BEYOND SOFTMAX AND ENTROPY: f-REGULARIZED POLICY GRADIENTS WITH COUPLED PARAMETRIZATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce f-PG, a new class of stochastic policy gradient methods regularized by a family of f-divergences, including entropy and Tsallis divergences. For each divergence, we employed a *coupled* parameterization, defined by f-softargmax, which allows us to establish the first explicit, non-asymptotic, last-iterate convergence rates for stochastic policy gradient. To derive our analysis, we prove that the f-regularized value function is smooth and satisfies a Polyak-Łojasiewicz inequality as a function of f-softargmax parameters. To establish the latter, we introduce a general policy improvement operator that restricts optimization to a well-defined policy space that excludes ill-behaved policies. In the case of softmax, this allows to escape the "gravitational pull" and yields the first *explicit* convergence guarantees for this parameterization, closing a gap in the literature. Finally, we leverage these rates to derive sample complexity bounds for the unregularized problem and show that f-PG with Tsallis divergences provides a provably better sample complexity/regularization bias trade-off compared to softmax-based policy gradient with entropy regularization.

1 INTRODUCTION

Regularization has become a cornerstone of modern Reinforcement Learning (RL), playing a central role in many of its key breakthroughs. A prominent example is the use of Kullback–Leibler (KL) penalization, which underlies algorithms such as Trust-Region Policy Optimization (Schulman et al., 2015) and Mirror Descent Policy Optimization (Tomar et al., 2022), where it ensures stability by constraining policy updates. Although the KL divergence is by far the most widely adopted regularizer in RL, recent advances have highlighted the benefits of alternative regularizers based on other f-divergences. For example, Lee et al. (2019) demonstrated empirically that Tsallis regularization leads to improved performance in continuous control tasks. On the theoretical front, the Tsallis-INF algorithm for multi-armed bandits (Zimmert & Seldin, 2021) achieves minimax optimality, leveraging Tsallis entropy regularization as a key component. This motivates the need for a deeper theoretical framework that goes beyond KL divergence.

In this paper, we propose new policy gradient methods for a family of f-divergence regularized RL problems, based on *coupled parametrizations*. As an example, the KL divergence is usually coupled with the softmax parameterization, improving convergence of different RL methods (Schulman et al., 2015; Haarnoja et al., 2017; 2018; Tomar et al., 2022). Global convergence rates of the policy gradient method under this coupling were derived in the deterministic setting Mei et al. (2020b), although without explicit constants. In this case, the *softmax* parameterization has a natural interpretation: the optimal parameters correspond to temperature-rescaled optimal Q-values up to a baseline function (Mei et al., 2020b). However, this is no longer the case when coupled with other types of regularization. Additionally, the *softmax* as a parametrization suffers from the *gravitational well* and the softmax *softmax dumping* effects (Mei et al., 2020a). This motivates the question: *"what is the appropriate parameterization for an f-regularized policy gradient method?"*

We argue that the appropriate parameterization follows from the problem's structure: every convex f-divergence induces a canonical parameterization via the *f-softargmax operator* (Blondel et al., 2020; Roulet et al., 2025). We call this the *coupled parametrization*, emphasizing the intrinsic link between divergence and parameterization. From this perspective, the familiar softmax–entropy pair is a special case, with the choice $f\colon u \mapsto u \log u - (u-1)$. To illustrate the strength of this coupling, we study the policy gradient method for a family of f-divergence regularized RL problems with *coupled parameterization*. We obtain last-iterate global convergence guarantees in the stochastic setting, in the tabular case, which is new even in the standard entropy-softmax case (Mei et al., 2020b; Agarwal et al., 2021; Cen et al., 2022; Müller & Cayci, 2024; Ding et al., 2025). Our results highlight that using alternatives to the entropy-softmax can lead to faster convergence. To derive these results, we introduce a novel *policy-improvement operator*, which discards ill-behaved policies for all types of parameterizations. This procedure, in the case of the *softmax* parameterization, escapes the *gravitational pull* (Mei et al., 2020a). Our contributions are threefold:

- **f-regularized RL via coupled parametrizations.** We introduce policies parameterized by the f-softargmax operator induced by the chosen divergence, thereby coupling regularization and parametrization. Leveraging the efficient implementation of f-softargmax (Roulet et al., 2025), our method is computationally efficient while allowing to naturally exploit the geometry of different regularizers.
- **Global convergence guarantees.** We prove global convergence rates, with *explicit constants*, for the stochastic policy gradient method with coupled parameterization. Our analysis relies on a novel characterization of the regularized value function's smoothness and the Polyak–Łojasiewicz property. These rates yield finite-sample complexity with convergence rates that depend on the choice of the f-regularizer.
- **Better sample complexity/regularization bias trade-off.** We establish that using divergences beyond entropy yields a better sample-complexity/regularization-bias trade-off than the classical softmax–entropy pair, highlighting the theoretical benefits of broader f-regularization. Importantly, in the case of α-Tsallis regularised PG, we show that the optimal α depends on the desired precision on the unregularised problem.

2 Related Work

KL regularization in policy gradients. Entropy and KL regularization are standard tools for stabilizing RL (Haarnoja et al., 2017; 2018; Nachum et al., 2017; Abdolmaleki et al., 2018; Vieillard et al., 2020), including in policy gradient methods (Schulman et al., 2015; Tomar et al., 2022). Under softmax parameterization, they enjoy global convergence guarantees in tabular settings (Mei et al., 2020b; Agarwal et al., 2021), but convergence can take exponential time in the worst case (Mei et al., 2020a; Li et al., 2023), revealing limitations of the softmax parameterization.

Alternatives to softmax. Outside of the RL literature, alternative parametrisations to softmax have been proposed and shown prominent results (Martins & Astudillo, 2016; Peters et al., 2019; Roulet et al., 2025). Yet in the RL setting, such alternatives are scarce, with the notable exception of the escort transform (Mei et al., 2020a; Liu et al., 2025), which does not exploit the geometry of the f-regularized objective.

f-Divergence Regularization. Beyond KL, general frameworks for f-divergence regularization in RL have been developed (Belousov & Peters, 2017; Geist et al., 2019). A notable particular case is Tsallis entropy regularization (Chow et al., 2018; Lee et al., 2019), which interpolates between mode-seeking and mode-covering exploration behaviors. Other works have applied f-divergence regularization in settings such as offline RL (Sikchi et al., 2024) and goal-conditioned RL (Agarwal et al., 2023). More recently, it has also proven useful for fine-tuning large language models (Go et al., 2023; Wang et al., 2024; Huang et al., 2025; Li et al., 2025), underscoring its practical relevance. Yet, unlike the KL case where softmax arises naturally, no principled parameterization is known for general f-regularization. This gap is the focus of our work.

3 BACKGROUND

Markov Decision Process. We consider a discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, \mathsf{P}, \mathsf{r}, \rho)$ with finite $\mathcal{S}, \mathcal{A}$, discount $\gamma \in (0,1)$, transition kernel $\mathsf{P}(s'|s,a)$, reward $\mathsf{r}(s,a) \in [0,1]$, and initial distribution ρ. A stationary policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ induces $\mathsf{P}_\pi(s'|s) = \sum_a \mathsf{P}(s'|s,a)\pi(a|s)$. For $v : \mathcal{S} \to \mathbb{R}$, set $\mathsf{P}v(s,a) = \sum_{s'} \mathsf{P}(s'|s,a)v(s')$ and $\mathsf{P}_\pi v(s) = \sum_{s'} \mathsf{P}_\pi(s'|s)v(s')$. The value function is

$$v_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathsf{r}(S_t, A_t) \,\middle|\, S_0 = s\right], \tag{1}$$

with $A_t \sim \pi(\cdot|S_t)$, $S_{t+1} \sim \mathsf{P}(\cdot|S_t, A_t)$. The Bellman operator is $\mathsf{T}_\pi v(s) = \mathsf{r}_\pi(s) + \gamma \mathsf{P}_\pi v(s)$, where $\mathsf{r}_\pi(s) = \sum_a \pi(a|s)\mathsf{r}(s,a)$. It is a γ-contraction with unique fixed point v_π. For $v \in \mathbb{R}^{|\mathcal{S}|}$, define Q-function $q(s,a) = \mathsf{r}(s,a) + \gamma \mathsf{P}v(s,a)$, yielding $\mathsf{T}_\pi v(s) = \langle \pi(\cdot|s), q(s,\cdot)\rangle$. The Bellman optimality operator $\mathsf{T}_\star v = \max_\pi \mathsf{T}_\pi v$ has fixed point $v_\star$ with optimal Q-function $q_\star$. For $\rho \in \mathcal{P}(\mathcal{S})$, define $v_\pi(\rho) = \sum_s \rho(s) v_\pi(s)$. The objective is to maximize $v_\pi(\rho)$; if ρ has full support, this reduces to finding an optimal policy $\pi_\star$ such that $v_{\pi_\star} = v_\star$, with a support of a subset of $\arg\max_a q_\star(s,a)$ for all $s \in \mathcal{S}$.

f-Regularized MDP. Let $f\colon (0,\infty) \to \mathbb{R}$ be a strictly convex *generator* with $f(1) = 0$. Its *adjoint* (or reverse generator) is $f_\dagger(u) := u\, f(1/u)$, $u > 0$, which is convex, strictly convex if f is, and satisfies $f_\dagger(1) = 0$. Boundary conventions are $f(0) := \lim_{u \downarrow 0} f(u) \in (-\infty, \infty]$ and $f_\dagger(0) := \lim_{u \downarrow 0} f_\dagger(u) = \lim_{t \uparrow \infty} f(t)/t \in (-\infty, \infty]$. For $p, q \in \mathcal{P}(\mathcal{A})$ over finite $\mathcal{A}$, the f–divergence is

$$\mathrm{D}^f(p\|q) \;:=\; \sum_{a \in \mathcal{A}: q(a) > 0} q(a) f(p(a)/q(a)) + f_\dagger(0) \sum_{a \in \mathcal{A}: q(a) = 0} p(a), \tag{2}$$

with conventions: $q(a)f(0)$ if $q(a) > 0, p(a) = 0$, and 0 if $p(a) = q(a) = 0$ (Rényi, 1961; Csiszár, 1967; Liese & Vajda, 2006). f–divergences satisfy $\mathrm{D}^f(p\|q) \in [0,\infty]$, are jointly convex (Csiszár, 1967), vanish iff $p = q$, and are not symmetric ($\mathrm{D}^{f_\dagger}(p\|q) = \mathrm{D}^f(p\|q)$).

Let $\pi_{\mathrm{ref}}\colon \mathcal{S} \to \mathcal{P}(\mathcal{A})$ be a reference policy. For $\lambda > 0$, the *regularized Bellman operator* is $\mathsf{T}_\pi^f v(s) := \mathsf{T}_\pi v(s) - \lambda\, \mathrm{D}^f(\pi(\cdot|s)\|\pi_{\mathrm{ref}}(\cdot|s))$, $s \in \mathcal{S}$. (Geist et al., 2019, Prop. 2) ensures a unique fixed point v_π^f, the regularized value function, with

$$v_\pi^f(s) = \textstyle\sum_a \pi(a|s)\mathsf{r}(s,a) - \lambda\, \mathrm{D}^f(\pi(\cdot|s)\|\pi_{\mathrm{ref}}(\cdot|s)) + \gamma \sum_{a,s'} \pi(a|s)\mathsf{P}(s'|s,a)\, v_\pi^f(s'). \tag{3}$$

The *regularized optimal Bellman operator* is $\mathsf{T}_\star^f v(s) = \max_\pi\{\langle \pi(\cdot|s), q(s,\cdot)\rangle - \lambda\, \mathrm{D}^f(\pi(\cdot|s)\|\pi_{\mathrm{ref}}(\cdot|s))\}$, with unique fixed point $v_\star^f$ and associated Q–function $q_\star^f$ given by:

$$v_\star^f(s) = \max_{\nu \in \mathcal{P}(\mathcal{A})}\{\langle \nu, q_\star^f(s,\cdot)\rangle - \lambda\, \mathrm{D}^f(\nu\|\pi_{\mathrm{ref}}(\cdot|s))\}, \quad q_\star^f(s,a) = \mathsf{r}(s,a) + \gamma \mathsf{P} v_\star^f(s,a) \tag{4}$$

$$\pi_\star^f(\cdot|s) := \arg\max_{\nu \in \mathcal{P}(\mathcal{A})}\{\langle \nu, q_\star^f(s,\cdot)\rangle - \lambda\, \mathrm{D}^f(\nu\|\pi_{\mathrm{ref}}(\cdot|s))\}. \tag{5}$$

As in the unregularized case, the objective is to maximize $v_\pi^f(\rho) = \sum_s \rho(s) v_\pi^f(s)$ over π.

Operators generated by f-divergences. For $x \in \mathbb{R}^{|\mathcal{A}|}$ and $q \in \mathcal{P}(\mathcal{A})$, Roulet et al. (2025) define the following general operators

$$\mathrm{softmax}^f(x,q) = \max_{\nu \in \mathcal{P}(\mathcal{A})} \big\{\langle \nu, x\rangle - \mathrm{D}^f(\nu\|q)\big\}, \quad \mathrm{softargmax}^f(x,q) = \arg\max_{\nu \in \mathcal{P}(\mathcal{A})} \big\{\langle \nu, x\rangle - \mathrm{D}^f(\nu\|q)\big\}.$$

For simplicity, we assume in the sequel that q has *full support*, i.e. that $q(a) > 0$ for all $a \in \mathcal{P}(\mathcal{A})$. Since $\mathrm{D}^f(p\|q)$ is strictly convex in its first argument on the simplex (Csiszár, 1967), the output of the $\mathrm{softargmax}^f$ operator is well defined and unique, as it corresponds to the $\arg\max$ of a strictly concave

function over a compact set. These definitions are tightly connected to the Fenchel-Legendre conjugate of $\Psi := \mathrm{D}^f(\cdot\|q) + \iota_{\mathcal{P}(\mathcal{A})}$, where $\iota_{\mathcal{P}(\mathcal{A})}$ denotes the convex indicator function of the probability simplex. Using these definitions, (4) and (5) can be rewritten as

$$v^f_\star(s) = \tau \cdot \mathrm{softmax}^f(q^f_\star(s,\cdot)/\tau, \pi_{\mathrm{ref}}(\cdot|s)), \quad \pi^f_\star(\cdot|s) = \mathrm{softargmax}^f(q^f_\star(s,\cdot)/\tau, \pi_{\mathrm{ref}}(\cdot|s)).$$

The efficient way to compute $\mathrm{softmax}^f$ and $\mathrm{softargmax}^f$ follow from (Roulet et al., 2025, Proposition 1) that (see also lemma B.1)

Proposition 3.1 (Follows from Roulet et al. 2025, Proposition 1)**.** *Assume f is strictly convex and C^1 on $(0,\infty)$ with $\lim_{u\downarrow 0^+} f'(u) = -\infty$, and $q(a) > 0$ for all a. Then for every $x \in \mathbb{R}^{|\mathcal{A}|}$, $a \in \mathcal{A}$:*

$$\mathrm{softargmax}^f(x,q)[a] = q(a)\left[f'\right]^{-1}\big(x(a) - \mu_x\big), \tag{6}$$

where μ_x is the unique root of $F(\mu) := \sum_{a\in\mathcal{A}} q(a)\left[f'\right]^{-1}\big(x(a)-\mu\big) - 1 = 0$ Moreover, F is strictly decreasing so μ_x can be computed by bisection in the bracket is $\mu_x \in (\max_a x(a) - f'(1/q_{\min}),\ \max_a x(a) - f'(1))$, with $q_{\min} := \min_a q(a)$. Moreover, denoting by f^* the convex conjugate of f,

$$\mathrm{softmax}^f(x,q) = \min_{\mu\in\mathbb{R}} \Big\{\mu + \sum_a q(a)\, f^*(x(a)-\mu)\Big\}, \qquad \nabla_x \mathrm{softmax}^f(x;q) = \mathrm{softargmax}^f(x;q). \tag{7}$$

Corollary 3.2 (KL case)**.** *For the (generalized) KL-divergence, the generator is $f(u) = u\log u - (u-1)$, $\mathrm{softmax}^{\mathrm{KL}}(x,q) = \log\sum_{a\in\mathcal{A}} q(a)\mathrm{e}^{x(a)}$ and $\mathrm{softargmax}^{\mathrm{KL}}(x,q)[a] = q(a)\,\mathrm{e}^{x(a)}/\sum_{b\in\mathcal{A}} q(b)\mathrm{e}^{x(b)}$.*

With uniform $q = (1/|\mathcal{A}|,\dots,1/|\mathcal{A}|)$, we recover the classical `mellowmax` and `softmax` operators (Asadi & Littman, 2017). For a real parameter $\alpha \neq 1$ (the entropic index), define $\exp_\alpha(x) = [1+(\alpha-1)x]_+^{1/(\alpha-1)}$ and $\log_\alpha(y) = (y^{\alpha-1}-1)/(\alpha-1)$ where $[z]_+ = \max\{z,0\}$ denotes truncation at zero and for $\alpha \in (0,1)$, $\exp_\alpha$ is defined only for $x < 1/(1-\alpha)$.

Corollary 3.3 (Tsallis, $0 < \alpha < 1$)**.** *For the α-Tsallis divergence generator $f_\alpha(u) = (u\log_\alpha(u) - (u-1))/\alpha$, $\mathrm{softargmax}^{f_\alpha}(x,q)[a] = q(a)\exp_\alpha(x(a) - \mu_\alpha(x,q))$, where $\mu_\alpha(x;q)$ is the unique root of $\sum_{a\in\mathcal{A}} q(a)\exp_\alpha(x(a)-\mu) = 1$; see (Lee et al., 2019; Roulet et al., 2025) and Appendix B.1.*

4 Regularized Value Function with Coupled Parametrization

In this work, we aim to optimize the f-regularized value function over a parameterized policy class,

$$\max_{\theta\in\Theta}\big\{J^f(\theta) := v^f_{\pi_\theta}(\rho)\big\}\ , \tag{8}$$

where the policy is given by a parameterized $\theta \mapsto \pi_\theta \in \mathcal{P}(\mathcal{A})^{|\mathcal{S}|}$.

f-regularized RL and coupled parameterization. The analytic form of the optimal policy (5) naturally suggests a *coupled parameterization* based on the f-softargmax operator. We define the policy induced by the coupled parameterization, for any $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and state $s \in \mathcal{S}$, as

$$\pi^f_\theta(\cdot|s) := \mathrm{softargmax}^f(\theta(s,\cdot), \pi_{\mathrm{ref}}(\cdot|s)) = \arg\max_{\nu\in\mathcal{P}(\mathcal{A})} \Big\{\langle \nu, \theta(s,\cdot)\rangle - \mathrm{D}^f(\nu\|\pi_{\mathrm{ref}}(\cdot|s))\Big\}\ . \tag{9}$$

This choice aligns directly with the structure of the f-regularized solution. In particular, under this parameterization, solutions to (8) can be written as

$$\theta^f_\star(s,a) = q^f_\star(s,a)/\lambda + b(s)\ ,$$

where $q^f_\star(s,a)$ is the regularized optimal Q-function defined in (4) and $b\colon \mathcal{S}\to\mathbb{R}$ is an arbitrary baseline function. Hence, learning the parameters boils down to recovering the regularized optimal Q-values, up to a baseline and rescaling by $1/\lambda$, and the induced policy satisfies $\pi^f_{\theta^\star} = \pi^f_\star$, where $\pi^f_\star$ is the regularized optimal policy, defined in (5). When applied to KL regularization, we retrieve Mei et al. (2020b).

***Remark* 4.1** (Connection with (Lazy) Mirror Descent.) *We stress that, despite similarities, f-PG is fundamentally different from mirror descent. With $\Phi(\pi) = \sum_{s\in\mathcal{S}} \mathrm{D}^f(\pi(\cdot|s)\|\pi_{\mathrm{ref}}(\cdot|s))$, mirror descent is*

$$\nabla\Phi(\widetilde{\pi}_{t+1}) = \nabla\Phi(\widetilde{\pi}_t) + \eta\nabla_\pi v^f_\pi(\rho)|_{\pi=\pi_t}\,, \quad \pi_{t+1} = \arg\min_{\pi\in\Pi}\{\Phi(\pi) - \langle\nabla\Phi(\widetilde{\pi}_{t+1}), \pi - \widetilde{\pi}_{t+1}\rangle\}\,. \tag{10}$$

where $\Pi = \mathcal{P}(\mathcal{A})^{|\mathcal{S}|}$ is a policy space. Denoting $\theta_t = \nabla\Phi(\widetilde{\pi}_t)$, one obtains updates that resemble (14) (without $\mathcal{T}$), with one key difference: the gradient in (10) is taken with respect to the policy π whereas in (14) it is computed w.r.t the "dual" parameter θ (in the mirror descent terminology). Moreover, the update (10) can be expressed as, by the chain rule, $\theta_{t+1} = \theta_t + \eta[\frac{\partial\pi_\theta}{\partial\theta}|_{\theta=\theta_t}]^{-1}\nabla_\theta J^f(\theta_t)$, which have an additional preconditioning term given by the inverse of the policy Jacobian. See Appendix H for more details.

Properties of f-regularized objective under coupled parameterization. We now establish two key properties of the f-regularized value with coupled parametrization: smoothness of the objective, and a Polyak-Łojasiewicz inequality. We derive these properties under following two assumptions on f and π_{ref}.

Assumption P($\underline{\pi_{\mathrm{ref}}}$). *There exists $\underline{\pi_{\mathrm{ref}}} > 0$ such that $\min_{(s,a)}\pi_{\mathrm{ref}}(a|s) > \underline{\pi_{\mathrm{ref}}}$.*

Assumption A$_f$($\underline{\pi_{\mathrm{ref}}}$). *The function f satisfies the following properties.*

(i) f is bounded and strictly convex on $[0; 1/\underline{\pi_{\mathrm{ref}}}]$, $f(1) = 0$, and is thrice differentiable on $(0, 1/\underline{\pi_{\mathrm{ref}}})$;

(ii) $\lim_{u\downarrow 0^+} f'(u) = -\infty$, and $\lim_{u\to 0}|f'(u)/f''(u)| < \infty$;

(iii) there exists $1 \le \omega_f < \infty$, and $\kappa_f < \infty$, such that for any $u \in [0; 1/\underline{\pi_{\mathrm{ref}}}]$, we have

$$1/(uf''(u)) \le \omega_f \quad , \quad \text{and} \quad |f'''(u)/f''(u)^2| \le \kappa_f \quad ;$$

(iv) there exists $1 \ge \iota_f > 0$ such that f'' decreases on $[0; \iota_f]$ and for any $u \in [\iota_f; 1/\underline{\pi_{\mathrm{ref}}}]$, $f''(\iota_f) \ge f''(u)$.

These conditions are met by a broad class of divergences commonly used in practice, like the KL and Tsallis divergences. These divergences admit finite constants ω_f,κ_f, and ζ_f; see Appendix F for explicit computations. In contrast, Tsallis divergences with $\alpha > 1$ violate condition (ii): since $f'(0)$ is finite, the induced policies may place zero mass on some actions, leading to sparsity and consequently to non-Lipschitz gradients. Similarly, for reverse KL regularization with $f(u) = -\log(u)$, the function f is unbounded as $u \to 0$, which prevents us from controlling the effect of regularization in the worst case.

Under **P**($\underline{\pi_{\mathrm{ref}}}$) and **A**$_f$($\underline{\pi_{\mathrm{ref}}}$), we can define the logits $\mathrm{w}^f_\theta(a|s)$ and their sum $\mathrm{W}^f_\theta(s)$, defined as

$$\mathrm{w}^f_\theta(a|s) = \frac{1}{\mathrm{W}^f_\theta(s)}\frac{\pi_{\mathrm{ref}}(a|s)}{f''(\pi^f_\theta(a|s)/\pi_{\mathrm{ref}}(a|s))}\,, \quad \text{with} \quad \mathrm{W}^f_\theta(s) = \sum_{a\in\mathcal{A}}\frac{\pi_{\mathrm{ref}}(a|s)}{f''(\pi^f_\theta(a|s)/\pi_{\mathrm{ref}}(a|s))}\,, \tag{11}$$

which will play a central role in our analysis. In the KL divergence case, we recover simple expressions $\mathrm{w}^f_\theta(a|s) = \pi^f_\theta(a|s)$ and $\mathrm{W}^f_\theta(s) = 1$. Using the notations in (11), the Jacobian of the policy w.r.t. θ is

$$\frac{\partial\pi^f_\theta(a|s)}{\partial\theta(s',b)} = 1_{s'}(s)\,\mathrm{W}^f_\theta(s)\left[1_b(a)\,\mathrm{w}^f_\theta(a|s) - \mathrm{w}^f_\theta(a|s)\,\mathrm{w}^f_\theta(b|s)\right]\,.$$

We also defined the following three quantities, which are bounded under **P**($\underline{\pi_{\mathrm{ref}}}$) and **A**$_f$($\underline{\pi_{\mathrm{ref}}}$).

$$\mathrm{y}_f := \max_{(s,\nu)\in\mathcal{S}\times\mathcal{P}(\mathcal{A})}\sum_{a\in\mathcal{A}}\pi_{\mathrm{ref}}(a|s)\frac{|f'(\nu(a)/\pi_{\mathrm{ref}}(a|s))|}{f''(\nu(a)/\pi_{\mathrm{ref}}(a|s))}\,, \quad \mathrm{d}_f := \max_{(s,\nu)\in\mathcal{S}\times\mathcal{P}(\mathcal{A})}\mathrm{D}^f(\nu\|\pi_{\mathrm{ref}}(\cdot|s))\,, \tag{12}$$

$$0 < \zeta_f := \min_{(s,\nu)\in\mathcal{S}\times\mathcal{P}(\mathcal{A})}\sum_{a\in\mathcal{A}}\frac{\pi_{\mathrm{ref}}(a|s)}{f''(\nu(a)/\pi_{\mathrm{ref}}(a|s))}\,. \tag{13}$$

Given **P**($\underline{\pi_{\mathrm{ref}}}$) and **A**$_f$($\underline{\pi_{\mathrm{ref}}}$), we prove that the regularized objective value function with coupled parametrization is smooth and satisfies a Łojasiewicz-type condition.

Algorithm 1 f-Regularized Policy Gradient with Improvement Projection

Initialization: Learning rate $\eta > 0$, initial parameter θ_0, divergence generator f.
for $t = 0$ to $T - 1$ **do**
  Collect B trajectories of length H: $Z_t := (S^b_{t,0:H-1}, A^b_{t,0:H-1})_{b=0}^{B-1}$ using $\pi^f_{\theta_t}$
  Compute $\bar{\theta}_{t+1} = \theta_t + \eta \mathrm{g}^f_{Z_t}(\theta_t)$ where $\mathrm{g}^f_{Z_t}(\theta_t)$ is computed using (15)
  Perform projection: $\theta_{t+1} = \mathcal{T}_{\tau_{\lambda,f}}(\bar{\theta}_{t+1})$ where $\mathcal{T}_{\tau_{\lambda,f}}$ (see Section 5 and Appendix D)

Theorem 4.2. *Assume that, for some $\underline{\pi_{\mathrm{ref}}} > 0$, $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\mathrm{ref}}})$ hold. For any $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, we have*

$$\left\|\nabla^2 v^f_\theta(\rho)\right\|_2 \leq L_{\lambda,f} \quad , \quad \text{where } L_{\lambda,f} \approx \frac{\omega_f[\max(\omega_f,\kappa_f) + \lambda \max(\omega_f \mathrm{d}_f, \kappa_f \mathrm{d}_f, \mathrm{y}_f, \kappa_f \mathrm{y}_f/\omega_f, 1)]}{(1-\gamma)^3} \quad .$$

We give explicit constants and prove this theorem in Appendix B.3's Theorem B.12. This guarantees that the regularized value function $v^f_\theta(\rho)$ is a smooth function of the parameter θ. Now, we derive a Łojasiewicz inequality, provided that the classical sufficient exploration assumption holds.

Assumption $\mathbf{A}_\rho$. *The smallest coefficient $\rho_{\min} := \min_{s\in\mathcal{S}} \rho(s)$ of the initial distribution ρ satisfies $\rho_{\min} > 0$.*

Theorem 4.3. *Assume that, for some $\underline{\pi_{\mathrm{ref}}} > 0$, $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\mathrm{ref}}})$ hold. Assume in addition that the initial distribution ρ satisfies $\boldsymbol{A}_\rho$. Then, it holds that*

$$\left\|\frac{\partial v^f_\theta(\rho)}{\partial \theta}\right\|_2^2 \geq \mu_{\lambda,f}(\theta)\left(v^f_\star(\rho) - v^f_\theta(\rho)\right) \quad , \quad \text{where } \mu_{\lambda,f}(\theta) := \lambda(1-\gamma)\rho^2_{\min}(\zeta_f/\omega_f)^2 \min_{s,a} \mathrm{w}^f_\theta(a|s)^2 \quad .$$

We prove this theorem in Appendix C.à Note that the correspondence $\theta^f_\star = q^f_\star/\lambda + b$ is crucial to establish the Polyak-Łojasiewicz condition. For KL-regularization, we retrieve a property outlined for KL-softmax by Mei et al. (2020b), but our proof, based on the properties of Fenchel-Legendre conjugation, is much simpler.

5 Convergence Analysis of f-PG

The f-PG algorithm. We introduce f-PG (Algorithm 1), a regularized policy gradient method with coupled parameterization. At each iteration of f-PG, the agent samples a batch of independent truncated trajectories of length H from $\nu(\pi;\cdot)$ defined for $z = (s_h, a_h)_{h=0}^{H-1} \in (\mathcal{S}\times\mathcal{A})^H$ by $\nu(\pi;z) = \rho(s_0)\pi(a_0|s_0)\prod_{h=0}^{H-1} \mathsf{P}(s_h \mid s_{h-1}, a_{h-1})\pi(a_h|s_h)$. The agent then performs the update

$$\theta_{t+1} = \mathcal{T}_\tau\left(\theta_t + \eta \cdot \mathrm{g}^f_{Z_t}(\theta_t)\right) \quad , \quad \text{for } t \geq 0 \quad , \tag{14}$$

where $\eta > 0$ is a learning rate, $\mathcal{T}_\tau \colon \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ is a projection-like operator, and $\mathrm{g}^f_{Z_t}(\theta_t)$ is a REINFORCE-like estimator (Williams, 1992) of $\nabla v^f_{\theta_t}(\rho)$ that uses a batch of B independent trajectories $Z_t \sim [\nu(\theta_t)]^{\otimes B}$. For a batch of trajectories $z = (s^b_{0:H-1}, a^b_{0:H-1})_{b=0}^{B-1}$, we define this estimator as

$$\mathrm{g}^f_z(\theta) = \frac{1}{B}\sum_{b=0}^{B-1}\sum_{h=0}^{H-1}\left\{\sum_{\ell=0}^{h}\frac{\partial \log \pi^f_\theta(a^b_\ell|s^b_\ell)}{\partial\theta}\gamma^h\left(\mathrm{r}(s^b_h,a^b_h) - \lambda\,\mathrm{D}^f(\pi^f_\theta(\cdot|s^b_h)\|\pi_{\mathrm{ref}}(\cdot|s^b_h))\right) - \lambda\gamma^h \mathrm{F}^f_\theta(s^b_h)\right\}, \tag{15}$$

where F^f_θ is a vector of size $|\mathcal{S}|\times|\mathcal{A}|$ defined by

$$[\mathrm{F}^f_\theta(s)]_{(s',b)} = \mathbf{1}_{s'}(s)\,\mathrm{W}^f_\theta(s)\,\mathrm{w}^f_\theta(b|s)\left[f'\left(\tfrac{\pi^f_\theta(b|s)}{\pi_{\mathrm{ref}}(b|s)}\right) - \sum_{a\in\mathcal{A}}\mathrm{w}^f_\theta(a|s) f'\left(\tfrac{\pi^f_\theta(a|s)}{\pi_{\mathrm{ref}}(a|s)}\right)\right] \quad . \tag{16}$$

Furthermore, for $z \in (\mathcal{S}\times\mathcal{A})^H$, we can derive the following upper bounds on the bias and the variance of the gradient estimator $\mathrm{g}^f_z(\cdot)$, that we prove in Appendix E.1.

Lemma 5.1. *Assume that, for some $\underline{\pi_{\text{ref}}} > 0$, f satisfy $\boldsymbol{A}_f(\underline{\pi_{\text{ref}}})$. For any parameter $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, we have*

$$\left\|\mathrm{g}^f(\theta) - \frac{\partial v_\theta^f(\rho)}{\partial \theta}\right\|_2 \leq \beta_{H,\lambda} \quad , \quad \mathbb{E}_{Z\sim[\nu(\theta)]^{\otimes B}}\left[\left\|\mathrm{g}^f(\theta) - \mathrm{g}_Z^f(\theta)\right\|_2^2\right] \leq \frac{\sigma_{\lambda,f}^2}{B} \quad .$$

where $\beta_{H,\lambda} \approx \frac{\gamma^H(H+1)}{(1-\gamma)^2}\omega_f(1+\lambda\max(\mathrm{d}_f,\mathrm{y}_f))$ and $\sigma_{\lambda,f}^2 \approx \frac{1}{(1-\gamma)^4}\omega_f^3\left[1+\lambda^2\max(\mathrm{d}_f^2,\mathrm{y}_f^2)\right]$.

Choice of the projection-like operator. The operator $\mathcal{T}_\tau$ used in the parameter updates (14) plays a similar role to the projection operator in projected gradient descent (Goldstein, 1964; Levitin & Polyak, 1966; Bertsekas, 2003; Shamir & Zhang, 2013). This operator allows for constraining the optimization process in a region of interest, discarding ill-behaved policies, where the divergence regularization is large. Given a policy π and $0 < \tau < 1/(2\underline{\pi_{\text{ref}}})$, we define the following operator $\mathcal{U}_\tau$, which for every $(s,a) \in \mathcal{S}\times\mathcal{A}$ gives

$$\mathcal{U}_\tau(\pi)(a|s) \;=\; \begin{cases} \pi_{\text{ref}}(a|s)\tau, & \text{if } \pi(a|s) \leq \pi_{\text{ref}}(a|s)\tau/2, \\ \pi(a|s) - \sum_{b\in\mathcal{A}_\tau^\pi(s)}\left(\pi_{\text{ref}}(b|s)\tau - \pi(b|s)\right), & \text{if } a = a_{\max}^\pi(s), \\ \pi(a|s), & \text{otherwise,} \end{cases}$$

where for any policy π and state $s \in \mathcal{S}$, we defined used the notation

$$\mathcal{A}_\tau^\pi(s) := \{a \in \mathcal{A}, \pi(a|s)/\pi_{\text{ref}}(a|s) \leq \tau/2\} \quad , \quad a_{\max}^\pi(s) = \arg\max_{a\in\mathcal{A}}\{\pi(a|s)/\pi_{\text{ref}}(a|s)\} \quad ,$$

choosing at random in the $\arg\max$ in the case of ties. This operator prevents policies from becoming "too deterministic": for any $s,a \in \mathcal{S}\times\mathcal{A}$, if $\pi(a|s)$ gets too close to zero, it is increased above a threshold that depends on the parameter τ and the reference policy π_{ref}. For a proper choice of τ, applying this operator on a policy returns a policy with a higher regularized value.

Lemma 5.2. *Assume that, for some $\underline{\pi_{\text{ref}}} > 0$, f and π_{ref} satisfy $\boldsymbol{A}_f(\underline{\pi_{\text{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\text{ref}}})$, and that the initial distribution ρ satisfies $\boldsymbol{A}_\rho$. Let $\tau_{\lambda,f} := \min([f']^{-1}(-\frac{16+8\gamma\lambda\mathrm{d}_f}{\lambda(1-\gamma)^2\rho_{\min}}), [f']^{-1}(-4|f'(\frac{1}{2})|), \frac{1}{2}\underline{\pi_{\text{ref}}})$. Then, for any policy π and for $\widetilde{\pi} = \mathcal{U}_{\tau_{\lambda,f}}(\pi)$, it holds that $v_{\widetilde{\pi}}^f(\rho) \geq v_\pi^f(\rho)$ and that $\widetilde{\pi}(a|s) \geq \underline{\pi_{\text{ref}}}\tau_{\lambda,f}$.*

Note that this operator $\mathcal{U}_\tau$ operates in the space of policies. We thus use it to define the operator $\mathcal{T}_\tau$ which updates θ such that $\pi_{\mathcal{T}_\tau\theta} = \mathcal{U}_\tau\pi_\theta$ (see Appendix D for an explicit construction of this operator).

Convergence analysis. We now present our main result, which gives a convergence rate for f-PG *with explicit constants* for our general class of parameterization. This result is based on the regularity properties of the regularized value function, which we developed in Section 4, in combination with an appropriate choice of policy improvement operator. First, we show that with the choice of threshold for $\mathcal{T}_\tau$, from Lemma 5.2, we can give a uniform lower bound on the constant $\mu_{\lambda,f}$ introduced in Theorem 4.3.

Lemma 5.3. *Assume that, for some $\underline{\pi_{\text{ref}}} > 0$, f and π_{ref} satisfy $\boldsymbol{A}_f(\underline{\pi_{\text{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\text{ref}}})$ and that ρ satisfies $\boldsymbol{A}_\rho$. If $\lambda = O(\frac{1}{(1-\gamma)^2\rho_{\min}}\min(\frac{4}{|f'(\iota_f)|}, \frac{1}{|f'(\frac{1}{2})|}, \frac{4}{|f'(\frac{1}{2}\underline{\pi_{\text{ref}}})|}))$, and $\tau_{\lambda,f}$ is set as in Lemma 5.2. then*

$$\inf_{\theta\in\mathbb{R}^d}\mu_{\lambda,f}(\mathcal{T}_{\tau_{\lambda,f}}\theta) \geq \underline{\mu}_{\lambda,f} := \frac{\lambda(1-\gamma)\rho_{\min}^2\zeta_f^2}{\omega_f^2}\underline{\pi_{\text{ref}}}^2(f^\star)''\left(-\frac{16+8\gamma\lambda\mathrm{d}_f}{\lambda(1-\gamma)^2\rho_{\min}}\right)^2 \quad .$$

This result is a consequence of Theorem 4.3, combined with the choice of policy improvement operator, which guarantees that the policies do not become too close to a deterministic policy. We give a proof of this Lemma in Appendix E.2. We get the following convergence rates for the regularized problem.

Theorem 5.4. *Assume that, for some $\underline{\pi_{\text{ref}}} > 0$, f and π_{ref} satisfy $\boldsymbol{A}_f(\underline{\pi_{\text{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\text{ref}}})$, and that ρ satisfies $\boldsymbol{A}_\rho$. Fix $\eta \leq 1/2L_{\lambda,f}$, and λ and $\tau_{\lambda,f}$ as in Lemma 5.3. Then, for any $t \geq 0$, the iterates of f-PG satisfy*

$$v_\star^f(\rho) - \mathbb{E}\left[v_{\theta_t}^f(\rho)\right] \leq (1-\underline{\mu}_{\lambda,f}\eta/4)^t(v_\star^f(\rho) - v_{\theta_0}^f(\rho)) + \frac{6\eta\sigma_{\lambda,f}^2}{B\underline{\mu}_{\lambda,f}} + \frac{6\beta_{H,\lambda}^2}{\underline{\mu}_{\lambda,f}} \quad ,$$

where the expressions of $\sigma_{\lambda,f}^2$ and $\beta_{H,\lambda}^2$ are given in Lemma 5.1.

