)$ and constructs $\pi^{\tau, \sigma} := \mathcal{G}_{\tau} q_{\sigma}^*$. In particular, it requires that $\sigma/\tau \rightarrow 0$ as $\tau \rightarrow 0$.

At any $\tau > 0$, decoupled-temperature policies $\pi^{\tau, \sigma}$ are necessarily *not* optimal for the τ -regularized problem. Nevertheless, unlike $\pi^{\tau, \star}$, the policies $\pi^{\tau, \sigma}$ produced by the temperature decoupling gambit realize long-term advantages: they have convergence guarantees in the vanishing temperature limit, and they recover an interpretable reference-optimal policy.

Definition 3.8. Let q^* denote the optimal action-value function in a given MDP, and let $\pi^{\text{ref}} \in \mathcal{K}(X, \mathcal{P}(A))$. The *optimality-filtered* reference policy $\pi^{\text{ref}, \star}$ is defined by

$$\pi_x^{\text{ref}, \star} \propto \pi_x^{\text{ref}} \odot \chi_{N_{\text{ref}}^*(x)} \quad \text{where} \quad N_{\text{ref}}^*(x) := \{a \in A : q^*(x, a) = \text{ess sup}_{\pi_x^{\text{ref}}} q^*(x, \cdot)\}.$$

Here χ_Y is the characteristic or indicator function for the measurable set Y .

Heuristically, the optimality-filtered reference $\pi_x^{\text{ref}, \star}$ is the *restriction* of π_x^{ref} onto the set of expected-value-optimal actions in the state x .³ When π^{ref} is the uniform random policy, that is, $\pi_x^{\text{ref}} = \mathcal{U}(A)$

³ This is exact when $q^* = q_{\text{ref}}^*$.

for all $x \in \mathcal{X}$, we see that $\pi_x^{\text{ref},*} = \mathcal{U}(\mathcal{N}_{\text{ref}}^*(x))$ —the *uniform policy on optimal actions*. In a sense, $\pi^{\text{ref},*}$ is the *most diverse* (reference-)optimal policy; it does not discriminate between optimal actions.

In general, even when $\pi^{\tau,*}$ does converge as τ converges to zero, its limit is different from $\pi^{\text{ref},*}$. We demonstrate this explicitly in Section 3.1. On the other hand, our next result proves that the temperature decoupling gambit enables convergence to $\pi^{\text{ref},*}$.⁴

Theorem 3.9. *Under Assumption 3.4, if $\sigma = \sigma(\tau)$ is such that $\lim_{\tau \rightarrow 0} \sigma/\tau = 0$, then $\pi_x^{\tau,\sigma} \rightarrow \pi_x^{\text{ref},*}$ as $\tau \rightarrow 0$, for all $x \in \mathcal{X}$, in TV if A is discrete and weakly if A is continuous.* [Proof]

At the heart of the proof of Theorem 3.9 is the following inequality (a direct consequence of Theorem 3.6 and Lemma B.10), which relates the BG policies at temperature τ and potentials q_σ^* and q_{ref}^* :

$$\lim_{\tau \rightarrow 0} \sup_x \|(\mathcal{G}_\tau q_\sigma^*)_x - (\mathcal{G}_\tau q_{\text{ref}}^*)_x\|_{\text{TV}} \lesssim -\lim_{\tau \rightarrow 0} \frac{\sigma}{\tau} \log p_{\text{ref}}.$$

This inequality reduces questions of convergence of $\mathcal{G}_\tau q_\sigma^*$ to those of $\mathcal{G}_\tau q_{\text{ref}}^*$ (the vanishing temperature limit of a BG policy with a fixed potential is well-studied). Note that the smaller the fraction σ/τ is, the closer these two policies are. For instance, taking $\sigma(\tau) = \tau^3$ ensures that $\mathcal{G}_\tau q_\sigma^*$ is more like $\mathcal{G}_\tau q_{\text{ref}}^*$ than taking $\sigma(\tau) = \tau^2$. In particular, it is from this inequality that the temperature decoupling gambit’s requirement that $\sigma/\tau \rightarrow 0$ as $\tau \rightarrow 0$ arises.

Beyond enabling policy convergence in the vanishing temperature limit, the temperature decoupling gambit also ensures return distribution function convergence.

Theorem 3.10. *Suppose A is discrete and Assumption 3.4 holds. If $\sigma = \sigma(\tau)$ is such that $\sigma/\tau \rightarrow 0$ as $\tau \rightarrow 0$, then, for any $p, p' \in [1, \infty)$ and $\omega \in \mathcal{P}(\mathcal{X} \times A)$, as $\tau \rightarrow 0$, the return distribution functions $\zeta^{\tau,\sigma}$ of the temperature-decoupled policies $\pi^{\tau,\sigma}$ satisfy $d_{p;p',\omega}(\zeta^{\tau,\sigma}, \zeta^{\pi^{\text{ref},*}}) \rightarrow 0$.* [Proof]

While Theorem 3.10 does not yet provide an algorithm for approximating ζ^* , this result serves as inspiration for such developments in Section 4.

3.1 Numerical Demonstration

In this section, we demonstrate that the policies learned via the temperature decoupling gambit differ from those learned in ERL, even in the presence of stochastic updates.

Figure 3.1 shows a given tristate MDP with two actions (blue: a_1 ; green: a_2), as well as learned policies $\hat{\pi}^{\tau,*}$ and $\hat{\pi}^{\tau,\sigma}$ estimated with soft Q-learning [18]. Here $\pi_x^{\text{ref}} = \mathcal{U}(A)$ for all $x \in \mathcal{X}$ and $\gamma = 0.9$. As this MDP is tabular, Theorem 2.5 implies that the policies $\pi^{\tau,*}$ converge as $\tau \rightarrow 0$. Thus, the temperature decoupling gambit is not necessary to guarantee convergence. Yet we see different limiting behavior. As predicted by Theorem 3.9, the estimates $\hat{\pi}^{\tau,\sigma}$ converge to $\pi^{\text{ref},*}$, as $\tau \rightarrow 0$. With uniform π^{ref} , this is the policy that samples all optimal actions, given a state, with equal probability. As $\tau \rightarrow 0$, the estimates $\hat{\pi}_{x_0}^{\tau,*}$ do converge to a different optimal policy. This difference is in x_0 , where $\hat{\pi}_{x_0}^{\tau,*}$ collapse to δ_{a_1} . We take $\sigma = \tau^2$, in line with Definition 3.7. The two optimal policies found emphasize different notions of diversity. The limit of $\pi^{\tau,*}$ filters out optimal actions in order to play actions more uniformly on average with respect to state occupancy in the long term, while the limit of $\pi^{\tau,\sigma}$ looks to maximize state-wise action diversity.

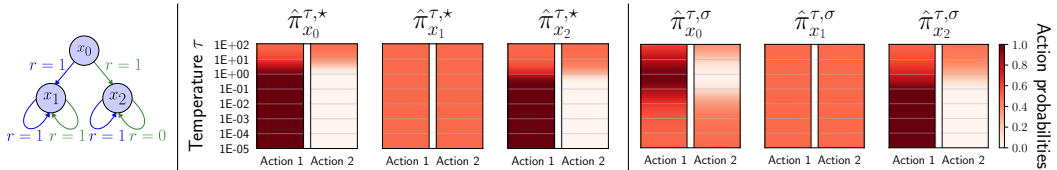


Figure 3.1: Differences between $\hat{\pi}^{\tau,*}$ and $\hat{\pi}^{\tau,\sigma}$, approximated with soft Q-learning. **Left:** Graphical model of the MDP; arrow colors encode actions. **Center:** Depiction of the estimated policies $\hat{\pi}^{\tau,*}$ at each state, as $\tau \rightarrow 0$. **Right:** Depiction of the estimated policies $\hat{\pi}^{\tau,\sigma}$ at each state, as $\tau \rightarrow 0$. **Summary:** Learned policies differ in x_0 , but are otherwise the same.

⁴ We discuss the benefits of this optimal policy in Appendix D.

4 Convergent Approximation of Optimal Return Distributions

In this section, we formalize a new branch of DRL and introduce distributional ERL (DERL).⁵ Our main results in this section, Theorems 4.5, 4.6, and 4.7, establish convergent iterative schemes for approximate (reference-)optimal return distribution estimation. In Section 4.1, we introduce novel soft distributional Bellman operators, for evaluation and for control, and establish the convergence of their iterates. The behavior of the resulting return distribution approximations in the vanishing temperature limit is treated in Section 4.2. To conclude, a simulation is presented in Section 4.3 to illustrate the resulting optimal return distribution approximations.

4.1 Entropy-Regularized Distributional Reinforcement Learning

We begin by defining a *soft distributional Bellman operator*, as an analogue to the distributional Bellman operator [5, 35]. It, under certain conditions, computes

$$\bar{\zeta}_{x,a}^{\tau,\pi} := \text{law} \left(r(X_0^\pi, A_0^\pi) + \sum_{t \geq 1} \gamma^t \left(r(X_t^\pi) - \tau \text{KL}(\pi_{X_t^\pi} \parallel \pi_{X_t^\pi}^{\text{ref}}) \right) \mid X_0^\pi = x, A_0^\pi = a \right).$$

Notationally, for any $\pi \in \mathcal{K}(\mathcal{X}, \mathcal{P}(\mathcal{A}))$, we define $\text{k1}[\pi] : \mathcal{X} \rightarrow \mathbb{R}$ via $\text{k1}[\pi](x) = \text{KL}(\pi_x \parallel \pi_x^{\text{ref}})$.

Definition 4.1. For any $\tau > 0$, $\gamma < 1$, and $\pi \in \mathcal{K}(\mathcal{X}, \mathcal{P}(\mathcal{A}))$, the *soft distributional Bellman operator* $\mathcal{T}_\tau^\pi$ is given by

$$(\mathcal{T}_\tau^\pi \bar{\zeta})_{x,a} := (\mathbf{b}_{r(x,a),\gamma} \circ \text{proj}^\mathbb{R} - \gamma \tau \text{k1}[\pi] \circ \text{proj}^\mathcal{X})_\# (\bar{\zeta}_{-, -} \otimes \check{P}_{x,a}^\pi).$$

Theorem 4.2. If $r \in M_b(\mathcal{X} \times \mathcal{A})$, $\gamma < 1$, and $\pi \in \mathcal{K}(\mathcal{X}, \mathcal{P}(\mathcal{A}))$ is such that

$$\sup_{x,a} \|\tau \text{k1}[\pi]\|_{L^p(P_{x,a})} < \infty, \quad (4.1)$$

the soft distributional Bellman operator $\mathcal{T}_\tau^\pi$ is a γ -contraction in $\bar{d}_p$ for every $\tau \geq 0$. Thus, it has a unique solution to the fixed point equation $\bar{\zeta} = \mathcal{T}_\tau^\pi \bar{\zeta}$, which we denote by $\bar{\zeta}^{\pi,\tau}$. [Proof]

Next, we move to *policy improvement*. In ERL, improving the action-value function q involves policy evaluation with the policy $\mathcal{G}_\tau q$. We leverage this insight to enable control.

Definition 4.3. For any $\tau > 0$, the *soft distributional optimality operator* $\mathcal{T}_\tau^*$ is given by

$$(\mathcal{T}_\tau^* \bar{\zeta})_{x,a} := (\mathcal{T}_\tau^{\mathcal{G}_\tau \bar{\zeta}} \bar{\zeta})_{x,a} \equiv (\mathbf{b}_{r(x,a),\gamma} \circ \text{proj}^\mathbb{R} - \gamma \tau \text{k1}[\mathcal{G}_\tau \bar{\zeta}] \circ \text{proj}^\mathcal{X})_\# (\bar{\zeta}_{-, -} \otimes \check{P}_{x,a}^{\mathcal{G}_\tau \bar{\zeta}})$$

where $\mathcal{Q} : \mathcal{K}(\mathcal{X} \times \mathcal{A}, \mathcal{P}(\mathbb{R})) \rightarrow M(\mathcal{X} \times \mathcal{A})$ is such that $(\mathcal{Q}\zeta)(x, a) := \mathbb{E}_{Z \sim \zeta_{x,a}}[Z]$.

We proceed by establishing a simple, but useful algebraic property.

Lemma 4.4. For any $\tau > 0$, $\mathcal{Q}\mathcal{T}_\tau^* = \mathcal{B}_\tau^* \mathcal{Q}$. [Proof]

Now we prove that iterates of $\mathcal{T}_\tau^*$ converge, unlike iterates of $\mathcal{T}^*$ [5].

Theorem 4.5. For any $\bar{\zeta} \in \bar{\mathcal{K}}^p(\mathcal{X} \times \mathcal{A}, \mathcal{P}(\mathbb{R}))$ and temperature $\tau > 0$ define the iterates $(\bar{\zeta}^n)_{n \in \mathbb{N}}$ given by $\bar{\zeta}^{n+1} = \mathcal{T}_\tau^* \bar{\zeta}^n$ for $\bar{\zeta}^0 = \mathcal{T}_\tau^* \bar{\zeta}$. Then, for $\bar{\zeta}^{\tau,*} := \bar{\zeta}^{\tau, \pi^{\tau,*}}$,

$$\bar{d}_p(\bar{\zeta}^n, \bar{\zeta}^{\tau,*}) \leq C_{p,\tau,\gamma} n \gamma^{n/p} \bar{d}_p(\bar{\zeta}^0, \bar{\zeta}^{\tau,*}) \quad \text{and} \quad \bar{d}_1(\bar{\zeta}^n, \bar{\zeta}^{\tau,*}) \leq \frac{1}{(1-\gamma)\sqrt{\tau}} C n \gamma^n \bar{d}_1(\bar{\zeta}^0, \bar{\zeta}^{\tau,*}),$$

where $C, C_{p,\tau,\gamma} < \infty$ are constants depending on $\|r\|_{\text{sup}}$, $(p, \tau, \gamma, \|r\|_{\text{sup}})$ respectively. [Proof]

Theorem 4.5 leads to stability in entropy-regularized optimal return distribution estimation. In Figure 4.1, we demonstrate the stability of $\mathcal{T}_\tau^*$ and the instability of $\mathcal{T}^*$. The iterates defined in Theorem 4.5 converge to *soft return distributions*, which are influenced by stepwise regularization penalties and correspond to policies that are optimal in ERL. To estimate *optimal* return distributions, we must consider vanishing temperature limits.

⁵ Independently and concurrently, similar results were established by [26] in the fixed-temperature regime, but only with discrete action spaces and π^{ref} being the uniform policy.

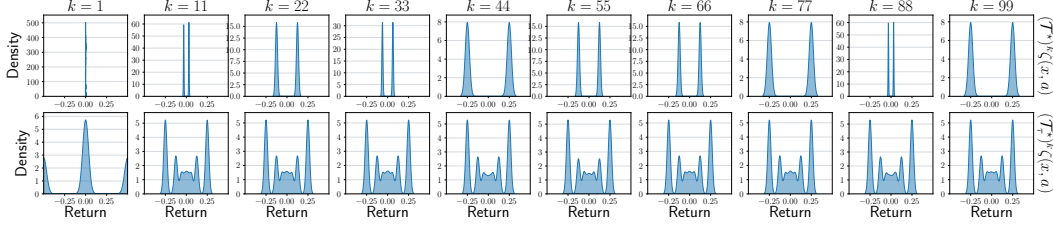


Figure 4.1: Evolution of the *soft* optimality iterates $(\mathcal{T}_\tau^*)^k \zeta(x, a)$ (bottom row) and the iterates of the distributional optimality operator $(\mathcal{T}^*)^k \zeta(x, a)$ (top row). Video of entire iterate sequence is available at <https://harwiltz.github.io/assets/stable-return-distributions/>.

4.2 Convergent Optimal Return Distribution Estimation in the Vanishing Temperature Limit

In this section, we instantiate the first methods for computing iterates that approximate reference-optimal return distribution functions in a stable manner.

Theorem 4.6. *Suppose Assumption 3.4 holds. Let $p, p' \in [1, \infty)$ and $\omega \in \mathcal{P}(\mathbf{X} \times \mathbf{A})$. For any $\epsilon, \delta > 0$, there exists a $\tau > 0$ for which $d_{p;p',\omega}(\bar{\zeta}^{\tau,\pi^{\tau,*}}, \zeta^{\pi^{\tau,*}}) \leq \delta/2$ and $q^{\pi^{\tau,*}}$ is $\epsilon/2$ -reference-optimal. In turn, an $n_{\epsilon,\delta} = n_{\epsilon,\delta}(\tau) \in \mathbb{N}$ exists for which*

$$d_{p;p',\omega}(\bar{\zeta}^n, \zeta^{\pi^{\tau,*}}) \leq \delta \quad \text{and} \quad \mathcal{G}_\tau \bar{\zeta}^n \text{ is } \epsilon\text{-reference-optimal} \quad \forall n \geq n_{\epsilon,\delta}$$

where $\bar{\zeta}^{n+1} = \mathcal{T}_\tau^* \bar{\zeta}^n$ and $\bar{\zeta}^0 = \mathcal{T}_\tau^* \bar{\zeta}$ for any $\bar{\zeta} \in \bar{\mathcal{K}}^p(\mathbf{X} \times \mathbf{A}, \mathcal{P}(\mathbb{R}))$. [Proof]

Theorem 4.6 is the first example of a convergent iterative scheme for approximating the return distribution of a (reference-)optimal policy. While it ensures convergence to a ϵ -reference-optimal return distribution, it is still not possible a priori to characterize which return distribution will be learned. As $\epsilon \rightarrow 0$, there may be no stable trend in the return distribution that will be estimated because $\pi^{\tau,*}$ may not converge. To achieve (characterizable) convergence to a reference-optimal return distribution, we turn back to the temperature decoupling gambit.

Theorem 4.7. *Suppose Assumption 3.4 holds and $\mathbf{A}$ is discrete. Let $p, p' \in [1, \infty)$ and $\omega \in \mathcal{P}(\mathbf{X} \times \mathbf{A})$. For any $\epsilon, \delta > 0$ and $\bar{\zeta}^0 \in \bar{\mathcal{K}}^p(\mathbf{X} \times \mathbf{A}, \mathcal{P}(\mathbb{R}))$, there exists $\tau > 0$, a decoupled $\sigma_\tau > 0$ and $n_{\text{opt}}, n_{\text{eval}} \in \mathbb{N}$ such that*

$$d_{p;p',\omega}(\hat{\zeta}^{n_{\text{eval}}}, \zeta^{\pi^{\text{ref},*}}) \leq \delta \quad \text{and} \quad \mathcal{G}_\tau \bar{\zeta}^{n_{\text{eval}}} \text{ is } \epsilon\text{-reference-optimal}$$

where $\bar{\zeta}^{n+1} = \mathcal{T}_\sigma^* \bar{\zeta}^n$, $\hat{\pi}^{\tau,\sigma} = \mathcal{G}_\tau \bar{\zeta}^{n_{\text{opt}}}$, and $\hat{\zeta}^{n+1} = \mathcal{T}_{\hat{\pi}^{\tau,\sigma}}^{\tau,\sigma} \hat{\zeta}^n$, for $\hat{\zeta}^0 = \bar{\zeta}^{n_{\text{opt}}}$. [Proof]

Theorem 4.7 outlines an algorithm for estimating $\zeta^{\pi^{\text{ref},*}}$. First, approximate $\bar{\zeta}^{\sigma,*}$ via n_{opt} applications of $\mathcal{T}_\sigma^*$ (control). Second, extract the mean: $\hat{q}_\sigma^* \approx q_\sigma^*$. Finally, apply $\mathcal{T}_\tau^{\pi^{\tau,\sigma}}$ n_{eval} times, with $\pi = \mathcal{G}_\tau \hat{q}_\sigma^*$ (evaluation). If $\tau \ll 1$, $\sigma = \tau^2$, for example, and $n_{\text{opt}}, n_{\text{eval}} \gg 1$, then the resulting return distribution is as desired. This ensures convergence, (reference-)optimality, and interpretability of the final iterate.

4.3 Numerical Demonstration

Here we validate that $\bar{\zeta}^{\tau,\sigma}$ approximates $\zeta^{\pi^{\text{ref},*}}$. We consider the MDP given in Figure 4.2. Arrow colors correspond to different actions. Dashed lines represent transitions that occur with probability $1/2$. In this MDP, different optimal policies have distinct return distributions. From x_1 , the blue action yields return of $2\gamma(1-\gamma)^{-1}$, while the green action achieves return $4\gamma(1-\gamma)^{-1}\text{Bernoulli}(1/2)$. In Figures 4.3 and 4.4, we compute estimates $\hat{\zeta}^{\tau,*} \approx \bar{\zeta}^{\tau,*}$ and $\hat{\zeta}^{\tau,\sigma} \approx \bar{\zeta}^{\tau,\sigma}$ by (soft) distributional dynamic programming using 64-bit precision and

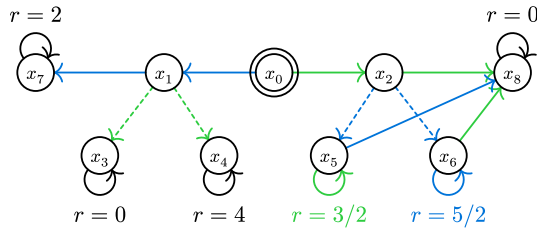


Figure 4.2: An illustrative MDP.

32-bit precision respectively. 32-bit precision is the default in many scientific computing libraries, such as Jax [8]. Here $\gamma = 1/2$, $\pi_x^{\text{ref}} = \mathcal{U}(\mathcal{A})$ for all $x \in \mathcal{X}$, and $\sigma = \tau^2$. We consider $\tau \in \{10^{-(2m+1)} : m = 0, 1, 2, 3, 4\}$. Our simulation is a practical implementation of Theorem 4.7. First, we approximate $n_{\text{opt}} = 1000$ iterative applications of our soft Bellman optimality operator at τ

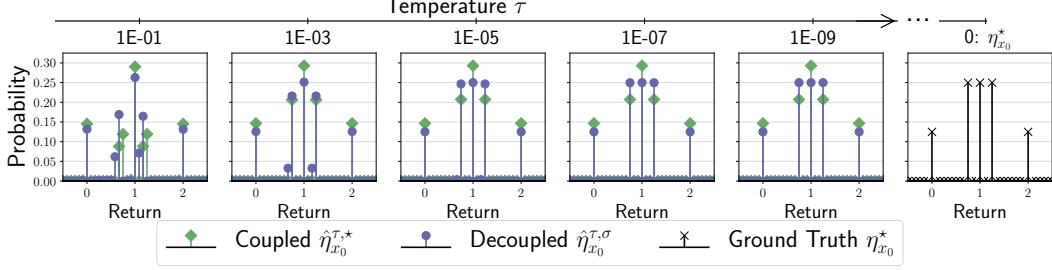


Figure 4.3: Estimates of return distributions via soft distributional dynamic programming— $\hat{\eta}^{\tau,\sigma}$ using the temperature-decoupling gambit and $\hat{\eta}^{\tau,*}$ without—as $\tau \rightarrow 0$. As the temperature vanishes, $\eta^{\tau,\sigma}$ recovers the return distribution of $\pi^{\text{ref},*}$, shown on the right.

