Proximal Point Nash Learning from Human Feedback

Daniil Tiapkin[1,2] **Daniele Calandriello**[3] **Denis Belomestny**[4,5] **Éric Moulines**[6,7]
Alexey Naumov[5] **Kashif Rasul**[8] **Michal Valko**[9] **Pierre Ménard**[10]
[1]CMAP, CNRS, École Polytechnique, IPP [2]LMO, Université Paris-Saclay [3]Google DeepMind
[4]Duisburg-Essen University [5]HSE University [6]Mohamed Bin Zayed University of AI
[7]LRE EPITA [8]Hugging Face [9]Isara Labs [10]ENS Lyon
daniil.tiapkin@polytechnique.edu dcalandriello@google.com
denis.belomestny@uni-due.de eric.moulines@mbzuai.ac.ae
anaumov@hse.ru kashif.rasul@gmail.com
michal@isara.io pierre.menard@ens-lyon.fr

Abstract

Traditional Reinforcement Learning from Human Feedback (RLHF) often relies on reward models, frequently assuming preference structures like the Bradley–Terry model, which may not accurately capture the complexities of real human preferences (e.g., intransitivity). Nash Learning from Human Feedback (NLHF) offers a more direct alternative by framing the problem as finding a Nash equilibrium of a game defined by these preferences. While many works study the Nash learning problem directly in the policy space, we instead consider it under a more realistic policy parametrization setting. We first analyze a simple self-play policy gradient method, which is equivalent to Online IPO. We establish high-probability last-iterate convergence guarantees for this method, but our analysis also reveals a possible stability limitation of the underlying dynamics. Motivated by this, we embed the self-play updates into a proximal point framework, yielding a stabilized algorithm. For this combined method, we prove high-probability last-iterate convergence and discuss its more practical version, which we call Nash Prox. Finally, we apply this method to post-training of large language models and validate its empirical performance.

1 Introduction

Aligning powerful pre-trained Large Language Models (LLMs) with complex and often subjective human preferences and values is a central challenge for safe and beneficial AI. Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017) addresses this by learning from human preference signals rather than sparse or hand-engineered reward functions. RLHF has been successfully used to fine-tune LLMs for tasks such as summarization (Stiennon et al., 2020), dialogue, and question answering (Ziegler et al., 2019; Ouyang et al., 2022; Bai et al., 2022).

A common approach within RLHF, rooted in contextual dueling bandits (Yue et al., 2012; Zoghi et al., 2014; Bengs et al., 2021), is to posit an underlying reward model. The most prevalent choice is the Bradley–Terry (BT) model (Zermelo, 1929; Bradley & Terry, 1952), which assigns each action a scalar reward and models pairwise preferences as a function of reward differences. Under BT, the goal reduces to learning the reward and choosing an action that is preferred, on average, to alternatives. In Social Choice Theory, this corresponds to a Condorcet winner.

However, reward-model approaches can be misspecified (Dudík et al., 2015; Munos et al., 2023). In particular, BT imposes transitivity: if a is preferred to b and b to c, then a must be preferred to c. Real human judgments often violate this property, exhibiting intransitivity (Gardner, 1970; Tversky, 1969;

Preprint.

Klimenko, 2015). Moreover, even when individual preferences are transitive, group aggregation can create cycles (May, 1954; Kreweras, 1965). Such non-transitive preferences can preclude a Condorcet winner and, more broadly, a consistent scalar reward function matching all comparisons.

Nash Learning from Human Feedback (NLHF). To avoid assuming a consistent reward or a Condorcet winner, Dudík et al. (2015) proposed a preference-based approach for dueling bandits, later termed Nash Learning from Human Feedback (NLHF) by Munos et al. (2023). NLHF models pairwise preferences via a symmetric two-player game in which each player proposes an action. The objective is to find a symmetric Nash equilibrium (NE, von Neumann 1928; Nash Jr 1950), called a von Neumann winner (VNW) in the dueling bandits literature (Dudík et al., 2015). Unlike a Condorcet winner, a VNW is a mixed policy and remains well-defined under intransitive preferences.