We provide the proof of this result in Appendix E.2. A crucial feature of this theorem is that it is *fully explicit*, as all the terms that appear can be expressed using problem-dependent constants. This allows us to derive the two following sample complexity results for optimizing the regularized value function and the consequences on its non-regularized counterpart.

Corollary 5.5. *Let $\epsilon > 0$. With Theorem 5.4's assumptions, f-PG gives $v_\star^f(\rho) - \mathbb{E}[v_{\theta_T}^f(\rho)] \leq \epsilon$ in $T = O\big(\max(\frac{L_{\lambda,f}}{\underline{\mu}_{\lambda,f}}, \frac{\sigma_{\lambda,f}^2}{\epsilon B \underline{\mu}_{\lambda,f}^2}) \ln(\frac{v_\star^f(\rho) - v_{\theta_0}^f(\rho)}{\epsilon})\big)$ iterations, $H = O(\frac{\ln(1/\underline{\mu}_{\lambda,f})}{(1-\gamma)^2})$ with $\eta = O(\min(\frac{1}{L_{\lambda,f}}, \frac{\epsilon B \underline{\mu}_{\lambda,f}}{\sigma_{\lambda,f}^2}))$.*

Corollary 5.6. *Let $\epsilon > 0$ such that $\epsilon < \frac{16}{(1-\gamma)^3 \rho_{\min}} \min(\frac{4}{|f'(\iota_f)|}, \frac{1}{|f'(\frac{1}{2})|}, \frac{4}{|f'(\frac{1}{2}\pi_{\text{ref}})|})$, and set $\lambda = (1-\gamma)\epsilon / 4\mathrm{c}_f$. Assume $\zeta_f = O(1)$, $\kappa_f = O(1)$, $\zeta_f = O(1)$, and $\mathrm{c}_f \leq \min(1/\mathrm{d}_f, 1/\mathrm{y}_f, 1)$. Under Theorem 5.4's assumptions, f-PG gives $v_\star(\rho) - \mathbb{E}\left[v_{\theta_T}(\rho)\right] \leq \epsilon$ in*

$$T = O\Big(\frac{(f^\star)''(\frac{-1}{\epsilon \mathrm{c}_f (1-\gamma)^3 \rho_{\min}})^{-2}}{\epsilon (1-\gamma)^2 \rho_{\min}^2 \underline{\pi_{\text{ref}}}^2} \max\Big(\frac{1}{(1-\gamma)^3}, \frac{(f^\star)''(\frac{-1}{\epsilon \mathrm{c}_f (1-\gamma)^3 \rho_{\min}})^{-2}}{\epsilon^2 (1-\gamma)^6 B \rho_{\min}^2 \underline{\pi_{\text{ref}}}^2}\Big) \log\Big(\frac{6(v_\star^f(\rho) - v_{\theta_0}^f(\rho))}{\epsilon}\Big)\Big) \ ,$$

iterations, $H = O(\frac{\ln(1/\underline{\mu}_{\lambda,f})}{(1-\gamma)^2})$ with $\eta = O(\min((1-\gamma)^3, \epsilon^2 (f^\star)''(\frac{-1}{\epsilon \mathrm{c}_f (1-\gamma)^3 \rho_{\min}})^2 (1-\gamma)^6 B \rho_{\min}^2 \underline{\pi_{\text{ref}}}^2)$.

We give a more precise statement and a proof of these two corollaries in Corollary E.10. These results give an explicit sample complexity for the f-PG algorithm. Corollary 5.5 shows that f-PG enjoy convergence rates close to the ones of any gradient method: r $O(\kappa \log(1/\epsilon))$ for a low-variance regime, where $\kappa > 0$ is a condition number, and $O(1/\epsilon \log(1/\epsilon))$ in general. Selecting an appropriate λ and using the explicit expression for $\underline{\mu}_{\lambda,f}$ from Lemma 5.3 gives the results from the second corollary for unregularized problem.

Sample complexity for specific choices of divergence. We now provide a more complete interpretation of these results by stating sample complexity bounds for specific choices of divergences.

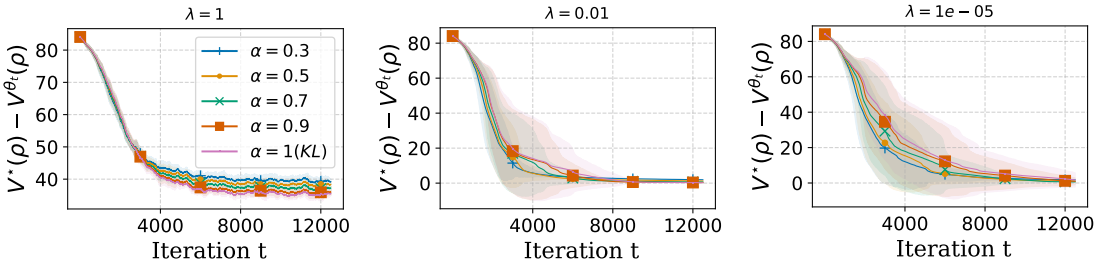
Corollary 5.7 (Complexity for Entropy regularization)**.** *Let f be the Kullback-Leibler divergence generator. Let $\epsilon > 0$. Under the assumptions of Theorem 5.4, f-PG achieves $v_\star(\rho) - \mathbb{E}[v_{\theta_T}(\rho)] \leq \epsilon$ in $THB = O(\frac{|\log(\pi_{\text{ref}})|^3}{\epsilon^4 (1-\gamma)^{12} \rho_{\min}^5 \underline{\pi_{\text{ref}}}^4} \exp(\frac{|\log(\pi_{\text{ref}})|}{\epsilon (1-\gamma)^3 \rho_{\min}}))$ samples.*

Corollary 5.8 (Complexity for Tsallis regularization)**.** *Let f be the α-Csiszár–Cressie–Read divergence generator for $0 < \alpha < 1$. Let $\epsilon > 0$. Under the assumptions of Theorem 5.4, f-PG achieves $v_\star(\rho) - \mathbb{E}[v_{\theta_T}(\rho)] \leq \epsilon$ in $TBH = O(\frac{\pi_{\text{ref}}^{7\alpha-7} |\log(\pi_{\text{ref}})|^3}{\epsilon^4 \alpha^6 (1-\gamma)^{12} \rho_{\min}^5 \underline{\pi_{\text{ref}}}^4} \exp_\alpha(-\frac{|\log(\pi_{\text{ref}})|}{\epsilon \alpha^2 (1-\gamma)^3 \rho_{\min}})^{4\alpha-8})$ samples, where we recall that $\exp_\alpha(x) = (1 + (\alpha-1)x)^{1/(\alpha-1)}$ for $x < 1/(1-\alpha)$.*

We give detailed versions and prove these corollaries in Appendix F, where we also give sample complexity results for the regularized problems. These corollaries show that Tsallis regularizers allow for faster learning, reducing the dependency on $(1-\gamma)$ from exponential in Corollary 5.7 to polynomial in Corollary 5.8, in essence overcoming an exponential lower bound of Li et al. (2023). Furthermore, we can show that the best choice of α to achieve fast convergence (according to our bounds) for Tsallis regularization is the following.

Corollary 5.9. *Assume the same condition of Corollary F.6. The optimal choice $\alpha^\star(\epsilon)$ to minimize the sample complexity from Corollary F.6 is $\alpha^\star(\epsilon) = 11/(2\log(1/\epsilon)) + o\left(1/\log(1/\epsilon)\right)$.*

We prove this corollary in Appendix F. This results shows that the best choice of α is not $\alpha = 0$ nor $\alpha = 1$, but depends on the desired precision level. This corroborates results from the bandit literature (Zimmert & Seldin, 2021), and gives strong evidence that Tsallis regularization with coupled policy parameterization has the potential to accelerate reinforcement learning algorithms.



(a) $\lambda = 1$ (b) $\lambda = 0.01$ (c) $\lambda = 10^{-5}$

Figure 1: Plot of the suboptimality gap of the unregularized objective, i.e. $v_\star(\rho) - v_{\theta_t}(\rho)$ during the learning process of f-PG for different choices for α-Tsallis divergence generators and different temperatures λ.

6 Experiments

We now evaluate the empirical performance of f-PG when the regularizer is chosen as the α-Tsallis divergence generator. Our primary goal is to identify which value of α achieves the best tradeoff between convergence speed and performance in the unregularised problem. We measure the number of iterations required for the algorithm to reach a good approximation of the optimal solution of the *unregularized objective*, with the step size fixed to $\eta = 0.01$.

Environment. We build upon the `GridWorld` environment (Domingues et al., 2021). The agent operates in a 5×5 grid starting from state $(0, 0)$. A small reward of $+0.1$ is obtained upon reaching and staying in state $(0, 1)$. Alternatively, by navigating a long path, the agent can collect a larger reward of $+1$ at state $(4, 4)$. At each step, the agent selects one of the four cardinal directions. The chosen action succeeds with probability 0.8 unless blocked by a wall; otherwise, the agent transitions uniformly at random to a neighboring state with probability 0.2. If the intended action leads into a wall, the agent remains in place.

Finding the best α. Figure 1 reports the performance of f-PG across different values of the temperature parameter λ and for various α-Tsallis regularizers. Recall from Corollary 5.6 that λ controls the bias–variance tradeoff: larger λ induces higher bias but faster convergence, whereas smaller λ reduces bias at the expense of slower learning. We observe precisely this tradeoff empirically. For large λ, all methods converge rapidly but towards biased solutions. In this regime, entropy-regularized PG converges slightly faster and with less bias than the other variants. However, in the high-accuracy regime (small λ), Tsallis regularization with $\alpha < 1$ consistently outperforms entropy regularization, confirming the theoretical prediction in Corollary 5.9 that moving beyond KL is beneficial.

7 Conclusion

We proposed a new class of policy parameterizations *coupled* with f-divergence regularizers, which naturally extend the classical softmax. This coupling bridges policy gradients with mirror descent, while remaining simple enough to apply in deep RL by replacing softmax with its f-generalization. Our analysis establishes the first convergence guarantees for f-regularized policy gradients under such parameterizations. Notably, our results show that entropy regularization is not always optimal, extending insights from bandits to a general reinforcement learning setting. We hope this perspective opens the door to designing algorithms where the choice of parameterization and regularization are treated jointly rather than in isolation. Extending our results to adversarial settings would further close the gap between bandits and reinforcement learning in applications of non-entropy regularizers.

REFERENCES

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=S1ANxQW0b.

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22 (98):1–76, 2021. URL http://jmlr.org/papers/v22/19-736.html.

Siddhant Agarwal, Ishan Durugkar, Peter Stone, and Amy Zhang. f-policy gradients: A general framework for goal-conditioned rl using f-divergences. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 12100–12123. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/27f4d95417bb722201597bf4d67cbacc-Paper-Conference.pdf.

Kavosh Asadi and Michael L. Littman. An alternative softmax operator for reinforcement learning. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 243–252. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/asadi17a.html.

Boris Belousov and Jan Peters. f-divergence constrained policy improvement. *arXiv preprint arXiv:1801.00056*, 2017.

Dimitri P Bertsekas. On the goldstein-levitin-polyak gradient projection method. *IEEE Transactions on automatic control*, 21(2):174–184, 2003.

Mathieu Blondel, André' F.T. Martins, and Vlad Niculae. Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020. URL http://jmlr.org/papers/v21/19-021.html.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578, 2022. doi: 10.1287/opre.2021.2151. URL https://doi.org/10.1287/opre.2021.2151.

Yinlam Chow, Ofir Nachum, and Mohammad Ghavamzadeh. Path consistency learning in Tsallis entropy regularized MDPs. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 979–988. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/chow18a.html.

Imre Csiszár. On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

John M. Danskin. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4): 641–664, 1966. doi: 10.1137/0114053. URL https://doi.org/10.1137/0114053.

Yuhao Ding, Junzi Zhang, Hyunin Lee, and Javad Lavaei. Beyond exact gradients: Convergence of stochastic soft-max policy gradient methods with entropy regularization. *IEEE Transactions on Automatic Control*, pp. 1–16, 2025.

Omar Darwiche Domingues, Yannis Flet-Berliac, Edouard Leurent, Pierre Ménard, Xuedong Shang, and Michal Valko. rlberry - A Reinforcement Learning Library for Research and Education, 10 2021. URL https://github.com/rlberry-py/rlberry.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2160–2169. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/geist19a.html.

Dongyoung Go, Tomasz Korbak, Germàn Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. Aligning language models with preferences through f-divergence minimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11546–11583. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/go23a.html.

Alan A. Goldstein. Convex programming in Hilbert space. *Bulletin of the American Mathematical Society*, 70(5):709 – 710, 1964.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1352–1361. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/haarnoja17a.html.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D. Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J Foster. Correcting the mythos of KL-regularization: Direct alignment without overoptimization via chi-squared preference optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=hXm0Wu2U9K.

Anatoli Juditsky, Joon Kwon, and Éric Moulines. Unifying mirror descent and dual averaging. *Mathematical Programming*, 199(1):793–830, May 2023. ISSN 1436-4646. doi: 10.1007/s10107-022-01850-3. URL https://doi.org/10.1007/s10107-022-01850-3.

Kyungjae Lee, Sungyub Kim, Sungbin Lim, Sungjoon Choi, and Songhwai Oh. Tsallis reinforcement learning: A unified framework for maximum entropy reinforcement learning. *arXiv:1902.00137*, 2019.

Evgeny S Levitin and Boris T Polyak. Constrained minimization methods. *USSR Computational mathematics and mathematical physics*, 6(5):1–50, 1966.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. *Mathematical Programming*, 201(1):707–802, Sep 2023. ISSN 1436-4646. doi: 10.1007/s10107-022-01920-6. URL https://doi.org/10.1007/s10107-022-01920-6.

Long Li, Jiaran Hao, Jason Klein Liu, Zhijian Zhou, Xiaoyu Tan, Wei Chu, Zhe Wang, Shirui Pan, Chao Qu, and Yuan Qi. The choice of divergence: A neglected key to mitigating diversity collapse in reinforcement learning with verifiable reward, 2025. URL https://arxiv.org/abs/2509.07430.

Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

Jiacai Liu, Jinchi Chen, and Ke Wei. On the linear convergence of policy gradient under hadamard parameterization. *Information and Inference: A Journal of the IMA*, 14(1):iaaf003, 2025.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1614–1623, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/martins16.html.

Jincheng Mei, Chenjun Xiao, Bo Dai, Lihong Li, Csaba Szepesvari, and Dale Schuurmans. Escaping the gravitational pull of softmax. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21130–21140. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f1cf2a082126bf02de0b307778ce73a7-Paper.pdf.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6820–6829. PMLR, 13–18 Jul 2020b. URL https://proceedings.mlr.press/v119/mei20b.html.

Arthur Mensch and Mathieu Blondel. Differentiable dynamic programming for structured prediction and attention. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3462–3471. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/mensch18a.html.

Johannes Müller and Semih Cayci. Essentially sharp estimates on the entropy regularization error in discounted markov decision processes. In *ICML 2024 Workshop: Foundations of Reinforcement Learning and Control–Connections and Perspectives*, 2024.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/facf9f743b083008a894eee7baa16469-Paper.pdf.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Springer Science & Business Media, 2004. doi: 10.1007/978-1-4419-8853-9.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120 (1):221–259, Aug 2009. ISSN 1436-4646. doi: 10.1007/s10107-007-0149-x. URL https://doi.org/10.1007/s10107-007-0149-x.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1504–1519, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1146. URL https://aclanthology.org/P19-1146/.

Martin L. Puterman. *Discounted Markov Decision Problems*, chapter 6, pp. 142–276. John Wiley and Sons, Ltd, 1994. ISBN 9780470316887. doi: https://doi.org/10.1002/9780470316887.ch6. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470316887.ch6.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pp. 547–562. University of California Press, 1961.

Vincent Roulet, Tianlin Liu, Nino Vieillard, Michael Eli Sander, and Mathieu Blondel. Loss functions and operators generated by f-divergences. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=V1YfPJDliw.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/schulman15.html.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 71–79, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/shamir13.html.

Harshit Sikchi, Qinqing Zheng, Amy Zhang, and Scott Niekum. Dual RL: Unification and new methods for reinforcement and imitation learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xt9Bu66rqv.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aBO5SvgSt1.

Nino Vieillard, Tadashi Kozuno, Bruno Scherrer, Olivier Pietquin, Remi Munos, and Matthieu Geist. Leverage the average: an analysis of KL regularization in reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12163–12174. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/8e2c381d4dd04f1c55093f22c59c3a08-Paper.pdf.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2cRzmWXK9N.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992. ISSN 1573-0565. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Lin Xiao. Dual averaging method for regularized stochastic learning and online optimization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper_files/paper/2009/file/7cce53cf90577442771720a370c3c723-Paper.pdf.

Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021. URL http://jmlr.org/papers/v22/19-753.html.

Appendix

Table of Contents

A Notations

The state–action sequence $(s_t, a_t)_{t\geq 0}$ defines a stochastic process on the canonical space $(\mathcal{S}\times\mathcal{A})^{\mathbb{N}}$. For any initial state $s_0 \in \mathcal{S}$, we denote by $\mathbb{P}^{\pi}_{s_0}$ the law of this process. That is, for any $n \in \mathbb{N}$ and any subset $B \subset (\mathcal{S}\times\mathcal{A})^n$,

$$\mathbb{P}^{\pi}_{s_0}(B) = \sum_{(a_0,\dots,a_{n-1})\in\mathcal{A}^n} \sum_{(s_1,\dots,s_{n-1})\in\mathcal{S}^{n-1}} \mathbb{1}_B\big((s_0,a_0),\dots,(s_{n-1},a_{n-1})\big) \prod_{i=0}^{n-1} \pi(a_i \mid s_i)\, \mathsf{P}(s_{i+1} \mid s_i, a_i),$$

with the convention s_0 is the given initial state. We denote by $\mathbb{E}^{\pi}_{s_0}$ the corresponding expectation operator. In particular, the state sequence $(s_t)_{t\geq 0}$ defines a Markov reward process (Section 2.1.6 in Puterman (1994)) with transition kernel

$$\mathsf{P}_\pi(s' \mid s) = \sum_{a\in\mathcal{A}} \mathsf{P}(s' \mid s, a)\, \pi(a \mid s) \ .$$

Symbols	Meaning	Definition
$\mathcal{S}$	State space	Section 3
$\mathcal{A}$	Action space	Section 3
γ	Discount factor	Section 3
P	Transition kernel	Section 3
r	Reward function	Section 3
π	Policy	Section 3
π_{ref}	Reference policy used in the regularization problem	Section 3
f	Divergence generator	Section 3
λ	Temperature of the regularization	Section 3
ρ	Initial state distribution	Section 3
κ_f	Upper bound on $\lvert f'''(x)/f''(x)^2\rvert$	$\mathbf{A}_f(\pi_{\mathrm{ref}})$
ω_f	Upper bound on $1/(xf''(x)^2)$	$\mathbf{A}_f(\pi_{\mathrm{ref}})$
d_f	Upper bound on a set of divergences	(12)
y_f	Upper bound on a quantity that depends on f'' and f'	(12)
ζ_f	Lower bound on a quantity that depends on f''	(13)
v_π	Value function of a policy π	(1)
v_π^f	Regularized Value function of a policy π	(3)
P_π	Transition kernel induced by policy π	Section 3
q_π^f	Regularized Q-function of a policy π	(49)
d_ρ^π	discounted state visitation of a policy π	(50)
θ	Parameter of the policy (element of $\mathbb{R}^{\lvert\mathcal{S}\rvert\times\lvert\mathcal{A}\rvert}$)	Section 4
π_θ^f	The soft-f-argmax policy associated with θ	(9)
w_θ^f	A matrix of size $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ such that for any $s\in\mathcal{S}$, $\mathrm{w}_\theta^f(\cdot\lvert s)\in\mathcal{P}(\mathcal{A})$	(11)
$f'_\theta(a\lvert s)$	shorthand notation for $f'(\pi_\theta^f(a\lvert s)/\pi_{\mathrm{ref}}(a\lvert s))$	(38)
$f''_\theta(a\lvert s)$	shorthand notation for $f''(\pi_\theta^f(a\lvert s)/\pi_{\mathrm{ref}}(a\lvert s))$	(39)
$f'''_\theta(a\lvert s)$	short hand notation for $f'''(\pi_\theta^f(a\lvert s)/\pi_{\mathrm{ref}}(a\lvert s))$	(40)
$\mathrm{W}_\theta^f(s)$	A function of $f''_\theta(\cdot\lvert s)$	(41)
$\mathrm{Y}_\theta^f(s)$	A function of $f'_\theta(\cdot\lvert s)$ and $f''_\theta(\cdot\lvert s)$	(41)
T	Number of iterations performed by f-PG	Algorithm 1
$\mathrm{g}_Z^f(\theta)$	Stochastic estimator of the gradient at θ	(15)
H	Truncation horizon in f-PG	Algorithm 1
$\mathrm{g}_Z^f(\theta)$	Stochastic estimator of the gradient at θ	(15)
$\beta_{H,\lambda}$	Bias of the stochastic estimator at θ	Lemma 5.1
$\sigma_{\lambda,f}$	Variance of the stochastic estimator at θ	Lemma 5.1
$L_{\lambda,f}$	Local smoothness of the objective at θ	Theorem 4.2
$\mathcal{P}(\mathcal{A})$	Set of probability measures over $\mathcal{A}$	Section 3
$\mathrm{D}^f(p\Vert q)$	f-divergence between two probability measures p and q	(2)

For $x\in\mathbb{R}^d$, we define the norms

$$\|x\|_\infty = \max_{i\in\{1,\dots,d\}} |x_i| \ , \quad \|x\|_1 = \sum_{i=1}^d |x_i| \ , \quad \|x\|_2 = \left(\sum_{i=1}^d |x_i|^2\right)^{1/2} .$$

For a $d \times d$ matrix M, we denote by $\|M\|_\infty$, and $\|M\|_2$ respectively the max *row* sum, and the spectral norm:

$$\|M\|_\infty = \sup_{x\neq 0}\{\|Mx\|_\infty / \|x\|_\infty\} = \sup_{i\in\{1,\dots,d\}} \sum_{j=1}^{d} |M_{i,j}| \ , \quad \|M\|_2 = \sup_{x\neq 0}\{\|Mx\|_2 / \|x\|_2\} \ . \tag{17}$$

Recall that, for any $x \in \mathbb{R}^d$, $\|Mx\|_\infty \leq \|M\|_\infty \|x\|_\infty$ and $\|Mx\|_2 \leq \|M\|_2 \|x\|_2$.

For notational convenience, we also view P as a $(|\mathcal{S}|\cdot|\mathcal{A}|) \times |\mathcal{S}|$ matrix with entries $\mathsf{P}_{(s,a),s'} = \mathsf{P}(s' \mid s, a)$. Similarly, v_π^f is a vector of size $|\mathcal{S}|$ and q_π^f a vector of size $|\mathcal{S}| \times |\mathcal{A}|$. Finally, we identify the parameter $\theta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ with its matrix representation $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, indexed by $(s, a) \in \mathcal{S} \times \mathcal{A}$. This slight abuse of notation allows us to conveniently switch between functional and matrix views.

B Smoothness of the objective

In this section, we establish the smoothness of the regularized value function $v_\theta^f := v_{\pi_\theta}^f(\rho)$ with respect to the parameter θ. As a first step, we show that the policy π_θ^f is smooth under suitable assumptions on the divergence generator f and compute its first and second derivatives. To do so, we start by studying the properties of the soft-f-argmax operator and then apply the obtained results to derive properties of the policy π_θ^f.

B.1 Properties of the soft-f-argmax

In this Section, we compute the derivative of $\nu_x^f(\cdot) := \mathrm{softargmax}^f(x, \nu_{\mathrm{ref}})$ and $\mathrm{softmax}^x(\cdot) := \mathrm{softmax}^f(x, \nu_{\mathrm{ref}})$, where

$$\mathrm{softmax}^f(x, \nu_{\mathrm{ref}}) := \max_{\nu\in\mathcal{P}(\mathcal{A})} \left\{ \langle \nu, x\rangle - \mathrm{D}^f(\nu\|\nu_{\mathrm{ref}}) \right\} \ . \tag{18}$$

The function $\mathrm{softmax}^x(\cdot)$ is the Fenchel–Legendre transform of $\mathrm{D}^f(\cdot\|\nu_{\mathrm{ref}})$, and the results in this section are therefore standard results from convex analysis, statements of which can be found in various forms in Hiriart-Urruty & Lemaréchal (2004); Mensch & Blondel (2018); Geist et al. (2019); Roulet et al. (2025).

In this section, we fix a reference probability distribution $\nu_{\mathrm{ref}} \in \mathcal{P}(\mathcal{A})$ such that for any $a \in \mathcal{A}$, we have $\nu_{\mathrm{ref}}(a) > \underline{\nu_{\mathrm{ref}}}$ for some $\min_{a\in\mathcal{A}} \nu_{\mathrm{ref}}(a) > \underline{\nu_{\mathrm{ref}}} > 0$. For a given $x \in \mathbb{R}^{|\mathcal{A}|}$, we define

$$\nu_x^f(\cdot) := \mathrm{softargmax}^f(x, \nu_{\mathrm{ref}}) = \operatorname*{arg\,max}_{\nu\in\mathcal{P}(\mathcal{A})} \left\{ \langle \nu, x\rangle - \mathrm{D}^f(\nu\|\nu_{\mathrm{ref}}) \right\} \ . \tag{19}$$

The following result is a simplified version of (Roulet et al., 2025, Proposition 1). For the sake of completeness, we provide a full proof.

Lemma B.1. *Assume that f is strictly convex on $[0, 1/\underline{\nu_{ref}}]$, differentiable on $(0, 1/\underline{\nu_{ref}})$, with $\lim_{x\downarrow 0^+} f'(u) = -\infty$, for some $\underline{\nu_{ref}} > 0$. Let ν_{ref} be a policy such that $\min_{a\in\mathcal{A}} \nu_{\mathrm{ref}}(a) > \underline{\nu_{ref}}$. For any $x \in \mathbb{R}^{|\mathcal{A}|}$ and $a \in \mathcal{A}$, we have $0 < \nu_x^f(a)$. Moreover, for all $x \in \mathbb{R}^{|\mathcal{A}|}$, there exists a unique $\mu_x \in I_{\underline{\nu_{ref}}}(x)$, where*

$$I_{\underline{\nu_{ref}}}(x) := (\max_{a\in\mathcal{A}} x(a) - f'(1/\underline{\nu_{ref}}), \max_{a\in\mathcal{A}} x(a) - f'(1)), \tag{20}$$

such that

$$\nu_x^f(a) = \nu_{\mathrm{ref}}(a)[f']^{-1}(x(a) - \mu_x) \, . \tag{21}$$

Moreover, $\mu_x \in \mathbb{R}$ is the unique root of the equation

$$F(x,\mu) := \sum_{a\in\mathcal{A}} \nu_{\mathrm{ref}}(a)[f']^{-1}(x(a)-\mu) - 1 = 0 \quad \text{for } \mu \in I_{\underline{\nu_{ref}}}(x). \tag{22}$$

Proof. Under the stated assumption, the map f' is strictly increasing on $(0, 1/\underline{\nu_{\mathrm{ref}}}]$, hence injective; therefore it is invertible *onto its image*. Moreover,

$$\mathrm{Dom}\left([f']^{-1}\right) = f'\left((0, 1/\underline{\nu_{\mathrm{ref}}}]\right) = \left(-\infty,\ f'(1/\underline{\nu_{\mathrm{ref}}})\right], \tag{23}$$

which is an interval and the function $[f']^{-1}$ is strictly increasing.

Fix $x \in \mathcal{A}$, and $(a, b) \in \mathcal{A} \times \mathcal{A}$. Recall from (19) the definition of the soft-f-argmax ,

$$\nu_x^f = \operatorname*{arg\,max}_{\nu\in\mathcal{P}(\mathcal{A})} \left\{ \langle \nu, x\rangle - \mathrm{D}^f(\nu\|\nu_{\mathrm{ref}}) \right\}$$

This is a strictly concave optimization problem over the probability simplex $\mathcal{P}(\mathcal{A})$ so it admits a unique maximizer. We now characterize the maximizer via the KKT conditions. Introduce multipliers $\mu \in \mathbb{R}$ for the equality constraint $\sum_{c\in\mathcal{A}} \nu(c) = 1$, and for every $c \in \mathcal{A}$, $\lambda(c) \in \mathbb{R}^+$ for the non-negativity constraints $\nu(c) \geq 0$. The Lagrangian reads

$$L(\nu, \mu, \{\lambda(c)\}_{c\in\mathcal{A}}) = \sum_{c\in\mathcal{A}} \nu(c)\, x(c) - \mathrm{D}^f(\nu\|\nu_{\mathrm{ref}}) + \mu\Big(1 - \sum_{c\in\mathcal{A}} \nu(c)\Big) - \sum_{c\in\mathcal{A}} \lambda(c)\, \nu(c) \ .$$

By the differentiability of f, differentiating the Lagrangian with respect to $\nu(a)$ gives

$$\frac{\partial L}{\partial \nu(a)} = x(a) - f'\left(\frac{\nu(a)}{\nu_{\mathrm{ref}}(a)}\right) - \mu - \lambda(a) \ . \tag{24}$$

At the optimum $(\nu_x^f, \mu_x, \{\lambda_x(c)\}_{c\in\mathcal{A}})$, the KKT conditions yield:

$$\nu_x^f(c)\, \lambda_x(c) = 0, \qquad \forall c \in \mathcal{A}, \tag{25}$$

$$x(c) - f'\left(\frac{\nu_x^f(c)}{\nu_{\mathrm{ref}}(c)}\right) - \mu_x - \lambda_x(c) = 0, \qquad \forall c \in \mathcal{A}. \tag{26}$$

Under the stated assumptions, $\lim_{x\to 0^+} f'(x) = -\infty$. Hence, if $\nu_x^f(c) = 0$ for some c, the stationarity condition (26) cannot hold with finite multipliers. Therefore $\nu_x^f(c) > 0$ for all $c \in \mathcal{A}$, which by (25) implies $\lambda_x(c) = 0$. Thus, for each $c \in \mathcal{A}$ the stationarity condition reduces to

$$x(c) - f'\left(\frac{\nu_x^f(c)}{\nu_{\mathrm{ref}}(c)}\right) = \mu_x \ . \tag{27}$$

Note also that (27) also implies that for all $a \in \mathcal{A}$, $x(a) - \mu_x \in \mathrm{Dom}(f')$, which implies, using (23) that $\max_{a\in\mathcal{A}} x(a) - f'(1/\underline{\nu_{\mathrm{ref}}}) \leq \mu_x$. Together with (27), this shows (21). Note that μ_x is a root of (22), since using (21),

$$\sum_{a\in\mathcal{A}} \nu_{\mathrm{ref}}(a)[f']^{-1}(x(a)-\mu_x) = \sum_{a\in\mathcal{A}} \nu_x^f(a) = 1.$$

Because $[f']^{-1}$ is strictly increasing on $(-\infty, f'(1/\underline{\nu_{\mathrm{ref}}})]$, for each $x \in \mathbb{R}^{|\mathcal{A}|}$, the function $\mu \mapsto F(x,\mu)$ (see (22)) is strictly decreasing on $(-\infty, f'(1/\underline{\nu_{\mathrm{ref}}})]$. Strict monotonicity gives the uniqueness of μ_x. Note finally that if $\mu > \max_{a\in\mathcal{A}} x(a) - f'(1)$, then $x(a) - \mu < f'(1)$, and since $[f']^{-1}$ is strictly increasing, $[f']^{-1}(x(a)-\mu) < 1$, showing that $F(x,\mu) < 0$, which concludes the proof. □

***Remark* B.2** *Let $f : (0,\tau) \to \mathbb{R}$ be a strictly convex and differentiable on $(0,\tau)$, $\mathrm{Dom}(f') = (0,\tau)$, and f' is strictly increasing and continuous on $(0,\tau)$. Let $\alpha := \lim_{x\to 0} f'(x) \in [-\infty, +\infty)$ and $\beta := \lim_{x\to\tau} f'(x) \in (-\infty, +\infty]$. Then $f'\big((0,\tau)\big) = (\alpha,\beta)$, i.e., $f' : (0,\tau) \to (\alpha,\beta)$ is a strictly increasing bijection (hence admits a continuous inverse $[f']^{-1} : (\alpha,\beta) \to (0,\tau)$). Define the convex conjugate of f,*

$$f^*(y) := \sup_{x\in(0,\tau)} \{xy - f(x)\} \ . \tag{28}$$

The two following properties hold,

(i) For every $y \in (\alpha,\beta)$, the supremum is attained at a unique point $x = (f')^{-1}(y)$.

(ii) $\mathrm{Dom}(f^*) \subseteq [\alpha,\beta]$ (with the convention that an infinite endpoint is excluded). Specifically, $\mathrm{Dom}(f^*) \cap (\alpha,\beta) = (\alpha,\beta)$, and $f^*(y) = +\infty$ for $y \notin [\alpha,\beta]$. At an endpoint $y = \alpha$ (resp. $y = \beta$), if finite, $f^*(y) = \lim_{x\downarrow 0}(yx - f(x))$ (resp. $\lim_{x\uparrow\tau}(yx - f(x))$), so $f^*(\alpha)$ (resp. $f^*(\beta)$) is finite iff the corresponding one–sided limit is finite.

On the open interval (α,β), the function f^* is differentiable and

$$(f^*)'(y) = [f']^{-1}(y), \qquad y \in (\alpha,\beta),$$

We retrieve the statement on (Roulet et al., 2025, Proposition 1) by replacing $[f']^{-1}$ by $(f^*)'$. In most examples, there is no need to resort to the convex conjugate to compute the inverse.

Lemma B.3. *Assume, in addition to Lemma B.1, that f is two-times continuously differentiable on $(0, \underline{\nu_{ref}})$. Then, the function $x \mapsto \mu_x$ is continuously differentiable on $\mathbb{R}^{|\mathcal{A}|}$, and for all $a \in \mathcal{A}$,*

$$\frac{\partial}{\partial x(a)}\mu_x = -\frac{\partial}{\partial x(a)}F(x,\mu_x) \Big/ \frac{\partial}{\partial \mu}F(x,\mu_x) \tag{29}$$

Proof. The function $(x,\mu) \mapsto F(x,\mu)$ is two-times continuously differentiable on the open set $U := \{(x,\mu) : x \in \mathbb{R}^{|\mathcal{A}|}, \mu \in I_{\underline{\nu_{\mathrm{ref}}}}(x)\} \subset \mathbb{R}^{|\mathcal{A}|} \times \mathbb{R}$. Let $x_0 \in \mathbb{R}^{|\mathcal{A}|}$. Since $[f']^{-1}$ is strictly increasing,

$$\frac{\partial}{\partial \mu}F(x,\mu_x) = \sum_{a\in\mathcal{A}} \nu_{\mathrm{ref}}(a) \frac{1}{f''([f']^{-1}(x_0(a) - \mu_{x_0}))} > 0.$$

Hence, we may apply the implicit function theorem, which shows that there exists an open neighborhood V_{x_0} and a unique function $x \mapsto \mu_x$ on V_{x_0}, such that for all $x \in V_{x_0}$, $F(x,\mu_x) = 0$ and (29) holds. □

We now introduce some compact notations that will be used throughout the sequel. For any $a \in \mathcal{A}$, define the first three derivatives of f evaluated at the probability ratio $\nu_x^f(a)/\nu_{\mathrm{ref}}(a)$:

$$f'_x(a) := f'\left(\frac{\nu_x^f(a)}{\nu_{\mathrm{ref}}(a)}\right) \ , \quad f''_x(a) := f''\left(\frac{\nu_x^f(a)}{\nu_{\mathrm{ref}}(a)}\right) \ , \quad f'''_x(a) := f'''\left(\frac{\nu_x^f(a)}{\nu_{\mathrm{ref}}(a)}\right) \ . \tag{30}$$

In addition, we introduce the quantities

$$\mathrm{W}_x^f := \sum_{a\in\mathcal{A}} \frac{\nu_{\mathrm{ref}}(a)}{f''_x(a)} \ . \tag{31}$$

Importantly, as f is strictly convex, its second derivative is strictly positive, making the preceding quantity well-defined. For any $a \in \mathcal{A}$, we also define the following normalized weights

$$\mathrm{w}_x^f(a) = \frac{1}{\mathrm{W}_x^f} \frac{\nu_{\mathrm{ref}}(a)}{f''_x(a)} \ . \tag{32}$$

We now specify the particular forms that these quantities take under three choices of the function f: the KL divergence ($f(u) = u\log u$), and the α-Tsallis entropies for $0 < \alpha < 1$.

KL case ($f(u) = u \log u$). Since $f'(u) = \log u + 1$, $f''(u) = 1/u$, and $[f']^{-1}(y) = \exp(y-1)$, the soft-f operators specialize as follows (with base measure $\nu_{\mathrm{ref}} \in \mathcal{P}(\mathcal{A})$):

$$\nu_x^{\mathrm{KL}}(a) = \nu_{\mathrm{ref}}(a)[f']^{-1}\big(x(a) - \mu_x\big) = \frac{\nu_{\mathrm{ref}}(a)\, \exp(x(a))}{\sum_{b\in\mathcal{A}} \nu_{\mathrm{ref}}(b)\, \exp(x(b))}\,, \qquad a \in \mathcal{A},$$

where the normalizer is

$$\mu_x \;=\; -1 + \log\Big(\sum_{b\in\mathcal{A}} \nu_{\mathrm{ref}}(b)\, \exp(x(b))\Big).$$

The associated softmax function is the log-partition:

$$\mathrm{softmax}^{\mathrm{KL}}(x, \nu_{\mathrm{ref}}) = \log\Big(\sum_{a\in\mathcal{A}} \nu_{\mathrm{ref}}(a)\, \exp(x(a))\Big).$$

For the curvature quantities in (31)–(32), since

$$f''\Big(\frac{\nu_x^{\mathrm{KL}}(a)}{\nu_{\mathrm{ref}}(a)}\Big) = \Big(\frac{\nu_x^{\mathrm{KL}}(a)}{\nu_{\mathrm{ref}}(a)}\Big)^{-1} = \frac{\nu_{\mathrm{ref}}(a)}{\nu_x^{\mathrm{KL}}(a)}\,,$$

we obtain

$$\mathrm{W}_x^{\mathrm{KL}} = \sum_{a\in\mathcal{A}} \frac{\nu_{\mathrm{ref}}(a)}{f_x''(a)} = \sum_{a\in\mathcal{A}} \nu_x^{\mathrm{KL}}(a) = 1$$

and the corresponding normalized weights are

$$\mathrm{w}_x^{\mathrm{KL}}(a) = \frac{1}{\mathrm{W}_x^{\mathrm{KL}}} \frac{\nu_{\mathrm{ref}}(a)}{f_x''(a)} = \nu_x^{\mathrm{KL}}(a)\,, \qquad a \in \mathcal{A}.$$

Tsallis-α case ($0 < \alpha < 1$). Let $f(u) = \dfrac{u^\alpha - \alpha u + \alpha - 1}{\alpha(\alpha-1)}$, so that

$$f'(u) = \frac{u^{\alpha-1} - 1}{\alpha - 1}, \qquad f''(u) = u^{\alpha-2}, \qquad [f']^{-1}(y) = \big[1 + (\alpha-1)y\big]^{\frac{1}{\alpha-1}}.$$

The soft-f operators specialize (with base measure $\nu_{\mathrm{ref}} \in \mathcal{P}(\mathcal{A})$) to

$$\nu_x^{\mathrm{TS}}(a) = \nu_{\mathrm{ref}}(a)[f']^{-1}(x(a) - \mu_x) = \nu_{\mathrm{ref}}(a)\, \big[1 + (\alpha-1)\big(x(a) - \mu_x\big)\big]^{1/(\alpha-1)}, \qquad a \in \mathcal{A},$$

where μ_x is the unique normalizer satisfying the constraint

$$\sum_{a\in\mathcal{A}} \nu_{\mathrm{ref}}(a)\, \big[1 + (\alpha-1)\big(x(a) - \mu_x\big)\big]^{\frac{1}{\alpha-1}} = 1,$$

with the domain condition $1 + (\alpha-1)\big(x(a) - \mu_x\big) > 0$ for all $a \in \mathcal{A}$.