(control). Then, we extract $\hat{q}_\tau^*$, an approximation of q_τ^* , and construct two policies: the BG policy at τ and the BG policy at $\tau^{1/2}$, both with potential $\hat{q}_\tau^*$. Next we approximate $n_{\text{eval}} = 1000$ iterative applications of our soft Bellman operator (policy evaluation) at temperature τ with the first policy and at temperature $\tau^{1/2}$ with the second policy. These yield approximations of $\bar{\zeta}^{\tau,*}$ and $\bar{\zeta}^{\tau,\sigma}$, respectively. Figures 4.3 and 4.4 depict the policy-averaged return distributions $\hat{\eta}_{x_0}^{\tau,*}$ and $\hat{\eta}_{x_0}^{\tau,\sigma}$ compared to the

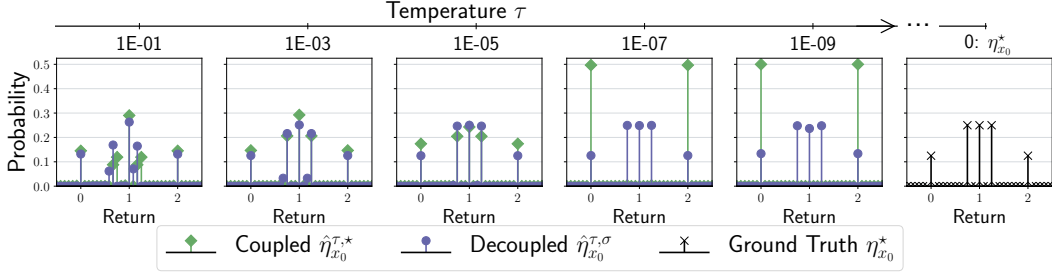


Figure 4.4: Return distribution estimation with vanishing temperature using soft distributional dynamic programming, with 32-bit floating point precision.

baseline $\eta_{x_0}^* := \text{proj}_{\#}^{\mathbb{R}}(\zeta_{x_0,-}^* \otimes \pi_{x_0}^{\text{ref},*})$. The iterates are approximated via categorical representations [5, 34] supported on 121 uniformly-spaced atoms on $[-2, 8]$, and MMD projections [43] with the energy distance kernel $\mathcal{E}_{3/2}$. In both figures, we see that the sequence of temperature-decoupled return distribution estimates approximate the return distribution associated to $\pi^{\text{ref},*}$ (right). Return distributions estimates of $\bar{\zeta}^{\tau,*}$ also converge to those of optimal policies, as predicted by Theorem 4.6, but we find reach *different* return distributions in each case. While the temperature-decoupling gambit is not impervious to precision issues, it stabilizes BG policy estimation.

5 Related Work

Entropy regularization in RL was introduced by [48] for *inverse* RL, where it is necessary to disambiguate optimal policies and identify the most likely reward function to explain demonstrated behavior. ERL with π^{ref} as the uniform policy—termed *maximum entropy* or *MaxEnt* RL, has been highly influential in deep reinforcement learning. Heuristically, MaxEnt RL encourages policies to be more uniform, thereby enhancing exploration, sample-efficiency, behavioral diversity [29, 18, 17], as well as robustness [16, 2, 11, 12]. Heuristic approaches to adaptive temperature schemes in deep

MaxEnt RL have been effective in practice [19, 47]. Policy optimization in MaxEnt RL has been shown to be equivalent to a form of inference, conditional on a notion of behavioral optimality, in a certain graphical model [24, 14], and further characterizations of MaxEnt RL have lead to principled algorithms for efficient exploration [30, 39]. Alternative forms of regularized RL objectives and optimizers have been proposed and analyzed [25, 31, 36, 4, 37, 15].

Policy optimization algorithms for entropy-regularization in general are presented and analyzed by [28]—these methods apply to tabular MDPs and fixed nonzero temperature. [27] provide improved convergence rates for entropy-regularized policy optimization. They also derive convergence results in the vanishing temperature limit, but only in the bandit setting. Exceptionally, [23], based on the work of [1], studies global convergence of policy gradient methods in continuous entropy-regularized MDPs, for fixed and vanishing temperature, with neural network policies via mean-field analysis. However, their analysis requires an extra regularization term to a distribution over neurons, precluding convergence to an optimum of RL. To the best of our knowledge, our work is the first to introduce a convergent policy optimization scheme for general MDPs in the vanishing temperature limit.

Entropy regularization in DRL is largely unexplored. [22] experimented with an adaptation of Rainbow [21] to MaxEntRL, but without analysis or formalism. The concurrent work of [26] also introduced soft distributional Bellman operators, but did not study vanishing temperature limits, and did not establish convergence rates for iterates of $\mathcal{T}_\tau^*$ even for fixed τ . Moreover, the work of [26] established convergence only in the case of discrete A , and only for a uniform reference policy. Works have investigated the challenges of estimating optimal return distributions [5, 42], and more generally, the influence of particular tractable distribution representations on learning dynamics and fixed point accuracy [45, 46, 43, 3]. In [6], the authors show that distributional analogues of $\mathcal{B}^*$ produce iterates that converge when there is a unique (deterministic) optimal policy. The interplay between policy optimization stability and return distributions was studied in [33]. Their empirical study found that distributions of returns following stochastic policy gradient updates tend to have long left tails, and called for methods to guide policies into smoother regions (“quiet” neighborhoods) of the *return landscape*, the manifold of policy returns across parameters. This study focused primarily on deterministic policy gradient methods.

6 Discussion

In this work, we have investigated policy and return distribution convergence as the temperature vanishes in ERL. Our findings motivate iterative schemes for achieving convergence results beyond expected returns. However, they come with several limitations. In particular, while we have established policy convergence via the temperature-decoupling gambit, this convergence qualitative. As a consequence, our ability to derive approximation algorithms for ζ^π with $\pi = \pi^{\text{ref},*}$ is limited; it is a priori unclear which temperatures are required for $\zeta^{\tau,\sigma}$ to be an ϵ -approximation of ζ^π with $\pi = \pi^{\text{ref},*}$ in $d_{p;p,\omega}$ and, therefore, to deploy for iterative applications of $\mathcal{T}_\tau^*$ or $\mathcal{T}_\tau^\pi$ with $\pi = \mathcal{G}_\tau q_\sigma^*$. At the moment, however, our results ensure that by progressively annealing τ , the scheme discussed in Theorem 4.7 will approach ζ^π with $\pi = \pi^{\text{ref},*}$. Nevertheless, quantifying Theorem 3.10 is an exciting direction for future work. Another exciting direction for future work is to try to incorporate the temperature-decoupling gambit into the many algorithms in ERL/RL.

Acknowledgments and Disclosure of Funding

The authors wish to thank Wesley Chung, Mark Rowland, Jesse Farebrother, Arnav Kumar Jain, Siddarth Venkatraman, Athanasios Vasileiadis, Aditya Mahajan, and Doina Precup for helpful comments and discussions. HW was supported by the National Sciences and Engineering Research Council of Canada (NSERC) and the Fonds de Recherche du Québec. MGB was supported by the Canada CIFAR AI Chair program and NSERC. This work was supported in part by DARPA HR0011-23-9-0050.

References

- [1] A. Agazzi and J. Lu. Global optimality of softmax policy gradient with single hidden layer neural networks in the mean-field regime. *arXiv preprint arXiv:2010.11858*, 2020.

- [2] Z. Ahmed, N. Le Roux, M. Norouzi, and D. Schuurmans. Understanding the impact of entropy on policy optimization. In *International Conference on Machine Learning (ICML)*, 2019.
- [3] J. Alhosh, H. Wiltzer, and D. Meger. Tractable representations for convergent approximation of distributional HJB equations. *Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*, 2025.
- [4] K. Asadi and M. L. Littman. An alternative softmax operator for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2017.
- [5] M. G. Bellemare, W. Dabney, and R. Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2017.
- [6] M. G. Bellemare, W. Dabney, and M. Rowland. *Distributional reinforcement learning*. MIT Press, 2023.
- [7] V. I. Bogachev and M. A. S. Ruas. *Measure theory*, volume 1. Springer, 2007.
- [8] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: composable transformations of Python+NumPy programs, 2018.
- [9] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 2023.
- [10] R. Dadashi, A. A. Taiga, N. Le Roux, D. Schuurmans, and M. G. Bellemare. The value function polytope in reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2019.
- [11] B. Eysenbach and S. Levine. If MaxEnt RL is the answer, what is the question? *arXiv preprint arXiv:1910.01913*, 2019.
- [12] B. Eysenbach and S. Levine. Maximum entropy RL (provably) solves some robust RL problems. In *International Conference on Learning Representations (ICLR)*, 2021.
- [13] E. A. Feinberg and A. Shwartz. *Handbook of Markov decision processes: methods and applications*, volume 40. Springer Science & Business Media, 2012.
- [14] M. Fellows, A. Mahajan, T. G. J. Rudner, and S. Whiteson. VIREL: A variational inference framework for reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [15] R. Fox. Toward provably unbiased temporal-difference value estimation. In *Optimization Foundations for Reinforcement Learning workshop (OPTRL @ NeurIPS)*, 2019.
- [16] R. Fox, A. Pakman, and N. Tishby. Taming the noise in reinforcement learning via soft updates. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2015.
- [17] D. Garg, J. Hejna, M. Geist, and S. Ermon. Extreme Q-learning: MaxEnt RL without entropy. In *International Conference on Learning Representations (ICLR)*, 2023.
- [18] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning (ICML)*, 2017.
- [19] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, 2018.
- [20] O. Hernández-Lerma and J. B. Lasserre. *Discrete-time Markov control processes: basic optimality criteria*, volume 30. Springer Science & Business Media, 2012.
- [21] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver. Rainbow: Combining improvements in deep reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2018.
- [22] D. Hu, P. Abbeel, and R. Fox. Count-based temperature scheduling for maximum entropy reinforcement learning. In *Deep RL Workshop @ (NeurIPS)*, 2021.
- [23] J.-M. Leahy, B. Kerimkulov, D. Siska, and L. Szpruch. Convergence of policy gradient for entropy regularized mdps with neural network approximation in the mean-field regime. In *International Conference on Machine Learning (ICML)*, 2022.

- [24] S. Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.
- [25] M. L. Littman and C. Szepesvári. A generalized reinforcement-learning model: Convergence and applications. In *International Conference on Machine Learning (ICML)*, 1996.
- [26] X. Ma, J. Chen, L. Xia, J. Yang, Q. Zhao, and Z. Zhou. DSAC: Distributional soft actor-critic for risk-sensitive reinforcement learning. *Journal of Artificial Intelligence Research*, 83, 2025.
- [27] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning (ICML)*, 2020.
- [28] G. Neu, A. Jonsson, and V. Gómez. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.
- [29] B. O’Donoghue, R. Munos, K. Kavukcuoglu, and V. Mnih. Combining policy gradient and Q-learning. In *International Conference on Learning Representations (ICLR)*, 2016.
- [30] B. O’Donoghue, I. Osband, and C. Ionescu. Making sense of reinforcement learning and probabilistic inference. In *International Conference on Learning Representations (ICLR)*, 2020.
- [31] J. Peters, K. Mulling, and Y. Altun. Relative entropy policy search. In *AAAI Conference on Artificial Intelligence*, 2010.
- [32] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- [33] N. Rahn, P. D’Oro, H. Wiltzer, P.-L. Bacon, and M. Bellemare. Policy optimization in a noisy neighborhood: On return landscapes in continuous control. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [34] M. Rowland, M. Bellemare, W. Dabney, R. Munos, and Y. W. Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2018.
- [35] M. Rowland, R. Dadashi, S. Kumar, R. Munos, M. G. Bellemare, and W. Dabney. Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2019.
- [36] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International Conference on Machine Learning (ICML)*, 2015.
- [37] Z. Song, R. E. Parr, and L. Carin. Revisiting the softmax bellman operator: New benefits and new perspective. In *International Conference on Machine Learning (ICML)*, 2019.
- [38] U. Syed, M. Bowling, and R. E. Schapire. Apprenticeship learning using linear programming. In *International Conference on Machine Learning (ICML)*, 2008.
- [39] J. Tarbouriech, T. Lattimore, and B. O’Donoghue. Probabilistic inference in reinforcement learning done right. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [40] C. Villani. *Optimal transport: old and new*, volume 338. Springer, 2008.
- [41] H. S. Wilf. *generatingfunctionology*. CRC press, 2005.
- [42] H. Wiltzer, M. G. Bellemare, D. Meger, P. Shafiro, and Y. Jhaveri. Action gaps and advantages in continuous-time distributional reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.
- [43] H. Wiltzer, J. Farebrother, A. Gretton, and M. Rowland. Foundations of multivariate distributional reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.
- [44] H. Wiltzer, J. Farebrother, A. Gretton, Y. Tang, A. Barreto, W. Dabney, M. G. Bellemare, and M. Rowland. A distributional analogue to the successor representation. In *International Conference on Machine Learning (ICML)*, 2024.
- [45] H. E. Wiltzer, D. Meger, and M. G. Bellemare. Distributional Hamilton-Jacobi-Bellman equations for continuous-time reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2022.
- [46] R. Wu, M. Uehara, and W. Sun. Distributional offline policy evaluation with predictive error guarantees. In *International Conference on Machine Learning (ICML)*, 2023.

- [47] Y. Xu, D. Hu, L. Liang, S. M. McAleer, P. Abbeel, and R. Fox. Target entropy annealing for discrete soft actor-critic. In *Deep RL Workshop @ (NeurIPS)*, 2021.
- [48] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2008.

A Entropy-Regularized RL in Continuous MDPs

Here we prove Theorem 2.3 as well as a collection of supporting and related results that generalize well-known results in tabular MDPs.

We start with a characterization the geometry of the space of occupancy measures. The following result extends the well-known counterpart in tabular MDPs [13, 38, 10] to continuous MDPs. While certain parts of this result are proved by [20], not all connections are made, which we state here for the first time.

Theorem A.1. *Let $\mathcal{O}(\nu_0) = \{\mu^\pi : \pi \in \mathcal{K}(\mathcal{X}, \mathcal{P}(\mathcal{A}))\}$ the space of all occupancy measures under the initial state distribution $\nu_0 \in \mathcal{P}(\mathcal{X})$. Then $\mathcal{O}(\nu_0)$ is equivalent to the space of all $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{A})$ that satisfy*

$$\text{proj}_{\#}^{\mathcal{X}} \mu(\mathcal{E}) = (1 - \gamma)\nu_0(\mathcal{E}) + \gamma \int P_{x,a}(\mathcal{E}) d\mu(x, a) \quad \forall \mathcal{E} \subset \mathcal{X} \text{ Borel.} \quad (\text{A.1})$$

The space $\mathcal{O}(\nu_0)$ is convex, it is closed under setwise convergence.

Before proceeding with the proof of Theorem A.1, we recall the *state occupancy measures* ν^π , given by

$$\nu^\pi := (1 - \gamma) \sum_{t \geq 0} \gamma^t \nu_t^\pi,$$

where $(\nu_t^\pi)_{t \geq 1}$ is the sequence of laws generated by $\hat{P}^\pi$ starting at ν_0 .

Proposition A.2. *Let ν_t^π and μ_t^π , for $t \geq 1$ denote the laws generated by $\hat{P}^\pi$ and $\check{P}^\pi$ starting at ν_0 and $\mu_0^\pi = \pi_- \otimes \nu_0$. Then $\mu_t^\pi = \pi_- \otimes \nu_t^\pi$, for all $t \geq 1$. Hence, given π and ν_0 , the state marginal of the associated occupancy measure μ^π is the associated state occupancy measure ν^π .*

The proof of this proposition will use the following lemma.

Lemma A.3. *Under the hypotheses of Proposition A.2, for every $t \geq 1$, the conditional probabilities of μ_t^π with respect to its state marginal are π_- .*

Proof. It suffices to prove that the conditional probabilities of μ_1^π are π_- . Let ν_1 denote the state marginal of μ_1^π . By definition,

$$\int \psi(x') d\nu_1(x') = \int \psi(x') d\mu_1^\pi(x', a') = \int \left[\int \psi(x') dP_{x,a}(x') \right] d\mu_0^\pi(x, a).$$

Thus, for any $\varphi \in M_b(\mathcal{X} \times \mathcal{A})$, with $\psi(x') := \int \varphi(x', a') d\pi_{x'}(a')$, observe that

$$\begin{aligned} \int \left[\int \varphi(x', a') d\pi_{x'}(a') \right] d\nu_1(x') &= \int \psi(x') d\nu_1(x') \\ &= \int \left[\int \psi(x') dP_{x,a}(x') \right] d\mu_0^\pi(x, a) \\ &= \int \left[\int \left[\int \varphi(x', a') d\pi_{x'}(a') \right] dP_{x,a}(x') \right] d\mu_0^\pi(x, a) \\ &= \int \left[\int \varphi(x', a') d\check{P}_{x,a}^\pi(x') \right] d\mu_0^\pi(x, a) \\ &= \int \varphi(x, a) d\mu_1^\pi(x, a). \end{aligned}$$

So the conditional probabilities of μ_1^π with respect to ν_1 are π_x , as desired. $\square$

Proof of Proposition A.2. By Lemma A.3, it suffices to show that the state marginal of μ_1^π is ν_1^π . This holds:

$$\begin{aligned} \int \psi(x') d\nu_1^\pi(x') &= \int \left[\int \psi(x') d\hat{P}_x^\pi(x') \right] d\nu_0(x) \\ &= \int \left[\int \left[\int \psi(x') dP_{x,a}(x') \right] d\pi_x(a) \right] d\nu_0(x) \\ &= \int \left[\int \psi(x') dP_{x,a}(x') \right] d(\pi_x \otimes \nu_0)(x, a) \\ &= \int \left[\int \psi(x') dP_{x,a}(x') \right] d\mu_0^\pi(x, a). \end{aligned}$$

By this computation and Lemma A.3 applied successively to each pair $(\mu_{t+1}^\pi, \mu_t^\pi)$ for every $t \geq 1$, we deduce that $\mu_t^\pi = \pi_- \otimes \nu_t^\pi$, for all $t \geq 1$. Finally, by the linearity of the integral, we conclude. Indeed,

$$\mu^\pi := (1 - \gamma) \sum_{t \geq 0} \gamma^t \mu_t^\pi = (1 - \gamma) \sum_{t \geq 0} \gamma^t (\pi_- \otimes \nu_t^\pi) = \pi_- \otimes (1 - \gamma) \sum_{t \geq 0} \gamma^t \nu_t^\pi =: \pi_- \otimes \nu^\pi.$$

□

Proof of Theorem A.1. We prove this theorem in three steps.

Step 1: $\mathcal{O}(\nu_0) = \mathcal{F}(\nu_0)$. First, recall that $\text{proj}_\#^\mathbf{X} \mu^\pi = \nu^\pi$ for any policy π , by Proposition A.2. Thus, we have that for any π and any Borel $E \subset \mathbf{X}$,

$$\begin{aligned} \text{proj}_\#^\mathbf{X} \mu^\pi(E) &= \nu^\pi(E) = (1 - \gamma) \nu_0(E) + \gamma(1 - \gamma) \sum_{t \geq 0} \gamma^t \nu_{t+1}^\pi(E) \\ &= (1 - \gamma) \nu_0(E) + \gamma(1 - \gamma) \sum_{t \geq 0} \gamma^t \int P_{x,a}(E) d\mu_t^\pi(x, a) \\ &= (1 - \gamma) \nu_0(E) + \gamma \int P_{x,a}(E) d\mu^\pi(x, a). \end{aligned}$$

This shows that $\mathcal{O}(\nu_0) \subset \mathcal{F}(\nu_0)$. It remains to show that $\mathcal{F}(\nu_0) \subset \mathcal{O}(\nu_0)$. Let $\mu \in \mathcal{F}(\nu_0)$, and let π^μ denote its conditional action probabilities with respect to its state marginal ν^μ —that is, $\mu = \pi^\mu \otimes \nu^\mu$. Moreover, let ϕ_0 be any bounded measurable function. By the definition of P , we note that (A.1) can be written as

$$\int \phi_0(x_0) d\nu^\mu(x_0) = (1 - \gamma) \int \phi_0(x_0) d\nu_0(x_0) + \gamma \int \left[\int \phi_0(x_1) d\hat{P}_{x_0}^{\pi^\mu}(x_1) \right] d\nu^\mu(x_0).$$

Defining $\phi_1(x) = \int_{\mathbf{X}} \phi_0(x') d\hat{P}_x^{\pi^\mu}(x')$, the rightmost term $\int_{\mathbf{X}} \phi_1(x_0) d\nu^\mu(x_0)$ can be again expanded via (A.1),

$$\begin{aligned} \int \phi_0(x_0) d\nu^\mu(x_0) &= (1 - \gamma) \int (\phi_0(x_0) + \gamma \phi_1(x_0)) d\nu_0(x_0) \\ &\quad + \gamma^2 \int \left[\int \phi_1(x_0) d\hat{P}_{x_0}^{\pi^\mu}(x_1) \right] d\nu^\mu(x_0). \end{aligned}$$

Continuing, we define $\phi_{n+1}(x) = \int_{\mathbf{X}} \phi_n(x') d\hat{P}_x^{\pi^\mu}(x')$, which is bounded and measurable for each $n \in \mathbb{N}$, yielding

$$\begin{aligned} \int \phi_0(x_0) d\nu^\mu(x_0) &= (1 - \gamma) \underbrace{\int \sum_{k=0}^n \gamma^k \phi_k(x_0) d\nu_0(x_0)}_{\text{I}_n} \\ &\quad + \underbrace{\gamma^{n+1} \int \left[\int \phi_n(x_0) d\hat{P}_{x_0}^{\pi^\mu}(x_1) \right] d\nu^\mu(x_0)}_{\text{II}_n}. \end{aligned}$$

By the definition of ϕ_n , we have that

$$\begin{aligned}
I_n &= (1 - \gamma) \int \phi_0(x_0) d\nu_0(x_0) + (1 - \gamma)\gamma \int \left[\int \phi_0(x_1) d\hat{P}_{x_0}^{\pi^\mu}(x_1) \right] d\nu_0(x_0) + \dots \\
&= (1 - \gamma) \int \phi_0(x_0) d\nu_0(x_0) + (1 - \gamma)\gamma \int \phi_0(x_0) d\nu_1^{\pi^\mu}(x_0) + \dots \\
&= \int \phi_0(x_0) (1 - \gamma) \sum_{k=0}^n \gamma^k d\nu_k^{\pi^\mu}(x_0).
\end{aligned}$$

Moreover, by the boundedness of ϕ_n , we deduce that $I_n \rightarrow 0$. Substituting, we have

$$\begin{aligned}
\int \phi_0(x_0) d\nu^\mu(x_0) &= \lim_{n \rightarrow \infty} I_n + \lim_{n \rightarrow \infty} II_n \\
&= \int \phi_0(x_0) \lim_{n \rightarrow \infty} (1 - \gamma) \sum_{k=0}^n \gamma^k d\nu_k^{\pi^\mu}(x_0) \\
&= \int \phi_0(x_0) d\nu^{\pi^\mu}(x_0).
\end{aligned}$$

Since ϕ_0 was an arbitrary bounded and measurable function, it follows that $\nu^\mu = \nu^{\pi^\mu}$. Thus, $\mu = \pi_- \otimes \nu^\mu = \mu^{\pi^\mu}$ —the occupancy measure for the policy π^μ . Consequently, any $\mu \in \mathcal{F}(\nu_0)$ is a member of $\mathcal{O}(\nu_0)$.