Regularization and the parameterized setting. In LLM post-training, we want to align to preferences while staying close to a fixed reference policy (for example, an instruction-following model). We therefore consider the NE of a *regularized* preference game, adding a penalty proportional to the Kullback–Leibler (KL) divergence from the current policy to the reference policy. While many works analyze algorithms for regularized NLHF by exploiting convexity in the policy space (Sokota et al., 2023; Munos et al., 2023; Cen et al., 2024), less is known when the policy is *parameterized* (as in neural networks) and trained via stochastic gradient methods.

Our approach. We close this gap by analyzing the simplest self-play policy gradient (SPG) under parameterization. Under a leave-one-out advantage estimator, SPG iterations coincide with Online IPO (Calandriello et al., 2024). We prove high-probability last-iterate convergence and identify a stability condition on regularization; beyond it, the dynamics may be unstable.

Motivated by this observation, we wrap self-play updates in an (inexact) proximal-point outer loop (Martinet, 1970). The resulting method uses two anchors: the reference policy (to limit drift from the base model) and the previous outer iterate (to stabilize dynamics). This yields high-probability last-iterate convergence without a stability restriction on the regularization parameter. We also propose a practical variant, `Nash Prox`, which approximates the proximal anchor via an exponential moving average (EMA), and show competitive performance on synthetic preference games and LLM post-training.

Contributions. We summarize our contributions as follows.

- We analyze SPG (equivalently, Online IPO under leave-one-out advantage estimator) for regularized NLHF under general policy parameterization, which includes softmax parametrization, proving high-probability last-iterate convergence under a stability condition on the regularized game.
- We embed SPG into an inexact proximal-point framework (PP-SPG), achieving high-probability last-iterate convergence in the parameterized setting without extra stability conditions.
- We develop a practical deep-learning implementation of PP-SPG called `Nash Prox`, based on an EMA target anchor, and show competitive results on synthetic games and LLM post-training.

2 Related work

Nash Learning from Human Feedback. The NLHF framework was introduced by Munos et al. (2023), building on the formulation of contextual dueling bandits as a symmetric two-player game by Dudík et al. (2015). They proposed the `NashMD` algorithm and showed that it enjoys last-iterate convergence to the VNW of a regularized preference game at a polynomial rate.

Subsequent work studied how to approximate a VNW in the *unregularized* preference game, e.g., via regret-minimization tools (Swamy et al., 2024; Wu et al., 2025a), optimistic mirror descent variants (Zhang et al., 2025a; Wu et al., 2025b), and meta-algorithms that ensure asymptotic convergence (Liu et al., 2024b). A separate line of work improved guarantees in the regularized setting (Zhang et al., 2025b; Wang et al., 2025; Tang et al., 2025), but these methods typically operate in the policy space via direct distributional updates. A notable exception is the work by Zhou et al. (2025), which provides a policy-gradient interpretation but requires samples from a uniform distribution over the action space, which is impractical in the LLM setting where actions are sentences.

Separately, Calandriello et al. (2024) showed that the online version of IPO (Gheshlaghi Azar et al., 2024) converges to the VNW, but without explicit rates. In this work, we give high-probability

last-iterate convergence guarantees for Online IPO in the parameterized setting. The result follows from its equivalence to SPG under a leave-one-out advantage estimator and holds under a condition on the regularization coefficient. We also introduce a stabilized proximal-point variant that uses Online IPO as a building block.

Most of the above work focuses on single-step preference feedback. Addressing multi-turn decision making, Shani et al. (2025) studied preference feedback in settings requiring planning, proposes a self-play mirror descent based algorithm, and proves convergence to a Nash equilibrium. Another line of work extends the NLHF to a Stackelberg principal-agent formulation, breaking the symmetry between max and min-players (Choi et al., 2025; Pásztor et al., 2025).

Policy gradient methods. A large body of work studies policy gradient (PG) methods for unregularized reinforcement learning with general parameterized policy classes. A widely used set of sufficient conditions combines regularity of the score function with Fisher non-degeneracy and (approximate) compatibility (Sutton et al., 1999), enabling global convergence and finite-sample guarantees beyond the tabular setting (Papini et al., 2018; Huang et al., 2020; Liu et al., 2020; Agarwal et al., 2021; Yuan et al., 2022; Ding et al., 2022; Fatkhullin et al., 2023; Lu et al., 2024). These analyses typically consider a *fixed* objective and quantify progress using smoothness and gradient-dominance conditions on the induced objective in parameter space.