The associated softmax is

$$\mathrm{softmax}^{\mathrm{TS}}(x, \nu_{\mathrm{ref}}) = \mu_x \;-\; \frac{1}{\alpha} \;+\; \frac{1}{\alpha} \sum_{a\in\mathcal{A}} \nu_{\mathrm{ref}}(a)\, \big[1 + (\alpha-1)\big(x(a) - \mu_x\big)\big]^{\frac{\alpha}{\alpha-1}}.$$

For the curvature quantities in (31)–(32), set

$$u_x(a) \;:=\; \frac{\nu_x^{\mathrm{TS}}(a)}{\nu_{\mathrm{ref}}(a)} = \big[1 + (\alpha-1)\big(x(a) - \mu_x\big)\big]^{\frac{1}{\alpha-1}}.$$

Since $f''\big(u_x(a)\big) = u_x(a)^{\alpha-2}$, we have

$$\mathrm{W}_x^{\mathrm{TS}} = \sum_{a\in\mathcal{A}} \frac{\nu_{\mathrm{ref}}(a)}{f_x''(a)} = \sum_{a\in\mathcal{A}} \nu_{\mathrm{ref}}(a)\, u_x(a)^{\,2-\alpha} = \sum_{a\in\mathcal{A}} \nu_{\mathrm{ref}}(a)\left[1+(\alpha-1)\big(x(a)-\mu_x\big)\right]^{(2-\alpha)/(\alpha-1)},$$

and the corresponding normalized weights are, for $a \in \mathcal{A}$,

$$\mathrm{w}_x^{\mathrm{TS}}(a) = \frac{1}{\mathrm{W}_x^f}\frac{\nu_{\mathrm{ref}}(a)}{f_x''(a)} = \frac{\nu_{\mathrm{ref}}(a)\, u_x(a)^{\,2-\alpha}}{\sum_{c\in\mathcal{A}} \nu_{\mathrm{ref}}(c)\, u_x(c)^{\,2-\alpha}} = \frac{\nu_{\mathrm{ref}}(a)\left[1+(\alpha-1)\big(x(a)-\mu_x\big)\right]^{(2-\alpha)/(\alpha-1)}}{\sum_{c\in\mathcal{A}} \nu_{\mathrm{ref}}(c)\left[1+(\alpha-1)\big(x(c)-\mu_x\big)\right]^{(2-\alpha)/(\alpha-1)}}\,.$$

Lemma B.4. *Assume $\boldsymbol{A}_f(\nu_{ref})$. Then the soft-$f$-argmax ν_x^f is twice continuously differentiable with respect to x. Moreover, for any $x \in \mathbb{R}^{|\mathcal{A}|}$, and $(a,b) \in \mathcal{A}^2$, we have*

$$\frac{1}{\mathrm{W}_x^f}\cdot\frac{\partial \nu_x^f(a)}{\partial x(b)} = 1_b(a)\,\mathrm{w}_x^f(a) - \mathrm{w}_x^f(a)\,\mathrm{w}_x^f(b)\ ,$$

In addition, for any $(a,b,c) \in \mathcal{A}^3$, the second derivative satisfies

$$\begin{aligned}\frac{1}{\mathrm{W}_x^f}\cdot\frac{\partial \nu_x^f(a)}{\partial x(b)\partial x(c)} &= -1_b(a)1_c(a)\frac{f_x'''(a)}{f_x''(a)^2}\cdot \mathrm{w}_x^f(a) + 1_c(b)\,\mathrm{w}_x^f(a)\,\mathrm{w}_x^f(b)\frac{f_x'''(b)}{f_x''(b)^2}\\ &\quad + 1_b(a)\,\mathrm{w}_x^f(a)\,\mathrm{w}_x^f(c)\frac{f_x'''(a)}{f_x''(a)^2} + 1_c(a)\,\mathrm{w}_x^f(a)\,\mathrm{w}_x^f(b)\frac{f_x'''(a)}{f_x''(a)^2}\\ &\quad - \mathrm{w}_x^f(a)\cdot\mathrm{w}_x^f(b)\cdot\mathrm{w}_x^f(c)\cdot\left[\frac{f_x'''(a)}{f_x''(a)^2}+\frac{f_x'''(b)}{f_x''(b)^2}+\frac{f_x'''(c)}{f_x''(c)^2}\right]\\ &\quad + \mathrm{w}_x^f(a)\cdot\mathrm{w}_x^f(b)\cdot\mathrm{w}_x^f(c)\cdot\sum_{d\in\mathcal{A}}\mathrm{w}_x^f(d)\cdot\frac{f_x'''(d)}{f_x''(d)^2}\ .\end{aligned}$$

Proof. Fix $x \in \mathbb{R}^{|\mathcal{A}|}$ and $(a,b,c) \in \mathcal{A}^3$. Define

$$F(x;\mu) = \sum_{a\in\mathcal{A}} \nu_{\mathrm{ref}}(a)[f']^{-1}(x(a)-\mu_x) - 1\ .$$

First derivative. Importantly, using Lemma B.1 we have that

$$F(x;\mu_x) = 0\ .$$

Differentiating the previous identity with respect to $x(b)$, yields

$$\frac{\partial F(x;\mu_x)}{\partial x(b)} = \sum_{a\in\mathcal{A}} \nu_{\mathrm{ref}}(a)\frac{1}{f''([f']^{-1}(x(a)-\mu_x))}\left(1_b(a) - \frac{\partial \mu_x}{\partial x(b)}\right)\ ,$$

where we used that the derivative of $[f']^{-1}$ is $1/f''([f']^{-1})$. Next, using from Lemma B.1 that $\nu_x^f(a) = \nu_{\mathrm{ref}}(a)[f']^{-1}(x(a)-\mu_x)$ yields

$$\frac{\partial F(x;\mu_x)}{\partial x(b)} = \sum_{a\in\mathcal{A}} \nu_{\mathrm{ref}}(a)\frac{1}{f_x''(a)}\left(1_b(a) - \frac{\partial \mu_x}{\partial x(b)}\right) = 0\ ,$$

where $f_x''(a)$ is defined in (30). This implies

$$\frac{\partial \mu_x}{\partial x(b)} = \mathrm{w}_x^f(b)\ , \tag{33}$$

where $\mathrm{w}_x^f(b)$ is defined in (32). Now that we have computed the derivative of the normalization factor μ_x, we can compute the derivative of the policy. Starting from Lemma B.1, we have that

$$\nu_x^f(a) = \nu_{\mathrm{ref}}(a)[f']^{-1}(x(a) - \mu_x) \ .$$

Differentiating the previous identity with respect to $x(b)$, yields

$$\frac{\partial \nu_x^f(a)}{\partial x(b)} = \frac{\nu_{\mathrm{ref}}(a)}{f_x''(a)}\left[1_b(a) - \frac{\partial \mu_x}{\partial x(b)}\right] = \frac{\nu_{\mathrm{ref}}(a)}{f_x''(a)}\left[1_b(a) - \mathrm{w}_x^f(b)\right] \ ,$$

where in the last identity, we used the expression of the derivative of μ_x given in (33). Finally, using the definition of W_x^f given in (31) gives

$$\boxed{\frac{\partial \nu_x^f(a)}{\partial x(b)} = 1_b(a)\frac{\nu_{\mathrm{ref}}(a)}{f_x''(a)} - \frac{\nu_{\mathrm{ref}}(a)}{f_x''(a)}\frac{\nu_{\mathrm{ref}}(b)}{f_x''(b)} \cdot \frac{1}{\mathrm{W}_x^f} \ ,} \tag{34}$$

Second derivative From (34), we aim to differentiate once more with respect to $x(c)$. First, note that it holds that

$$\frac{\partial f_x''(a)}{\partial x(c)} = \frac{f_x'''(a)}{\nu_{\mathrm{ref}}(a)} \cdot \frac{\partial \nu_x^f(a)}{\partial x(c)} = 1_c(a)\frac{f_x'''(a)}{f_x''(a)} - \frac{\nu_{\mathrm{ref}}(c) f_x'''(a)}{f_x''(a)\, f_x''(c)} \cdot \frac{1}{\mathrm{W}_x^f} \ , \tag{35}$$

$$\frac{\partial \mathrm{W}_x^f}{\partial x(c)} = -\sum_{d\in\mathcal{A}} \frac{\nu_{\mathrm{ref}}(d)}{f_x''(d)^2}\frac{\partial f_x''(d)}{\partial x(c)} = -\frac{\nu_{\mathrm{ref}}(c) f_x'''(c)}{f_x''(c)^3} + \frac{1}{\mathrm{W}_x^f}\sum_{d\in\mathcal{A}} \frac{\nu_{\mathrm{ref}}(c)\nu_{\mathrm{ref}}(d) f_x'''(d)}{f_x''(d)^3\, f_x''(c)} \ , \tag{36}$$

Now computing the second derivative of ν_x^f gives

$$\begin{aligned}\frac{\partial \nu_x^f(a)}{\partial x(b)\partial x(c)} &= -1_b(a)\frac{\nu_{\mathrm{ref}}(a)}{f_x''(a)^2}\frac{\partial f_x''(a)}{\partial x(c)} + \frac{\nu_{\mathrm{ref}}(a)\nu_{\mathrm{ref}}(b)}{f_x''(a)^2 f_x''(b)} \cdot \frac{1}{\mathrm{W}_x^f}\frac{\partial f_x''(a)}{\partial x(c)} \\ &\quad + \frac{\nu_{\mathrm{ref}}(a)\nu_{\mathrm{ref}}(b)}{f_x''(a) f_x''(b)^2} \cdot \frac{1}{\mathrm{W}_x^f}\frac{\partial f_x''(b)}{\partial x(c)} + \frac{\nu_{\mathrm{ref}}(a)\nu_{\mathrm{ref}}(b)}{f_x''(a) f_x''(b)} \cdot \frac{1}{(\mathrm{W}_x^f)^2}\frac{\partial \mathrm{W}_x^f}{\partial x(c)}\end{aligned}$$

Plugging in (35), and (36) in the preceding inequality yields

$$\begin{aligned}\frac{\partial \nu_{(}^{\theta} a)}{\partial x(b)\partial x(c)} &= -1_b(a)\frac{\nu_{\mathrm{ref}}(a)}{f_x(a)^2}\left[1_c(a)\frac{f_x'''(a)}{f_x''(a)} - \frac{\nu_{\mathrm{ref}}(c) f_x'''(a)}{f_x''(a)\, f_x''(c)} \cdot \frac{1}{\mathrm{W}_x^f}\right] \\ &\quad + \frac{\nu_{\mathrm{ref}}(a)\pi_{\mathrm{ref}}(b)}{f_x''(a)^2 f_x''(b)} \cdot \frac{1}{\mathrm{W}_x^f}\left[1_c(a)\frac{f_x'''(a)}{f_x''(a)} - \frac{\nu_{\mathrm{ref}}(c) f_x'''(a)}{f_x''(a)\, f_x''(c)} \cdot \frac{1}{\mathrm{W}_x^f}\right] \\ &\quad + \frac{\nu_{\mathrm{ref}}(a)\pi_{\mathrm{ref}}(b)}{f_x''(a) f_x''(b)^2} \cdot \frac{1}{\mathrm{W}_x^f}\left[1_c(b)\frac{f_x'''(b)}{f_x''(b)} - \frac{\nu_{\mathrm{ref}}(c) f_x'''(b)}{f_x''(b)\, f_x''(c)} \cdot \frac{1}{\mathrm{W}_x^f}\right] \\ &\quad + \frac{\nu_{\mathrm{ref}}(a)\nu_{\mathrm{ref}}(b)}{f_x''(a) f_x''(b)} \cdot \frac{1}{(\mathrm{W}_x^f)^2}\left[-\frac{\nu_{\mathrm{ref}}(c) f_x'''(c)}{f_x''(c)^3} + \frac{1}{\mathrm{W}_x^f}\sum_{d\in\mathcal{A}} \frac{\nu_{\mathrm{ref}}(c)\nu_{\mathrm{ref}}(d) f_x'''(d)}{f_x''(d)^3\, f_x''(c)}\right] \ ,\end{aligned}$$

which concludes the proof. □

The following lemma links the gradients $\mathrm{softmax}^f$ and $\mathrm{softargmax}^f$ operators.

Lemma B.5. *Assume $\boldsymbol{A}_f(\nu_{ref})$. For any $x \in \mathbb{R}^{|\mathcal{A}|}$, it holds that*

$$\frac{\partial \,\mathrm{softmax}^f(x, \nu_{\mathrm{ref}})}{\partial x} = \mathrm{softargmax}^f(x, \nu_{\mathrm{ref}}) \quad \textit{and} \quad \left\| \frac{\partial^2 \,\mathrm{softmax}^f(x, \nu_{\mathrm{ref}})]}{\partial x^2} \right\|_2 \leq 2\,\mathrm{W}_x^f \ .$$

Proof. For any $x \in \mathcal{A}$ and $\nu \in \mathcal{P}(\mathcal{A})$, define

$$h^f(x,\nu) = \langle \nu, x \rangle - \mathrm{D}^f(\nu \| \nu_{\mathrm{ref}}) \ . \tag{37}$$

Fix $b \in \mathcal{A}$ and note that

$$\frac{\partial h^f(x,\nu)}{\partial x(b)} = \nu(b) \ .$$

It holds that $\mathrm{softmax}^f(x) = \max_{\nu \in \mathcal{P}(\mathcal{A})} h^f(x,\nu)$. As h^f is continuous in its two variables, $\mathcal{P}(\mathcal{A})$ is a compact set, and for every $x \in \mathbb{R}^{|\mathcal{A}|}$, the function $h^f(x,\cdot)$ admits a unique optimizer in $\mathcal{P}(\mathcal{A})$, then by Danskin's theorem (Lemma G.5)

$$\frac{\partial \, \mathrm{softmax}^f(x,\nu_{\mathrm{ref}})}{\partial x(b)} = \frac{\partial h^f(x,\nu^\star(x))}{\partial x} = \nu^\star(x) \ , \text{ where } \nu^\star(x) = \operatorname*{arg\,max}_{\nu \in \mathcal{P}(\mathcal{A})} h^f(x,\nu) \ .$$

Finally, using that

$$\nu^\star(x) = \mathrm{softargmax}^f(x,\nu_{\mathrm{ref}}) \ ,$$

establishes the first identity of the lemma. Using the fact that for any matrix $A \in \mathbb{R}^{d \times d}$, we have $\|A\|_2 \leq \sum_{i=1}^d \sum_{j=1}^d |a_{i,j}|$, implies

$$\left\| \frac{\partial^2 \, \mathrm{softmax}^f(x,\nu_{\mathrm{ref}})]}{\partial x^2} \right\|_2 \leq \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{A}} \left| \frac{\partial \nu_x^f(a)}{\partial x(b)} \right| \leq \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{A}} \mathrm{W}_x^f \left| 1_b(a) \, \mathrm{w}_x^f(a) - \mathrm{w}_x^f(a) \, \mathrm{w}_x^f(b) \right|$$

where in the last equality, we used Lemma B.4. Finally, applying the triangle inequality establishes the second claim of the lemma. □

B.2 Derivatives of the policy

Next, we exploit the expression of the derivatives of the soft-f-argmax derived in Lemma B.4 to compute the first and second derivatives of the policy. We begin by extending the notations defined in (30), (31), and (32) to encompass a dependence on the state. For any pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, define the first three derivatives of f evaluated at the likelihood ratio $\pi_\theta^f(a|s)/\pi_{\mathrm{ref}}(a|s)$:

$$f'_\theta(a|s) \ := \ f'_{\theta(s,\cdot)}(a) = f'\left(\frac{\pi_\theta^f(a|s)}{\pi_{\mathrm{ref}}(a|s)} \right) \ , \tag{38}$$

$$f''_\theta(a|s) \ := \ f''_{\theta(s,\cdot)}(a) = f''\left(\frac{\pi_\theta^f(a|s)}{\pi_{\mathrm{ref}}(a|s)} \right) \ , \tag{39}$$

$$f'''_\theta(a|s) \ := \ f'''_{\theta(s,\cdot)}(a) = f'''\left(\frac{\pi_\theta^f(a|s)}{\pi_{\mathrm{ref}}(a|s)} \right) \ . \tag{40}$$

In addition, for every $s \in \mathcal{S}$ we introduce the quantities

$$\mathrm{W}_\theta^f(s) \ := \ \mathrm{W}_{\theta(s,\cdot)}^f = \sum_{a \in \mathcal{A}} \frac{\pi_{\mathrm{ref}}(a|s)}{f''_\theta(a|s)}, \quad \text{and} \quad \mathrm{Y}_\theta^f(s) \ := \ \sum_{a \in \mathcal{A}} \frac{\pi_{\mathrm{ref}}(a|s)}{f''_\theta(a|s)} \, |f'_\theta(a|s)| \ . \tag{41}$$

For every $(s,a) \in \mathcal{S} \times \mathcal{A}$, we define the normalized weights

$$\mathrm{w}_\theta^f(a|s) = \frac{1}{\mathrm{W}_\theta^f(s)} \frac{\pi_{\mathrm{ref}}(a|s)}{f''_\theta(a|s)} \tag{42}$$

The following lemma provides bounds on several key quantities that will appear in the appendix.

Lemma B.6. *Assume that, for some $\underline{\pi_{\text{ref}}} > 0$, f and π_{ref} satisfy $\boldsymbol{A}_f(\underline{\pi_{\text{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\text{ref}}})$, respectively. For any parameter $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, it holds that*

$$\left\|\mathrm{W}_\theta^f\right\|_\infty \le \omega_f \ , \quad \left\|\mathrm{Y}_\theta^f\right\|_\infty \le \mathrm{y}_f \ , \quad \max_{s\in\mathcal{S}} \mathrm{D}^f(\pi_\theta^f(\cdot|s)\|\pi_{\text{ref}}(\cdot|s)) \le \mathrm{d}_f \ , \quad \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\mathrm{w}_\theta^f(a|s)}{\pi_\theta^f(a|s)} \le \omega_f \ .$$

Proof. The proof immediately follows from the definition of the different quantities and $\mathbf{A}_f(\underline{\pi_{\text{ref}}})$. □

Using Lemma B.4, we get the following expression for the derivatives of the policy.

Corollary B.7. *Assume that, for some $\underline{\pi_{\text{ref}}} > 0$, f and π_{ref} satisfy $\boldsymbol{A}_f(\underline{\pi_{\text{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\text{ref}}})$, respectively. Then the policy π_θ^f is twice continuously differentiable with respect to θ. Additionally, for all $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and $s \in \mathcal{S}$, there exists a unique $\mu_\theta(s) \in \mathbb{R}$ such that for any $a \in \mathcal{A}$, we have*

$$\pi_\theta^f(a|s) = \pi_{\text{ref}}(a|s)[f']^{-1}\left(\theta(s,a) - \mu_\theta(s)\right). \tag{43}$$

For any $s \in \mathcal{S}$, the $\theta \mapsto \mu_\theta(s)$ is continuously differentiable. For any $s' \neq s$, $\partial/\partial\theta(s',\cdot)\mu_\theta(s) = 0$ and

$$\frac{\partial \mu_\theta(s)}{\partial\theta(s,\cdot)} = \mathrm{w}_\theta^f(\cdot|s).$$

Moreover, for any $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, $s \in \mathcal{S}$, and $(a,b) \in \mathcal{A}\times\mathcal{A}$, we have

$$\frac{1}{\mathrm{W}_\theta^f(s)} \cdot \frac{\partial \pi_\theta^f(a|s)}{\partial\theta(s,b)} = 1_b(a)\,\mathrm{w}_\theta^f(a|s) - \mathrm{w}_\theta^f(a|s)\,\mathrm{w}_\theta^f(b|s) \ ,$$

In addition, for any $(a,b,c) \in \mathcal{A}^3$, the second derivative satisfies

$$\begin{aligned}
&\frac{1}{\mathrm{W}_\theta^f(s)} \cdot \frac{\partial \pi_\theta^f(a|s)}{\partial\theta(s,b)\partial\theta(s,c)} = -1_b(a)1_c(a)\frac{f_\theta'''(a|s)}{f_\theta''(a|s)^2} \cdot \mathrm{w}_\theta^f(a|s) + 1_c(b)\,\mathrm{w}_\theta^f(a|s)\,\mathrm{w}_\theta^f(b|s)\frac{f_\theta'''(b|s)}{f_\theta''(b|s)^2} \\
&+ 1_b(a)\,\mathrm{w}_\theta^f(a|s)\,\mathrm{w}_\theta^f(c|s)\frac{f_\theta'''(a|s)}{f_\theta''(a|s)^2} + 1_c(a)\,\mathrm{w}_\theta^f(a|s)\,\mathrm{w}_\theta^f(b|s)\frac{f_\theta'''(a|s)}{f_\theta''(a|s)^2} \\
&- \mathrm{w}_\theta^f(a|s)\cdot\mathrm{w}_\theta^f(b|s)\cdot\mathrm{w}_\theta^f(c|s)\cdot\left[\frac{f_\theta'''(a|s)}{f_\theta''(a|s)^2} + \frac{f_\theta'''(b|s)}{f_\theta''(b|s)^2} + \frac{f_\theta'''(c|s)}{f_\theta''(c|s)^2}\right] \\
&+ \mathrm{w}_\theta^f(a|s)\cdot\mathrm{w}_\theta^f(b|s)\cdot\mathrm{w}_\theta^f(c|s)\cdot\sum_{d\in\mathcal{A}}\mathrm{w}_\theta^f(d|s)\cdot\frac{f_\theta'''(d|s)}{f_\theta''(d|s)^2} \ .
\end{aligned}$$

Lemma B.8. *Assume that, for some $\underline{\pi_{\text{ref}}} > 0$, f and π_{ref} satisfy $\boldsymbol{A}_f(\underline{\pi_{\text{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\text{ref}}})$, respectively. Then, it holds that*

$$\sum_{(a,b)\in\mathcal{A}^2}\left|\frac{\partial\pi_\theta^f(a|s)}{\partial\theta(s,b)}\right| \le 2\,\mathrm{W}_\theta^f(s) \ , \quad \sum_{(a,b,c)\in\mathcal{A}^3}\left|\frac{\partial^2\pi_\theta^f(a|s)}{\partial\theta(s,b)\partial\theta(s,c)}\right| \le 8\kappa_f\,\mathrm{W}_\theta^f(s) \ .$$

Proof. Using the expression of the derivative of the policy provided in Corollary B.7, we have by the triangle inequality

$$\sum_{b\in\mathcal{A}}\left|\frac{\partial\pi_\theta^f(a|s)}{\partial\theta(s,b)}\right| = \sum_{b\in\mathcal{A}}\mathrm{W}_\theta^f(s)\,\mathrm{w}_\theta^f(a|s)\left|1_b(a) - \mathrm{w}_\theta^f(b|s)\right| = 2\,\mathrm{W}_\theta^f(s)\,\mathrm{w}_\theta^f(a|s)(1-\mathrm{w}_\theta^f(a|s)) \ .$$

where we used that

$$\sum_{b\in\mathcal{A}}\left|1_b(a)-\mathrm{w}_\theta^f(b|s)\right| = 2(1-\mathrm{w}_\theta^f(a|s)) \ . \tag{44}$$

Hence, we have

$$\sum_{(a,b)\in\mathcal{A}^2}\left|\frac{\partial\pi_\theta^f(a|s)}{\partial\theta(s,b)}\right| \le 2\,\mathrm{W}_\theta^f(s) \ .$$

Fix $a \in \mathcal{A}$. By using the expression of the second derivative of the policy provided in Corollary B.7 combined with the triangle inequality and (44), we get

$$\sum_{(b,c)\in\mathcal{A}^2}\left|\frac{\partial^2\pi_\theta^f(a|s)}{\partial\theta(s,b)\partial\theta(s,c)}\right| \le 4\,\mathrm{W}_\theta^f(s)\frac{f_\theta'''(a|s)}{f_\theta''(a|s)^2}\,\mathrm{w}_\theta^f(a|s) + 4\,\mathrm{W}_\theta^f(s)\,\mathrm{w}_\theta^f(a|s)\sum_{b\in\mathcal{A}}\frac{f_\theta'''(b|s)}{f_\theta''(b|s)^2}\,\mathrm{w}_\theta^f(b|s) \ .$$

Next combining $\mathbf{A}_f(\underline{\pi_{\mathrm{ref}}})$and (44), we obtain

$$\sum_{(b,c)\in\mathcal{A}^2}\left|\frac{\partial^2\pi_\theta^f(a|s)}{\partial\theta(s,b)\partial\theta(s,c)}\right| \le 8\,\mathrm{W}_\theta^f(s)\,\mathrm{w}_\theta^f(a|s)\kappa_f \ .$$

Finally, summing over the actions concludes the proof. □

B.3 Smoothness of Objective

In order to prove the smoothness of the objective, we will prove that the all the second-order directional derivatives are bounded. Denote $\theta_\alpha = \theta + \alpha u$ where $\alpha \in \mathbb{R}$ and $u \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. By Equation (3), for any $s \in \mathcal{S}$ it holds that

$$v_{\theta_\alpha}^f(s) = \mathbf{e}_s^\top M(\alpha)\mathsf{r}_{\theta_\alpha}^f \ , \tag{45}$$

where $M(\alpha)$ is a matrix of $\mathbb{R}^{\mathcal{S}\times\mathcal{S}}$ that satisfies for any $(s,s') \in \mathcal{S}^2$

$$M(\alpha) = (\mathrm{Id} - \gamma\mathsf{P}_{\theta_\alpha})^{-1} \ ,$$

and where $\mathsf{r}_{\theta_\alpha}^f$, and $\mathsf{P}_{\theta_\alpha}$ are defined in Section 3. Taking the derivative of (45) with respect to α yields

$$\frac{\partial v_{\theta_\alpha}^f(s)}{\partial\alpha} = \gamma\mathbf{e}_s^\top M(\alpha)\frac{\partial\mathsf{P}_{\theta_\alpha}}{\partial\alpha}M(\alpha)\mathsf{r}_{\theta_\alpha}^f + \mathbf{e}_s^\top M(\alpha)\frac{\partial\mathsf{r}_{\theta_\alpha}^f}{\partial\alpha} \ .$$

Taking the derivative of the preceding equation with respect to α gives

$$\begin{aligned}\frac{\partial^2 v_{\theta_\alpha}^f(s)}{\partial\alpha^2} &= 2\gamma^2\mathbf{e}_s^\top M(\alpha)\frac{\partial\mathsf{P}_{\theta_\alpha}}{\partial\alpha}M(\alpha)\frac{\partial\mathsf{P}_{\theta_\alpha}}{\partial\alpha}M(\alpha)\mathsf{r}_{\theta_\alpha}^f + \gamma\mathbf{e}_s^\top M(\alpha)\frac{\partial^2\mathsf{P}_{\theta_\alpha}}{\partial^2\alpha}M(\alpha)\mathsf{r}_{\theta_\alpha}^f \\ &\quad + 2\gamma\mathbf{e}_s^\top M(\alpha)\frac{\partial\mathsf{P}_{\theta_\alpha}}{\partial\alpha}M(\alpha)\frac{\partial\mathsf{r}_{\theta_\alpha}^f}{\partial\alpha} + \mathbf{e}_s^\top M(\alpha)\frac{\partial^2\mathsf{r}_{\theta_\alpha}^f}{\partial^2\alpha} \ .\end{aligned} \tag{46}$$

In order to control the second-order directional derivative of the regularised value function, we establish first several properties of the quantities that appear in the preceding equality.

Lemma B.9. *Assume that, for some* $\underline{\pi_{\text{ref}}} > 0$, f *and* π_{ref} *satisfy* $\boldsymbol{A}_f(\underline{\pi_{\text{ref}}})$ *and* $\boldsymbol{P}(\underline{\pi_{\text{ref}}})$, *respectively. We have*

$$\left\| \frac{\partial \mathsf{P}_{\theta_\alpha}}{\partial \alpha}\bigg|_{\alpha=0} \right\|_\infty \leq 2 \max_s \{\mathrm{W}_\theta^f(s)\} \, \|u\|_2 \quad ,$$

Similarly, we have

$$\left\| \frac{\partial^2 \mathsf{P}_{\theta_\alpha}}{\partial^2 \alpha}\bigg|_{\alpha=0} \right\|_\infty \leq 8\kappa_f \max_s \{\mathrm{W}_\theta^f(s)\} \, \|u\|_2^2 \quad .$$

Proof. **Bounding the first-order directional derivative.** The derivative with respect to α is

$$\left[\frac{\partial \mathsf{P}_{\theta_\alpha}}{\partial \alpha}\bigg|_{\alpha=0} \right]_{s,s'} = \sum_{a \in \mathcal{A}} \left[\frac{\partial \pi^f_{\theta_\alpha}(a|s)}{\partial \alpha}\bigg|_{\alpha=0} \right] \mathsf{P}(s'|s,a) \quad .$$

Fix $s \in \mathcal{S}$. Because $\pi^f_{\theta_\alpha}(a|s)$ depends only on $\theta(s,\cdot)$, by the chain rule

$$\sum_{a \in \mathcal{A}} \left| \frac{\partial \pi^f_{\theta_\alpha}(a|s)}{\partial \alpha}\bigg|_{\alpha=0} \right| = \sum_{a \in \mathcal{A}} \left| \langle \frac{\partial \pi^f_\theta(a|s)}{\partial \theta(s,\cdot)}, u(s,\cdot) \rangle \right| \leq \sum_{a \in \mathcal{A}} \left\| \frac{\partial \pi^f_\theta(a|s)}{\partial \theta(s,\cdot)} \right\|_1 \|u(s,\cdot)\|_2 \quad ,$$

where in the last inequality we used Cauchy-Schwarz inequality and the fact that the L_1 norm dominates the L_2 norm. Now using Lemma B.8, we get that

$$\sum_{a \in \mathcal{A}} \left\| \frac{\partial \pi^f_\theta(a|s)}{\partial \theta(s,\cdot)} \right\|_1 \|u(s,\cdot)\|_2 \leq 2 \, \|u\|_2 \, \mathrm{W}_\theta^f(s) \quad .$$

Bounding the second-order directional derivative. Similarly, taking the second derivative with respect to α yields

$$\left[\frac{\partial^2 \mathsf{P}_{\theta_\alpha}}{\partial^2 \alpha}\bigg|_{\alpha=0} \right]_{s,s'} = \sum_{a \in \mathcal{A}} \left[\frac{\partial^2 \pi^f_{\theta_\alpha}(a|s)}{\partial^2 \alpha}\bigg|_{\alpha=0} \right] \mathsf{P}(s'|s,a) \quad .$$

Fix $s \in \mathcal{S}$. It holds that

$$\sum_{a \in \mathcal{A}} \left| \frac{\partial^2 \pi^f_{\theta_\alpha}(a|s)}{\partial^2 \alpha}\bigg|_{\alpha=0} \right| = \sum_{a \in \mathcal{A}} \left| \langle \frac{\partial^2 \pi^f_\theta(a|s)}{\partial^2 \theta(s,\cdot)} u(s,\cdot), u(s,\cdot) \rangle \right| \leq \sum_{a \in \mathcal{A}} \left\| \frac{\partial^2 \pi^f_\theta(a|s)}{\partial^2 \theta(s,\cdot)} \right\|_2 \|u(s,\cdot)\|_2^2 \quad ,$$

where in the last inequality, we used the Cauchy-Schwarz inequality and the definition of the matrix operator norm. Additionally, using that for any matrix $A \in \mathbb{R}^{d \times d}$, we have

$$\|A\|_2 \leq \sum_{i=1}^d \sum_{j=1}^d |a_{i,j}|$$

combined with Lemma B.8, we get that

$$\sum_{a \in \mathcal{A}} \left| \frac{\partial^2 \pi^f_{\theta_\alpha}(a|s)}{\partial^2 \alpha}\bigg|_{\alpha=0} \right| \leq 8 \, \|u(s,\cdot)\|_2^2 \, \kappa_f \max_{s \in \mathcal{S}} \mathrm{W}_\theta^f(s)$$

□

Lemma B.10. *Assume that, for some $\underline{\pi_{\mathrm{ref}}} > 0$, f and π_{ref} satisfy $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\mathrm{ref}}})$, respectively. Then, the regularized reward satisfies*

$$\|\mathsf{r}_\theta^f\|_\infty \leq 1 + \lambda \max_{s\in\mathcal{S}} \mathrm{D}^f(\pi_\theta^f(\cdot|s)\|\pi_{\mathrm{ref}}(\cdot|s)) \ .$$

Additionally, we have that

$$\left\| \frac{\partial \mathsf{r}^f_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} \right\|_\infty \leq 2 \max_{s\in\mathcal{S}} \left\{ \mathrm{W}_\theta^f(s) + \lambda\, \mathrm{Y}_\theta^f(s) \right\} \|u\|_2 \ ,$$

and that

$$\left\| \frac{\partial^2 \mathsf{r}^f_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0} \right\|_\infty \leq \max_{s\in\mathcal{S}} \left\{ 4(2\kappa_f + \lambda)\, \mathrm{W}_\theta^f(s) + 8\lambda\kappa_f\, \mathrm{Y}_\theta^f(s) \right\} \|u\|_2^2 \ .$$

Proof. The bound on $\|\mathsf{r}_\theta^f\|_\infty$ is immediate.