Step 2: $\mathcal{O}(\nu_0)$ is convex. The convexity of $\mathcal{O}(\nu_0)$ follows immediately from the structure of $\mathcal{F}(\nu_0)$. Consider any $\mu_0, \mu_1 \in \mathcal{O}(\nu_0)$ any $\alpha \in [0, 1]$, and define $\mu_\alpha = \alpha\mu_0 + (1 - \alpha)\mu_1$. For any Borel $E \subset \mathbf{X}$, we have that

$$\text{proj}_{\#}^{\mathbf{X}} \mu_\alpha(E) = \alpha \text{proj}_{\#}^{\mathbf{X}} \mu_0(E) + (1 - \alpha) \text{proj}_{\#}^{\mathbf{X}} \mu_1(E)$$

Since $\mu_0, \mu_1 \in \mathcal{F}(\nu_0)$, they solve (A.1), so we expand the RHS,

$$\begin{aligned}
\text{proj}_{\#}^{\mathbf{X}} \mu_\alpha(E) &= \alpha(1 - \gamma)\nu_0(E) + \alpha\gamma \int P_{x,a}(E) d\mu_0(x, a) \\
&\quad + (1 - \alpha)(1 - \gamma)\nu_0(E) + (1 - \alpha)\gamma \int P_{x,a}(E) d\mu_1(x, a) \\
&= (1 - \gamma)\nu_0(E) + \gamma \int P_{x,a}(E) (\alpha d\mu_0(x, a) + (1 - \alpha) d\mu_1(x, a)) \\
&= (1 - \gamma)\nu_0(E) + \gamma \int P_{x,a}(E) d\mu_\alpha(x, a).
\end{aligned}$$

So $\mu_\alpha \in \mathcal{F}(\nu_0) = \mathcal{O}(\nu_0)$, as desired.

Step 3: $\mathcal{O}(\nu_0)$ is closed under setwise convergence. Let $(\mu_k)_{k \in \mathbb{N}} \subset \mathcal{F}(\nu_0)$ be a sequence that converges setwise to μ . Since $(x, a) \mapsto P_{x,a}(E)$ is bounded and measurable for any Borel $E \subset \mathbf{X}$,

$$\int P_{x,a}(E) d\mu_k(x, a) \rightarrow \int P_{x,a}(E) d\mu(x, a). \quad (\text{A.2})$$

Likewise,

$$\text{proj}_{\#}^{\mathbf{X}} \mu_k(E) = \mu_k(E \times \mathbf{A}) \rightarrow \mu(E \times \mathbf{A}) = \text{proj}_{\#}^{\mathbf{X}} \mu(E), \quad (\text{A.3})$$

as $\mu_k \rightarrow \mu$ setwise. Consequently, we have that

$$\begin{aligned}
\text{proj}_{\#}^{\mathbf{X}} \mu(E) &= \lim_{k \rightarrow \infty} \text{proj}_{\#}^{\mathbf{X}} \mu_k(E) \\
&= \lim_{k \rightarrow \infty} \left[(1 - \gamma)\nu_0(E) + \gamma \int P_{x,a}(E) d\mu_k(x, a) \right] \\
&= (1 - \gamma)\nu_0(E) + \gamma \int P_{x,a}(E) d\mu(x, a),
\end{aligned}$$

where the first equality follows from (A.3), the second follows as $\mu_k \in \mathcal{F}(\nu_0)$, and the final equality follows from (A.2). Thus, we see that $\mu \in \mathcal{F}(\nu_0) = \mathcal{O}(\nu_0)$. $\square$

Now we prove Lemma 2.1.

Lemma 2.1. *The functional $\mathcal{R} : \mathcal{P}(\mathbf{X} \times \mathbf{A}) \rightarrow \mathbb{R}$ is strictly convex.*

[Source]

Proof. Observe that

$$\mathcal{R}(\mu) = \text{KL}(\mu \parallel \bar{\pi}_- \otimes \nu^\mu).$$

We prove this in two steps. First, for every Borel $f : \mathbf{X} \times \mathbf{A} \rightarrow [0, \infty)$, we have that

$$\begin{aligned} \int f(x, a) \frac{d\pi_x^\mu}{d\pi_x^{\text{ref}}}(a) d(\pi_-^{\text{ref}} \otimes \nu^\mu)(x, a) &= \int \left[\int f(x, a) \frac{d\pi_x^\mu}{d\pi_x^{\text{ref}}} d\pi_x^{\text{ref}}(a) \right] d\nu^\mu(x) \\ &= \int \left[\int f(x, a) d\pi_x^\mu(a) \right] d\nu^\mu(x) \\ &= \int f(x, a) d(\pi_-^\mu \otimes \nu^\mu)(x, a). \end{aligned}$$

Hence, $\mu = \pi_-^\mu \otimes \nu^\mu \ll \pi_-^{\text{ref}} \otimes \nu^\mu$ if $\pi_x^\mu \ll \pi_x^{\text{ref}}$ for ν^μ -almost every x , and

$$\frac{d\mu}{d(\pi_-^{\text{ref}} \otimes \nu^\mu)}(x, a) = \frac{d\pi_x^\mu}{d\pi_x^{\text{ref}}}(a).$$

Second, $\mu = \pi_-^\mu \otimes \nu^\mu \ll \pi_-^{\text{ref}} \otimes \nu^\mu$ implies that $\pi_x^\mu \ll \pi_x^{\text{ref}}$ for ν^μ -almost every x . Indeed, suppose that a set $S \subset \mathbf{X}$ exists such that $\nu^\mu(S) > 0$ and for each $x \in S$, we have that

$$\pi_x^\mu(\mathbf{B}_x) > 0 \quad \text{but} \quad \pi_x^{\text{ref}}(\mathbf{B}_x) = 0.$$

Let

$$\mathbf{E} := \bigcup_{x \in S} \{x\} \times \mathbf{B}_x.$$

Then,

$$(\pi_-^{\text{ref}} \otimes \nu^\mu)(\mathbf{E}) = \int_S \pi_x^{\text{ref}}(\mathbf{B}_x) d\nu^\mu(x) = 0 \quad \text{and} \quad (\pi_-^\mu \otimes \nu^\mu)(\mathbf{E}) = \int_S \pi_x^\mu(\mathbf{B}_x) d\nu^\mu(x) > 0.$$

This is a contradiction. And so,

$$\begin{aligned} \mathcal{R}(\mu) &= \int \left[\int \log \left(\frac{d\pi_x^\mu}{d\pi_x^{\text{ref}}}(a) \right) d\pi_x^\mu(a) \right] d\nu^\mu(x) \\ &= \int \left[\int \log \left(\frac{d\mu}{d(\pi_-^{\text{ref}} \otimes \nu^\mu)}(x, a) \right) d\pi_x^\mu(a) \right] d\nu^\mu(x) \\ &= \int \log \left(\frac{d\mu}{d(\pi_-^{\text{ref}} \otimes \nu^\mu)}(x, a) \right) d\mu(x, a) \\ &= \text{KL}(\mu \parallel \bar{\pi}_- \otimes \nu^\mu), \end{aligned}$$

as desired.

Now recall that

$$\text{KL}(t\mu_1 + (1-t)\mu_0 \parallel t\mu'_1 + (1-t)\mu'_0) \leq t\text{KL}(\mu_1 \parallel \mu'_1) + (1-t)\text{KL}(\mu_0 \parallel \mu'_0).$$

Moreover, note that

$$\nu^{t\mu_1 + (1-t)\mu_0} = t\nu^{\mu_1} + (1-t)\nu^{\mu_0}.$$

In turn,

$$\begin{aligned} \mathcal{R}(t\mu_1 + (1-t)\mu_0) &= \text{KL}(t\mu_1 + (1-t)\mu_0 \parallel \pi_-^{\text{ref}} \otimes \nu^{t\mu_1 + (1-t)\mu_0}) \\ &= \text{KL}(t\mu_1 + (1-t)\mu_0 \parallel \pi_-^{\text{ref}} \otimes (t\nu^{\mu_1} + (1-t)\nu^{\mu_0})) \\ &= \text{KL}(t\mu_1 + (1-t)\mu_0 \parallel t(\pi_-^{\text{ref}} \otimes \nu^{\mu_1}) + (1-t)(\pi_-^{\text{ref}} \otimes \nu^{\mu_0})) \\ &\leq t\text{KL}(\mu_1 \parallel \pi_-^{\text{ref}} \otimes \nu^{\mu_1}) + (1-t)\text{KL}(\mu_0 \parallel \pi_-^{\text{ref}} \otimes \nu^{\mu_0}) \\ &= t\mathcal{R}(\mu_1) + (1-t)\mathcal{R}(\mu_0). \end{aligned}$$

Thus, $\mathcal{R}$ is convex. In particular, $\mathcal{R}$ is strictly convex as KL is strictly convex in its first argument. $\square$

With Theorem A.1 and Lemma 2.1 in hand, we use the direct method from the Calculus of Variations to prove the well-posedness of τ -ERL, in the tabular setting.

Remark A.4. The space $M_b(\mathsf{X} \times \mathsf{A})$ endowed with the supnorm is a Banach space. Note that $M_b(\mathsf{X} \times \mathsf{A})^* \cong ba(\mathsf{X} \times \mathsf{A})$, where $ba(\mathsf{X} \times \mathsf{A})$ denotes the set of finitely additive set functions on $\mathcal{B}(\mathsf{X} \times \mathsf{A})$ equipped with the total variation norm. Note that the set of probability measures on $\mathsf{X} \times \mathsf{A}$ is a subset of the closed unit ball in $ba(\mathsf{X} \times \mathsf{A})$, which is weak* compact, by Banach–Alaoglu. The duality pairing for any $\mu \in \mathcal{P}(\mathsf{X} \times \mathsf{A})$ and for any $\varphi \in M_b(\mathsf{X} \times \mathsf{A})$ is given by integration: $\langle \mu, \varphi \rangle := \int \varphi d\mu$. In other words, weak* convergence is setwise convergence when $\mathcal{P}(\mathsf{X} \times \mathsf{A})$ is considered as a subset of the dual of $ba(\mathsf{X} \times \mathsf{A})$.

Theorem A.5. Suppose that $r \in M_b(\mathsf{X} \times \mathsf{A})$, $\mathsf{X} \times \mathsf{A}$ is finite, and let $\nu_0 \in \mathcal{P}(\mathsf{X})$. A $\mu_\tau^* \in \mathcal{O}(\nu_0)$ that achieves the supremum in (2.2) exists. Moreover, no other occupancy measure does so.

Proof. Let the supremum in (2.2) be denoted by ϑ_τ^* and $(\mu_k)_{k \in \mathbb{N}} \subset \mathcal{O}(\nu_0)$ be such that

$$\vartheta_\tau^* - \frac{1}{k} < \mathcal{J}_\tau(\mu_k) \leq \vartheta_\tau^*.$$

In other words, let $(\mu_k)_{k \in \mathbb{N}} \subset \mathcal{O}(\nu_0)$ be a maximizing sequence. By Remark A.4, owing to the fact that $M_b(\mathsf{X} \times \mathsf{A})$ is separable (since $\mathsf{X} \times \mathsf{A}$ is finite), let $(\mu_{k_\ell})_{\ell \in \mathbb{N}}$ be a weakly* convergent subsequence, with weak* limit μ_∞ . In particular, $\mu_{k_\ell} \rightarrow \mu_\infty$ setwise. As $\mathcal{O}(\nu_0)$ is closed under setwise convergence, by Theorem A.1, we have that $\mu_\infty \in \mathcal{O}(\nu_0)$. Furthermore, $\pi_-^{\text{ref}} \otimes \nu^{\mu_{k_\ell}} \rightarrow \pi_-^{\text{ref}} \otimes \nu^{\mu_\infty}$ setwise as well. As setwise convergence implies weak convergence and as the $\text{KL}(\mu \parallel \mu')$ is lower-semicontinuous in the pair (μ, μ') in the weak topology, we find that

$$\begin{aligned} \vartheta^{\tau,*} &\leq \limsup_{\ell \rightarrow \infty} \int r d\mu_{k_\ell} - \tau \liminf_{\ell \rightarrow \infty} \text{KL}(\mu_{k_\ell} \parallel \pi_-^{\text{ref}} \otimes \nu^{\mu_{k_\ell}}) \\ &\leq \limsup_{\ell \rightarrow \infty} \int r d\mu_{k_\ell} - \tau \text{KL}(\mu_\infty \parallel \pi_-^{\text{ref}} \otimes \nu^{\mu_\infty}) \\ &= \int r d\mu_\infty - \tau \text{KL}(\mu_\infty \parallel \pi_-^{\text{ref}} \otimes \nu^{\mu_\infty}) \\ &= \mathcal{J}_\tau(\mu_\infty). \end{aligned}$$

The penultimate equality uses that r is bounded. Thus, $\mathcal{J}_\tau(\mu_\infty) = \vartheta_\tau^*$. The previous argument applies to any sub-sequential weak* limit of our maximizing sequence. But as $\mathcal{R}$ is strictly convex, by Lemma 2.1, and $\mathcal{O}(\nu_0)$ is convex, by Theorem A.1, only one such limit exists. $\square$

We now move to prove Theorem 2.3. To do so, we state and prove some helpful results. We begin with policy evaluation.

For any $\pi \in \mathsf{K}(\mathsf{X}, \mathcal{P}(\mathsf{A}))$, define $q_\tau^\pi : \mathsf{X} \times \mathsf{A} \rightarrow \mathbb{R} \cup \{-\infty\}$ by

$$q_\tau^\pi(x, a) := \mathbf{E} \left[r(X_0^\pi, A_0^\pi) + \sum_{t \geq 1} \gamma^t \left(r(X_t^\pi, A_t^\pi) - \tau \text{KL}(\pi_{X_t^\pi} \parallel \pi_{X_t^\pi}^{\text{ref}}) \right) \middle| (X_0^\pi, A_0^\pi) = (x, a) \right].$$

By the tower property of condition expectation, we have that

$$q_\tau^\pi(x, a) = r(x, a) + \gamma \int q_\tau^\pi(x', a') - \tau \text{KL}(\pi_{x'} \parallel \pi_{x'}^{\text{ref}}) d\check{P}_{x,a}^\pi(x', a').$$

It is convenient to be able to evaluate a policy π (find q_τ^π) in an iterative fashion. This can be done via the *soft Bellman operator* $\mathcal{B}_\tau^\pi : M(\mathsf{X} \times \mathsf{A}) \rightarrow M(\mathsf{X} \times \mathsf{A})$ defined by

$$(\mathcal{B}_\tau^\pi q)(x, a) := r(x, a) + \gamma \int q(x', a') - \tau \text{KL}(\pi_{x'} \parallel \pi_{x'}^{\text{ref}}) d\check{P}_{x,a}^\pi(x', a'),$$

but only on a restricted collection of policies.

Lemma A.6. If $r \in M_b(\mathsf{X} \times \mathsf{A})$, $\gamma < 1$, and π is such that (4.1) holds with $p = 1$, then the $\mathcal{B}_\tau^\pi$ is contractive on $M_b(\mathsf{X} \times \mathsf{A})$ endowed with the supnorm. Its unique fixed point is q_τ^π .

Proof. Observe that

$$\|\mathcal{B}_\tau^\pi q\|_{\sup} \leq \|r\|_{\sup} + \gamma \|q\|_{\sup} + \gamma \sup_{x,a} \|\tau \text{KL}(\pi_- \| \pi_-^{\text{ref}})\|_{L^1(P_{x,a})} < \infty,$$

by (4.1), and

$$\|\mathcal{B}_\tau^\pi q - \mathcal{B}_\tau^\pi q'\|_{\sup} \leq \gamma \|\mathcal{V}_\tau q - \mathcal{V}_\tau q'\|_{\sup} \leq \gamma \|q - q'\|_{\sup}.$$

□

Next, we proceed with policy improvement.

Lemma A.7. *If $r \in M_b(\mathsf{X} \times \mathsf{A})$ and $\gamma < 1$, then the soft Bellman optimality operator is a contraction on $M_b(\mathsf{X} \times \mathsf{A})$ endowed with the supremum norm. Thus, it has a unique fixed point q_τ^* .*

Proof. Observe that

$$\|\mathcal{B}_\tau^* q\|_{\sup} \leq \|r\|_{\sup} + \gamma \|q\|_{\sup} < \infty$$

and

$$\|\mathcal{B}_\tau^* q - \mathcal{B}_\tau^* q'\|_{\sup} \leq \gamma \|\mathcal{V}_\tau q - \mathcal{V}_\tau q'\|_{\sup} \leq \gamma \|q - q'\|_{\sup}.$$

□

Lemma A.8. *The following equality holds true: $q_\tau^{\mathcal{G}_\tau q_\tau^*} = q_\tau^*$.*

Proof. Observe that

$$\begin{aligned} \mathcal{B}_\tau^{\mathcal{G}_\tau q_\tau^*} q_\tau^* &= r(x, a) + \gamma \int q_\tau^* - \tau \text{KL}((\mathcal{G}_\tau q_\tau^*)_+ \| \pi_-^{\text{ref}}) d\tilde{P}_{x,a}^{\mathcal{G}_\tau q_\tau^*} \\ &= r(x, a) + \gamma \int \mathcal{V}_\tau q_\tau^* dP_{x,a} \\ &= \mathcal{B}_\tau^* q_\tau^* \\ &= q_\tau^*. \end{aligned}$$

In words, q_τ^* is a fixed point of the soft Bellman (policy evaluation) operator with $\pi = \mathcal{G}_\tau q_\tau^*$. As $\mathcal{G}_\tau q_\tau^*$ is a Boltzmann–Gibbs policy with a bounded potential, by Lemma A.6 and the preceding note, this operator is a contraction with a unique fixed point. Hence,

$$q_\tau^* = q_\tau^{\mathcal{G}_\tau q_\tau^*},$$

the unique fixed point of $\mathcal{B}_\tau^\pi$ with $\pi = \mathcal{G}_\tau q_\tau^*$, as desired. □

Lemma A.9. *For every $\pi \in \mathsf{K}(\mathsf{X}, \mathcal{P}(\mathsf{A}))$, we have that*

$$q_\tau^* \geq q_\tau^\pi.$$

Proof. First, we prove that

$$\mathcal{B}_\tau^* q_\tau^\pi \geq q_\tau^\pi. \tag{A.4}$$

By definition and the Donsker–Varadhan variational principle,

$$\begin{aligned} q_\tau^\pi(x, a) &= r(x, a) + \gamma \int \left[\int q_\tau^\pi(x', a') d\pi_{x'}(a') - \tau \text{KL}(\pi'_x \| \pi_{x'}^{\text{ref}}) \right] dP_{x,a}(x') \\ &\leq r(x, a) + \gamma \int (\mathcal{V}_\tau q_\tau^\pi)(x') dP_{x,a}(x') \\ &= (\mathcal{B}_\tau^* q_\tau^\pi)(x, a). \end{aligned}$$

Now we conclude. Let $q_{\tau,0}^\pi := \max\{q_\tau^\pi, 0\}$. By (A.4) and since $\mathcal{B}_\tau^*$ is a monotone operator,

$$q_\tau^\pi \leq \mathcal{B}_\tau^* q_\tau^\pi \leq \mathcal{B}_\tau^* q_{\tau,0}^\pi \leq \mathcal{B}_\tau^*(\mathcal{B}_\tau^* q_{\tau,0}^\pi) \leq \dots \leq \lim_{n \rightarrow \infty} (\mathcal{B}_\tau^*)^n q_{\tau,0}^\pi = q_\tau^*,$$

where the final equality holds by Lemma A.7, noting that $\|q_{\tau,0}^\pi\|_{\sup} < \infty$. □

Finally, we prove Theorem 2.3.