Regularized policy gradients. Less is known about *Kullback-Leibler (KL)- or entropy-regularized* objectives under parameterization, despite their prominence in RLHF and preference optimization. For entropy-regularized MDPs, Mei et al. (2020) proved global convergence of softmax PG and showed gradient dominance is non-uniform without control of the minimum action probability. Subsequent work analyzes regularized gradient flows and stability (Leahy et al., 2022), and uses projection/truncation to maintain a minimum action probability (Zhang et al., 2021; Labbi et al., 2025); see also Liu et al. (2024a) for convergence over a wider range of step sizes. For two-player zero-sum Markov games, Zeng et al. (2022) studied regularized gradient descent–ascent with softmax policies under partially decoupled updates (alternating steps and unequal step sizes). In contrast, our self-play policy gradient uses a single timescale with simultaneous updates and a shared learning rate, so the opponent evolves at the same rate each step, making the analysis more challenging.

3 Setting

We consider a contextual dueling bandit setting $(\mathcal{X}, \mathcal{Y}, \mathcal{P})$, where $\mathcal{X}$ is a context space, $\mathcal{Y}$ is a finite action space, and $\mathcal{P}(y \succ y' \mid x) \in [0,1]$ is the probability that action $y \in \mathcal{Y}$ is preferred to $y' \in \mathcal{Y}$ given context $x \in \mathcal{X}$, which satisfies the symmetry condition: $\mathcal{P}(y \succ y' \mid x) = 1 - \mathcal{P}(y' \succ y \mid x)$ for all x, y, y'. A policy $\pi \colon \mathcal{X} \to \Delta_{\mathcal{Y}}$ maps contexts to probability distributions over actions, where $\Delta_{\mathcal{Y}}$ is the probability simplex over $\mathcal{Y}$. Let Π be the space of all such policies. For a context $x \in \mathcal{X}$, we define the expected preference of a policy $\pi \in \Pi$ over π', as:

$$\mathcal{P}(\pi \succ \pi' | x) \triangleq \mathbb{E}_{y \sim \pi(\cdot|x), y' \sim \pi'(\cdot|x)} [\mathcal{P}(y \succ y' | x)] \,.$$

For a context distribution $\rho \in \Delta_{\mathcal{X}}$ we define the expected preference and Kullback-Leibler (KL) divergence as $\mathcal{P}(\pi \succ \pi') = \mathbb{E}_{x \sim \rho}[\mathcal{P}(\pi \succ \pi' | x)]$ and $\mathrm{KL}_\rho(\pi \| \pi') = \mathbb{E}_{x \sim \rho}[\mathrm{KL}(\pi(x) \| \pi'(x)]$.

Regularized preference game. For a fixed reference policy $\pi^{\mathrm{ref}} \in \Pi$ we define a *β-regularized preference* as $\mathcal{P}_\beta(\pi \succ \pi') \triangleq \mathcal{P}(\pi \succ \pi') - \beta \,\mathrm{KL}(\pi \| \pi^{\mathrm{ref}}) + \beta \,\mathrm{KL}(\pi' \| \pi^{\mathrm{ref}})$, where $\beta > 0$ is a regularization parameter. The β-regularized preference function induces a β-regularized preference game (Munos et al., 2023), which is symmetric, so it admits a symmetric Nash Equilibrium (NE) $(\pi^\star_\beta, \pi^\star_\beta)$ (von Neumann, 1928; Nash Jr, 1950). We call a policy $\pi^\star_\beta$ a β-regularized von Neumann winner (VNW, Dudík et al. 2015) and in particular it satisfies

$$\pi^\star_\beta \in \arg\max_{\pi \in \Pi} \min_{\pi' \in \Pi} \mathcal{P}_\beta(\pi \succ \pi') \,.$$