Bounding the first derivative. It holds that

$$\left\| \frac{\partial \mathsf{r}^f_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} \right\|_\infty = \max_{s\in\mathcal{S}} \left| \langle \frac{\partial \mathsf{r}_\theta^f(s)}{\partial \theta}, u \rangle \right| = \max_{s\in\mathcal{S}} \left| \langle \frac{\partial \mathsf{r}_\theta^f(s)}{\partial \theta(s,\cdot)}, u(s,\cdot) \rangle \right| \leq \max_{s\in\mathcal{S}} \left\| \frac{\partial \mathsf{r}_\theta^f(s)}{\partial \theta(s,\cdot)} \right\|_1 \|u\|_\infty \ .$$

Computing the derivative of $\mathsf{r}_\theta^f(s)$ with respect to $\theta(s,b)$ yields

$$\frac{\partial \mathsf{r}_\theta^f(s)}{\partial \theta(s,b)} = \sum_{a\in\mathcal{A}} \frac{\partial \pi_\theta^f(a|s)}{\partial \theta(s,b)} \mathsf{r}(s,a) - \lambda \frac{\partial \pi_\theta^f(a|s)}{\partial \theta(s,b)} f'\left(\frac{\pi_\theta^f(a|s)}{\pi_{\mathrm{ref}}(a|s)} \right) \ .$$

Plugging in the expression of the derivative of the policy of Corollary B.7 in the preceding identity yields

$$\begin{aligned} \frac{\partial \tilde{\mathsf{r}}_\theta(s)}{\partial \theta(s,b)} &= \frac{\pi_{\mathrm{ref}}(b|s)}{f_\theta''(b|s)} \mathsf{r}(s,b) - \lambda \frac{\pi_{\mathrm{ref}}(b|s)}{f_\theta''(b|s)} f_\theta'(b|s) - \sum_{a\in\mathcal{A}} \frac{\pi_{\mathrm{ref}}(a|s)\pi_{\mathrm{ref}}(b|s)}{f_\theta''(a|s) f_\theta''(b|s)} \cdot \frac{\mathsf{r}(s,a)}{\mathrm{W}_\theta^f(s)} \\ &\quad + \lambda \frac{1}{\mathrm{W}_\theta^f(s)} \sum_{a\in\mathcal{A}} \frac{\pi_{\mathrm{ref}}(a|s)\pi_{\mathrm{ref}}(b|s)}{f_\theta''(a|s) f_\theta''(b|s)} f_\theta'(a|s) \ . \end{aligned} \tag{47}$$

Taking the absolute value, applying the triangle inequality, and using that the rewards are bounded by 1 gives

$$\sum_{b\in\mathcal{A}} \left| \frac{\partial \mathsf{r}_\theta^f(s)}{\partial \theta(s,b)} \right| \leq 2\, \mathrm{W}_\theta^f(s) + 2\lambda\, \mathrm{Y}_\theta^f(s) \ .$$

Bounding the second derivative. It holds that

$$\left\| \frac{\partial^2 \mathsf{r}^f_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0} \right\|_\infty = \max_{s\in\mathcal{S}} \left| \frac{\partial^2 \mathsf{r}^f_{\theta_\alpha}(s)}{\partial \alpha^2}\Big|_{\alpha=0} \right| \leq \max_{s\in\mathcal{S}} \left\| \frac{\partial^2 \mathsf{r}_\theta^f(s)}{\partial \theta(s,\cdot)^2} \right\|_2 \|u\|_2^2 \ ,$$

where in the last inequality, we used the Cauchy-Schwarz inequality and the definition of the matrix operator norm.. We now compute the second derivative of $\mathsf{r}_\theta^f(s)$. Starting from (47), we get

$$\frac{\partial^2 \mathsf{r}_\theta^f(s)}{\partial \theta(s,b) \partial \theta(s,c)} = -\frac{\pi_{\mathrm{ref}}(b|s)}{f_\theta''(b|s)^2} \left[1_c(b) \frac{f_\theta'''(b|s)}{f_\theta''(b|s)} - \frac{\pi_{\mathrm{ref}}(c|s) f_\theta'''(b|s)}{f_\theta''(b|s)\, f_\theta''(c|s)} \cdot \frac{1}{\mathrm{W}_\theta^f(s)} \right] \mathsf{r}(s,b)$$

$$
\begin{aligned}
&+\lambda\frac{\pi_{\mathrm{ref}}(b|s)f'_\theta(b|s)}{f''_\theta(b|s)^2}\left[1_c(b)\frac{f'''_\theta(b|s)}{f''_\theta(b|s)}-\frac{\pi_{\mathrm{ref}}(c|s)f'''_\theta(b|s)}{f''_\theta(b|s)\,f''_\theta(c|s)}\cdot\frac{1}{\mathrm{W}^f_\theta(s)}\right]\\
&-\lambda\left[1_b(c)\frac{\pi_{\mathrm{ref}}(b|s)}{f''_\theta(b|s)}-\frac{\pi_{\mathrm{ref}}(b|s)\pi_{\mathrm{ref}}(c|s)}{f''_\theta(b|s)f''_\theta(c|s)}\cdot\frac{1}{\mathrm{W}^f_\theta(s)}\right]\\
&+\sum_{a\in\mathcal{A}}\frac{\pi_{\mathrm{ref}}(a|s)\pi_{\mathrm{ref}}(b|s)}{f''_\theta(a|s)^2f''_\theta(b|s)}\cdot\frac{\mathsf{r}(s,a)}{\mathrm{W}^f_\theta(s)}\left[1_c(a)\frac{f'''_\theta(a|s)}{f''_\theta(a|s)}-\frac{\pi_{\mathrm{ref}}(c|s)f'''_\theta(a|s)}{f''_\theta(a|s)\,f''_\theta(c|s)}\cdot\frac{1}{\mathrm{W}^f_\theta(s)}\right]\\
&+\sum_{a\in\mathcal{A}}\frac{\pi_{\mathrm{ref}}(a|s)\pi_{\mathrm{ref}}(b|s)}{f''_\theta(a|s)f''_\theta(b|s)^2}\cdot\frac{\mathsf{r}(s,a)}{\mathrm{W}^f_\theta(s)}\left[1_c(b)\frac{f'''_\theta(b|s)}{f''_\theta(b|s)}-\frac{\pi_{\mathrm{ref}}(c|s)f'''_\theta(b|s)}{f''_\theta(b|s)\,f''_\theta(c|s)}\cdot\frac{1}{\mathrm{W}^f_\theta(s)}\right]\\
&+\sum_{a\in\mathcal{A}}\frac{\pi_{\mathrm{ref}}(a|s)\pi_{\mathrm{ref}}(b|s)}{f''_\theta(a|s)f''_\theta(b|s)}\cdot\frac{\mathsf{r}(s,a)}{\mathrm{W}^f_\theta(s)^2}\left[-\frac{\pi_{\mathrm{ref}}(c|s)f'''_\theta(c|s)}{f''_\theta(c|s)^3}+\frac{1}{\mathrm{W}^f_\theta(s)}\sum_{d\in\mathcal{A}}\frac{\pi_{\mathrm{ref}}(c|s)\pi_{\mathrm{ref}}(d|s)f'''_\theta(d|s)}{f''_\theta(d|s)^3\,f''_\theta(c|s)}\right]\\
&-\lambda\sum_{a\in\mathcal{A}}\frac{\pi_{\mathrm{ref}}(a|s)\pi_{\mathrm{ref}}(b|s)}{f''_\theta(a|s)f''_\theta(b|s)}\frac{f'_\theta(b|s)}{\mathrm{W}^f_\theta(s)^2}\left[-\frac{\pi_{\mathrm{ref}}(c|s)f'''_\theta(c|s)}{f''_\theta(c|s)^3}+\frac{1}{\mathrm{W}^f_\theta(s)}\sum_{d\in\mathcal{A}}\frac{\pi_{\mathrm{ref}}(c|s)\pi_{\mathrm{ref}}(d|s)f'''_\theta(d|s)}{f''_\theta(d|s)^3\,f''_\theta(c|s)}\right]\\
&-\lambda\sum_{a\in\mathcal{A}}\frac{\pi_{\mathrm{ref}}(a|s)\pi_{\mathrm{ref}}(b|s)}{f''_\theta(a|s)^2f''_\theta(b|s)}\frac{f'_\theta(b|s)}{\mathrm{W}^f_\theta(s)}\left[1_c(a)\frac{f'''_\theta(a|s)}{f''_\theta(a|s)}-\frac{\pi_{\mathrm{ref}}(c|s)f'''_\theta(a|s)}{f''_\theta(a|s)\,f''_\theta(c|s)}\cdot\frac{1}{\mathrm{W}^f_\theta(s)}\right]\\
&-\lambda\sum_{a\in\mathcal{A}}\frac{\pi_{\mathrm{ref}}(a|s)\pi_{\mathrm{ref}}(b|s)}{f''_\theta(a|s)f''_\theta(b|s)^2}\frac{f'_\theta(b|s)}{\mathrm{W}^f_\theta(s)}\left[1_c(b)\frac{f'''_\theta(b|s)}{f''_\theta(b|s)}-\frac{\pi_{\mathrm{ref}}(c|s)f'''_\theta(b|s)}{f''_\theta(b|s)\,f''_\theta(c|s)}\cdot\frac{1}{\mathrm{W}^f_\theta(s)}\right]\\
&+\lambda\sum_{a\in\mathcal{A}}\frac{\pi_{\mathrm{ref}}(a|s)}{f''_\theta(a|s)}\frac{1}{\mathrm{W}^f_\theta(s)}\left[1_b(c)\frac{\pi_{\mathrm{ref}}(b|s)}{f''_\theta(b|s)}-\frac{\pi_{\mathrm{ref}}(b|s)\pi_{\mathrm{ref}}(c|s)}{f''_\theta(b|s)f''_\theta(c|s)}\cdot\frac{1}{\mathrm{W}^f_\theta(s)}\right]\,.
\end{aligned}
$$

Taking the absolute value, applying the triangle inequality, using that under $\mathbf{A}_f(\underline{\pi_{\mathrm{ref}}})$, for all $x\in\mathbb{R}+$, we have $|f'''(x)/f''(x)^2|\le\kappa_f$, and using that the rewards are bounded by 1 gives

$$
\sum_{b\in\mathcal{A}}\sum_{c\in\mathcal{A}}\left|\frac{\partial^2\mathsf{r}^f_{\theta_0}(s)}{\partial\theta(s,b)\partial\theta(s,c)}\right|\le 8\kappa_f\,\mathrm{W}^f_\theta(s)+8\lambda\kappa_f\,\mathrm{Y}^f_\theta(s)+4\lambda\,\mathrm{W}^f_\theta(s)\,,
$$

which concludes the proof. □

Lemma B.11. *Assume that, for some $\underline{\pi_{\mathrm{ref}}}>0$, f and π_{ref} satisfy $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\mathrm{ref}}})$ respectively. Then, for any $\theta\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and $u\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$,*

$$
\left|u^\top\frac{\partial^2 v^f_\theta(s)}{\partial\theta^2}u\right|\le\left(\sum_{i=1}^3\frac{L^{(i)}_{\lambda,f}(\theta)}{(1-\gamma)^i}\right)\|u\|_2^2\,,
$$

where for $i\in\{1,2,3\}$, $L^{(i)}_{\lambda,f}(\theta)$, are defined as

$$
\begin{aligned}
L^{(1)}_{\lambda,f}(\theta)&:=4(2\kappa_f+\lambda)\left\|\mathrm{W}^f_\theta\right\|_\infty+8\lambda\kappa_f\left\|\mathrm{Y}^f_\theta\right\|_\infty\,,\\
L^{(2)}_{\lambda,f}(\theta)&:=8\gamma\left\|\mathrm{W}^f_\theta\right\|_\infty\left(\kappa_f\{1+\lambda\max_{s\in\mathcal{S}}\mathrm{D}^f(\pi^f_\theta(\cdot|s)\|\pi_{\mathrm{ref}}(\cdot|s))\}+\left\|\mathrm{W}^f_\theta\right\|_\infty+\lambda\left\|\mathrm{Y}^f_\theta\right\|_\infty\right)\,,\\
L^{(3)}_{\lambda,f}(\theta)&:=8\gamma^2\left\|\mathrm{W}^f_\theta\right\|^2_\infty(1+\lambda\max_{s\in\mathcal{S}}\mathrm{D}^f(\pi^f_\theta(\cdot|s)\|\pi_{\mathrm{ref}}(\cdot|s))\,.
\end{aligned}
$$

Proof. By construction, we get

$$\left| u^\top \frac{\partial^2 v^f_\theta(s)}{\partial \theta^2} u \right| = \left| \frac{\partial^2 v^f_{\theta_\alpha}(s)}{\partial \alpha^2}\Big|_{\alpha=0} \right| \ .$$

Using (46), we get that

$$\left| \frac{\partial^2 v^f_{\theta_\alpha}(s)}{\partial \alpha^2}\Big|_{\alpha=0} \right| \le \underbrace{\left| 2\gamma^2 \mathbf{e}_s^\top M(0) \frac{\partial \mathsf{P}_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} M(0) \frac{\partial \mathsf{P}_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} M(0) \mathsf{r}^f_\theta \right|}_{(\mathbf{A})} + \underbrace{\left| \gamma \mathbf{e}_s^\top M(0) \frac{\partial^2 \mathsf{P}_{\theta_\alpha}}{\partial^2 \alpha}\Big|_{\alpha=0} M(0) \mathsf{r}^f_\theta \right|}_{(\mathbf{B})}$$
$$+ \underbrace{\left| 2\gamma \mathbf{e}_s^\top M(0) \frac{\partial \mathsf{P}_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} M(0) \frac{\partial \mathsf{r}^f_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} \right|}_{(\mathbf{C})} + \underbrace{\left| \mathbf{e}_s^\top M(0) \frac{\partial^2 \mathsf{r}^f_{\theta_\alpha}}{\partial^2 \alpha}\Big|_{\alpha=0} \right|}_{(\mathbf{D})} \ .$$

We now bound each of these terms separately

Bounding $(\mathbf{A})$. First note that, for any vector $x \in \mathbb{R}^{\mathcal{S}}$ and $\alpha \in \mathbb{R}$, we have

$$\|M(\alpha)x\|_\infty \le \frac{1}{1-\gamma}\|x\|_\infty \ , \tag{48}$$

This yields

$$(\mathbf{A}) \le 2\gamma^2 \left\| M(0) \frac{\partial \mathsf{P}_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} M(0) \frac{\partial \mathsf{P}_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} M(0) \mathsf{r}^f_\theta \right\|_\infty \le \frac{2\gamma^2}{(1-\gamma)^3} \left\| \frac{\partial \mathsf{P}_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} \right\|_\infty^2 \|\mathsf{r}^f_\theta\|_\infty$$

By using again (48), Lemma B.9, and Lemma B.10 we get

$$(\mathbf{A}) \le \frac{8\gamma^2 \max_s \{\mathrm{W}^f_\theta(s)\}^2 \|u\|_2^2}{(1-\gamma)^3} (1 + \lambda \max_{s\in\mathcal{S}} \mathrm{D}^f(\pi^f_\theta(\cdot|s) \| \pi_{\mathrm{ref}}(\cdot|s))) \ .$$

Bounding $(\mathbf{B})$. Using (48), Lemma B.9, and Lemma B.10 we get

$$\begin{aligned}(\mathbf{B}) &\le \frac{\gamma}{(1-\gamma)^2} \left\| \frac{\partial^2 \mathsf{P}_{\theta_\alpha}}{\partial^2 \alpha}\Big|_{\alpha=0} \right\|_\infty \|\mathsf{r}^f_\theta\|_\infty \\ &\le \frac{8\gamma}{(1-\gamma)^2} \kappa_f \max_s \{\mathrm{W}^f_\theta(s)\} \|u\|_2^2 \{1 + \lambda \max_{s\in\mathcal{S}} \mathrm{D}^f(\pi^f_\theta(\cdot|s) \| \pi_{\mathrm{ref}}(\cdot|s))\} \ .\end{aligned}$$

Bounding $(\mathbf{C})$. Similarly, using (48), Lemma B.9, and Lemma B.10 we get

$$(\mathbf{C}) \le \frac{8\gamma}{(1-\gamma)^2} \max_{s\in\mathcal{S}} \{\mathrm{W}^f_\theta(s)\} \max_{s\in\mathcal{S}} \left\{ \mathrm{W}^f_\theta(s) + \lambda \mathrm{Y}^f_\theta(s) \right\} \|u\|_2^2 \ .$$

Bounding $(\mathbf{D})$. Using (48), Lemma B.9, and Lemma B.10 we get

$$(\mathbf{D}) \le \frac{1}{1-\gamma} \left\| \frac{\partial^2 \mathsf{r}^f_{\theta_\alpha}}{\partial^2 \alpha}\Big|_{\alpha=0} \right\|_\infty \le \frac{1}{1-\gamma} \max_{s\in\mathcal{S}} \left\{ 4(2\kappa_f + \lambda) \mathrm{W}^f_\theta(s) + 8\lambda\kappa_f \mathrm{Y}^f_\theta(s) \right\} \|u\|_2^2 \ .$$

The proof is completed by collecting these upper bounds. □

Theorem B.12. *Assume that, for some $\underline{\pi_{\mathrm{ref}}} > 0$, f and π_{ref} satisfy $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\mathrm{ref}}})$ respectively. Then for any $\theta, \theta' \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, it holds that*

$$\left| v^f_{\theta'}(\rho) - v^f_\theta(\rho) - \langle \frac{\partial v^f_\theta(\rho)}{\partial \theta}, \theta' - \theta \rangle \right| \leq \frac{L_{\lambda,f}}{2} \left\| \theta' - \theta \right\|_2^2 \ .$$

where

$$L_{\lambda,f} := \frac{8\omega_f \left(\gamma\omega_f + (1-\gamma)\kappa_f\right)}{(1-\gamma)^3} + 4\lambda \frac{2\gamma^2\omega_f^2 \mathrm{d}_f + 2\gamma(1-\gamma)\omega_f \left[\kappa_f \mathrm{d}_f + \mathrm{y}_f\right] + (1-\gamma)^2 \left[\omega_f + 2\kappa_f \mathrm{y}_f\right]}{(1-\gamma)^3} \ .$$

Proof. Fix any vector $u \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$.Using Lemma B.11, it holds that

$$\left| u^\top \frac{\partial^2 v^f_\theta(s)}{\partial \theta^2} u \right| \leq \left(\sum_{i=1}^3 \frac{L^{(i)}_{\lambda,f}(\theta)}{(1-\gamma)^i} \right) \left\| u \right\|_2^2 \ ,$$

where for $i \in \{1,2,3\}$, $L^{(i)}_{\lambda,f}(\theta)$, are defined as

$$L^{(1)}_{\lambda,f}(\theta) := 4(2\kappa_f + \lambda) \left\| \mathrm{W}^f_\theta \right\|_\infty + 8\lambda\kappa_f \left\| \mathrm{Y}^f_\theta \right\|_\infty \ ,$$

$$L^{(2)}_{\lambda,f}(\theta) := 8\gamma \left\| \mathrm{W}^f_\theta \right\|_\infty \left(\kappa_f \{1 + \lambda \max_{s\in\mathcal{S}} \mathrm{D}^f(\pi^f_\theta(\cdot|s) \| \pi_{\mathrm{ref}}(\cdot|s))\} + \left\| \mathrm{W}^f_\theta \right\|_\infty + \lambda \left\| \mathrm{Y}^f_\theta \right\|_\infty \right) \ ,$$

$$L^{(3)}_{\lambda,f}(\theta) := 8\gamma^2 \left\| \mathrm{W}^f_\theta \right\|^2_\infty (1 + \lambda \max_{s\in\mathcal{S}} \mathrm{D}^f(\pi^f_\theta(\cdot|s) \| \pi_{\mathrm{ref}}(\cdot|s)) \ .$$

Using Lemma B.6 combined with Lemma G.1 concludes the proof. □

C Non-Uniform Łojasiewicz inequality

Firstly, define respectively q^f_θ and d^θ_ρ as the regularized Q-function and discounted state visitation associated with the policy π^f_θ, i.e.

$$q^f_\theta(s,a) = \mathsf{r}(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathsf{P}(s'|s,a) q^f_\theta(s') \ , \tag{49}$$

$$d^\theta_\rho(s) = (1-\gamma) \sum_{t=0}^\infty \gamma^t \rho \mathsf{P}^t_{\pi^f_\theta}(s) \ . \tag{50}$$

The goal of this section is to prove that the global objective satisfies a non-uniform Łojasiewicz inequality, i.e we aim to show the following theorem

Theorem C.1. *Assume that, for some $\underline{\pi_{\mathrm{ref}}} > 0$, f and π_{ref} satisfy $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\mathrm{ref}}})$ respectively. Assume in addition that the initial distribution ρ satisfy $\boldsymbol{A}_\rho$. Then, it holds that*

$$\left\| \frac{\partial v^f_\theta(\rho)}{\partial \theta} \right\|_2^2 \geq \mu_{\lambda,f}(\theta) \left(v^f_\star(\rho) - v^f_\theta(\rho) \right) \ ,$$

where

$$\mu_{\lambda,f}(\theta) := \frac{\lambda(1-\gamma)\rho^2_{\min}\zeta^2_f}{\omega^2_f} \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathrm{w}^f_\theta(a|s)^2 \ .$$

One of the main challenges in establishing such an inequality lies in connecting global information (the suboptimality gap) to local information (the gradient norm). Recall from Section 3 that if

$$\theta(s,a) = q_\star^f(s,a)/\lambda, \quad \forall\, a \in \mathcal{A},$$

then $\pi_\theta^f = \pi_\star^f$. This observation highlights that, under this parametrization and regularization, the key quantity is the closeness between θ and q_θ^f/λ. Formally, we will show that both the suboptimality gap and the gradient norm can be upper and lower bounded, respectively, by a quantity proportional to $\|\zeta_\theta(s)\|_2$, where we define

$$\zeta_\theta^f(s) := q_\theta^f(s,\cdot)/\lambda - \theta(s,\cdot) - K_\theta^f(s)\,1_{|\mathcal{A}|}, \tag{51}$$

$$K_\theta^f(s) := \frac{\langle q_\theta^f(s,\cdot)/\lambda - \theta(s,\cdot), 1_{|\mathcal{A}|}\rangle}{|\mathcal{A}|}\ . \tag{52}$$

Note that ζ_θ^f is the projection of $q_\theta^f(s,\cdot)/\lambda - \theta(s,\cdot)$ onto the subspace orthogonal to $1_{|\mathcal{A}|}$.

The proof proceeds in three steps:

1. Derive an explicit expression for the gradient of the objective and establish a lower bound in terms of $\|\zeta_\theta^f\|$.
2. Upper bound the suboptimality gap by a quantity directly related to $\|\zeta_\theta^f\|$.
3. Combine these two bounds to identify the corresponding non-uniform PL coefficient.

We now detail each step in turn.

C.1 Lower Bounding the Norm of the Gradient

Before deriving a lower bound on the norm of the gradient, we start by deriving an expression for the latter.

Lemma C.2. *Assume that, for some $\underline{\pi_{\mathrm{ref}}} > 0$, f and π_{ref} satisfy $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\mathrm{ref}}})$ respectively. For any $s \in \mathcal{S}$ and $b \in \mathcal{A}$, we have*

$$\frac{1}{\mathrm{W}_\theta^f(s)}\frac{\partial v_\theta^f(\rho)}{\partial\theta(s,b)} = \frac{d_\rho^\theta(s)}{1-\gamma}\,\mathrm{w}_\theta^f(b|s)\left[q_\theta^f(s,b) - \lambda\theta(s,b) - \sum_{a\in\mathcal{A}}\mathrm{w}_\theta^f(a|s)\left[q_\theta^f(s,a) - \lambda\theta(s,a)\right]\right]\ .$$

Proof. Fix $s \in \mathcal{S}$ and $b \in \mathcal{A}$. Additionally fix any $\tilde{s} \in \mathcal{S}$. Using (3), we have

$$v_\theta^f(\tilde{s}) = \sum_{a\in\mathcal{A}}\pi_\theta^f(a|\tilde{s})\mathsf{r}(\tilde{s},a) - \lambda\,\mathrm{D}^f(\pi_\theta^f(\cdot|\tilde{s})\|\pi_{\mathrm{ref}}(\cdot|\tilde{s})) + \gamma\sum_{a\in\mathcal{A}}\sum_{\tilde{s}'\in\mathcal{S}}\pi_\theta^f(a|\tilde{s})\mathsf{P}(\tilde{s}'|\tilde{s},a)v_\theta^f(\tilde{s}')$$

Deriving the preceding recursion with respect to $\theta(s,b)$ yields

$$\frac{\partial v_\theta^f(\tilde{s})}{\partial\theta(s,b)} = \underbrace{\sum_{a\in\mathcal{A}}\frac{\partial\pi_\theta^f(a|\tilde{s})}{\partial\theta(s,b)}\left[\mathsf{r}(\tilde{s},a) - \lambda f'\left(\frac{\pi_\theta^f(a|\tilde{s})}{\pi_{\mathrm{ref}}(a|\tilde{s})}\right) + \gamma\sum_{\tilde{s}'\in\mathcal{S}}\mathsf{P}(\tilde{s}'|\tilde{s},a)v_\theta^f(\tilde{s}')\right]}_{Z(\tilde{s})}$$
$$+\gamma\sum_{a\in\mathcal{A}}\sum_{\tilde{s}'\in\mathcal{S}}\pi_\theta^f(a|\tilde{s})\mathsf{P}(\tilde{s}'|\tilde{s},a)\frac{\partial v_\theta^f(\tilde{s}')}{\partial\theta(s,b)}\ .$$

Using the definition of the regularized Q-function and writing the preceding recursion in a vector form yields

$$\frac{\partial v_\theta^f(\cdot)}{\partial \theta(s,b)} = Z(\cdot) + \gamma \mathsf{P}_\theta \frac{\partial v_\theta^f(\cdot)}{\partial \theta(s,b)} \ .$$

which implies

$$\rho^\top \frac{\partial v_\theta^f(\cdot)}{\partial \theta(s,b)} = \frac{\partial v_\theta^f(\rho)}{\partial \theta(s,b)} = \rho^\top (\mathrm{Id} - \gamma \mathsf{P}_\theta)^{-1} Z(\cdot) \ .$$

Next, using the definition of the discounted state visitation (50) and the regularized Q-function (49) implies

$$\frac{\partial v_\theta^f(\rho)}{\partial \theta(s,b)} = \frac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} d_\rho^\theta(s') \sum_{a \in \mathcal{A}} \frac{\partial \pi_\theta^f(a|s')}{\partial \theta(s,b)} \left[q_\theta^f(s',a) - \lambda f' \left(\frac{\pi_\theta^f(a|s')}{\pi_{\mathrm{ref}}(a|s)} \right) \right] \ . \tag{53}$$

Using that $\sum_{a \in \mathcal{A}} \frac{\partial \pi_\theta^f(a|s)}{\partial \theta(s,b)} = 0$ and that for $s \neq s'$, we have $\frac{\partial \pi_\theta^f(a|s')}{\partial \theta(s,b)} = 0$ yields

$$\frac{\partial v_\theta^f(\rho)}{\partial \theta(s,b)} = \frac{1}{1-\gamma} d_\rho^\theta(s) \sum_{a \in \mathcal{A}} \frac{\partial \pi_\theta^f(a|s)}{\partial \theta(s,b)} \left[q_\theta^f(s,a) - \lambda f' \left(\frac{\pi_\theta^f(a|s)}{\pi_{\mathrm{ref}}(a|s)} \right) - \lambda \mu_\theta(s) \right] \ .$$

where $\mu_\theta(s)$ is defined in Corollary B.7 and satisfies for any $a \in \mathcal{A}$

$$\theta(s,a) - f' \left(\frac{\pi_\theta^f(a|s)}{\pi_{\mathrm{ref}}(a|s)} \right) = \mu_\theta(s) \ .$$

Thus, we obtain

$$\frac{\partial v_\theta^f(\rho)}{\partial \theta(s,b)} = \frac{1}{1-\gamma} d_\rho^\theta(s) \sum_{a \in \mathcal{A}} \frac{\partial \pi_\theta^f(a|s)}{\partial \theta(s,b)} \left[q_\theta^f(s,a) - \lambda \theta(s,a) \right] \ ,$$

Finally, plugging in the expression of the derivative of the policy derived in Corollary B.7 in the previous equality concludes the proof. □

Using the previous lemma, we prove the following lower bound for the norm of the gradient.

Lemma C.3. *Assume that, for some $\underline{\pi_{\mathrm{ref}}} > 0$, f and π_{ref} satisfy $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\mathrm{ref}}})$ respectively. Assume in addition that the initial distribution ρ satisfy $\boldsymbol{A}_\rho$. We have*

$$\left\| \frac{\partial v_\theta^f(\rho)}{\partial \theta} \right\|_2^2 \geq \lambda^2 \rho_{\min}^2 \min_{(s,a) \in \mathcal{S} \times \mathcal{A}} \{ \mathrm{w}_\theta^f(a|s)^2 \} \min_{s \in \mathcal{S}} \{ \mathrm{W}_\theta^f(s)^2 \} \sum_{s \in \mathcal{S}} \| \zeta_\theta(s) \|_2^2 \ .$$

Proof. It holds that

$$\left\| \frac{\partial v_\theta^f(\rho)}{\partial \theta} \right\|_2^2 = \sum_{s \in \mathcal{S}} \left\| \frac{\partial v_\theta^f(\rho)}{\partial \theta(s,\cdot)} \right\|_2^2 \ .$$

Fix $s \in \mathcal{S}$. Using Lemma C.2, we observe that

$$\frac{1}{\mathrm{W}_\theta^f(s)} \frac{\partial v_\theta^f(\rho)}{\partial \theta(s,\cdot)} = \frac{d_\rho^\theta(s)}{1-\gamma} H(\mathrm{w}_\theta^f(\cdot|s)) \left[q_\theta^f(s,\cdot) - \lambda \theta(s,\cdot) \right] \ .$$

where for any vector $u \in \mathbb{R}^{|\mathcal{A}|}$, we define $H(u) := \mathrm{diag}(u) - uu^\top$. Thus, we get that

$$\begin{aligned}
\frac{1-\gamma}{\mathrm{W}_\theta^f(s)} \left\| \frac{\partial v_\theta^f(\rho)}{\partial \theta(s,\cdot)} \right\|_2 &= d_\rho^\theta(s) \left\| H(\mathrm{w}_\theta^f(\cdot|s)) \left[q_\theta^f(s,\cdot) - \lambda\theta(s,\cdot) \right] \right\|_2 \\
&= \lambda d_\rho^\theta(s) \left\| H(\mathrm{w}_\theta^f(\cdot|s)) \left[q_\theta^f(s,\cdot)/\lambda - \theta(s,\cdot) - K_\theta^f(s) 1_{|\mathcal{A}|} \right] \right\|_2 \quad \text{(using that } H(u)1_{|\mathcal{A}|} = 0 \text{ and (52))} \\
&\geq \lambda d_\rho^\theta(s) \min_{a \in \mathcal{A}} \mathrm{w}_\theta^f(a|s) \left\| \zeta_\theta(s) \right\|_2 \quad \text{(where } \zeta_\theta(s) \text{ is defined in (51) and using Lemma G.4)} \ .
\end{aligned}$$

Finally, using $d_\rho^\theta(s) \geq (1-\gamma)\rho(s)$ and $\mathbf{A}_\rho$ concludes the proof. □

C.2 Bounding the Suboptimality Gap

The first step is to connect the suboptimality gap to information localized at θ. This is achieved via the performance difference lemma for the regularized value function yields (see Lemma G.3)

$$v_\star^f(\rho) - v_\theta^f(\rho) = \sum_{s \in \mathcal{S}} \frac{\lambda d_\rho^{\pi_\star^f}(s)}{1-\gamma} \underbrace{\left[\sum_{a \in \mathcal{A}} \pi_\star^f(a|s) q_\theta^f(s,a)/\lambda - \mathrm{D}^f(\pi_\star^f(\cdot|s) \| \pi_{\mathrm{ref}}(\cdot|s)) - v_\theta^f(s)/\lambda \right]}_{(\mathbf{A}(\mathbf{s}))} . \tag{54}$$

Fix $s \in \mathcal{S}$. Using the definition of the regularized value functions and Q-functions combined with Equation (3), we have

$$v_\theta^f(s) = \langle \pi_\theta^f(\cdot|s), q_\theta^f(s,\cdot) \rangle - \lambda \, \mathrm{D}^f(\pi_\theta^f(\cdot|s) \| \pi_{\mathrm{ref}}(\cdot|s)) \ .$$

This implies $\mathbf{A}(s) = \mathbf{A}_i(s) - \mathbf{A}_3(s) - \mathbf{A}_3(s)$ where

$$\begin{aligned}
\mathbf{A}_1(s) &= \langle \pi_\star^f(\cdot|s), q_\theta^f(s,\cdot)/\lambda \rangle - \mathrm{D}^f(\pi_\star^f(\cdot|s) \| \pi_{\mathrm{ref}}(\cdot|s)) \\
\mathbf{A}_2(s) &= \langle \pi_\theta^f(\cdot|s), \theta(s,\cdot) \rangle - \mathrm{D}^f(\pi_\theta^f(\cdot|s) \| \pi_{\mathrm{ref}}(\cdot|s)) \\
\mathbf{A}_3(s) &= \langle \pi_\theta^f(\cdot|s), q_\theta^f(s,\cdot)/\lambda - \theta(s,\cdot) \rangle \ .
\end{aligned} \tag{55}$$

Using (9) and (18), $\mathbf{A}_1(s) \leq \mathrm{softmax}^f(q_\theta^f(s,\cdot)/\lambda, \pi_{\mathrm{ref}}(\cdot|s))$ and $\mathbf{A}_2(s) = \mathrm{softmax}^f(\theta(s,\cdot), \pi_{\mathrm{ref}}(\cdot|s))$. Thus, we have

$$\begin{aligned}
\mathbf{A}(s) &\leq \mathrm{softmax}^f(q_\theta^f(s,\cdot)/\lambda, \pi_{\mathrm{ref}}(\cdot|s)) - \mathrm{softmax}^f(\theta(s,\cdot), \pi_{\mathrm{ref}}(\cdot|s)) - \mathbf{A}_3(s) \\
&= \mathrm{softmax}^f(q_\theta^f(s,\cdot)/\lambda, \pi_{\mathrm{ref}}(\cdot|s)) - \mathrm{softmax}^f(\theta(s,\cdot) + K_\theta^f(s) 1_{|\mathcal{A}|}) \\
&\quad - \langle \pi_\theta^f(\cdot|s), q_\theta^f(s,\cdot)/\lambda - \theta(s,\cdot) - K_\theta^f(s) 1_{|\mathcal{A}|} \rangle \ ,
\end{aligned} \tag{56}$$

where in the last equality we used that, for any $x \in \mathbb{R}^{|\mathcal{A}|}$ and $\alpha \in \mathbb{R}$,

$$\mathrm{softmax}^f(x + \alpha 1_{|\mathcal{A}|}, \pi_{\mathrm{ref}}(\cdot|s)) = \mathrm{softmax}^f(x, \pi_{\mathrm{ref}}(\cdot|s)) + \alpha \ .$$

The structure of $(\mathbf{A}(\mathbf{s}))$ closely resembles that of a first-order Taylor expansion as by Lemma B.5, we have that

$$\frac{\partial \, \mathrm{softmax}^f(\theta(s,\cdot), \pi_{\mathrm{ref}}(\cdot|s))}{\partial \theta(s,\cdot)} = \mathrm{softargmax}^f(\theta(s,\cdot), \pi_{\mathrm{ref}}(\cdot|s)) = \pi_\theta^f(\cdot|s) \ . \tag{57}$$

Lemma C.4. *Assume* $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$. *It holds that*

$$v_\star^f(\rho) - v_\theta^f(\rho) \leq \frac{\lambda}{1-\gamma} \sup_{\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}} \{\|W_\theta\|_\infty\} \sum_{s\in\mathcal{S}} \left\|\zeta_\theta^f(s)\right\|_2^2$$

Proof. In this proof, we denote by $g_s(x) = \mathrm{softmax}^f(x, \pi_{\mathrm{ref}}(\cdot|s))$. Combining (54) and (56) yields

$$v_\star^f(\rho) - v_\theta^f(\rho) \leq \frac{\lambda}{1-\gamma} \sum_{s\in\mathcal{S}} d_\rho^{\pi_\star^f}(s) B(s) \ , \tag{58}$$

where we have defined

$$B(s) = \left[g_s\left(q_\theta^f(s,\cdot)/\lambda\right) - g_s\left(\theta(s,\cdot) + K_\theta^f(s)1_{|\mathcal{A}|}\right) - \langle \pi_\theta^f(\cdot|s), q_\theta^f(s,\cdot)/\lambda - \theta(s,\cdot) - K_\theta^f(s)1_{|\mathcal{A}|}\rangle \right]$$

Next, by Lemma B.5, it holds that

$$\left.\frac{\partial g_s(x)}{\partial x}\right|_{x=\theta(s,\cdot)+K_\theta(s)1_{|\mathcal{A}|}} = \pi_\theta^f(\cdot|s) \ .$$

Standard one-dimensional Taylor theorem with Lagrange remainder shows that there exists $y \in \mathbb{R}^{|\mathcal{A}|}$ which belongs to the segment between $\theta(s,\cdot) + K_\theta(s)1_{\mathcal{A}}$ and $q_\theta^f(s,\cdot)/\lambda$ such that

$$B(s) = \frac{1}{2}\langle \left.\frac{\partial^2 g_s(x)}{\partial x^2}\right|_{x=y} \zeta_\theta^f(s), \zeta_\theta^f(s)\rangle$$

Using the bound on the spectral norm of the Hessian of g_s derived in Lemma B.5, we obtain

$$B(s) = \frac{1}{2}\langle \left.\frac{\partial^2 g_s(x)}{\partial x^2}\right|_{x=y} \zeta_\theta^f, \zeta_\theta^f\rangle \leq \frac{1}{2}\left\|\left.\frac{\partial^2 g_s(x)}{\partial x^2}\right|_{x=y}\right\|_2 \left\|\zeta_\theta^f(s)\right\|_2^2 \leq \mathrm{W}_y^f(s)\left\|\zeta_\theta^f(s)\right\|_2^2 \ .$$

Finally, bounding the discounted state visitation measure in (58) by 1 and plugging in the preceding bound on $B(s)$ concludes the proof. □

The proof of Theorem C.1 follows immediately from Lemma C.3, Lemma C.4, Lemma B.6, and (13).

D Monotone Improvement Operators

A key challenge in analyzing stochastic policy gradient methods is that the Łojasiewicz inequality depends on θ and degenerates whenever the probability of an action becomes small. The goal of this section is therefore to show the existence of an operator $\textsc{Imp}^f$ with two crucial properties: (i) for any policy, applying this operator produces a new policy with higher objective value, and (ii) every policy generated by this operator assigns at least a fixed minimum probability to every action. The main idea is to build the improvement operator such that it slightly augments the smallest probability weights, such that for any state action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$ the probability ratio $\pi(a|s)/\pi_{\mathrm{ref}}(a|s)$ stays above a certain threshold. We will show below that this procedure improves the global objective while keeping the probabilities uniformly bounded away from 0 when the threshold is properly chosen. Let $\underline{\pi_{\mathrm{ref}}} > 0$ be such that $\mathbf{A}_f(\underline{\pi_{\mathrm{ref}}})$ and $\mathbf{P}(\underline{\pi_{\mathrm{ref}}})$ hold. For any policy π, state $s \in \mathcal{S}$, $\tau < 1/(1\underline{\pi_{\mathrm{ref}}})$, we respectively define $\mathcal{A}_\tau^\pi(s)$, and $a_{\max}^\pi(s)$ as

$$\mathcal{A}_\tau^\pi(s) := \{a \in \mathcal{A}, \pi(a|s)/\pi_{\mathrm{ref}}(a|s) \leq \tau/2\} \quad , \quad a_{\max}^\pi(s) = \arg\max_{a\in\mathcal{A}} \{\pi(a|s)/\pi_{\mathrm{ref}}(a|s)\} \ ,$$

where the $\arg\max$ is chosen at random in the case of ties. Note that the definition of $\tau_{\lambda,f}$ ensures that $a^\pi_{\max}(s)$ does not belong to the set $\mathcal{A}^\pi_\tau(s)$ as

$$\max_{a\in\mathcal{A}} \frac{\pi(a|s)}{\pi_{\mathrm{ref}}(a|s)} \geq 1 \ .$$

Finally, we define the improvement operator as follows:

$$\begin{aligned} \mathcal{U}_\tau : \mathcal{P}(\mathcal{A})^{\mathcal{S}} &\longrightarrow \mathcal{P}(\mathcal{A})^{\mathcal{S}}, \\ \pi &\longmapsto \mathcal{U}_\tau(\pi), \end{aligned} \tag{59}$$

where for every $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\mathcal{U}_\tau(\pi)(a|s) \ = \ \begin{cases} \pi_{\mathrm{ref}}(a|s)\tau, & \text{if } \pi(a|s) \leq /\pi_{\mathrm{ref}}(a|s)\tau/2, \\ \pi(a|s) - \displaystyle\sum_{b\in\mathcal{A}^\pi_\tau(s)} \big(\pi_{\mathrm{ref}}(b|s)\tau - \pi(b|s)\big), & \text{if } a = a^\pi_{\max}(s), \\ \pi(a|s), & \text{otherwise.} \end{cases}$$

The operator $\mathcal{U}_\tau$ builds $\mathcal{U}_\tau(\pi)(a|s)$ by (statewise) raising each $a\in\mathcal{A}^\pi_\tau(s)$ to $\pi_{\mathrm{ref}}(a|s)\tau$, substracting the total added mass from the single action $a^\pi_{\max}(s)$, and leaving other actions unchanged. If $\mathcal{A}^\pi_\tau(s)=\emptyset$, for all $s\in\mathcal{S}$, then $\textsc{Imp}^f(\pi)=\pi$. Note that mass conservation is immediate from the definition and the fact that $\tau < 1/(2\underline{\pi_{\mathrm{ref}}})$. Non-negativity of $\mathcal{U}_\tau(\pi)(a^\pi_{\max}(s)|s)$ follows because the removed mass is

$$\sum_{a\in\mathcal{A}^\pi_\tau(s)} \{\pi_{\mathrm{ref}}(a|s)\tau - \pi(a|s)\} \leq \tau \sum_{a\in\mathcal{A}^\pi_\tau(s)} \pi_{\mathrm{ref}}(a|s) \leq \tau$$

Since $\pi(a^\pi_{\max}(s)|s) \geq \pi_{\mathrm{ref}}(a^\pi_{\max}(s)|s) \geq \underline{\pi_{\mathrm{ref}}}$, and $\tau \leq \underline{\pi_{\mathrm{ref}}}/2$, we get that $\mathcal{U}_\tau(\pi)(a^\pi_{\max}(s)|s) \geq \tau/2$. This in particular shows that $\mathcal{U}_\tau(\pi)$ is a policy. As by $\mathbf{A}_f(\underline{\pi_{\mathrm{ref}}})$ we have $\lim_{x\to 0^+} f'(x) = -\infty$, we consider $[f']^{-1} : (-\infty, f'(1/\underline{\pi_{\mathrm{ref}}})] \mapsto \mathbb{R}_+$ the inverse of f' and define

$$\tau_{\lambda,f} := \min\left([f']^{-1}\left(-\frac{16+8\gamma\lambda \mathrm{d}_f}{\lambda(1-\gamma)^2\rho_{\min}} \right), [f']^{-1}\left(-4\left| f'\left(\frac{1}{2}\right)\right| \right), \frac{1}{2}\underline{\pi_{\mathrm{ref}}} \right) \ . \tag{60}$$

The following lemma establishes the crucial improvement property when $\tau=\tau_{\lambda,f}$.