Theorem 2.3. Let $\tau > 0$. The policy $\pi^{\tau,*} := \mathcal{G}_\tau q_\tau^*$ is optimal, and uniquely so. More precisely, for all $\nu_0, \nu'_0 \in \mathcal{P}(\mathbf{X})$, we have that $\arg \max_{\mathcal{O}(\nu_0)} \mathcal{J}_\tau = \pi^{\tau,*} = \arg \max_{\mathcal{O}(\nu'_0)} \mathcal{J}_\tau$. [Source]

Proof. For any $\pi \in \mathbf{K}(\mathbf{X}, \mathcal{P}(\mathbf{A}))$, let

$$v_\tau^\pi(x) := \int q_\tau^\pi(x, a) d\pi_x(a) - \tau \text{KL}(\pi_x \| \pi_x^{\text{ref}}).$$

Note that

$$\mathcal{J}_\tau(\mu^\pi) = (1 - \gamma) \int v_\tau^\pi d\nu_0$$

if $\mu^\pi \in \mathcal{O}(\nu_0)$. Hence, it suffices to show that

$$v_\tau^{\mathcal{G}_\tau q_\tau^*} \geq \sup_\pi v_\tau^\pi. \quad (\text{A.5})$$

Observe, by Lemma A.8,

$$v_\tau^{\mathcal{G}_\tau q_\tau^*}(x) = \int q_\tau^*(x, a) d(\mathcal{G}_\tau q_\tau^*)_x(a) - \tau \text{KL}((\mathcal{G}_\tau q_\tau^*)_x \| \pi_x^{\text{ref}}) = (\mathcal{V}_\tau q_\tau^*)(x).$$

Thus, by the Donsker–Varadhan variational principle,

$$\begin{aligned} v_\tau^{\mathcal{G}_\tau q_\tau^*}(x) - v_\tau^\pi(x) &= (\mathcal{V}_\tau q_\tau^*)(x) - \int q_\tau^\pi(x, a) d\pi_x(a) + \tau \text{KL}(\pi_x \| \pi_x^{\text{ref}}) \\ &\geq (\mathcal{V}_\tau q_\tau^*)(x) - (\mathcal{V}_\tau q_\tau^\pi)(x). \end{aligned}$$

Finally, by Lemma A.9, we have that

$$\mathcal{V}_\tau q_\tau^* - \mathcal{V}_\tau q_\tau^\pi \geq 0,$$

for all π , as desired. □

To conclude this section, we prove Theorem 2.5.

Theorem 2.5. Suppose that $r \in M_b(\mathbf{X} \times \mathbf{A})$ and that $\mathbf{X} \times \mathbf{A}$ is finite. For every $\tau > 0$, let μ_τ^* be the maximizer of $\mathcal{J}_\tau$ over $\mathcal{O}(\nu_0)$. If Assumption 2.4 holds, the sequence $(\mu_\tau^*)_{\tau>0}$ has a unique setwise limit as τ tends to zero. This limit μ_0^* is the minimizer of $\mathcal{R}$ over $\arg \sup_{\mathcal{O}(\nu_0)} \mathcal{J}_0$. [Source]

Proof. Let $\mu^* \in \{\arg \sup_{\mathcal{O}(\nu_0)} \mathcal{J}_0\} \cap \{\mathcal{R} < \infty\}$. Then,

$$0 \leq \mathcal{J}_0(\mu^*) - \mathcal{J}_0(\mu_\tau^*) \leq \tau(\mathcal{R}(\mu^*) - \mathcal{R}(\mu_\tau^*)) < \infty.$$

In turn, for all $\tau > 0$, we deduce that $\mathcal{R}(\mu_\tau^*) \leq \mathcal{R}(\mu^*)$.

Now let μ_0 be any limit of any setwise convergent subsequence of $(\mu_\tau^*)_{\tau>0}$ (cf. the proof of Theorem A.5 and Remark A.4). As $\mathcal{R}$ is weakly lower semi-continuous we find that

$$\mathcal{R}(\mu_0) \leq \liminf_{\tau \rightarrow 0} \mathcal{R}(\mu_\tau^*) \leq \mathcal{R}(\mu^*).$$

Moreover, since $\mathcal{R}(\mu^*) < \infty$, by Lemma B.2, and as $r \in M_b(\mathbf{X} \times \mathbf{A})$, we deduce that

$$\lim_{\tau \rightarrow 0} \tau(\mathcal{R}(\mu^*) - \mathcal{R}(\mu_\tau^*)) = 0 \quad \text{and} \quad \mathcal{J}_0(\mu_0) = \mathcal{J}_0(\mu^*).$$

Therefore, $\mu_0 \in \arg \sup_{\mathcal{O}(\nu_0)} \mathcal{J}_0$ and minimizes $\mathcal{R}$ over $\arg \sup_{\mathcal{O}(\nu_0)} \mathcal{J}_0$.

Since $\mathcal{R}$ is strictly convex, by Lemma 2.1, and the set $\arg \sup_{\mathcal{O}(\nu_0)} \mathcal{J}_0$ is convex, $\mathcal{R}$ has at most one minimizer among this set. In turn, only one such limit μ_0 exists, call it μ_0^* . Hence, $\mu_\tau^* \rightarrow \mu_0^*$ setwise, as desired. □

B Proofs for Section 3

Before proving the results from Section 3, we introduce some helpful notation. For any $q : \mathsf{X} \times \mathsf{A} \rightarrow \mathbb{R}$, we define

$$(\mathcal{M}_\tau q)(x) := \int q(x, \cdot) d(\mathcal{G}_\tau q)_x.$$

Additionally, we will define $M_\tau : L^\infty(\mathsf{X} \times \mathsf{A}) \rightarrow \mathbb{R} \cup \{\infty\}$ according to

$$M_\tau(q) := \sup_x \{\text{ess sup}_{\pi_x^{\text{ref}}} q(x, \cdot) - \mathcal{V}_\tau q(x)\}.$$

We start by proving that $\mathcal{B}_{\text{ref}}^*$ is contractive on $M_b(\mathsf{X} \times \mathsf{A})$.

Lemma 3.1. *Let $r \in M_b(\mathsf{X} \times \mathsf{A})$, $\gamma < 1$, and $\mathcal{B}_{\text{ref}}^* : M(\mathsf{X} \times \mathsf{A}) \rightarrow M(\mathsf{X} \times \mathsf{A})$ be defined by*

$$(\mathcal{B}_{\text{ref}}^* q)(x, a) := r(x, a) + \gamma \int \text{ess sup}_{\pi_{x'}^{\text{ref}}} q(x', \cdot) dP_{x,a}(x').$$

Then $\mathcal{B}_{\text{ref}}^$ is a contraction on $M_b(\mathsf{X} \times \mathsf{A})$. Thus, it has a unique fixed point q_{ref}^* .* [Source]

Proof. First, observe that

$$\|\mathcal{B}_{\text{ref}}^* q\|_{\text{sup}} \leq \|r\|_{\text{sup}} + \gamma \|q\|_{\text{sup}}.$$

Second,

$$\begin{aligned} \|\mathcal{B}_{\text{ref}}^* q - \mathcal{B}_{\text{ref}}^* q'\|_{\text{sup}} &\leq \gamma \sup_{x'} |\text{ess sup}_{\pi_{x'}^{\text{ref}}} q(x', \cdot) - \text{ess sup}_{\pi_{x'}^{\text{ref}}} q'(x', \cdot)| \\ &\leq \gamma \sup_{x'} (\text{ess sup}_{\pi_{x'}^{\text{ref}}} |q(x', \cdot) - q'(x', \cdot)|) \\ &\leq \gamma \|q - q'\|_{\text{sup}}. \end{aligned}$$

The lemma follows by the Banach fixed point theorem. □

Next we prove value function convergence.

Theorem 3.2. *We have that $q_\tau^* \rightarrow q_{\text{ref}}^*$ monotonically as $\tau \rightarrow 0$.* [Source]

Proof. Since q_τ^* is bounded (as the fixed point of a contractive operator on $M_b(\mathsf{X} \times \mathsf{A})$), there exists $q_0 : \mathsf{X} \times \mathsf{A} \rightarrow \mathbb{R}$ such that $q_\tau^* \rightarrow q_0$ monotonically and pointwise as $\tau \rightarrow 0$, as a direct consequence of Lemma B.1. Therefore, by the monotone convergence theorem,

$$\lim_{\sigma \rightarrow 0} \mathcal{V}_\tau q_\sigma^*(x) = \lim_{\sigma \rightarrow 0} \log \|\exp(q_\sigma^*(x, \cdot))\|_{L^{1/\tau}(\pi_x^{\text{ref}})} = \log \|\exp(q_0(x, \cdot))\|_{L^{1/\tau}(\pi_x^{\text{ref}})}.$$

Consequently,

$$\begin{aligned} \lim_{\tau \rightarrow 0} \lim_{\sigma \rightarrow 0} \mathcal{V}_\tau q_\sigma^*(x) &= \lim_{\tau \rightarrow 0} \log \|\exp(q_0(x, \cdot))\|_{L^{1/\tau}(\pi_x^{\text{ref}})} \\ &= \log \|\exp(q_0(x, \cdot))\|_{L^\infty(\pi_x^{\text{ref}})} \\ &= \text{ess sup}_{\pi_x^{\text{ref}}} q_0(x, \cdot). \end{aligned}$$

The second step holds since for any $f \in L^\infty$, $\|f\|_p$ converges up to $\|f\|_\infty$ as $p \rightarrow \infty$. So, since the sequence $(\mathcal{V}_\tau q_\sigma^*(x))_{\tau, \sigma \geq 0}$ is monotone and bounded, its limit exists, and coincides with that computed above:

$$\lim_{\tau \rightarrow 0} \mathcal{V}_\tau q_\tau^*(x) = \text{ess sup}_{\pi_x^{\text{ref}}} q_0(x, \cdot).$$

Since q_τ^* is the unique fixed point of $\mathcal{B}_\tau^*$, by the monotone convergence theorem, we have

$$\begin{aligned} q_0(x, a) &= \lim_{\tau \rightarrow 0} q_\tau^*(x, a) \\ &= \lim_{\tau \rightarrow 0} (\mathcal{B}_\tau^* q_\tau^*)(x, a) \\ &= r(x, a) + \gamma \int \lim_{\tau \rightarrow 0} \mathcal{V}_\tau q_\tau^*(x') dP_{x,a}(x') \\ &= r(x, a) + \gamma \int \text{ess sup}_{\pi_{x'}^{\text{ref}}} q_0(x', \cdot) dP_{x,a}(x'), \end{aligned}$$

so that q_0 is a fixed point of $\mathcal{B}_{\text{ref}}^*$. Since the $\mathcal{B}_{\text{ref}}^*$ has a fixed point q_{ref}^* , it follows that $q_0 = q_{\text{ref}}^*$. □

Now we prove our core estimate.

Theorem 3.6. *Let $q, q' \in M(\mathbb{X} \times \mathbb{A})$. For any $\tau > 0$ and any $x \in \mathbb{X}$,*

$$\begin{aligned} & \|(\mathcal{G}_\tau q)_x - (\mathcal{G}_\tau q')_x\|_{\text{TV}} \\ & \leq \min \left\{ \sqrt{\tau^{-1} \|q(x, \cdot) - q'(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})}}, \frac{1}{2} \sinh(4\tau^{-1} \|q(x, \cdot) - q'(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})}) \right\}. \end{aligned}$$

In particular,

$$\|(\mathcal{G}_\tau q)_x - (\mathcal{G}_\tau q')_x\|_{\text{TV}} \leq \frac{2e-3}{4} \tau^{-1} \|q(x, \cdot) - q'(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})},$$

if $\|q(x, \cdot) - q'(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})} < \tau/2$.

[Source]

Proof. Let $\pi := \mathcal{G}_\tau q, \pi' := \mathcal{G}_\tau q'$. By Lemma B.6, we have

$$\|\pi_x - \pi'_x\|_{\text{TV}} \leq \sqrt{\tau^{-1} \|q(x, \cdot) - q'(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})}}.$$

Moreover, by Lemma B.9,

$$\|\pi_x - \pi'_x\|_{\text{TV}} \leq \frac{1}{2} \sinh(4\tau^{-1} \|q(x, \cdot) - q'(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})}).$$

This concludes the proof of the first claim. Next, we recall that

$$\sinh(y) = \sum_{k=0}^{\infty} \frac{y^{2k+1}}{(2k+1)!},$$

which is convergent for any $y \in \mathbb{C}$. Therefore, for $y \in (0, 1)$, we have

$$\begin{aligned} \sinh(y) & \leq y + \frac{y^3}{3!} + \frac{y^5}{5!} + \dots \\ & \leq y \left(1 + e^y - \frac{5}{2} \right) \\ & \leq y \left(e - \frac{3}{2} \right). \end{aligned}$$

So, when $\|q(x, \cdot) - q'(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})} < \tau/2$, it follows that

$$\begin{aligned} \|\pi_x - \pi'_x\|_{\text{TV}} & \leq \frac{1}{2} \left(e - \frac{3}{2} \right) (4\tau^{-1} \|q(x, \cdot) - q'(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})}) \\ & = \frac{2e-3}{4} \tau^{-1} \|q(x, \cdot) - q'(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})}. \end{aligned}$$

□

Finally, we prove policy and return distribution convergence.

Theorem 3.9. *Under Assumption 3.4, if $\sigma = \sigma(\tau)$ is such that $\lim_{\tau \rightarrow 0} \sigma/\tau = 0$, then $\pi_x^{\tau, \sigma} \rightarrow \pi_x^{\text{ref}, \star}$ as $\tau \rightarrow 0$, for all $x \in \mathbb{X}$, in TV if $\mathbb{A}$ is discrete and weakly if $\mathbb{A}$ is continuous.*

[Source]

Proof. Recall $\pi^{\tau, \sigma} := \mathcal{G}_\tau q_\sigma^*$. By Theorem 3.6 and Lemma B.10,

$$\limsup_{\tau \rightarrow 0} \sup_x \|(\mathcal{G}_\tau q_\sigma^*)_x - (\mathcal{G}_\tau q_{\text{ref}}^*)_x\|_{\text{TV}} \lesssim - \lim_{\tau \rightarrow 0} \frac{\sigma}{\tau} \log p_{\text{ref}}. \quad (\text{B.1})$$

Consequently, $\pi_x^{\tau, \sigma} \rightarrow \pi_x^{\text{ref}, \star}$ if and only if $(\mathcal{G}_\tau q_{\text{ref}}^*)_x \rightarrow \pi_x^{\text{ref}, \star}$, and in whatever sense the later convergence occurs. In particular, if $\mathbb{A}$ is continuous, this is in the weak sense. While, if $\mathbb{A}$ is discrete, this is in total variation. □

Theorem 3.10. Suppose $\mathbf{A}$ is discrete and Assumption 3.4 holds. If $\sigma = \sigma(\tau)$ is such that $\sigma/\tau \rightarrow 0$ as $\tau \rightarrow 0$, then, for any $p, p' \in [1, \infty)$ and $\omega \in \mathcal{P}(\mathbf{X} \times \mathbf{A})$, as $\tau \rightarrow 0$, the return distribution functions $\zeta^{\tau, \sigma}$ of the temperature-decoupled policies $\pi^{\tau, \sigma}$ satisfy $d_{p; p', \omega}(\zeta^{\tau, \sigma}, \zeta^{\pi^{\text{ref}, *}}) \rightarrow 0$. [Source]

Proof. By the distributional Bellman equation [5], we have that

$$\begin{aligned} & d_p(\zeta_{x,a}^*, \zeta_{x,a}^{\tau, \sigma}) \\ & \leq \int d_p \left((\mathbf{b}_{r(x,a), \gamma} \circ \text{proj}^{\mathbb{R}})_{\#}(\zeta_{x',-}^* \otimes \pi_{x',-}^{\text{ref}, *}), (\mathbf{b}_{r(x,a), \gamma} \circ \text{proj}^{\mathbb{R}})_{\#}(\zeta_{x',-}^{\tau, \sigma} \otimes \pi_{x',-}^{\tau, \sigma}) \right) dP_{x,a}(x') \\ & = \int d_p \left((\mathbf{b}_{0,\gamma} \circ \text{proj}^{\mathbb{R}})_{\#}(\zeta_{x',-}^* \otimes \pi_{x',-}^{\text{ref}, *}), (\mathbf{b}_{0,\gamma} \circ \text{proj}^{\mathbb{R}})_{\#}(\zeta_{x',-}^{\tau, \sigma} \otimes \pi_{x',-}^{\tau, \sigma}) \right) dP_{x,a}(x') \\ & = \gamma \int I_{\tau, \sigma} dP_{x,a} \end{aligned}$$

where

$$I_{\tau, \sigma}(x') = d_p \left((\mathbf{b}_{0,1} \circ \text{proj}^{\mathbb{R}})_{\#}(\zeta_{x',-}^* \otimes \pi_{x',-}^{\text{ref}, *}), (\mathbf{b}_{0,1} \circ \text{proj}^{\mathbb{R}})_{\#}(\zeta_{x',-}^{\tau, \sigma} \otimes \pi_{x',-}^{\tau, \sigma}) \right).$$

We now derive a bound on $I_{\tau, \sigma}$. Starting with a triangle inequality,

$$\begin{aligned} I_{\tau, \sigma}(x') & \leq d_p \left((\mathbf{b}_{0,1} \circ \text{proj}^{\mathbb{R}})_{\#}(\zeta_{x',-}^* \otimes \pi_{x',-}^{\text{ref}, *}), (\mathbf{b}_{0,1} \circ \text{proj}^{\mathbb{R}})_{\#}(\zeta_{x',-}^{\tau, \sigma} \otimes \pi_{x',-}^{\text{ref}, *}) \right) \\ & \quad + d_p \left((\mathbf{b}_{0,1} \circ \text{proj}^{\mathbb{R}})_{\#}(\zeta_{x',-}^{\tau, \sigma} \otimes \pi_{x',-}^{\text{ref}, *}), (\mathbf{b}_{0,1} \circ \text{proj}^{\mathbb{R}})_{\#}(\zeta_{x',-}^{\tau, \sigma} \otimes \pi_{x',-}^{\tau, \sigma}) \right) \\ & \stackrel{(a)}{\leq} \int d_p(\zeta_{x',a'}^*, \zeta_{x',a'}^{\tau, \sigma}) d\pi_{x',-}^{\text{ref}, *}(a') \\ & \quad + d_p \left((\mathbf{b}_{0,1} \circ \text{proj}^{\mathbb{R}})_{\#}(\zeta_{x',-}^{\tau, \sigma} \otimes \pi_{x',-}^{\text{ref}, *}), (\mathbf{b}_{0,1} \circ \text{proj}^{\mathbb{R}})_{\#}(\zeta_{x',-}^{\tau, \sigma} \otimes \pi_{x',-}^{\tau, \sigma}) \right) \\ & \stackrel{(b)}{\leq} \int d_p(\zeta_{x',a'}^*, \zeta_{x',a'}^{\tau, \sigma}) d\pi_{x',-}^{\text{ref}, *}(a') \\ & \quad + 2^{\frac{p-1}{p}} \left(\int \left[\int |z|^p d\zeta_{x',a'}^{\tau, \sigma}(z) \right] d|\pi_{x',-}^{\text{ref}, *} - \pi_{x',-}^{\tau, \sigma}|(a') \right)^{1/p} \\ & \stackrel{(c)}{\leq} \int d_p(\zeta_{x',a'}^*, \zeta_{x',a'}^{\tau, \sigma}) d\pi_{x',-}^{\text{ref}, *}(a') + \frac{2^{\frac{p-1}{p}}}{1-\gamma} \|r\|_{\text{sup}} \|\pi_{x',-}^{\text{ref}, *} - \pi_{x',-}^{\tau, \sigma}\|_{\text{TV}}^{1/p}, \end{aligned}$$

where (a) follows by the convexity of the Wasserstein metrics [40, 6], (b) applies [40, Theorem 6.15], and (c) leverages that the support $\zeta_{x',a'}^{\pi_{x',-}^{\text{ref}, *}}$ lives in a ball of radius $\|r\|_{\text{sup}}/(1-\gamma)$, for any π and $(x', a') \in \mathbf{X} \times \mathbf{A}$.

So, thus far, we have shown that

$$d_p(\zeta_{x,a}^*, \zeta_{x,a}^{\tau, \sigma}) \leq \gamma \int d_p(\zeta_{x',a'}^*, \zeta_{x',a'}^{\tau, \sigma}) d\check{P}_{x,a}^{\pi^{\text{ref}, *}}(x', a') + C\gamma \int \|\pi_{x',-}^{\text{ref}, *} - \pi_{x',-}^{\tau, \sigma}\|_{\text{TV}}^{1/p} dP_{x,a}(x').$$

By Theorem 3.9, the total variation term tends to zero as τ tends to zero. Thus, defining $\iota(x', a') := \limsup_{\tau \rightarrow 0} d_p(\zeta_{x',a'}^*, \zeta_{x',a'}^{\tau, \sigma})$, this implies that

$$\iota(x', a') \leq \gamma \int \iota(y, b) d\check{P}_{x,a'}^{\pi^{\text{ref}, *}}(y, b),$$

In turn, $\sup \iota \leq \gamma \sup \iota$, implying that $\iota \equiv 0$. Therefore, $d_p(\zeta_{x,a}^*, \zeta_{x,a}^{\tau, \sigma}) \rightarrow 0$ pointwise over $\mathbf{X} \times \mathbf{A}$, so by the dominated convergence theorem, $d_{p; p', \omega}(\zeta^*, \zeta^{\tau, \sigma}) \rightarrow 0$ for any $p' \in [1, \infty)$ and $\omega \in \mathcal{P}(\mathbf{X} \times \mathbf{A})$. $\square$

B.1 Supplemental Lemmas for Section 3

The following lemma translates immediately from the corresponding result in tabular MDPs; we prove it here for completeness.

