Revisiting Actor-Critic Methods in Discrete Action Off-Policy Reinforcement Learning

Reza Asad
Simon Fraser University
rasad@sfu.ca

Reza Babanezhad
Samsung AI - Montreal
babanezhad@gmail.com

Sharan Vaswani
Simon Fraser University
vaswani.sharan@gmail.com

Abstract

Value-based approaches such as DQN are the default methods for off-policy reinforcement learning with discrete-action environments such as Atari. Common policy-based methods are either on-policy and do not effectively learn from off-policy data (e.g. PPO), or have poor empirical performance in the discrete-action setting (e.g. SAC). Consequently, starting from discrete SAC (DSAC), we revisit the design of actor-critic methods in this setting. First, we determine that the coupling between the actor and critic entropy is the primary reason behind the poor performance of DSAC. We demonstrate that by merely decoupling these components, DSAC can have comparable performance as DQN. Motivated by this insight, we introduce a flexible off-policy actor-critic framework that subsumes DSAC as a special case. Our framework allows using an m-step Bellman operator for the critic update, and enables combining standard policy optimization methods with entropy regularization to instantiate the resulting actor objective. Theoretically, we prove that the proposed methods can guarantee convergence to the optimal regularized value function in the tabular setting. Empirically, we demonstrate that these methods can approach the performance of DQN on standard Atari games, and do so even without entropy regularization or explicit exploration.

1 Introduction

Value-based algorithms such as DQN [Mnih et al., 2013] and its derivatives (Rainbow [Hessel et al., 2018], IQN [Dabney et al., 2018], M-DQN [Vieillard et al., 2020b]) are commonly used in deep reinforcement learning. These methods can effectively learn from off-policy data, making them sample-efficient in complex environments. These value-based algorithms are especially well-suited for environments with discrete actions and have achieved strong performance on large-scale benchmarks such as Atari 2600 [Bellemare et al., 2013].

In contrast, common policy-based methods such as PPO [Schulman et al., 2017], TRPO [Schulman, 2015] are on-policy and do not effectively reuse the data collected by past policies, making them sample-inefficient. Although prior work has developed off-policy variants of these popular RL algorithms to improve sample efficiency [Queeney et al., 2021, Meng et al., 2023, Gan et al., 2024], these methods have not been thoroughly evaluated in discrete-action settings. Similarly, soft actor-critic (SAC) [Haarnoja et al., 2018] also supports off-policy learning, is sample-efficient, and achieves strong empirical performance for continuous control tasks. However, in the discrete action-space

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

SAC's extensions and prior off-policy policy optimization work [Wang et al., 2016, Christodoulou, 2019, Xu et al., 2021, Zhou et al., 2022] introduce multiple interacting components (such as off-policy Retrace [Munos et al., 2016], bias correction, a dueling network architecture, ad-hoc actor regularization, and double-averaged Q-clipping) which complicate implementation and can limit practical performance. Motivated by this gap between off-policy policy-based and value-based methods, we ask the question: *can we design simple actor-critic methods that can approach the performance of DQN in the off-policy discrete-action setting?*

Given the recent interest in applications such as reinforcement learning with human feedback [Ouyang et al., 2022], the development of better actor-critic methods for discrete action environments is important. Doing so can also help address the limitations of value-based methods. For example, DQN and its successors lack a principled mechanism for exploration, often relying on ϵ-greedy strategies that require manual tuning and are known to be brittle [Hessel et al., 2018]. On the other hand, actor-critic methods such as SAC rely on entropy regularization in both the actor and critic updates. Importantly, these methods do not require explicit exploration making them potentially easier to tune on a new environment. Furthermore, when used with complex function approximation, value-based methods are prone to the *delusional bias* [Lu et al., 2018], which arises from performing a greedy update per next state independently. This ignores the joint distribution over states and actions and may yield a target Q-function that is not realizable by the current function class. In contrast, SAC avoids the delusional bias by backing up the value function under the current actor policy rather than performing greedy max-based backups.

To progress towards our goal, we first investigate the poor empirical performance of the discrete variant of SAC (denoted as DSAC) in Christodoulou [2019] on various Atari games. Previous work has attributed this poor performance to the use of a fixed target entropy [Xu et al., 2021] and/or Q-value underestimation bias [Zhou et al., 2022]. To mitigate these issues, Xu et al. [2021] introduce an adaptive entropy target, while Zhou et al. [2022] propose a variant of SAC which incorporates entropy coefficient regularization and a clipped double-averaged Q-learning objective. While such modifications improve the algorithm's stability, they also introduce additional hyper-parameters, complicating the method without matching the performance of DQN. Furthermore, these modifications lack any theoretical justification even in the simple tabular setting.

Contribution 1: We perform an extensive ablation study on DSAC. Our results demonstrate that by simply disabling the entropy regularization in the critic update (corresponding to using a hard Bellman operator for policy evaluation) and keeping all other components (entropy regularized soft actor update, automatic entropy tuning) fixed, yields a stable variant of DSAC. The resulting variant does not require double Q-learning, does not introduce any additional hyper-parameters and is competitive with DQN across various Atari games (see Fig. 1). Alternatively, we find that DSAC with a carefully tuned (on a per-game basis) entropy coefficient in the critic can also attain similar performance as DQN. Hence, we conclude that the entropy coefficient in the critic update substantially impacts the empirical performance of DSAC.

In order to explain the good empirical performance of the proposed DSAC variant, and to systematically design related algorithms, it is necessary to develop a more general approach. Prior works [Vieillard et al., 2020a, Xiao, 2022] have proposed actor-critic frameworks in the discrete action setting. While Vieillard et al. [2020a] show that SAC falls within their general framework, the proposed DSAC variant can not be captured by this framework. In particular, in Vieillard et al. [2020a], the actor and critic entropy is closely coupled, and the framework cannot support using entropy regularization asymmetrically (e.g. entropy regularization for actor, but not for the critic). On the other hand, Xiao [2022] propose a policy gradient framework and analyze the actor objective in SAC. However, they assume that the Q functions are estimated using a black-box procedure, and do not instantiate the critic update.

Contribution 2: In Section 3, we develop a general off-policy actor-critic framework for the discrete-action setting, introducing new objectives, with variants of DSAC arising as special cases. In particular, in the policy evaluation step for the critic, we use a look-ahead target formed by either the soft *critic entropy*-regularized or hard Bellman operator. The policy optimization step for the actor consists of two stages: (i) computing an intermediate policy by using either the NPG [Kakade, 2001] or SPMA [Asad et al., 2024] updates, and (ii) projecting (using either the forward or reverse

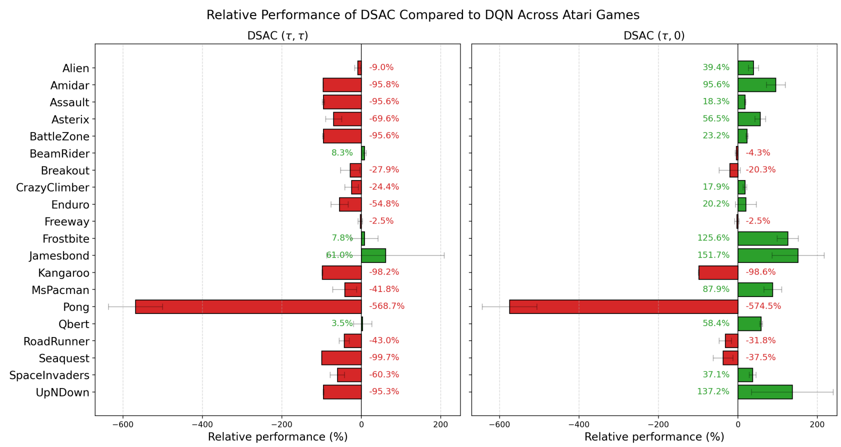

Figure 1: Performance of DSAC relative to DQN across 20 Atari games, with and without critic entropy. The left plot shows that the default DSAC underperforms DQN on most games, whereas incorporating a hard Bellman operator during policy evaluation (right plot) substantially improves DSAC's performance.

KL divergence) the intermediate policy onto the class of realizable policies while simultaneously maximizing a proximal entropy regularization term (referred to as the *actor entropy*). Importantly, our framework decouples the choice of using a critic entropy from that of the actor entropy and is consequently more flexible.

Contribution 3: In Section 4, we analyze the proposed actor-critic framework in the simplified tabular setting. In particular, in Theorem 1, we reduce the problem of analyzing the sub-optimality in the entropy-regularized value function to (i) bounding the policy evaluation error for the critic and (ii) bounding the regret for an online convex optimization problem related to the policy optimization step for the actor. For (i), we analyze the error incurred by a finite number of applications of the hard or soft Bellman operator (Corollary 1). For (ii), we bound the regret for actor entropy-regularized variants of both NPG and SPMA (Corollary 2). Our modular framework can be used to analyze different combinations of the actor and critic. For example, in Corollary 13, we use our framework to provide a theoretical guarantee for the NPG update with actor entropy and m steps of the hard Bellman operator for policy evaluation, a variant that matches the performance of DQN. Compared to Vieillard et al. [2020a] that couples the actor and critic entropy, and only analyzes one application of the Bellman operator, our framework is more flexible.

Contribution 4: In Section 5, we empirically evaluate the objectives derived from our actor-critic framework and compare them against DQN. Our results reveal three key findings. First, consistent with our observations on DSAC, our objectives benefit from using the hard Bellman operator for policy evaluation, leading to improved performance. Second, we demonstrate that unlike DSAC, the proposed objectives can achieve performance competitive with DQN even without entropy regularization or explicit exploration. Third, the choice of forward vs reverse KL in projecting the intermediate policy onto the class of realizable policies does not matter in most cases.

2 Preliminaries

We consider an infinite-horizon discounted Markov decision process (MDP), defined as $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho, \gamma \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ is the transition probability function, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $\rho \in \Delta_{\mathcal{S}}$ is the initial state distribution, and $\gamma \in [0, 1)$ is the discount factor. Throughout this paper, we assume that the state and action spaces are finite but potentially large.

For a fixed $s \in \mathcal{S}$, the policy π induces a probability distribution $\pi(\cdot|s)$ over the actions. The action-value function $q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ of policy π is defined as $q^\pi(s, a) := \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$, with $s_t \sim p(\cdot|s_{t-1}, a_{t-1})$ and $a_t \sim \pi(\cdot|s_t)$. Given an initial state $s \sim \rho$, the corresponding value function is defined as $v^\pi(s) := \mathbb{E}_{a \sim \pi(\cdot|s)}[q^\pi(s, a)]$. The advantage function

$\mathfrak{a}^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ induced by π is represented by $\mathfrak{a}^\pi(s,a) := q^\pi(s,a) - v^\pi(s)$. We define $J(\pi) := v^\pi(\rho) = \mathbb{E}_{s\sim\rho} v^\pi(s)$ as the expected discounted cumulative reward.
Given the Shannon entropy function $\mathcal{H}(\pi) = -\sum_a \pi(a)\,\ln(\pi(a))$ and if $\tau \geq 0$ is the entropy regularization, then, the soft (entropy-regularized) counterparts of the above functions [Liu et al., 2024, Vieillard et al., 2020a] are defined as: $v^\pi_\tau(s) := v^\pi(s) + \tau \sum_{t=0}^{\infty} \gamma^t \left[\mathcal{H}(\pi(\cdot|s_t)) \mid s_0 = s\right]$, $J_\tau(\pi) := \mathbb{E}_{s\sim\rho}[v^\pi_\tau(s)]$, $q^\pi_\tau(s,a) := \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[r(s,a) + \gamma v^\pi_\tau(s')\right]$ and $\mathfrak{a}^\pi_\tau(s,a) := q^\pi_\tau(s,a) - v^\pi_\tau(s) - \tau\,\ln(\pi(a|s))$. We note that the soft value and action-value functions, $v^\pi_\tau(s)$ and $q^\pi_\tau(s,a)$, are bounded and lie within the interval $[0, H_\tau]$, where $H_\tau := \frac{1+\tau\log(A)}{1-\gamma}$ [Liu et al., 2024]. Furthermore, for a fixed $s \in \mathcal{S}$ and an arbitrary pair of policies π_1 and π_2, the *soft* Bellman operator T^π_τ is defined such that: $(T^{\pi_1}_\tau v^{\pi_2}_\tau)(s) = \mathbb{E}_{a\sim\pi_1(\cdot|s)}\left[q^{\pi_2}_\tau(s,a) - \tau\,\ln(\pi_1(a|s))\right] = (T^{\pi_1} v^{\pi_2}_\tau)(s) + \tau\mathcal{H}(\pi_1(\cdot|s))$ and $(T^{\pi_1}_\tau q^{\pi_2}_\tau)(s,a) = (T^{\pi_1} q^{\pi_2}_\tau)(s,a) + \tau\,\mathcal{H}(\pi_1(\cdot|s))$. If $\tau = 0$, we refer to the corresponding operator as the *hard* Bellman operator. The objective is to $\max_{\pi\in\Pi} J_\tau(\pi)$, where Π is the set of feasible policies. We denote the optimal entropy-regularized policy by $\pi^*_\tau := \arg\max_\pi J_\tau(\pi)$ and its corresponding value function as v^*_τ.

3 A General Off-policy Actor-critic Framework

Building on the empirical findings from Section 1, we present a general off-policy actor-critic framework that subsumes variants of `DSAC` and enables the design of new actor objectives. In Section 3.1, we focus on the policy evaluation step in the tabular setting, and the subsequent critic objective with function approximation. In Section 3.2, we focus on two alternative policy updates in the tabular setting, and the corresponding actor objectives in the function approximation setting. We show how these objectives relate to existing methods, including `DSAC` as a special case.

3.1 Policy Evaluation

At iteration $t \in [K]$ of the actor-critic framework, we evaluate the current policy π_t using the m-step entropy-regularized Bellman operator. The coefficient, $\zeta \geq 0$ controls the strength of entropy regularization in the policy evaluation step and is referred to as the *critic entropy*. The corresponding estimate of the entropy-regularized q function at iteration t, denoted by q^t_ζ is computed recursively as:

$$q^0_\zeta = q^{\pi_0}_\zeta \quad ; \quad \forall t \geq 1, \quad q^t_\zeta = \mathbb{P}_{[0,H_\tau]}[(T^{\pi_t}_\zeta)^m q^{t-1}_\zeta]\,, \tag{1}$$

where $\mathbb{P}_{[0,H_\tau]}$ projects each entry onto the $[0, H_\tau]$ interval and is used only for theoretical purposes. As $m \to \infty$, q^t_ζ converges to the fixed point $q^{\pi_t}_\zeta$ and the algorithm exactly evaluates the policy π_t. In this special case, setting $\zeta = 0$ recovers the standard q function.

Handling function approximation: Eq. (1) requires updating the state-action value function for each state and action. In settings where the state or action space is large, this is not computationally feasible and we aim to approximate this update. For this, we use the standard squared loss and define the critic objective similar to Haarnoja et al. [2018]. Specifically, consider the *replay buffer* $\mathcal{D}_t$ consisting of (state, action, next state, reward) pairs obtained by the policies in the previous iterations, define ϕ as the parameters of a model (typically a neural network) parameterizing the critic and q_ϕ as the corresponding function. For our purposes, the details of the model are irrelevant and are implicit in the q_ϕ notation. We use Eq. (1) to construct a *one-step look-ahead target* (corresponding to $m = 1$) and define the critic objective $\mathcal{L}_t(\phi)$ given by:

$$\mathbb{E}_{\substack{(s,a,s',r(s,a))\sim\mathcal{D}_t \\ a'\sim\pi_t(\cdot|s')}} \left\| q_\phi(s,a) - \mathbb{P}_{[0,H_\tau]}[r(s,a) + \gamma\,[q^{t-1}_\zeta(s',a') - \zeta\,\ln(\pi_t(a'|s'))]\right\|_2^2. \tag{2}$$

With a slight abuse of notation, by $(s,a,s',r(s,a)) \sim \mathcal{D}_t$, we mean that the tuple is sampled from a discrete distribution over $\mathcal{D}_t$. We set $q^t_\zeta = q_{\phi_t}$ where $\phi_t := \arg\min \mathcal{L}_t(\phi)$. In practice, we do not project the look-ahead target, the optimization is done over the $(s,a,s',r(s,a))$ tuples in a randomly-sampled batch from $\mathcal{D}_t$ and q^{t-1}_ζ is parametrized using a separate target network, whose parameters are maintained as an exponential moving average of those in the critic model [Haarnoja et al., 2018, Christodoulou, 2019, Mnih et al., 2013].

3.2 Policy Optimization

The policy optimization step at iteration $t \in [K]$ uses the q function estimates from Section 3.1 and updates the policy. In particular, for a state $s \in \mathcal{S}$, we first compute an intermediate policy $\pi_{t+1/2}$ using two alterative ways: i) using the natural policy gradient (`NPG`) or policy mirror descent (PMD) [Kakade, 2001, Xiao, 2022] update or ii) the recently proposed `SPMA` update in Asad et al. [2024]. Given the step-size η_t,

$$\pi_{t+1/2}(a|s) \propto \pi_t(a|s)\ \exp(\eta_t\, q^t_\zeta(s,a))\textbf{(NPG)} \tag{3}$$

$$\pi_{t+1/2}(a|s) \propto \pi_t(a|s)\ \left[1 + \eta_t\ \left(q^t_\zeta(s,a) - v^t_\zeta(s)\right)\right]\ \textbf{(SPMA)} \tag{4}$$

In the special case, when $\zeta = 0$, we note that the `NPG` update requires an explicit normalization across actions to ensure that $\pi_{t+1/2}$ is a valid probability distribution. On the other hand, the `SPMA` update can use a sufficiently small step-size to avoid an explicit normalization across the actions [Asad et al., 2024]. We now use a proximal update to include entropy-regularization in the policy optimization step. Given $\pi_{t+1/2}$, if $\tau \geq 0$, then the updated policy π_{t+1} can be computed in two alternative ways that aim to find the "closest" policy π to $\pi_{t+1/2}$ while encouraging the resulting policy to have sufficiently high entropy (referred to as the *actor entropy*[1]). Specifically, if $\tau_t := \tau\,\eta_t \geq 0$ is the entropy regularization parameter at iteration t, we use either the forward KL (FKL) or reverse KL (RKL) divergence to measure the proximity between policies. The resulting updates are given by:

$$\pi_{t+1}(\cdot|s) = \arg\min_{\pi \in \Delta} \text{KL}(\pi_{t+1/2}(\cdot|s) \| \pi(\cdot|s)) - \tau_t \mathcal{H}(\pi(\cdot|s)) \qquad \textbf{(FKL)} \tag{5a}$$

$$\pi_{t+1}(\cdot|s) = \arg\min_{\pi \in \Delta} \text{KL}(\pi(\cdot|s) \| \pi_{t+1/2}(\cdot|s)) - \tau_t \mathcal{H}(\pi(\cdot|s)) \qquad \textbf{(RKL)} \tag{5b}$$

Note that in the special case when $\tau_t = 0$, $\pi_t = \pi_{t+1/2}$. Furthermore, the objective in Eq. (5b) is convex in π and the resulting update can be obtained in closed form where $\pi_{t+1}(a|s) \propto [\pi_{t+1/2}(a|s)]^{\frac{1}{1+\tau_t}}$. Combining Eqs. (5a) and (5b) with Eqs. (3) and (4) gives rise to four possible ways of updating the policy. We instantiate the corresponding actor objectives in the function approximation setting below.

Handling function approximation: Eqs. (5a) and (5b) require updating the state-action value function for each state and action. In order to scale to large state-action spaces, we aim to handle function approximation in the policy space. Consequently, given the actor model parameterized by θ, we define $\Pi_\theta = \{\pi | \exists\, \theta \text{ s.t } \pi = \pi(\theta)\}$ as the set of realizable policies. Similar to Section 3.1, the choice of the model is implicit in the $\pi(\theta)$ notation. Following [Haarnoja et al., 2018], we modify Eqs. (5a) and (5b) to optimize (i) only over the states in the replay buffer $\mathcal{D}_t$ and (ii) over the restricted policy class Π_θ to get the following updates: $\pi_{t+1} = \arg\min_{\pi \in \Pi_\theta} \sum_{s \sim \mathcal{D}_t} \left[\text{KL}(\pi_{t+1/2}(\cdot|s) \| \pi(\cdot|s)) - \tau_t \mathcal{H}(\pi(\cdot|s))\right]$ and $\pi_{t+1}(\cdot|s) = \arg\min_{\pi \in \Pi_\theta} \sum_{s \sim \mathcal{D}_t} \left[\text{KL}(\pi(\cdot|s) \| \pi_{t+1/2}(\cdot|s)) - \tau_t \mathcal{H}(\pi(\cdot|s))\right]$ for the FKL and RKL variants respectively. Following prior work [Vaswani et al., 2021, Lavington et al., 2023, Tomar et al., 2020, Xiong et al., 2024], we convert the above projection problem into an unconstrained optimization over the parameters, and form $\ell_t(\theta)$, the corresponding actor objective. For each variant, we append the postfix (τ, ζ) to denote its dependence on the actor and critic entropy. The actor objective $\ell_t(\theta)$ can be defined as one of four possible choices:

$$\mathbb{E}_{s \sim \mathcal{D}_t}\left[\mathbb{E}_{a \sim \pi_\theta(.|s)}\left[q^t_\zeta(s,a) - \tau\ \ln(\pi_\theta(a|s))\right] - \frac{1}{\eta_t}\,\text{KL}(\pi_\theta(\cdot|s)\,||\,\pi_t(\cdot|s))\right] \qquad (\textbf{NPG-RKL}(\tau,\zeta))$$

$$\mathbb{E}_{s \sim \mathcal{D}_t}\left[\mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[\ln\left(1 + \eta_t(q^t_\zeta(s,a) - v^t_\zeta(s))\right) - \tau_t\ \ln \pi_\theta(a|s)] - \text{KL}(\pi_\theta(\cdot|s)||\pi_t(\cdot|s))\right] \qquad (\textbf{SPMA-RKL}(\tau,\zeta))$$

$$\mathbb{E}_{s \sim \mathcal{D}_t}\left[\mathbb{E}_{a \sim \pi_t(\cdot|s)}\left[\frac{\exp(\eta_t\, q^t_\zeta(s,a))}{\sum_{a'} \pi_t(a')\ \exp(\eta_t\, q^t_\zeta(s,a'))}\ \ln\left(\pi_\theta(a|s)\right)\right] + \tau_t \mathcal{H}(\pi_\theta(.|s))\right] \qquad (\textbf{NPG-FKL}(\tau,\zeta))$$

[1] The actor entropy coefficient τ is independent from the critic entropy coefficient ζ

$$\mathbb{E}_{s\sim\mathcal{D}_t}\left[\mathbb{E}_{a\sim\pi_t(\cdot|s)}\left[\frac{(1+\eta_t(q^t_\zeta(s,a)-v^t_\zeta(s)))}{\sum_{a'}\pi_t(a')\,(1+\eta_t(q^t_\zeta(s,a')-v^t_\zeta(s)))}\;\ln\left(\pi_\theta(a|s)\right)\right]+\tau_t\mathcal{H}(\pi_\theta(.|s))\right] \tag{SPMA-FKL(τ,ζ)}$$

We define $\pi_{t+1} = \pi(\theta_{t+1})$ where $\theta_{t+1} = \arg\max_\theta \ell_t(\theta)$. Note that the NPG−RKL and SPMA−RKL objectives differ in the first term, which is linear in the q function for NPG−RKL, while it is logarithmic in the advantage for SPMA−RKL. Similarly, the FKL variants also differ in the first term. Crucially, in the special case of zero critic entropy, SPMA−FKL$(\tau, 0)$ yields a valid probability distribution by choosing $\eta \leq 1-\gamma$ and without requiring an explicit normalization over the actions [Asad et al., 2024], making it easier to implement in practice. Finally, we note that for the RKL variants, the expectation is over the actions sampled from π_θ. From an implementation perspective, optimizing the objective either involves calculating the full expectation or using the reparameterization trick. On the other hand, the FKL variants involve an expectation over the actions sampled from π_t, simplifying the resulting implementation. In Algorithm 1 in Appendix B, we present the entire pseudo-code in the function approximation setting.

Comparison to existing methods: First, considering the RKL variants, observe that NPG−RKL(τ, τ) recovers the off-policy variant of MDPO studied in [Tomar et al., 2020]. In the limit that $\eta_t \to \infty$, NPG−RKL(τ, τ) recovers the original SAC objective in Haarnoja et al. [2018]. This is intuitive since SAC can be viewed as soft policy iteration. Importantly, NPG−RKL$(\tau, 0)$ and $\eta_t \to \infty$ recovers the DSAC variant that demonstrated good empirical performance in Fig. 1. To the best of our knowledge, the SPMA−RKL(τ, ζ) objective is novel, and has not been studied in the previous literature. Now, considering the FKL variants, note that in the tabular setting, NPG−FKL(τ, τ) is the same as the objective proposed by Mei et al. [2019], but under function approximation, the two objectives differ. On the other hand, SPMA−FKL$(0, 0)$ can be interpreted as an off-policy extension of the surrogate introduced in Asad et al. [2024]. More generally, in the tabular setting, when $\zeta = \tau$ (i.e. the actor and critic entropy is coupled), using fixed coefficients for entropy and KL regularization, the NPG−RKL variant is the same as that proposed in Vieillard et al. [2020a]. Moreover, by sampling the states from d^{π_t}, the distribution induced by policy π_t (instead of $\mathcal{D}_t$), setting $\tau = \zeta = 0$ and $m = \infty$ for the critic (corresponding to exact policy evaluation) we can recover the framework in Vaswani et al. [2021], and its instantiations [Tomar et al., 2020, Asad et al., 2024]. Hence, in this sense, our framework generalizes that in Vieillard et al. [2020a] and Vaswani et al. [2021]. In the next section, we show that the RKL variants of the above objectives have theoretical guarantees in the simplified tabular setting.

4 Theoretical Guarantee in the Tabular Setting

In this section, we prove a bound on the sub-optimality of the entropy-regularized value function when using decoupled critic and actor entropy. For the critic, we consider estimating the q functions using Eq. (1) and allow for $\zeta = 0$, whereas for the actor, we consider using the RKL variant in Eq. (5b) with $\tau \neq 0$ for both the NPG and SPMA updates in Eqs. (3) and (4). We denote the corresponding variants as *soft NPG* and *soft SPMA* respectively.