Lemma D.1. *Assume that, for some $\underline{\pi_{\mathrm{ref}}} > 0$, f and π_{ref} satisfy $\mathbf{A}_f(\underline{\pi_{\mathrm{ref}}})$ and $\mathbf{P}(\underline{\pi_{\mathrm{ref}}})$ respectively. Assume in addition that the initial distribution ρ satisfies $\mathbf{A}_\rho$. For any policy π, it holds that*

$$v^f_{\mathcal{U}_{\tau_{\lambda,f}}(\pi)}(\rho) \geq v^f_\pi(\rho) \ .$$

Additionally, for any policy π, we have that

$$\mathcal{U}_{\tau_{\lambda,f}}(\pi)(a|s) \geq \underline{\pi_{\mathrm{ref}}}\tau_{\lambda,f} \ .$$

Proof. Set an arbitrary policy π. For avoiding heavy notations, we will, through this proof, denote by $A^\pi_\tau = A^\pi_{\tau_{\lambda,f}}$. We consider the case where there is $s\in\mathcal{S}$ such that $\mathcal{A}^\pi_\tau(s)\neq\emptyset$ (alternatively $\mathcal{U}_{\tau_{\lambda,f}}(\pi)=\pi$, which makes the previous inequality immediately valid). Define $\tilde\pi = \mathcal{U}_{\tau_{\lambda,f}}(\pi)$. The following applies

$$\begin{aligned} v^f_{\tilde\pi}(\rho) - v^f_\pi(\rho) = &\sum_{s\in\mathcal{S}} d^{\tilde\pi}_\rho(s) \sum_{a\in\mathcal{A}} \left[\tilde\pi(a|s)\mathsf{r}(s,a) - \lambda\pi_{\mathrm{ref}}(a|s) f\left(\frac{\tilde\pi(a|s)}{\pi_{\mathrm{ref}}(a|s)}\right)\right] \\ &- \sum_{s\in\mathcal{S}} d^{\pi}_\rho(s) \sum_{a\in\mathcal{A}} \left[\pi(a|s)\mathsf{r}(s,a) - \lambda\pi_{\mathrm{ref}}(a|s) f\left(\frac{\pi(a|s)}{\pi_{\mathrm{ref}}(a|s)}\right)\right] \end{aligned}$$

$$= \underbrace{\sum_{s\in\mathcal{S}} \left(d_\rho^{\tilde{\pi}}(s) - d_\rho^{\pi}(s)\right) \sum_{a\in\mathcal{A}} \left[\tilde{\pi}(a|s)\mathsf{r}(s,a) - \lambda\pi_{\mathrm{ref}}(a|s) f\left(\frac{\tilde{\pi}(a|s)}{\pi_{\mathrm{ref}}(a|s)}\right)\right]}_{(\mathbf{I})}$$
$$+ \underbrace{\sum_{s\in\mathcal{S}} d_\rho^{\pi}(s) \sum_{a\in\mathcal{A}} (\tilde{\pi}(a|s) - \pi(a|s))\mathsf{r}(s,a)}_{(\mathbf{II})}$$
$$+ \underbrace{\lambda \sum_{s\in\mathcal{S}} d_\rho^{\pi}(s) \sum_{a\in\mathcal{A}} \pi_{\mathrm{ref}}(a|s) \left[f\left(\frac{\pi(a|s)}{\pi_{\mathrm{ref}}(a|s)}\right) - f\left(\frac{\tilde{\pi}(a|s)}{\pi_{\mathrm{ref}}(a|s)}\right)\right]}_{(\mathbf{III})} .$$

We now lower-bound each of the three terms separately.

Bounding (I). Using Lemma G.2, we have

$$\begin{aligned}
(\mathbf{I}) &\geq -\left\|d_\rho^{\tilde{\pi}} - d_\rho^{\pi}\right\|_1 \max_{s\in\mathcal{S}} \left|\sum_{a\in\mathcal{A}} \left[\tilde{\pi}(a|s)\mathsf{r}(s,a) - \lambda\pi_{\mathrm{ref}}(a|s) f\left(\frac{\tilde{\pi}(a|s)}{\pi_{\mathrm{ref}}(a|s)}\right)\right]\right| \\
&\geq -\frac{\gamma}{1-\gamma} \sup_{s\in\mathcal{S}} \|\tilde{\pi}(\cdot|s) - \pi(\cdot|s)\|_1 \sup_{s\in\mathcal{S}} \left[1 + \lambda\, \mathrm{D}^f(\tilde{\pi}(\cdot|s)\|\pi_{\mathrm{ref}}(\cdot|s))\right] \\
&\geq -\frac{2\gamma}{1-\gamma}\tau_{\lambda,f} \max_{s\in\mathcal{S}} \left\{\sum_{a\in A_\tau^\pi(s)} \pi_{\mathrm{ref}}(a|s)\right\} \sup_{\nu\in\mathcal{P}(\mathcal{A})} \sup_{s\in\mathcal{S}} \left[1 + \lambda\, \mathrm{D}^f(\nu\|\pi_{\mathrm{ref}}(\cdot|s))\right] ,
\end{aligned}$$

where in the last inequality we used that (because we increase the probability of the actions in $A_\tau^\pi(s)$ by $\tau_{\lambda,f}$ and remove the total added mass from the probability of $\pi(a_{\max}^\pi(s))$

$$\sup_{s\in\mathcal{S}} \|\tilde{\pi}(\cdot|s) - \pi(\cdot|s)\|_1 \leq 2 \max_{s\in\mathcal{S}} \left\{\sum_{a\in A_\tau^\pi(s)} \pi_{\mathrm{ref}}(a|s)\right\} \tau_{\lambda,f} .$$

Bounding (II). Using the triangle inequality yields

$$(\mathbf{II}) \geq -\sup_{s\in\mathcal{S}} \|\tilde{\pi}(\cdot|s) - \pi(\cdot|s)\|_1 \geq -2 \max_{s\in\mathcal{S}} \left\{\sum_{a\in A_\tau^\pi(s)} \pi_{\mathrm{ref}}(a|s)\right\} \tau_{\lambda,f} .$$

Bounding (III). All the state-action pairs on which the original π allocates the same probability then the policy $\tilde{\pi}$ are equal to 0 in (**III**) allowing us to simplify this term

$$\begin{aligned}
(\mathbf{III}) &= \lambda \sum_{s\in\mathcal{S}} d_\rho^{\pi}(s) \sum_{a\in\mathcal{A}} \pi_{\mathrm{ref}}(a|s) \left[f\left(\frac{\pi(a|s)}{\pi_{\mathrm{ref}}(a|s)}\right) - f\left(\frac{\tilde{\pi}(a|s)}{\pi_{\mathrm{ref}}(a|s)}\right)\right] \\
&= \lambda \sum_{s\in\mathcal{S}} d_\rho^{\pi}(s) \sum_{a\in\mathcal{A}_\tau^\pi(s)} \pi_{\mathrm{ref}}(a|s) \left[f\left(\frac{\pi(a|s)}{\pi_{\mathrm{ref}}(a|s)}\right) - f\left(\frac{\tilde{\pi}(a|s)}{\pi_{\mathrm{ref}}(a|s)}\right)\right] \\
&+ \lambda \sum_{s\in\mathcal{S}} \mathbb{1}(\mathcal{A}_\tau^\pi(s) \neq \emptyset) d_\rho^{\pi}(s) \pi_{\mathrm{ref}}(a_{\max}^\pi(s)|s) \left[f\left(\frac{\pi(a_{\max}^\pi(s)|s)}{\pi_{\mathrm{ref}}(a_{\max}^\pi(s)|s)}\right) - f\left(\frac{\tilde{\pi}(a_{\max}^\pi(s)|s)}{\pi_{\mathrm{ref}}(a_{\max}^\pi(s)|s)}\right)\right] .
\end{aligned}$$

Since f is convex, for all $u, v \in [0; 1/\underline{\pi_{\text{ref}}}]$, $f(u) - f(v) \geq f'(v)(u - v)$, we have

$$\begin{aligned}(\mathbf{III}) &\geq \lambda \sum_{s\in\mathcal{S}} d_\rho^\pi(s) \sum_{a\in\mathcal{A}_\tau^\pi(s)} (\pi(a|s) - \tilde{\pi}(a|s)) f'(\tau_{\lambda,f}) \qquad (\text{since } \tilde{\pi}(a|s)/\pi_{\text{ref}}(a|s) = \tau_{\lambda,f}) \\ &+ \lambda \sum_{s\in\mathcal{S}} \mathbb{1}(\mathcal{A}_\tau^\pi(s) \neq \emptyset) d_\rho^\pi(s) \left[\pi(a_{\max}^\pi(s)|s) - \tilde{\pi}(a_{\max}^\pi(s)|s)\right] f'\left(\frac{\tilde{\pi}(a_{\max}^\pi(s)|s)}{\pi_{\text{ref}}(a_{\max}^\pi(s)|s)}\right) \ ,\end{aligned}$$

Next, using that

$$\frac{\tilde{\pi}(a_{\max}^\pi(s)|s)}{\pi_{\text{ref}}(a_{\max}^\pi(s)|s)} \geq \frac{\pi(a_{\max}^\pi(s)|s) - \tau_{\lambda,f}}{\pi_{\text{ref}}(a_{\max}^\pi(s)|s)} \geq \frac{\pi(a_{\max}^\pi(s)|s) - \underline{\pi_{\text{ref}}}/2}{\pi_{\text{ref}}(a_{\max}^\pi(s)|s)} \geq 1 - \frac{1}{2} = \frac{1}{2} \ ,$$

combined with the monotonicity of f' and the fact that $\pi(a_{\max}^\pi(s)|s) - \tilde{\pi}(a_{\max}^\pi(s)|s) \geq 0$ yields

$$\begin{aligned}(\mathbf{III}) &\geq \lambda \sum_{s\in\mathcal{S}} d_\rho^\pi(s) \sum_{a\in\mathcal{A}_\tau^\pi(s)} (\pi(a|s) - \tilde{\pi}(a|s)) f'(\tau_{\lambda,f}) \qquad (\text{since } \tilde{\pi}(a|s)/\pi_{\text{ref}}(a|s) = \tau_{\lambda,f}) \\ &+ \lambda \sum_{s\in\mathcal{S}} \mathbb{1}(\mathcal{A}_\tau^\pi(s) \neq \emptyset) d_\rho^\pi(s) \left[\pi(a_{\max}^\pi(s)|s) - \tilde{\pi}(a_{\max}^\pi(s)|s)\right] f'\left(\frac{\tilde{\pi}(a_{\max}^\pi(s)|s)}{\pi_{\text{ref}}(a_{\max}^\pi(s)|s)}\right) \ ,\end{aligned}$$

Additionally, since

$$0 \leq \pi(a_{\max}^\pi(s)|s) - \tilde{\pi}(a_{\max}^\pi(s)|s) \leq \sum_{a\in\mathcal{A}_\tau^\pi(s)} (\pi(a|s) - \tilde{\pi}(a|s)) \leq \tau_{\lambda,f} \sum_{a\in\mathcal{A}_\tau^\pi(s)} \pi_{\text{ref}}(a|s) \ ,$$

implies

$$\begin{aligned}(\mathbf{III}) &\geq -\frac{\lambda}{2} \sum_{s\in\mathcal{S}} d_\rho^\pi(s) \mathbb{1}(\mathcal{A}_\tau^\pi(s) \neq \emptyset) \left(\sum_{a\in\mathcal{A}_\tau^\pi(s)} \pi_{\text{ref}}(a|s)\right) \tau_{\lambda,f} f'(\tau_{\lambda,f}) \\ &+ \lambda \sum_{s\in\mathcal{S}} d_\rho^\pi(s) \mathbb{1}(\mathcal{A}_\tau^\pi(s) \neq \emptyset) \left(\sum_{a\in\mathcal{A}_\tau^\pi(s)} \pi_{\text{ref}}(a|s)\right) \tau_{\lambda,f} f'(1/2) \ , \\ &\geq -\frac{\lambda}{4} \sum_{s\in\mathcal{S}} d_\rho^\pi(s) \mathbb{1}(\mathcal{A}_\tau^\pi(s) \neq \emptyset) \left(\sum_{a\in\mathcal{A}_\tau^\pi(s)} \pi_{\text{ref}}(a|s)\right) \tau_{\lambda,f} f'(\tau_{\lambda,f}) \ ,\end{aligned}$$

where in the last inequality, we used that $f'(\tau_{\lambda,f}) \leq -4|f'(1/2)|$. Hence, by using $\mathbf{A}_\rho$, we can lower bound this term as follows

$$(\mathbf{III}) \geq -\frac{\lambda}{4}(1-\gamma) \min_{s\in\mathcal{S}}\{\rho(s)\} \max_{s\in\mathcal{S}} \left\{\sum_{a\in\mathcal{A}_\tau^\pi(s)} \pi_{\text{ref}}(a|s)\right\} \tau_{\lambda,f} f'(\tau_{\lambda,f}) \ .$$

Collecting these lower bounds and using that

$$f'(\tau_{\lambda,f}) \leq -\frac{16 + 8\gamma\lambda \mathrm{d}_f}{\lambda(1-\gamma)^2 \rho_{\min}}$$

concludes the proof. □

Finally, we define the operator that maps each policy to one corresponding parameter

$$M^f : \Pi \to \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$$

by

$$M^f(\pi)(s,a) := f'\left(\frac{\pi(a \mid s)}{\pi_{\mathrm{ref}}(a \mid s)}\right) - f'\left(\frac{\pi(a_{|\mathcal{A}|}|s)}{\pi_{\mathrm{ref}}(a_{|\mathcal{A}|}|s)}\right), \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}. \tag{61}$$

Finally, we define the improvement operator on the logitspace as

$$\mathcal{T}_\tau := M^f \circ \mathcal{U}_\tau \ .$$

The following lemma shows that M^f successfully recovers a parameter that gives the policy and that $\mathcal{T}_\tau$ improves the value of the objective when $\lambda = \tau_{\lambda,f}$.

Lemma D.2. *Assume that, for some $\underline{\pi_{\mathrm{ref}}} > 0$, f and π_{ref} satisfy $\mathbf{A}_f(\underline{\pi_{\mathrm{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\mathrm{ref}}})$ respectively. Assume in addition that the initial distribution ρ satisfies $\mathbf{A}_\rho$. For any policy π, it holds that*

$$\pi^f_{M^f(\pi)} = \pi \ ,$$

Additionally, for any $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have that

$$v^f_{\mathcal{T}_{\tau_{\lambda,f}}(\theta)} \geq v^f_\theta \ , \quad \pi^f_{\mathcal{T}_{\tau_{\lambda,f}}(\theta)} \geq \underline{\pi_{\mathrm{ref}}}\tau_{\lambda,f} \ .$$

Proof. The proof follows immediately from a combination of equality (43) in Corollary B.7, (61), and Lemma D.1. □

E Convergence analysis of Stochastic Policy Gradient

In this section, we aim to derive under $\mathbf{A}_f(\underline{\pi_{\mathrm{ref}}})$ and $\mathbf{A}_\rho$ non-asymptotic convergence rates for f-PG. First, we establish a bound on the bias and variance of the REINFORCE estimator defined in (15).

E.1 Bounding the bias and variance of the stochastic estimator

First, recall the expression of the stochastic estimator of the gradient

$$\begin{aligned}\mathrm{g}^f_z(\theta) = &\frac{1}{B}\sum_{b=0}^{B-1}\sum_{h=0}^{H-1}\sum_{\ell=0}^{h}\frac{\partial \log \pi^f_\theta(a_\ell|s_\ell)}{\partial\theta}\gamma^h \mathsf{r}(s_h,a_h) \\ &- \lambda\frac{1}{B}\sum_{b=0}^{B-1}\sum_{h=0}^{H-1}\sum_{\ell=0}^{h-1}\frac{\partial \log \pi^f_\theta(a_\ell|s_\ell)}{\partial\theta}\gamma^h \mathrm{D}^f(\pi^f_\theta(\cdot|s_h)\|\pi_{\mathrm{ref}}(\cdot|s_h)) - \lambda\frac{1}{B}\sum_{b=0}^{B-1}\sum_{h=0}^{H-1}\gamma^h \mathrm{F}^f_\theta(s_h) \ ,\end{aligned} \tag{62}$$

where $z = (s^b_{0:H-1}, a^b_{0:H-1})_{b=0}^{B-1} \in (\mathcal{S}\cdot\mathcal{A})^{H\cdot B}$, and we recall that for any $s \in \mathcal{S}$, $\mathrm{F}^f_\theta(s)$ is a vector of size $|\mathcal{S}|\times|\mathcal{A}|$ defined in (16) as

$$[\mathrm{F}^f_\theta(s)]_{(s',b)} = 1_s(s')\,\mathrm{W}^f_\theta(s)\,\mathrm{w}^f_\theta(b|s)\left[f'(\frac{\pi^f_\theta(b|s)}{\pi_{\mathrm{ref}}(b|s)}) - \sum_{a\in\mathcal{A}}\mathrm{w}^f_\theta(a|s)f'(\frac{\pi^f_\theta(a|s)}{\pi_{\mathrm{ref}}(a|s)})\right] \ ,$$

and where W^f_θ, Y^f_θ, and w^f_θ are defined in (41) and (42). Finally, define the expected gradient estimator as

$$\mathrm{g}^f(\theta) := \mathbb{E}_{Z\sim[\nu(\theta)]^{\otimes B}}\left[\mathrm{g}^f_Z(\theta)\right] \ , \tag{63}$$

Before bounding the bias and the variance, we give an explicit expression of the derivative of the log probability that appears in the expression of our stochastic gradient estimator. We also provide a bound on the derivative of the log probabilities and on the matrix $\mathrm{F}^f_\theta(s)$ for any state $s \in \mathcal{S}$.

Lemma E.1. *Assume that, for some* $\underline{\pi_{\mathrm{ref}}} > 0$, f *satisfy* $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$. *For any* $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, $(s, s', a, b) \in \mathcal{S}^2 \times \mathcal{A}^2$, *we have*

$$\frac{\partial \log \pi_\theta^f(a|s)}{\partial \theta(s', b)} = 1_{s'}(s) \frac{\mathrm{W}_\theta^f(s)}{\pi_\theta^f(a|s)} \left[1_b(a)\, \mathrm{w}_\theta^f(a|s) - \mathrm{w}_\theta^f(a|s)\, \mathrm{w}_\theta^f(b|s)\right] \ .$$

Additionally, we have that

$$\left\| \frac{\partial \log \pi_\theta^f(a|s)}{\partial \theta} \right\|_2 \leq \frac{2\, \mathrm{W}_\theta^f(s)\, \mathrm{w}_\theta^f(a|s)}{\pi_\theta^f(a|s)} \ , \quad \left\| \mathrm{F}_\theta^f(s) \right\|_2 \leq 2\, \mathrm{W}_\theta^f(s)\, \mathrm{Y}_\theta^f(s) \ .$$

Proof. The proof follows from the log-derivative trick and the expression of the derivative of the policy provided in Corollary B.7. □

Next, we establish a REINFORCE-type formula for the gradient of the objective.

Lemma E.2. *Assume that, for some* $\underline{\pi_{\mathrm{ref}}} > 0$, f *satisfy* $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$. *It holds that*

$$\begin{aligned}
\frac{\partial v_\theta^f(\rho)}{\partial \theta(s,b)} &= \mathbb{E}\left[\sum_{t=0}^{\infty} \sum_{\ell=0}^{t} \frac{\partial \log \pi_\theta^f(A_\ell|S_\ell)}{\partial \theta(s,b)} \gamma^t \mathsf{r}(S_t, A_t)\right] \\
&\quad - \lambda \mathbb{E}\left[\sum_{t=0}^{\infty} \sum_{\ell=0}^{t-1} \frac{\partial \log \pi_\theta^f(A_\ell|S_\ell)}{\partial \theta(s,b)} \gamma^t\, \mathrm{D}^f(\pi_\theta^f(\cdot|S_t) \| \pi_{\mathrm{ref}}(\cdot|S_t))\right] \\
&\quad - \lambda\, \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t 1_s(S_t)\, \mathrm{W}_\theta^f(s)\, \mathrm{w}_\theta^f(b|s) \left[f'\Big(\frac{\pi_\theta^f(b|s)}{\pi_{\mathrm{ref}}(b|s)}\Big) - \sum_{a\in\mathcal{A}} \mathrm{w}_\theta^f(a|s) f'\Big(\frac{\pi_\theta^f(a|s)}{\pi_{\mathrm{ref}}(a|s)}\Big)\right]\right] \ .
\end{aligned}$$

Proof. Fix a parameter $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, a horizon T, and a divergence generator f. For any truncated trajectory $z = (s_t, a_t)_{t=0}^{T-1} \in (\mathcal{S}\times\mathcal{A})^T$, we define its probability as

$$\nu_T^f(\theta; z) \ = \ \rho(s_0)\pi_\theta^f(a_0|s_0) \prod_{t=1}^{T-1} \mathsf{P}(s_t|s_{t-1}, a_{t-1})\pi_\theta^f(a_t|s_t) \ ,$$

and the regularized return

$$R_{\theta,T}^f(z) \ = \ \sum_{t=0}^{T-1} \gamma^t \Big(\mathsf{r}(s_t, a_t) - \lambda \frac{\pi_{\mathrm{ref}}(a_t|s_t)}{\pi_\theta^f(a_t|s_t)} f\Big(\frac{\pi_\theta^f(a_t|s_t)}{\pi_{\mathrm{ref}}(a_t|s_t)}\Big)\Big). \tag{64}$$

The finite-horizon objective is

$$J_T^f(\theta) = \sum_{z\in(\mathcal{S}\times\mathcal{A})^T} \nu_T^f(\theta; z) R_{\theta,T}^f(z) \ .$$

Fix $(s, b) \in \mathcal{S}\times\mathcal{A}$. Differentiating this finite-horizon objective gives

$$\frac{\partial J_T^f(\theta)}{\partial \theta(s,b)} = \underbrace{\sum_{z\in(\mathcal{S}\times\mathcal{A})^T} \frac{\partial \nu_T^f(\theta; z)}{\partial \theta(s,b)} R_{\theta,T}^f(z)}_{(\mathbf{A})} + \underbrace{\sum_{z\in(\mathcal{S}\times\mathcal{A})^T} \nu_T^f(\theta; z) \frac{\partial R_{\theta,T}^f(z)}{\partial \theta(s,b)}}_{(\mathbf{B})} \ . \tag{65}$$

We now treat these two terms separately.

Term (A). Using the log-derivative trick and the fact that the only terms that depend on θ in $\nu_T^f(\theta; z)$ are the ones that depend on the policy itself, we obtain

$$\frac{\partial \nu_T^f(\theta; z)}{\partial \theta(s,b)} = \nu_T^f(\theta; z) \sum_{t=0}^{T-1} \frac{\partial \log \pi_\theta^f(a_t|s_t)}{\partial \theta(s,b)} \quad . \tag{66}$$

Plugging expressions (64) and (66) in (A) then gives

$$(\mathbf{A}) = \sum_{z \in (\mathcal{S}\times\mathcal{A})^T} \nu_T^f(\theta; z) \sum_{t=0}^{T-1} \sum_{\ell=0}^{T-1} \frac{\partial \log \pi_\theta^f(a_\ell|s_\ell)}{\partial \theta(s,b)} \gamma^t \Big(\mathsf{r}(s_t,a_t) - \lambda \frac{\pi_{\text{ref}}(a_t|s_t)}{\pi_\theta^f(a_t|s_t)} f\big(\frac{\pi_\theta^f(a_t|s_t)}{\pi_{\text{ref}}(a_t|s_t)}\big)\Big) \quad .$$

Now observe that for $\ell > t$, the sum of the log-gradient derivatives over $z \in (\mathcal{S}\times\mathcal{A})^T$ is 0. Therefore (A) reduces to

$$(\mathbf{A}) = \sum_{z \in (\mathcal{S}\times\mathcal{A})^T} \nu_T^f(\theta; z) \sum_{t=0}^{T-1} \sum_{\ell=0}^{t} \frac{\partial \log \pi_\theta^f(a_\ell|s_\ell)}{\partial \theta(s,b)} \gamma^t \Big(\mathsf{r}(s_t,a_t) - \lambda \frac{\pi_{\text{ref}}(a_t|s_t)}{\pi_\theta^f(a_t|s_t)} f\big(\frac{\pi_\theta^f(a_t|s_t)}{\pi_{\text{ref}}(a_t|s_t)}\big)\Big) \quad . \tag{67}$$

Term (B). Taking the derivative of (64), we have

$$\frac{\partial R_{\theta,T}^f(z)}{\partial \theta(s,b)} = \lambda \sum_{t=0}^{T-1} \gamma^t \Big(\frac{\partial \pi_\theta^f(a_t|s_t)}{\partial \theta(s,b)} \frac{\pi_{\text{ref}}(a_t|s_t)}{\pi_\theta^f(a_t|s_t)^2} f\big(\frac{\pi_\theta^f(a_t|s_t)}{\pi_{\text{ref}}(a_t|s_t)}\big) - \frac{\partial \pi_\theta^f(a_t|s_t)}{\partial \theta(s,b)} \frac{1}{\pi_\theta^f(a_t|s_t)} f'\big(\frac{\pi_\theta^f(a_t|s_t)}{\pi_{\text{ref}}(a_t|s_t)}\big)\Big) \quad .$$

Summing over $z \in (\mathcal{S}\times\mathcal{A})^T$ and using the log-derivative trick gives

$$(\mathbf{B}) = \lambda \sum_{z \in (\mathcal{S}\times\mathcal{A})^T} \nu_T^f(\theta; z) \sum_{t=0}^{T-1} \gamma^t \frac{\partial \log \pi_\theta^f(a_t|s_t)}{\partial \theta(s,b)} \left[\frac{\pi_{\text{ref}}(a_t|s_t)}{\pi_\theta^f(a_t|s_t)} f\big(\frac{\pi_\theta^f(a_t|s_t)}{\pi_{\text{ref}}(a_t|s_t)}\big) - f'\big(\frac{\pi_\theta^f(a_t|s_t)}{\pi_{\text{ref}}(a_t|s_t)}\big) \right] \quad . \tag{68}$$

Plugging the expressions (67) and (68) in (65) gives

$$\begin{aligned}
\frac{\partial J_T^f(\theta)}{\partial \theta(s,b)} &= \mathbb{E}_{Z\sim\nu_T^f(\theta)} \left[\sum_{t=0}^{T-1} \sum_{\ell=0}^{t} \frac{\partial \log \pi_\theta^f(A_\ell|S_\ell)}{\partial \theta(s,b)} \gamma^t \left(\mathsf{r}(S_t,A_t) - \lambda \frac{\pi_{\text{ref}}(A_t|S_t)}{\pi_\theta^f(A_t|S_t)} f\big(\frac{\pi_\theta^f(A_t|S_t)}{\pi_{\text{ref}}(A_t|S_t)}\big) \right) \right] \\
&\quad + \lambda\, \mathbb{E}_{Z\sim\nu_T^f(\theta)} \left[\sum_{t=0}^{T-1} \gamma^t \frac{\partial \log \pi_\theta^f(A_t|S_t)}{\partial \theta(s,b)} \left(\frac{\pi_{\text{ref}}(A_t|S_t)}{\pi_\theta^f(A_t|S_t)} f\big(\frac{\pi_\theta^f(A_t|S_t)}{\pi_{\text{ref}}(A_t|S_t)}\big) - f'\big(\frac{\pi_\theta^f(A_t|S_t)}{\pi_{\text{ref}}(A_t|S_t)}\big) \right) \right] \quad .
\end{aligned}$$

The previous term can be rewritten as

$$\begin{aligned}
\frac{\partial J_T^f(\theta)}{\partial \theta(s,b)} &= \mathbb{E}_{Z\sim\nu_T^f(\theta)} \left[\sum_{t=0}^{T-1} \sum_{\ell=0}^{t} \frac{\partial \log \pi_\theta^f(A_\ell|S_\ell)}{\partial \theta(s,b)} \gamma^t \mathsf{r}(S_t,A_t) \right] \\
&\quad - \lambda \mathbb{E}_{Z\sim\nu_T^f(\theta)} \left[\sum_{t=0}^{T-1} \sum_{\ell=0}^{t-1} \frac{\partial \log \pi_\theta^f(A_\ell|S_\ell)}{\partial \theta(s,b)} \gamma^t \frac{\pi_{\text{ref}}(A_t|S_t)}{\pi_\theta^f(A_t|S_t)} f\big(\frac{\pi_\theta^f(A_t|S_t)}{\pi_{\text{ref}}(A_t|S_t)}\big) \right] \\
&\quad - \lambda\, \mathbb{E}_{Z\sim\nu_T^f(\theta)} \left[\sum_{t=0}^{T-1} \gamma^t \frac{\partial \log \pi_\theta^f(A_t|S_t)}{\partial \theta(s,b)} f'\big(\frac{\pi_\theta^f(A_t|S_t)}{\pi_{\text{ref}}(A_t|S_t)}\big) \right] \quad .
\end{aligned}$$

Applying the tower property by taking the conditional expectation with respect to $\mathcal{G}_t := \sigma(S_0, A_0, \ldots, S_t)$ on the second expectation and using Lemma E.1 in the third expectation, gives

$$\frac{\partial J_T^f(\theta)}{\partial \theta(s,b)} = \mathbb{E}_{Z\sim\nu_T^f(\theta)} \left[\sum_{t=0}^{T-1} \sum_{\ell=0}^{t} \frac{\partial \log \pi_\theta^f(A_\ell|S_\ell)}{\partial \theta(s,b)} \gamma^t \mathsf{r}(S_t,A_t) \right]$$

$$- \lambda \mathbb{E}_{Z\sim\nu_T^f(\theta)}\left[\sum_{t=0}^{T-1}\sum_{\ell=0}^{t-1}\frac{\partial \log \pi_\theta^f(A_\ell|S_\ell)}{\partial\theta(s,b)}\gamma^t \,\mathrm{D}^f(\pi_\theta^f(\cdot|S_t)\|\pi_{\mathrm{ref}}(\cdot|S_t))\right]$$
$$- \lambda\, \mathbb{E}_{Z\sim\nu_T^f(\theta)}\left[\sum_{t=0}^{T-1}\gamma^t \mathrm{1}_s(S_t)\frac{\mathrm{W}_\theta^f(S_t)\,\mathrm{w}_\theta^f(b|S_t)}{\pi_\theta^f(A_t|S_t)}\left[\mathrm{1}_b(A_t) - \mathrm{w}_\theta^f(A_t|S_t)\right] f'(\frac{\pi_\theta^f(A_t|S_t)}{\pi_{\mathrm{ref}}(A_t|S_t)})\right] \ .$$

Applying again the tower property by taking the conditional expectation with respect to $\mathcal{G}_t$ in the third expectation gives

$$\frac{\partial J_T^f(\theta)}{\partial\theta(s,b)} = \mathbb{E}_{Z\sim\nu_T^f(\theta)}\left[\sum_{t=0}^{T-1}\sum_{\ell=0}^{t}\frac{\partial \log \pi_\theta^f(A_\ell|S_\ell)}{\partial\theta(s,b)}\gamma^t \mathsf{r}(S_t,A_t)\right]$$
$$- \lambda \mathbb{E}_{Z\sim\nu_T^f(\theta)}\left[\sum_{t=0}^{T-1}\sum_{\ell=0}^{t-1}\frac{\partial \log \pi_\theta^f(A_\ell|S_\ell)}{\partial\theta(s,b)}\gamma^t \,\mathrm{D}^f(\pi_\theta^f(\cdot|S_t)\|\pi_{\mathrm{ref}}(\cdot|S_t))\right]$$
$$- \lambda\, \mathbb{E}_{Z\sim\nu_T^f(\theta)}\left[\sum_{t=0}^{T-1}\gamma^t \mathrm{1}_s(S_t)\,\mathrm{W}_\theta^f(S_t)\,\mathrm{w}_\theta^f(b|S_t)\left[f'(\frac{\pi_\theta^f(b|S_t)}{\pi_{\mathrm{ref}}(b|S_t)}) - \sum_{a\in\mathcal{A}}\mathrm{w}_\theta^f(a|S_t)f'(\frac{\pi_\theta^f(a|S_t)}{\pi_{\mathrm{ref}}(a|S_t)})\right]\right] \ .$$

Taking $T\to+\infty$ and applying the dominated convergence theorem concludes the proof. □

The following lemma establishes a bound on the variance and bias of the stochastic estimator.

Lemma E.3. *Assume that, for some $\underline{\pi_{\mathrm{ref}}} > 0$, f satisfy $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$. There exists a constant $\beta_{H,\lambda} \geq 0$ such that, for any parameter $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, we have*

$$\left\|\mathrm{g}^f(\theta) - \frac{\partial v_\theta^f(\rho)}{\partial\theta}\right\|_2 \leq \beta_{H,\lambda} \ ,$$

where $\beta_{H,\lambda}$ is an upper bound on the bias defined as

$$\beta_{H,\lambda} := \frac{2\gamma^H(H+1)}{(1-\gamma)^2}\omega_f\left[2 + 2\lambda \mathrm{d}_f + \lambda(1-\gamma)\mathrm{y}_f\right] \ ,$$

where ω_f is defined in $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$ and d_f, and y_f are defined in (12).

Proof. Using the expression of the gradient truncated at H from (62) and (63), of the true gradient from Lemma E.2, and the triangle inequality, we have

$$\left\|\mathrm{g}^f(\theta) - \frac{\partial v_\theta^f(\rho)}{\partial\theta}\right\|_2 \leq \sum_{t=H}^{\infty}\sum_{\ell=0}^{t}\gamma^t\left\|\mathbb{E}_\rho^{\pi_\theta^f}\left[\frac{\partial \log \pi_\theta^f(A_\ell|S_\ell)}{\partial\theta}\mathsf{r}(S_t,A_t)\right]\right\|_2$$
$$+ \sum_{t=H}^{\infty}\sum_{\ell=0}^{t-1}\lambda\gamma^t\left\|\mathbb{E}_\rho^{\pi_\theta^f}\left[\frac{\partial \log \pi_\theta^f(A_\ell|S_\ell)}{\partial\theta}\,\mathrm{D}^f(\pi_\theta^f(\cdot|S_t)\|\pi_{\mathrm{ref}}(\cdot|S_t))\right]\right\|_2$$
$$+ \lambda\sum_{t=H}^{\infty}\gamma^t\left\|\mathbb{E}_\rho^{\pi_\theta^f}\left[\mathrm{F}_\theta^f(S_t)\right]\right\|_2 \ .$$

Next, applying Lemma E.1 combined with the triangle inequality yields

$$\left\|\mathrm{g}^f(\theta) - \frac{\partial v_\theta^f(\rho)}{\partial\theta}\right\|_2 \leq \sum_{t=H}^{\infty}\sum_{\ell=0}^{t}\gamma^t\mathbb{E}_\rho^{\pi_\theta^f}\left[\frac{2\,\mathrm{W}_\theta^f(S_\ell)\,\mathrm{w}_\theta^f(A_\ell|S_\ell)}{\pi_\theta^f(A_\ell|S_\ell)}\,|\mathsf{r}(S_t,A_t)|\right]$$

$$+\sum_{t=H}^{\infty}\sum_{\ell=0}^{t-1}\lambda\gamma^t\mathbb{E}_\rho^{\pi_\theta^f}\left[\frac{2\,\mathrm{W}_\theta^f(S_\ell)\,\mathrm{w}_\theta^f(A_\ell|S_\ell)}{\pi_\theta^f(A_\ell|S_\ell)}\,\mathrm{D}^f(\pi_\theta^f(\cdot|S_t)\|\pi_{\mathrm{ref}}(\cdot|S_t))\right]+\lambda\sum_{t=H}^{\infty}\gamma^t\mathbb{E}_\rho^{\pi_\theta^f}\left[2\,\mathrm{W}_\theta^f(S_t)\,\mathrm{Y}_\theta^f(S_t)\right]\ .$$

We define the following filtration, for $t\geq 0$,

$$\mathcal{G}_t=\sigma(S_0,A_0,\dots,S_t)\ ,$$

Next, applying the tower property of the conditional expectation by conditioning on $\mathcal{G}_t$, bounding the reward and the divergence respectively by 1, and $\max_{s\in\mathcal{S}}\sup_{\nu\in\mathcal{P}(\mathcal{A})}\mathrm{D}^f(\nu\|\pi_{\mathrm{ref}}(\cdot|s))\}$, and using that $\mathrm{w}_\theta^f(\cdot|s)\in\mathcal{P}(\mathcal{A})$ for any $s\in\mathcal{S}$ yields

$$\begin{aligned}\left\|\mathrm{g}^f(\theta)-\frac{\partial v_\theta^f(\rho)}{\partial\theta}\right\|_2&\leq 2\sum_{t=H}^{\infty}\sum_{\ell=0}^{t}\gamma^t\left\|\mathrm{W}_\theta^f\right\|_\infty\\&+2\lambda\sum_{t=H}^{\infty}\sum_{\ell=0}^{t-1}\gamma^t\left\|\mathrm{W}_\theta^f\right\|_\infty\sup_{(s,\nu)\in\mathcal{S}\times\mathcal{P}(\mathcal{A})}\mathrm{D}^f(\nu\|\pi_{\mathrm{ref}}(\cdot|s))\}+2\lambda\sum_{t=H}^{\infty}\gamma^t\left\|\mathrm{W}_\theta^f\right\|_\infty\left\|\mathrm{Y}_\theta^f\right\|_\infty\ .\end{aligned}$$

Finally, using that

$$\sum_{t=H}^{\infty}\gamma^t\leq\frac{\gamma^H}{1-\gamma}\ ,\quad\sum_{t=H}^{\infty}\gamma^t(t-1)\leq 2\gamma^H\frac{H}{(1-\gamma)^2}\ ,\quad\sum_{t=H}^{\infty}\gamma^t t\leq 2\gamma^H\frac{H+1}{(1-\gamma)^2}\ ,$$

combined with Lemma B.6 completes the proof. □

Lemma E.4. *Assume that, for some $\underline{\pi_{\mathrm{ref}}}>0$, f satisfy $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$. For any $\theta\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, it holds that*

$$\mathbb{E}_{Z\sim[\nu(\theta)]^{\otimes B}}\left[\left\|\mathrm{g}^f(\theta)-\mathrm{g}_Z^f(\theta)\right\|_2^2\right]\leq\frac{\sigma_{\lambda,f}^2}{B}\ ,$$

where we have defined

$$\sigma_{\lambda,f}^2:=\frac{12}{(1-\gamma)^4}\left[\omega_f^3+\lambda^2\gamma^2\omega_f^3\mathrm{d}_f^2+\lambda^2(1-\gamma)^2\omega_f^2\mathrm{y}_f^2\right]\ ,$$

and where ω_f, d_f, and y_f are defined in $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$ and (12).