Lemma B.1. *If $\tau \leq \sigma$, then $q_\sigma^* \leq q_\tau^*$.*

Proof. By the monotonicity of $\mathcal{B}_\tau^*$,

$$q_\sigma^* = r + \gamma \int \mathcal{V}_\sigma q_\sigma^* dP_{-, -} \leq r + \gamma \int \mathcal{V}_\tau q_\sigma^* dP_{-, -} = \mathcal{B}_\tau^* q_\sigma^* \leq \dots \leq q_\tau^*$$

(cf. the proof of Lemma A.9). $\square$

Lemma B.2. *Let $\sigma = \sigma(\tau)$ and suppose $\sigma \rightarrow 0$ as $\tau \rightarrow 0$. Then*

$$\tau \text{KL}((\mathcal{G}_\tau q_\sigma^*)_x \parallel \pi_x^{\text{ref}}) \xrightarrow{\tau \downarrow 0} 0.$$

Proof. Expanding the KL, we have

$$\begin{aligned} \tau \text{KL}((\mathcal{G}_\tau q_\sigma^*)_x \parallel \pi_x^{\text{ref}}) &= \int_{\mathbf{A}} (q_\sigma^*(x, \cdot) - \mathcal{V}_\tau q_\sigma^*(x)) d(\mathcal{G}_\tau q_\sigma^*)_x \\ &\leq \text{ess sup}_{\pi_x^{\text{ref}}} q_\sigma^*(x, \cdot) - (\mathcal{V}_\tau q_\sigma^*)(x) \\ &\leq v_{\text{ref}}^*(x) - (\mathcal{V}_\tau q_\sigma^*)(x). \end{aligned}$$

where the final inequality holds by Lemma B.1. Since $\sigma = \sigma(\tau) \leq \tau$, we have

$$\mathcal{V}_\tau q_\sigma^*(x) = \log \|\exp(q_\sigma^*(x, \cdot))\|_{L^{1/\tau}(\pi_x^{\text{ref}})} \geq \log \|\exp(q_\tau^*(x, \cdot))\|_{L^{1/\tau}(\pi_x^{\text{ref}})} = \mathcal{V}_\tau q_\tau^*,$$

where the inequality again is due to Lemma B.1. Consequently, we have

$$\limsup_{\tau \rightarrow 0} \tau \text{KL}(\pi_x^{\tau, \sigma} \parallel \pi_x^{\text{ref}}) \leq v_{\text{ref}}^*(x) - \liminf_{\tau \rightarrow 0} \mathcal{V}_\tau q_\tau^*(x) = v_{\text{ref}}^*(x) - v_{\text{ref}}^*(x) = 0,$$

where the penultimate step is due to the fact that $\mathcal{V}_\tau q_\tau^* \rightarrow v_{\text{ref}}^*$ monotonically, as shown in the proof of Theorem 3.2. $\square$

Lemma B.3. *For every $q \in M_b(\mathbf{X} \times \mathbf{A})$, with the notation [above](#), $M_\tau(q) \rightarrow 0$ as $\tau \rightarrow 0$. If Assumption 3.4 is satisfied, then for any $\sigma > 0$,*

$$M_\tau(q_\sigma^*) \leq -\tau \log p_{\text{ref}}.$$

Proof. First, we observe that

$$\lim_{\tau \rightarrow 0} \mathcal{V}_\tau q(x) = \lim_{\tau \rightarrow 0} \log \|\exp(q(x, \cdot))\|_{L^{1/\tau}(\pi_x^{\text{ref}})} = \log \|\exp(q(x, \cdot))\|_{L^\infty(\pi_x^{\text{ref}})} = \text{ess sup}_{\pi_x^{\text{ref}}} q(x, \cdot).$$

This is a monotone limit in τ , as it is known that for any $f \in L^\infty$, $\|f\|_p$ converges up to $\|f\|_{L^\infty}$ as $p \rightarrow \infty$. Thus, we see that

$$M_\tau(q) = \sup_x \left(\text{ess sup}_{\pi_x^{\text{ref}}} q(x, \cdot) - \mathcal{V}_\tau q(x) \right) \rightarrow \sup_x \left(\text{ess sup}_{\pi_x^{\text{ref}}} q(x, \cdot) - \text{ess sup}_{\pi_x^{\text{ref}}} q(x, \cdot) \right) = 0$$

as claimed. Now, under Assumption 3.4, we have

$$\begin{aligned} M_\tau(q_\sigma^*) &= \sup_x \left(\text{ess sup}_{\pi_x^{\text{ref}}} q_\sigma^*(x, \cdot) - \mathcal{V}_\tau q_\sigma^*(x) \right) \\ &= \sup_x \left(\text{ess sup}_{\pi_x^{\text{ref}}} q_\sigma^*(x, \cdot) - \tau \log \int \exp(\tau^{-1} q_\sigma^*(x, \cdot)) d\pi_x^{\text{ref}} \right) \\ &= \sup_x \left(-\tau \log \int \exp(\tau^{-1} (q_\sigma^*(x, \cdot) - \text{ess sup}_{\pi_x^{\text{ref}}} q_\sigma^*(x, \cdot))) d\pi_x^{\text{ref}} \right). \end{aligned}$$

Let $\mathbf{B}_x = \{a \in \mathbf{A} : q_\sigma^*(x, a) = \text{ess sup}_{\pi_x^{\text{ref}}} q_\sigma^*(x, \cdot)\}$. Then,

$$\begin{aligned} M_\tau(q_\sigma^*) &= \sup_x \left[-\tau \log \left(\int_{\mathbf{B}_x} \exp(\tau^{-1} (q_\sigma^*(x, \cdot) - \text{ess sup}_{\pi_x^{\text{ref}}} q_\sigma^*(x, \cdot))) d\pi_x^{\text{ref}} \right. \right. \\ &\quad \left. \left. + \int_{\mathbf{A} \setminus \mathbf{B}_x} \exp(\tau^{-1} (q_\sigma^*(x, \cdot) - \text{ess sup}_{\pi_x^{\text{ref}}} q_\sigma^*(x, \cdot))) d\pi_x^{\text{ref}} \right) \right] \\ &= \sup_x \left[-\tau \log \left(\pi_x^{\text{ref}}(\mathbf{B}_x) + \int_{\mathbf{A} \setminus \mathbf{B}_x} \exp(\tau^{-1} (q_\sigma^*(x, \cdot) - \text{ess sup}_{\pi_x^{\text{ref}}} q_\sigma^*(x, \cdot))) d\pi_x^{\text{ref}} \right) \right] \\ &\leq \sup_x -\tau \log \pi_x^{\text{ref}}(\mathbf{B}_x) \\ &\leq \tau \log p_{\text{ref}}, \end{aligned}$$

where the final inequality invokes Assumption 3.4. $\square$

Lemma B.4. For all $\tau > 0$ and any $q \in L^\infty(\mathsf{X} \times \mathsf{A})$,

$$\mathcal{B}_{\text{ref}}^* q \geq \mathcal{B}_\tau^* q,$$

where $\mathcal{B}_{\text{ref}}^*$ denotes the Bellman optimality operator (cf. Lemma 3.1).

Proof. A direct calculation gives

$$\begin{aligned} \mathcal{V}_\tau q(x) &= \tau \log \left(\int \exp(\tau^{-1} q(x, a)) \, d\pi_x^{\text{ref}}(a) \right) \\ &\leq \tau \log \left(\int \exp(\tau^{-1} \text{ess sup}_{\pi_x^{\text{ref}}} q(x, \cdot)) \, d\pi_x^{\text{ref}} \right) \\ &= \text{ess sup}_{\pi_x^{\text{ref}}} q(x, \cdot). \end{aligned}$$

Therefore, it immediate follows that

$$\begin{aligned} (\mathcal{B}_{\text{ref}}^* q)(x, a) &= r(x, a) + \gamma \int \text{ess sup}_{\pi_{x'}^{\text{ref}}} q(x', \cdot) \, dP_{x,a}(x') \\ &\geq r(x, a) + \gamma \int \mathcal{V}_\tau q(x') \, dP_{x,a}(x') \\ &= (\mathcal{B}_\tau^* q)(x, a). \end{aligned}$$

□

The follow proof is essentially the performance difference bound in [37].

Lemma B.5. For all $n \geq 1$ and any $\tau > 0$,

$$(\mathcal{B}_{\text{ref}}^*)^n q_\tau^* - q_\tau^* \leq \sum_{k=1}^n \gamma^k M_\tau(q_\tau^*).$$

If, additionally, Assumption 3.4 is satisfied, then

$$(\mathcal{B}_{\text{ref}}^*)^n q_\tau^* - q_\tau^* \leq -\tau \log p_{\text{ref}} \sum_{k=1}^n \gamma^k.$$

Proof. We begin with the first statement. Recall that q_τ^* is the fixed point of $\mathcal{B}_\tau^*$, so that $q_\tau^* = \mathcal{B}_\tau^* q_\tau^*$. We will proceed by induction on n . For $n = 1$, we observe that

$$\begin{aligned} (\mathcal{B}_{\text{ref}}^* q_\tau^*)(x, a) - q_\tau^*(x, a) &= (\mathcal{B}_{\text{ref}}^* q_\tau^*)(x, a) - (\mathcal{B}_\tau^* q_\tau^*)(x, a) \\ &= \gamma \int \left(\text{ess sup}_{\pi_{x'}^{\text{ref}}} q_\tau^*(x', \cdot) - \mathcal{V}_\tau q_\tau^*(x') \right) \, dP_{x,a}(x') \\ &\leq \gamma M_\tau(q_\tau^*), \end{aligned}$$

recalling the notation established above. This proves the base case. Now, assume the statement holds for all $m \leq n$. We have

$$\begin{aligned} (\mathcal{B}_{\text{ref}}^*)^{n+1} q_\tau^* - q_\tau^* &= (\mathcal{B}_{\text{ref}}^*)^{n+1} q_\tau^* - \mathcal{B}_\tau^{*n+1} q_\tau^* \\ &\leq \mathcal{B}_{\text{ref}}^* \left(\mathcal{B}_\tau^{*n} q_\tau^* + \sum_{k=1}^n \gamma^k M_\tau(q_\tau^*) \right) - \mathcal{B}_\tau^{*n+1} q_\tau^* \\ &= \mathcal{B}_{\text{ref}}^* q_\tau^* + \sum_{k=1}^n \gamma^{k+1} M_\tau(q_\tau^*) - \mathcal{B}_\tau^{*n+1} q_\tau^* \\ &\leq \gamma M_\tau(q_\tau^*) + \sum_{k=2}^{n+1} \gamma^k M_\tau(q_\tau^*) \\ &= \sum_{k=1}^{n+1} \gamma^k M_\tau(q_\tau^*), \end{aligned}$$

where the first inequality invokes the induction hypothesis, and the second inequality is due to the base case. Thus, we have shown that the claimed statement holds for any $n \in \mathbb{N}$.

When Assumption 3.4 is satisfied, by Lemma B.3, we have $M_\tau(q_\tau^*) \leq -\tau \log p_{\text{ref}}$, and the second statement follows. $\square$

Lemma B.6. *Let $q, q' \in L^\infty(\mathsf{X} \times \mathsf{A})$. Then for any $\tau > 0$ and any $x \in \mathsf{X}$,*

$$\|(\mathcal{G}_\tau q)_x - (\mathcal{G}_\tau q')_x\|_{\text{TV}} \leq \sqrt{\tau^{-1} \|q(x, \cdot) - q'(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})}}.$$

Proof. Let $\pi = \mathcal{G}_\tau q$ and let $\pi' = \mathcal{G}_\tau q'$. By Pinsker's inequality, we have

$$\|\pi_x - \pi'_x\|_{\text{TV}} \leq \sqrt{\frac{1}{2} \text{KL}(\pi_x \parallel \pi'_x)}.$$

Since $q, q' \in L^\infty(\mathsf{X} \times \mathsf{A})$, π_x, π'_x are mutually absolutely continuous. Expanding the KL divergence, we have

$$\begin{aligned} \text{KL}(\pi_x \parallel \pi'_x) &= \int_{\mathsf{A}} \log \frac{\pi_x}{\pi'_x} d\pi_x \\ &= \int_{\mathsf{A}} \log \frac{\pi_x^{\text{ref}}(a) \exp(\tau^{-1}(q(x, a) - \mathcal{V}_\tau q(x)))}{\pi_x^{\text{ref}}(a) \exp(\tau^{-1}(q'(x, a) - \mathcal{V}_\tau q'(x)))} d\pi_x(a) \\ &= \int_{\mathsf{A}} \tau^{-1} (q(x, a) - \mathcal{V}_\tau q(x) - q'(x, a) + \mathcal{V}_\tau q'(x)) d\pi_x(a) \\ &\leq \tau^{-1} \|q(x, \cdot) - q'(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})} + \tau^{-1} \|\mathcal{V}_\tau q(x) - \mathcal{V}_\tau q'(x)\|_{L^\infty(\pi_x^{\text{ref}})} \\ &\leq 2\tau^{-1} \|q(x, \cdot) - q'(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})}, \end{aligned}$$

where the last inequality holds since $\mathcal{V}_\tau$ is 1-Lipschitz, as shown in the proof of Lemma B.4. Substituting back into Pinsker's inequality, we have

$$\|\pi_x - \pi'_x\|_{\text{TV}} \leq \sqrt{\tau^{-1} \|q(x, \cdot) - q'(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})}},$$

as claimed. $\square$

Lemma B.7. *Let $\pi, \pi' \in \mathcal{P}(\mathsf{Y})$ for some measurable space Y be mutually absolutely continuous. Then*

$$\|\pi - \pi'\|_{\text{TV}} \leq \frac{1}{4} \left(\text{ess sup}_{\pi'} \frac{d\pi}{d\pi'} - \text{ess inf}_{\pi'} \frac{d\pi}{d\pi'} \right).$$

Proof. Define $h := \frac{d\pi}{d\pi'}$, and write $M := \text{ess sup}_{\pi'} h, m := \text{ess inf}_{\pi'} h$. Note that

$$0 = \int_{\mathsf{Y}} (d\pi - d\pi') = \int_{\mathsf{Y}} (h - 1) d\pi' = \int_{\mathsf{E}} (h - 1) d\pi' + \int_{\mathsf{Y} \setminus \mathsf{E}} (h - 1) d\pi',$$

for any measurable $\mathsf{E} \subset \mathsf{Y}$. Consequently, we have

$$\int_{\mathsf{E}} (h - 1) d\pi' = \int_{\mathsf{Y} \setminus \mathsf{E}} (1 - h) d\pi'.$$

Now, we derive the following upper bounds,

$$\begin{aligned} \pi(\mathsf{E}) - \pi'(\mathsf{E}) &= \int_{\mathsf{E}} (h - 1) d\pi' \leq (M - 1) \pi'(\mathsf{E}) \\ \pi(\mathsf{E}) - \pi'(\mathsf{E}) &= \int_{\mathsf{Y} \setminus \mathsf{E}} (1 - h) d\pi' \leq (1 - m) \pi'(\mathsf{Y} \setminus \mathsf{E}). \end{aligned}$$

Multiplying these inequalities by $\pi'(\mathsf{Y} \setminus \mathsf{E})$ and $\pi'(\mathsf{E})$, respectively, and adding the results, we have

$$\begin{aligned} (\pi(\mathsf{E}) - \pi'(\mathsf{E}))(\pi'(\mathsf{Y} \setminus \mathsf{E}) + \pi'(\mathsf{E})) &\leq ((M - 1) + (1 - m)) \pi'(\mathsf{E}) \pi'(\mathsf{Y} \setminus \mathsf{E}) \\ \therefore \pi(\mathsf{E}) - \pi'(\mathsf{E}) &\leq (M - m) \pi'(\mathsf{E}) \pi'(\mathsf{Y} \setminus \mathsf{E}). \end{aligned}$$

In fact, the same bound can be achieved for $\pi'(E) - \pi(E)$; to see this, note that

$$\begin{aligned}\pi'(E) - \pi(E) &= \int_E (1 - h) d\pi' \leq (1 - m)\pi'(E) \\ \pi'(E) - \pi(E) &= \int_{Y \setminus E} (h - 1) d\pi' \leq (M - 1)\pi'(E),\end{aligned}$$

so by the same procedure as above, $\pi'(E) - \pi(E) \leq (M - m)\pi'(E)\pi'(Y \setminus E)$. Therefore, we have shown that

$$|\pi(E) - \pi'(E)| \leq (M - m)\pi'(E)\pi'(Y \setminus E)$$

for any measurable $E \subset Y$. Since $\pi'(E)\pi'(Y \setminus E)$ is maximized at $\pi'(E) = \pi'(Y \setminus E) = 1/2$, we have

$$\|\pi - \pi'\|_{\text{TV}} = \sup_E |\pi(E) - \pi'(E)| \leq \frac{1}{4}(M - m),$$

as claimed. $\square$

Lemma B.8. *Let $u, w \in L^\infty(Y)$ for some measurable space Y , and let λ be a measure on Y . Define $\pi^u, \pi^w \in \mathcal{P}(Y)$ absolutely continuous with respect to λ such that $\frac{d\pi^\bullet}{d\lambda} \propto e^{-\bullet}$ for $\bullet \in \{u, w\}$. Then*

$$\|\pi^u - \pi^w\|_{\text{TV}} \leq \frac{1}{2} \sinh(2\|u - w\|_{L^\infty(\lambda)}).$$

Proof. Firstly, since $u, w \in L^\infty(Y)$, it follows that π^u, π^w are mutually absolutely continuous. Now, define $h := \frac{d\pi^u}{d\pi^w}$, with $M := \text{ess sup}_\lambda h$ and $m := \text{ess inf}_\lambda h$. Note that

$$d\pi^u(x) = \frac{e^{-u(x)}}{Z_u} d\lambda(x), \quad d\pi^w(x) = \frac{e^{-w(x)}}{Z_w} d\lambda(x),$$

where $Z_u, Z_w \in \mathbb{R}$ are normalizing constants. Defining $f := u - w$, we have

$$h(x) = \frac{Z_w}{Z_u} e^{-f(x)}.$$

Additionally, we have

$$\frac{Z_w}{Z_u} = \frac{\int_Y e^{-w(x)} d\lambda(x)}{\int_Y e^{-w(x)} e^{-f(x)} d\lambda(x)} = \frac{1}{\mathbb{E}_{\pi^w}[e^{-f}]}.$$

Consequently, it holds that $\text{ess inf}_\lambda h \geq e^{\text{ess inf}_\lambda f - \text{ess sup}_\lambda f}$ and $\text{ess sup}_\lambda h \leq e^{\text{ess sup}_\lambda f - \text{ess inf}_\lambda f}$. So, by the definition of f , we have $m \geq e^{-2\|u - w\|_{L^\infty(\lambda)}}$ and $M \leq e^{2\|u - w\|_{L^\infty(\lambda)}}$. Then, invoking Lemma B.7, we have

$$\|\pi^u - \pi^w\|_{\text{TV}} \leq \frac{1}{4} \left(e^{2\|u - w\|_{L^\infty(\lambda)}} - e^{-2\|u - w\|_{L^\infty(\lambda)}} \right) = \frac{1}{2} \sinh(2\|u - w\|_{L^\infty(\lambda)}).$$

$\square$

Lemma B.9. *Let $q, q' \in L^\infty(X \times A)$. Then for any $\tau > 0$ and any $x \in X$,*

$$\|(\mathcal{G}_\tau q)_x - (\mathcal{G}_\tau q')_x\|_{\text{TV}} \leq \frac{1}{2} \sinh(2\tau^{-1}\|q(x, \cdot) - q'(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})}).$$

Proof. Note that, for any $q \in M_b(X \times A)$, we have

$$d(\mathcal{G}_\tau q)_x \propto \exp(\tau^{-1}q(x, \cdot)) d\pi_x^{\text{ref}}.$$

So, invoking Lemma B.8 with $u = -\tau^{-1}q(x, \cdot)$, $v = -\tau^{-1}q'(x, \cdot)$, and $\lambda = \pi_x^{\text{ref}}$, we have

$$\|(\mathcal{G}_\tau q)_x - (\mathcal{G}_\tau q')_x\|_{\text{TV}} \leq \frac{1}{2} \sinh(2\tau^{-1}\|q - q'\|_{L^\infty(\pi_x^{\text{ref}})}).$$

$\square$

Lemma B.10. For every $\tau > 0$, recalling the notation [above](#),

$$0 \leq q_{\text{ref}}^* - q_\tau^* \leq \frac{\gamma}{1-\gamma} M_\tau(q_\tau^*).$$

If Assumption [3.4](#) is satisfied, then

$$M_\tau(q_\tau^*) \leq -\tau \log p_{\text{ref}},$$

and q_τ^* converges uniformly up to q_{ref}^* .

Proof. By Lemma [B.4](#), we have that for any $q \in L^\infty(\mathbf{X} \times \mathbf{A})$,

$$q_{\text{ref}}^* - q_\tau^* = \lim_{n \rightarrow \infty} (\mathcal{B}_{\text{ref}}^*)^n q - \lim_{n \rightarrow \infty} \mathcal{B}_\tau^{*n} q \geq 0.$$

Then, by Lemma [B.5](#), we have

$$\begin{aligned} q_{\text{ref}}^* - q_\tau^* &= \lim_{n \rightarrow \infty} ((\mathcal{B}_{\text{ref}}^*)^n q_\tau^* - \mathcal{B}_\tau^{*n} q_\tau^*) \\ &\leq \lim_{n \rightarrow \infty} M_\tau(q_\tau^*) \sum_{k=1}^n \gamma^k \\ &= \frac{\gamma}{1-\gamma} M_\tau(q_\tau^*), \end{aligned}$$

proving the first claim. When Assumption [3.4](#) is satisfied, we have $M_\tau(q_\tau^*) \leq -\tau \log p_{\text{ref}}$ by Lemma [B.3](#), so that $M_\tau(q_\tau^*)$ converges down to 0, and consequently q_τ^* converges up to q^* . $\square$

C Proofs from Section 4

Theorem 4.2. If $r \in M_b(\mathbf{X} \times \mathbf{A})$, $\gamma < 1$, and $\pi \in \mathcal{K}(\mathbf{X}, \mathcal{P}(\mathbf{A}))$ is such that

$$\sup_{x,a} \|\tau \mathbf{kl}[\pi]\|_{L^p(P_{x,a})} < \infty, \quad (4.1)$$

the soft distributional Bellman operator $\mathcal{T}_\tau^\pi$ is a γ -contraction in $\bar{d}_p$ for every $\tau \geq 0$. Thus, it has a unique solution to the fixed point equation $\bar{\zeta} = \mathcal{T}_\tau^\pi \bar{\zeta}$, which we denote by $\bar{\zeta}^{\pi,\tau}$. [\[Source\]](#)

Proof. To begin, let us show that $\mathcal{T}_\tau^\pi$ maps elements of $\bar{K}^p(\mathbf{X} \times \mathbf{A}, \mathcal{P}(\mathbb{R}))$ to $\bar{K}^p(\mathbf{X} \times \mathbf{A}, \mathcal{P}(\mathbb{R}))$. For any $\zeta \in \bar{K}^p(\mathbf{X} \times \mathbf{A}, \mathcal{P}(\mathbb{R}))$, observe that

$$\begin{aligned} &\sup_{x,a} \left(\int |z|^p d(\mathcal{T}_\tau^\pi \zeta)_{x,a}(z) \right)^{1/p} \\ &= \sup_{x,a} \left(\int \left[|r(x,a) - \gamma \tau \mathbf{KL}(\pi_{x'} \| \pi_{x'}^{\text{ref}}) + \gamma z|^p d\zeta_{x',a'}(z) \right] d\check{P}_{x,a}^\pi(x',a') \right)^{1/p} \\ &\leq \|r\|_{\text{sup}} + \gamma \sup_{x,a} \|\tau \mathbf{kl}[\pi]\|_{L^p(P_{x,a})} + \gamma \left(\sup_{x',a'} \int |z|^p d\zeta_{x',a'}(z) \right)^{1/p} \\ &< \infty, \end{aligned}$$

by assumption, as desired.