In Theorem 1, we first reduce the problem of bounding the sub-optimality to a per-state online convex optimization problem. This reduction is independent of both the actor and critic updates.

Theorem 1 (Generic Reduction with Actor Entropy)**.** *If π^*_τ is the optimal entropy-regularized policy whose value function is equal to v^*_τ, for an estimate of $q^{\pi_t}_\tau$ obtained via the policy evaluation scheme in Eq. (1) at iteration t s.t. $\epsilon_t = q^t_\zeta - q^{\pi_t}_\tau$ and for any sequence of policies $\{\pi_0, \pi_1, \ldots, \pi_K\}$ if $\bar{\pi}_K$ is the corresponding mixture policy, then,*

$$\left\|v^*_\tau - v^{\bar{\pi}_K}_\tau\right\|_\infty \leq \frac{\|Regret(K)\|_\infty}{K\,(1-\gamma)} + \frac{2\sum_{t\in[K]}\|\boldsymbol{\epsilon}_t\|_\infty}{K\,(1-\gamma)}\,,$$

where $(Regret(K))(s) := \sum_{t=0}^{K-1}\left[\langle \pi^*_\tau(\cdot|s) - \pi_t(\cdot|s), q^t_\zeta(s,\cdot)\rangle + \tau\;[\mathcal{H}(\pi^*_\tau(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))]\right]$ is the regret incurred on an online convex optimization problem for each state $s \in \mathcal{S}$.

The above result shows that the sub-optimality of a mixture policy obtained by using a generic policy optimization method can be bounded in terms of the regret incurred by the method, and the sum of errors incurred in estimating $q_\tau^{\pi_t}$, the entropy-regularized q function corresponding to policy π_t. Since ϵ_t depends on π_t, it implicitly depends on the policy optimization method. We now bound the *policy evaluation errors* ϵ_t for both soft NPG and soft SPMA.

Corollary 1 (Policy Evaluation Error). *Using the policy evaluation update in Eq. (1), setting $\eta_t = \frac{1}{c+\tau\,(t+1)}$ for a constant c to be determined later, and defining $\delta(\tau,\zeta) := \frac{|\tau-\zeta|\,\ln(A)}{1-\gamma}$, the error ϵ_t when using soft NPG and soft SPMA can be bounded for all $t \in [K]$ as:*

$$\epsilon_t := \|\boldsymbol{\epsilon}_t\|_\infty = O\left(\frac{\gamma^m}{(1-\gamma)^3}\left[\ln(K)\,\left(\frac{1}{\sqrt{t}} + \gamma^{\frac{t}{2}}\right) + \frac{1}{\sqrt{K}}\right] + \frac{\delta(\tau,\zeta)}{1-\gamma}\right)$$

Note that as m (the number of steps of the Bellman operator) increases, the policy is evaluated more accurately, and ϵ_t decreases. As $m \to \infty$, $\epsilon_t \to O(\delta(\tau,\zeta))$, a term that quantifies the mismatch between the actor and critic entropy and is equal to zero when $\zeta = \tau$.

Next, we show that both soft NPG and soft SPMA can control the regret for the online optimization problem defined in Theorem 1.

Corollary 2 (Regret Bounds with Actor Entropy). *Suppose $\pi_0(\cdot|s)$ is the uniform distribution over actions for each state s, and let $\eta_t = \frac{1}{c+\tau\,(t+1)}$ for some constant $c \geq 0$ to be determined later. For any sequence $\{q_\zeta^t\}_{t=0}^{K-1}$ satisfying $\left\|q_\zeta^t\right\|_\infty \leq H_\tau$, the regret bound for soft NPG and soft SPMA is:*

$$\max_s \left|\sum_{t=0}^{K-1}\left[\langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s),\, q_\zeta^t(s,\cdot)\rangle + \tau\left[\mathcal{H}(\pi_\tau^*(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right]\right]\right| = O(\ln(K))$$

Since Eq. (1) ensures that $\left\|q_\zeta^t\right\|_\infty \leq H_\tau$, we can use the above result. The actor entropy term in Eq. (5b) makes the online functions strongly-convex and hence both methods incur only a logarithmic regret [Orabona, 2019]. Combining the results in Corollaries 1 and 2 with Theorem 1, we obtain our main theorem for both soft NPG and soft SPMA.

Theorem 2 (Sub-optimality of Soft NPG/Soft SPMA). *Let π_τ^* denote the optimal entropy-regularized policy with value function v_τ^*. Consider the soft NPG and soft SPMA updates with step size $\eta_t = \frac{1}{c+\tau\,(t+1)}$, constant c defined in Theorems 5 and 7, $\delta(\tau,\zeta) := \frac{|\tau-\zeta|\,\ln(A)}{1-\gamma}$. Let $\pi_0(\cdot|s)$ be the uniform policy over actions for all $s \in \mathcal{S}$, and assume the policy evaluation step is performed using Eq. (1). Then the resulting mixture policy $\bar{\pi}_K$ satisfies the following sub-optimality bound:*

$$\left\|v_\tau^{\bar{\pi}_K} - v_\tau^*\right\|_\infty = O\left(\frac{\ln(K)}{K\,(1-\gamma)^3} + \frac{\gamma^m}{K\,(1-\gamma)^4}\left[\ln(K)\left(\sqrt{K} + \frac{1}{1-\sqrt{\gamma}}\right) + \sqrt{K}\right] + \frac{\delta(\tau,\zeta)}{1-\gamma}\right)$$

Hence, the sub-optimality can be bounded as $O\left(\frac{1}{K} + \frac{\gamma^m}{\sqrt{K}} + \delta(\tau,\zeta)\right)$. Therefore, for exact policy evaluation ($m = \infty$) with coupled actor and critic entropy ($\zeta = \tau$), both soft NPG and SPMA have a $O(1/K)$ convergence for the resulting mixture policy. When using the one-step Bellman operator ($m = 1$) with $\zeta = \tau$, both methods have an $O(1/\sqrt{K})$ convergence. Finally, for the practical variant that sets $m = 1$ and $\zeta = 0$, the above result can achieve an $O(1/\sqrt{K})$ convergence to a $O(\tau)$ neighbourhood of v_τ^*. All proofs are deferred to Appendix C, where we also explore the case without actor entropy i.e. $\tau = 0$ and prove an $O(1/\sqrt{K})$ rate.

Comparison to existing results: We note that Vieillard et al. [2020a] consider the special case of soft NPG with coupled actor and critic entropy ($\zeta = \tau$) and $m = 1$, and establish a convergence rate of $O(1/K)$. Our result above is weaker ($O(1/K)$ vs $O(1/\sqrt{K})$ convergence) in this case. On the other hand Xiao [2022, Theorem 4.4] analyze soft NPG with coupled actor and critic entropy ($\zeta = \tau$) and $m = \infty$, and prove an $O(1/K)$ convergence rate. In this case, our result above can match this convergence rate. However, we note that our framework is more flexible – we can support $m \in (1,\infty)$ (left as an open question in Vieillard et al. [2020a]), support policy optimization methods beyond NPG (for example, SPMA) that have sublinear regret and Lipschitz policy updates (see the proof of Theorem 3 in Appendix C).

In the next section, we demonstrate that our framework encompasses methods (soft NPG and soft SPMA with decoupled actor and critic entropy and $m = 1$) that have good empirical performance.

5 Empirical Evaluation

In this section, we empirically evaluate the proposed off-policy policy optimization methods on various Atari 2600 games [Bellemare et al., 2013]. We follow the experimental protocol of Tomar et al. [2020], Asad et al. [2024], utilizing the `stable-baselines3` framework [Raffin et al., 2021] and running each experiment with five random seeds. Results are reported as average expected returns with 95% confidence intervals. Additional details regarding the experimental setup and the hyper-parameters used for all algorithms are provided in Appendix E.1. We first present an ablation study of `DSAC` to identify the important factors that affect its empirical performance. Subsequently, we investigate the impact of actor and critic entropy and the direction of the KL divergence on the empirical performance of `DSAC` and the other objectives in our framework.

Conclusion 1: Using $\zeta = 0$ Improves DSAC Performance: In Fig. 2, we present an ablation study for `DSAC` (with a single Q-network and $\tau \neq 0$), and compare the performance of $\zeta = 0$ and $\zeta = \tau$. For setting τ, we compare the performance of a fixed entropy coefficient and the adaptive entropy regularization in Christodoulou [2019]. Our results indicate that using $\zeta = 0$ and adaptive τ yields strong performance comparable to `DQN` (green vs. blue curve). Note that we observed this same trend across the 20 Atari games as presented (see Fig. 1). Moreover, Fig. 2 shows that both $\zeta = 0$ and $\zeta = \tau$ can result in good empirical performance when using fixed well-tuned value of τ (red and purple curves vs. blue). However, since this tuning needs to be done on a per-environment basis, we conclude that it is beneficial to set $\zeta = 0$ and completely avoid the need for manual entropy coefficient tuning. For this simple configuration ($\zeta = 0$ and adaptive τ), we test the hypothesis in [Zhou et al., 2022] that double Q-learning [Fujimoto et al., 2018] adversely affects the `DSAC` performance. In Fig. 4 in Appendix E.2, we compare single and double Q-learning across 8 Atari games and observe that, under our setup, double Q-learning yields similarly strong results and does not degrade performance. We also evaluate the hypothesis in Xu et al. [2021] that a fixed target entropy used for auto-tuning the entropy coefficient adversely affects the `DSAC` performance. We observe that our proposed `DSAC` configuration significantly outperforms those reported in Xu et al. [2021], despite their use of an adaptive target entropy (see Figure 2 in Xu et al. [2021]).

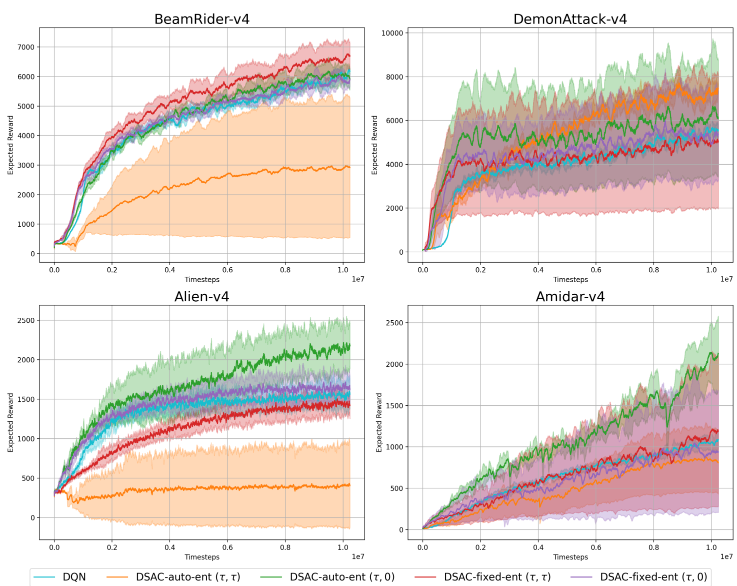

Figure 2: `DSAC` ablation study. Training `DSAC` with the adaptive entropy coefficient loss from the original paper with $\zeta = 0$ achieves consistently stronger performance than `DQN` across games (green vs. blue curves). In contrast, using $\zeta = \tau$ without a well-tuned fixed τ can lead to substantially worse performance (orange curve).

Conclusion 2: All Proposed Objectives with $\zeta = 0$ Have Similar Good Performance: From Fig. 5 in Appendix E we observe that all proposed actor objectives (RKL and FKL variants of both NPG and SPMA) defined in Section 3 have relatively (compared to DQN) bad performance with adaptive $\zeta = \tau$, and similar good performance with $\zeta = 0$ and adaptive τ. Moreover, similar to $\texttt{DSAC}(\tau, 0)$, these methods have good empirical performance even with 1 step of gradient descent (on a randomly sampled batch from the replay buffer) for both the actor and critic objectives. For the subsequent results, we set $\zeta = 0$ and use adaptive τ.

Conclusion 3: Entropy Regularization is Sometimes Important: We examine the effect of completely disabling entropy regularization ($\tau = \zeta = 0$) when the actor objective $\ell_t(\theta)$ is approximately optimized by n gradient steps. For $n = 1$, $\texttt{DSAC}(0, 0)$, and all proposed objectives exhibit substantially degraded performance. We attribute this to insufficient exploration, supported by the observed collapse in policy entropy. In particular, as shown in Fig. 6, the Shannon entropy drops sharply and fails to recover across all methods and games.
However, increasing n significantly mitigates this effect. On most games, using $n = 10$ results in relatively larger values of policy entropy, suggesting improved exploration and stability. In terms of performance, Fig. 3 shows that, with large n and $\tau = \zeta = 0$, all proposed objectives (except $\texttt{SPMA-RKL}(0, 0)$) achieve competitive results with DQN on most games. In contrast, $\texttt{DSAC}(0, 0)$ fails to match DQN even with large n. These findings in Fig. 8 indicate that several of our proposed objectives can achieve competitive performance with DQN without explicit entropy regularization, provided that the actor objective is sufficiently optimized.

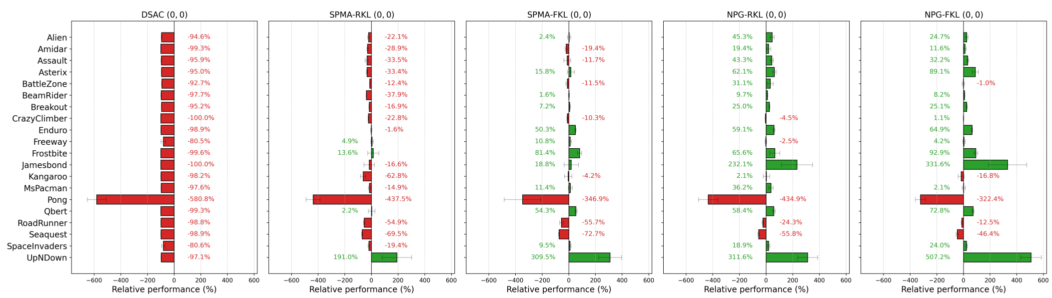

Figure 3: In contrast to $\texttt{DSAC}(0, 0)$, all our objectives (except $\texttt{SPMA-RKL}(0, 0)$) achieve performance comparable to DQN without entropy regularization, when increasing the number of actor gradient updates n to 10.

Conclusion 4: Direction of KL Mostly Does Not Matter: Prior work [Chan et al., 2022] has examined the role of the direction of the KL divergence when the intermediate policy is a Boltzmann distribution i.e. $\pi_{t+1/2} \propto \exp(q_\tau^t(s, a)/\tau)$ and reported no significant differences in the discrete-action setting. We revisit this question in the context of our actor objectives, evaluating SPMA and NPG as intermediate policies under two regimes: (i) with $\zeta = \tau = 0$ (e.g., $\texttt{NPG-RKL}(0, 0)$) and (ii) with $\tau \neq 0$ and $\zeta = 0$ (e.g., $\texttt{NPG-RKL}(\tau, 0)$). When $\zeta = \tau = 0$ and using $n = 10$ (since $n = 1$ has poor performance), forward KL provides a clear advantage for SPMA, whereas NPG does not exhibit a consistent trend (Figs. 11 and 12). When $\tau \neq 0$ and $\zeta = 0$, either the value of n or choice of the intermediate policy yields no consistent evidence that one KL direction systematically outperforms the other (see Figs. 13 to 16 in Appendix E).

6 Conclusion and Future Work

We revisited the design and implementation of off-policy actor-critic methods in the discrete-action setting, where value-based methods such as DQN and its derivatives dominate. By decoupling actor and critic entropy, we identified a variant of the discrete extension of SAC that achieves performance comparable to DQN. Building on this insight, we introduced a flexible off-policy actor-critic framework that separates entropy regularization, enables the actor to incorporate two distinct intermediate policies followed by forward or reverse KL projections, and subsumes variants of DSAC. For future work, we plan to: (i) investigate strategies for automatically selecting the actor step size η_t rather than relying on grid search; (ii) thoroughly investigate the behavior of the proposed objectives in the absence of entropy regularization, and (iii) evaluate the performance of the proposed methods in continuous-action settings.

Acknowledgments

We would like to thank Michael Liu for helpful feedback on the paper. This work was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2022-04816, and enabled in part by support provided by the Digital Research Alliance of Canada (alliancecan.ca).

7 Bibliography

Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702. PMLR, 2019.

Reza Asad, Reza Babanezhad, Issam Laradji, Nicolas Le Roux, and Sharan Vaswani. Fast convergence of softmax policy mirror ascent. *arXiv preprint arXiv:2411.12042*, 2024.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47: 253–279, 2013.

Alan Chan, Hugo Silva, Sungsu Lim, Tadashi Kozuno, A Rupam Mahmood, and Martha White. Greedification operators for policy optimization: Investigating forward and reverse kl divergences. *Journal of Machine Learning Research*, 23(253):1–79, 2022.

Petros Christodoulou. Soft actor-critic for discrete action settings. *arXiv preprint arXiv:1910.07207*, 2019.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018.

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

Yaozhong Gan, Renye Yan, Xiaoyang Tan, Zhe Wu, and Junliang Xing. Transductive off-policy proximal policy optimization. *arXiv preprint arXiv:2406.03894*, 2024.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Jonathan Wilder Lavington, Sharan Vaswani, Reza Babanezhad, Mark Schmidt, and Nicolas Le Roux. Target-based surrogates for stochastic optimization. *arXiv preprint arXiv:2302.02607*, 2023.

Jiacai Liu, Wenye Li, and Ke Wei. Elementary analysis of policy gradient methods. *arXiv preprint arXiv:2404.03372*, 2024.

Tyler Lu, Dale Schuurmans, and Craig Boutilier. Non-delusional q-learning and value-iteration. *Advances in neural information processing systems*, 31, 2018.

Jincheng Mei, Chenjun Xiao, Ruitong Huang, Dale Schuurmans, and Martin Müller. On principled entropy exploration in policy optimization. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3130–3136, 2019.

Wenjia Meng, Qian Zheng, Gang Pan, and Yilong Yin. Off-policy proximal policy optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9162–9170, 2023.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and efficient off-policy reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

James Queeney, Yannis Paschalidis, and Christos G Cassandras. Generalized proximal policy optimization with sample reuse. *Advances in Neural Information Processing Systems*, 34:11909–11919, 2021.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

Igal Sason. Entropy bounds for discrete random variables via maximal coupling. *IEEE Transactions on Information Theory*, 59(11):7118–7131, 2013.

John Schulman. Trust region policy optimization. *arXiv preprint arXiv:1502.05477*, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. *arXiv preprint arXiv:2005.09814*, 2020.

Sharan Vaswani, Olivier Bachem, Simone Totaro, Robert Müller, Shivam Garg, Matthieu Geist, Marlos C Machado, Pablo Samuel Castro, and Nicolas Le Roux. A general class of surrogate functions for stable and efficient reinforcement learning. *arXiv preprint arXiv:2108.05828*, 2021.

Nino Vieillard, Tadashi Kozuno, Bruno Scherrer, Olivier Pietquin, Rémi Munos, and Matthieu Geist. Leverage the average: an analysis of kl regularization in reinforcement learning. *Advances in Neural Information Processing Systems*, 33:12163–12174, 2020a.

Nino Vieillard, Olivier Pietquin, and Matthieu Geist. Munchausen reinforcement learning. *Advances in Neural Information Processing Systems*, 33:4235–4246, 2020b.

Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Remi Munos, Koray Kavukcuoglu, and Nando De Freitas. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*, 2016.

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022.

Zhihan Xiong, Maryam Fazel, and Lin Xiao. Dual approximation policy optimization. *arXiv preprint arXiv:2410.01249*, 2024.

Yaosheng Xu, Dailin Hu, Litian Liang, Stephen McAleer, Pieter Abbeel, and Roy Fox. Target entropy annealing for discrete soft actor-critic. *arXiv preprint arXiv:2112.02852*, 2021.

Haibin Zhou, Zichuan Lin, Junyou Li, Qiang Fu, Wei Yang, and Deheng Ye. Revisiting discrete soft actor-critic. *arXiv preprint arXiv:2209.10081*, 2022.

Supplementary Material

Organization of the Appendix

A Actor Objective Instantiations

In this section, we instantiate our forward and reverse KL-based objectives in Equations 5a and 5b using soft `NPG`, and soft `SPMA` in the function approximation setting.

A.1 `NPG-FKL`(τ, ζ)

$$\pi_{t+1} = \operatorname*{arg\,min}_{\pi \in \Pi} \mathbb{E}_{s \sim d_t} \left[\mathrm{KL}(\pi_{t+1/2}(\cdot|s) || \pi_\theta(\cdot|s)) - \tau_t \mathcal{H}(\pi_\theta(\cdot|s))\right]$$

$$= \arg\min_{\pi \in \Pi} \mathbb{E}_{s \sim d_t} \left[\mathbb{E}_{a \sim \pi_t(\cdot|s)} \left[- \frac{\exp(\eta_t\, q^t_\zeta(s,a))}{\mathcal{Z}_t(s)} \ln \left(\frac{\pi_\theta(a|s)}{\frac{\pi_t(a|s)\, \exp(\eta_t q^t_\zeta(s,a))}{\mathcal{Z}_t(s)}} \right)\right] - \tau_t \mathcal{H}(\pi_\theta(\cdot|s))\right]$$

(Using the NPG update in Eq. (3))

$$= \operatorname*{arg\,min}_{\pi \in \Pi} \mathbb{E}_{s \sim d_t} \left[\mathbb{E}_{a \sim \pi_t(\cdot|s)} \left[- \frac{\exp(\eta_t\, q^t_\zeta(s,a))}{\mathcal{Z}_t(s)} \ln \left(\frac{\pi(a|s)}{\pi_t(a|s)} \right)\right] - \tau_t \mathcal{H}(\pi_\theta(\cdot|s))\right]$$

(dropping the constant w.r.t π)

$$= \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_t} \left[\mathbb{E}_{a \sim \pi_t(\cdot|s)} \left[\frac{\exp(\eta_t\, q^t_\zeta(s,a))}{\sum_{a'} \pi_t(a')\, \exp(\eta_t\, q^t_\zeta(s,a))} \ln \left(\frac{\pi_\theta(a|s)}{\pi_t(a|s)} \right)\right] + \tau_t \mathcal{H}(\pi_\theta(\cdot|s))\right]$$

A.2 `SPMA-FKL`(τ, ζ)

$$\pi_{t+1} = \operatorname*{arg\,min}_{\pi \in \Pi} \mathbb{E}_{s \sim d_t} \left[\mathrm{KL}(\pi_{t+1/2}(\cdot|s) || \pi_\theta(\cdot|s)) - \tau_t \mathcal{H}(\pi_\theta(\cdot|s))\right]$$

$$= \arg\min_{\pi \in \Pi} \mathbb{E}_{s \sim d_t} \left[\mathbb{E}_{a \sim \pi_t(\cdot|s)} \left[- \frac{\left(1 + \eta_t (q^t_\zeta(s,a) - v_\zeta(s)^t)\right)}{\mathcal{Z}_t(s)} \ln \left(\frac{\pi_\theta(a|s)}{\frac{\pi_t(a|s)\, (1 + \eta_t\, (q^t_\zeta(s,a) - v^t_\zeta(s)))}{\mathcal{Z}_t(s)}} \right)\right.\right]$$

$$- \tau_t \mathcal{H}(\pi_\theta(\cdot|s))\Bigg] \qquad \text{(Using the SPMA update in Eq. (4))}$$

$$= \arg\min_{\pi \in \Pi} \mathbb{E}_{s \sim d_t} \left[\mathbb{E}_{a \sim \pi_t(\cdot|s)} \left[- \frac{\left(1 + \eta_t (q_\zeta^t(s,a) - v_\zeta^t(s))\right)}{\mathcal{Z}_t(s)} \ln \left(\frac{\pi_\theta(a|s)}{\pi_t(a|s)} \right) \right] \right.$$

$$\left. - \tau_t \mathcal{H}(\pi_\theta(\cdot|s)) \right] \qquad \text{(dropping the constant w.r.t } \pi)$$

$$= \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_t} \left[\mathbb{E}_{a \sim \pi_t(\cdot|s)} \left[\frac{\left(1 + \eta_t (q_\zeta^t(s,a) - v_\zeta^t(s))\right)}{\sum_{a'} \pi_t(a'|s) \left(1 + \eta_t (q_\zeta^t(s,a') - v_\zeta^t(s))\right)} \ln \left(\frac{\pi(a|s)}{\pi_t(a|s)} \right) \right] \right.$$

$$\left. + \tau_t \mathcal{H}(\pi_\theta(\cdot|s)) \right]$$

A.3 **NPG-RKL**(τ, ζ)

$$\pi_{t+1} = \arg\min_{\pi_\theta \in \Pi} \mathbb{E}_{s \sim d_t} \left[\text{KL}(\pi_\theta(\cdot|s) || \pi_{t+1/2}(\cdot|s)) - \tau_t \mathcal{H}(\pi_\theta(\cdot|s)) \right]$$

$$= \arg\min_{\pi_\theta \in \Pi} \mathbb{E}_{s \sim d_t} \mathbb{E}_{a \sim \pi_\theta(.|s)} \left[(1 + \tau_t) \ln(\pi_\theta(a|s)) - \ln(\pi_{t+1/2}(a|s)) \right]$$

$$= \arg\min_{\pi \in \Pi} \mathbb{E}_{s \sim d_t} \mathbb{E}_{a \sim \pi_\theta(.|s)} \left[(1 + \tau_t) \ln(\pi_\theta(a|s)) - \ln \left(\pi_t(a|s) \exp(\eta_t \, q_\zeta^t(s,a)) \right) \right]$$