The β-regularized suboptimality (also known as exploitability gap) of π against π' is $\mathrm{SubOpt}_\beta(\pi, \pi') \triangleq \frac{1}{2} - \mathcal{P}_\beta(\pi \succ \pi')$. The worst-case β-regularized suboptimality of π is:

$$\mathrm{SubOpt}_\beta(\pi) \triangleq \tfrac{1}{2} - \min_{\pi' \in \Pi} \mathcal{P}_\beta(\pi \succ \pi') \,. \tag{1}$$

A policy π is an ε-VNW in the β-regularized game if $\mathrm{SubOpt}_\beta(\pi) \le \varepsilon$. Our goal is to learn an ε-VNW given (i) sampling access to the context distribution ρ, (ii) the ability to sample from context-conditioned policies, and (iii) a stochastic estimate of $\mathcal{P}(y \succ y' \mid x)$ (e.g., from pairwise comparisons).

We measure efficiency via the iteration complexity $N_{\text{iter}}(\varepsilon)$, i.e., the number of algorithmic updates, and the sample complexity $N_{\text{sample}}(\varepsilon)$, i.e., the number of queries to a comparison oracle.

Value and best-response. Due to the symmetry of the game, it is enough to consider only the point of view of the min-player, the value of its policy $\pi \in \Pi$ against a competitor policy $\mu \in \Pi$ is

$$V_\beta(\pi;\mu) \triangleq \mathcal{P}(\mu \succ \pi) + \beta\,\mathrm{KL}_\rho(\pi\|\pi^{\text{ref}})\,. \tag{2}$$

Thus, we can define a best-response policy $\nu^\star_\beta(\mu) \in \arg\min_{\pi\in\Pi} V_\beta(\pi;\mu)$ and its value as $V^\star_\beta(\mu)$. We can rewrite the worst-case suboptimality in terms of value as $\mathrm{SubOpt}_\beta(\pi) = V_\beta(\pi;\pi) - V^\star_\beta(\pi)$.

Additional notations. For a vector $x \in \mathbb{R}^d$ we define a span seminorm of x as $\|x\|_{\text{sp}} = \inf_{c\in\mathbb{R}}\|x + c\mathbf{1}\|_\infty$, where $\mathbf{1} = (1,\dots,1)^\intercal$ is a vector of all ones. For $M > 0$ define $\mathrm{clip}_{[-M,M]}(x) = \max\{-M, \min\{x, M\}\}$. For a distribution $\rho \in \Delta_{\mathcal{X}}$ and two functions $f\colon \mathcal{X} \to \mathbb{R}^d$, define $\|f\|^2_{1,\rho} \triangleq \mathbb{E}_{x\sim\rho}[\|f(x)\|_1^2]$ and $\|f\|^2_{\text{sp},\rho} \triangleq \mathbb{E}_{x\sim\rho}[\|f(x)\|^2_{\text{sp}}]$.

4 Self-Play Policy Gradients

In this section, we consider a simple algorithm for finding a VNW of a β-regularized preference game with a general parameterized policy class and a finite action space. We call this algorithm the Self-Play Policy Gradient (SPG) method.

For a parameter $\theta \in \Theta = \mathbb{R}^d$, we define $\theta \mapsto \pi_\theta \in \Pi$ as a differentiable policy parameterization. Define a parametrized value as $J_\beta(\theta;\mu) \triangleq V_\beta(\pi_\theta;\mu)$ where μ is a competitor policy. Given this definition and an initial parameter $\theta_0 \in \Theta$, we define the iterates of SPG as follows

$$\theta_{t+1} = \mathcal{T}(\theta_t - \gamma_t g_t), \qquad \pi_{t+1} = \pi_{\theta_{t+1}}\,, \tag{3}$$

where g_t is a stochastic gradient estimator of $\nabla J_\beta(\theta_t;\pi_t)$, and $\mathcal{T}$ is an optional projection-like policy improvement operator (e.g., clipping). As a main example, we consider the following mini-batch pairwise REINFORCE gradient estimator (Williams, 1992), defined as $g_t = (1/B_t)\sum_{j=1}^{B_t} G_j(\theta_t)$ for