Next, by the convexity of the Wasserstein metric [\[6, 40\]](#), we have

$$\begin{aligned} &d_p((\mathcal{T}_\tau^\pi \bar{\zeta})_{x,a}, (\mathcal{T}_\tau^\pi \bar{\zeta}')_{x,a}) \\ &\leq \int d_p \left((\mathbf{b}_{r(x,a) - \gamma \tau \mathbf{KL}(\pi_{x'} \| \pi_{x'}^{\text{ref}}), \gamma}) \# \bar{\zeta}_{x',a'}, (\mathbf{b}_{r(x,a) - \gamma \tau \mathbf{KL}(\pi_{x'} \| \pi_{x'}^{\text{ref}}), \gamma}) \# \bar{\zeta}'_{x',a'} \right) d\check{P}_{x,a}^\pi(x',a') \\ &\leq \gamma \int d_p(\bar{\zeta}_{x',a'}, \bar{\zeta}'_{x',a'}) d\check{P}_{x,a}^\pi(x',a') \\ &\leq \gamma \sup_{x',a'} d_p(\bar{\zeta}_{x',a'}, \bar{\zeta}'_{x',a'}) \\ &= \gamma \bar{d}_p(\bar{\zeta}, \bar{\zeta}'), \end{aligned}$$

where the second inequality holds since the common transformation $\mathbf{b}_{r(x,a) - \gamma\tau \text{KL}(\pi_{x'} \parallel \pi_{x'}^{\text{ref}}), \gamma}$ is affine. As a consequence, we have that

$$\bar{d}_p(\mathcal{T}_\tau^\pi \bar{\zeta}, \mathcal{T}_\tau^\pi \bar{\zeta}') \leq \gamma \bar{d}_p(\bar{\zeta}, \bar{\zeta}'),$$

which validates that $\mathcal{T}_\tau^\pi$ is a γ -contraction in $\bar{d}_p$. Consequently, since $(\bar{K}^p(\mathbf{X} \times \mathbf{A}, \mathcal{P}(\mathbb{R})), \bar{d}_p)$ is complete and separable [6], it follows that $\mathcal{T}_\tau^\pi$ has a unique fixed point. That $\bar{\zeta}^{\pi, \tau}$ coincides with this fixed point follows precisely by [6, Proposition 4.9]. $\square$

Lemma 4.4. For any $\tau > 0$, $\mathcal{Q}\mathcal{T}_\tau^\star = \mathcal{B}_\tau^\star \mathcal{Q}$.

[Source]

Proof. For any $\bar{\zeta} \in \bar{K}^p(\mathbf{X} \times \mathbf{A}, \mathcal{P}(\mathbb{R}))$, we have

$$\begin{aligned} (\mathcal{Q}\mathcal{T}_\tau^\star \bar{\zeta})(x, a) &= \int \int (r(x, a) - \gamma\tau \text{KL}((\mathcal{G}_\tau \mathcal{Q}\bar{\zeta})_{x'} \parallel \pi_{x'}^{\text{ref}}) + \gamma z) d\bar{\zeta}_{x', a'}(z) d\check{P}_{x, a}^{\mathcal{G}_\tau \mathcal{Q}\bar{\zeta}}(x', a') \\ &= \int \left(r(x, a) - \gamma\tau \text{KL}((\mathcal{G}_\tau \mathcal{Q}\bar{\zeta})_{x'} \parallel \pi_{x'}^{\text{ref}}) + \gamma \int z d\bar{\zeta}_{x', a'}(z) \right) d\check{P}_{x, a}^{\mathcal{G}_\tau \mathcal{Q}\bar{\zeta}}(x', a') \\ &= \int (r(x, a) - \gamma\tau \text{KL}((\mathcal{G}_\tau \mathcal{Q}\bar{\zeta})_{x'} \parallel \pi_{x'}^{\text{ref}}) + \gamma(\mathcal{Q}\bar{\zeta})(x', a')) d\check{P}_{x, a}^{\mathcal{G}_\tau \mathcal{Q}\bar{\zeta}}(x', a') \end{aligned}$$

Defining $q := \mathcal{Q}\bar{\zeta}$, this is equivalent to

$$(\mathcal{Q}\mathcal{T}_\tau^\star \bar{\zeta})(x, a) = \int (r(x, a) - \gamma\tau \text{KL}((\mathcal{G}_\tau q)_{x'} \parallel \pi_{x'}^{\text{ref}}) + \gamma q(x', a')) d\check{P}_{x, a}^{\mathcal{G}_\tau q}(x', a')$$

Moreover, note that

$$\begin{aligned} \text{KL}((\mathcal{G}_\tau q)_x \parallel \pi_x^{\text{ref}}) &= \int \log \frac{d(\mathcal{G}_\tau q)_x}{d\pi_x^{\text{ref}}} d(\mathcal{G}_\tau q)_x \\ &= \tau^{-1} \int (q(x, a) - \mathcal{V}_\tau q(x)) d(\mathcal{G}_\tau q)_x(a) \\ &= \tau^{-1} \left(\int q(x, a) d(\mathcal{G}_\tau q)_x(a) - \mathcal{V}_\tau q(x) \right). \end{aligned}$$

Substituting, we have shown that

$$\begin{aligned} (\mathcal{Q}\mathcal{T}_\tau^\star \bar{\zeta})(x, a) &= \int \left(r(x, a) - \gamma \int q(x', a'') d\check{P}_{x, a}^{\mathcal{G}_\tau q} + \gamma \mathcal{V}_q^\tau(x') + \gamma q(x, a) \right) d\check{P}_{x, a}^{\mathcal{G}_\tau q}(x', a') \\ &= \int (r(x, a) + \gamma \mathcal{V}_q^\tau(x')) d\check{P}_{x, a}^{\mathcal{G}_\tau q}(x', a') \\ &\equiv \mathcal{B}_\tau^\star q(x, a) \\ &= \mathcal{B}_\tau^\star \mathcal{Q}\bar{\zeta}(x, a). \end{aligned}$$

$\square$

Theorem 4.5. For any $\bar{\zeta} \in \bar{K}^p(\mathbf{X} \times \mathbf{A}, \mathcal{P}(\mathbb{R}))$ and temperature $\tau > 0$ define the iterates $(\bar{\zeta}^n)_{n \in \mathbb{N}}$ given by $\bar{\zeta}^{n+1} = \mathcal{T}_\tau^\star \bar{\zeta}^n$ for $\bar{\zeta}^0 = \mathcal{T}_\tau^\star \bar{\zeta}$. Then, for $\bar{\zeta}^{\tau, \star} := \bar{\zeta}^{\tau, \pi^{\tau, \star}}$,

$$\bar{d}_p(\bar{\zeta}^n, \bar{\zeta}^{\tau, \star}) \leq C_{p, \tau, \gamma} n \gamma^{n/p} \bar{d}_p(\bar{\zeta}^0, \bar{\zeta}^{\tau, \star}) \quad \text{and} \quad \bar{d}_1(\bar{\zeta}^n, \bar{\zeta}^{\tau, \star}) \leq \frac{1}{(1 - \gamma)\sqrt{\tau}} C_n \gamma^n \bar{d}_1(\bar{\zeta}^0, \bar{\zeta}^{\tau, \star}),$$

where $C, C_{p, \tau, \gamma} < \infty$ are constants depending on $\|r\|_{\text{sup}}$, $(p, \tau, \gamma, \|r\|_{\text{sup}})$ respectively. [Source]

Proof. We begin by defining some helper notation. For any $\bar{\zeta} \in \bar{K}^p(X \times A, \mathcal{P}(\mathbb{R}))$, we define $\xi^{\bar{\zeta}} \in \bar{K}^p(X, \mathcal{P}(\mathbb{R} \times A))$ where

$$\xi_x^{\bar{\zeta}} := (\mathbf{b}_{-\tau \text{KL}((\mathcal{G}_\tau \Omega \bar{\zeta})_x \parallel \pi_x^{\text{ref}}), 1} \circ \text{proj}^{\mathbb{R}})_{\#}(\bar{\zeta}_{x,-} \otimes (\mathcal{G}_\tau \Omega \bar{\zeta})_x). \quad (\text{C.1})$$

In turn,

$$(\mathcal{I}_\tau^* \bar{\zeta})_{x,a} = (\mathbf{b}_{r(x,a), \gamma} \circ \text{proj}^{\mathbb{R}})_{\#}(\xi_x^{\bar{\zeta}} \otimes P_{x,a}). \quad (\text{C.2})$$

Next, we define the following helpers,

$$\pi^n := \mathcal{G}_\tau \Omega \bar{\zeta}^n, \quad \xi^n := \xi^{\bar{\zeta}^n}, \quad \xi^* := \xi^{\bar{\zeta}^{\tau,*}}.$$

By [40, Theorem 4.8], we have that for any $(x, a) \in X \times A$,

$$d_p(\bar{\zeta}_{x,a}^{n+1}, \bar{\zeta}^{\tau,*}) \leq \gamma \int d_p(\xi_{x'}^n, \xi_{x'}^*) dP_{x,a}(x'). \quad (\text{C.3})$$

Invoking the triangle inequality together with the expansion of the ξ terms by definition, we have that for any $x \in X$,

$$\begin{aligned} d_p(\xi_x^n, \xi_x^*) &= d_p((\text{proj}^{\mathbb{R}} - \tau \text{KL}(\pi_x^n \parallel \pi_x^{\text{ref}}))_{\#}(\bar{\zeta}_{x,-}^n \otimes \pi_x^n), (\text{proj}^{\mathbb{R}} - \tau \text{KL}(\pi_x^{\tau,*} \parallel \pi_x^{\text{ref}}))_{\#}(\bar{\zeta}_{x,-}^{\tau,*} \otimes \pi_x^{\tau,*})) \\ &\leq \overbrace{d_p((\text{proj}^{\mathbb{R}} - \tau \text{KL}(\pi_x^n \parallel \pi_x^{\text{ref}}))_{\#}(\bar{\zeta}_{x,-}^n \otimes \pi_x^n), (\text{proj}^{\mathbb{R}} - \tau \text{KL}(\pi_x^n \parallel \pi_x^{\text{ref}}))_{\#}(\bar{\zeta}_{x,-}^{\tau,*} \otimes \pi_x^n))}^{I_n} \\ &\quad + \overbrace{d_p((\text{proj}^{\mathbb{R}} - \tau \text{KL}(\pi_x^n \parallel \pi_x^{\text{ref}}))_{\#}(\bar{\zeta}_{x,-}^{\tau,*} \otimes \pi_x^n), (\text{proj}^{\mathbb{R}} - \tau \text{KL}(\pi_x^{\tau,*} \parallel \pi_x^{\text{ref}}))_{\#}(\bar{\zeta}_{x,-}^{\tau,*} \otimes \pi_x^{\tau,*}))}^{II_n}. \end{aligned}$$

Since the measures being compared in I_n are both translated by the same pushforward map, another application of [40, Theorem 4.8] yields the following inequality:

$$I_n \leq \int d_p(\bar{\zeta}_{x,a}^n, \bar{\zeta}_{x,a}^{\tau,*}) d\pi_x^n(a) \leq \bar{d}_p(\bar{\zeta}^n, \bar{\zeta}^{\tau,*}).$$

Next, we bound II_n . Let $\mathcal{C}(\rho_1, \rho_2)$ be the set of couplings between measures ρ_1, ρ_2 . Then

$$\begin{aligned} II_n &\leq \inf_{\kappa \in \mathcal{C}(\bar{\zeta}_{x,-}^{\tau,*} \otimes \pi_x^n, \bar{\zeta}_{x,-}^{\tau,*} \otimes \pi_x^{\tau,*})} \left(\int \left| \mathbf{b}_{-\tau \text{KL}(\pi_x^n \parallel \pi_x^{\text{ref}}), 1}(z) - \mathbf{b}_{-\tau \text{KL}(\pi_x^{\tau,*} \parallel \pi_x^{\text{ref}}), 1}(z') \right|^p d\kappa \right)^{1/p} \\ &\stackrel{(a)}{\leq} \inf_{\kappa \in \mathcal{C}(\bar{\zeta}_{x,-}^{\tau,*} \otimes \pi_x^n, \bar{\zeta}_{x,-}^{\tau,*} \otimes \pi_x^{\tau,*})} \left(\int |z - z'|^p d\kappa \right)^{1/p} + \tau |\text{KL}(\pi_x^n \parallel \pi_x^{\text{ref}}) - \text{KL}(\pi_x^{\tau,*} \parallel \pi_x^{\text{ref}})| \\ &\stackrel{(b)}{\leq} C \gamma^{n/p} \|\Omega \bar{\zeta}^0 - q_\tau^*\|_{\text{sup}}^{1/p}, \end{aligned}$$

for some constant C depending on $\tau, p, \gamma, \|r\|_{\text{sup}}$ where (a) applies Minkowski's inequality, noting that the KL terms are independent of κ , and (b) invokes Lemma C.5 and Lemma C.6. Indeed, for n large enough, Lemmas C.5 and C.6 assert that $C \lesssim \tau^{-1}$ for fixed p , and more generally that $C \lesssim \tau^{-1/2}$ for any n (and fixed p). Substituting back into (C.3), we see that

$$\bar{d}_p(\bar{\zeta}^{n+1}, \bar{\zeta}^{\tau,*}) \leq \gamma \bar{d}_p(\bar{\zeta}^n, \bar{\zeta}^{\tau,*}) + C \gamma^{1+n/p} \|\Omega \bar{\zeta}^0 - q_\tau^*\|_{\text{sup}}^{1/p}.$$

Let $a_n := \bar{d}_p(\bar{\zeta}^n, \bar{\zeta}^{\tau,*})$. We have shown that $a_{n+1} \leq \gamma a_n + C' \gamma^{1+n/p}$, where $C' = C \|\Omega \bar{\zeta}^0 - q_\tau^*\|_{\text{sup}}^{1/p}$ is a constant depending on p and τ . We will apply techniques of *generatingfunctionology* [41] to bound this sequence. We define $A : \mathbb{R} \rightarrow \mathbb{R}$ as the *formal power series* given by

$$A(y) = \sum_{n=0}^{\infty} a_n y^n,$$

and we will pick off the coefficients a_n from the power series representation of A . Our recurrence above, upon multiplying through by y^n and summing over n yields

$$\begin{aligned} \sum_{n=0}^{\infty} a_{n+1} y^n &\leq \gamma \sum_{n=0}^{\infty} a_n y^n + C' \gamma \sum_{n=0}^{\infty} \gamma^{n/p} y^n \\ \therefore \frac{1}{y} A(y) - a_0 &\leq \gamma A(y) + C' \gamma \frac{1}{1 - \gamma^{1/p} y} \\ \therefore A(y) &\leq \frac{a_0 y}{1 - \gamma y} + \frac{C' \gamma y}{(1 - \gamma^{1/p} y)(1 - \gamma y)}, \end{aligned}$$

where $a_0 = \bar{d}_p(\bar{\zeta}^0, \bar{\zeta}^{\tau,*})$. Now, the formal power series expansion gives

$$A(y) \leq \begin{cases} \sum_{n=1}^{\infty} \left[a_0 \gamma^n + C' \gamma \frac{\gamma^{n/p} - \gamma^n}{\gamma^{1/p} - \gamma} \right] y^n & p \neq 1 \\ \sum_{n=1}^{\infty} [a_0 \gamma^n + C' n \gamma^n] y^n & p = 1. \end{cases}$$

Combining, we have

$$\bar{d}_p(\bar{\zeta}^n, \bar{\zeta}^{\tau,*}) = a_n \leq (1 + \bar{d}_p(\bar{\zeta}^0, \bar{\zeta}^{\tau,*})) C'' n \gamma^{n/p}$$

where $C'' = C'$ when $p = 1$, and $C'' = C' / (\gamma^{1/p} - \gamma)$ otherwise—in any case, C'' is a constant depending only on p, τ, γ , and the proof is complete. $\square$

Theorem 4.6. *Suppose Assumption 3.4 holds. Let $p, p' \in [1, \infty)$ and $\omega \in \mathcal{P}(\mathbf{X} \times \mathbf{A})$. For any $\epsilon, \delta > 0$, there exists a $\tau > 0$ for which $d_{p;p',\omega}(\bar{\zeta}^{\tau,\pi^{\tau,*}}, \zeta^{\pi^{\tau,*}}) \leq \delta/2$ and $q^{\pi^{\tau,*}}$ is $\epsilon/2$ -reference-optimal. In turn, an $n_{\epsilon,\delta} = n_{\epsilon,\delta}(\tau) \in \mathbb{N}$ exists for which*

$$d_{p;p',\omega}(\bar{\zeta}^n, \zeta^{\pi^{\tau,*}}) \leq \delta \quad \text{and} \quad \mathcal{G}_{\tau} \mathcal{Q} \bar{\zeta}^n \text{ is } \epsilon\text{-reference-optimal} \quad \forall n \geq n_{\epsilon,\delta}$$

where $\bar{\zeta}^{n+1} = \mathcal{T}_{\tau}^* \bar{\zeta}^n$ and $\bar{\zeta}^0 = \mathcal{T}_{\tau}^* \bar{\zeta}$ for any $\bar{\zeta} \in \bar{\mathcal{K}}^p(\mathbf{X} \times \mathbf{A}, \mathcal{P}(\mathbb{R}))$. [Source]

Proof. By Lemma B.10 and under Assumption 3.4,

$$\|q_{\tau}^* - q_{\text{ref}}^*\|_{\text{sup}} \leq \frac{\gamma \log p_{\text{ref}}^{-1}}{1 - \gamma} \tau \leq \frac{\epsilon}{2},$$

choosing $\tau \leq \tau_{\epsilon} := \frac{\epsilon(1-\gamma)}{2\gamma \log p_{\text{ref}}^{-1}}$. Hence, by Lemmas 4.4 and A.7

$$\|\mathcal{Q} \bar{\zeta}^{n_{\epsilon}} - q_{\text{ref}}^*\|_{\text{sup}} \leq \gamma^{n_{\epsilon}} \|\mathcal{Q} \bar{\zeta}^0 - q_{\tau}^*\|_{\text{sup}} + \|q_{\tau}^* - q_{\text{ref}}^*\|_{\text{sup}} \leq \gamma^{n_{\epsilon}} \|\mathcal{Q} \bar{\zeta}^0 - q_{\text{ref}}^*\|_{\text{sup}} + \frac{\epsilon}{2} \leq \epsilon,$$

which holds when $n_{\epsilon} \geq (\log \gamma)^{-1} \log \frac{\epsilon}{2\|\mathcal{Q} \bar{\zeta}^0 - q_{\text{ref}}^*\|_{\text{sup}}}$.

Next, we will show that the soft return distribution estimates will approximate $\zeta^{\pi^{\tau,*}}$. For notational simplicity, define $X_t := X_t^{\pi^{\tau,*}}$ and $A_t := A_t^{\pi^{\tau,*}}$ for $t \in \mathbb{N}$. Recall that

$$\bar{\zeta}_{x,a}^{\tau,\pi^{\tau,*}} = \text{law} \left(r(x, a) + \sum_{t \geq 1} \gamma^t (r(X_t, A_t) - \tau \text{KL}(\pi_{X_t}^{\tau,*} \parallel \pi_{X_t}^{\text{ref}})) \mid X_0 = x, A_0 = a \right).$$

Moreover, we define $\tilde{\zeta}_{x,a}^{\tau,\pi^{\tau,*}} := (-\tau \text{KL}(\pi_{x,a}^{\tau,*} \parallel \pi_{x,a}^{\text{ref}}) + \text{id})_{\#} \bar{\zeta}_{x,a}^{\tau,\pi^{\tau,*}}$, so that

$$\tilde{\zeta}^{\tau,\pi^{\tau,*}} = \text{law} \left(\sum_{t \geq 0} \gamma^t (r(X_t, A_t) - \tau \text{KL}(\pi_{X_t}^{\tau,*} \parallel \pi_{X_t}^{\text{ref}})) \mid X_0 = x, A_0 = a \right).$$

Now, by the triangle inequality, we have

$$d_{p;p',\omega}^p(\bar{\zeta}^{\tau,\pi^{\tau,*}}, \zeta^{\pi^{\tau,*}}) \leq 2^{p'-1} \int \left[\underbrace{d_p^{p'}(\bar{\zeta}_{x,a}^{\tau,\pi^{\tau,*}}, \tilde{\zeta}_{x,a}^{\tau,\pi^{\tau,*}})}_{\text{I}_{\tau}(x,a)} + \underbrace{d_p^{p'}(\tilde{\zeta}_{x,a}^{\tau,\pi^{\tau,*}}, \zeta_{x,a}^{\pi^{\tau,*}})}_{\text{II}_{\tau}(x,a)} \right] d\omega(x, a). \quad (\text{C.4})$$

We proceed by analyzing I_τ . By coupling states and actions, we immediately have

$$I_\tau(x, a) \leq (\tau \text{KL}(\pi_x^{\tau,*} \parallel \pi_x^{\text{ref}}))^{p'},$$

and so, since $\pi^{\tau,*} = \mathcal{G}_\tau q_\tau^*$, by virtue of Lemma B.2 we have

$$\limsup_{\tau \rightarrow 0} I_\tau(x, a) = 0.$$

Next, we bound Π_τ . Denote by $r_{\pi,\tau} : \mathbf{X} \times \mathbf{A} \rightarrow \mathbb{R}$ the reward function defined by

$$r_{\pi,\tau}(x, a) = r(x, a) - \tau \text{KL}(\pi_x^{\tau,*} \parallel \pi_x^{\text{ref}}).$$

The work of [44] shows that, for any policy π , there is a unique $\mathbb{T}^\pi \in \mathcal{K}(\mathbf{X} \times \mathbf{A}, \mathcal{P}(\mathcal{P}(\mathbf{X} \times \mathbf{A})))$ for which $(\mu \mapsto (1 - \gamma)^{-1}(\mu r)(x, a))_\# \mathbb{T}_{x,a}^\pi = \zeta_{x,a}^{\pi,r}$, where $\zeta^{\pi,r}$ denotes the return distribution function associated to the policy π for the reward function r . Noting that $\tilde{\zeta}^{\tau,\pi^{\tau,*}} = \zeta^{\pi^{\tau,*},r_{\pi,\tau}}$, we have

$$\begin{aligned} \Pi_\tau(x, a) &= d_p^{p'} \left(\left(\mu \mapsto \frac{1}{1 - \gamma} (\mu r_{\pi,\tau})(x, a) \right)_\# \mathbb{T}_{x,a}^{\pi^{\tau,*}}, \left(\mu \mapsto \frac{1}{1 - \gamma} (\mu r)(x, a) \right)_\# \mathbb{T}_{x,a}^{\pi^{\tau,*}} \right) \\ &\leq \frac{1}{1 - \gamma} \left(\int \left[\int |r_{\pi,\tau}(x', a') - r(x', a')|^p d\mu(x', a') \right] d\mathbb{T}_{x,a}^{\pi^{\tau,*}}(\mu) \right)^{p'/p} \\ &= \frac{1}{1 - \gamma} \left(\int \left[\int \tau \text{KL}(\pi_{x'}^{\tau,*} \parallel \pi_{x'}^{\text{ref}})^p d\mu(x', a') \right] d\mathbb{T}_{x,a}^{\pi^{\tau,*}}(\mu) \right)^{p'/p} \end{aligned}$$

where the penultimate step is simply a coupling argument (coupling the samples of $\mathbb{T}^{\pi^{\tau,*}}$). Once again, since $\limsup_{\tau \rightarrow 0} \tau \text{KL}(\pi_x^{\tau,*} \parallel \pi_x^{\text{ref}}) = 0$, and $\text{KL}(\pi_x^{\tau,*} \parallel \pi_x^{\text{ref}})$ is bounded by Lemma B.2, the dominated convergence theorem asserts that $\lim_{\tau \rightarrow 0} \Pi_\tau(x, a) = 0$ pointwise.