(Using the NPG update in Eq. (3) and since $\mathcal{Z}_t$ can be marginalized out)

$$= \arg\min_{\pi \in \Pi} \mathbb{E}_{s \sim d_t} \mathbb{E}_{a \sim \pi_\theta(.|s)} \left[(1 + \tau_t) \ln(\pi_\theta(a|s)) - \ln(\pi_t(a|s)) - \eta_t \, q_\zeta^t(s,a) \right]$$

$$= \arg\min_{\pi_\theta \in \Pi} \mathbb{E}_{s \sim d_t} \mathbb{E}_{a \sim \pi_\theta(.|s)} \left[(1 + \tau \eta_t) \ln(\pi_\theta(a|s)) - \eta_t \left(q_\zeta^t(s,a) + \frac{1}{\eta_t} \ln(\pi_t(a|s)) \right) \right]$$

(Since $\tau_t = \eta_t \, \tau$)

$$= \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_t} \mathbb{E}_{a \sim \pi_\theta(.|s)} \left[q_\zeta^t(s,a) - \tau \ln(\pi_\theta(a|s)) - \frac{1}{\eta_t} \ln \left(\frac{\pi_\theta(a|s)}{\pi_t(a|s)} \right) \right]$$

$$= \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_t} \left[\mathbb{E}_{a \sim \pi_\theta(.|s)} \left[q_\zeta^t(s,a) - \tau \ln(\pi_\theta(a|s)) \right] - \frac{1}{\eta_t} \text{KL}(\pi_\theta(\cdot|s) \, || \, \pi_t(\cdot|s)) \right]$$

Given an estimate of the entropy-regularized q function, `DSAC` is a special of `NPG-RKL` by setting $\eta_t = \infty$ resulting in the following surrogate loss:

$$\pi_{t+1} = \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_t} \mathbb{E}_{a \sim \pi(.|s)} \left[\left(q_\zeta^t(s,a) \right) - \tau \ln(\pi(a|s)) \right]$$

A.4 **SPMA-RKL**(τ, ζ)

$$\pi_{t+1} = \arg\min_{\pi \in \Pi} \mathbb{E}_{s \sim d_t} \left[\text{KL}(\pi_\theta(\cdot|s) || \pi_{t+1/2}(\cdot|s)) - \tau_t \mathcal{H}(\pi_\theta(\cdot|s)) \right]$$

$$= \arg\min_{\pi \in \Pi} \mathbb{E}_{s \sim d_t} \mathbb{E}_{a \sim \pi_\theta(.|s)} \left[(1 + \tau_t) \ln(\pi(a|s)) - \ln(\pi_{t+1/2}(a|s)) \right]$$

$$= \arg\min_{\pi \in \Pi} \mathbb{E}_{s \sim d_t} \mathbb{E}_{a \sim \pi_\theta(.|s)} \left[(1 + \tau_t) \ln(\pi_\theta(a|s)) - \ln \left(\pi_t(a|s) \left[1 + \eta_t \left(q_\zeta^t(s,a') - v_\zeta^t(s) \right) \right] \right) \right]$$

(Using the SPMA update in Eq. (4))

$$= \arg\max_{\pi_\theta \in \Pi} \mathbb{E}_{s \sim d_t} \left[\mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \ln \left(1 + \eta_t (q_\zeta^t(s,a) - v_\zeta^t(s)) \right) - \text{KL}(\pi_\theta || \pi_t) - \tau_t \ln \pi_\theta(a|s) \right]$$

B General Off-Policy Actor-Critic Pseudocode

Algorithm 1: General Off-Policy Actor-Critic Framework

Input: θ_0 (π_0's parameters), ϕ_0 (q_ζ^0's parameters), π_θ (function approximation for actor), q_ϕ (function approximation for critic), $\mathcal{L}_t$ (critic loss), ℓ_t (actor loss), K (total iterations), N (number of environment steps), n (number of policy optimization steps), α (inner-loop step-size)
for $t = 0$ **to** $K - 1$ **do**
  Interact with the environment for N steps to collect data using π_t:
  $D_t \leftarrow D_t \cup \{s_i, a_i, r(s_i, a_i), s_{i+1}\}_{i=1}^N$
  $\phi_t = \arg\min \mathcal{L}_t(\phi)$; $q_\zeta^t = q_{\phi_t}$
  Initialize inner-loop: $\omega_0 = \theta_t$
  for $j = 0$ **to** $n - 1$ **do**
    $\omega_{j+1} = \omega_j - \alpha \nabla_\omega \ell_t(\omega_j)$
  end for
  $\theta_{t+1} = \omega_n$
  $\pi_{t+1}(\cdot|s) = \pi_{\theta_{t+1}}(\cdot|s)$
end for
Return: θ_K

C Theoretical Results

C.1 Proof of Theorem 1

Theorem 1 (Generic Reduction with Actor Entropy). *If π_τ^* is the optimal entropy-regularized policy whose value function is equal to v_τ^*, for an estimate of $q_\tau^{\pi_t}$ obtained via the policy evaluation scheme in Eq. (1) at iteration t s.t. $\epsilon_t = q_\zeta^t - q_\tau^{\pi_t}$ and for any sequence of policies $\{\pi_0, \pi_1, \dots, \pi_K\}$ if $\bar{\pi}_K$ is the corresponding mixture policy, then,*

$$\left\| v_\tau^* - v_\tau^{\bar{\pi}_K} \right\|_\infty \leq \frac{\|Regret(K)\|_\infty}{K\,(1-\gamma)} + \frac{2\sum_{t \in [K]} \|\boldsymbol{\epsilon}_t\|_\infty}{K\,(1-\gamma)},$$

where $(Regret(K))(s) := \sum_{t=0}^{K-1} \left[\langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s,\cdot) \rangle + \tau\, [\mathcal{H}(\pi_\tau^*(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))] \right]$ is the regret incurred on an online convex optimization problem for each state $s \in \mathcal{S}$.

Proof.

$$\begin{aligned}
v_\tau^{\pi_\tau^*} - v_\tau^\pi &= \mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_\tau^*} - v_\tau^\pi && \text{(Since } v_\tau^{\pi_\tau^*} \text{ is a fixed point of } \mathcal{T}_\tau^{\pi_\tau^*}) \\
&= [\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^\pi - v_\tau^\pi] + [\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi^*} - \mathcal{T}_\tau^{\pi_\tau^*} v_\tau^\pi] && \text{(Add/subtract } \mathcal{T}_\tau^{\pi_\tau^*} v_\tau^\pi) \\
&= [\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^\pi - v_\tau^\pi] + \gamma \mathcal{P}_{\pi_\tau^*}(v_\tau^{\pi^*} - v_\tau^\pi) && \text{(Using the definition of } \mathcal{T}_\tau^{\pi_\tau^*}) \\
\implies v_\tau^{\pi_\tau^*} - v_\tau^\pi &= \left(I - \gamma \mathcal{P}_{\pi_\tau^*}\right)^{-1} \left[\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^\pi - \mathcal{T}_\tau^\pi\, v_\tau^\pi \right]
\end{aligned}$$

Summing up from $t = 0$ to $t = K - 1$ and dividing by K,

$$v_\tau^* - \frac{\sum_{t=0}^{K-1} v_\tau^{\pi_t}}{K} = \frac{1}{K} \left(I - \gamma \mathcal{P}_{\pi_\tau^*}\right)^{-1} \sum_{t=0}^{K-1} \left[\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_t} - v_\tau^{\pi_t} \right]$$

(By definition of v_τ^*)

$$\begin{aligned}
\implies v_\tau^* - v_\tau^{\bar{\pi}_K} &= \frac{1}{K}(I - \gamma \mathcal{P}_{\pi_\tau^*})^{-1} \sum_{t=0}^{K-1} \left[\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_t} - v_\tau^{\pi_t} \right] && \text{(Since } v_\tau^{\bar{\pi}_K} = \tfrac{\sum_{t=0}^{K-1} v_\tau^{\pi_t}}{K}) \\
&= \frac{1}{K}\,(I - \gamma \mathcal{P}_{\pi_\tau^*})^{-1} \sum_{t=0}^{K-1} \left[\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_t} - \mathcal{T}_\tau^{\pi_t}\, v_\tau^{\pi_t} \right] && \text{(Since } v_\tau^\pi = \mathcal{T}_\tau^\pi v_\tau^\pi) \\
\implies \left\| v_\tau^* - v_\tau^{\bar{\pi}_K} \right\|_\infty &= \frac{1}{K} \left\| (I - \gamma \mathcal{P}_{\pi_\tau^*})^{-1} \sum_{t=0}^{K-1} \left[\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_t} - \mathcal{T}_\tau^{\pi_t}\, v_\tau^{\pi_t} \right] \right\|_\infty
\end{aligned}$$

$$\leq \frac{1}{K} \left\|(I - \gamma \mathcal{P}_{\pi_\tau^*})^{-1}\right\|_\infty \left\| \sum_{t=0}^{K-1} \left[\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_t} - \mathcal{T}_\tau^{\pi_t} v_\tau^{\pi_t} \right] \right\|_\infty$$

(By definition of matrix norm)

$$\leq \frac{1}{K(1-\gamma)} \left\| \sum_{t=0}^{K-1} \left[\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_t} - \mathcal{T}_\tau^{\pi_t} v_\tau^{\pi_t} \right] \right\|_\infty$$

(Since $\left\|(I - \gamma \mathcal{P}_\pi)^{-1}\right\|_\infty = \|\sum_{t=0}^\infty [\mathcal{P}_\pi]^t\|_\infty \leq \sum_{t=0}^\infty \gamma^t = \frac{1}{1-\gamma}$)

Let us calculate $\left[\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_t} - \mathcal{T}_\tau^{\pi_t} v_\tau^{\pi_t} \right](s)$.

$$\begin{aligned}
\left[\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_t} - \mathcal{T}_\tau^{\pi_t} v_\tau^{\pi_t} \right](s) &= (\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_t})(s) - (\mathcal{T}_\tau^{\pi_t} v_\tau^{\pi_t})(s) \\
&= \mathbb{E}_{a \sim \pi_\tau^*} \left[q_\tau^{\pi_t}(s,a) - \tau \ln(\pi_\tau^*(a|s)) \right] \\
&\quad - \mathbb{E}_{a \sim \pi_t} \left[q_\tau^{\pi_t}(s,a) - \tau \ln(\pi_t(a|s)) \right] && \text{(By definition of } \mathcal{T}_\tau^{\pi_1} v_\tau^{\pi_2}\text{)} \\
&= \langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s), q_\tau^{\pi_t}(s,\cdot) \rangle \\
&\quad - \tau \left[\langle \pi_\tau^*(\cdot|s), \ln(\pi_\tau^*(\cdot|s)) \rangle - \langle \pi_t(\cdot|s), \ln(\pi_t(\cdot|s)) \rangle \right] \\
&= \langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s), q_\tau^{\pi_t}(s,\cdot) \rangle + \tau \left[\mathcal{H}(\pi_\tau^*(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s)) \right] \\
& && \text{(By definition of } \mathcal{H}(\pi(\cdot|s))\text{)} \\
&= \langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s,\cdot) \rangle + \tau \left[\mathcal{H}(\pi_\tau^*(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s)) \right] \\
&\quad + \langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s), q_\tau^{\pi_t}(s,\cdot) - q_\zeta^t(s,\cdot) \rangle \\
& && \text{(Add/Subtract } \langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s,\cdot) \rangle\text{)} \\
&\leq \langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s,\cdot) \rangle + \tau \left[\mathcal{H}(\pi_\tau^*(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s)) \right] \\
&\quad + \|\pi_\tau^*(\cdot|s) - \pi_t(\cdot|s)\|_1 \left\| q_\tau^{\pi_t}(s,\cdot) - q_\zeta^t(s,\cdot) \right\|_\infty \\
& && \text{(By Holder's inequality)} \\
&\leq \langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s,\cdot) \rangle + \tau \left[\mathcal{H}(\pi_\tau^*(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s)) \right] \\
&\quad + 2 \left\| q_\tau^{\pi_t}(s,\cdot) - q_\zeta^t(s,\cdot) \right\|_\infty && \text{(Since } \|\pi_\tau^*(\cdot|s) - \pi_t(\cdot|s)\|_1 \leq 2\text{)}
\end{aligned}$$

Define $\text{Regret}(K,u,s) = \sum_{t=0}^{K-1} \left[\langle u(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s,\cdot) \rangle + \tau \left[\mathcal{H}(u(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s)) \right] \right]$ as the regret incurred for state s when the comparator is policy u. Hence,

$$\sum_{t=0}^{K-1} \left[\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_t} - \mathcal{T}_\tau^{\pi_t} v_\tau^{\pi_t} \right](s) \leq \text{Regret}(K, \pi_\tau^*, s) + 2 \sum_{t=0}^{K} \left\| q_\tau^{\pi_t}(s,\cdot) - q_\zeta^t(s,\cdot) \right\|_\infty$$

$$\left\| \sum_{t=0}^{K-1} \left[\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_t} - \mathcal{T}_\tau^{\pi_t} v_\tau^{\pi_t} \right] \right\|_\infty \leq \max_s \text{Regret}(K, \pi_\tau^*, s) + 2 \sum_{t=0}^{K} \|\boldsymbol{\epsilon}_t\|_\infty \qquad \text{(By definition of } \boldsymbol{\epsilon}_t\text{)}$$

Using the definition of $\text{Regret}(K) = \left[\text{Regret}(K, \pi_\tau^*, s_i)\right]_{i=1}^S \in \mathbb{R}^S$,

$$\left\| v_\tau^* - v_\tau^{\bar{\pi}_K} \right\|_\infty \leq \frac{\|\text{Regret}(K)\|_\infty}{K(1-\gamma)} + \frac{2 \sum_{t \in [K]} \|\boldsymbol{\epsilon}_t\|_\infty}{K(1-\gamma)}$$

□

In the special case $(0, \zeta)$, i.e., no entropy regularization from the actor side, the problem reduces to online linear optimization, as in Politex [Abbasi-Yadkori et al., 2019], yielding the following corollary.

Corollary 3 (Generic Reduction without Actor Entropy)**.** *If π^* is the optimal policy whose value function is equal to v^*, for an estimate of $q_\tau^{\pi_t}$ at iteration t s.t. $\boldsymbol{\epsilon}_t = q_\zeta^t - q_\tau^{\pi_t}$ and for any sequence of policies $\{\pi_0, \pi_1, \ldots, \pi_K\}$, if $\bar{\pi}_K$ is the corresponding mixture policy, then,*

$$\left\| v^* - v^{\bar{\pi}_K} \right\|_\infty \leq \frac{\|\text{Regret}(K)\|_\infty}{K(1-\gamma)} + \frac{2 \sum_{t \in [K]} \|\boldsymbol{\epsilon}_t\|_\infty}{K(1-\gamma)},$$

where $(\text{Regret}(K))(s) := \sum_{t=0}^{K-1} \left[\langle \pi^*(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s,\cdot) \rangle \right]$ is the regret incurred on an online linear optimization problem for each state $s \in \mathcal{S}$.

C.2 Proof of Corollary 1

C.2.1 Generic Policy Evaluation

Theorem 3 (Generic policy evaluation)**.** *Using the policy evaluation scheme in Eq. (1), if* $\delta(\tau,\zeta) := \frac{|\tau-\zeta|\,\ln(A)}{1-\gamma}$*, and* $TV_i := \|\pi_i(\cdot|s) - \pi_{i-1}(\cdot|s)\|_1$*, then, for all* $t \in [K]$*,*

$$\begin{aligned}\epsilon_t := \|\boldsymbol{\epsilon}_t\|_\infty &= \left\|q_\tau^{\pi_t} - q_\zeta^t\right\|_\infty \\ &\leq \frac{H_\tau\,\gamma^m}{1-\gamma}\sum_{i=1}^{t}(\gamma^m)^{t-i}\max_s\left[TV_i + \tau(1-\gamma)\left|\mathcal{H}(\pi_i(\cdot|s)) - \mathcal{H}(\pi_{i-1}(\cdot|s))\right|\right] \\ &\quad + \sum_{i=0}^{t}(\gamma^m)^{t-i}\,\delta(\tau,\zeta)\end{aligned}$$

Furthermore, if for all $i \in [t]$*,* $TV_i \leq \frac{1}{2}$*, then, for any constant* $C \in (0, 1/2)$*,*

$$\begin{aligned}\epsilon_t &\leq \frac{H_\tau\,\gamma^m}{(1-\gamma)}\sum_{i=1}^{t}(\gamma^m)^{t-i}\,\max_s\left[TV_i + \tau(1-\gamma)\,\left[TV_i\,\ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A)}{2} + \sqrt{2}\right)\sqrt{C}\right]\right] \\ &\quad + \sum_{i=0}^{t}(\gamma^m)^{t-i}\,\delta(\tau,\zeta)\end{aligned}$$

Proof.

$$\begin{aligned}\epsilon_t &= \left\|q_\tau^{\pi_t} - q_\zeta^t\right\|_\infty = \left\|q_\tau^{\pi_t} - \mathcal{P}_{[0,H_\tau]}[(T_\zeta^{\pi_t})^m q_\zeta^{t-1}]\right\|_\infty && \text{(Using the update)} \\ &= \left\|\mathcal{P}_{[0,H_\tau]}[q_\tau^{\pi_t}] - \mathcal{P}_{[0,H_\tau]}[(T_\zeta^{\pi_t})^m q_\zeta^{t-1}]\right\|_\infty && \text{(Since } q_\tau^{\pi_t} \in [0,H_\tau]\text{)} \\ &= \left\|\mathcal{P}_{[0,H_\tau]}[(T_\tau^{\pi_t})^m q_\tau^{\pi_t}] - \mathcal{P}_{[0,H_\tau]}[(T_\zeta^{\pi_t})^m q_\zeta^{t-1}]\right\|_\infty && \text{(Since } q_\tau^{\pi_t} = T_\tau^{\pi_t} q_\tau^{\pi_t}\text{)} \\ &\leq \left\|(T_\tau^{\pi_t})^m q_\tau^{\pi_t} - (T_\zeta^{\pi_t})^m q_\zeta^{t-1}\right\|_\infty && \text{(Since projections are non-expansive)} \\ &\leq \left\|(T_\tau^{\pi_t})^m q_\tau^{\pi_t} - (T_\tau^{\pi_t})^m q_\zeta^{t-1}\right\|_\infty + \left\|(T_\tau^{\pi_t})^m q_\zeta^{t-1} - (T_\zeta^{\pi_t})^m q_\zeta^{t-1}\right\|_\infty \\ & && \text{(Add/Subtract } (T_\tau^{\pi_t})^m q_\zeta^{t-1} \text{ and using triangle inequality)} \\ &\leq \left\|(T_\tau^{\pi_t})^m q_\tau^{\pi_t} - (T_\tau^{\pi_t})^m q_\zeta^{t-1}\right\|_\infty + \underbrace{\frac{|\zeta-\tau|}{1-\gamma}\,\ln(A)}_{:=\delta(\tau,\zeta)} && \text{(Using Lemma 3)} \\ &= \left\|(T_\tau^{\pi_t})^m q_\tau^{\pi_t} - (T_\tau^{\pi_t})^m q_\zeta^{t-1}\right\|_\infty + \delta(\tau,\zeta) \\ &\leq \gamma^m\,\left\|q_\tau^{\pi_t} - q_\zeta^{t-1}\right\|_\infty + \delta(\tau,\zeta) && \text{(Since } T_\tau^\pi \text{ is a } \gamma \text{ contraction)} \\ &\leq \gamma^m\,\|q_\tau^{\pi_t} - q_\tau^{\pi_{t-1}}\|_\infty + \gamma^m\,\left\|q_\tau^{\pi_{t-1}} - q_\zeta^{t-1}\right\|_\infty + \delta(\tau,\zeta) \\ & && \text{(Add/subtract } q_\tau^{\pi_{t-1}} \text{ and using triangle inequality)} \\ &= \gamma^m\,\|q_\tau^{\pi_t} - q_\tau^{\pi_{t-1}}\|_\infty + \gamma^m\,\epsilon_{t-1} + \delta(\tau,\zeta)\end{aligned}$$

The first term is the difference in the q_τ functions between two consecutive policies, and we bound it next. For all (s,a),

$$\begin{aligned}q_\tau^{\pi_t}(s,a) - q_\tau^{\pi_{t-1}}(s,a) &= \gamma\sum_{s'}\mathcal{P}(s'|s,a)v_\tau^{\pi_t}(s') - \gamma\sum_{s'}\mathcal{P}(s'|s,a)v_\tau^{\pi_{t-1}}(s') \\ &= \gamma\,\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}[v_\tau^{\pi_t}(s') - v_\tau^{\pi_{t-1}}(s')] \\ \implies \|q_\tau^{\pi_t} - q_\tau^{\pi_{t-1}}\|_\infty &\leq \gamma\,\|v_\tau^{\pi_t} - v_\tau^{\pi_{t-1}}\|_\infty\end{aligned}$$

Let us now bound the difference in the v_τ functions between two consecutive policies.

$$\|v_\tau^{\pi_t} - v_\tau^{\pi_{t-1}}\|_\infty = \|T_\tau^{\pi_t}v_\tau^{\pi_t} - T_\tau^{\pi_t}v_\tau^{\pi_{t-1}} + T_\tau^{\pi_t}v_\tau^{\pi_{t-1}} - T_\tau^{\pi_{t-1}}v_\tau^{\pi_{t-1}}\|_\infty$$

(Since $T_\tau^\pi v_\tau^\pi = v_\tau^\pi$ and add/subtract $T_\tau^{\pi_t}v_\tau^{\pi_{t-1}}$)

$$\leq \|T_\tau^{\pi_t} v_\tau^{\pi_t} - T_\tau^{\pi_t} v_\tau^{\pi_{t-1}}\|_\infty + \|T_\tau^{\pi_t} v_\tau^{\pi_{t-1}} - T_\tau^{\pi_{t-1}} v_\tau^{\pi_{t-1}}\|_\infty \quad \text{(Triangle inequality)}$$

$$\leq \gamma \ \|v_\tau^{\pi_t} - v_\tau^{\pi_{t-1}}\|_\infty + \|T_\tau^{\pi_t} v_\tau^{\pi_{t-1}} - T_\tau^{\pi_{t-1}} v_\tau^{\pi_{t-1}}\|_\infty \quad \text{(Since } T_\tau^\pi \text{ is a } \gamma\text{-contraction)}$$

$$\implies \|v_\tau^{\pi_t} - v_\tau^{\pi_{t-1}}\|_\infty \leq \frac{1}{1-\gamma} \ \|T_\tau^{\pi_t} v_\tau^{\pi_{t-1}} - T_\tau^{\pi_{t-1}} v_\tau^{\pi_{t-1}}\|_\infty$$

In order to bound $\left\|T_\tau^{\pi_t} v_\tau^{\pi_{t-1}} - T_\tau^{\pi_{t-1}} v_\tau^{\pi_{t-1}}\right\|_\infty$, consider a fixed state s. By definition of T_τ^π,

$$\begin{aligned}
T_\tau^{\pi_t} v_\tau^{\pi_{t-1}}(s) - T_\tau^{\pi_{t-1}} v_\tau^{\pi_{t-1}}(s) &= \langle \pi_t(\cdot|s) - \pi_{t-1}(\cdot|s), r(s,\cdot)\rangle \\
&\quad + \gamma \sum_a [\pi_t(a|s) - \pi_{t-1}(a|s)] \, \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)} v_\tau^{\pi_{t-1}}(s') \\
&\quad + \tau[\mathcal{H}(\pi_t(\cdot|s)) - \mathcal{H}(\pi_{t-1}(\cdot|s))] \\
&\leq \|\pi_t(\cdot|s) - \pi_{t-1}\|_1 \ \|r(s,\cdot)\|_\infty \\
&\quad + \gamma \ \|\pi_t(\cdot|s) - \pi_{t-1}\|_1 \ \|v_\tau^{\pi_{t-1}}\|_\infty \\
&\quad + \tau[\mathcal{H}(\pi_t(\cdot|s)) - \mathcal{H}(\pi_{t-1}(\cdot|s))] \quad \text{(By Holder's inequality)} \\
&\leq (1 + \gamma \, H_\tau) \ \|\pi_t(\cdot|s) - \pi_{t-1}(\cdot|s)\|_1 \\
&\quad + \tau|\mathcal{H}(\pi_t(\cdot|s)) - \mathcal{H}(\pi_{t-1}(\cdot|s))| \\
&\qquad \text{(Since rewards are in } [0,1] \text{ and } v_\tau^{\pi_{t-1}}(s) \leq H_\tau) \\
&\leq H_\tau \ \|\pi_t(\cdot|s) - \pi_{t-1}(\cdot|s)\|_1 + \tau|\mathcal{H}(\pi_t(\cdot|s)) - \mathcal{H}(\pi_{t-1}(\cdot|s))| \\
&\qquad \text{(Since } 1 \leq 1 + \tau\ln(A) = (1-\gamma)\,H_\tau)
\end{aligned}$$

$$\implies \|T_\tau^{\pi_t} v_\tau^{\pi_{t-1}} - T_\tau^{\pi_{t-1}} v_\tau^{\pi_{t-1}}\|_\infty \leq H_\tau \ \left(\max_s \left[TV_i + \tau(1-\gamma)\,|\mathcal{H}(\pi_t(\cdot|s)) - \mathcal{H}(\pi_{t-1}(\cdot|s))|\right]\right) \quad \text{(Since } 1 \leq (1-\gamma)\,H_\tau)$$

Combining the above inequalities,

$$\|v_\tau^{\pi_t} - v_\tau^{\pi_{t-1}}\|_\infty \leq \frac{H_\tau}{(1-\gamma)} \ \max_s \left[TV_i + \tau(1-\gamma)\,|\mathcal{H}(\pi_t(\cdot|s)) - \mathcal{H}(\pi_{t-1}(\cdot|s))|\right]$$

$$\implies \|q_\tau^{\pi_t} - q_\tau^{\pi_{t-1}}\|_\infty \leq \frac{H_\tau}{(1-\gamma)} \ \max_s \left[TV_i + \tau(1-\gamma)\,|\mathcal{H}(\pi_t(\cdot|s)) - \mathcal{H}(\pi_{t-1}(\cdot|s))|\right]$$

$$\implies \epsilon_t \leq \underbrace{\frac{H_\tau\,\gamma^m}{(1-\gamma)} \ \max_s \left[TV_i + \tau(1-\gamma)\,|\mathcal{H}(\pi_t(\cdot|s)) - \mathcal{H}(\pi_{t-1}(\cdot|s))|\right]}_{:=B_t} + \gamma^m \,\epsilon_{t-1} + \delta(\tau,\zeta)$$

$$\implies \epsilon_t \leq B_t + \gamma^m \,\epsilon_{t-1} + \delta(\tau,\zeta)$$