Proof. Firstly, define for $\xi=(s_h,a_h)_{h=0}^{H-1}\in(\mathcal{S}\times\mathcal{A})^H$

$$\begin{aligned}u_\xi(\theta):=&\sum_{h=0}^{H-1}\sum_{\ell=0}^{h}\frac{\partial\log\pi_\theta^f(a_\ell|s_\ell)}{\partial\theta}\gamma^h\mathsf{r}(s_h,a_h)\\&-\lambda\sum_{h=0}^{H-1}\sum_{\ell=0}^{h-1}\frac{\partial\log\pi_\theta^f(a_\ell|s_\ell)}{\partial\theta}\gamma^h\,\mathrm{D}^f(\pi_\theta^f(\cdot|s_h)\|\pi_{\mathrm{ref}}(\cdot|s_h))-\lambda\sum_{h=0}^{H-1}\gamma^h\mathrm{F}_\theta^f(s_h)\ ,\end{aligned}$$

Importantly, for a given $Z\sim[\nu(\theta)]^{\otimes B}$, denoting by $Z=(Z_0,\dots,Z_{B-1})$, it holds that

$$\mathrm{g}_Z^f(\theta)=\frac{1}{B}\sum_{b=0}^{B-1}u_{Z_b}(\theta)\ ,\quad\text{and }\mathrm{g}^f(\theta)=\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[u_{\mathfrak{T}}(\theta)\right]\ .$$

Using that the variables $(Z_0,\dots,Z_{B-1})$ are independent and identically distributed, we get

$$\mathbb{E}_{Z\sim[\nu(\theta)]^{\otimes B}}\left[\left\|\mathrm{g}^f(\theta)-\mathrm{g}_Z^f(\theta)\right\|_2^2\right]=\mathbb{E}_{Z\sim[\nu(\theta)]^{\otimes B}}\left[\left\|\frac{1}{B}\sum_{b=0}^{B-1}u_{Z_b}(\theta)-\mathrm{g}^f(\theta)\right\|_2^2\right]$$

$$= \frac{1}{B}\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[\left\|u_{\mathfrak{T}}(\theta) - \mathrm{g}^f(\theta)\right\|_2^2\right] \leq \frac{1}{B}\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[\|u_{\mathfrak{T}}(\theta)\|_2^2\right] \quad , \quad (69)$$

where in the last inequality, we used that the second moment of a random variable dominates its variance. Next, using Jensen's inequality combined with the convexity of the square function, we have

$$\begin{aligned}
\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[\|u_{\mathfrak{T}}(\theta)\|_2^2\right] &\leq 3\mathbb{E}_{\mathfrak{T}}\left[\left\|\sum_{h=0}^{H-1}\sum_{\ell=0}^{h}\frac{\partial \log \pi_\theta^f(A_\ell|S_\ell)}{\partial\theta}\gamma^h \mathsf{r}(S_h, A_h)\right\|_2^2\right] \\
&\quad + 3\lambda^2\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[\left\|\sum_{h=0}^{H-1}\sum_{\ell=0}^{h-1}\frac{\partial \log \pi_\theta^f(A_\ell|S_\ell)}{\partial\theta}\gamma^h \,\mathrm{D}^f(\pi_\theta^f(\cdot|S_h)\|\pi_{\mathrm{ref}}(\cdot|S_h))\right\|_2^2\right] \\
&\quad + 3\lambda^2\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[\left\|\sum_{h=0}^{H-1}\gamma^h \mathrm{F}_\theta^f(S_h)\right\|_2^2\right] \quad ,
\end{aligned}$$

Applying the triangle inequality and the fact that the reward and the divergence are positive yields

$$\begin{aligned}
\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[\|u_{\mathfrak{T}}(\theta)\|_2^2\right] &\leq 3\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[\left(\sum_{h=0}^{H-1}\sum_{\ell=0}^{h}\gamma^h\left\|\frac{\partial \log \pi_\theta^f(A_\ell|S_\ell)}{\partial\theta}\right\|_2 \mathsf{r}(S_h, A_h)\right)^2\right] \\
&\quad + 3\lambda^2\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[\left(\sum_{h=0}^{H-1}\sum_{\ell=0}^{h-1}\gamma^h\left\|\frac{\partial \log \pi_\theta^f(A_\ell|S_\ell)}{\partial\theta}\right\|_2 \mathrm{D}^f(\pi_\theta^f(\cdot|S_h)\|\pi_{\mathrm{ref}}(\cdot|S_h))\right)^2\right] \\
&\quad + 3\lambda^2\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[\left(\sum_{h=0}^{H-1}\gamma^h\left\|\mathrm{F}_\theta^f(S_h)\right\|_2\right)^2\right] \quad ,
\end{aligned}$$

Combining Lemma E.1 and the fact that the reward is bounded between 0 and 1 gives

$$\begin{aligned}
\mathbb{E}_{\mathfrak{T}}\left[\|u_{\mathfrak{T}}(\theta)\|_2^2\right] &\leq 3\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[\left(\sum_{h=0}^{H-1}\sum_{\ell=0}^{h}2\gamma^{h/2}\cdot\gamma^{h/2}\frac{\mathrm{W}_\theta^f(S_\ell)\,\mathrm{w}_\theta^f(A_\ell|S_\ell)}{\pi_\theta^f(A_\ell|S_\ell)}\right)^2\right] \\
&\quad + 3\lambda^2\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[\left(\sum_{h=0}^{H-1}\sum_{\ell=0}^{h-1}2\gamma^{h/2}\cdot\gamma^{h/2}\frac{\mathrm{W}_\theta^f(S_\ell)\,\mathrm{w}_\theta^f(A_\ell|S_\ell)}{\pi_\theta^f(A_\ell|S_\ell)}\sup_{(s,\nu)\in\mathcal{S}\times\mathcal{P}(\mathcal{A})}\mathrm{D}^f(\nu\|\pi_{\mathrm{ref}}(\cdot|s))\right)^2\right] \\
&\quad + 3\lambda^2\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[\left(\sum_{h=0}^{H-1}2\gamma^{h/2}\cdot\gamma^{h/2}\left\|\mathrm{W}_\theta^f\right\|_\infty\left\|\mathrm{Y}_\theta^f\right\|_\infty\right)^2\right] \quad ,
\end{aligned}$$

Next, applying the Cauchy-Schwarz inequality gives

$$\begin{aligned}
\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[\|u_{\mathfrak{T}}(\theta)\|_2^2\right] &\leq 3\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[\left(\sum_{h=0}^{H-1}\sum_{\ell=0}^{h}4\gamma^h\right)\left(\sum_{h=0}^{H-1}\sum_{\ell=0}^{h}\gamma^h\frac{\mathrm{W}_\theta^f(S_\ell)^2\,\mathrm{w}_\theta^f(A_\ell|S_\ell)^2}{\pi_\theta^f(A_\ell|S_\ell)^2}\right)\right] \\
&\quad + 3\lambda^2\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[\left(\sum_{h=0}^{H-1}\sum_{\ell=0}^{h-1}4\gamma^h\right)\left(\sum_{h=0}^{H-1}\sum_{\ell=0}^{h-1}\gamma^h\frac{\mathrm{W}_\theta^f(S_\ell)^2\,\mathrm{w}_\theta^f(A_\ell|S_\ell)^2}{\pi_\theta^f(A_\ell|S_\ell)^2}\sup_{(s,\nu)\in\mathcal{S}\times\mathcal{P}(\mathcal{A})}\mathrm{D}^f(\nu\|\pi_{\mathrm{ref}}(\cdot|s))^2\right)\right] \\
&\quad + 3\lambda^2\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[\left(\sum_{h=0}^{H-1}4\gamma^h\right)\left(\sum_{h=0}^{H-1}\gamma^h\left\|\mathrm{W}_\theta^f\right\|_\infty^2\left\|\mathrm{Y}_\theta^f\right\|_\infty^2\right)\right] \quad ,
\end{aligned}$$

We define the following filtration, for $t \geq 0$,

$$\mathcal{G}_t = \sigma(S_0, A_0, \dots, S_t) \ ,$$

Next, applying the tower property of the conditional expectation by conditioning on $\mathcal{G}_t$, and using that $\mathrm{w}_\theta^f(\cdot|s) \in \mathcal{P}(\mathcal{A})$ for any $s \in \mathcal{S}$ yields

$$\begin{aligned}
\mathbb{E}_{\mathfrak{T}\sim\nu(\theta)}\left[\|u_{\mathfrak{T}}(\theta)\|_2^2\right] &\leq 12\left\|\mathrm{W}_\theta^f\right\|_\infty^2 \max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left\{\frac{\mathrm{w}_\theta^f(a|s)}{\pi_\theta^f(a|s)}\right\}\left(\sum_{h=0}^{H-1}\gamma^h(h+1)\right)^2 \\
&+ 12\lambda^2\left\|\mathrm{W}_\theta^f\right\|_\infty^2 \max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left\{\frac{\mathrm{w}_\theta^f(a|s)}{\pi_\theta^f(a|s)}\right\} \sup_{(s,\nu)\in\mathcal{S}\times\mathcal{P}(\mathcal{A})}\left\{\mathrm{D}^f(\nu\|\pi_{\mathrm{ref}}(\cdot|s))^2\right\}\left(\sum_{h=0}^{H-1}\gamma^h h\right)^2 \\
&+ 12\lambda^2\left\|\mathrm{W}_\theta^f\right\|_\infty^2\left\|\mathrm{Y}_\theta^f\right\|_\infty^2\left(\sum_{h=0}^{H-1}\gamma^h\right)^2 \ .
\end{aligned}$$

Next using

$$\sum_{t=0}^{H-1}\gamma^t \leq \frac{1}{1-\gamma} \ , \quad \sum_{t=0}^{H-1}\gamma^t t \leq \frac{\gamma}{(1-\gamma)^2} \ , \sum_{t=0}^{H-1}\gamma^t(t+1) \leq \frac{1}{(1-\gamma)^2} \ ,$$

and plugging in the obtained bound in (69), combined with Lemma B.6 concludes the proof. □

E.2 Sample complexity of Stochastic f-PG

We now derive convergence rates for f-PG. First, we define the following quantity, which will be the Polyak-Łojasiewicz constant of our function over the optimization space, where policies are guaranteed not to be too ill-conditioned, i.e., all their entries are larger than the $\tau_{\lambda,f}$ defined in (60),

$$\underline{\mu}_{\lambda,f} := \frac{\lambda(1-\gamma)\rho_{\min}^2\zeta_f^2}{\omega_f^2}\underline{\pi_{\mathrm{ref}}}^2 \min_{x\in[\tau_{\lambda,f},\frac{1}{\underline{\pi_{\mathrm{ref}}}}]} f''(x)^{-2} \ . \tag{70}$$

As we will prove in this subsection, this quantity represents a lower bound of the non-uniform Łojasiewicz coefficient along the trajectory. In the following lemma, we give a simpler lower bound of $\underline{\mu}_{\lambda,f}$ provided that λ is not too large.

Lemma E.5. *Assume that, for some $\underline{\pi_{\mathrm{ref}}} > 0$, f and π_{ref} satisfy $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\mathrm{ref}}})$ respectively. Assume in addition that the initial distribution ρ satisfies $\boldsymbol{A}_\rho$ and that λ satisfies*

$$\lambda \leq \frac{4}{(1-\gamma)^2\rho_{\min}}\min\left(\frac{4}{|f'(\iota_f)|}, \frac{1}{\left|f'(\frac{1}{2})\right|}, \frac{4}{\left|f'(\frac{1}{2}\underline{\pi_{\mathrm{ref}}})\right|}\right) \ .$$

In this case, it holds that

$$\underline{\mu}_{\lambda,f} = \frac{\lambda(1-\gamma)\rho_{\min}^2\zeta_f^2}{\omega_f^2}\underline{\pi_{\mathrm{ref}}}^2(f^\star)''\left(-\frac{16+8\gamma\lambda\mathrm{d}_f}{\lambda(1-\gamma)^2\rho_{\min}}\right)^2 \ .$$

Additionally if $\lambda \leq 1/\mathrm{d}_f$, then it holds that

$$\underline{\mu}_{\lambda,f} \geq \frac{\lambda(1-\gamma)\rho_{\min}^2\zeta_f^2}{\omega_f^2}\underline{\pi_{\mathrm{ref}}}^2(f^\star)''\left(-\frac{24}{\lambda(1-\gamma)^2\rho_{\min}}\right)^2 \ .$$

Proof. First note that the first condition on λ implies that $\tau_{\lambda,f} < \iota_f$, with ι_f defined in $\mathbf{A}_f(\underline{\pi}_{\mathrm{ref}})$, and thus

$$\min_{x\in[\tau_{\lambda,f},\frac{1}{\underline{\pi}_{\mathrm{ref}}}]} f''(x)^{-2} = f''(\tau_{\lambda,f})^{-2}.$$

Additionally, the second and third conditions on λ guarantee that the minimum of $\tau_{\lambda,f}$ in (60) is attained in the first term, that is

$$\tau_{\lambda,f} = [f']^{-1}\left(-\frac{16+8\gamma\lambda \mathrm{d}_f}{\lambda(1-\gamma)^2\rho_{\min}}\right) .$$

Finally, we recall that the convex conjugate of f defined in (28) satisfies, for any $y \in (-\infty, f'(\frac{1}{\underline{\pi}_{\mathrm{ref}}}))$,

$$(f^\star)''(y) = \frac{1}{f''([f']^{-1}(y))} .$$

Thus, we obtain that $f''(\tau_{\lambda,f})^{-2} = (f^\star)''(\frac{-16+8\gamma\lambda \mathrm{d}_f}{\lambda(1-\gamma)^2\rho_{\min}})^2$ which concludes the proof. □

In the following, we define the filtration adapted to the iterates of f-PG as

$$\mathcal{F}_t := \sigma\Big(Z_t : t \in \{0,\dots,T-1\}\Big) .$$

The following theorem gives convergence rates of f-PG.

Theorem E.6. *Assume that, for some $\underline{\pi}_{\mathrm{ref}} > 0$, f and π_{ref} satisfy $\mathbf{A}_f(\underline{\pi}_{\mathrm{ref}})$ and $\mathbf{P}(\underline{\pi}_{\mathrm{ref}})$ respectively. Assume in addition that the initial distribution ρ satisfies $\mathbf{A}_\rho$. Fix $\eta \le 1/2L_{\lambda,f}$, a given temperature λ, and consider the iterates $(\theta_t)_{t=0}^\infty$ of the algorithm f-PG. It holds almost surely that*

$$\inf_{t\ge 0} \mu_{\lambda,f}(\theta_t) \ge \underline{\mu}_{\lambda,f} . \tag{71}$$

Additionally, for any $t \ge 0$ we have that

$$v^f_\star(\rho) - \mathbb{E}\left[v^f_{\theta_t}(\rho)\right] \le (1-\underline{\mu}_{\lambda,f}\eta/4)^t (v^f_\star(\rho) - v^f_{\theta_0}(\rho)) + \frac{6\eta\sigma^2_{\lambda,f}}{B\underline{\mu}_{\lambda,f}} + \frac{6\beta^2_{H,\lambda}}{\underline{\mu}_{\lambda,f}} .$$

Proof. Recall that for any $t \in [T]$, we have that

$$\theta_t = \widetilde{\mathrm{IMP}}^f(\bar{\theta}_t) .$$

Hence, by Lemma D.2, for any $t \in [T]$ it holds that

$$\pi^f_{\theta_t} \ge \tau_{\lambda,f}\underline{\pi}_{\mathrm{ref}} .$$

Combining the previous inequality with the expression of the coefficient $\mu_{\lambda,f}$ provided in Theorem C.1 proves the first statement of the lemma. Next, using Theorem B.12 gives

$$v^f_{\theta_{t+1}}(\rho) \ge v^f_{\theta_t}(\rho) + 2\eta\langle \nabla v^f_{\theta_t}(\rho), \mathrm{g}^f_{Z_t}(\theta_t)\rangle - \frac{\eta^2 L_{\lambda,f}}{2}\left\|\mathrm{g}^f_{Z_t}(\theta_t)\right\|_2^2 .$$

Next, taking the conditional expectation with respect to $\mathcal{F}_t$ and adding and subtracting $\nabla v^f_{\theta_t}(\rho)$ in the dot product gives

$$\begin{aligned}\mathbb{E}\left[v^f_{\theta_{t+1}}(\rho)\middle|\mathcal{F}_t\right] \ge v^f_{\theta_t}(\rho) + 2\eta\left\|\nabla v^f_{\theta_t}(\rho)\right\|^2 + \underbrace{2\eta\langle \nabla v^f_{\theta_t}(\rho), \mathrm{g}^f(\theta_t) - \nabla v^f_{\theta_t}(\rho)\rangle}_{(\mathbf{K_1})} \\ - \frac{\eta^2 L_{\lambda,f}}{2}\underbrace{\mathbb{E}\left[\left\|\mathrm{g}^f_{Z_t}(\theta_t)\right\|_2^2\middle|\mathcal{F}_t\right]}_{(\mathbf{K_2})} .\end{aligned} \tag{72}$$

We now bound each of these terms separately.

Bounding $\mathbf{K}_1$. Using the Cauchy-Schwarz inequality, yields

$$2\eta\langle\nabla v^f_{\theta_t}(\rho), \mathrm{g}^f(\theta_t) - \nabla v^f_{\theta_t}(\rho)\rangle \geq -2\eta \left\|\nabla v^f_{\theta_t}(\rho)\right\|_2 \left\|\mathrm{g}^f(\theta_t) - \nabla v^f_{\theta_t}(\rho)\right\|_2$$
$$= -2\cdot\eta^{1/2}\left\|\nabla v^f_{\theta_t}(\rho)\right\|_2 \cdot \eta^{1/2}\beta_{H,\lambda} \ ,$$

where in the last inequality we used Lemma E.3. Next, using Young's inequality gives

$$2\eta\langle\nabla v^f_{\theta_t}(\rho), \mathrm{g}^f(\theta_t) - \nabla v^f_{\theta_t}(\rho)\rangle \geq -\eta\left\|\nabla v^f_{\theta_t}(\rho)\right\|_2^2 - \eta\beta^2_{H,\lambda} \ . \tag{73}$$

Bounding $\mathbf{K}_2$. Using the convexity of the square function with Jensen's inequality gives

$$\mathbb{E}\left[\left\|\mathrm{g}^f_{Z_t}(\theta_t)\right\|_2^2 \middle| \mathcal{F}_t\right] = \mathbb{E}\left[\left\|\mathrm{g}^f_{Z_t}(\theta_t) - \mathrm{g}^f Z_t(\theta_t) + \mathrm{g}^f Z_t(\theta_t) - \nabla v^f_{\theta_t}(\rho) + \nabla v^f_{\theta_t}(\rho)\right\|_2^2 \middle| \mathcal{F}_t\right]$$
$$\leq 3\beta^2_{H,\lambda} + 3\left\|\nabla v^f_{\theta_t}(\rho)\right\|_2^2 + \frac{3\sigma^2_{\lambda,f}}{B} \ , \tag{74}$$

where we used Lemma E.3 and Lemma E.4. Plugging in the bounds (73) on $\mathbf{K}_1$ and (74) on $\mathbf{K}_2$ in (72) gives

$$\mathbb{E}\left[v^f_{\theta_{t+1}}(\rho)\middle|\mathcal{F}_t\right] \geq v^f_{\theta_t}(\rho) + \eta\left\|\nabla v^f_{\theta_t}(\rho)\right\|^2 - \left(\eta + \frac{3\eta^2 L_{\lambda,f}}{2}\right)\beta^2_{H,\lambda} - \frac{3\eta^2 L_{\lambda,f}}{2}\left\|\nabla v^f_{\theta_t}(\rho)\right\|_2^2 - \frac{3\eta^2 L_{\lambda,f}\sigma^2_{\lambda,f}}{2B} \ .$$

Taking the expectation with respect to all the stochasticity, multiplying both sides by -1, and adding $v^f_\star(\rho)$ gives

$$v^f_\star(\rho) - \mathbb{E}\left[v^f_{\theta_{t+1}}(\rho)\right] \leq v^f_\star(\rho) - \mathbb{E}\left[v^f_{\theta_t}(\rho)\right] - \eta\left(1 - \frac{3\eta L_{\lambda,f}}{2}\right)\left\|\nabla v^f_{\theta_t}(\rho)\right\|^2$$
$$+ \left(\eta + \frac{3\eta^2 L_{\lambda,f}}{2}\right)\beta^2_{H,\lambda} + \frac{3\eta^2 L_{\lambda,f}\sigma^2_{\lambda,f}}{2B} \ .$$

Next using Theorem C.1 combined with (71) yields

$$v^f_\star(\rho) - \mathbb{E}\left[v^f_{\theta_{t+1}}(\rho)\right] \leq \left(1 - \eta\underline{\mu}_{\lambda,f}\left(1 - \frac{3\eta L_{\lambda,f}}{2}\right)\right)\left(v^f_\star(\rho) - \mathbb{E}\left[v^f_{\theta_t}(\rho)\right]\right)$$
$$+ 2\eta\beta^2_{H,\lambda} + \frac{3\eta^2 L_{\lambda,f}\sigma^2_{\lambda,f}}{2B} \ ,$$

where we used $3\eta L_{\lambda,f}/2 \leq 1$. Finally, using that $\eta \leq 1/2L_{\lambda,f}$ to bound $1 - 3\eta L_{\lambda,f}/2$ and unrolling the recursion concludes the proof. □

Next, we provide the sample complexity of f-PG for solving the f-regularized objective.

Corollary E.7. *Assume that, for some $\underline{\pi}_{\mathrm{ref}} > 0$, f and π_{ref} satisfy $\mathbf{A}_f(\underline{\pi}_{\mathrm{ref}})$ and $\boldsymbol{P}(\underline{\pi}_{\mathrm{ref}})$ respectively. Assume in addition that the initial distribution ρ satisfies $\mathbf{A}_\rho$. Fix any $\epsilon > 0$ and $\lambda > 0$. Setting*

$$H \geq \frac{4}{(1-\gamma)^2} + \frac{1}{1-\gamma}\ln\left(\frac{216\omega_f^2}{\epsilon\underline{\mu}_{\lambda,f}(1-\gamma)^4}\left[4 + 4\lambda^2\mathrm{d}_f^2 + \lambda^2(1-\gamma)^2\mathrm{y}_f^2\right]\right), \tag{75}$$

and

$$\eta \leq \min\left(\frac{1}{2L_{\lambda,f}}, \frac{\epsilon B\underline{\mu}_{\lambda,f}}{18\sigma^2_{\lambda,f}}\right) \ , \tag{76}$$

and

$$T \geq \frac{4}{\underline{\mu}_{\lambda,f}} \max\left(2L_{\lambda,f}, \frac{18\sigma^2_{\lambda,f}}{\epsilon B \underline{\mu}_{\lambda,f}}\right) \cdot \ln\left(\frac{3(v^f_\star(\rho) - v^f_{\theta_0}(\rho))}{\epsilon}\right) \quad , \tag{77}$$

guarantees that

$$v^f_\star(\rho) - \mathbb{E}\left[v^f_{\theta_t}(\rho)\right] \leq \epsilon \ .$$

Proof. As $\eta \leq 1/2L_{\lambda,f}$, then by using Theorem E.6, it holds that

$$v^f_\star(\rho) - \mathbb{E}\left[v^f_{\theta_T}(\rho)\right] \leq \underbrace{(1 - \underline{\mu}_{\lambda,f}\eta/4)^T (v^f_\star(\rho) - v^f_{\theta_0}(\rho))}_{(\mathbf{U})} + \underbrace{\frac{6\eta\sigma^2_{\lambda,f}}{B\underline{\mu}_{\lambda,f}}}_{(\mathbf{V})} + \underbrace{\frac{6\beta^2_{H,\lambda}}{\underline{\mu}_{\lambda,f}}}_{(\mathbf{W})} \ .$$

Next, we aim to show that under our conditions on T, H, and η, each of these terms is smaller than $\epsilon/3$.

Bounding V. We start with the term **V**, which gives a condition on the step size. In particular, setting

$$\eta \leq \frac{\epsilon B \underline{\mu}_{\lambda,f}}{18\sigma^2_{\lambda,f}} \ ,$$

guarantees that $\mathbf{V} \leq \epsilon/3$, which, together with $\eta \leq 1/(2L_{\lambda,f})$, gives the condition (76).

Bounding U. In order to ensure that **U** is smaller then $\epsilon/3$, we need T to satisfy

$$T \geq \frac{4}{\eta \underline{\mu}_{\lambda,f}} \ln\left(\frac{\epsilon}{3\left(v^f_\star(\rho) - v^f_{\theta_0}(\rho)\right)}\right) \ ,$$

which, combined with the inequality $\ln(1+x) \leq x$ for $x > -1$, ensures that **U** is smaller than $\epsilon/3$, and the condition (77) follows from (76).

Bounding W. Using Lemma E.3 and Jensen's inequality, it holds that

$$\mathbf{W} \leq \frac{6}{\underline{\mu}_{\lambda,f}} \cdot \frac{4\gamma^{2H}(H+1)^2}{(1-\gamma)^4} \omega^2_f \left[12 + 12\lambda^2 \mathrm{d}^2_f + 3\lambda^2(1-\gamma)^2 \mathrm{y}^2_f\right] \ ,$$

Next, we remark that for any $a > 0$, we have $a\gamma^{2H}(H+1)^2 \leq \epsilon/3$ for

$$H \geq \frac{1}{1-\gamma} \max\left(\frac{4}{1-\gamma}, \ln\left(\frac{3a}{\epsilon}\right)\right) \ .$$

Taking $a = \frac{6}{\underline{\mu}_{\lambda,f}} \cdot \frac{4}{(1-\gamma)^4} \omega^2_f \left[12 + 12\lambda^2 \mathrm{d}^2_f + 3\lambda^2(1-\gamma)^2 \mathrm{y}^2_f\right]$ gives $\mathbf{W} < \epsilon/3$, provided that (75) holds. □

E.3 Guarantees on the non-regularized problem

A key criterion for evaluating the quality of a reinforcement learning algorithm is its sample efficiency in solving the original, unregularized objective. To this end, we recall a result from Geist et al. (2019), which establishes a connection between regularized and unregularized value functions, and further characterizes the performance of the optimal f-regularized policy when evaluated in the original unregularized MDP.

Lemma E.8 (Proposition 3 and Theorem 2 of Geist et al. (2019)). *For any policy π, and state $s \in \mathcal{S}$ it holds that*

$$\left|v_\pi^f(s) - v_\pi(s)\right| \leq \frac{\lambda \mathrm{d}_f}{1-\gamma} \ .$$

Additionally, denote by $\pi_\star^f$ the optimal regularized policy. For any state $s \in \mathcal{S}$, It holds that

$$v_\star^f(s) - v_{\pi_\star^f} \leq \frac{\lambda \mathrm{d}_f}{1-\gamma} \ .$$

The following theorem gives the convergence rate for the non-regularised problem.

Theorem E.9. *Assume that, for some $\underline{\pi_{\mathrm{ref}}} > 0$, f and π_{ref} satisfy $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\mathrm{ref}}})$ respectively. Assume in addition that the initial distribution ρ satisfies $\boldsymbol{A}_\rho$. Fix $\eta \leq 1/2L_{\lambda,f}$, a given temperature λ, and consider the iterates $(\theta)_{t=0}^\infty$ of the algorithm f-PG. For any $t \geq 0$ we have that*

$$v_\star(\rho) - \mathbb{E}\left[v_{\theta_t}(\rho)\right] \leq (1 - \underline{\mu}_{\lambda,f}\eta/4)^t (v_\star^f(\rho) - v_{\theta_0}^f(\rho)) + \frac{6\eta\sigma_{\lambda,f}^2}{B\underline{\mu}_{\lambda,f}} + \frac{6\beta_{H,\lambda}^2}{\underline{\mu}_{\lambda,f}} + \frac{2\lambda \mathrm{d}_f}{1-\gamma} \ .$$

Proof. The proof holds from Theorem E.6 and Lemma E.8. □

Finally, we give the sample complexity of f-PG to solve the unregularised problem.

Corollary E.10. *Assume that, for some $\underline{\pi_{\mathrm{ref}}} > 0$, f and π_{ref} satisfy $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$ and $\boldsymbol{P}(\underline{\pi_{\mathrm{ref}}})$ respectively. Assume in addition that the initial distribution ρ satisfies $\boldsymbol{A}_\rho$. Consider any constant $\mathrm{c}_f > 0$ such that*

$$\mathrm{c}_f \leq \min(\frac{1}{\mathrm{d}_f}, \frac{1}{\mathrm{y}_f}, 1) \ .$$

Fix any $(1-\gamma)^{-1} > \epsilon > 0$ such that

$$\epsilon < \frac{16}{(1-\gamma)^3 \rho_{\min}} \min\left(\frac{4}{|f'(\iota_f)|}, \frac{1}{\left|f'(\frac{1}{2})\right|}, \frac{4}{\left|f'(\frac{1}{2}\underline{\pi_{\mathrm{ref}}})\right|}\right) \ , \quad \textit{and set } \lambda = \frac{(1-\gamma)\epsilon}{4}\mathrm{c}_f \ . \tag{78}$$

Define

$$d(\epsilon) = (f^\star)''\left(\frac{-96}{\epsilon \mathrm{c}_f (1-\gamma)^3 \rho_{\min}}\right)^2 \ , \quad \ell(\epsilon) = \log\left(\frac{6(v_\star^f(\rho) - v_{\theta_0}^f(\rho))}{\epsilon}\right) \ . \tag{79}$$

Additionally, define the three following constants which depend only on f as

$$C_f^{(1)} = \frac{16\omega_f^2}{\mathrm{c}_f \zeta_f^2} \ , \quad C_f^{(2)} = 48\omega_f\left(\gamma\omega_f + (1-\gamma)\kappa_f\right) \ , \quad C_f^{(3)} = \frac{3456\omega_f^5}{\zeta_f^2 \mathrm{c}_f} \ . \tag{80}$$

Setting

$$H \geq \frac{4}{(1-\gamma)^2} + \frac{1}{1-\gamma} \ln\left(\frac{297\omega_f^2 C_1^f}{\epsilon d(\epsilon)(1-\gamma)^6 \rho_{\min}^2 \underline{\pi_{\mathrm{ref}}}^2}\right) \ ,$$

and

$$\eta \leq \min\left(\frac{(1-\gamma)^3}{C_f^{(2)}}, \frac{\epsilon^2 d(\epsilon)(1-\gamma)^6 B \rho_{\min}^2 \underline{\pi_{\mathrm{ref}}}^2}{C_f^{(3)}}\right) \ , \tag{81}$$

and

$$T \geq \frac{16 C_f^{(1)} \ell(\epsilon)}{\epsilon d(\epsilon)(1-\gamma)^2 \rho_{\min}^2 \underline{\pi_{\text{ref}}}^2} \max\left(\frac{C_f^{(2)}}{(1-\gamma)^3}, \frac{C_f^{(3)}}{\epsilon^2 d(\epsilon)(1-\gamma)^6 B \rho_{\min}^2 \underline{\pi_{\text{ref}}}^2} \right) \quad , \tag{82}$$

guarantees that

$$v_\star(\rho) - \mathbb{E}\left[v_{\theta_t}(\rho)\right] \leq \epsilon \ .$$

Proof. First, note that the conditions (78) on ϵ and λ guarantee that

$$\lambda \leq \frac{4}{(1-\gamma)^2 \rho_{\min}} \min\left(\frac{4}{|f'(\iota_f)|}, \frac{1}{|f'(\frac{1}{2})|}, \frac{4}{|f'(\frac{1}{2}\underline{\pi_{\text{ref}}})|} \right) \quad ,$$

Thus, using Lemma E.5, and the fact that $\epsilon < (1-\gamma)^{-1}$, we have that

$$\underline{\mu}_{\lambda,f} \geq \frac{\lambda(1-\gamma)\rho_{\min}^2 \zeta_f^2}{\omega_f^2} \underline{\pi_{\text{ref}}}^2 (f^\star)'' \left(\frac{-24}{\lambda(1-\gamma)^2 \rho_{\min}} \right)^2 \ .$$

Plugging in the expression of λ from (78) yields the following simplified expression

$$\underline{\mu}_{\lambda,f} \geq \frac{\epsilon c_f (1-\gamma)^2 \rho_{\min}^2 \zeta_f^2}{4\omega_f^2} \underline{\pi_{\text{ref}}}^2 (f^\star)'' \left(\frac{-96}{\epsilon c_f (1-\gamma)^3 \rho_{\min}} \right)^2 = \frac{4\epsilon(1-\gamma)^2 \rho_{\min}^2 \underline{\pi_{\text{ref}}}^2}{C_f^{(1)}} d(\epsilon) \ , \tag{83}$$

where $d(\epsilon)$ and $C_f^{(1)}$ are defined respectively in (79) and (80). Additionally, combining Lemma E.3, Lemma E.4 and Theorem B.12 and the expression of λ yields

$$\sigma_{\lambda,f}^2 \leq \frac{24}{(1-\gamma)^4} \omega_f^3 \quad , \quad \beta_{H,\lambda} \leq \frac{6\gamma^H H}{(1-\gamma)^2} \omega_f \quad , \quad L_{\lambda,f} = \frac{13\omega_f \left(\omega_f + (1-\gamma)\kappa_f\right)}{(1-\gamma)^3} \quad , \tag{84}$$

where we used that under $\mathbf{A}_f(\underline{\pi_{\text{ref}}})$, we have $\omega_f \geq 1$. Using Theorem E.9 and the fact that $\lambda \leq (1-\gamma)\epsilon/4\mathrm{d}_f$, we have that

$$v_\star(\rho) - \mathbb{E}\left[v_{\theta_t}(\rho)\right] \leq \underbrace{(1 - \underline{\mu}_{\lambda,f}\eta/4)^t (v_\star^f(\rho) - v_{\theta_0}^f(\rho))}_{(\mathbf{U}')} + \underbrace{\frac{6\eta\sigma_{\lambda,f}^2}{B\underline{\mu}_{\lambda,f}}}_{(\mathbf{V}')} + \underbrace{\frac{6\beta_{H,\lambda}^2}{\underline{\mu}_{\lambda,f}}}_{(\mathbf{W}')} + \frac{\epsilon}{2} \ .$$

Next, we aim to show that under our conditions on T, H, and η, each of these terms is smaller than $\epsilon/6$.

Bounding V'. We start with the term **V'** which gives a condition on the step-size. Using (83) and (84), it holds that

$$\mathbf{V'} \leq \frac{144\eta\omega_f^3}{(1-\gamma)^4 B \underline{\mu}_{\lambda,f}} \leq \frac{576\eta\omega_f^5}{(1-\gamma)^6 B} \frac{1}{\epsilon c_f \rho_{\min}^2 \zeta_f^2 \underline{\pi_{\text{ref}}}^2 (f^\star)'' \left(\frac{-96}{\epsilon c_f (1-\gamma)^3 \rho_{\min}}\right)^2} = \frac{C_f^{(3)} \eta}{6\epsilon d(\epsilon)(1-\gamma)^6 B \rho_{\min}^2 \underline{\pi_{\text{ref}}}^2},$$

where $C_f^{(3)}$ is defined in (80). In particular, setting

$$\eta \leq \frac{\epsilon^2 d(\epsilon)(1-\gamma)^6 B \rho_{\min}^2 \underline{\pi_{\text{ref}}}^2}{C_f^{(3)}} \quad ,$$

guarantees that $\mathbf{V'}' \leq \epsilon/6$, which together with the condition $\eta \leq 1/(2L_{\lambda,f})$, gives the condition (81).

Bounding U'. In order to ensure that **U'** is smaller then $\epsilon/6$, we need T to satisfy

$$T \geq \frac{1}{\ln(1-\eta\underline{\mu}_{\lambda,f}/4)} \ln\left(\frac{\epsilon}{6\left(v_\star^f(\rho) - v_{\theta_0}^f(\rho)\right)}\right) \ ,$$

which combined with the inequality $\ln(1+x) \leq x$ for $x > -1$, ensures that $\mathbf{V'} < \epsilon/6$ and the condition (82) follows from (81).

Bounding W'. Using (84), it holds that

$$\mathbf{W'} \leq \frac{6}{\underline{\mu}_{\lambda,f}} \cdot \frac{36\gamma^{2H}(H+1)^2}{(1-\gamma)^4}\omega_f^2 \ . \ ,$$

Next, using that for any $a > 0$, we have $a\gamma^{2H}(H+1)^2 \leq \epsilon/6$ for

$$H \geq \frac{1}{1-\gamma} \max\left(\frac{4}{1-\gamma}, \ln\left(\frac{6a}{\epsilon}\right)\right) \ ,$$

shows that under our condition on H, we have $\mathbf{W'} < \epsilon/6$. □

F Application to common f-divergences

In this section, we apply the results of the preceding section to two commonly used f-divergences, which are Kullback-Leibler and α-Tsallis.

F.1 Kullback-Leibler

Lemma F.1. *Assume that, for some $\underline{\pi_{\mathrm{ref}}} > 0$, π_{ref} satisfy $\boldsymbol{P}(\underline{\pi_{\mathrm{ref}}})$. The function f defined by $f(u) = u\log(u)$ satisfies $\boldsymbol{A}_f(\underline{\pi_{\mathrm{ref}}})$, with*

$$\omega_f = 1 \ , \quad \kappa_f = 1 \ , \quad \iota_f = 1 \ .$$

Additionally, under the condition that

$$\lambda \leq \frac{4}{(1-\gamma)^2\rho_{\min}(\log(2/\underline{\pi_{\mathrm{ref}}}) + 1)} \ , \tag{85}$$

we have

$$\zeta_f = 1 \ , \quad \mathrm{d}_f \leq |\log(\underline{\pi_{\mathrm{ref}}})| \ , \quad \mathrm{y}_f \leq 1 + 2|\log(\underline{\pi_{\mathrm{ref}}})| \ , \quad L_{\lambda,f} = \frac{8}{(1-\gamma)^3} + 4\lambda\frac{3 + 4|\log(\underline{\pi_{\mathrm{ref}}})|}{(1-\gamma)^3} \ ,$$

$$\beta_{H,\lambda} = \frac{2\gamma^H(H+1)}{(1-\gamma)^2}\left[2 + \lambda + 4\lambda|\log(\underline{\pi_{\mathrm{ref}}})|\right] \ , \quad \sigma_{\lambda,f}^2 \leq \frac{12}{(1-\gamma)^4}\left[1 + \lambda^2\left(2 + 5|\log(\underline{\pi_{\mathrm{ref}}})|^2\right)\right] \ ,$$

$$\underline{\mu}_{\lambda,f} \geq \lambda(1-\gamma)\rho_{\min}^2\underline{\pi_{\mathrm{ref}}}^2 \exp\left(-\frac{32 + 16\gamma\lambda|\log(\underline{\pi_{\mathrm{ref}}})|}{\lambda(1-\gamma)^2\rho_{\min}}\right)/9 \ .$$

Proof. Firstly, note that we have

$$f(u) = u\log(u) \ , \quad f'(u) = \log(u) + 1 \ , \quad f''(u) = u^{-1} \ , \quad f'''(u) = -u^{-2} \ .$$

Satisfying $\mathbf{A}_f(\underline{\pi_{\text{ref}}})$. Observe that (i) and (ii) are immediately valid from the expression above of the derivatives of f. Moreover, we have

$$1/uf''(u) = 1 \ , \quad \frac{|f'''(u)|}{f''(u)^2} = 1$$

showing that (iii) of $\mathbf{A}_f(\underline{\pi_{\text{ref}}})$, is satisfied with $\omega_f = \kappa_f = 1$. Finally, as f'' is a strictly decreasing function on $\mathbb{R}_+$ then (iv) is valid with $\iota_f = 1$.