Altogether, we have shown that $\lim_{\tau \rightarrow 0} (I_\tau(x, a) + \Pi_\tau(x, a)) = 0$ pointwise, and is bounded as a consequence of Lemma B.2. Thus, by another application of the dominated convergence theorem together with (C.4), we have that

$$\lim_{\tau \rightarrow 0} d_{p;p',\omega}(\tilde{\zeta}^{\tau,\pi^{\tau,*}}, \zeta^{\pi^{\tau,*}}) = 0.$$

It follows that there exists some $\tau_\delta > 0$ for which $d_{p;p',\omega}(\tilde{\zeta}^{\tau,\pi^{\tau,*}}, \zeta^{\pi^{\tau,*}}) \leq \delta/2$ whenever $\tau \leq \tau_\delta$. For any such τ , by Theorem 4.5, there exists $n_\delta \in \mathbb{N}$ for which

$$d_{p;p',\omega}(\tilde{\zeta}^{n_\delta}, \tilde{\zeta}^{\tau,\pi^{\tau,*}}) \leq \bar{d}_p(\tilde{\zeta}^{n_\delta}, \tilde{\zeta}^{\tau,\pi^{\tau,*}}) \leq \frac{\delta}{2}.$$

For this choice of τ and n_δ , by the triangle inequality,

$$d_{p;p',\omega}(\tilde{\zeta}^{n_\delta}, \zeta^\pi) \leq \delta.$$

Altogether, taking $\tau = \min\{\tau_\epsilon, \tau_\delta\}$ and $n = \max\{n_\epsilon, n_\delta\}$, we have that

$$d_{p;p',\omega}(\tilde{\zeta}^{\tau,\pi^{\tau,*}}, \zeta^\pi) \leq \frac{\delta}{2} \quad \text{and} \quad \|q^{\pi^{\tau,*}} - q_{\text{ref}}^*\|_{\text{sup}} \leq \frac{\epsilon}{2},$$

as well as

$$d_{p;p',\omega}(\tilde{\zeta}^n, \zeta^\pi) \leq \delta \quad \text{and} \quad \|\mathcal{Q}\tilde{\zeta}^n - q_{\text{ref}}^*\|_{\text{sup}} \leq \epsilon.$$

To complete the proof, we note that

$$q^{\mathcal{G}_\tau \mathcal{Q}\tilde{\zeta}^n} \geq q^{\mathcal{G}_\tau \mathcal{Q}\zeta^n} \geq q_{\text{ref}}^* - \epsilon,$$

so that $\mathcal{G}_\tau \mathcal{Q}\tilde{\zeta}^n$ is ϵ -reference-optimal. $\square$

Theorem 4.7. Suppose Assumption 3.4 holds and $\mathbf{A}$ is discrete. Let $p, p' \in [1, \infty)$ and $\omega \in \mathcal{P}(\mathbf{X} \times \mathbf{A})$. For any $\epsilon, \delta > 0$ and $\tilde{\zeta}^0 \in \overline{\mathcal{K}}^p(\mathbf{X} \times \mathbf{A}, \mathcal{P}(\mathbb{R}))$, there exists $\tau > 0$, a decoupled $\sigma_\tau > 0$ and $n_{\text{opt}}, n_{\text{eval}} \in \mathbb{N}$ such that

$$d_{p;p',\omega}(\hat{\zeta}^{n_{\text{eval}}}, \zeta^{\pi^{\text{ref}},*}) \leq \delta \quad \text{and} \quad \mathcal{G}_\tau \mathcal{Q}\hat{\zeta}^{n_{\text{eval}}} \text{ is } \epsilon\text{-reference-optimal}$$

where $\tilde{\zeta}^{n+1} = \mathcal{T}_\sigma^* \tilde{\zeta}^n$, $\hat{\pi}^{\tau,\sigma} = \mathcal{G}_\tau \tilde{\zeta}^{n_{\text{opt}}}$, and $\hat{\zeta}^{n+1} = \mathcal{T}_\tau^{\hat{\pi}^{\tau,\sigma}} \hat{\zeta}^n$, for $\hat{\zeta}^0 = \tilde{\zeta}^{n_{\text{opt}}}$.

[Source]

Proof. Appealing to Theorem 3.10, for any $\delta > 0$, any temperature decoupling gambit yields a $\tau_\delta > 0$ and an associated decoupled temperature $\sigma_\delta = \sigma(\tau_\delta) > 0$ such that

$$d_{p;p',\omega}(\zeta^{\tau,\sigma}, \zeta^\star) \leq \delta/3$$

whenever $\tau \leq \tau_\delta$. Moreover, as shown in the proof of Theorem 4.6, for small enough τ'_δ ,

$$d_{p;p',\omega}(\zeta^{\tau,\sigma}, \bar{\zeta}^{\tau,\sigma}) \leq \delta/3$$

whenever $\tau \leq \tau'_\delta$ —here, we recall that $\bar{\zeta}^{\tau,\sigma}$ is the *entropy-regularized* return distribution function for the decoupled policy $\pi^{\tau,\sigma}$.

Now, define $\hat{\zeta}^{\tau,\sigma} = (\mathcal{T}_\tau^{\hat{\pi}^{\tau,\sigma}})^{n_{\text{eval}}} \hat{\zeta}^{\sigma,\star}$, and $\hat{\zeta}^{\sigma,\star} = (\mathcal{T}_\tau^\star)^{n_{\text{opt}}} \bar{\zeta}^0$. By the triangle inequality, we have

$$\begin{aligned} d_{p;p',\omega}(\hat{\zeta}^{\tau,\sigma}, \bar{\zeta}^{\tau,\sigma}) &\leq d_{p;p',\omega}((\mathcal{T}_\tau^{\hat{\pi}^{\tau,\sigma}})^{n_{\text{eval}}} \hat{\zeta}^{\sigma,\star}, (\mathcal{T}_\tau^{\pi^{\tau,\sigma}})^{n_{\text{eval}}} \hat{\zeta}^{\sigma,\star}) \\ &\quad + d_{p;p',\omega}((\mathcal{T}_\tau^{\pi^{\tau,\sigma}})^{n_{\text{eval}}} \hat{\zeta}^{\sigma,\star}, \bar{\zeta}^{\tau,\sigma}) \\ &\stackrel{(a)}{\leq} d_{p;p',\omega}((\mathcal{T}_\tau^{\hat{\pi}^{\tau,\sigma}})^{n_{\text{eval}}} \hat{\zeta}^{\sigma,\star}, (\mathcal{T}_\tau^{\pi^{\tau,\sigma}})^{n_{\text{eval}}} \hat{\zeta}^{\sigma,\star}) \\ &\quad + d_{p;p',\omega}((\mathcal{T}_\tau^{\pi^{\tau,\sigma}})^{n_{\text{eval}}} \hat{\zeta}^{\sigma,\star}, (\mathcal{T}_\tau^{\pi^{\tau,\sigma}})^{n_{\text{eval}}} \bar{\zeta}^{\tau,\sigma}) \\ &\stackrel{(b)}{\leq} d_{p;p',\omega}((\mathcal{T}_\tau^{\hat{\pi}^{\tau,\sigma}})^{n_{\text{eval}}} \hat{\zeta}^{\sigma,\star}, (\mathcal{T}_\tau^{\pi^{\tau,\sigma}})^{n_{\text{eval}}} \hat{\zeta}^{\sigma,\star}) \\ &\quad + \gamma^{n_{\text{eval}}} \bar{d}_p(\hat{\zeta}^{\tau,\sigma}, \bar{\zeta}^{\tau,\sigma}) \\ &\stackrel{(c)}{\lesssim} \gamma^{n_{\text{opt}}/2p} + \gamma^{n_{\text{eval}}}. \end{aligned}$$

Here, (a) leverages the fact that $\bar{\zeta}^{\tau,\sigma}$ is the fixed point of $\mathcal{T}_\tau^{\pi^{\tau,\sigma}}$ by definition, (b) invokes the contractivity of $\mathcal{T}_\tau^{\pi^{\tau,\sigma}}$ shown in Theorem 4.2 appealing to the fact that $\pi^{\tau,\sigma}$ is a BG policy for reference π^{ref} , and (c) follows by Lemma C.1. As a consequence, again since $|\gamma| < 1$, for sufficiently large $n_{\text{opt}}, n_{\text{eval}} \in \mathbb{N}$, we have

$$d_{p;p',\omega}(\hat{\zeta}^{\tau,\sigma}, \bar{\zeta}^{\tau,\sigma}) \leq \delta/3.$$

Altogether, by the triangle inequality once again, for the choices of $n_{\text{opt}}, n_{\text{eval}}, \tau, \sigma$ above,

$$\begin{aligned} d_{p;p',\omega}(\hat{\zeta}^{\tau,\sigma}, \zeta^\star) &\leq d_{p;p',\omega}(\hat{\zeta}^{\tau,\sigma}, \bar{\zeta}^{\tau,\sigma}) + d_{p;p',\omega}(\bar{\zeta}^{\tau,\sigma}, \zeta^{\tau,\sigma}) + d_{p;p',\omega}(\zeta^{\tau,\sigma}, \zeta^\star) \\ &\leq \frac{\delta}{3} + \frac{\delta}{3} + \frac{\delta}{3} \\ &= \delta. \end{aligned}$$

This completes the proof of the first claim. It remains to show now that $\mathcal{G}_\tau Q^{\hat{\zeta}^{n_{\text{eval}}}}$ is ϵ -reference-optimal. Towards this end, we note that by Theorem 4.6 and 4.7 that there exists $\tau_\epsilon > 0$, $n_\epsilon \in \mathbb{N}$ such that

$$\bar{d}_1(\hat{\zeta}^{n_{\text{eval}}}, \bar{\zeta}^{\tau,\sigma}) \leq \epsilon/3 \tag{C.5}$$

whenever $\max\{n_{\text{eval}}, n_{\text{opt}}\} \geq n_\epsilon$ and $\tau \leq \tau_\epsilon$. To proceed, we note that for any $(x, a) \in \mathbf{X} \times \mathbf{A}$,

$$\begin{aligned} |q_\tau^{\pi^{\tau,\sigma}}(x, a) - q_\tau^\star(x, a)| &\leq |q_\tau^{\pi^{\tau,\sigma}}(x, a) - q_\tau^\star(x, a)| + |q_\tau^\star(x, a) - q_\tau^\star(x, a)| \\ &\leq |q_\tau^{\pi^{\tau,\sigma}}(x, a) - q_\tau^\star(x, a)| + \epsilon/3, \end{aligned}$$

where the last inequality holds for small enough τ by Theorem 3.2. Continuing, we have

$$\begin{aligned} |q_\tau^{\pi^{\tau,\sigma}}(x, a) - q_\tau^\star(x, a)| &= |q_\tau^{\pi^{\tau,\sigma}}(x, a) - q_\tau^{\pi^{\tau,\star}}(x, a)| \\ &= \gamma \left| \int_{\mathbf{X}} (\mathcal{V}_\tau q_\sigma^\star - \mathcal{V}_\tau q_\tau^\star) dP_{x,a} \right| \\ &\leq \gamma \|\mathcal{V}_\tau q_\sigma^\star - \mathcal{V}_\tau q_\tau^\star\|_{\text{sup}} \\ &\leq \gamma \|q_\sigma^\star - q_\tau^\star\|_{\text{sup}}, \end{aligned}$$

where the final inequality holds since $\mathcal{V}_\tau$, as a log-sum-exp, is 1-Lipschitz. Now, again by Theorem 3.2, for small enough τ (inducing small enough σ), we have

$$\gamma \|q_\sigma^* - q_\tau^*\|_{\sup} \leq \gamma \|q_\sigma^* - q_{\text{ref}}^*\|_{\sup} + \gamma \|q_\tau^* - q_{\text{ref}}^*\|_{\sup} \leq \epsilon/6 + \epsilon/6 = \epsilon/3.$$

Altogether, we have that

$$\sup_{x,a} |q_\tau^{\pi^{\tau,\sigma}}(x, a) - q_{\text{ref}}^*(x, a)| \leq \epsilon/3 + \epsilon/3 = 2\epsilon/3.$$

Next, since $d_1(\rho_1, \rho_2) \geq \mathbf{E}_{(Z_1, Z_2) \sim \rho_1 \otimes \rho_2} [|Z_1 - Z_2|]$, we have that

$$\|\mathcal{Q}\hat{\zeta}^{n_{\text{eval}}} - q_\tau^{\pi^{\tau,\sigma}}\|_{\sup} \leq \bar{d}_1(\hat{\zeta}^{n_{\text{eval}}}, \bar{\zeta}^{\tau,\sigma}) \leq \epsilon/3,$$

by (C.5). Now, by yet another triangle inequality,

$$\begin{aligned} \|\mathcal{Q}\hat{\zeta}^{n_{\text{eval}}} - q_{\text{ref}}^*\|_{\sup} &\leq \|\mathcal{Q}\hat{\zeta}^{n_{\text{eval}}} - q_\tau^{\pi^{\tau,\sigma}}\|_{\sup} + \|q_\tau^{\pi^{\tau,\sigma}} - q_{\text{ref}}^*\|_{\sup} \\ &\leq \epsilon/2 + 2\epsilon/3 = \epsilon. \end{aligned}$$

Consequently, we have

$$q_{\mathcal{G}_\tau \mathcal{Q}\hat{\zeta}^{n_{\text{eval}}}} \geq q_\tau \mathcal{G}_\tau \mathcal{Q}\hat{\zeta}^{n_{\text{eval}}} \geq q_{\text{ref}}^* - \epsilon.$$

Thus, we have shown that $\mathcal{G}_\tau \mathcal{Q}\hat{\zeta}^{n_{\text{eval}}}$ is ϵ -reference-optimal, completing the proof. $\square$

Lemma C.1. Let $\zeta \in \bar{K}^p(\mathbf{X} \times \mathbf{A}, \mathcal{P}(\mathbb{R}))$, $\tau, \sigma > 0$, $n_{\text{eval}}, n_{\text{opt}} \in \mathbb{N}$ be given. Define $\hat{\zeta}^{\sigma,*} := (\mathcal{J}_\sigma^*)^{n_{\text{opt}}} \zeta$, and let $\hat{\pi}^{\sigma,*} = \mathcal{G}_\tau \hat{\zeta}^{\sigma,*}$. Then, we have

$$\bar{d}_p((\mathcal{J}_\tau^{\hat{\pi}^{\sigma,*}})^{n_{\text{eval}}} \hat{\zeta}^{\sigma,*}, (\mathcal{J}_\tau^{\pi^{\tau,\sigma}})^{n_{\text{eval}}} \hat{\zeta}^{\sigma,*}) \lesssim \gamma^{n_{\text{eval}}} + \gamma^{n_{\text{opt}}/2p}.$$

Proof. By Lemma C.2, we have

$$\begin{aligned} &\bar{d}_p((\mathcal{J}_\tau^{\hat{\pi}^{\sigma,*}})^{n_{\text{eval}}} \hat{\zeta}^{\sigma,*}, (\mathcal{J}_\tau^{\pi^{\tau,\sigma}})^{n_{\text{eval}}} \hat{\zeta}^{\sigma,*}) \\ &\lesssim (\tau^{-1} \|\mathcal{Q}\hat{\zeta}^{\sigma,*} - q_\sigma^*\|_{\sup})^{1/2p} + \sqrt{\tau^{-1} \|\mathcal{Q}\hat{\zeta}^{\sigma,*} - q_\sigma^*\|_{\sup}} + \|\mathcal{Q}\hat{\zeta}^{\tau,\sigma} - q_\sigma^*\|_{\sup}. \end{aligned}$$

It remains to bound $\|\mathcal{Q}\hat{\zeta}^{\sigma,*} - q_\sigma^*\|_{\sup}$. However, by Lemma 4.4 and the contractivity of $\mathcal{B}_\sigma^*$, we have that

$$\|\mathcal{Q}\hat{\zeta}^{\sigma,*} - q_\sigma^*\|_{\sup} \lesssim \gamma^{n_{\text{opt}}}.$$

Since $|\gamma| < 1$, it follows that

$$\bar{d}_p((\mathcal{J}_\tau^{\hat{\pi}^{\sigma,*}})^{n_{\text{eval}}} \hat{\zeta}^{\sigma,*}, (\mathcal{J}_\tau^{\pi^{\tau,\sigma}})^{n_{\text{eval}}} \hat{\zeta}^{\sigma,*}) \lesssim \gamma^{n_{\text{opt}}/2p}.$$

$\square$

Lemma C.2. Let $\zeta \in \bar{K}^p(\mathbf{X} \times \mathbf{A}, \mathcal{P}(\mathbb{R}))$, $\tau, \sigma > 0$, and $n_{\text{eval}} \in \mathbb{N}$ be given. Then

$$\bar{d}_p((\mathcal{J}_\tau^{\hat{\pi}})^{n_{\text{eval}}} \zeta, (\mathcal{J}_\tau^{\pi})^{n_{\text{eval}}} \zeta) \lesssim (\tau^{-1} \|\mathcal{Q}\zeta - q_\sigma^*\|_{\sup})^{1/2p} + \sqrt{\tau^{-1} \|\mathcal{Q}\zeta - q_\sigma^*\|_{\sup}} + \|\mathcal{Q}\zeta - q_\sigma^*\|_{\sup}.$$

Proof. For simplicity, we define $\hat{\pi} = \mathcal{G}_\tau \mathcal{Q}\zeta$ and $\pi = \mathcal{G}_\tau q_\sigma^*$. We want to bound

$$\bar{d}_p((\mathcal{J}_\tau^{\hat{\pi}})^{n_{\text{eval}}} \zeta, (\mathcal{J}_\tau^{\pi})^{n_{\text{eval}}} \zeta).$$

By Lemma C.3, we have

$$\begin{aligned} &\bar{d}_p((\mathcal{J}_\tau^{\hat{\pi}})^{n_{\text{eval}}} \zeta, (\mathcal{J}_\tau^{\pi})^{n_{\text{eval}}} \zeta) \\ &\leq \gamma \bar{d}_p((\mathcal{J}_\tau^{\hat{\pi}})^{n_{\text{eval}}-1} \zeta, (\mathcal{J}_\tau^{\pi})^{n_{\text{eval}}-1} \zeta) + 2\gamma \sup_{y,b} \left[\|\text{id}\|_{L^p(((\mathcal{J}_\tau^{\hat{\pi}})^{n_{\text{eval}}-1} \zeta)_{y,b})} c_1 + c_2 \right] \\ &\leq \gamma^2 \bar{d}_p((\mathcal{J}_\tau^{\hat{\pi}})^{n_{\text{eval}}-2} \zeta, (\mathcal{J}_\tau^{\pi})^{n_{\text{eval}}-2} \zeta) + 2\gamma^2 \sup_{y,b} \left[\|\text{id}\|_{L^p(((\mathcal{J}_\tau^{\hat{\pi}})^{n_{\text{eval}}-2} \zeta)_{y,b})} c_1 + c_2 \right] \\ &\quad + 2\gamma \sup_{y,b} \left[\|\text{id}\|_{L^p(((\mathcal{J}_\tau^{\hat{\pi}})^{n_{\text{eval}}-1} \zeta)_{y,b})} c_1 + c_2 \right] \\ &\leq \gamma^{n_{\text{eval}}} \bar{d}_p(\zeta, \zeta) + 2 \sum_{k=1}^{n_{\text{eval}}} \gamma^k \sup_{y,b} \left[\|\text{id}\|_{L^p(((\mathcal{J}_\tau^{\hat{\pi}})^{n_{\text{eval}}-k} \zeta)_{y,b})} c_1 + c_2 \right] \end{aligned}$$

where

$$c_1 := \sup_x \|\hat{\pi}_x - \pi_x\|_{\text{TV}}^{1/p} \quad \text{and} \quad c_2 := c_1^p \|q_\sigma^*\|_{\text{sup}} + 2\|\mathcal{Q}\zeta - q_\sigma^*\|_{\text{sup}}.$$

Now $((\mathcal{T}_\tau^{\mathcal{G}_\tau q_\sigma^*})^n \hat{\zeta}^{\sigma,*})_{n \in \mathbb{N}}$ is the sequence of return distributions generated by iterative applications of a contractive operator on $\bar{K}^p(\mathbf{X} \times \mathbf{A}, \mathcal{P}(\mathbb{R}))$. Thus,

$$\sup_n \sup_{y,b} \|\text{id}\|_{L^p(((\mathcal{T}_\tau^{\mathcal{G}_\tau q_\sigma^*})^n \hat{\zeta}^{\sigma,*})_{y,b})} \leq c_3 < \infty,$$

where c_3 is a constant depending only on $p, \gamma, \sigma, \tau, \|r\|_{\text{sup}}$. It remains to bound c_1 and c_2 . By Theorem 3.6, we have

$$\begin{aligned} c_1 &= \sup_x \|\mathcal{G}_\tau \mathcal{Q}\zeta_x^{\sigma,*} - \mathcal{G}_\tau q_\sigma^*\|_{\text{TV}}^{1/p} \\ &\leq (\tau^{-1} \|\mathcal{Q}\zeta - q_\sigma^*\|_{\text{sup}})^{1/2p}. \end{aligned}$$

Thus, since $\|q_\sigma^*\|_{\text{sup}}$ is uniformly bounded for any $\sigma > 0$, we have shown that

$$\bar{d}_p((\mathcal{T}_\tau^{\hat{\pi}})^{n_{\text{eval}}} \zeta, (\mathcal{T}_\tau^\pi)^{n_{\text{eval}}} \zeta) \lesssim (\tau^{-1} \|\mathcal{Q}\zeta - q_\sigma^*\|_{\text{sup}})^{1/2p} + \sqrt{\tau^{-1} \|\mathcal{Q}\zeta - q_\sigma^*\|_{\text{sup}} + \|\mathcal{Q}\zeta - q_\sigma^*\|_{\text{sup}}}.$$

□

Lemma C.3. Let $\tau > 0$, $p \in [1, \infty)$, $q, q' \in M_b(\mathbf{X} \times \mathbf{A})$, and $\zeta, \zeta' \in \bar{K}^p(\mathbf{X} \times \mathbf{A}, \mathcal{P}(\mathbb{R}))$. Then,

$$\begin{aligned} &\bar{d}_p(\mathcal{T}_\tau^{\mathcal{G}_\tau q} \zeta, \mathcal{T}_\tau^{\mathcal{G}_\tau q'} \zeta') \\ &\leq \gamma \bar{d}_p(\zeta, \zeta') \\ &\quad + 2\gamma \sup_{(x,y,b) \in \mathbf{X} \times \mathbf{X} \times \mathbf{A}} \left[\|\text{id}\|_{L^p(\zeta_{y,b})} \|\pi_x^q - \pi_x^{q'}\|_{\text{TV}}^{1/p} + c_{q,q'} \|\pi_x^q - \pi_x^{q'}\|_{\text{TV}} + 2\|q - q'\|_{\text{sup}} \right], \end{aligned}$$

where $c_{q,q'} := \min\{\|q\|_{\text{sup}}, \|q'\|_{\text{sup}}\}$.