Bounding ϵ_0,

$$\epsilon_0 = q_\tau^{\pi_0} - q_\zeta^{\pi_0} = \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[r(s,a) + \gamma v_\tau^\pi(s')\right] - \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[r(s,a) + \gamma v_\zeta^\pi(s')\right] \quad \text{(By definition)}$$

$$\begin{aligned}
&= \gamma \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}[v_\tau^{\pi_0}(s') - v_\zeta^{\pi_0}(s')] \\
&= \gamma \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\Bigg[[v^{\pi_0}(s') + \tau \ \sum_{t=0}^\infty \gamma^t \,[\mathcal{H}(\pi_0(\cdot|s_t))|s_0 = s'] \\
&\qquad - [v^{\pi_0}(s') + \zeta \ \sum_{t=0}^\infty \gamma^t \,[\mathcal{H}(\pi_0(\cdot|s_t))|s_0 = s']]\Bigg] \quad \text{(By definition)} \\
&\leq \frac{|\tau - \zeta| \ \ln(A)}{1-\gamma} = \delta(\tau,\zeta) \quad \text{(Since } \mathcal{H}(\pi(\cdot|s)) \leq \ln(A))
\end{aligned}$$

For a fixed $t \in [K]$, recursing from $i = t-1$ to $i = 1$ and using that

$$\begin{aligned}
\epsilon_t &\leq (\gamma^m)^t \epsilon_0 + \sum_{i=1}^{t} (\gamma^m)^{t-i} \left(B_i + \delta(\tau, \zeta)\right) \\
&\leq \frac{H_\tau \, \gamma^m}{(1-\gamma)} \sum_{i=1}^{t} (\gamma^m)^{t-i} \, \max_s \left[TV_i + \tau(1-\gamma) \left|\mathcal{H}(\pi_i(\cdot|s)) - \mathcal{H}(\pi_{i-1}(\cdot|s))\right|\right] \\
&\quad + \sum_{i=0}^{t} (\gamma^m)^{t-i} \, \delta(\tau, \zeta)
\end{aligned}$$

Furthermore, if $TV_i \leq \frac{1}{2}$ for all $i \in [t]$, using Lemma 1, we can further upper-bound $|\mathcal{H}(\pi_i(\cdot|s)) - \mathcal{H}(\pi_{i-1}(\cdot|s))|$ to get that for any constant $C \in (0, 1/2)$,

$$\left|\mathcal{H}(\pi_i(\cdot|s)) - \mathcal{H}(\pi_{i-1}(\cdot|s))\right| \leq TV_i \, \ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A-1)}{2} + \sqrt{2}\right) \sqrt{C}$$

Combining the above relations, in this case, we get that,

$$\begin{aligned}
\epsilon_t \leq \frac{H_\tau \, \gamma^m}{(1-\gamma)} \sum_{i=1}^{t} (\gamma^m)^{t-i} \, \max_s \left[TV_i + \tau(1-\gamma) \left[TV_i \, \ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A)}{2} + \sqrt{2}\right) \sqrt{C}\right]\right] \\
+ \sum_{i=0}^{t} (\gamma^m)^{t-i} \, \delta(\tau, \zeta)
\end{aligned}$$

□

Corollary 4. *Using the policy evaluation scheme in Eq. (1) with $\zeta = \tau$, for all $t \in [K]$, if for all $i \in [t]$, $TV_i := \|\pi_i(\cdot|s) - \pi_{i-1}(\cdot|s)\|_1 \leq \frac{1}{2}$, then, for any constant $C \in (0, 1/2)$,*

$$\epsilon_t \leq \frac{H_\tau \, \gamma^m}{(1-\gamma)} \sum_{i=1}^{t} (\gamma^m)^{t-i} \, \max_s \left[TV_i + \tau(1-\gamma) \left[TV_i \, \ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A)}{2} + \sqrt{2}\right) \sqrt{C}\right]\right]$$

Proof. Setting $\zeta = \tau$ in Theorem 3. □

Corollary 5. *Using the policy evaluation scheme in Eq. (1) with $\zeta = 0$, for all $t \in [K]$, if for all $i \in [t]$, $TV_i := \|\pi_i(\cdot|s) - \pi_{i-1}(\cdot|s)\|_1 \leq \frac{1}{2}$, then, for any constant $C \in (0, 1/2)$,*

$$\begin{aligned}
\epsilon_t \leq \frac{H_\tau \, \gamma^m}{(1-\gamma)} \sum_{i=1}^{t} (\gamma^m)^{t-i} \, \max_s \left[TV_i + \tau(1-\gamma) \left[TV_i \, \ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A)}{2} + \sqrt{2}\right) \sqrt{C}\right]\right] \\
+ \frac{\tau \, \ln(A)}{(1-\gamma)^2}
\end{aligned}$$

Proof. Setting $\zeta = 0$ in Theorem 3. □

We note that the reverse KL-based objective in Eq. (5b) admits a closed-form solution, as established in the following proposition.

Proposition 1. *If $\alpha_t := \frac{1}{1+\tau_t}$, the closed-form solution for the proximal update in Eq. (5b) for any s, a is given as,*

$$\pi_{t+1}(a|s) = \frac{[\pi_{t+1/2}(a|s)]^{\alpha_t}}{\sum_{a'} [\pi_{t+1/2}(a'|s)]^{\alpha_t}} \tag{6}$$

Proof.

$$\begin{aligned}
\pi_{t+1}(\cdot|s) &= \arg\min_{\pi\in\Delta} \left[\text{KL}(\pi||\pi_{t+1/2}(\cdot|s)) - \tau_t \mathcal{H}(\pi)\right] \\
&= \arg\min_{\pi\in\Delta} \mathbb{E}_{a\sim\pi(.|s)} \left[(1+\tau_t)\, \ln(\pi(.|s)) - \ln(\pi_{t+1/2}(\cdot|s))\right] \\
& \qquad \text{(Using the definition of } H(\pi(\cdot|s))) \\
&= \arg\min_{\pi\in\Delta} \mathbb{E}_{a\sim\pi(.|s)} \left[\ln(\pi(.|s)) - \alpha_t\, \ln(\pi_{t+1/2}(\cdot|s))\right] \qquad \text{(Since } \alpha_t = \tfrac{1}{1+\tau_t}) \\
&= \arg\min_{\pi\in\Delta} \mathbb{E}_{a\sim\pi(.|s)} \left[\ln(\pi(.|s)) - \ln([\pi_{t+1/2}(\cdot|s)]^{\alpha_t})\right] \\
&= \arg\min_{\pi\in\Delta} \text{KL}(\pi(.|s)||\pi_{t+1/2}(\cdot|s)^{\alpha_t}) \qquad \text{(By definition of the KL divergence)}
\end{aligned}$$

Using the fact that KL projection onto the simplex results in normalization, we get that,

$$\pi_{t+1}(a|s) = \frac{[\pi_{t+1/2}(a|s)]^{\alpha_t}}{\sum_{a'} [\pi_{t+1/2}(a'|s)]^{\alpha_t}}$$

□

Note that when $\tau = 0$, we have $\tau_t = 0$ and $\alpha_t = 1$ for all t, recovering the standard unregularized updates for both `NPG` and `SPMA`. For $\tau > 0$, the soft updates can be expressed as: for any s, a, with $\alpha_t = \frac{1}{1+\tau_t}$ and $\tau_t = \eta_t\, \tau$,

$$\pi_{t+1}(a|s) = \frac{[\pi_t(a|s)]^{\alpha_t}\, \exp(\eta_t\, \alpha_t\, q^t_\zeta(s,a))}{\mathcal{Z}_t}, \tag{7}$$

$$\text{with} \quad \mathcal{Z}_t = \sum_{a'} [\pi_t(a'|s)]^{\alpha_t}\, \exp(\eta_t\, \alpha_t\, q^t_\zeta(s,a')), \qquad \textbf{(Soft-NPG)}$$

$$\pi_{t+1}(a|s) = \frac{[\pi_t(a|s)]^{\alpha_t} \left[1 + \eta_t \left(q^t_\zeta(s,a) - v^t_\zeta(s)\right)\right]^{\alpha_t}}{\mathcal{Z}_t}, \tag{8}$$

$$\text{with} \quad \mathcal{Z}_t = \sum_{a'} [\pi_t(a'|s)]^{\alpha_t} \left[1 + \eta_t \left(q^t_\zeta(s,a') - v^t_\zeta(s)\right)\right]^{\alpha_t}, \qquad \textbf{(Soft-SPMA)}$$

C.2.2 Policy Error Bound for (soft) NPG

In the next corollary, we use Theorem 3 with $\zeta = \tau$ to instantiate the policy error bound for soft `NPG` with entropy-regularized policy evaluation.

Corollary 6 (Policy evaluation with $\zeta = \tau$)**.** *Using the policy evaluation update in Eq. (1) with $\zeta = \tau$, for soft NPG with $\eta_t = \frac{1}{c+\tau\,(t+1)}$ for a constant $c \geq \max\left\{\frac{8\,(1+\tau\,\log(A))}{(1-\gamma)}, 32\,\tau\, \ln(A), \frac{2\,(1+\tau\,\ln(A))^2}{(1-\gamma)^2\,\tau\ln(A)}\right\}$, for all $t \in [K]$, ϵ_t can be bounded as:*

$$\epsilon_t := \|\boldsymbol{\epsilon}_t\|_\infty \leq \frac{4(1+\tau\log(A))\,\gamma^m}{(1-\gamma)^3} \left[(1+\tau\, \ln(A\,K))\, \sqrt{\ln(A)}\, (\frac{1}{\sqrt{t}} + (\gamma^m)^{t/2}) + \tau\, (\ln(A)+1)\, \frac{1}{\sqrt{K}}\right]$$

Proof. For a fixed iteration $t \in [K]$ and state $s \in \mathcal{S}$, let us first bound $\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1$.

$$\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \leq \left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1 + \left\|\pi_{t+1/2}(\cdot|s) - \pi_t(\cdot|s)\right\|_1$$

(Triangle inequality)

We first bound $\left\|\pi_{t+1/2}(\cdot|s) - \pi_t(\cdot|s)\right\|_1$. Using the mirror descent view of `NPG` [Xiao, 2022], the update can be written as:

$$\pi_{t+1/2}(\cdot|s) = \arg\min_{\pi\in\Delta} \left[-\eta_t \langle q^t_\tau(s,\cdot), \pi(\cdot|s)\rangle + \text{KL}(\pi(\cdot|s)||\pi_t(\cdot|s))\right]$$

$$\implies -\eta_t \langle q^t_\tau(s,\cdot), \pi_{t+1/2}(\cdot|s)\rangle + \text{KL}(\pi_{t+1/2}(\cdot|s)||\pi_t(\cdot|s)) \leq -\eta_t \langle q^t_\tau(s,\cdot), \pi_t(\cdot|s)\rangle + \text{KL}(\pi_t(\cdot|s)||\pi_t(\cdot|s))$$

$$\implies \frac{1}{2}\left\|\pi_{t+1/2}(\cdot|s) - \pi_t(\cdot|s)\right\|_1^2 \leq \text{KL}(\pi_{t+1/2}(\cdot|s)||\pi_t(\cdot|s)) \leq \eta_t \langle q^t_\tau(s,\cdot), \pi_{t+1/2}(\cdot|s) - \pi_t(\cdot|s)\rangle \quad \text{(By Pinsker's inequality)}$$

$$\leq \eta_t \left\|q^t_\tau(s,\cdot)\right\|_\infty \left\|\pi_{t+1/2}(\cdot|s) - \pi_t(\cdot|s)\right\|_1 \quad \text{(By Holder's inequality)}$$

$$\implies \left\|\pi_{t+1/2}(\cdot|s) - \pi_t(\cdot|s)\right\|_1 \leq 2\eta_t H_\tau \quad \text{(Since } \|q^t_\tau(s,\cdot)\|_\infty \leq H_\tau\text{)}$$

In order to bound $\left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1$, we use the Eq. (5b) update. Specifically,

$$\pi_{t+1}(\cdot|s) = \arg\min_{\pi\in\Delta}\left[\text{KL}(\pi(\cdot|s)||\pi_{t+1/2}(\cdot|s)) - \tau_t \mathcal{H}(\pi(\cdot|s))\right]$$

$$\implies \text{KL}(\pi_{t+1}(\cdot|s)||\pi_{t+1/2}(\cdot|s)) - \tau_t\mathcal{H}(\pi_{t+1}(\cdot|s)) \leq \text{KL}(\pi_{t+1/2}(\cdot|s)||\pi_{t+1/2}(\cdot|s)) - \tau_t\mathcal{H}(\pi_{t+1/2}(\cdot|s))$$

$$\frac{1}{2}\left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1^2 \leq \text{KL}(\pi_{t+1}||\pi_{t+1/2}(\cdot|s)) \quad \text{(By Pinsker's inequality)}$$

$$\leq \tau_t\mathcal{H}(\pi_{t+1}(\cdot|s)) - \tau_t\mathcal{H}(\pi_{t+1/2}(\cdot|s))$$

$$\leq \tau_t \ln(A) \qquad \text{(Since } \mathcal{H}(\pi) \in [0, \ln(A)]\text{)}$$

$$\implies \left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1 \leq \sqrt{2\,\tau\,\eta_t\,\ln(A)} \qquad \text{(Since } \tau_t = \tau\,\eta_t\text{)}$$

Combining the above relations,

$$\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \leq \sqrt{2\,\tau\,\eta_t\,\ln(A)} + 2\,\eta_t\,H_\tau \tag{9}$$

Using Eq. (9) in Theorem 3, for the special case $\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \leq \frac{1}{2}$ for all $t \in [K]$, η_t must satisfy the following conditions:

- $2\,\eta_t\,H_\tau \leq \frac{1}{4} \implies \eta_t \leq \frac{(1-\gamma)}{8\,(1+\tau\,\log(A))}$
- $\sqrt{2\,\tau\,\eta_t\,\ln(A)} \leq \frac{1}{4} \implies \eta_t \leq \frac{1}{32\,\tau\,\ln(A)}$

For $\eta_t = \frac{1}{c+\tau(t+1)} \leq \frac{1}{c}$, it is thus sufficient to ensure that,

$$c \geq \max\left\{\frac{8\,(1+\tau\,\log(A))}{(1-\gamma)}, 32\,\tau\,\ln(A)\right\}$$

Given this constraint on c, we use Theorem 3 with $\zeta = \tau$ and $TV_i := \|\pi_i(\cdot|s) - \pi_{i-1}(\cdot|s)\|_1$, to get that, for any constant $C \in (0, 1/2)$, for all $t \in [K]$,

$$\epsilon_t \leq \frac{H_\tau\,\gamma^m}{(1-\gamma)}\sum_{i=1}^{t}(\gamma^m)^{t-i}\max_s\left[TV_i + \tau(1-\gamma)\left[TV_i\,\ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A)}{2} + \sqrt{2}\right)\sqrt{C}\right]\right]$$

In order to bound $\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1$, from Eq. (9), we know that,

$$\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \leq \sqrt{2\,\tau\,\eta_t\,\ln(A)} + 2\,\eta_t\,H_\tau$$

For $c \geq \frac{2\,(1+\tau\,\ln(A))^2}{(1-\gamma)^2\,\tau\,\ln(A)}$, we obtain $2\eta_t H_\tau \leq \sqrt{2\tau\,\eta_t\,\ln(A)}$. Hence, using $c \geq \max\left\{\frac{8\,(1+\tau\,\log(A))}{(1-\gamma)}, 32\,\tau\,\ln(A), \frac{2\,(1+\tau\,\ln(A))^2}{(1-\gamma)^2\,\tau\,\ln(A)}\right\}$, combined with the above inequalities, we get,

$$\epsilon_t \leq \frac{H_\tau\,\gamma^m}{(1-\gamma)}\sum_{i=1}^{t}(\gamma^m)^{t-i}\left[\left(1 + \tau\,(1-\gamma)\,\ln\left(\frac{A}{C}\right)\right)\left(2\sqrt{2\,\tau\,\eta_i\,\ln(A)}\right) + \tau(1-\gamma)\left(\frac{\ln(A)}{2} + \sqrt{2}\right)\sqrt{C}\right]$$

Setting $C = \frac{1}{K}$,

$$\leq \frac{H_\tau\, \gamma^m}{(1-\gamma)} \Bigg[(1 + \tau\, (1-\gamma)\, \ln{(A\, K)})\, \left(2\, \sqrt{2\, \ln(A)}\right) \sum_{i=1}^{t} \frac{(\gamma^m)^{t-i}}{\sqrt{(i+1)}} + \frac{\tau(1-\gamma)}{1-\gamma^m} \left(\frac{\ln(A)}{2} + \sqrt{2}\right) \frac{1}{\sqrt{K}} \Bigg]$$

$$\leq \frac{H_\tau\, \gamma^m}{(1-\gamma)} \Bigg[(1 + \tau\, (1-\gamma)\, \ln{(A\, K)})\, \left(2\, \sqrt{2\, \ln(A)}\right) \sum_{i=1}^{t} \frac{(\gamma^m)^{t-i}}{\sqrt{(i+1)}} + \tau \left(\frac{\ln(A)}{2} + \sqrt{2}\right) \frac{1}{\sqrt{K}} \Bigg] \qquad \text{(Since } \gamma < 1\text{)}$$

Using Lemma 2, we can bound $\sum_{i=1}^{t} \frac{(\gamma^m)^{t-i}}{\sqrt{(i+1)}} \leq \frac{\sqrt{2}}{1-\gamma^m} \left(\frac{1}{\sqrt{t}} + (\gamma^m)^{t/2}\right) \leq \frac{\sqrt{2}}{1-\gamma} \left(\frac{1}{\sqrt{t}} + (\gamma^m)^{t/2}\right)$.

$$\implies \epsilon_t \leq \frac{H_\tau\, \gamma^m}{(1-\gamma)} \Bigg[(1 + \tau\, (1-\gamma)\, \ln{(A\, K)}) \left(\frac{4\, \sqrt{\ln(A)}}{1-\gamma}\right) \left(\frac{1}{\sqrt{t}} + (\gamma^m)^{t/2}\right) + \tau \left(\frac{\ln(A)}{2} + \sqrt{2}\right) \frac{1}{\sqrt{K}} \Bigg]$$

$$\leq \frac{4 H_\tau\, \gamma^m}{(1-\gamma)^2} \left[(1 + \tau\, \ln{(A\, K)})\, \sqrt{\ln(A)} \left(\frac{1}{\sqrt{t}} + (\gamma^m)^{t/2}\right) + \tau\, (\ln(A) + 1)\, \frac{1}{\sqrt{K}} \right] \qquad \text{(Since } \gamma < 1\text{)}$$

□

In the next corollary, we use Theorem 3 with $\zeta = 0$ to instantiate the policy error bound for (soft) NPG with entropy-regularized policy evaluation.

Corollary 7 (Policy evaluation with $\zeta = 0$)**.** *Using the policy evaluation update in Eq. (1) with $\zeta = 0$, ϵ_t can be bounded as:*

- ***Soft NPG**: If $\eta_t = \frac{1}{c + \tau\, (t+1)}$ for a constant $c \geq \max\left\{\frac{8\,(1+\tau\, \log(A))}{(1-\gamma)}, 32\, \tau\, \ln(A), \frac{2\,(1+\tau\, \ln(A))^2}{(1-\gamma)^2\, \tau \ln(A)}\right\}$, then, for all $t \in [K]$,*

$$\epsilon_t := \|\boldsymbol{\epsilon}_t\|_\infty \leq \frac{4(1 + \tau \ln(A))\, \gamma^m}{(1-\gamma)^3} \left[(1 + \tau\, \ln{(A\, K)})\, \sqrt{\ln(A)} \left(\frac{1}{\sqrt{t}} + (\gamma^m)^{t/2}\right) + \tau\, (\ln(A) + 1)\, \frac{1}{\sqrt{K}} \right] + \frac{\tau}{(1-\gamma)^2} \ln{(A)}$$

- ***NPG**: If $\eta_t = \eta = \frac{(1-\gamma)\, \sqrt{\ln(A)}}{\sqrt{K}}$, then, for all $t \in [K]$,*

$$\epsilon_t := \|\boldsymbol{\epsilon}_t\|_\infty \leq \frac{2\, \sqrt{\ln(A)}\, \gamma^m}{(1-\gamma)^3} \frac{1}{\sqrt{K}}$$

Proof. Using Theorem 3 for soft NPG and Corollary 5 for NPG, and following the same proof as Corollary 6. □

C.2.3 Policy Error Bound for (soft) SPMA

In the next corollary, we use Theorem 3 with $\zeta = \tau$ to instantiate the policy error bound for soft SPMA with entropy-regularized policy evaluation.

Corollary 8 (Policy evaluation with $\zeta = \tau$). *Using the policy evaluation update in Eq. (1) with $\zeta = \tau$, for soft SPMA, ϵ_t can be bounded as: if $\eta_t = \frac{1}{c+\tau\,(t+1)}$ for a constant $c \geq \max\left\{\frac{4(1+\tau\,\ln(A))}{(1-\gamma)}, 32\,\tau\,\ln(A), \frac{(1+\tau\,\ln(A))^2}{(1-\gamma)^2\,2\tau\ln(A)}\right\}$, then, for all $t \in [K]$,*

$$\epsilon_t := \|\boldsymbol{\epsilon}_t\|_\infty \leq \frac{4(1+\tau\,\log(A))\,\gamma^m}{(1-\gamma)^3}\Bigg[(1+\tau\,\ln(A\,K))\,\sqrt{\ln(A)}\,(\frac{1}{\sqrt{t}}+(\gamma^m)^{t/2}) + \tau\,(\ln(A)+1)\,\frac{1}{\sqrt{K}}\Bigg]$$

Proof. For a fixed iteration $t \in [K]$ and state $s \in \mathcal{S}$, let us first bound $\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1$.

$$\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \leq \left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1 + \left\|\pi_{t+1/2}(\cdot|s) - \pi_t(\cdot|s)\right\|_1 \quad \text{(Triangle inequality)}$$

$$\leq \left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1 + \eta_t \sum_a \pi_t(a)\left[q^t_\tau(s,a) - v^t_\tau(s)\right] \quad \text{(By the SPMA update in Eq. (4))}$$

$$\implies \|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \leq \left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1 + \eta_t H_\tau \quad \text{(Since } |q^t_\tau(s,a) - v^t_\tau(s)| \leq H_\tau\text{)}$$

In order to bound $\left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1$, we use the Eq. (5b) update. Specifically,

$$\pi_{t+1}(\cdot|s) = \arg\min_{\pi\in\Delta}\left[\text{KL}(\pi||\pi_{t+1/2}(\cdot|s)) - \tau_t\mathcal{H}(\pi(\cdot|s))\right]$$

$$\implies \text{KL}(\pi_{t+1}(\cdot|s)||\pi_{t+1/2}(\cdot|s)) - \tau_t\mathcal{H}(\pi_{t+1}(\cdot|s)) \leq \text{KL}(\pi_{t+1/2}(\cdot|s)||\pi_{t+1/2}(\cdot|s)) - \tau_t\mathcal{H}(\pi_{t+1/2}(\cdot|s))$$

$$\frac{1}{2}\left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1^2 \leq \text{KL}(\pi_{t+1}(\cdot|s)||\pi_{t+1/2}(\cdot|s)) \quad \text{(By Pinsker's inequality)}$$

$$\leq \tau_t\mathcal{H}(\pi_{t+1}(\cdot|s)) - \tau_t\mathcal{H}(\pi_{t+1/2}(\cdot|s))$$

$$\leq \tau_t\ln(A) \qquad \text{(Since } \mathcal{H}(\pi) \in [0, \ln(A)]\text{)}$$

$$\implies \left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1 \leq \sqrt{2\,\tau\,\eta_t\,\ln(A)} \qquad \text{(Since } \tau_t = \tau\,\eta_t\text{)}$$

Combining the above relations,

$$\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \leq \sqrt{2\,\tau\,\eta_t\,\ln(A)} + \eta_t\,H_\tau \tag{10}$$

In order to use Theorem 3, we need to ensure that $\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \leq \frac{1}{2}$ for all $t \in [K]$. Using Eq. (10), it is sufficient to ensure that η_t satisfies the following relations:

- $\eta_t\,H_\tau \leq \frac{1}{4} \implies \eta_t \leq \frac{(1-\gamma)}{4\,(1+\tau\,\log(A))}$
- $\sqrt{2\,\tau\,\eta_t\,\ln(A)} \leq \frac{1}{4} \implies \eta_t \leq \frac{1}{32\,\tau\,\ln(A)}$

For $\eta_t = \frac{1}{c+\tau(t+1)} \leq \frac{1}{c}$, it is thus sufficient to ensure that,

$$c \geq \max\left\{\frac{4\,(1+\tau\,\log(A))}{(1-\gamma)}, 32\,\tau\,\ln(A)\right\}$$

Given this constraint on c, we use Theorem 3 with $\zeta = \tau$ and $TV_i := \|\pi_i(\cdot|s) - \pi_{i-1}(\cdot|s)\|_1$, to get that, for any constant $C \in (0, 1/2)$, for all $t \in [K]$,

$$\epsilon_t \leq \frac{H_\tau\,\gamma^m}{(1-\gamma)}\sum_{i=1}^t(\gamma^m)^{t-i}\max_s\left[TV_i + \tau(1-\gamma)\left(TV_i\,\ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A)}{2}+\sqrt{2}\right)\sqrt{C}\right)\right]$$

In order to bound $\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1$, from Eq. (10), we know that,

$$\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \leq \sqrt{2\,\tau\,\eta_t\,\ln(A)} + \eta_t\,H_\tau$$

For $c \geq \frac{(1+\tau\,\ln(A))^2}{(1-\gamma)^2\,2\tau\ln(A)}$, we have $\eta_t H_\tau \leq \sqrt{2\tau\,\eta_t\,\ln(A)}$. Hence, with $c \geq \max\left\{\frac{4(1+\tau\,\log(A))}{(1-\gamma)}, 32\,\tau\,\ln(A), \frac{(1+\tau\,\ln(A))^2}{(1-\gamma)^2\,2\tau\ln(A)}\right\}$, combining the above inequalities, we get that,

$$\epsilon_t \leq \frac{H_\tau\,\gamma^m}{(1-\gamma)}\sum_{i=1}^{t}(\gamma^m)^{t-i}\left[\left(1+\tau\,(1-\gamma)\,\ln\left(\frac{A}{C}\right)\right)\left(2\,\sqrt{2\,\tau\,\eta_i\,\ln(A)}\right) + \tau(1-\gamma)\left(\frac{\ln(A)}{2}+\sqrt{2}\right)\sqrt{C}\right]$$

Setting $C = \frac{1}{K}$,

$$\leq \frac{H_\tau\,\gamma^m}{(1-\gamma)}\left[(1+\tau\,(1-\gamma)\,\ln(A\,K))\left(2\,\sqrt{2\,\ln(A)}\right)\sum_{i=1}^{t}\frac{(\gamma^m)^{t-i}}{\sqrt{(i+1)}} + \tau\left(\frac{\ln(A)}{2}+\sqrt{2}\right)\frac{1}{\sqrt{K}}\right]$$

Using Lemma 2, we can bound $\sum_{i=1}^{t}\frac{(\gamma^m)^{t-i}}{\sqrt{(i+1)}} \leq \frac{\sqrt{2}}{1-\gamma^m}\left(\frac{1}{\sqrt{t}} + (\gamma^m)^{t/2}\right) \leq \frac{\sqrt{2}}{1-\gamma}\left(\frac{1}{\sqrt{t}} + (\gamma^m)^{t/2}\right)$.

$$\implies \epsilon_t \leq \frac{H_\tau\,\gamma^m}{(1-\gamma)}\left[(1+\tau\,(1-\gamma)\,\ln(A\,K))\left(\frac{4\,\sqrt{\ln(A)}}{1-\gamma}\right)\left(\frac{1}{\sqrt{t}} + (\gamma^m)^{t/2}\right) + \tau\left(\frac{\ln(A)}{2}+\sqrt{2}\right)\frac{1}{\sqrt{K}}\right]$$

$$\leq \frac{4H_\tau\,\gamma^m}{(1-\gamma)^2}\left[(1+\tau\,\ln(A\,K))\,\sqrt{\ln(A)}\left(\frac{1}{\sqrt{t}} + (\gamma^m)^{t/2}\right) + \tau\,(\ln(A)+1)\,\frac{1}{\sqrt{K}}\right]$$

□

In the next corollary, we use Theorem 3 with $\zeta = 0$ to instantiate the policy error bound for (soft) `SPMA` without entropy-regularized policy evaluation.