Bounding the constants. Next, we bound sequentially each of the constants that appear in the statement of the lemma. For any $s \in \mathcal{S}$ and $\nu \in \mathcal{P}(\mathcal{A})$, we have that

$$\sum_{a\in\mathcal{A}} \frac{\pi_{\text{ref}}(a|s)}{f''(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)})} = \sum_{a\in\mathcal{A}} \nu(a) = 1 \ ,$$

Thus using (13), we have that $\zeta_f = 1$. It holds that

$$\begin{aligned}
\mathrm{d}_f &= \max_{(s,\nu)\in\mathcal{S}\times\mathcal{P}(\mathcal{A})} \sum_{a\in\mathcal{A}} \pi_{\text{ref}}(a|s) f(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)}) \\
&= \max_{(s,\nu)\in\mathcal{S}\times\mathcal{P}(\mathcal{A})} \sum_{a\in\mathcal{A}} \nu(a) \log(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)}) \\
&\leq \max_{(s,\nu)\in\mathcal{S}\times\mathcal{P}(\mathcal{A})} \sum_{a\in\mathcal{A}} \nu(a)\log(\nu(a)) - \nu(a)\log(\underline{\pi_{\text{ref}}}) \ ,
\end{aligned}$$

which gives $\mathrm{d}_f \leq -\log(\underline{\pi_{\text{ref}}})$. Next, we have

$$\begin{aligned}
\mathrm{y}_f &= \max_{(s,\nu)\in\mathcal{S}\times\mathcal{P}(\mathcal{A})} \sum_{a\in\mathcal{A}} \frac{\pi_{\text{ref}}(a|s)}{f''\left(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)}\right)} \left| f'\left(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)}\right)\right| \\
&= \max_{(s,\nu)\in\mathcal{P}(\mathcal{A})} \sum_{a\in\mathcal{A}} \nu(a) \left|\log\left(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)}\right) + 1\right| \\
&= 1 - \log(\underline{\pi_{\text{ref}}}) + \max_{\nu\in\mathcal{P}(\mathcal{A})} -\sum_{a\in\mathcal{A}} \nu(a)\log(\nu(a)) \\
&\leq 1 + 2|\log(\underline{\pi_{\text{ref}}})| \ ,
\end{aligned}$$

where in the last inequality, we used that the entropy of a distribution on $\mathcal{A}$ is bounded by $\log(|\mathcal{A}|)$ and the fact that $\underline{\pi_{\text{ref}}} \leq 1/|\mathcal{A}|$. Next, using Theorem B.12 and the bounds above, we have

$$\begin{aligned}
L_{\lambda,f} &= \frac{8\omega_f\left(\gamma\omega_f + (1-\gamma)\kappa_f\right)}{(1-\gamma)^3} + 4\lambda\frac{2\gamma^2\omega_f^2\mathrm{d}_f + 2\gamma(1-\gamma)\omega_f\left[\kappa_f\mathrm{d}_f + \mathrm{y}_f\right] + (1-\gamma)^2\left[\omega_f + 2\kappa_f\mathrm{y}_f\right]}{(1-\gamma)^3} \\
&\leq \frac{8}{(1-\gamma)^3} + 4\lambda\frac{3 + 4|\log(\underline{\pi_{\text{ref}}})|}{(1-\gamma)^3} \ .
\end{aligned}$$

Using Lemma E.3 gives

$$\beta_{H,\lambda} = \frac{2\gamma^H(H+1)}{(1-\gamma)^2}\omega_f\left[2 + 2\lambda\mathrm{d}_f + \lambda(1-\gamma)\mathrm{y}_f\right] \leq \frac{2\gamma^H(H+1)}{(1-\gamma)^2}\left[2 + \lambda + 4\lambda|\log(\underline{\pi_{\text{ref}}})|\right] \ .$$

Using Lemma E.4 gives

$$\sigma^2_{\lambda,f} = \frac{12}{(1-\gamma)^4}\left[\omega_f^3 + \lambda^2\gamma^2\omega_f^3\mathrm{d}_f^2 + \lambda^2(1-\gamma)^2\omega_f^2\mathrm{y}_f^2\right] \leq \frac{12}{(1-\gamma)^4}\left[1 + \lambda^2(2 + 5|\log(\underline{\pi_{\text{ref}}})|^2))\right] \ .$$

Next, note that (85), guarantees that we have

$$\lambda \leq \frac{4}{(1-\gamma)^2\rho_{\min}} \min\left(\frac{4}{|f'(\iota_f)|}, \frac{1}{\left|f'(\frac{1}{2})\right|}, \frac{4}{\left|f'(\frac{1}{2}\underline{\pi_{\mathrm{ref}}})\right|}\right)$$

Thus, using Lemma E.5, we have

$$\begin{aligned}\underline{\mu}_{\lambda,f} &= \frac{\lambda(1-\gamma)\rho_{\min}^2\zeta_f^2}{\omega_f^2}\underline{\pi_{\mathrm{ref}}}^2(f^\star)''\left(-\frac{16+8\gamma\lambda \mathrm{d}_f}{\lambda(1-\gamma)^2\rho_{\min}}\right)^2 \\ &\geq \lambda(1-\gamma)\rho_{\min}^2\underline{\pi_{\mathrm{ref}}}^2\exp\left(-\frac{32+16\gamma\lambda|\log(\underline{\pi_{\mathrm{ref}}})|}{\lambda(1-\gamma)^2\rho_{\min}}\right)/9 \ ,\end{aligned}$$

where in the last equality, we used that the convex conjugate of $f(u) = u\log(u)$ is $f^\star(y) = \exp(y-1)$ and that $\exp(-2) \geq 1/9$. □

In the next two corollaries, we apply Corollary E.7 and Corollary E.10 to get more explicitly the sample complexity of f-PG with entropy regularization.

Corollary F.2. *Assume that, for some $1/4 \geq \underline{\pi_{\mathrm{ref}}} > 0$, π_{ref} satisfy $\boldsymbol{P}(\underline{\pi_{\mathrm{ref}}})$. Assume in addition that the initial distribution ρ satisfies $\boldsymbol{A}_\rho$. Fix any $(1-\gamma)^{-1} \geq \epsilon > 0$, $\lambda > 0$ and B such that*

$$\lambda \leq \min\left(\frac{4}{(1-\gamma)^2\rho_{\min}(\log(2/\underline{\pi_{\mathrm{ref}}})+1)}, 1\right) \ , \quad B \leq \frac{216|\log(\underline{\pi_{\mathrm{ref}}})|}{\epsilon\lambda(1-\gamma)^2\underline{\pi_{\mathrm{ref}}}^2\rho_{\min}^2}\exp\left(\frac{48|\log(\underline{\pi_{\mathrm{ref}}})|}{\lambda(1-\gamma)^2\rho_{\min}}\right) \ .$$

Setting

$$H \geq \frac{1}{1-\gamma}\log\left(\frac{29160\log(\underline{\pi_{\mathrm{ref}}})^2}{\epsilon\lambda(1-\gamma)^5\rho_{\min}^2\underline{\pi_{\mathrm{ref}}}^2}\right) + \frac{52|\log(\underline{\pi_{\mathrm{ref}}})|}{\lambda(1-\gamma)^3\rho_{\min}}, \tag{86}$$

and

$$\eta \leq \frac{\epsilon B\lambda(1-\gamma)^5\rho_{\min}^2\underline{\pi_{\mathrm{ref}}}^2}{15552|\log(\underline{\pi_{\mathrm{ref}}})|^2}\exp\left(-\frac{48|\log(\underline{\pi_{\mathrm{ref}}})|}{\lambda(1-\gamma)^2\rho_{\min}}\right) \ ,$$

and

$$T \geq \frac{559872|\log(\underline{\pi_{\mathrm{ref}}})|^2}{\lambda^2\epsilon B(1-\gamma)^6\rho_{\min}^4\underline{\pi_{\mathrm{ref}}}^4}\exp\left(\frac{96|\log(\underline{\pi_{\mathrm{ref}}})|}{\lambda(1-\gamma)^2\rho_{\min}}\right) \ ,$$

guarantees that

$$v_\star^f(\rho) - \mathbb{E}\left[v_{\theta_t}^f(\rho)\right] \leq \epsilon \ ,$$

where f is the Kullback-Leibler divergence generator. Thus, the sample complexity of f-PG to learn an ϵ-solution of the entropy regularized problem is

$$TBH \approx \frac{1}{\lambda^3\epsilon B}\frac{|\log(\underline{\pi_{\mathrm{ref}}})|^3}{(1-\gamma)^9\rho_{\min}^5\underline{\pi_{\mathrm{ref}}}^4}\exp\left(\frac{|\log(\underline{\pi_{\mathrm{ref}}})|}{\lambda(1-\gamma)^2\rho_{\min}}\right)$$

Proof. To prove this corollary, we show that the assumptions of Corollary E.7 hold. First, $\mathbf{A}_f(\underline{\pi_{\mathrm{ref}}})$ holds as a consequence of Lemma F.1. Then, we show that the condition (75) in Corollary E.7 holds, that is $H \leq \frac{4}{(1-\gamma)^2} + \frac{1}{1-\gamma}\log\left(\frac{216\omega_f^2}{\epsilon\underline{\mu}_{\lambda,f}(1-\gamma)^4}\left[4+4\lambda^2\mathrm{d}_f^2+\lambda^2(1-\gamma)^2\mathrm{y}_f^2\right]\right)$. To this end, we remark that

$$\frac{4}{(1-\gamma)^2} + \frac{1}{1-\gamma}\log\left(\frac{216\omega_f^2}{\epsilon\underline{\mu}_{\lambda,f}(1-\gamma)^4}\left[4+4\lambda^2\mathrm{d}_f^2+\lambda^2(1-\gamma)^2\mathrm{y}_f^2\right]\right)$$

$$\leq \frac{4}{(1-\gamma)^2} + \frac{1}{1-\gamma}\log\left(\frac{216}{\epsilon \underline{\mu}_{\lambda,f}(1-\gamma)^4}\left[4+\lambda^2(5+6\log(\underline{\pi}_{\text{ref}})^2)\right]\right)$$
$$\leq \frac{4}{(1-\gamma)^2} + \frac{1}{1-\gamma}\log\left(\frac{1944\left[4+\lambda^2(5+6\log(\underline{\pi}_{\text{ref}})^2)\right]}{\epsilon\lambda(1-\gamma)^5\rho_{\min}^2\underline{\pi}_{\text{ref}}^2}\right) + \frac{32+16\gamma\lambda|\log(\underline{\pi}_{\text{ref}})|}{\lambda(1-\gamma)^3\rho_{\min}}$$
$$\leq \frac{1}{1-\gamma}\log\left(\frac{29160\log(\underline{\pi}_{\text{ref}})^2}{\epsilon\lambda(1-\gamma)^5\rho_{\min}^2\underline{\pi}_{\text{ref}}^2}\right) + \frac{52|\log(\underline{\pi}_{\text{ref}})|}{\lambda(1-\gamma)^3\rho_{\min}} \leq H \ ,$$

where we used the lower bound on $\underline{\mu}_{\lambda,f}$ provided in Lemma F.1 in the second inequality, as well as $\lambda < 1$ and $\underline{\pi}_{\text{ref}} \leq 1/4$ in the last two inequalities. Furthermore, Lemma F.1 with $\lambda \leq 1$ gives

$$L_{\lambda,f} \leq \frac{36|\log(\underline{\pi}_{\text{ref}})|}{(1-\gamma)^3} \ , \tag{87}$$

$$\sigma^2_{\lambda,f} \leq \frac{96|\log(\underline{\pi}_{\text{ref}})|^2}{(1-\gamma)^4} \ , \tag{88}$$

$$\underline{\mu}_{\lambda,f} \geq \frac{\lambda(1-\gamma)\rho_{\min}^2\underline{\pi}_{\text{ref}}^2}{9}\exp\left(-\frac{48|\log(\underline{\pi}_{\text{ref}})|}{\lambda(1-\gamma)^2\rho_{\min}}\right) \ . \tag{89}$$

Using these three bounds on smoothness, variance and Polyak-Łojasiewicz coefficients, we obtain

$$\min\left(\frac{1}{2L_{\lambda,f}}, \frac{\epsilon B\underline{\mu}_{\lambda,f}}{18\sigma^2_{\lambda,f}}\right) \geq \min\left(\frac{(1-\gamma)^3}{72|\log(\underline{\pi}_{\text{ref}})|}, \frac{\epsilon B\lambda(1-\gamma)^5\rho_{\min}^2\underline{\pi}_{\text{ref}}^2\exp\left(-\frac{48|\log(\underline{\pi}_{\text{ref}})|}{\lambda(1-\gamma)^2\rho_{\min}}\right)}{15552|\log(\underline{\pi}_{\text{ref}})|^2}\right)$$
$$= \frac{\epsilon B\lambda(1-\gamma)^5\rho_{\min}^2\underline{\pi}_{\text{ref}}^2}{15552|\log(\underline{\pi}_{\text{ref}})|^2}\exp\left(-\frac{48|\log(\underline{\pi}_{\text{ref}})|}{\lambda(1-\gamma)^2\rho_{\min}}\right) \ , \tag{90}$$

where in the last identity, we used the fact that $\epsilon < (1-\gamma)^{-1}$ and that

$$B \leq \frac{216|\log(\underline{\pi}_{\text{ref}})|}{\epsilon\lambda(1-\gamma)^2\underline{\pi}_{\text{ref}}^2\rho_{\min}^2}\exp\left(\frac{48|\log(\underline{\pi}_{\text{ref}})|}{\lambda(1-\gamma)^2\rho_{\min}}\right) \ .$$

This shows that our condition on η guarantees that the one set in Corollary E.7 is satisfied. Finally using (90) and (89), we have

$$\frac{4}{\underline{\mu}_{\lambda,f}}\max\left(2L_{\lambda,f}, \frac{18\sigma^2_{\lambda,f}}{\epsilon B\underline{\mu}_{\lambda,f}}\right) \leq \frac{4}{\underline{\mu}_{\lambda,f}}\frac{15552|\log(\underline{\pi}_{\text{ref}})|^2}{\epsilon B\lambda(1-\gamma)^5\rho_{\min}^2\underline{\pi}_{\text{ref}}^2}\exp\left(\frac{48|\log(\underline{\pi}_{\text{ref}})|}{\lambda(1-\gamma)^2\rho_{\min}}\right)$$
$$\leq \frac{559872|\log(\underline{\pi}_{\text{ref}})|^2}{\lambda^2\epsilon B(1-\gamma)^6\rho_{\min}^4\underline{\pi}_{\text{ref}}^4}\exp\left(\frac{96|\log(\underline{\pi}_{\text{ref}})|}{\lambda(1-\gamma)^2\rho_{\min}}\right) \ ,$$

which concludes the proof. □

Corollary F.3. *Assume that, for some $1/4 \geq \underline{\pi}_{\text{ref}} > 0$, π_{ref} satisfy $\boldsymbol{P}(\underline{\pi}_{\text{ref}})$. Assume in addition that the initial distribution ρ satisfies $\boldsymbol{A}_\rho$ and fix f to be the Kullback-Leibler divergence generator, i.e. $f(u) = u\log(u)$. Fix any $(1-\gamma)^{-1} \geq \epsilon > 0$, such that*

$$\epsilon < \frac{16}{(1-\gamma)^3\rho_{\min}(\log(2/\underline{\pi}_{\text{ref}})+1)} \ , \quad \textit{and set } \lambda = \frac{(1-\gamma)\epsilon}{12\log(|\underline{\pi}_{\text{ref}}|)} \ . \tag{91}$$

Additionally set any B such that

$$B \leq \frac{1}{\epsilon^2(1-\gamma)^3\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\exp\left(\frac{576|\log(\underline{\pi_{\text{ref}}})|}{\epsilon(1-\gamma)^3\rho_{\min}}\right) \quad . \tag{92}$$

Setting

$$H \geq \frac{1}{1-\gamma}\log\left(\frac{|\log(\underline{\pi_{\text{ref}}})|}{\epsilon(1-\gamma)^6\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\right) + \frac{586|\log(\underline{\pi_{\text{ref}}})|}{\epsilon(1-\gamma)^4\rho_{\min}} \quad ,$$

and

$$\eta \leq \frac{\epsilon^2(1-\gamma)^6 B\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}{93312|\log(\underline{\pi_{\text{ref}}})|}\exp\left(\frac{-576|\log(\underline{\pi_{\text{ref}}})|}{\epsilon(1-\gamma)^3\rho_{\min}}\right) \quad , \tag{93}$$

and

$$T \geq \frac{644972544|\log(\underline{\pi_{\text{ref}}})|^2}{\epsilon^3(1-\gamma)^8\rho_{\min}^4\underline{\pi_{\text{ref}}}^4 B}\exp\left(\frac{1152|\log(\underline{\pi_{\text{ref}}})|}{\epsilon(1-\gamma)^3\rho_{\min}}\right)\log\left(\frac{6(v_\star^f(\rho)-v_{\theta_0}^f(\rho))}{\epsilon}\right) \quad , \tag{94}$$

guarantees that

$$v_\star(\rho) - \mathbb{E}\left[v_{\theta_t}(\rho)\right] \leq \epsilon \quad .$$

Thus, the sample complexity of f−PG, where f is the Kullback-Leibler divergence generator, to learn an ϵ-solution of the non-regularized problem is

$$TBH \approx \frac{|\log(\underline{\pi_{\text{ref}}})|^3}{\epsilon^4(1-\gamma)^{12}\rho_{\min}^5\underline{\pi_{\text{ref}}}^4}\exp\left(\frac{|\log(\underline{\pi_{\text{ref}}})|}{\epsilon(1-\gamma)^3\rho_{\min}}\right)$$

Proof. This result follows from Corollary E.10, whose assumptions we check now. First, note that (91) implies that

$$\epsilon < \frac{16}{(1-\gamma)^3\rho_{\min}}\min\left(\frac{4}{|f'(\iota_f)|}, \frac{1}{\left|f'(\frac{1}{2})\right|}, \frac{4}{\left|f'(\frac{1}{2}\underline{\pi_{\text{ref}}})\right|}\right) \quad .$$

Additionally, we can rewrite the constants from Corollary E.10 using Lemma F.1, which gives

$$C_f^{(1)} \leq 48|\log(\underline{\pi_{\text{ref}}})|, \;\; C_f^{(2)} \leq 48, \;\; C_f^{(3)} = 10368|\log(\underline{\pi_{\text{ref}}})|, \;\; d(\epsilon) \geq \exp\left(\frac{-576|\log(\underline{\pi_{\text{ref}}})|}{\epsilon(1-\gamma)^3\rho_{\min}}\right)/9, \tag{95}$$

where $C_f^{(1)}, C_f^{(2)}, C_f^{(3)}$, and $d(\epsilon)$ are defined in (80) and (79). Next, the condition on H in Corollary E.10 holds since

$$\begin{aligned}
&\frac{4}{(1-\gamma)^2} + \frac{1}{1-\gamma}\log\left(\frac{297\omega_f^2 C_1^f}{\epsilon d(\epsilon)(1-\gamma)^6\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\right) \\
&\leq \frac{4}{(1-\gamma)^2} + \frac{1}{1-\gamma}\log\left(\frac{128304|\log(\underline{\pi_{\text{ref}}})|}{\epsilon(1-\gamma)^6\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\right) + \frac{576|\log(\underline{\pi_{\text{ref}}})|}{\epsilon(1-\gamma)^4\rho_{\min}} \\
&\leq \frac{1}{1-\gamma}\log\left(\frac{|\log(\underline{\pi_{\text{ref}}})|}{\epsilon(1-\gamma)^6\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\right) + \frac{586|\log(\underline{\pi_{\text{ref}}})|}{\epsilon(1-\gamma)^4\rho_{\min}} \leq H \quad ,
\end{aligned}$$

where in the second to last inequality, we used that $\pi_{\text{ref}} < 1/4$ and that $\log(128304) \leq 6$. Using (95), we have

$$\begin{aligned}
&\min\left(\frac{(1-\gamma)^3}{C_f^{(2)}}, \frac{\epsilon^2 d(\epsilon)(1-\gamma)^6 B \rho_{\min}^2 \underline{\pi_{\text{ref}}}^2}{C_f^{(3)}}\right) \\
&\geq \min\left(\frac{(1-\gamma)^3}{48}, \frac{\epsilon^2(1-\gamma)^6 B \rho_{\min}^2 \underline{\pi_{\text{ref}}}^2}{93312|\log(\underline{\pi_{\text{ref}}})|}\exp\left(\frac{-576|\log(\underline{\pi_{\text{ref}}})|}{\epsilon(1-\gamma)^3\rho_{\min}}\right)\right) \\
&\geq \frac{\epsilon^2(1-\gamma)^6 B \rho_{\min}^2 \underline{\pi_{\text{ref}}}^2}{93312|\log(\underline{\pi_{\text{ref}}})|}\exp\left(\frac{-576|\log(\underline{\pi_{\text{ref}}})|}{\epsilon(1-\gamma)^3\rho_{\min}}\right) \quad , \qquad (96)
\end{aligned}$$

where in the last inequality, we used the condition on B introduced in (92). Hence our condition on the step size ensures that the one assumed in Corollary E.10 is satisfied. Next, using (95) and (96) yields

$$\begin{aligned}
&\frac{16 C_f^{(1)}\ell(\epsilon)}{\epsilon d(\epsilon)(1-\gamma)^2\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\max\left(\frac{C_f^{(2)}}{(1-\gamma)^3}, \frac{C_f^{(3)}}{\epsilon^2 d(\epsilon)(1-\gamma)^6 B\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\right) \\
&\leq \frac{16 C_f^{(1)}\ell(\epsilon)}{\epsilon d(\epsilon)(1-\gamma)^2\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\cdot\frac{93312|\log(\underline{\pi_{\text{ref}}})|}{\epsilon^2(1-\gamma)^6 B\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\exp\left(\frac{576|\log(\underline{\pi_{\text{ref}}})|}{\epsilon(1-\gamma)^3\rho_{\min}}\right) \\
&\leq \frac{6912|\log(\underline{\pi_{\text{ref}}})|\ell(\epsilon)}{\epsilon(1-\gamma)^2\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\cdot\frac{93312|\log(\underline{\pi_{\text{ref}}})|}{\epsilon^2(1-\gamma)^6 B\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\exp\left(\frac{1152|\log(\underline{\pi_{\text{ref}}})|}{\epsilon(1-\gamma)^3\rho_{\min}}\right) \\
&\leq \frac{644972544|\log(\underline{\pi_{\text{ref}}})|^2}{\epsilon^3(1-\gamma)^8\rho_{\min}^4\underline{\pi_{\text{ref}}}^4 B}\exp\left(\frac{1152|\log(\underline{\pi_{\text{ref}}})|}{\epsilon(1-\gamma)^3\rho_{\min}}\right)\log\left(\frac{6(v_\star^f(\rho)-v_{\theta_0}^f(\rho))}{\epsilon}\right) \quad ,
\end{aligned}$$

which proves that under our condition on T the one assumed by Corollary E.10 is satisfied. □

F.2 α-TSALLIS

Lemma F.4. *Assume that, for some $1/4 > \underline{\pi_{\text{ref}}} > 0$, π_{ref} satisfy $\boldsymbol{P}(\underline{\pi_{\text{ref}}})$. For any $\alpha \in (0;1)$, the function f_α defined by*

$$f_\alpha(u) = \frac{u^\alpha - \alpha u + \alpha - 1}{\alpha(\alpha-1)} \quad ,$$

satisfies $\boldsymbol{A}_f(\underline{\pi_{\text{ref}}})$, with

$$\omega_f = \underline{\pi_{\text{ref}}}^{\alpha-1} \quad , \quad \kappa_f = 2\underline{\pi_{\text{ref}}}^{\alpha-1} \quad , \quad \iota_f = 1 \quad .$$

Additionally, under the condition that

$$\lambda \leq \frac{4}{(1-\gamma)^2\rho_{\min}}\cdot\frac{1-\alpha}{(\pi_{\text{ref}}/2)^{\alpha-1}-1} \quad , \qquad (97)$$

we have

$$\zeta_f = 1 \ , \ \mathrm{d}_f \leq \frac{4|\log(\underline{\pi_{\text{ref}}})|}{\alpha^2} \ , \ \mathrm{y}_f \leq 4|\log(\underline{\pi_{\text{ref}}})| \ , \ L_{\lambda,f} = \frac{16\underline{\pi_{\text{ref}}}^{\alpha-1}}{(1-\gamma)^3} + 180\lambda\frac{\underline{\pi_{\text{ref}}}^{2\alpha-2}|\log(\underline{\pi_{\text{ref}}})|}{\alpha^2(1-\gamma)^3} \ ,$$

$$\beta_{H,\lambda} = \frac{4\gamma^H(H+1)}{(1-\gamma)^2}\left[1+6\lambda\frac{|\log(\underline{\pi_{\text{ref}}})|}{\alpha^2}\right] \quad , \quad \sigma_{\lambda,f}^2 \leq \frac{12\underline{\pi_{\text{ref}}}^{3\alpha-3}}{(1-\gamma)^4}\left[1+\frac{16\lambda^2}{\alpha^4}\log(\underline{\pi_{\text{ref}}}|^2)\right] \ ,$$

$$\underline{\mu}_{\lambda,f} \geq \lambda(1-\gamma)\underline{\pi_{\text{ref}}}^{2-2\alpha}\rho_{\min}^2\underline{\pi_{\text{ref}}}^2\exp_\alpha\left(-\frac{16+32\gamma\lambda|\log(\underline{\pi_{\text{ref}}})|/\alpha^2}{\lambda(1-\gamma)^2\rho_{\min}}\right)^{4-2\alpha} \quad .$$

Proof. Fix any $\alpha \in (0,1)$ and set $f = f_\alpha$. Firstly, note that we have

$$f(u) = \frac{u^\alpha - \alpha u + \alpha - 1}{\alpha(\alpha-1)} \ , \quad f'(u) = \frac{u^{\alpha-1}-1}{\alpha-1} \ , \quad f''(u) = u^{\alpha-2} \ , \quad f'''(u) = (\alpha-2)u^{\alpha-3} \ .$$

Satisfying $\mathbf{A}_f(\underline{\pi_{\text{ref}}})$. Observe that (i) and (ii) are immediately valid from the expression above of the derivatives of f. Moreover, we have

$$1/uf''_\alpha(u) = u^{1-\alpha} \ , \quad \frac{|f'''(u)|}{f''(u)^2} = |\alpha-2|u^{1-\alpha} \ ,$$

showing that (iii) of $\mathbf{A}_f(\underline{\pi_{\text{ref}}})$, is satisfied with $\omega_f = \underline{\pi_{\text{ref}}}^{\alpha-1}$, and $\kappa_f = 2\underline{\pi_{\text{ref}}}^{\alpha-1}$. Finally,as f'' is a strictly decreasing function on $\mathbb{R}_+$ then (iv) is valid with $\iota_f = 1$. Next, we bound sequentially each of the constants that appear in the statement of the lemma.

Bounding ζ_f. For any $s \in \mathcal{S}$ and $\nu \in \mathcal{P}(\mathcal{A})$, using Jensen's inequality we have that

$$\sum_{a\in\mathcal{A}} \frac{\pi_{\text{ref}}(a|s)}{f''(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)})} = \sum_{a\in\mathcal{A}} \pi_{\text{ref}}(a|s)\left(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)}\right)^{2-\alpha} \geq \left(\sum_{a\in\mathcal{A}} \pi_{\text{ref}}(a|s)\frac{\nu(a)}{\pi_{\text{ref}}(a|s)}\right)^{2-\alpha} = 1 \ ,$$

Thus using (13), we have that $\zeta_f = 1$.

Bounding d_f. For any state $s \in \mathcal{S}$ and $\nu \in \mathcal{P}(\mathcal{A})$, it holds that

$$\mathrm{D}^f(\nu\|\pi_{\text{ref}}(\cdot|s)) = \frac{1}{\alpha(\alpha-1)}\sum_{a\in\mathcal{A}} \pi_{\text{ref}}(a|s)\left[\left(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)}\right)^\alpha - \alpha\frac{\nu(a)}{\pi_{\text{ref}}(a|s)} - (1-\alpha)\right] \ .$$

Next, for $y \in]0; \frac{1}{\underline{\pi_{\text{ref}}}}]$, define the function $p_y \colon [\alpha; 1] \to \mathbb{R}$ which satisfies

$$p_y(\beta) = y^\beta - \beta y \ , \quad p'_y(\beta) = \log(y)y^\beta - y \ .$$

It holds that

$$\frac{y^\alpha - \alpha y - (1-\alpha)}{\alpha-1} = \frac{p_y(\alpha) - p_y(1)}{\alpha-1} \leq \sup_{\beta\in[\alpha,1]} |p'_y(\beta)| \leq y + y|\log(y)|1_{y\geq 1} + |\log(y)|y^\alpha 1_{y\leq 1} \ .$$

Applying the previous inequality with $y = \nu(a)/\pi_{\text{ref}}(a|s)$ yields

$$\begin{aligned}
\mathrm{D}^f(\nu\|\pi_{\text{ref}}(\cdot|s)) &\leq \frac{1}{\alpha}\sum_{a\in\mathcal{A}} \pi_{\text{ref}}(a|s)\Bigg[\frac{\nu(a)}{\pi_{\text{ref}}(a|s)} + \frac{\nu(a)}{\pi_{\text{ref}}(a|s)}|\log(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)})|1_{\nu(a)/\pi_{\text{ref}}(a|s)\geq 1} \\
&\qquad + |\log(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)})|\left(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)}\right)^\alpha 1_{\nu(a)/\pi_{\text{ref}}(a|s)\leq 1}\Bigg] \\
&\leq \frac{1}{\alpha} + \frac{1}{\alpha}\sum_{a\in\mathcal{A}} \nu(a)|\log(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)})| + \frac{1}{\alpha}\max_{a\in\mathcal{A}}|\log(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)})|\left(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)}\right)^\alpha 1_{\nu(a)/\pi_{\text{ref}}(a|s)\leq 1} \\
&\leq \frac{1}{\alpha} + \frac{1}{\alpha}\sum_{a\in\mathcal{A}} \nu(a)|\log(\nu(a))| + \frac{1}{\alpha}\sum_{a\in\mathcal{A}} \nu(a)|\log(\pi_{\text{ref}}(a|s))| + \frac{1}{\alpha}\max_{x\in[0;1]}|\log(x)|x^\alpha \ , \\
&\leq \frac{1}{\alpha} + \frac{1}{\alpha}\log(|\mathcal{A}|) + \frac{1}{\alpha}|\log(\underline{\pi_{\text{ref}}})| + \frac{1}{\alpha}\max_{x\in[0;1]}|\log(x)|x^\alpha \ ,
\end{aligned}$$

where in the last inequality, we used that the entropy of a probability measure on $\mathcal{A}$ is bounded by $|\log(\underline{\pi_{\text{ref}}})|$. Next using that $\max_{x\in[0;1]} |\log(x)|x^\alpha \leq e^{-1}/\alpha$ combined with $\max(1, \log(|\mathcal{A}|)) \leq |\log(\underline{\pi_{\text{ref}}})|$ gives

$$\mathrm{D}^f(\nu\|\pi_{\text{ref}}(\cdot|s)) \leq \frac{3|\log(\underline{\pi_{\text{ref}}})|}{\alpha} + \frac{1}{\alpha^2} \leq \frac{4|\log(\underline{\pi_{\text{ref}}})|}{\alpha^2} \ .$$

Thus, it holds that

$$\mathrm{d}_f \leq \frac{4|\log(\underline{\pi_{\text{ref}}})|}{\alpha^2} \ .$$

Bounding y_f. For any policy $\nu \in \mathcal{P}(\mathcal{A})$ and $s \in \mathcal{S}$, we have

$$\sum_{a\in\mathcal{A}} \frac{\pi_{\text{ref}}(a|s)}{f''(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)})} \left|f'(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)})\right| = \left|\frac{1}{1-\alpha}\right| \sum_{a\in\mathcal{A}} \frac{\nu(a)^{2-\alpha}}{\pi_{\text{ref}}(a|s)^{1-\alpha}} \left|\left(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)}\right)^{\alpha-1} - 1\right| \ .$$

Next, define the function g_y for $y \in]0; \frac{1}{\underline{\pi_{\text{ref}}}}]$ which satisfies

$$g_y(\beta) = y^{\beta-1} \ , \quad g_y'(\beta) = \log(y)y^{\beta-1} \ .$$

It holds that

$$\left|\frac{y^{\alpha-1}-1}{1-\alpha}\right| = \left|\frac{g_y(\alpha)-g_y(1)}{\alpha-1}\right| \leq \sup_{\beta\in[\alpha,1]} \left|g_y'(\beta)\right| \leq |\log(y)|\,1_{y\leq 1} + \left|\log(y)y^{\alpha-1}\right| 1_{y\geq 1} \ ,$$

Hence, applying the previous inequality with $y = \nu(a)/\pi_{\text{ref}}(a|s)$ gives

$$\begin{aligned}
&\sum_{a\in\mathcal{A}} \frac{\pi_{\text{ref}}(a|s)}{f''(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)})} \left|f'(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)})\right| \\
&\quad \leq \sum_{a\in\mathcal{A}} \frac{\nu(a)^{2-\alpha}}{\pi_{\text{ref}}(a|s)^{1-\alpha}} \left[|\log(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)})|1_{\nu(a)/\pi_{\text{ref}}(a|s)\leq 1} + |\log(\frac{\pi(a|s)}{\pi_{\text{ref}}(a|s)})| \left(\frac{\pi_{\text{ref}}(a|s)}{\nu(a)}\right)^{1-\alpha}\right] \\
&\quad \leq \sum_{a\in\mathcal{A}} \left(\frac{\nu(a)}{\pi_{\text{ref}}(a)}\right)^{1-\alpha} \nu(a)|\log(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)})|1_{\nu(a)/\pi_{\text{ref}}(a|s)\leq 1} + \sum_{a\in\mathcal{A}} \nu(a)|\log(\frac{\pi(a|s)}{\pi_{\text{ref}}(a|s)})| \\
&\quad \leq \sum_{a\in\mathcal{A}} \nu(a)|\log(\frac{\nu(a)}{\pi_{\text{ref}}(a|s)})| + 2|\log(\underline{\pi_{\text{ref}}})| \ ,
\end{aligned}$$

where in the last inequality, we used for the first term that for any $u \in \mathbb{R}, u1_{u\leq 1} \leq 1$ that the entropy of a probability distribution on $\mathcal{A}$ is bounded by $\log(|\mathcal{A}|)$, the fact that $\underline{\pi_{\text{ref}}} \leq 1/|\mathcal{A}|$. Using the same argument again to bound the first term gives

$$\mathrm{y}_f \leq 4|\log(\underline{\pi_{\text{ref}}})| \ .$$

.

Bounding $L_{\lambda,f}$. Next, using Theorem B.12 and the bound on the constants previously computed, we have

$$\begin{aligned}
L_{\lambda,f} &= \frac{8\omega_f\left(\gamma\omega_f + (1-\gamma)\kappa_f\right)}{(1-\gamma)^3} + 4\lambda\frac{2\gamma^2\omega_f^2\mathrm{d}_f + 2\gamma(1-\gamma)\omega_f\left[\kappa_f\mathrm{d}_f + \mathrm{y}_f\right] + (1-\gamma)^2\left[\omega_f + 2\kappa_f\mathrm{y}_f\right]}{(1-\gamma)^3} \\
&\leq \frac{16\underline{\pi_{\text{ref}}}^{\alpha-1}}{(1-\gamma)^3} + 180\lambda\frac{\underline{\pi_{\text{ref}}}^{2\alpha-2}|\log(\underline{\pi_{\text{ref}}})|}{\alpha^2(1-\gamma)^3} \ ,
\end{aligned}$$

where in the last inequality, we used that $\underline{\pi_{\text{ref}}} < 1/4$.

Bounding $\beta_{H,\lambda}$. Using Lemma E.3 gives

$$\beta_{H,\lambda} = \frac{2\gamma^H(H+1)}{(1-\gamma)^2}\omega_f\left[2 + 2\lambda \mathrm{d}_f + \lambda(1-\gamma)\mathrm{y}_f\right] \leq \frac{4\gamma^H(H+1)}{(1-\gamma)^2}\left[1 + 6\lambda\frac{|\log(\underline{\pi_{\mathrm{ref}}})|}{\alpha^2}\right] .$$

Bounding $\sigma_{\lambda,f}$. Using Lemma E.4 gives

$$\sigma^2_{\lambda,f} = \frac{12}{(1-\gamma)^4}\left[\omega_f^3 + \lambda^2\gamma^2\omega_f^3\mathrm{d}_f^2 + \lambda^2(1-\gamma)^2\omega_f^2\mathrm{y}_f^2\right] \leq \frac{12{\underline{\pi_{\mathrm{ref}}}}^{3\alpha-3}}{(1-\gamma)^4}\left[1 + \frac{16\lambda^2}{\alpha^4}\log({\underline{\pi_{\mathrm{ref}}}}^2)\right] .$$

Bounding $\underline{\mu}_{\lambda,f}$. Next, note that as f'_α is an increasing function then $f'_\alpha(\underline{\pi_{\mathrm{ref}}}/2) \leq f'_\alpha(1/2) \leq f'_\alpha(\iota_f) = f'_\alpha(1) = 0$. Thus, we have $|f'_\alpha(\iota_f)| \leq |f'_\alpha(1/2)| \leq |f'_\alpha(\underline{\pi_{\mathrm{ref}}}/2)|$. This proves that (97), guarantees that

$$\lambda \leq \frac{4}{(1-\gamma)^2\rho_{\min}}\min\left(\frac{4}{|f'(\iota_f)|}, \frac{1}{\left|f'(\frac{1}{2})\right|}, \frac{4}{\left|f'(\frac{1}{2}\underline{\pi_{\mathrm{ref}}})\right|}\right)$$

Thus, using Lemma E.5, we have

$$\underline{\mu}_{\lambda,f} = \frac{\lambda(1-\gamma)\rho^2_{\min}\zeta_f^2}{\omega_f^2}{\underline{\pi_{\mathrm{ref}}}}^2(f^\star)''\left(-\frac{16 + 8\gamma\lambda\mathrm{d}_f}{\lambda(1-\gamma)^2\rho_{\min}}\right)^2 . \tag{98}$$

Next, recall from proposition 8 of Roulet et al. (2025) that

$$f^*_\alpha(x) = \frac{(1+(\alpha-1)x)^{\frac{\alpha}{\alpha-1}} - 1}{\alpha} , \quad \text{for } x \leq \frac{1}{1-\alpha} .$$

Thus, we have

$$(f^\star_\alpha)'(x) = \left(1+(\alpha-1)x\right)^{\frac{1}{\alpha-1}} = \exp_\alpha(x) ,$$

where we have originally defined $\exp_\alpha$ in Section 3. Finally, it holds that

$$(f^\star_\alpha)''(x) = \exp_\alpha(x)^{2-\alpha} .$$

Thus,

$$\underline{\mu}_{\lambda,f} \geq \lambda(1-\gamma){\underline{\pi_{\mathrm{ref}}}}^{2-2\alpha}\rho^2_{\min}{\underline{\pi_{\mathrm{ref}}}}^2\exp_\alpha\left(-\frac{16 + 32\gamma\lambda|\log(\underline{\pi_{\mathrm{ref}}})|/\alpha^2}{\lambda(1-\gamma)^2\rho_{\min}}\right)^{4-2\alpha} ,$$

□

In the next two corollaries, we apply Corollary E.7 and Corollary E.10 to get more explicitly the sample complexity of f-PG with entropy regularization.