Proof. Observe

$$\begin{aligned} d_p(\mathcal{T}_\tau^{\mathcal{G}_\tau q} \zeta, \mathcal{T}_\tau^{\mathcal{G}_\tau q'} \zeta') &\leq d_p(\mathcal{T}_\tau^{\mathcal{G}_\tau q} \zeta, \mathcal{T}_\tau^{\mathcal{G}_\tau q'} \zeta) + d_p(\mathcal{T}_\tau^{\mathcal{G}_\tau q'} \zeta, \mathcal{T}_\tau^{\mathcal{G}_\tau q'} \zeta') \\ &\leq d_p(\mathcal{T}_\tau^{\mathcal{G}_\tau q} \zeta, \mathcal{T}_\tau^{\mathcal{G}_\tau q'} \zeta) + \gamma d_p(\zeta, \zeta'). \end{aligned}$$

So by Lemma C.4, we conclude. □

Lemma C.4. Let $\zeta \in K(\mathbf{X} \times \mathbf{A}, \mathcal{P}(\mathbb{R}))$ and $q, q' \in M_b(\mathbf{X} \times \mathbf{A})$. For any $\tau > 0$, defining $\pi^\bullet = \mathcal{G}_\tau \bullet$ for $\bullet \in \{q, q'\}$, denoting $c_{q,q'} = \min\{\|q'\|_{\text{sup}}, \|q\|_{\text{sup}}\}$, we have

$$\begin{aligned} &\bar{d}_p(\mathcal{T}_\tau^{\mathcal{G}_\tau q} \zeta, \mathcal{T}_\tau^{\mathcal{G}_\tau q'} \zeta) \\ &\leq 2\gamma \sup_{(x,y,b) \in \mathbf{X} \times \mathbf{X} \times \mathbf{A}} \left[\|\text{id}\|_{L^p(\zeta_{y,b})} \|\pi_x^q - \pi_x^{q'}\|_{\text{TV}}^{1/p} + c_{q,q'} \|\pi_x^q - \pi_x^{q'}\|_{\text{TV}} + 2\|q - q'\|_{\text{sup}} \right] \end{aligned}$$

Proof. For notational simplicity, we define

$$\xi_x^{\zeta,q} = (\text{proj}^{\mathbb{R}} - \tau \text{kl}[\mathcal{G}_\tau q] \circ \text{proj}^{\mathbf{X}})_\#(\zeta_{x,-} \otimes (\mathcal{G}_\tau q)_x).$$

Then, by the definition of $\mathcal{T}_\tau^\pi$, we have

$$(\mathcal{T}_\tau^{\mathcal{G}_\tau q} \zeta)_{x,a} = (\mathbf{b}_{r(x,a),\gamma} \circ \text{proj}^{\mathbb{R}})_\#(\xi_x^{\zeta,q} \otimes P_{x,a})$$

Following, by [40, Theorem 4.8], we have

$$d_p((\mathcal{T}_\tau^{\mathcal{G}_\tau q} \zeta)_{x,a}, (\mathcal{T}_\tau^{\mathcal{G}_\tau q'} \zeta)_{x,a}) \leq \gamma \int d_p(\xi_x^{\zeta,q}, \xi_x^{\zeta,q'}) dP_{x,a}.$$

We will now estimate the integrand above. By the definition of $\xi^{\zeta,q}$, for any $x \in \mathsf{X}$, denoting $\pi^q := \mathcal{G}_\tau q$, we have

$$\begin{aligned} d_p(\xi_x^{\zeta,q}, \xi_x^{\zeta,q'}) &= d_p\left((b_{-\tau \text{KL}(\pi_x^q \parallel \pi_x^{\text{ref}}), 1} \circ \text{proj}^{\mathbb{R}})_{\#}(\zeta_{x,-} \otimes \pi_x^q), (b_{-\tau \text{KL}(\pi_x^{q'} \parallel \pi_x^{\text{ref}}), 1} \circ \text{proj}^{\mathbb{R}})_{\#}(\zeta_{x,-} \otimes \pi_x^{q'})\right) \\ &\leq \underbrace{\inf_{\kappa_x \in \mathcal{C}(\zeta_x \otimes \pi_x^q, \zeta_x \otimes \pi_x^{q'})} \left(\int |z - z'|^p d\kappa_x \right)^{1/p}}_{\text{I}(x)} + \underbrace{\tau |\text{KL}(\pi_x^q \parallel \pi_x^{\text{ref}}) - \text{KL}(\pi_x^{q'} \parallel \pi_x^{\text{ref}})|}_{\text{II}(x)}. \end{aligned}$$

The inequality is due to a technique employed in the proof of Theorem 4.5. Next, by [40, Theorem 6.15], we bound I via

$$\text{I}(x) \leq 2^{\frac{p-1}{p}} \sup_{x', a'} \|\text{id}\|_{L^p(\zeta_{x', a'})} \|\pi_x^q - \pi_x^{q'}\|_{\text{TV}}^{1/p} \leq 2 \sup_{x', a'} \|\text{id}\|_{L^p(\zeta_{x', a'})} \|\pi_x^q - \pi_x^{q'}\|_{\text{TV}}^{1/p}$$

Now, for II, we have

$$\text{II}(x) \leq \min\{\|q'\|_{\text{sup}}, \|q\|_{\text{sup}}\} \|\pi_x^q - \pi_x^{q'}\|_{\text{TV}} + 2\|q - q'\|_{\text{sup}},$$

as shown in the proof of Lemma C.5. Therefore, we have shown that

$$\begin{aligned} d_p((\mathcal{T}_\tau^{\mathcal{G}_\tau q} \zeta)_{x,a}, (\mathcal{T}_\tau^{\mathcal{G}_\tau q'} \zeta)_{x,a}) &\leq \gamma \int \left[\sup_{x', a'} \|\text{id}\|_{L^p(\zeta_{x', a'})} \|\pi_x^q - \pi_x^{q'}\|_{\text{TV}}^{1/p} + c_{q,q'} \|\pi_x^q - \pi_x^{q'}\|_{\text{TV}} + 2\|q - q'\|_{\text{sup}} \right] dP_{x,a}. \end{aligned}$$

□

C.1 Supplemental Lemmas for Section 4

Lemma C.5. Let $\bar{\zeta} \in \bar{\mathsf{K}}^1(\mathsf{X} \times \mathsf{A}, \mathcal{P}(\mathbb{R}))$, and for any $n \in \mathbb{N}$, define $\bar{\zeta}^{n+1} = \mathcal{T}_\tau^* \bar{\zeta}^n$, with $\bar{\zeta}^0 = \mathcal{T}_\tau^* \zeta$. Also, define $\pi^n := \mathcal{G}_\tau \bar{\zeta}^n$. Then for any $x \in \mathsf{X}$, denoting $C_x := \|\mathcal{Q}\bar{\zeta}(x, \cdot) - q_\tau^*(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})}$,

$$\tau |\text{KL}(\pi_x^n \parallel \pi_x^{\text{ref}}) - \text{KL}(\pi_x^\tau \parallel \pi_x^{\text{ref}})| \leq (2 + C_1 \sqrt{\tau}) \max\{\gamma^n C_x, \sqrt{\gamma^n C_x}\},$$

where $C_1 < \infty$ is a constant. If $\tau \geq 2\gamma^n C_x$, then for a constant $C_2 < \infty$, we have

$$\tau |\text{KL}(\pi_x^n \parallel \pi_x^{\text{ref}}) - \text{KL}(\pi_x^{\tau, \star} \parallel \pi_x^{\text{ref}})| \leq (2 + C_2 \tau^{-1}) C_x \gamma^n.$$

Proof. First, observe that

$$\begin{aligned} &\tau |\text{KL}(\pi_x^n \parallel \pi_x^{\text{ref}}) - \text{KL}(\pi_x^{\tau, \star} \parallel \pi_x^{\text{ref}})| \\ &= \left| \int \mathcal{Q}\bar{\zeta}^n(x, a) d\pi_x^n(a) - \mathcal{V}_\tau \mathcal{Q}\bar{\zeta}^n(x) - \int q_\tau^*(x, a) d\pi_x^{\tau, \star}(a) + \mathcal{V}_\tau q_\tau^*(x) \right| \\ &\leq \left| \int \mathcal{Q}\bar{\zeta}^n(x, a) d\pi_x^n(a) - \int q_\tau^*(x, a) d\pi_x^{\tau, \star}(a) \right| + |\mathcal{V}_\tau \mathcal{Q}\bar{\zeta}^n(x) - \mathcal{V}_\tau q_\tau^*(x)| \\ &\leq \|q_\tau^*(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})} \|\pi_x^n - \pi_x^{\tau, \star}\|_{\text{TV}} + 2\|\mathcal{Q}\bar{\zeta}^n(x, \cdot) - q_\tau^*(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})} \end{aligned}$$

By Lemma 4.4 and the γ -contractivity of $\mathcal{B}_\tau^*$, we note that

$$\|\mathcal{Q}\bar{\zeta}^n(x, \cdot) - q_\tau^*(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})} \leq \gamma^n \|\mathcal{Q}\bar{\zeta}^0(x, \cdot) - q_\tau^*(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})}.$$

Then, by Theorem 3.6, we have

$$\|\pi_x^n - \pi_x^{\tau, \star}\|_{\text{TV}} \leq \begin{cases} \frac{2e-3}{4} \gamma^n \tau^{-1} \|\mathcal{Q}\bar{\zeta}(x, \cdot) - q_\tau^*(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})} & \|\mathcal{Q}\bar{\zeta}(x, \cdot) - q_\tau^*(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})} \leq \frac{1}{2} \gamma^{-n} \tau \\ \sqrt{\tau^{-1} \gamma^n \|\mathcal{Q}\bar{\zeta} - q_\tau^*\|_{L^\infty(\pi_x^{\text{ref}})}} & \text{otherwise.} \end{cases}$$

Note that $\|q_\tau^*\|_{\text{sup}} \leq \|r\|_{\text{sup}}/(1 - \gamma)$. Indeed, the upper bound is free; the lower bound comes from comparing q_τ^* with q_τ^π for $\pi = \pi_x^{\text{ref}}$. Altogether, we have that

$$\tau |\text{KL}(\pi_x^n \parallel \pi_x^{\text{ref}}) - \text{KL}(\pi_x^{\tau, \star} \parallel \pi_x^{\text{ref}})| \leq \left(2 + \frac{\|r\|_{\text{sup}} \tau^{-1/2}}{1 - \gamma} \right) \max\{\gamma^n C_x, \sqrt{\gamma^n C_x}\}$$

for $C_x = \|\mathcal{Q}\bar{\zeta}(x, \cdot) - q_\tau^*(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})}$.

If $\tau \geq 2\gamma^n C_x$, then we have the stronger bound

$$\tau |\text{KL}(\pi_x^n \parallel \pi_x^{\text{ref}}) - \text{KL}(\pi_x^{\tau, \star} \parallel \pi_x^{\text{ref}})| \leq (2 + C' \tau^{-1}) C_x \gamma^n,$$

where $C' = (2e - 3)\|r\|_{\text{sup}}/4(1 - \gamma)$. $\square$

Lemma C.6. *Let $\bar{\zeta} \in \bar{\mathcal{K}}^p(\mathsf{X} \times \mathsf{A}, \mathcal{P}(\mathbb{R}))$. For any $n \in N$, define $\bar{\zeta}^{n+1} = \mathcal{T}_\tau^* \bar{\zeta}^n$, with $\bar{\zeta}^0 = \mathcal{T}_\tau^* \bar{\zeta}$. Denoting by $\mathcal{C}(\rho_1, \rho_2)$ the space of all couplings between the measures ρ_1, ρ_2 , for all $x \in \mathsf{X}$ we have*

$$\inf_{\kappa \in \mathcal{C}(\bar{\zeta}_{x, -}^{\tau, \star} \otimes \pi_x^n, \bar{\zeta}_{x, -}^{\tau, \star} \otimes \pi_x^{\tau, \star})} \int |z - z'|^p d\kappa \leq C_p \frac{\gamma^{n/2}}{(1 - \gamma)^p} \sqrt{\tau^{-1} \|\mathcal{Q}\bar{\zeta}(x, \cdot) - q_\tau^*(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})}},$$

where $\pi^n := \mathcal{G}_\tau \mathcal{Q}\bar{\zeta}^n$ and $C_p < \infty$ is a constant depending only on p and $\|r\|_{\text{sup}}$. Moreover, when $n > \log \gamma^{-1} (\log 2 \|\mathcal{Q}\bar{\zeta}(x, \cdot) - q_\tau^*(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})} - \log \tau)$, we have

$$\inf_{\kappa \in \mathcal{C}(\bar{\zeta}_{x, -}^{\tau, \star} \otimes \pi_x^n, \bar{\zeta}_{x, -}^{\tau, \star} \otimes \pi_x^{\tau, \star})} \int |z - z'|^p d\kappa \leq C'_p \frac{\gamma^n}{(1 - \gamma)^p} \tau^{-1} \|\mathcal{Q}\bar{\zeta}(x, \cdot) - q_\tau^*(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})},$$

where $C'_p = (2e - 3)C_p/4$.

Proof. For notational convenience, define $\varpi_x^\bullet := \bar{\zeta}_{x, -}^{\tau, \star} \otimes \pi_x^\bullet$, for $\bullet \in \{n, (\tau, \star)\}$. Then,

$$\begin{aligned} W_n^p &:= \inf_{\kappa \in \mathcal{C}(\varpi_x^n, \varpi_x^{\tau, \star})} \int |z - z'|^p d\kappa = d_p^p(\varpi_x^n, \varpi_x^{\tau, \star}) \\ &\stackrel{(a)}{\leq} 2^{p-1} \int |z|^p d|\varpi_x^n - \varpi_x^{\tau, \star}| \\ &= 2^{p-1} \int \left[\int |z|^p d\bar{\zeta}_{x, a}^{\tau, \star}(z) \right] d|\pi_x^n - \pi_x^{\tau, \star}|(a) \\ &\stackrel{(b)}{\leq} \frac{3^{2p-1} \|r\|_{\text{sup}}^p}{(1 - \gamma)^p} \|\pi_x^n - \pi_x^{\tau, \star}\|_{\text{TV}}, \end{aligned}$$

where (a) applies [40, Theorem 6.15], and (b) uses that the support of $\bar{\zeta}^{\tau, \star}$ is contained in a ball of radius $3\|r\|_{\text{sup}}/(1 - \gamma)$. By Lemma B.6, it follows that

$$\begin{aligned} W_n^p &\leq \sqrt{\tau^{-1} \|\mathcal{Q}\bar{\zeta}^n(x, \cdot) - q_\tau^*(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})}} \\ &\leq C_p \frac{\gamma^{n/2}}{(1 - \gamma)^p} \sqrt{\tau^{-1} \|\mathcal{Q}\bar{\zeta}(x, \cdot) - q_\tau^*(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})}}, \end{aligned}$$

where the last inequality holds by Lemma 4.4 and the γ -contractivity of $\mathcal{B}_\tau^*$ with $C_p := 3^{2p-1} \|r\|_{\text{sup}}^p$.

Moreover, if $n > \log \gamma^{-1} (\log 2 \|\mathcal{Q}\bar{\zeta}(x, \cdot) - q_\tau^*(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})} - \log \tau)$, then

$$\|\mathcal{Q}\bar{\zeta}^n(x, \cdot) - q_\tau^*(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})} < \tau/2$$

for each $x \in \mathsf{X}$, so by Theorem 3.6,

$$W_n^p \leq \frac{2e - 3}{4} C_p \frac{\gamma^n}{(1 - \gamma)^p} \tau^{-1} \|\mathcal{Q}\bar{\zeta}(x, \cdot) - q_\tau^*(x, \cdot)\|_{L^\infty(\pi_x^{\text{ref}})}.$$

$\square$

D Comparison between vanishing temperature limits of ERL with and without temperature decoupling

In this section, we compare and contrast the properties of vanishing temperature limits of standard ERL (assuming they exist) with those achieved by the temperature decoupling gambit. As we showed in Theorem 2.3 and Theorem 3.9, both schemes achieve reference-optimality in the limit; yet, their

limits may be notably distinct according to criteria beyond the RL objective, as we saw in Sections 3.1 and 4.3.

In the remainder of this section, we will define $\zeta^{\text{ref},*} := \zeta^{\pi^{\text{ref},*}}$ as the return distribution function corresponding to the limiting temperature-decoupled policy, and $\zeta^{\text{ERL},*} := \zeta^{\pi^{\text{ERL},*}}$ as the return distribution function corresponding to the limiting ERL policy $\pi^{\text{ERL},*}$, assuming such a limit exists.

A very nice property of $\pi^{\text{ref},*}$ is that it is easy to characterize as the *optimality-filtered reference*, as per Definition 3.8. In particular, $\pi^{\text{ref},*}$ is characterized *entirely* in terms of the optimal action-value function q^* and the reference policy π^{ref} . On the other hand, as we see explicitly in Section 3.1, $\pi^{\text{ERL},*}$ *does not* have such a simple characterization: it is influenced also by the transition dynamics of the MDP (as well as the q^* and π^{ref}).

A notable consequence of this fact is that one can reason about $\pi^{\text{ref},*}$ generically across MDPs, which is not the case for $\pi^{\text{ERL},*}$. For instance, in *any* MDP, if π^{ref} is the uniform policy, $\pi^{\text{ref},*}$ is the uniform policy on optimal actions. Thus, one can say definitively that all actions leading to optimal behavior are played equally under $\pi^{\text{ref},*}$. But this is not true of $\pi^{\text{ERL},*}$; in general, it is difficult to characterize exactly how $\pi^{\text{ERL},*}$ behaves: among a set of MDPs with equal q^* , the corresponding $\pi^{\text{ERL},*}$ can vary significantly.

Similarly, this property of $\pi^{\text{ref},*}$ enables one to easily influence the optimal policy that is achieved via temperature decoupling by intervening on π^{ref} . Again, this is possible due to the simple characterization of $\pi^{\text{ref},*}$ as the optimality-filtered reference. Suppose, for example, there exists a particular action a^{sary} that you want to avoid whenever possible (e.g., certain controversial phrases in language generation). It may be undesirable to filter this action out completely (say, by choosing π^{ref} to never play a^{sary}), because perhaps from some states this action is necessary to achieve optimal return. Instead, with temperature-decoupling, you can choose π^{ref} to play this action with very low probability (e.g., $\pi_x^{\text{ref}}(a^{\text{sary}}) = p_{\text{ref}}$ for each x). By Theorem 3.9, a^{sary} will only ever be played when it achieves optimal returns, and moreover, as long as other actions exist that achieve optimal returns, a^{sary} will be played with much lower probability.

The same logic *does not* hold, in general, for $\pi^{\text{ERL},*}$. As we saw in Section 3.1, $\pi^{\text{ERL},*}$ may continue to play a^{sary} with high probability even if π^{ref} plays it with low probability. Suppose, for instance, that after playing a^{sary} in state x , it is optimal to play π^{ref} subsequently for the rest of the episode. Then $\pi^{\text{ERL},*}$ may strongly prefer to play a^{sary} from state x , even if other actions can achieve the same expected return. In fact, depending on the transition kernel, the scale of the rewards, and the discount factor, $\pi^{\text{ERL},*}$ may play a^{sary} from state x with arbitrarily high probability.