Corollary 9 (Policy evaluation with $\zeta = 0$)**.** *Using the policy evaluation update in Eq.* (1) *with (soft)* `SPMA`*, ϵ_t can be bounded as:*

- ***Soft*** ***`SPMA`****: if* $\eta_t = \frac{1}{c+\tau\,(t+1)}$ *for a constant* $c \geq \max\left\{\frac{4(1+\tau\,\log(A))}{(1-\gamma)}, 32\,\tau\,\ln(A), \frac{(1+\tau\,\ln(A))^2}{(1-\gamma)^2\,2\tau\ln(A)}\right\}$*, then, for all* $t \in [K]$*,*

$$\epsilon_t := \|\boldsymbol{\epsilon}_t\|_\infty \leq \frac{4(1+\tau\,\log(A))\,\gamma^m}{(1-\gamma)^3}\left[(1+\tau\,\ln(A\,K))\,\sqrt{\ln(A)}\left(\frac{1}{\sqrt{t}} + (\gamma^m)^{t/2}\right) + \tau\,(\ln(A)+1)\,\frac{1}{\sqrt{K}}\right] + \frac{\tau}{(1-\gamma)^2}\ln(A)$$

- ***`SPMA`****: if* $\eta_t = \eta = \min\left\{\frac{1-\gamma}{2}, \frac{(1-\gamma)\,\sqrt{\ln(A)}}{\sqrt{K}}\right\}$*, then, for all* $t \in [K]$*,*

$$\epsilon_t := \|\boldsymbol{\epsilon}_t\|_\infty \leq \frac{\sqrt{\ln(A)}\,\gamma^m}{(1-\gamma)^3}\,\frac{1}{\sqrt{K}}$$

Proof. Using Theorem 3 for soft `SPMA` and Corollary 5 for `SPMA`, and following the same proof as Corollary 8. Note that the requirement for $\eta \leq \frac{1-\gamma}{2}$ is to ensure $\ln(1+\eta_t\,\Delta^t(s,\cdot))$ in the `SPMA` update is well-defined (see the proof in Corollary 11 for details) □

C.3 Proof of Corollary 2

C.3.1 Generic Regret Bound

Theorem 4 (Generic Regret Bound)**.** *Consider a sequence of linear functions $f_t(\pi) := \langle \pi, d_t \rangle$ for a sequence of vectors $\{d_0, d_1, \ldots, d_{K-1}\}$ s.t. $\|d_t\|_\infty \leq D_t$. Consider the following update at iteration $t \in [K]$, if η_t is a step-size sequence, $\tau_t = \eta_t \, \tau$, π_0 is the uniform distribution and*

$$\pi_{t+1} = \arg\min_{\pi \in \Delta_A} \Big\{ \langle \pi, d_t \rangle + KL(\pi || \pi_t) + \mathcal{R}_t(\pi) \Big\}$$
$$\mathcal{R}_t(\pi) := \tau_t \, \mathcal{R}(\pi)$$
$$\mathcal{R}(\pi) := \ln(A) - \mathcal{H}(\pi) \geq 0$$

then, for any comparator $u \in \Delta_A$,

$$\sum_{t=0}^{K-1} \left[\frac{\langle \pi_t - u, \, d_t \rangle}{\eta_t} + \tau \left[\mathcal{H}(u) - \mathcal{H}(\pi_t) \right] \right] \leq \sum_{t=0}^{K-1} \left[\frac{KL(u||\pi_t)}{\eta_t} - \frac{KL(u||\pi_{t+1})}{\eta_t} - \tau KL(u||\pi_{t+1}) \right] + \sum_{t=0}^{K-1} \frac{D_t^2}{2 \, \eta_t}$$

Proof. The following properties will be helpful in proving the theorem. For policies π, π' and comparator u,

$$\mathcal{R}_t(\pi) - \mathcal{R}_t(\pi') = \tau_t \, \langle \ln(\pi), \pi - \pi' \rangle - \tau_t \, \text{KL}(\pi' || \pi) \qquad \text{(Entropy property)}$$
$$\langle u - \pi', \ln(\pi') - \ln(\pi) \rangle = \text{KL}(u||\pi) - \text{KL}(u||\pi') - \text{KL}(\pi'||\pi) \qquad \text{(3 point property)}$$
$$\langle \pi - \pi_{t+1}, d_t + \ln(\pi_{t+1}) - \ln(\pi_t) + \tau_t \ln(\pi_{t+1}) \rangle \geq 0 \qquad \text{(Optimality condition)}$$

$$\begin{aligned}
[f_t(\pi_t) - f_t(u)] + \mathcal{R}_t(\pi_{t+1}) - \mathcal{R}_t(u) &= \langle \pi_t - u, \, d_t \rangle + \langle \tau_t \ln(\pi_{t+1}), \, \pi_{t+1} - u \rangle \\
&\quad - \tau_t \text{KL}(u||\pi_{t+1}) \\
&\qquad \text{(Entropy property with } \pi = \pi_{t+1}, \pi' = u) \\
&= \langle \pi_{t+1} - u, \, d_t \rangle + \langle \pi_t - \pi_{t+1}, \, d_t \rangle \\
&\quad + \langle \tau_t \ln(\pi_{t+1}), \, \pi_{t+1} - u \rangle - \tau_t \text{KL}(u||\pi_{t+1}) \\
&= \underbrace{\langle \pi_{t+1} - u, \, d_t + \tau_t \ln(\pi_{t+1}) - \ln(\pi_t) + \ln(\pi_{t+1}) \rangle}_{\leq 0 \text{ by the optimality condition for } \pi = u} \\
&\quad + \langle \pi_{t+1} - u, \, \ln(\pi_t) - \ln(\pi_{t+1}) \rangle + \langle \pi_t - \pi_{t+1}, \, d_t \rangle \\
&\quad - \tau_t \text{KL}(u||\pi_{t+1}) \qquad \text{(Dropping the negative term)} \\
&\leq \langle \pi_{t+1} - u, \, \ln(\pi_t) - \ln(\pi_{t+1}) \rangle + \langle \pi_t - \pi_{t+1}, \, d_t \rangle \\
&\quad - \tau_t \text{KL}(u||\pi_{t+1}) \\
&= \text{KL}(u||\pi_t) - \text{KL}(u||\pi_{t+1}) - \text{KL}(\pi_{t+1}||\pi_t) \\
&\quad + \langle \pi_t - \pi_{t+1}, \, d_t \rangle - \tau_t \text{KL}(u||\pi_{t+1}) \\
&\qquad \text{(3 point property with } u = u, \pi = \pi_t, \pi' = \pi_{t+1}) \\
&\leq \text{KL}(u||\pi_t) - \text{KL}(u||\pi_{t+1}) - \text{KL}(\pi_{t+1}||\pi_t) \\
&\quad - \tau_t \text{KL}(u||\pi_{t+1}) + \frac{1}{2} ||\pi_t - \pi_{t+1}||_1^2 + \frac{1}{2} ||d_t||_\infty^2 \\
&\qquad \text{(Fenchel-Young inequality)} \\
&\leq \text{KL}(u||\pi_t) - \text{KL}(u||\pi_{t+1}) - \text{KL}(\pi_{t+1}||\pi_t) \\
&\quad - \tau_t \text{KL}(u||\pi_{t+1}) + \text{KL}(\pi_{t+1}||\pi_t) + \frac{1}{2} ||d_t||_\infty^2 \\
&\qquad \text{(Pinsker's inequality)} \\
&= \text{KL}(u||\pi_t) - \text{KL}(u||\pi_{t+1}) - \tau_t \text{KL}(u||\pi_{t+1}) + \frac{1}{2} ||d_t||_\infty^2
\end{aligned}$$

$$\leq \mathrm{KL}(u||\pi_t) - \mathrm{KL}(u||\pi_{t+1}) - \tau_t \mathrm{KL}(u||\pi_{t+1}) + \frac{D_t^2}{2}$$
(Since $\|d_t\|_\infty \leq D_t$)

$$= \mathrm{KL}(u||\pi_t) - \mathrm{KL}(u||\pi_{t+1}) - \eta_t \tau \mathrm{KL}(u||\pi_{t+1}) + \frac{D_t^2}{2}$$
(Since $\tau_t = \eta_t\, \tau$)

Rearranging and dividing both-sides by η_t we get

$$\frac{\langle \pi_t - u,\, d_t \rangle}{\eta_t} + \mathcal{R}(\pi_{t+1}) - \mathcal{R}(u) \leq \frac{\mathrm{KL}(u||\pi_t)}{\eta_t} - \frac{\mathrm{KL}(u||\pi_{t+1})}{\eta_t} - \tau \mathrm{KL}(u||\pi_{t+1}) + \frac{D_t^2}{2\,\eta_t}$$

$$\frac{\langle \pi_t - u,\, d_t \rangle}{\eta_t} + \mathcal{R}(\pi_t) - \mathcal{R}(u) \leq [\mathcal{R}(\pi_t) - \mathcal{R}(\pi_{t+1})] + \frac{\mathrm{KL}(u||\pi_t)}{\eta_t} - \frac{\mathrm{KL}(u||\pi_{t+1})}{\eta_t} - \tau \mathrm{KL}(u||\pi_{t+1}) + \frac{D_t^2}{2\,\eta_t}$$

Summing from $t = 0$ to $K - 1$,

$$\sum_{t=0}^{K-1} \left[\frac{\langle \pi_t - u,\, d_t \rangle}{\eta_t} + \mathcal{R}(\pi_t) - \mathcal{R}(u)\right] \leq [\mathcal{R}(\pi_0) - \mathcal{R}(\pi_K)] + \sum_{t=0}^{K-1} \left[\frac{\mathrm{KL}(u||\pi_t)}{\eta_t} - \frac{\mathrm{KL}(u||\pi_{t+1})}{\eta_t} - \tau \mathrm{KL}(u||\pi_{t+1}) + \frac{D_t^2}{2\,\eta_t}\right]$$

$$\leq \sum_{t=0}^{K-1} \left[\frac{\mathrm{KL}(u||\pi_t)}{\eta_t} - \frac{\mathrm{KL}(u||\pi_{t+1})}{\eta_t} - \tau \mathrm{KL}(u||\pi_{t+1})\right] + \sum_{t=0}^{K-1} \frac{D_t^2}{2\,\eta_t}$$
(Since $\mathcal{R}(\pi_0) = \ln(A) - \mathcal{H}(\pi_0) = 0$ and $\mathcal{R}(\pi_K) \geq 0$)

□

C.3.2 Regret Bound for (soft) NPG

In this section, we instantiate the regret and policy evaluation bounds for (soft) NPG

Corollary 10 (Regret Bounds)**.** *Suppose $\pi_0(\cdot|s)$ is the uniform distribution over actions for each state s. For any sequence $\{q_\zeta^t\}_{t=0}^{K-1}$ satisfying $\left\|q_\zeta^t\right\|_\infty \leq H_\tau$, the regret for (soft) NPG can be bounded as:*

- ***Soft NPG**: Setting $\eta_t = \frac{1}{c+\tau\,(t+1)}$ for a constant $c \geq 0$ to be determined later, guarantees that,*

$$\max_s \left|\sum_{t=0}^{K-1} \left[\langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s),\, q_\zeta^t(s,\cdot)\rangle + \tau\left[\mathcal{H}(\pi_\tau^*(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right]\right]\right| \leq \frac{H_\tau^2}{2\,\tau}\left[1 + \ln(K)\right] + (c+\tau)\,\ln(A)$$

- ***NPG**: Setting $\eta_t = \eta = \frac{\sqrt{2}\,(1-\gamma)\,\sqrt{\ln(A)}}{\sqrt{K}}$ guarantees that,*

$$\max_s \left|\sum_{t=0}^{K-1} \left[\langle \pi^*(\cdot|s) - \pi_t(\cdot|s),\, q_\zeta^t(s,\cdot)\rangle\right]\right| \leq \frac{\sqrt{2\ln(A)}\,\sqrt{K}}{1-\gamma}$$

Proof. First note that by using the mirror descent view of the (soft) NPG update [Xiao, 2022], it can be equivalently written as: for all $s \in \mathcal{S}$,

$$\pi_{t+1}(\cdot|s) = \operatorname*{arg\,min}_{\pi \in \Delta} \left[-\eta_t \langle q_\zeta^t(s,\cdot), \pi(\cdot|s)\rangle + \mathrm{KL}(\pi(\cdot|s)||\pi_t(\cdot|s)) - \tau_t \mathcal{H}(\pi(\cdot|s))\right]$$

By comparing to the update in Theorem 4, we note that $d_t = -\eta_t \, q^t_\zeta(s, \cdot)$ and $\|d_t\|_\infty = \eta_t \, \left\| q^t_\zeta \right\|_\infty \leq \eta_t \, H_\tau$. If $\tau_t = \eta_t \tau$ and $\pi_0(\cdot|s)$ is a uniform distribution for each state s, we can instantiate Theorem 4 for each state s, and obtain the following regret bound for the comparator u.

$$
\begin{aligned}
&\sum_{t=0}^{K-1} \left[\langle u(\cdot|s) - \pi_t(\cdot|s), \, q^t_\zeta(s, \cdot)\rangle + \tau \left[\mathcal{H}(u(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right]\right] \\
&\quad \leq \sum_{t=0}^{K-1} \left[\frac{\text{KL}(u(\cdot|s)||\pi_t(\cdot|s))}{\eta_t} - \frac{\text{KL}(u(\cdot|s)||\pi_{t+1}(\cdot|s))}{\eta_t} - \tau \text{KL}(u(\cdot|s)||\pi_{t+1}(\cdot|s))\right] \\
&\quad\quad + \frac{H_\tau^2}{2} \sum_{t=0}^{K-1} \eta_t
\end{aligned}
$$

Now we consider two cases corresponding to `NPG` and its soft variant.

Soft `NPG`: Using that $\tau \neq 0$, setting $u = \pi^*_\tau$ and bounding the RHS in the above inequality,

$$
\begin{aligned}
&\sum_{t=0}^{K-1} \left[\langle \pi^*_\tau(\cdot|s) - \pi_t(\cdot|s), \, q^t_\zeta(s, \cdot)\rangle + \tau \left[\mathcal{H}(\pi^*_\tau(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right]\right] \\
&\quad \leq \sum_{t=1}^{K-1} \text{KL}(\pi^*_\tau(\cdot|s)||\pi_t(\cdot|s)) \left[\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \tau\right] + \frac{1}{\eta_0} \text{KL}(\pi^*_\tau(\cdot|s)||\pi_0(\cdot|s)) + \frac{H_\tau^2}{2} \sum_{t=0}^{K-1} \eta_t \\
&\quad = \frac{H_\tau^2}{2} \sum_{t=0}^{K-1} \frac{1}{c + \tau\,(t+1)} + (c + \tau) \, \text{KL}(\pi^*_\tau(\cdot|s)||\pi_0(\cdot|s)) \qquad \left(\text{Setting } \eta_t = \tfrac{1}{c + \tau\,(t+1)}\right) \\
&\quad \leq \frac{H_\tau^2}{2} \sum_{t=0}^{K-1} \frac{1}{\tau\,(t+1)} + (c + \tau) \, \ln(A) \qquad (\text{Since } \pi_0(\cdot|s) \text{ is a uniform distribution for all } s) \\
&\quad \leq \frac{H_\tau^2}{2\,\tau} \left[1 + \ln(K)\right] + (c + \tau) \, \ln(A) \qquad \left(\text{Since } \textstyle\sum_{t=1}^{K} 1/t \leq 1 + \ln(K)\right)
\end{aligned}
$$

Since the above bound holds for all s,

$$
\begin{aligned}
&\max_s \left| \sum_{t=0}^{K-1} \left[\langle \pi^*_\tau(\cdot|s) - \pi_t(\cdot|s), \, q^t_\zeta(s, \cdot)\rangle + \tau \left[\mathcal{H}(\pi^*_\tau(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right]\right] \right| \\
&\quad \leq \frac{H_\tau^2}{2\,\tau} \left[1 + \ln(K)\right] + (c + \tau) \, \ln(A)
\end{aligned}
$$

`NPG`: Using $u = \pi^*$ and a constant step-size i.e. $\eta_t = \eta$ for all t, in which case the regret bound can be simplified as:

$$
\begin{aligned}
\sum_{t=0}^{K-1} \left[\langle \pi^*(\cdot|s) - \pi_t(\cdot|s), \, q^t_\zeta(s, \cdot)\rangle\right] &\leq \frac{1}{\eta} \text{KL}(\pi^*(\cdot|s)||\pi_0(\cdot|s) + \frac{\eta K}{2\,(1-\gamma)^2} \\
&= \frac{\ln(A)}{\eta} + \frac{\eta K}{2\,(1-\gamma)^2} \qquad (\text{Since } \pi_0 \text{ is the uniform distribution}) \\
&\leq \frac{\sqrt{2\ln(A)}\,\sqrt{K}}{1-\gamma} \qquad \left(\text{Setting } \eta = \tfrac{\sqrt{2}\,(1-\gamma)\,\sqrt{\ln(A)}}{\sqrt{K}}\right)
\end{aligned}
$$

Since the above bound holds for all s,

$$
\max_s \left| \sum_{t=0}^{K-1} \left[\langle \pi^*(\cdot|s) - \pi_t(\cdot|s), \, q^t_\zeta(s, \cdot)\rangle\right] \right| \leq \frac{\sqrt{2\ln(A)}\,\sqrt{K}}{1-\gamma}.
$$

□

C.3.3 Regret Bound for (soft) SPMA

In this section, we instantiate the regret and policy evaluation bounds for (soft) SPMA.

Corollary 11 (Regret Bounds)**.** *Suppose $\pi_0(\cdot|s)$ is the uniform distribution over actions for each state s, and let $\eta_t = \frac{1}{c+\tau\,(t+1)}$ for some constant $c \geq 0$ to be determined later. For any sequence $\{q^t_\zeta\}_{t=0}^{K-1}$ satisfying $\left\|q^t_\zeta\right\|_\infty \leq H_\tau$, the regret for (soft) SPMA can be bounded as:*

- ***Soft SPMA**: Setting $\eta_t = \frac{1}{c+\tau\,(t+1)}$ for a constant $c \geq 2\,\max\{H_\tau, \zeta\ln(A)\}$, guarantees that,*

$$\max_s \left|\sum_{t=0}^{K-1} \left[\langle \pi^*_\tau(\cdot|s) - \pi_t(\cdot|s),\, q^t_\zeta(s,\cdot)\rangle + \tau\left[\mathcal{H}(\pi^*_\tau(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right]\right]\right|$$
$$\leq \frac{3\,H_\tau^2}{\tau}\left[1+\ln(K)\right] + (c+\tau)\,\ln(A)$$

- ***SPMA**: Setting $\eta_t = \eta = \min\left\{\frac{1-\gamma}{2}, \frac{\sqrt{2}\,(1-\gamma)\,\sqrt{\ln(A)}}{\sqrt{K}}\right\}$ guarantees that,*

$$\max_s \left|\sum_{t=0}^{K-1} \left[\langle \pi^*(\cdot|s) - \pi_t(\cdot|s),\, q^t_\zeta(s,\cdot)\rangle\right]\right| \leq \frac{7\,\sqrt{\ln(A)}\,\sqrt{K}}{\sqrt{2}\,(1-\gamma)} + \frac{2}{1-\gamma}\,\ln(A).$$

Proof. For a fixed state $s \in \mathcal{S}$, first note that the (soft) SPMA update in Eq. (5b) can be equivalently be written as follows: if $\Delta^t(s,a) := q^t_\zeta(s,a) - v^t_\zeta(s)$ for $d_t := -\ln(1+\eta_t\,\Delta^t(s,\cdot))$,

$$\pi_{t+1}(\cdot|s) = \arg\min_{\pi\in\Delta}\left[\text{KL}(\pi(\cdot|s)||\pi_t(\cdot|s)\ \left[1+\eta_t\ \left(\Delta^t(s,\cdot)\right)\right]) - \tau_t\mathcal{H}(\pi(\cdot|s))\right]$$
$$= \arg\min_{\pi\in\Delta}\left[\langle d_t, \pi(\cdot|s)\rangle + \text{KL}(\pi(\cdot|s)||\pi_t(\cdot|s)) - \tau_t\mathcal{H}(\pi(\cdot|s))\right]\ ,$$

where we require that $1+\eta_t\Delta^t(s,\cdot)) \geq 0$. Note that since $\left\|q^t_\zeta(s,\cdot)\right\|_\infty \leq H_\tau$, $\|\Delta^t(s,\cdot))\|_\infty \leq H_\tau$ and we require that $\eta_t \leq \frac{1}{2H_\tau}$. With this choice, $|\eta_t\,\Delta^t(s,a)| \leq \frac{1}{2}$. By comparing to the update in Theorem 4, we note that $d_t = -\ln(1+\eta_t\,\Delta^t(s,\cdot))$. Since $|\ln(1+x)| \leq 2|x|$ for all $x \geq -\frac{1}{2}$,

$$|-\ln(1+\eta_t\,\Delta^t(s,a))| \leq 2\,\eta_t\,|\Delta^t(s,a)| \leq 2\,\eta_t\,H_\tau \implies \|d_t\|_\infty \leq 2\,\eta_t\,H_\tau$$

Hence, $D_t = 2\,\eta_t\,H_\tau$. If $\tau_t = \eta_t\tau$ and $\pi_0(\cdot|s)$ is a uniform distribution for each state s, we can instantiate Theorem 4 for each state s, and obtain the following regret bound for the comparator u,

$$\sum_{t=0}^{K-1}\left[\frac{\langle u(\cdot|s) - \pi_t(\cdot|s),\, \ln(1+\eta_t\,\Delta^t(s,\cdot))\rangle}{\eta_t} + \tau\left[\mathcal{H}(u(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right]\right]$$
$$\leq \sum_{t=0}^{K-1}\left[\frac{KL(u(\cdot|s)||\pi_t(\cdot|s))}{\eta_t} - \frac{KL(u(\cdot|s)||\pi_{t+1}(\cdot|s))}{\eta_t} - \tau KL(u(\cdot|s)||\pi_{t+1}(\cdot|s))\right]$$
$$+ 2H_\tau^2\sum_{t=0}^{K-1}\eta_t$$

In order to simplify the above expression, first note that,

$$\langle \pi_t(\cdot|s),\, \ln(1+\eta_t\,\Delta^t(s,\cdot))\rangle \leq \ln(1-\eta_t\,\zeta\,\mathcal{H}(\pi_t(\cdot|s)))$$
(Using Jensen's inequality and the fact $\sum_a \pi_t(a|s)\Delta^t(s,a) = -\zeta\,\mathcal{H}(\pi_t(\cdot|s))$)

If η_t is chosen such that $\eta_t\,\zeta\,\mathcal{H}(\pi_t(\cdot|s)) \leq \frac{1}{2}$, and since $\mathcal{H}(\pi) \in [0, \ln(A)]$, it suffices to ensure $\eta_t \leq \frac{1}{2\zeta\,\ln(A)}$, and use the fact $\ln(1+x) \leq x$ for $x > -1$ to guarantee:

$$\langle \pi_t(\cdot|s),\, \ln(1+\eta_t\,\Delta^t(s,\cdot))\rangle \leq -\eta_t\,\zeta\,\mathcal{H}(\pi_t(\cdot|s)) \tag{11}$$

On the other hand, since choosing $\eta_t \leq \frac{1}{2H_\tau}$ ensures $\ln(1 + \eta_t \Delta^t(s, \cdot))$ is well-defined,

$$\langle u(\cdot|s),\, \ln(1 + \eta_t \Delta^t(s, \cdot))\rangle \geq [\langle u(\cdot|s),\, \eta_t \Delta^t(s, \cdot)\rangle - \langle u(\cdot|s),\, \eta_t^2 [\Delta^t(s, \cdot)]^2\rangle]$$

(since $\ln(1 + x) \geq x - x^2$ for $x > -1/2$)

$$= \eta_t(\langle u(\cdot|s),\, q_\zeta^t(s, \cdot)\rangle - v_\zeta^t(s)) - \eta_t^2 \,\langle u(\cdot|s),\, [\Delta^t(s, \cdot)]^2\rangle$$

$$= \eta_t(\langle u(\cdot|s),\, q_\zeta^t(s, \cdot)\rangle - \langle \pi_t,\, q_\zeta^t(s, \cdot)\rangle - \eta_t\, \zeta \mathcal{H}(\pi_t(\cdot|s))) - \eta_t^2 \langle u(\cdot|s),\, [\Delta^t(s, \cdot)]^2\rangle$$

(since $v_\zeta^t(s) = \langle \pi_t(\cdot|s),\, q_\zeta^t\rangle + \zeta \mathcal{H}(\pi_t(\cdot|s))$)

$$\geq \eta_t(\langle u(\cdot|s),\, q_\zeta^t(s, \cdot)\rangle - \langle \pi_t(\cdot|s),\, q_\zeta^t(s, \cdot)\rangle - \eta_t\, \zeta \mathcal{H}(\pi_t(\cdot|s))) - \eta_t^2 H_\tau^2$$

(since $\|\Delta^t(s, \cdot))\|_\infty \leq H_\tau$)

Combining the above inequalities with Eq. (11),

$$\frac{\langle u(\cdot|s) - \pi_t(\cdot|s),\, \ln(1 + \eta_t\, \Delta^t(s, \cdot))\rangle}{\eta_t} \geq \langle u(\cdot|s),\, q_\zeta^t(s, \cdot)\rangle - \langle \pi_t,\, q_\zeta^t(s, \cdot)\rangle - \eta_t\, H_\tau^2$$
$$= \langle u(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s, \cdot)\rangle - \eta_t H_\tau^2$$

Using the above relation with the regret expression,

$$\sum_{t=0}^{K-1} \left[\langle u(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s, \cdot)\rangle + \tau \left[\mathcal{H}(u(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right]\right]$$
$$\leq \sum_{t=0}^{K-1} \left[\frac{KL(u(\cdot|s)||\pi_t(\cdot|s))}{\eta_t} - \frac{KL(u(\cdot|s)||\pi_{t+1}(\cdot|s))}{\eta_t} - \tau KL(u(\cdot|s)||\pi_{t+1}(\cdot|s))\right]$$
$$+ (\zeta\, \ln(A) + 3\, H_\tau^2)\, \sum_{t=0}^{K-1} \eta_t$$

Note that we require $\eta_t \leq \frac{1}{2H_\tau}$ and $\eta_t \leq \frac{1}{2\zeta \ln(A)}$ simultaneously for all t. Hence, it is sufficient to ensure that $\eta_t \leq \frac{1}{2\, \max\{H_\tau, \zeta \ln(A)\}}$ for all t, and hence require that $c \geq 2\, \max\{H_\tau, \zeta \ln(A)\}$.
Now we consider two cases corresponding to `SPMA` and its soft variant.