$$G_j(\theta) \triangleq (\nabla_\theta \log\pi_\theta(y_j|x_j) - \nabla_\theta \log\pi_\theta(y'_j|x_j)) \cdot \mathrm{clip}_{[-M,M]}(A_j)\,, \tag{4}$$

where M is a clipping threshold and A_j is an advantage estimate defined as

$$A_j \triangleq \tfrac{1}{2} - p_j + \beta\Big(\log\tfrac{\pi_\theta(y|x)}{\pi^{\text{ref}}(y|x)} - \log\tfrac{\pi_\theta(y'|x)}{\pi^{\text{ref}}(y'|x)}\Big)\,, \tag{5}$$

and where $x_j \sim \rho$, $y_j, y'_j \sim \pi_t(x_j)$ and p_j is a sample from a Bernoulli distribution of parameter $\mathcal{P}(y_j \succ y'_j|x_j)$. This gradient estimator can be considered as a leave-one-out (LOO) advantage estimator (Kool et al., 2019; Ahmadian et al., 2024) with a group of size 2, which was also applied to estimate the KL-divergence.

Connection with Online IPO. Recall the Online IPO loss

$$\mathcal{L}_{\texttt{IPO}}(\theta) \triangleq \mathbb{E}_{x\sim\rho,y,y'\sim\texttt{sg}(\pi_\theta(\cdot|x)),p\sim\mathcal{B}\text{er}(\mathcal{P}(y\succ y'|x))}\left[\left(\log\left(\frac{\pi_\theta(y|x)}{\pi_\theta(y'|x)}\frac{\pi^{\text{ref}}(y'|x)}{\pi^{\text{ref}}(y|x)}\right) - \frac{p}{2\beta}\right)^2\right], \tag{6}$$

where sg as a stop-gradient operation. It is known (Calandriello et al., 2024, Proposition 4.2) that expected gradients of Online IPO loss are the same of the self-play updates. However, it is easy to check the sample-based version of (6) have exactly the same (stochastic) gradients as a pairwise REINFORCE estimator (4) used in our analysis, up to a constant scaling and clipping. In particular, it allows us to extend all our theoretical results to a parametrized version of online IPO, and also directly interpret Online IPO as a LOO-variance-reduced self-play policy gradient update with a group size of 2.

4.1 Theoretical Guarantees

We emphasize that the primary challenge in the parametrized setting lies in the intrinsic non-convexity of the value $J_\beta(\theta;\mu)$ with respect to its first argument. Consequently, in the absence of additional structural assumptions, global convergence of the method cannot be guaranteed.

Assumption 1 (Parametrization regularity, informal)**.** A pair $(\theta \mapsto \pi_\theta, \mathcal{T})$ satisfies the following properties: **(A1)** Lipschitz parametrisation $\|\pi_\theta - \pi_{\theta'}\|_{1,\rho} \le G\|\theta - \theta'\|_2$; **(A2)** L-smoothness of $\theta \mapsto J_\beta(\theta;\mu)$; **(A3)** (approx.) Polyak–Łojasiewicz (PL) inequality: $\|\nabla J_\beta(\theta;\pi_\theta)\|_2^2 + \varepsilon_{\mathrm{PL}} \ge 2m\,\mathrm{SubOpt}_\beta(\pi_\theta)$ for θ in the range of $\mathcal{T}$; **(A4)** $\mathcal{T}$ non-increase suboptimality; **(A5)** g_t has a bias $\le \varepsilon_{\mathrm{grad}}$ and is subgaussian with a variance-proxy $\propto M^2/B_t$, where M is a clipping threshold and B_t is a batch size.

These conditions hold for context-free softmax policies $\pi_\theta(y) \propto \exp(\theta_y)$ with a suitable choice of $\mathcal{T}$, and more generally for Fisher-nondegenerate compatible parameterizations (Yuan et al., 2022), where $\varepsilon_{\mathrm{PL}}$ reflects function-approximation error. Appendix C provides the full assumptions and verification. The constants may depend on β and the reference policy π^{ref}; in particular, the clipping level M depends on the bias $\varepsilon_{\mathrm{grad}}$ and regularity of π^{ref}.