Corollary F.5. *Assume that, for some $1/4 \geq \underline{\pi_{\mathrm{ref}}} > 0$, π_{ref} satisfy $\boldsymbol{P}(\underline{\pi_{\mathrm{ref}}})$. Assume in addition that the initial distribution ρ satisfies $\boldsymbol{A}_\rho$. Fix any $(1-\gamma)^{-1} \geq \epsilon > 0$, $\alpha \in (0,1)$, $\lambda > 0$, and B such that*

$$\lambda \leq \min\left(\frac{4}{(1-\gamma)^2\rho_{\min}}\cdot\frac{1-\alpha}{(\underline{\pi_{\mathrm{ref}}}/2)^{\alpha-1}-1}, 1\right), \quad B \leq \frac{1}{{\underline{\pi_{\mathrm{ref}}}}^2\rho^2_{\min}}\exp_\alpha\left(-\frac{48|\log(\underline{\pi_{\mathrm{ref}}})|\max(\lambda,\alpha^2)}{\lambda\alpha^2(1-\gamma)^2\rho_{\min}}\right)^{2\alpha-4} .$$

Setting

$$H \geq \frac{1}{1-\gamma}\log\left(\frac{28152{\underline{\pi_{\mathrm{ref}}}}^{4\alpha-4}|\log(\underline{\pi_{\mathrm{ref}}})|^2}{\epsilon\lambda\alpha^4(1-\gamma)^5\rho^2_{\min}{\underline{\pi_{\mathrm{ref}}}}^2}\max(\alpha^2,\lambda)^2\right) + \frac{196|\log(\underline{\pi_{\mathrm{ref}}})|}{\lambda\alpha^2(1-\gamma)^3\rho_{\min}}\max(\alpha^2,\lambda) \tag{99}$$

and

$$\eta \leq \frac{\epsilon B \lambda (1-\gamma)^5 \alpha^4 \underline{\pi_{\text{ref}}}^{5-5\alpha} \rho_{\min}^2 \underline{\pi_{\text{ref}}}^2}{3672 \log(\underline{\pi_{\text{ref}}})^2 \max(\alpha^2, \lambda)^2} \exp_\alpha\left(-\frac{48|\log(\underline{\pi_{\text{ref}}})|}{\lambda \alpha^2 (1-\gamma)^2 \rho_{\min}} \max(\lambda, \alpha^2)\right)^{4-2\alpha} ,$$

and

$$T \geq \frac{14688 \log(\underline{\pi_{\text{ref}}})^2 \max(\alpha^2, \lambda)^2}{\lambda^2 \epsilon B (1-\gamma)^6 \alpha^4 \underline{\pi_{\text{ref}}}^{7-7\alpha} \rho_{\min}^4 \underline{\pi_{\text{ref}}}^4} \exp_\alpha\left(-\frac{48|\log(\underline{\pi_{\text{ref}}})|}{\lambda \alpha^2 (1-\gamma)^2 \rho_{\min}} \max(\lambda, \alpha^2)\right)^{4\alpha-8} ,$$

guarantees that

$$v_\star^f(\rho) - \mathbb{E}\left[v_{\theta_t}^f(\rho)\right] \leq \epsilon ,$$

where $f = f_\alpha$ is the α-Csiszár–Cressie–Read divergence generator. Thus, the sample complexity of f-PG to learn an ϵ-solution of the α-Tsallis regularized problem is

$$TBH \approx \frac{|\log(\underline{\pi_{\text{ref}}})|^3 \max(\alpha^{-6}, \lambda^{-3})}{\epsilon B (1-\gamma)^9 \underline{\pi_{\text{ref}}}^{7-7\alpha} \rho_{\min}^5 \underline{\pi_{\text{ref}}}^4} \exp_\alpha\left(-\frac{|\log(\underline{\pi_{\text{ref}}})|}{\lambda \alpha^2 (1-\gamma)^2 \rho_{\min}} \max(\lambda, \alpha^2)\right)^{4\alpha-8} .$$

Proof. To prove this corollary, we will show that under the conditions of this corollary, the assumptions of Theorem E.6 holds. Firstly, note that by using Lemma F.4, the assumption $\mathbf{A}_f(\underline{\pi_{\text{ref}}})$holds. Secondly, using Lemma F.4 note that

$$\begin{aligned}
&\frac{4}{(1-\gamma)^2} + \frac{1}{1-\gamma} \log\left(\frac{216\omega_f^2}{\epsilon \underline{\mu}_{\lambda,f} (1-\gamma)^4}\left[4 + 4\lambda^2 \mathrm{d}_f^2 + \lambda^2 (1-\gamma)^2 \mathrm{y}_f^2\right]\right) \\
&\leq \frac{4}{(1-\gamma)^2} + \frac{1}{1-\gamma} \log\left(\frac{216 \underline{\pi_{\text{ref}}}^{2\alpha-2}}{\epsilon \underline{\mu}_{\lambda,f} (1-\gamma)^4}\left[4 + 32\lambda^2 \frac{|\log(\underline{\pi_{\text{ref}}})|^2}{\alpha^4}\right]\right) \\
&\leq \frac{4}{(1-\gamma)^2} + \frac{1}{1-\gamma} \log\left(\frac{864 \underline{\pi_{\text{ref}}}^{4\alpha-4}\left[1 + 32\lambda^2 |\log(\underline{\pi_{\text{ref}}})|^2/\alpha^4\right]}{\epsilon \lambda (1-\gamma)^5 \rho_{\min}^2 \underline{\pi_{\text{ref}}}^2}\right) \\
&- \frac{1}{1-\gamma} \log\left(\exp_\alpha\left(-\frac{16 + 32\gamma\lambda |\log(\underline{\pi_{\text{ref}}})|}{\lambda \alpha^2 (1-\gamma)^2 \rho_{\min}}\right)^{4-2\alpha}\right) ,
\end{aligned}$$

where in the last inequality, we used the lower bound on $\underline{\mu}_{\lambda,f}$ provided in Lemma F.4. Next using the definition of $\exp_\alpha$ (see Section 3) and the fact that $\lambda \leq 1$, we have

$$\begin{aligned}
&\frac{4}{(1-\gamma)^2} + \frac{1}{1-\gamma} \log\left(\frac{216\omega_f^2}{\epsilon \underline{\mu}_{\lambda,f} (1-\gamma)^4}\left[4 + 4\lambda^2 \mathrm{d}_f^2 + \lambda^2 (1-\gamma)^2 \mathrm{y}_f^2\right]\right) \\
&\leq \frac{4}{(1-\gamma)^2} + \frac{1}{1-\gamma} \log\left(\frac{28152 \underline{\pi_{\text{ref}}}^{4\alpha-4} |\log(\underline{\pi_{\text{ref}}})|^2}{\epsilon \lambda \alpha^4 (1-\gamma)^5 \rho_{\min}^2 \underline{\pi_{\text{ref}}}^2} \max(\alpha^2, \lambda)^2\right) \\
&+ \frac{4-2\alpha}{1-\alpha} \frac{1}{1-\gamma} \log\left(1 + (1-\alpha)\left(\frac{48|\log(\underline{\pi_{\text{ref}}})|}{\lambda \alpha^2 (1-\gamma)^2 \rho_{\min}} \max(\alpha^2, \lambda)\right)\right) \\
&\leq \frac{1}{1-\gamma} \log\left(\frac{28152 \underline{\pi_{\text{ref}}}^{4\alpha-4} |\log(\underline{\pi_{\text{ref}}})|^2}{\epsilon \lambda \alpha^4 (1-\gamma)^5 \rho_{\min}^2 \underline{\pi_{\text{ref}}}^2} \max(\alpha^2, \lambda)^2\right) + \frac{196|\log(\underline{\pi_{\text{ref}}})|}{\lambda \alpha^2 (1-\gamma)^3 \rho_{\min}} \max(\alpha^2, \lambda) ,
\end{aligned}$$

where in the last inequality, we used that for $x \geq 0$, we have $\log(1+x) \leq x$. This shows that our condition on H guarantees that the one set in Theorem E.9 is satisfied. Next, using again Lemma F.4 observe that

$$L_{\lambda,f} = 196\frac{\underline{\pi_{\text{ref}}}^{2\alpha-2}|\log(\underline{\pi_{\text{ref}}})|}{\alpha^2(1-\gamma)^3}\max(\alpha^2,\lambda) \quad , \quad \sigma^2_{\lambda,f} \leq \frac{204\underline{\pi_{\text{ref}}}^{3\alpha-3}\log(\underline{\pi_{\text{ref}}})^2}{\alpha^4(1-\gamma)^4}\max(\alpha^2,\lambda)^2 \tag{100}$$

$$\underline{\mu}_{\lambda,f} \geq \lambda(1-\gamma)\underline{\pi_{\text{ref}}}^{2-2\alpha}\rho^2_{\min}\underline{\pi_{\text{ref}}}^2\exp_\alpha\left(-\frac{48|\log(\underline{\pi_{\text{ref}}})|}{\lambda\alpha^2(1-\gamma)^2\rho_{\min}}\max(\lambda,\alpha^2)\right)^{4-2\alpha} . \tag{101}$$

Hence, we have that

$$\begin{aligned}
&\min\left(\frac{1}{2L_{\lambda,f}}, \frac{\epsilon B\underline{\mu}_{\lambda,f}}{18\sigma^2_{\lambda,f}}\right) \\
&\geq \min\left(\frac{\alpha^2(1-\gamma)^3\min(\alpha^{-2},\lambda^{-1})}{392\underline{\pi_{\text{ref}}}^{2\alpha-2}|\log(\underline{\pi_{\text{ref}}})|}, \right. \\
&\qquad \left. \frac{\epsilon B\lambda(1-\gamma)^5\alpha^4\underline{\pi_{\text{ref}}}^{5-5\alpha}\rho^2_{\min}\underline{\pi_{\text{ref}}}^2}{3672\log(\underline{\pi_{\text{ref}}})^2\max(\alpha^2,\lambda)^2}\exp_\alpha\left(-\frac{48|\log(\underline{\pi_{\text{ref}}})|}{\lambda\alpha^2(1-\gamma)^2\rho_{\min}}\max(\lambda,\alpha^2)\right)^{4-2\alpha}\right) \qquad (102)\\
&= \frac{\epsilon B\lambda(1-\gamma)^5\alpha^4\underline{\pi_{\text{ref}}}^{5-5\alpha}\rho^2_{\min}\underline{\pi_{\text{ref}}}^2}{3672\log(\underline{\pi_{\text{ref}}})^2\max(\alpha^2,\lambda)^2}\exp_\alpha\left(-\frac{48|\log(\underline{\pi_{\text{ref}}})|}{\lambda\alpha^2(1-\gamma)^2\rho_{\min}}\max(\lambda,\alpha^2)\right)^{4-2\alpha} , \qquad (103)
\end{aligned}$$

where in the last identity, we used the fact that $\epsilon < (1-\gamma)^{-1}$ and that

$$B \leq \frac{1}{\underline{\pi_{\text{ref}}}^2\rho^2_{\min}}\exp_\alpha\left(-\frac{48|\log(\underline{\pi_{\text{ref}}})|}{\lambda\alpha^2(1-\gamma)^2\rho_{\min}}\max(\lambda,\alpha^2)\right)^{2\alpha-4} .$$

This shows that our condition on η guarantees that the one set in Theorem E.9 is satisfied. Finally using (103) and (101), we have

$$\begin{aligned}
&\frac{4}{\underline{\mu}_{\lambda,f}}\max\left(2L_{\lambda,f}, \frac{18\sigma^2_{\lambda,f}}{\epsilon B\underline{\mu}_{\lambda,f}}\right) \\
&\quad \leq \frac{4}{\underline{\mu}_{\lambda,f}}\frac{3672\log(\underline{\pi_{\text{ref}}})^2\max(\alpha^2,\lambda)^2}{\epsilon B\lambda(1-\gamma)^5\alpha^4\underline{\pi_{\text{ref}}}^{5-5\alpha}\rho^2_{\min}\underline{\pi_{\text{ref}}}^2}\exp_\alpha\left(-\frac{48|\log(\underline{\pi_{\text{ref}}})|}{\lambda\alpha^2(1-\gamma)^2\rho_{\min}}\max(\lambda,\alpha^2)\right)^{2\alpha-4} \\
&\quad \leq \frac{14688\log(\underline{\pi_{\text{ref}}})^2\max(\alpha^2,\lambda)^2}{\lambda^2\epsilon B(1-\gamma)^6\alpha^4\underline{\pi_{\text{ref}}}^{7-7\alpha}\rho^4_{\min}\underline{\pi_{\text{ref}}}^4}\exp_\alpha\left(-\frac{48|\log(\underline{\pi_{\text{ref}}})|}{\lambda\alpha^2(1-\gamma)^2\rho_{\min}}\max(\lambda,\alpha^2)\right)^{4\alpha-8} ,
\end{aligned}$$

which concludes the proof. □

Corollary F.6. *Assume that, for some $1/4 \geq \underline{\pi_{\text{ref}}} > 0$, π_{ref} satisfy $\boldsymbol{P}(\underline{\pi_{\text{ref}}})$. Assume in addition that the initial distribution ρ satisfies $\boldsymbol{A}_\rho$ and fix f to be the α-Csiszár–Cressie–Read divergence generator, i.e.*

$$f(u) = f_\alpha(u) = \frac{u^\alpha - \alpha u + \alpha - 1}{\alpha(\alpha-1)} .$$

Fix any $(1-\gamma)^{-1} \geq \epsilon > 0$, such that

$$\epsilon < \frac{16}{(1-\gamma)^3\rho_{\min}}\cdot\frac{1-\alpha}{(\pi_{\text{ref}}/2)^{\alpha-1}-1} \quad , \quad \textit{and set } \lambda = \frac{(1-\gamma)\alpha^2\epsilon}{16|\log(\underline{\pi_{\text{ref}}})|} . \tag{104}$$

Additionally set any B such that

$$B \leq \frac{1}{\epsilon^2\alpha^2(1-\gamma)^3\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\exp_\alpha\left(-\frac{384|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^3\rho_{\min}}\right)^{2\alpha-4} \tag{105}$$

Setting

$$H \geq \frac{1}{1-\gamma}\log\left(\frac{19008\underline{\pi_{\text{ref}}}^{4\alpha-2}|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^6\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\right)+\frac{1540|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^4\rho_{\min}}\ ,$$

and

$$\eta \leq \frac{\epsilon^2\alpha^2(1-\gamma)^6B\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}{13824\underline{\pi_{\text{ref}}}^{5\alpha-5}|\log(\underline{\pi_{\text{ref}}})|}\exp_\alpha\left(-\frac{384|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^3\rho_{\min}}\right)^{4-2\alpha}\ , \tag{106}$$

and

$$T \geq \frac{12^7\underline{\pi_{\text{ref}}}^{7\alpha-7}|\log(\underline{\pi_{\text{ref}}})|^2}{\epsilon^3\alpha^4(1-\gamma)^8\rho_{\min}^4\underline{\pi_{\text{ref}}}^4B}\exp_\alpha\left(-\frac{384|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^3\rho_{\min}}\right)^{4\alpha-8}\log\left(\frac{6(v_\star^f(\rho)-v_{\theta_0}^f(\rho))}{\epsilon}\right)\ , \tag{107}$$

guarantees that

$$v_\star(\rho)-\mathbb{E}\left[v_{\theta_t}(\rho)\right]\leq\epsilon\ .$$

Thus, the sample complexity of f-PG, where f is theα-Csiszár–Cressie–Read divergence generator, to learn an ϵ-solution of the non-regularized problem is

$$TBH \approx \frac{\underline{\pi_{\text{ref}}}^{7\alpha-7}|\log(\underline{\pi_{\text{ref}}})|^3}{\epsilon^4\alpha^6(1-\gamma)^{12}\rho_{\min}^5\underline{\pi_{\text{ref}}}^4}\exp_\alpha\left(-\frac{384|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^3\rho_{\min}}\right)^{4\alpha-8}\ .$$

Proof. To prove this corollary, we will show that under the conditions of this corollary, the assumptions of Theorem E.9 holds. Firstly, note that (104) implies that

$$\epsilon<\frac{16}{(1-\gamma)^3\rho_{\min}}\min\left(\frac{4}{|f'(\iota_f)|},\frac{1}{|f'(\frac{1}{2})|},\frac{4}{|f'(\frac{1}{2}\underline{\pi_{\text{ref}}})|}\right)\ ,\quad \lambda=\frac{(1-\gamma)\epsilon}{4}\mathrm{c}_f\ ,$$

with $\mathrm{c}_f=\alpha^2/4|\log(\underline{\pi_{\text{ref}}})|$. Additionally, observe using Lemma F.4 that we have

$$C_f^{(1)}\leq\frac{64\underline{\pi_{\text{ref}}}^{2\alpha-2}|\log(\underline{\pi_{\text{ref}}})|}{\alpha^2},\ \ C_f^{(2)}\leq 96\underline{\pi_{\text{ref}}}^{2\alpha-2},\ \ C_f^{(3)}\leq\frac{13824\underline{\pi_{\text{ref}}}^{5\alpha-5}|\log(\underline{\pi_{\text{ref}}})|}{\alpha^2},$$
$$d(\epsilon)=\exp_\alpha\left(-\frac{384|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^3\rho_{\min}}\right)^{4-2\alpha}, \tag{108}$$

where $C_f^{(1)},C_f^{(2)},C_f^{(3)}$, and $d(\epsilon)$ are defined in (80) and (79). Next, observe that

$$\begin{aligned}
&\frac{4}{(1-\gamma)^2}+\frac{1}{1-\gamma}\log\left(\frac{297\omega_f^2C_1^f}{\epsilon d(\epsilon)(1-\gamma)^6\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\right)\\
&\leq\frac{4}{(1-\gamma)^2}+\frac{1}{1-\gamma}\log\left(\frac{19008\underline{\pi_{\text{ref}}}^{4\alpha-2}|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^6\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\right)+\frac{1}{1-\gamma}\log\left(\exp_\alpha\left(-\frac{384|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^3\rho_{\min}}\right)^{4-2\alpha}\right)\\
&\leq\frac{4}{(1-\gamma)^2}+\frac{1}{1-\gamma}\log\left(\frac{19008\underline{\pi_{\text{ref}}}^{4\alpha-2}|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^6\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\right)+\frac{1536|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^4\rho_{\min}}\ ,
\end{aligned}$$

where in the last inequality, we used that $\log(1+u) \leq x$ for $u > -1$. This shows that our condition on H implies the one assumed in Theorem E.9. Using (108), we have

$$\begin{aligned}
&\min\left(\frac{(1-\gamma)^3}{C_f^{(2)}}, \frac{\epsilon^2 d(\epsilon)(1-\gamma)^6 B \rho_{\min}^2 \underline{\pi_{\text{ref}}}^2}{C_f^{(3)}}\right) \\
&\geq \min\left(\frac{(1-\gamma)^3}{96\underline{\pi_{\text{ref}}}^{2\alpha-2}}, \frac{\epsilon^2\alpha^2(1-\gamma)^6 B\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}{13824\underline{\pi_{\text{ref}}}^{5\alpha-5}|\log(\underline{\pi_{\text{ref}}})|}\exp_\alpha\left(-\frac{384|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^3\rho_{\min}}\right)^{4-2\alpha}\right) \\
&\geq \frac{\epsilon^2\alpha^2(1-\gamma)^6 B\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}{13824\underline{\pi_{\text{ref}}}^{5\alpha-5}|\log(\underline{\pi_{\text{ref}}})|}\exp_\alpha\left(-\frac{384|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^3\rho_{\min}}\right)^{4-2\alpha}, \qquad (109)
\end{aligned}$$

where in the last inequality, we used the condition on B introduced in (105). Hence our condition on the step size ensures that the one assumed in Theorem E.9 is satisfied. Next, using (108) and (109) yields

$$\begin{aligned}
&\frac{16C_f^{(1)}\ell(\epsilon)}{\epsilon d(\epsilon)(1-\gamma)^2\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\max\left(\frac{C_f^{(2)}}{(1-\gamma)^3}, \frac{C_f^{(3)}}{\epsilon^2 d(\epsilon)(1-\gamma)^6 B\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\right) \\
&\leq \frac{16C_f^{(1)}\ell(\epsilon)}{\epsilon d(\epsilon)(1-\gamma)^2\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\cdot\frac{13824\underline{\pi_{\text{ref}}}^{5\alpha-5}|\log(\underline{\pi_{\text{ref}}})|}{\epsilon^2\alpha^2(1-\gamma)^6 B\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\exp_\alpha\left(-\frac{384|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^3\rho_{\min}}\right)^{2\alpha-4} \\
&\leq \frac{1024\underline{\pi_{\text{ref}}}^{2\alpha-2}|\log(\underline{\pi_{\text{ref}}})|\ell(\epsilon)}{\epsilon\alpha^2(1-\gamma)^2\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\cdot\frac{13824\underline{\pi_{\text{ref}}}^{5\alpha-5}|\log(\underline{\pi_{\text{ref}}})|}{\epsilon^2\alpha^2(1-\gamma)^6 B\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\exp_\alpha\left(-\frac{384|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^3\rho_{\min}}\right)^{4\alpha-8} \\
&\leq \frac{12^7\underline{\pi_{\text{ref}}}^{7\alpha-7}|\log(\underline{\pi_{\text{ref}}})|^2}{\epsilon^3\alpha^4(1-\gamma)^8\rho_{\min}^4\underline{\pi_{\text{ref}}}^4 B}\exp_\alpha\left(-\frac{384|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^3\rho_{\min}}\right)^{4\alpha-8}\log\left(\frac{6(v_\star^f(\rho)-v_{\theta_0}^f(\rho))}{\epsilon}\right),
\end{aligned}$$

which proves that under our condition on T the one assumed by Theorem E.9 is satisfied. □

Corollary F.7. *Assume the same condition of Corollary F.6. For any $(1-\gamma)^{-1} > \epsilon > 0$ and $\alpha \in (0,1)$, denote respectively by $T(\epsilon,\alpha)$, $B(\epsilon,\alpha)$, and $H(\epsilon,\alpha)$, the thresholds set in Corollary F.6 on T, B, and H, to learn an ϵ-solution of the unregularized problem. Finally, denote by $\alpha^\star(\epsilon)$ the minimizer of $T(\epsilon,\alpha)B(\epsilon,\alpha)H(\epsilon,\alpha)$. It holds that*

$$\alpha^\star(\epsilon) = \frac{11}{2}\cdot\frac{1}{\log(1/\epsilon)} + o\left(\frac{1}{\log(1/\epsilon)}\right).$$

Proof. Firstly, note that using Corollary F.6, we have

$$\begin{aligned}
\mathrm{S}(\epsilon,\alpha) &:= \log\left(T(\epsilon,\alpha)B(\epsilon,\alpha)H(\epsilon,\alpha)\right) \\
&= \log\left(\frac{12^7\underline{\pi_{\text{ref}}}^{7\alpha-7}|\log(\underline{\pi_{\text{ref}}})|^2}{\epsilon^3\alpha^4(1-\gamma)^8\rho_{\min}^4\underline{\pi_{\text{ref}}}^4}\right) \\
&+ \frac{4\alpha-8}{\alpha-1}\log\left(1+(1-\alpha)\left(\frac{384|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^3\rho_{\min}}\right)\right) \\
&+ \log\left(\log\left(\frac{6(v_\star^f(\rho)-v_{\theta_0}^f(\rho))}{\epsilon}\right)\right) \\
&+ \log\left(\frac{1}{1-\gamma}\log\left(\frac{19008\underline{\pi_{\text{ref}}}^{4\alpha-2}|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^6\rho_{\min}^2\underline{\pi_{\text{ref}}}^2}\right) + \frac{1540|\log(\underline{\pi_{\text{ref}}})|}{\epsilon\alpha^2(1-\gamma)^4\rho_{\min}}\right).
\end{aligned}$$

Firstly, observe that for any function $k(\epsilon)$ which does converge to a different value from 0, we have

$$\lim_{\varepsilon \to 0} \frac{S(\varepsilon, \varepsilon)}{S(\varepsilon, k(\varepsilon))} < 1 \ ,$$

which establishes that $\alpha^\star(\epsilon) \to 0$. This allows, to rewrite $\mathrm{S}(\epsilon, \alpha)$ as

$$\mathrm{S}(\epsilon, \alpha) = \log(\frac{1}{\epsilon^4 \alpha^6}) + \frac{8 - 4\alpha}{1 - \alpha} \log\left(\frac{1}{\epsilon \alpha^2}\right) + \psi(\alpha, \epsilon) \ ,$$

where $\psi(\alpha, \epsilon)$ is defined as

$$\begin{aligned}
\psi(\alpha, \epsilon) = \log &\left(\frac{12^7 \underline{\pi_{\mathrm{ref}}}^{7\alpha - 7} |\log(\underline{\pi_{\mathrm{ref}}})|^2}{(1-\gamma)^8 \rho_{\min}^4 \underline{\pi_{\mathrm{ref}}}^4}\right) \\
&+ \frac{4\alpha - 8}{\alpha - 1}\left[\log\left(1 + (1-\alpha)\left(\frac{384 |\log(\underline{\pi_{\mathrm{ref}}})|}{\epsilon \alpha^2 (1-\gamma)^3 \rho_{\min}}\right)\right) - \log\left(\frac{1}{\epsilon \alpha^2}\right)\right] \\
&+ \log\left(\log\left(\frac{6(v_\star^f(\rho) - v_{\theta_0}^f(\rho))}{\epsilon}\right)\right) \\
&+ \log\left(\frac{\epsilon \alpha^2}{1-\gamma} \log\left(\frac{19008 \underline{\pi_{\mathrm{ref}}}^{4\alpha - 2} |\log(\underline{\pi_{\mathrm{ref}}})|}{\epsilon \alpha^2 (1-\gamma)^6 \rho_{\min}^2 \underline{\pi_{\mathrm{ref}}}^2}\right) + \frac{1540 |\log(\underline{\pi_{\mathrm{ref}}})|}{(1-\gamma)^4 \rho_{\min}}\right) \ .
\end{aligned}$$

Importantly, observe that function dominated by $\log(1/\epsilon)$ when $(\alpha, \epsilon) \to 0$ and that

$$\frac{\partial \psi(\alpha, \epsilon)}{\partial \alpha} + o\left(\frac{1}{\alpha \log(\frac{1}{\epsilon \alpha})}\right) \ .$$

Computing the derivative of this function with respect to α yields

$$\begin{aligned}
\frac{\partial \mathrm{S}(\epsilon, \alpha)}{\partial \alpha} &= \frac{-6}{\alpha} + 4\frac{1}{(1-\alpha)^2} \log(\frac{1}{\epsilon}) + 4\frac{1}{(1-\alpha)^2} \log(\frac{1}{\alpha^2}) - \frac{8}{\alpha}\left(1 + \frac{1}{1-\alpha}\right) + \frac{\partial \psi(\epsilon, \alpha)}{\partial \alpha} && (110) \\
&= \frac{-22}{\alpha} + 4\log(\frac{1}{\epsilon}) + o\left(\frac{1}{\alpha \log(\frac{1}{\epsilon \alpha})}\right) \ . && (111)
\end{aligned}$$

As

$$\left.\frac{\partial \mathrm{S}(\epsilon, \alpha)}{\partial \alpha}\right|_{\alpha = \alpha^\star(\epsilon)} = 0 \ ,$$

Then this implies that

$$\alpha^\star(\epsilon) = \frac{11}{2} \cdot \frac{1}{\log(1/\epsilon)} + o\left(\frac{1}{\log(1/\epsilon)}\right) \ .$$

□

G TECHNICAL LEMMAS

Lemma G.1 (Lemma 1.2.3 in Nesterov (2004)). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable. Suppose there exists $L \geq 0$ such that for all $x \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$,*

$$|v^\top \nabla^2 f(x)\, v| \ \leq \ L\|v\|^2.$$

Then f has an L-Lipschitz continuous gradient (i.e., f is L-smooth); in particular,

$$\|\nabla f(y) - \nabla f(x)\| \;\leq\; L\|y - x\|,$$

and

$$f(y) \;\geq\; f(x) + \langle \nabla f(x), y - x\rangle - \tfrac{L}{2}\|y - x\|^2$$

for all $x, y \in \mathbb{R}^d$.

Lemma G.2. *Consider any two policies π_i, $i = 1, 2$. It holds that*

$$\left\|d_\rho^{\pi_1} - d_\rho^{\pi_2}\right\|_1 \leq \frac{\gamma}{1-\gamma} \sup_{s\in\mathcal{S}} \|\pi_1(\cdot|s) - \pi_2(\cdot|s)\|_1 \quad .$$

Proof. Let us start from the definition of flow conservation constraints for the discounted state visitation (Puterman, 1994), for $i \in \{1, 2\}$, we have

$$d_\rho^{\pi_i}(s) = (1-\gamma)\rho(s) + \gamma \sum_{s'} \mathsf{P}_{\pi_i}(s|s') d_\rho^{\pi_i}(s') \quad .$$

Then, we have

$$\begin{aligned}
\sum_{s\in\mathcal{S}} |d_\rho^{\pi_2}(s) - d_\rho^{\pi_1}(s)| &\leq \gamma \sum_{(s',a')} \sum_{s} \left|\mathsf{P}(s|s',a')\pi_2(a'|s')d_\rho^{\pi_2}(s') - \mathsf{P}(s|s',a')\pi_1(a'|s')d_\rho^{\pi_1}(s')\right| \\
&\leq \gamma \sum_{s',a'} \sum_{s} \mathsf{P}(s|s',a') \left|\pi_2(a'|s') - \pi_1(a'|s')\right| d_\rho^{\pi_2}(s') \\
&+ \gamma \sum_{s',a'} \sum_{s} \mathsf{P}(s|s',a')\pi_1(a'|s') \left|d_\rho^{\pi_1}(s') - d_\rho^{\pi_2}(s')\right| \\
&\leq \gamma \sup_{s\in\mathcal{S}} \|\pi_1(\cdot|s) - \pi_2(\cdot|s)\|_1 + \gamma \sum_{s'} |d_\rho^{\pi_1}(s') - d_\rho^{\pi_2}(s')| \quad ,
\end{aligned}$$

which concludes the proof. □

Lemma G.3 (Performance Difference Lemma)**.** *It holds that*

$$v_\star^f(\rho) - v_\theta^f(\rho) = \frac{1}{1-\gamma} \sum_{s\in\mathcal{S}} d_\rho^{\pi_\star^f}(s) \left[\sum_{a\in\mathcal{A}} \pi_\star^f(a|s) q_\theta^f(s,a) - \lambda\, \mathrm{D}^f(\pi_\star^f(\cdot|s)\|\pi_{\mathrm{ref}}(\cdot|s)) - v_{\pi_\theta}^f(s)\right] \quad .$$

Proof. Fix $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and any state $s \in \mathcal{S}$. It holds that

$$\begin{aligned}
v_\star^f(s) - v_{\pi_\theta}^f(s) &= \sum_{a\in\mathcal{A}} \pi_\star^f(a|s) q_\star^f(s,a) - \sum_{a\in\mathcal{A}} \pi_\theta^f(a|s) q_\theta^f(s,a) \\
&\quad - \lambda\, \mathrm{D}^f(\pi_\star^f(\cdot|s)\|\pi_{\mathrm{ref}}(\cdot|s)) + \lambda\, \mathrm{D}^f(\pi_\theta^f(\cdot|s)\|\pi_{\mathrm{ref}}(\cdot|s)) \\
&= \sum_{a\in\mathcal{A}} \pi_\star^f(a|s) \left(q_\star^f(s,a) - q_\theta^f(s,a)\right) + \sum_{a\in\mathcal{A}} \left(\pi_\star^f(a|s) - \pi_\theta^f(a|s)\right) q_\theta^f(s,a) \\
&\quad - \lambda\, \mathrm{D}^f(\pi_\star^f(\cdot|s)\|\pi_{\mathrm{ref}}(\cdot|s)) + \lambda\, \mathrm{D}^f(\pi_\theta^f(\cdot|s)\|\pi_{\mathrm{ref}}(\cdot|s)) \\
&= \gamma \sum_{a\in\mathcal{A}} \pi_\star^f(a|s) \sum_{s'\in\mathcal{S}} \mathsf{P}(s'|s,a) \left(v_\star^f(s') - v_\theta^f(s')\right) + \sum_{a\in\mathcal{A}} \left(\pi_\star^f(a|s) - \pi_\theta^f(a|s)\right) q_\theta^f(s,a) \\
&\quad - \lambda\, \mathrm{D}^f(\pi_\star^f(\cdot|s)\|\pi_{\mathrm{ref}}(\cdot|s)) + \lambda\, \mathrm{D}^f(\pi_\theta^f(\cdot|s)\|\pi_{\mathrm{ref}}(\cdot|s)) \quad ,
\end{aligned}$$

where in the last equality, we used the definition of the regularized Q-function (4). Expanding the recursion yields

$$
\begin{aligned}
v_\star^f(s) - v_{\pi_\theta}^f(s) &= \frac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} d_s^\star(s') \left[\sum_{a \in \mathcal{A}} \left(\pi_\star^f(a|s') - \pi_\theta^f(a|s') \right) q_\theta^f(s', a) \right] \\
&\quad + \frac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} d_s^\star(s') \left[\lambda \, \mathrm{D}^f(\pi_\theta^f(\cdot|s') \| \pi_{\mathrm{ref}}(\cdot|s')) - \lambda \, \mathrm{D}^f(\pi_\star^f(\cdot|s') \| \pi_{\mathrm{ref}}(\cdot|s')) \right] \\
&= \frac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} d_s^\star(s') \left[\sum_{a \in \mathcal{A}} \pi_\star^f(a|s') q_\theta^f(s', a) - \lambda \, \mathrm{D}^f(\pi_\star^f(\cdot|s') \| \pi_{\mathrm{ref}}(\cdot|s')) \right] \\
&\quad - \frac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} d_s^\star(s') \left[\sum_{a \in \mathcal{A}} \pi_\theta^f(a|s') q_\theta^f(s', a) - \lambda \, \mathrm{D}^f(\pi_\theta^f(\cdot|s') \| \pi_{\mathrm{ref}}(\cdot|s')) \right] \quad ,
\end{aligned}
$$

which concludes the proof. □

Lemma G.4 (Lemma 23 of Mei et al. (2020b))**.** *Let $\pi \in \mathcal{P}(\mathcal{A})$. Denote $H(\pi) = \mathrm{diag}(\pi) - \pi\pi^\top$. For any vector $x \in \mathbb{R}^{|\mathcal{A}|}$*

$$
\left\| H(\pi) \left(x - \frac{\langle x, 1_{|\mathcal{A}|} \rangle}{|\mathcal{A}|} 1_{|\mathcal{A}|} \right) \right\|_2 \geq \min_{a \in \mathcal{A}} \pi(a) \cdot \left\| x - \frac{\langle x, 1_{|\mathcal{A}|} \rangle}{|\mathcal{A}|} 1_{|\mathcal{A}|} \right\|_2 \quad .
$$

Lemma G.5 (Danskin (1966))**.** *Let $Z \subset \mathbb{R}^m$ be compact and let $\phi : \mathbb{R}^n \times Z \to \mathbb{R}$ be continuous. Define*

$$
f(x) \;=\; \max_{z \in Z} \phi(x, z), \qquad Z_0(x) \;=\; \arg\max_{z \in Z} \phi(x, z).
$$

Assume that for each fixed $z \in Z$, the map $x \mapsto \phi(x, z)$ is differentiable. If $Z_0(x) = \{\bar{z}\}$ and $x \mapsto \phi(x, \bar{z})$ is differentiable at x, then f is differentiable at x with

$$
\frac{\partial f(x)}{\partial x} \;=\; \frac{\partial \phi(x, \bar{z})}{\partial x}.
$$

H Links with Mirror Descent

To avoid overwhelming readers with technical details, we keep the discussion in this paragraph at a high level. There is a clear connection between the coupled parameterization we consider and mirror descent (MD) algorithms. The discussion below is informal, meant to highlight the key ideas. We stress that the proposed method is fundamentally different from mirror descent. Let us define a mapping $\Phi(\pi) = \sum_{s \in \mathcal{S}} \mathrm{D}^f(\pi(\cdot|s) \| \pi_{\mathrm{ref}}(\cdot|s))$. For the functions f that we consider, Φ is Legendre on the positive orthant and separable across states (Bubeck et al., 2015). In this case, the f-regularized value function $v_\pi^f(\rho)$ can be optimized directly in the policy space via the Lazy Mirror Descent algorithm (or dual averaging; see Nesterov (2009); Xiao (2009); Juditsky et al. (2023)) with Φ as mirror map. Denoting by π_t the policy at step t and $\widetilde{\pi}_t$ by the unnormalized policy at step t, the lazy MD updates reads:

$$
\nabla\Phi(\widetilde{\pi}_{t+1}) = \nabla\Phi(\widetilde{\pi}_t) + \eta \nabla_\pi v_\pi^f(\rho)|_{\pi=\pi_t} \,, \quad \pi_{t+1} = \arg\min_{\pi \in \Pi} B_\Phi(\pi \| \widetilde{\pi}_{t+1}) \,. \tag{112}
$$

where $\Pi = \mathcal{P}(\mathcal{A})^{|\mathcal{S}|}$ is a policy space and $B_\Phi(\pi \| \pi') = \Phi(\pi) - \Phi(\pi') - \langle \nabla\Phi(\pi'), \pi - \pi' \rangle$ is the corresponding Bregman divergence. Since Φ is separable over states, the Bregman projection can be written state-wise as $\pi_{t+1}(\cdot|s) = \mathrm{softargmax}^f(\nabla\Phi(\widetilde{\pi}_{t+1})(s, \cdot), \pi_{\mathrm{ref}}(\cdot|s))$.

By denoting $\theta_t = \nabla\Phi(\widetilde{\pi}_t)$, one obtains updates that resemble those of (14) (after the removal of $\mathcal{T}$), with one important difference: the gradient in (112) is taken with respect to the policy π whereas in (14) it is computed w.r.t the "dual" parameter θ (in the MD terminology). Even more important, the update (112) can be expressed as, by the chain rule

$$\theta_{t+1} = \theta_t + \eta \left[\frac{\partial \pi_\theta^f}{\partial \theta} \Big|_{\theta=\theta_t} \right]^{-1} \nabla_\theta J^f(\theta_t) \,,$$

which have an additional preconditioning term given by the inverse of the policy Jacobian.

A crucial feature of (14) is that it performs a gradient ascent in the "dual" space directly. This algorithm can be extended in the non-tabular setting directly, by parameterizing the function $\theta(s, a)$, allowing extensions to deep RL. This is in contrast with Lazy-MD methods (Nesterov, 2009; Xiao, 2009; Juditsky et al., 2023), due to preconditioning, which cannot be expressed as direct parameter-space gradient steps. This remark has several important implications, which we leave for future work.