Soft `SPMA`: Using that $\tau \neq 0$, setting $u = \pi_\tau^*$ and bounding the RHS in the above inequality,

$$\sum_{t=0}^{K-1} \left[\langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s),\, q_\zeta^t(s, \cdot)\rangle + \tau \left[\mathcal{H}(\pi_\tau^*(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right]\right]$$
$$\leq \sum_{t=1}^{K-1} \mathrm{KL}(\pi_\tau^*(\cdot|s)||\pi_t(\cdot|s)) \left[\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \tau\right] + \frac{1}{\eta_0} \mathrm{KL}(\pi_\tau^*(\cdot|s)||\pi_0(\cdot|s))$$
$$+ 3\, H_\tau^2 \sum_{t=0}^{K-1} \eta_t$$
$$= 3\, H_\tau^2\, \sum_{t=0}^{K-1} \frac{1}{c + \tau\, (t + 1)} + (c + \tau)\, \mathrm{KL}(\pi_\tau^*(\cdot|s)||\pi_0(\cdot|s)) \qquad \left(\text{Since } \eta_t = \tfrac{1}{c+\tau\,(t+1)}\right)$$
$$= 3\, H_\tau^2\, \sum_{t=0}^{K-1} \frac{1}{c + \tau\, (t + 1)} + (c + \tau)\, \ln(A)$$

(Since $\pi_0(\cdot|s)$ is a uniform distribution for all s)

$$\leq \frac{3\, H_\tau^2}{\tau}\, \sum_{t=0}^{K-1} \frac{1}{t + 1} + (c + \tau)\, \ln(A)$$
$$\leq \frac{3\, H_\tau^2}{\tau}\, [1 + \ln(K)] + (c + \tau)\, \ln(A) \qquad \left(\text{Since } \textstyle\sum_{t=1}^{K} 1/t \leq 1 + \ln(K)\right)$$

Since the above bound holds for all s,

$$\max_s \left| \sum_{t=0}^{K-1} \left[\langle \pi^*_\tau(\cdot|s) - \pi_t(\cdot|s),\, q^t_\zeta(s,\cdot) \rangle + \tau \left[\mathcal{H}(\pi^*_\tau(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s)) \right] \right] \right|$$
$$\leq \frac{3\, H_\tau^2}{\tau} \left[1 + \ln(K) \right] + (c + \tau)\, \ln(A)$$

SPMA: Using $u = \pi^*$, $\tau = \zeta = 0$, and a constant step-size i.e. $\eta_t = \eta$ for all t, in which case the regret bound can be simplified as:

$$\sum_{t=0}^{K-1} \left[\langle \pi^*(\cdot|s) - \pi_t(\cdot|s),\, q^t_\zeta(s,\cdot) \rangle \right] \leq \frac{1}{\eta} \text{KL}(\pi^*(\cdot|s) || \pi_0(\cdot|s)) + \frac{3\, \eta K}{(1-\gamma)^2}$$
$$= \frac{\ln(A)}{\eta} + \frac{3\, \eta K}{(1-\gamma)^2}$$

(Since $\pi_0(\cdot|s)$ is a uniform distribution for all s)

Recall that in the presence of entropy ensuring $\ln(1 + \eta_t \Delta^t(s,\cdot))$ is well-defined required us to choose $\eta_t \leq \frac{1}{2\, H_\tau}$. When $\tau = 0$ and $\eta_t = \eta$, the condition simplifies to $\eta \leq \frac{1-\gamma}{2}$. Setting $\eta = \min\left\{ \frac{\sqrt{2}\,(1-\gamma)\, \sqrt{\ln(A)}}{\sqrt{K}}, \frac{1-\gamma}{2} \right\}$ and using the fact that $\frac{1}{\min\{a,b\}} = \max\{1/a, 1/b\}$

$$\leq \ln(A)\, \max\left\{ \frac{\sqrt{K}}{\sqrt{2}\,(1-\gamma)\, \sqrt{\ln(A)}}, \frac{2}{1-\gamma} \right\}$$
$$+ \frac{3\, K}{(1-\gamma)^2}\, \min\left\{ \frac{\sqrt{2}\,(1-\gamma)\, \sqrt{\ln(A)}}{\sqrt{K}}, \frac{1-\gamma}{2} \right\}$$
$$\leq \frac{7\, \sqrt{\ln(A)}\, \sqrt{K}}{\sqrt{2}\,(1-\gamma)} + \frac{2}{1-\gamma}\, \ln(A)$$

(Since $\max\{a,b\} \leq a + b$, $\min\{a,b\} \leq a$)

Since the above bound holds for all s,

$$\max_s \left| \sum_{t=0}^{K-1} \left[\langle \pi^*(\cdot|s) - \pi_t(\cdot|s),\, q^t_\zeta(s,\cdot) \rangle \right] \right| \leq \frac{7\, \sqrt{\ln(A)}\, \sqrt{K}}{\sqrt{2}\,(1-\gamma)} + \frac{2}{1-\gamma}\, \ln(A).$$

□

C.4 Proof of Theorem 2

Finally, we put everything together, and in the following two subsections, we prove theorems that quantify the performance of (soft) NPG and (soft) SPMA when using the hard or soft Bellman operator (i.e., $\zeta = \tau$ or $\zeta = 0$).

C.4.1 Putting everything together for soft NPG

Theorem 5 (Sub-optimality of Soft NPG)**.** *Let π^*_τ denote the optimal entropy-regularized policy with value function v^*_τ. Consider the soft NPG update with step size $\eta_t = \frac{1}{c + \tau\,(t+1)}$, $c \geq \max\left\{ \frac{8\,(1+\tau\, \ln(A))}{(1-\gamma)}, 32\, \tau\, \ln(A), \frac{2\,(1+\tau\, \ln(A))^2}{(1-\gamma)^2\, \tau \ln(A)} \right\}$ and, $\delta(\tau,\zeta) := \frac{|\tau-\zeta|\, \ln(A)}{1-\gamma}$. Let $\pi_0(\cdot|s)$ be the uniform policy over actions for all $s \in \mathcal{S}$ and assume the policy evaluation step in Eq. (1). Then the resulting mixture policy $\bar{\pi}_K$ satisfies the following sub-optimality bound,*

$$\left\| v^{\bar{\pi}_K}_\tau - v^*_\tau \right\|_\infty \leq \frac{1}{K\,(1-\gamma)} \left[\frac{(1 + \tau\, \log(A))^2}{2\, \tau\, (1-\gamma)^2} \left[1 + \ln(K) \right] + (c + \tau)\, \ln(A) \right]$$
$$+ \frac{16\,(1 + \tau\, \log(A))\, \gamma^m}{(1-\gamma)^4\, K} \left[(1 + \tau\; \ln(A\,K))\, \sqrt{\ln(A)} \left(\sqrt{K} + \frac{1}{1 - \sqrt{\gamma}} \right) \right.$$

$$+\tau\ (\ln(A)+1)\ \sqrt{K}\Bigg]$$

$$+\frac{2\,\delta(\tau,\zeta)}{1-\gamma}$$

Proof. Plugging the regret bound in Corollary 10 for soft `NPG` into the regret part of Theorem 1 immediately gives the first part of the upper-bound:

$$\frac{\|\text{Regret}(K)\|_\infty}{K\,(1-\gamma)} \leq \frac{1}{K\,(1-\gamma)}\left[\frac{(1+\tau\,\log(A))^2}{2\,\tau\,(1-\gamma)^2}\,[1+\ln(K)]+(c+\tau)\,\ln(A)\right]$$

Using the result from Theorem 3 and Corollary 6 to upper-bound the error part in Theorem 1 we obtain:

$$\begin{aligned}
&\frac{2\,\sum_{k\in[K]}\|\boldsymbol{\epsilon}_k\|_\infty}{K\,(1-\gamma)} \\
&\leq \frac{8\,(1+\tau\,\log(A))\,\gamma^m}{(1-\gamma)^4\,K}\Bigg[(1+\tau\ \ln(A\,K))\ \sqrt{\ln(A)}\sum_{t=1}^{K}\left(\frac{1}{\sqrt{t}}+(\gamma^m)^{t/2}\right) \\
&\qquad +\tau\ (\ln(A)+1)\ \sqrt{K}\Bigg]+\frac{2\,\delta(\tau,\zeta)}{1-\gamma} \\
&\leq \frac{8\,(1+\tau\,\log(A))\,\gamma^m}{(1-\gamma)^4\,K}\Bigg[(1+\tau\ \ln(A\,K))\ \sqrt{\ln(A)}\left(2\,\sqrt{K}+\frac{1}{1-\sqrt{\gamma}}\right) \\
&\qquad +\tau\ (\ln(A)+1)\ \sqrt{K}\Bigg]+\frac{2\,\delta(\tau,\zeta)}{1-\gamma}
\end{aligned}$$

Where for obtain the inequality above using the fact $\sum_{t=1}^{K}\frac{1}{\sqrt{t}}\leq 2\sqrt{K}$ (using integration) and $\sum_{t=1}^{K}(\gamma^m)^{t/2}\leq\frac{1}{1-\gamma^{m/2}}\leq\frac{1}{1-\sqrt{\gamma}}$

$$\begin{aligned}
&\leq \frac{16\,(1+\tau\,\log(A))\,\gamma^m}{(1-\gamma)^4\,K}\Bigg[(1+\tau\ \ln(A\,K))\ \sqrt{\ln(A)}\left(\sqrt{K}+\frac{1}{1-\sqrt{\gamma}}\right) \\
&\qquad +\tau\ (\ln(A)+1)\ \sqrt{K}\Bigg]+\frac{2\,\delta(\tau,\zeta)}{1-\gamma}
\end{aligned}$$

□

Corollary 12 (Sub-optimality of Soft `NPG`)**.** *Let π^*_τ denote the optimal entropy-regularized policy with value function v^*_τ. Consider the soft `NPG` update with step size $\eta_t=\frac{1}{c+\tau\,(t+1)}$ and $c\geq\max\left\{\frac{8\,(1+\tau\,\ln(A))}{(1-\gamma)},32\,\tau\,\ln(A),\frac{2\,(1+\tau\,\ln(A))^2}{(1-\gamma)^2\,\tau\,\ln(A)}\right\}$. Let $\pi_0(\cdot|s)$ be the uniform policy over actions for all $s\in\mathcal{S}$ and assume the policy evaluation step in Eq. (1) with $\zeta=\tau$. Then the resulting mixture policy $\bar{\pi}_K$ satisfies the following sub-optimality bound,*

$$\begin{aligned}
\left\|v^{\bar{\pi}_K}_\tau-v^*_\tau\right\|_\infty &\leq \frac{1}{K\,(1-\gamma)}\left[\frac{(1+\tau\,\log(A))^2}{2\,\tau\,(1-\gamma)^2}\,[1+\ln(K)]+(c+\tau)\,\ln(A)\right] \\
&\quad+\frac{16\,(1+\tau\,\log(A))\,\gamma^m}{(1-\gamma)^4\,K}\Bigg[(1+\tau\ \ln(A\,K))\ \sqrt{\ln(A)}\left(\sqrt{K}+\frac{1}{1-\sqrt{\gamma}}\right)
\end{aligned}$$

$$+ \tau\,(\ln(A)+1)\,\sqrt{K}\Bigg]$$

Proof. Using Theorem 1 with Corollary 10 for soft NPG, Theorem 3, Corollary 6 and the result from Theorem 5. □

Corollary 13 (Sub-optimality of Soft NPG)**.** *Let π^*_τ denote the optimal entropy-regularized policy with value function v^*_τ. Consider the soft NPG update with step size $\eta_t = \frac{1}{c+\tau\,(t+1)}$ and $c \geq \max\left\{\frac{8\,(1+\tau\,\ln(A))}{(1-\gamma)}, 32\,\tau\,\ln(A), \frac{2\,(1+\tau\,\ln(A))^2}{(1-\gamma)^2\,\tau\ln(A)}\right\}$. Let $\pi_0(\cdot|s)$ be the uniform policy over actions for all $s \in \mathcal{S}$ and assume the policy evaluation step in Eq. (1) with $\zeta = 0$. Then the resulting mixture policy $\bar{\pi}_K$ satisfies the following sub-optimality bound,*

$$\begin{aligned}\left\|v^{\bar{\pi}_K}_\tau - v^*_\tau\right\|_\infty \leq \frac{1}{K\,(1-\gamma)}&\left[\frac{(1+\tau\,\log(A))^2}{2\,\tau\,(1-\gamma)^2}\,[1+\ln(K)] + (c+\tau)\,\ln(A)\right] \\ &+ \frac{16\,(1+\tau\,\log(A))\,\gamma^m}{(1-\gamma)^4\,K}\Bigg[(1+\tau\;\ln(A\,K))\,\sqrt{\ln(A)}\left(\sqrt{K}+\frac{1}{1-\sqrt{\gamma}}\right) \\ &\qquad + \tau\,(\ln(A)+1)\,\sqrt{K}\Bigg] \\ &+ \frac{2\,\tau\,\ln(A)}{(1-\gamma)^2}\end{aligned}$$

Proof. Using Theorem 1 with Corollary 10 for soft NPG, Theorem 3, Corollary 7 and the result from Theorem 5. □

Theorem 6 (NPG + policy evaluation without entropy regularization)**.** *If π^* is the optimal policy whose value function is equal to v^*, the NPG update with $\eta_t = \eta = \frac{\sqrt{2}\,(1-\gamma)\,\sqrt{\ln(A)}}{\sqrt{K}}$, $\pi_0(\cdot|s)$ as the uniform initial policy for each $s \in \mathcal{S}$ with the policy evaluation procedure in Eq. (1) with $\zeta = 0$ satisfies the following sub-optimality bound for the mixture policy $\bar{\pi}_K$,*

$$\left\|v^{\bar{\pi}_K} - v^*\right\|_\infty \leq \frac{\sqrt{2\ln(A)}}{\sqrt{K}\,(1-\gamma)^2} + \frac{4\,\sqrt{\ln(A)}\,\gamma^m}{\sqrt{K}\,(1-\gamma)^4}$$

Proof. Using Corollary 3 with Corollary 10 for NPG and Corollary 7. □

C.4.2 Putting everything together for soft SPMA

Theorem 7 (Sub-optimality of Soft SPMA)**.** *Let π^*_τ denote the optimal entropy-regularized policy with value function v^*_τ. Consider the soft SPMA update with step size $\eta_t = \frac{1}{c+\tau\,(t+1)}$, $c \geq \max\left\{\frac{4\,(1+\tau\,\ln(A))}{(1-\gamma)}, 32\,\tau\,\ln(A), \frac{(1+\tau\,\ln(A))^2}{(1-\gamma)^2\,2\tau\ln(A)}, \frac{2\,(1+\tau\ln(A))}{1-\gamma}\right\}$, and, $\delta(\tau,\zeta) := \frac{|\tau-\zeta|\,\ln(A)}{1-\gamma}$. Let $\pi_0(\cdot|s)$ be the uniform policy over actions for all $s \in \mathcal{S}$ and assume the policy evaluation step in Eq. (1). Then the resulting mixture policy $\bar{\pi}_K$ satisfies the following sub-optimality bound,*

$$\begin{aligned}\left\|v^{\bar{\pi}_K}_\tau - v^*_\tau\right\|_\infty \leq \frac{1}{K\,(1-\gamma)}&\left[\frac{3\,(1+\tau\,\log(A))^2}{2\,\tau\,(1-\gamma)^2}\,[1+\ln(K)] + (c+\tau)\,\ln(A)\right] \\ &+ \frac{16\,(1+\tau\,\log(A))\,\gamma^m}{(1-\gamma)^4\,K}\Bigg[(1+\tau\;\ln(A\,K))\,\sqrt{\ln(A)}\left(\sqrt{K}+\frac{1}{1-\sqrt{\gamma}}\right) \\ &\qquad + \tau\,(\ln(A)+1)\,\sqrt{K}\Bigg] \\ &+ \frac{2\,\delta(\tau,\zeta)}{1-\gamma}\end{aligned}$$

Proof. Using Theorem 1 with Corollary 11 for soft `SPMA`, Theorem 3, Corollary 8 and the facts from the proof of Theorem 5. □

Corollary 14 (Sub-optimality of Soft `SPMA`)**.** *Let π^*_τ denote the optimal entropy-regularized policy with value function v^*_τ. Consider the soft `SPMA` update with step size $\eta_t = \frac{1}{c+\tau\,(t+1)}$, $c \geq \max\left\{\frac{4\,(1+\tau\,\ln(A))}{(1-\gamma)}, 32\,\tau\,\ln(A), \frac{(1+\tau\,\ln(A))^2}{(1-\gamma)^2\,2\tau\,\ln(A)}, \frac{2\,(1+\tau\ln(A))}{1-\gamma}\right\}$. Let $\pi_0(\cdot|s)$ be the uniform policy over actions for all $s \in \mathcal{S}$ and assume the policy evaluation step in Eq. (1) with $\zeta = \tau$. Then the resulting mixture policy $\bar{\pi}_K$ satisfies the following sub-optimality bound,*

$$\left\|v^{\bar{\pi}_K}_\tau - v^*_\tau\right\|_\infty \leq \frac{1}{K\,(1-\gamma)}\left[\frac{3\,(1+\tau\,\log(A))^2}{2\,\tau\,(1-\gamma)^2}\,[1+\ln(K)] + (c+\tau)\,\ln(A)\right] + \frac{16\,(1+\tau\,\log(A))\,\gamma^m}{(1-\gamma)^4\,K}\left[(1+\tau\,\ln(A\,K))\,\sqrt{\ln(A)}\left(\sqrt{K} + \frac{1}{1-\sqrt{\gamma}}\right) + \tau\,(\ln(A)+1)\,\sqrt{K}\right]$$

Proof. Using Theorem 1 with Corollary 11 for soft `SPMA`, Theorem 3, Corollary 8, Theorem 7 and the facts from the proof of Theorem 5. □

Corollary 15 (Sub-optimality of Soft `SPMA`)**.** *Let π^*_τ denote the optimal entropy-regularized policy with value function v^*_τ. Consider the soft `SPMA` update with step size $\eta_t = \frac{1}{c+\tau\,(t+1)}$, $c \geq \max\left\{\frac{4\,(1+\tau\,\ln(A))}{(1-\gamma)}, 32\,\tau\,\ln(A), \frac{(1+\tau\,\ln(A))^2}{(1-\gamma)^2\,2\tau\,\ln(A)}, \frac{2\,(1+\tau\ln(A))}{1-\gamma}\right\}$. Let $\pi_0(\cdot|s)$ be the uniform policy over actions for all $s \in \mathcal{S}$ and assume the policy evaluation step in Eq. (1) with $\zeta = 0$. Then the resulting mixture policy $\bar{\pi}_K$ satisfies the following sub-optimality bound,*

$$\left\|v^{\bar{\pi}_K}_\tau - v^*_\tau\right\|_\infty \leq \frac{1}{K\,(1-\gamma)}\left[\frac{3\,(1+\tau\,\log(A))^2}{2\,\tau\,(1-\gamma)^2}\,[1+\ln(K)] + (c+\tau)\,\ln(A)\right] + \frac{16\,(1+\tau\,\log(A))\,\gamma^m}{(1-\gamma)^4\,K}\left[(1+\tau\,\ln(A\,K))\,\sqrt{\ln(A)}\left(\sqrt{K} + \frac{1}{1-\sqrt{\gamma}}\right) + \tau\,(\ln(A)+1)\,\sqrt{K}\right] + \frac{2\,\tau\,\ln(A)}{(1-\gamma)^2}$$

Proof. Using Theorem 1 with Corollary 11 for soft `SPMA`, Theorem 3, Corollary 9, Theorem 7 and the facts from the proof of Theorem 5. □

Theorem 8 (`SPMA` + policy evaluation without entropy regularization)**.** *If π^* is the optimal policy whose value function is equal to v^*, the `SPMA` update in Eq. (4) with $\eta_t = \eta = \min\left\{\frac{1-\gamma}{2}, \frac{\sqrt{2}\,(1-\gamma)\,\sqrt{\ln(A)}}{\sqrt{K}}\right\}$, $\pi_0(\cdot|s)$ as the uniform initial policy for each $s \in \mathcal{S}$ with the policy evaluation procedure in Eq. (1) with $\zeta = 0$ satisfies the following sub-optimality bound for the mixture policy $\bar{\pi}_K$,*

$$\left\|v^{\bar{\pi}_K} - v^*\right\|_\infty \leq \frac{1}{K\,(1-\gamma)}\left[\frac{7\,\sqrt{\ln(A)}\,\sqrt{K}}{\sqrt{2}\,(1-\gamma)} + \frac{2}{1-\gamma}\,\ln(A)\right] + \frac{2\,\sqrt{\ln(A)}\,\gamma^m}{\sqrt{K}\,(1-\gamma)^4}$$

Proof. Using Corollary 3 with Corollary 11 for `SPMA` and Corollary 9. □

D Helper Lemmas

Lemma 1. *For any constant $C \in (0, 1/2)$, if P and Q are discrete distributions with support A, if $\|P - Q\|_1 \leq \frac{1}{2}$, then,*

$$|\mathcal{H}(Q) - \mathcal{H}(P)| \leq \|Q - P\|_1 \ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A-1)}{2} + \sqrt{2}\right)\sqrt{C}$$

Proof. By Cover [1999, Theorem 17.3.3], if $\|Q - P\|_1 \leq \frac{1}{2}$, then,

$$|\mathcal{H}(Q) - \mathcal{H}(P)| \leq \|Q - P\|_1 \ln\left(\frac{A}{\|Q - P\|_1}\right) \tag{12}$$

Furthermore, using Sason [2013, Theorem 3], we also know that,

$$|\mathcal{H}(Q) - \mathcal{H}(P)| \leq \frac{\ln(A-1)}{2} \|Q - P\|_1 + h\left(\frac{1}{2} \|Q - P\|_1\right),$$

where h is a binary entropy, i.e. for $0 < x < 1$, $h(x) = -x\ln(x) - (1-x)\ln(1-x)$. Using the fact that $h(x) \leq 2\sqrt{x(1-x)}$,

$$\leq \frac{\ln(A-1)}{2} \|Q - P\|_1 + \sqrt{2}\sqrt{\|Q - P\|_1 \left(1 - \frac{\|Q - P\|_1}{2}\right)}$$
$$\leq \frac{\ln(A-1)}{2} \|Q - P\|_1 + \sqrt{2}\sqrt{\|Q - P\|_1}$$

Assuming $\|Q - P\|_1 \leq 2$

$$|\mathcal{H}(Q) - \mathcal{H}(P)| \leq \left(\frac{\ln(A-1)}{2} + \sqrt{2}\right)\sqrt{\|Q - P\|_1} \tag{13}$$

We will use each of these inequalities for different cases of $\|Q - P\|_1$, where $\|Q - P\|_1 \leq \frac{1}{2}$.
Case (1): For any constant $C \in (0, 1/2)$, if $\|Q - P\|_1 \geq C$, then, using Eq. (12),

$$|\mathcal{H}(Q) - \mathcal{H}(P)| \leq \|Q - P\|_1 \ln\left(\frac{A}{C}\right)$$

Case (2): On the other hand, if $\|Q - P\|_1 \leq C$, then we use Eq. (13) to get that,

$$|\mathcal{H}(Q) - \mathcal{H}(P)| \leq \left(\frac{\ln(A-1)}{2} + \sqrt{2}\right)\sqrt{C}$$

Combining both cases, if $\|Q - P\|_1 \leq \frac{1}{2}$, then, for a constant $C \in (0, 1/2)$,

$$|\mathcal{H}(Q) - \mathcal{H}(P)| \leq \|Q - P\|_1 \ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A-1)}{2} + \sqrt{2}\right)\sqrt{C}$$

□

Lemma 2. *For all $k \in [K]$,*

$$\sum_{i=1}^{k} \frac{\gamma^{k-i}}{\sqrt{(i+1)}} \leq \sqrt{\frac{2}{k}}\,\frac{1}{1-\gamma} + \frac{\gamma^{k/2}}{1-\gamma}$$

Furthermore for $k \geq \frac{1}{(\ln(1/\gamma))^2}$,

$$\sum_{i=1}^{k} \frac{\gamma^{k-i}}{\sqrt{(i+1)}} \leq \frac{1}{\sqrt{k}}\,\frac{4}{1-\gamma}$$

Proof. Define $j = k - i$, in which case, we need to bound,

$$\sum_{j=0}^{k-1} \frac{\gamma^j}{\sqrt{k-j+1}} = \underbrace{\sum_{j=0}^{\lfloor k/2 \rfloor} \frac{\gamma^j}{\sqrt{k-j+1}}}_{\text{Term (i)}} + \underbrace{\sum_{j=\lfloor k/2 \rfloor+1}^{k-1} \frac{\gamma^j}{\sqrt{k-j+1}}}_{\text{Term (ii)}}$$

For bounding Term (i), note that $j \leq \frac{k}{2} + 1$, meaning that $k - j + 1 \geq \frac{k}{2}$. Hence,

$$\text{Term (i)} = \sum_{j=0}^{\lfloor k/2 \rfloor} \frac{\gamma^j}{\sqrt{k-j+1}} \leq \sqrt{\frac{2}{k}} \sum_{j=0}^{\lfloor k/2 \rfloor} \gamma^j \leq \sqrt{\frac{2}{k}} \sum_{j=0}^{\infty} \gamma^j \leq \sqrt{\frac{2}{k}} \frac{1}{1-\gamma}$$

For bounding Term (ii), note that since $j \leq k$, $k - j + 1 \geq 1$. Hence,

$$\text{Term (ii)} = \sum_{j=\lfloor k/2 \rfloor+1}^{k-1} \frac{\gamma^j}{\sqrt{k-j+1}} \leq \sum_{j=\lfloor k/2 \rfloor+1}^{k-1} \gamma^j \leq \sum_{j=\lfloor k/2 \rfloor+1}^{\infty} \gamma^j \leq \frac{\gamma^{k/2}}{1-\gamma}$$

Combining both terms,

$$\sum_{j=0}^{k-1} \frac{\gamma^j}{\sqrt{k-j+1}} \leq \sqrt{\frac{2}{k}} \frac{1}{1-\gamma} + \frac{\gamma^{k/2}}{1-\gamma}$$

Furthermore, for large enough k s.t. $\frac{\ln(k)}{k} \leq \ln(1/\gamma)$, $\gamma^{k/2} \leq \frac{1}{\sqrt{k}}$. In this case,

$$\sum_{j=0}^{k-1} \frac{\gamma^j}{\sqrt{k-j+1}} \leq \frac{1}{\sqrt{k}} \frac{4}{1-\gamma}$$

Since $\ln(k) \leq \sqrt{k}$ for all $k \geq 1$, we can find k s.t. $\frac{\sqrt{k}}{k} \leq \ln(1/\gamma)$. Solving this we get $k \geq \frac{1}{(\ln(1/\gamma))^2}$. □

Lemma 3. *For any state-action value function q, for any $m \geq 1$,*

$$|(T_\tau^{\pi_1} q)^m(s,a) - (T_\zeta^{\pi_1} q)^m(s,a)| \leq \frac{|\tau - \zeta|}{1-\gamma} \ln(A)$$

Proof.

$$(T_\tau^{\pi_1} q)^m(s,a) = \mathbb{E}_{\substack{s_{t+1} \sim \mathcal{P}(\cdot|s_t,a_t) \\ a_t \sim \pi_1(\cdot|s_t)}} \Bigg[\sum_{t=0}^{m-1} \gamma^t \big(r(s_t,a_t) - \tau \ln \pi_1(a_t|s_t) \big) \tag{14}$$

$$+ \gamma^m q(s_m,a_m) \,\Big|\, s_0 = s,\ a_0 = a \Bigg] \qquad \text{(By definition of } T_\tau^{\pi_1}\text{)}$$

$$= \mathbb{E}\left[\sum_{t=0}^{m-1} \gamma^t (r(s_t,a_t)) + \gamma^m q(s_m,a_m)\right] - \tau\,\mathbb{E}\left[\sum_{t=0}^{m-1} \gamma^t \ln(\pi_1(a_t|s_t))\right] \tag{15}$$

$$+ \zeta\,\mathbb{E}\left[\sum_{t=0}^{m-1} \gamma^t \ln(\pi_1(a_t|s_t))\right] - \zeta\,\mathbb{E}\left[\sum_{t=0}^{m-1} \gamma^t \ln(\pi_1(a_t|s_t))\right]$$

(Add/Subtract $\zeta\,\mathbb{E}\left[\sum_{t=0}^{m-1} \gamma^t \ln(\pi_1(a_t|s_t))\right]$)

$$= (T_\zeta^{\pi_1} q)^m(s,a) - (\tau - \zeta)\,\mathbb{E}\left[\sum_{t=0}^{m-1} \gamma^t \ln(\pi_1(a_t|s_t))\right] \qquad \text{(By definition of } T_\zeta^{\pi_1}\text{)}$$

$$= (T_\zeta^{\pi_1} q)^m(s,a) + (\tau - \zeta)\,\mathbb{E}\left[\sum_{t=0}^{m-1} \gamma^t \mathcal{H}(\pi_1(.|s_t))\right]$$

(By definition of $\mathcal{H}(\pi_1(.|s_t))$)

$$\leq (T_\zeta^{\pi_1} q)^m(s,a) + \frac{|\tau - \zeta|}{1-\gamma} \ln(A) \qquad \text{(Since } \mathbb{E}[\mathcal{H}(\pi_1(.|s_t))] \leq \ln(A)\text{)}$$

□

The following proposition shows that the objective in Eq. (5b) admits a closed-form solution. Substituting the intermediate policies NPG and SPMA into the closed-form expression of Eq. (16) yields their entropy-regularized (actor entropy) counterparts, denoted soft NPG and soft SPMA.

Proposition 2. *If* $\alpha_t := \frac{1}{1+\tau_t}$*, the closed-form solution for the proximal update in Eq. (5b) for any* s, a *is given as,*

$$\pi_{t+1}(a|s) = \frac{[\pi_{t+1/2}(a|s)]^{\alpha_t}}{\sum_{a'} [\pi_{t+1/2}(a'|s)]^{\alpha_t}} \tag{16}$$

Proof.

$$\begin{aligned}
\pi_{t+1}(\cdot|s) &= \arg\min_{\pi \in \Delta} \left[\mathrm{KL}(\pi || \pi_{t+1/2}(\cdot|s)) - \tau_t \mathcal{H}(\pi)\right] \\
&= \arg\min_{\pi \in \Delta} \mathbb{E}_{a \sim \pi(.|s)} \left[(1+\tau_t) \ln(\pi(.|s)) - \ln(\pi_{t+1/2}(\cdot|s))\right] \\
&\qquad \text{(Using the definition of } H(\pi(\cdot|s))\text{)} \\
&= \arg\min_{\pi \in \Delta} \mathbb{E}_{a \sim \pi(.|s)} \left[\ln(\pi(.|s)) - \alpha_t \ln(\pi_{t+1/2}(\cdot|s))\right] \qquad \text{(Since } \alpha_t = \tfrac{1}{1+\tau_t}\text{)} \\
&= \arg\min_{\pi \in \Delta} \mathbb{E}_{a \sim \pi(.|s)} \left[\ln(\pi(.|s)) - \ln([\pi_{t+1/2}(\cdot|s)]^{\alpha_t})\right] \\
&= \arg\min_{\pi \in \Delta} \mathrm{KL}(\pi(.|s) || \pi_{t+1/2}(\cdot|s)^{\alpha_t}) \qquad \text{(By definition of the KL divergence)}
\end{aligned}$$

Using the fact that KL projection onto the simplex results in normalization, we get that,

$$\pi_{t+1}(a|s) = \frac{[\pi_{t+1/2}(a|s)]^{\alpha_t}}{\sum_{a'} [\pi_{t+1/2}(a'|s)]^{\alpha_t}}$$

□

E Experimental Details and Additional Results

E.1 Details for Stable Baselines Experiments

For our experiments, we follow Tomar et al. [2020], Asad et al. [2024] and use the default hyperparameters from `stable-baselines3` [Raffin et al., 2021] for each method. This choice is motivated by prior work, which focuses on evaluating the effectiveness of different surrogate losses rather than performing exhaustive hyperparameter searches. Such searches are particularly impractical for CNN-based actor and critic networks, where tuning multiple hyperparameters (e.g., framestack, buffer size, minibatch size, discount factor) is computationally expensive. The full list of hyperparameters for the Atari experiments is provided in Table 1. For our forward and reverse KL-based objectives, we set η_t to a constant selected via grid search over $[0.01, 0.1, 1]$.

E.2 Does Double Q Learning Hurt or Improve Performance?

In Fig. 4, we evaluate single vs. double Q-learning on 8 Atari games under our setup, which uses the hard Bellman operator, adaptive entropy via coefficient loss, and a fixed target entropy. Double Q-learning performs comparably to single Q-learning. Notably, for the bottom four games, our DSAC configuration substantially outperforms the results reported in Xu et al. [2021], despite their use of an adaptive target entropy (see Figure2 in Xu et al. [2021])

E.3 Our Objectives Benefit Similarly from the Hard Bellman Operator

Using our proposed objectives, we consistently observe the same trend as with DSAC: employing the soft Bellman operator ($\zeta = \tau$) with an adaptive entropy coefficient leads to poor performance compared to DQN. In contrast, switching to the hard Bellman operator yields substantial improvements (see left vs. right columns of Fig. 5).

E.4 Additional Results: Entropy Regularization and KL Direction

Hyperparameter	**FKL Objectives**	**RKL Objectives**	**DSAC**	**DQN**
Reward normalization	✗	✗	✗	✗
Observation normalization	✗	✗	✗	✗
Orthogonal weight initialization	✗	✗	✗	✗
Value function clipping	✗	✗	✗	✗
Gradient clipping	✗	✗	✗	✗
Probability ratio clipping	✗	✗	✗	✗
Entropy coefficient	auto	auto	auto	ϵ-greedy
Adam step-size	3×10^{-4}			
Buffer size	10^6			
Minibatch size	256			
Framestack	4			
Number of environment copies	8			
Discount factor	0.99			
Total number of timesteps	10^7			
Number of runs for plot averages	5			
Confidence interval for plot runs	$\sim 95\%$			

Table 1: Hyper-parameters for Atari experiments.

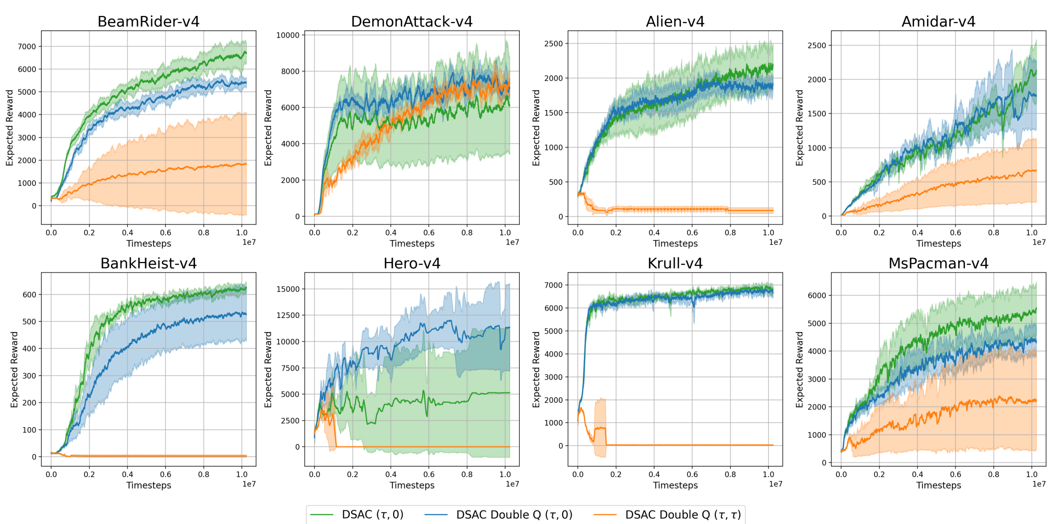

Figure 4: When critic entropy is disabled, Double-Q learning performs comparably to using a single Q, with no conclusive evidence that one approach consistently outperforms the other.

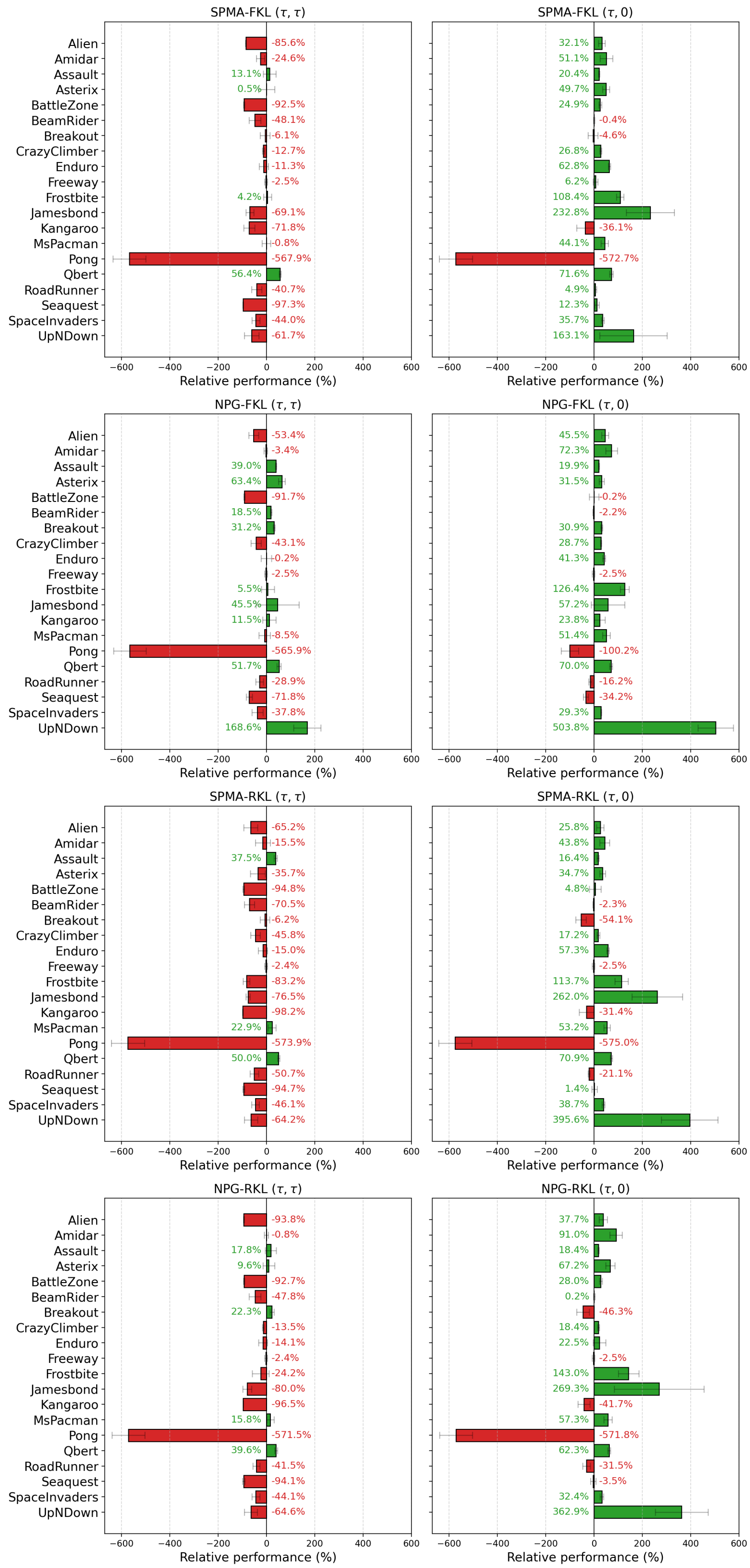

Figure 5: Similar to DSAC, disabling critic entropy for our forward- and reverse-KL-based actors leads to a substantial performance improvement (left vs. right columns).

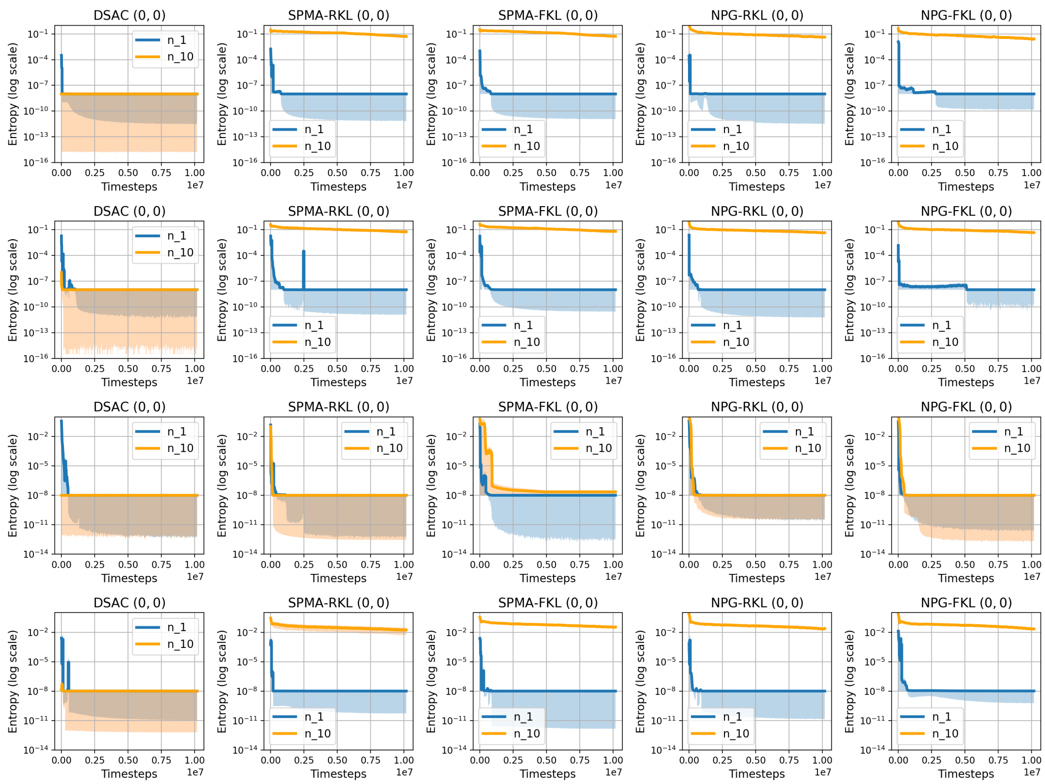

Figure 6: Policy Shannon entropy during training. Without entropy regularization, increasing n from 1 to 10 substantially increases and stabilizes the Shannon entropy for our forward and reverse KL-based methods, compared to `DSAC`, across four Atari games. The rows correspond to MsPacman-v4, Seaquest-v4, Freeway-v4, and Breakout-v4.

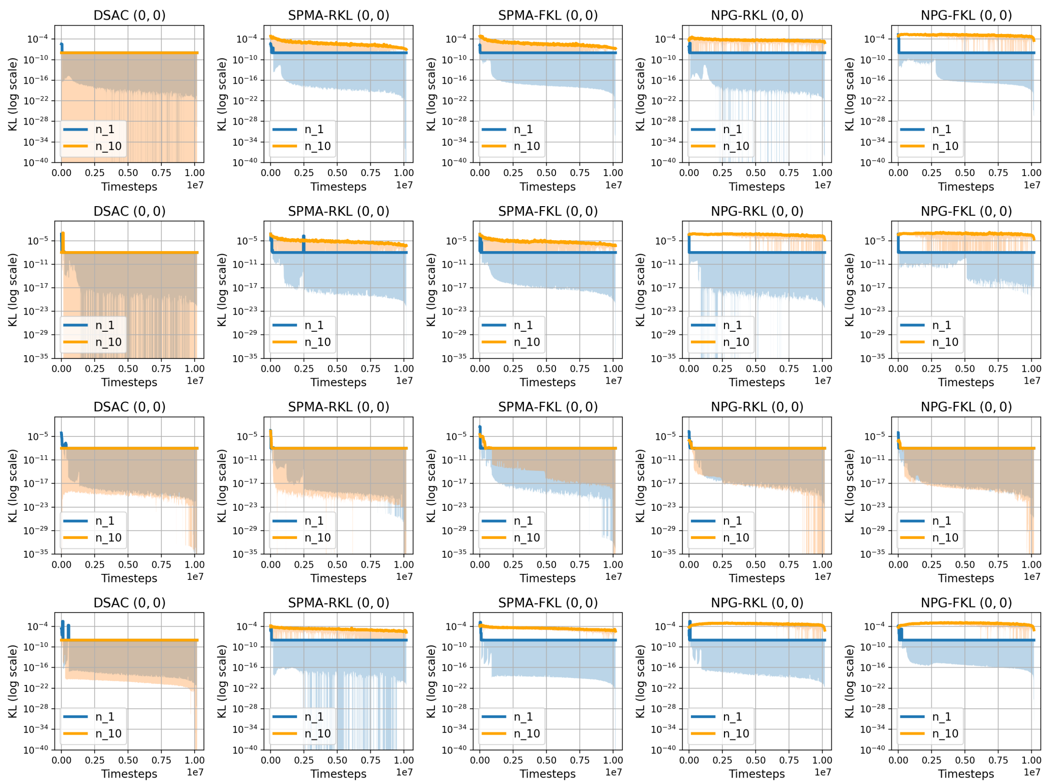

Figure 7: KL divergence during training. Consistent with the Shannon entropy results, increasing n to 10 without entropy regularization allows our forward- and reverse-KL-based actors to achieve higher KL divergence than DSAC, which also stabilizes over the course of training.

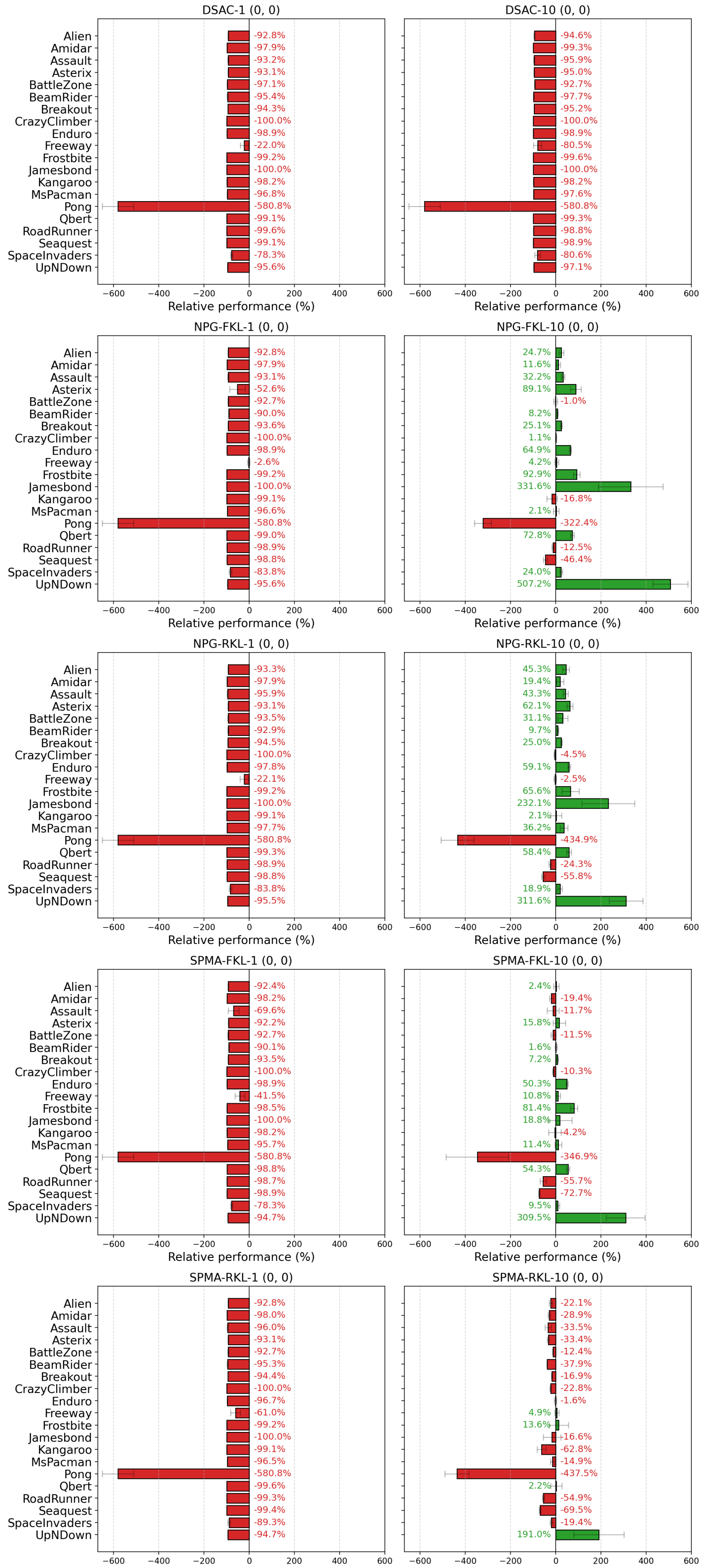

Figure 8: Effect of the number of actor optimization steps n without entropy regularization. Unlike DSAC, all our objectives (except SPMA−RKL$(0, 0)$) achieve performance comparable to DQN when n is increased from 1 to 10 (left vs. right columns).

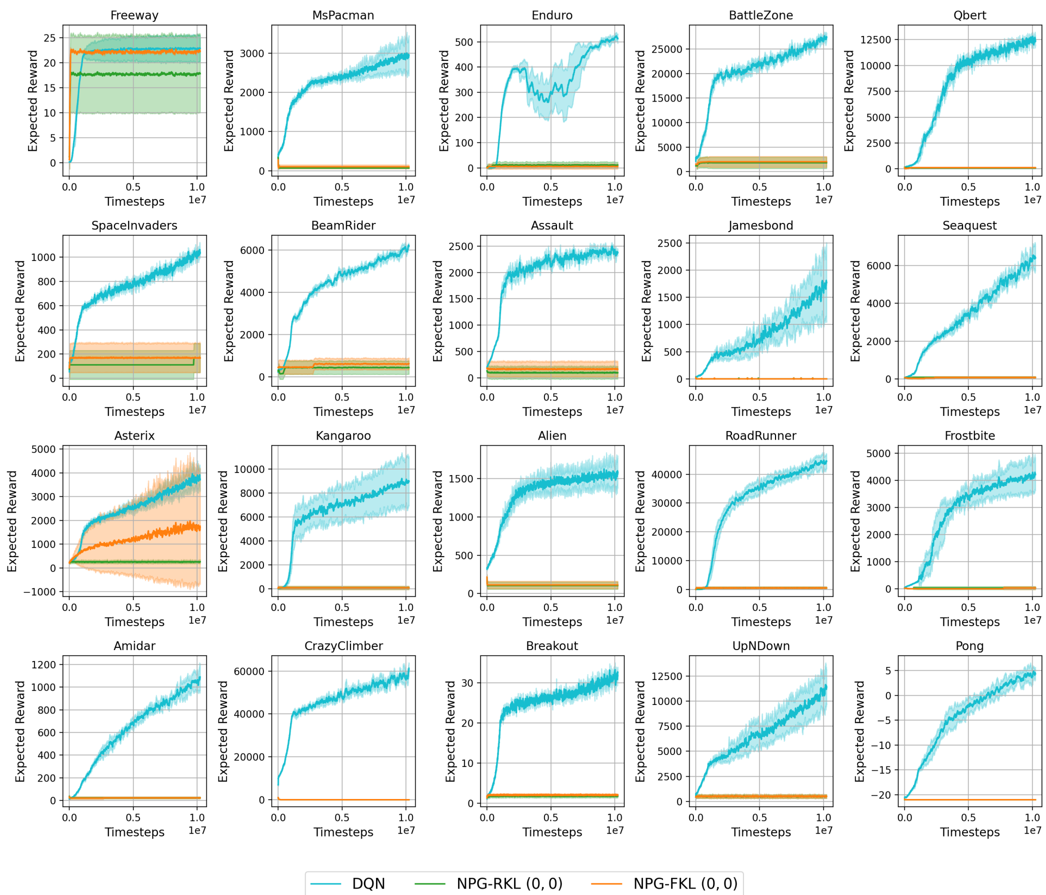

Figure 9: Direction of KL divergence without entropy regularization, using $n = 1$ and NPG as the intermediate policy. Across nearly all games, performance remains poor relative to DQN, and the KL direction has no conclusive effect.

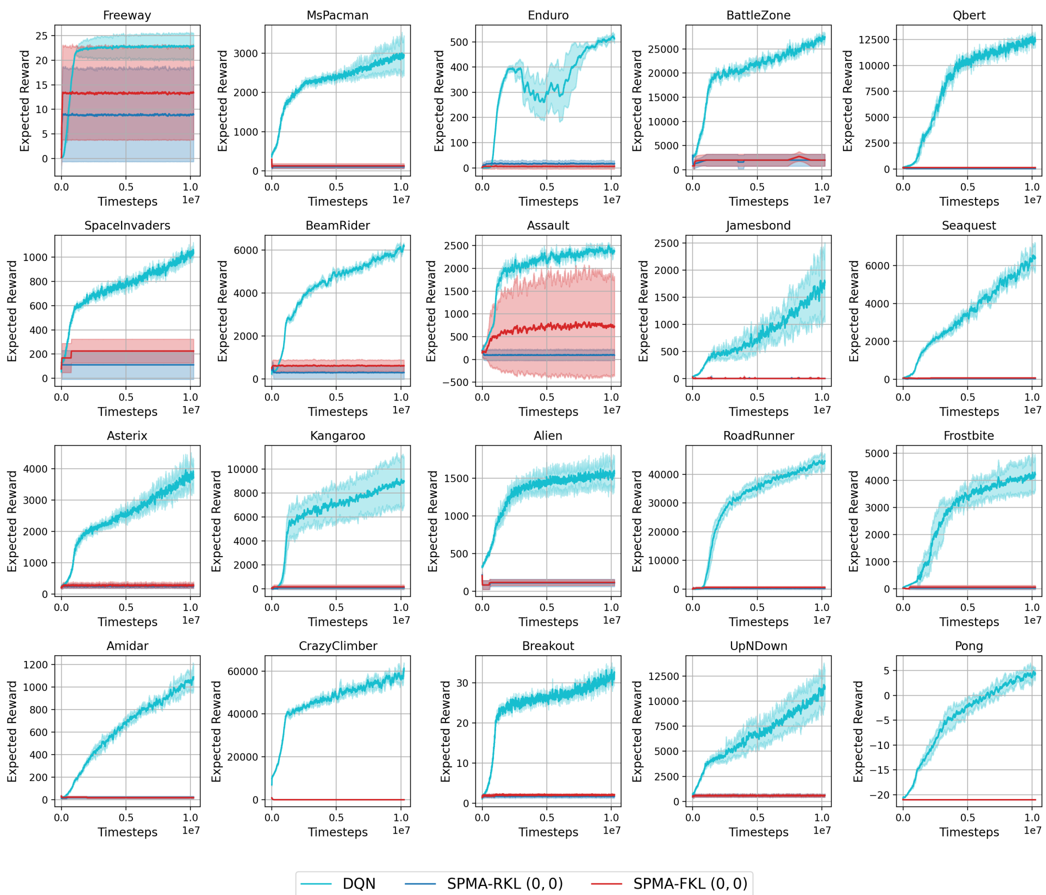

Figure 10: Direction of KL divergence without entropy regularization, using $n = 1$ and SPMA as the intermediate policy. Across nearly all games, performance remains poor relative to DQN, and the KL direction has no conclusive effect.

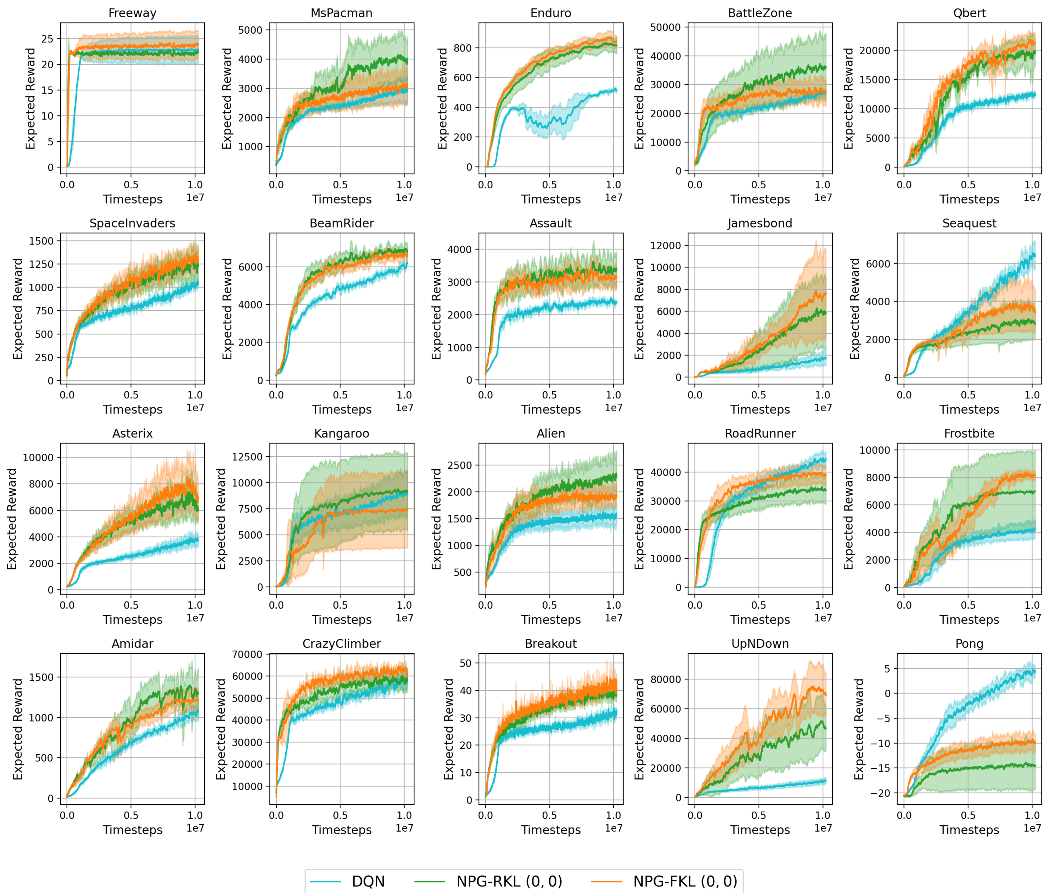

Figure 11: Direction of KL divergence without entropy regularization, using $n = 10$ and `NPG` as the intermediate policy. Both `NPG-RKL`$(0, 0)$ and `NPG-FKL`$(0, 0)$ achieve performance comparable to `DQN`, but no KL direction consistently outperforms the other.

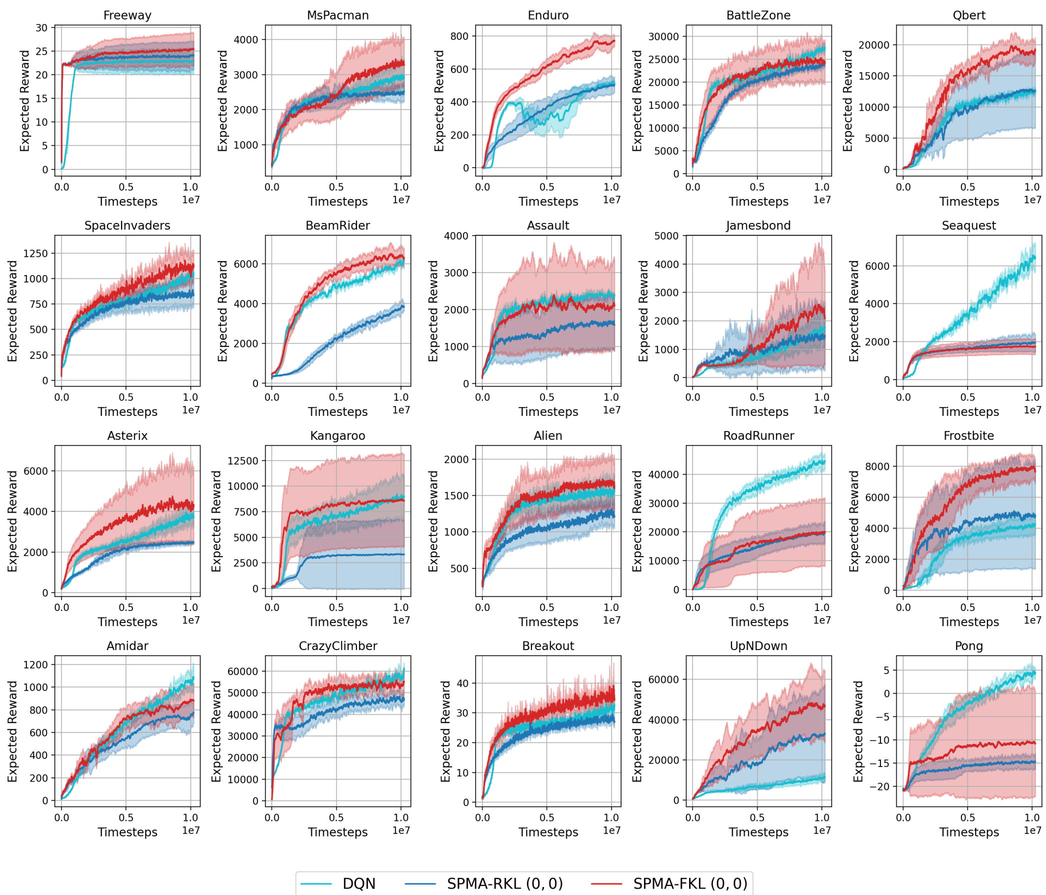

Figure 12: Direction of KL divergence without entropy regularization, using $n = 10$ and SPMA as the intermediate policy. SPMA-FKL$(0, 0)$ achieves performance comparable to DQN, and overall, the forward KL direction appears to outperform SPMA-RKL$(0, 0)$.

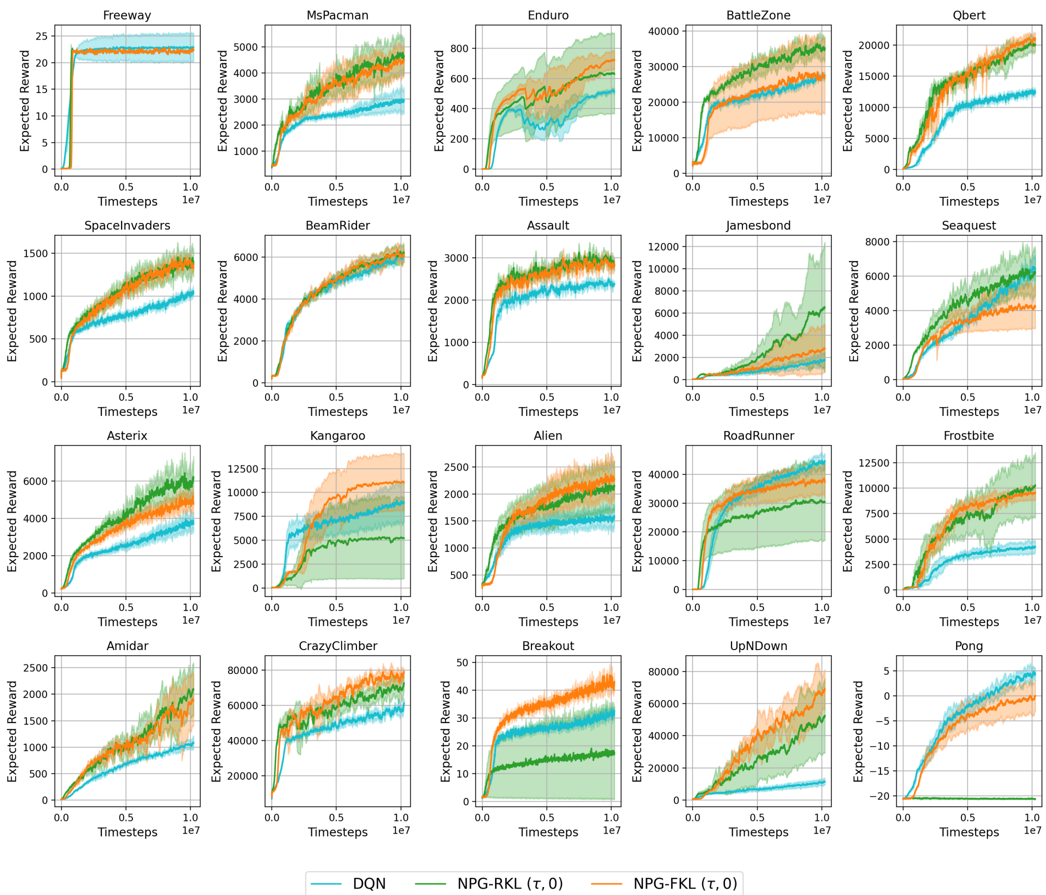

Figure 13: Direction of KL divergence with actor entropy regularization, using $n = 1$ and NPG as the intermediate policy. Both NPG-FKL$(\tau, 0)$ and NPG-RKL$(\tau, 0)$ achieve performance comparable to DQN, even with $m = 1$ but no KL direction consistently outperforms the other.

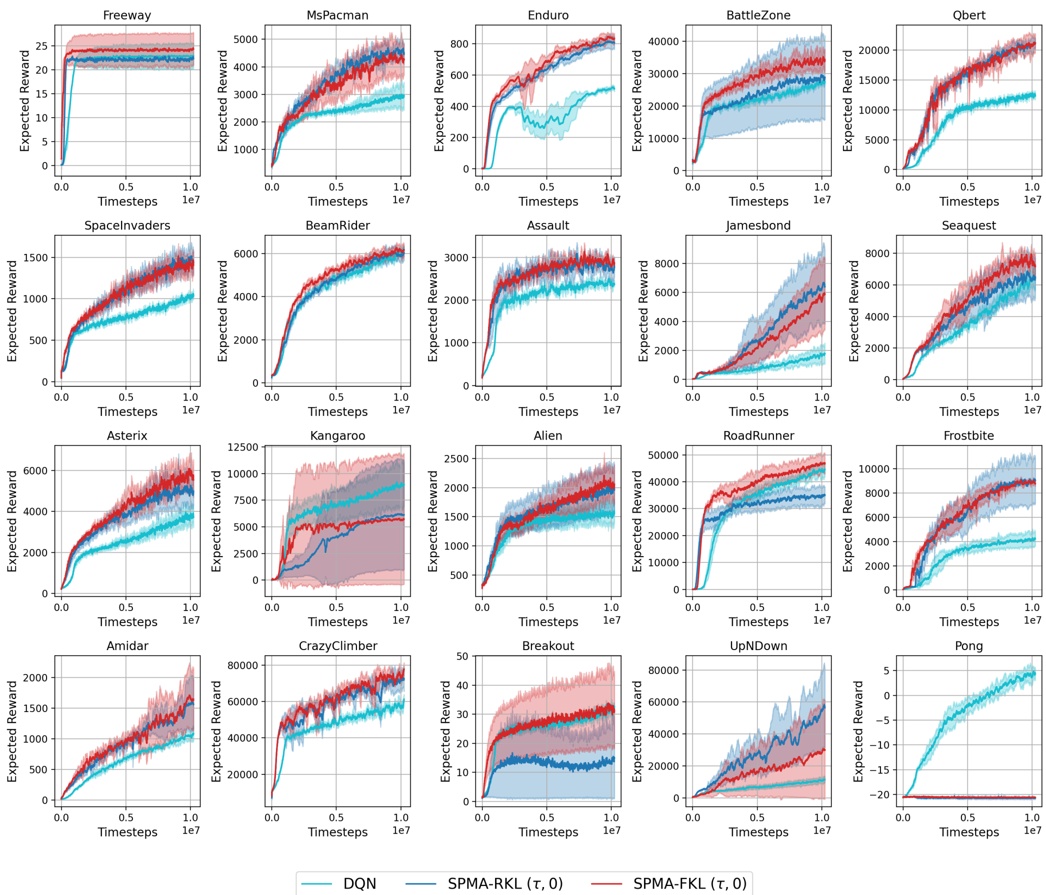

Figure 14: Direction of KL divergence with actor entropy regularization, using $n = 1$ and SPMA as the intermediate policy. In the presence of actor entropy, both SPMA-FKL$(\tau, 0)$ and SPMA-RKL$(\tau, 0)$ achieve performance comparable to DQN, even with $m = 1$, but no KL direction consistently outperforms the other.

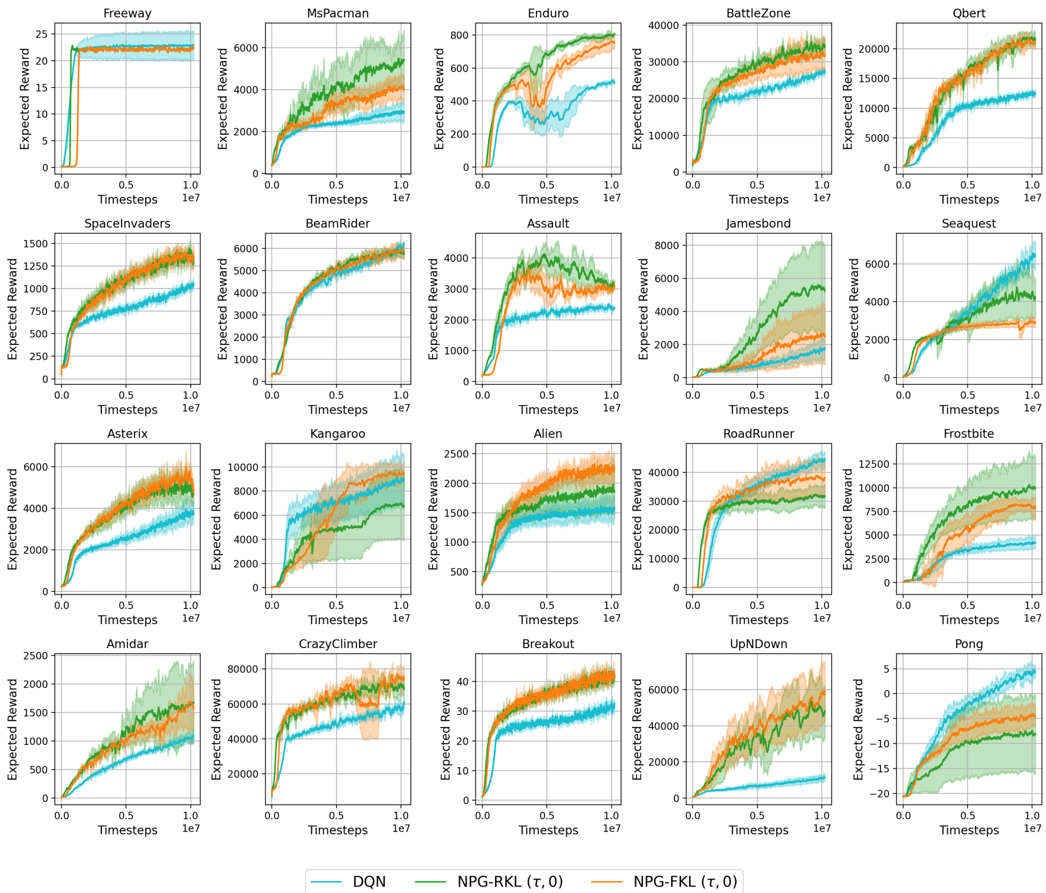

Figure 15: Direction of KL divergence with actor entropy regularization, using $n = 10$ and NPG as the intermediate policy. In the presence of actor entropy, both NPG-FKL$(\tau, 0)$ and NPG-RKL$(\tau, 0)$ achieve performance comparable to DQN, but no KL direction consistently outperforms the other.

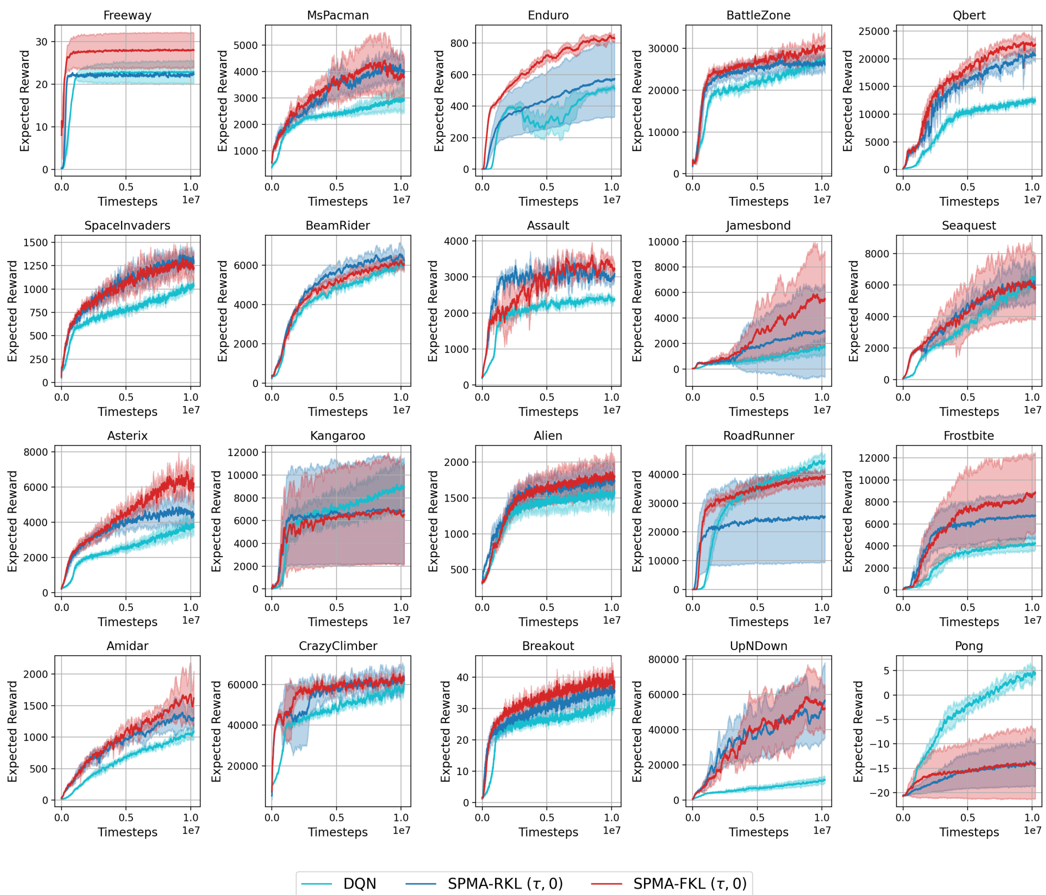

Figure 16: Direction of KL divergence with actor entropy regularization, using $n = 10$ and SPMA as the intermediate policy. In the presence of actor entropy, both SPMA-FKL$(\tau, 0)$ and SPMA-RKL$(\tau, 0)$ achieve performance comparable to DQN, but no KL direction consistently outperforms the other.