Theorem 1 (Convergence guarantees of SPG, informal)**.** *Assume **(A1)–(A5)** and $\beta m \ge G^2$. Let $\kappa \triangleq L/m$. For the iterates of self-play policy gradients (SPG, 3):*

***Deterministic:** with exact gradients and $\gamma_t \equiv 1/(2L)$,*

$$\mathrm{SubOpt}_\beta(\pi_t) \le (1-\gamma m/2)^t\ \mathrm{SubOpt}_\beta(\pi_0) + \mathcal{O}(\varepsilon_{\mathrm{PL}}/m).$$

***Stochastic:** with $\gamma_t = \Theta(1/(mt))$ and a growing batch size $B_t = \Theta(t/m)$, with probability at least $1-\delta$,*

$$\mathrm{SubOpt}_\beta(\pi_t) = \widetilde{\mathcal{O}}\Big(\kappa^2/t^2 + \kappa M^2/t^2 + G^2M^2/(\beta m\cdot t) + (\varepsilon_{\mathrm{PL}} + \varepsilon_{\mathrm{grad}}^2)/m\Big)\,.$$

We refer to Propositions 3 and 4 in Appendix C for complete statements and proofs. In the bound above, for the stochastic case, the first term corresponds to deterministic convergence, and the next two terms correspond to different noise terms: the first comes from standard PL-SGD arguments (Madden et al., 2024), whereas the second comes from a *moving best-response*. The last term corresponds to an accumulated bias from an inexact PL inequality or biased gradient updates. Before discussing the assumptions, we demonstrate the source of the $1/(\beta m\cdot t)$-term (moving best-response) in the stochastic bound.

Lemma 1 (Descent Lemma I)**.** *Assume **(A1)**,**(A2)**, and **(A4)**. Let $\tilde{L} \triangleq L + G^2/(4\beta)$. For the iterates of self-play policy gradients (3), for any sequence $(\gamma_t)_{t\ge 0}$ and for any $t \ge 0$, it holds that*

$$\mathrm{SubOpt}_\beta(\pi_{t+1}) \le \sqrt{\mathrm{SubOpt}_\beta(\pi_t) - \gamma_t\big\langle \nabla J_\beta(\theta_t;\pi_t), g_t\big\rangle + \tfrac{\gamma_t^2\tilde{L}}{2}\|g_t\|_2^2} + \sqrt{\tfrac{\gamma_t^2 G^2\|g_t\|_2^2}{8\beta}}\,.$$

We refer to Lemma 7 in Appendix C.1 for a complete proof. In the following, we provide a proof sketch to demonstrate where the influence of moving baselines appears. In fact, the additional term $\sqrt{\gamma_t^2\|g_t\|_2^2/\beta}$ is exactly the reason to introduce assumptions on β and a growing batch size.

Sketch of the proof. For simplicity of exposition, assume $\mathcal{T} \equiv \mathrm{Id}$. Then the smoothness condition **(A2)** implies for $f_t(\theta) \equiv J_\beta(\theta;\pi_t) \equiv V_\beta(\pi_\theta;\pi_t)$ and $\theta_{t+1} = \theta_t - \gamma_t g_t$

$$f_t(\theta_{t+1}) \le f_t(\theta_t) - \gamma_t\langle \nabla f_t(\theta_t), g_t\rangle + \tfrac{\gamma_t^2 L}{2}\|g_t\|_2^2\,.$$

Define $f_t^\star = V_\beta^\star(\pi_t)$, and using a relation $S_t^2 = f_t(\theta_t) - f_t^\star$ and simple algebraic manipulations

$$S_{t+1}^2 \le S_t^2 - \gamma_t\langle \nabla f_t(\theta_t), g_t\rangle + \tfrac{\gamma_t^2 L}{2}\|g_t\|_2^2 + \underbrace{[f_{t+1}(\theta_{t+1}) - f_t(\theta_{t+1})] + [f_t^\star - f_{t+1}^\star]}_{:=\mathcal{R}_t}\,.$$

For the term $\mathcal{R}_t$, we first notice that in the difference $f_{t+1}(\theta_{t+1}) - f_t(\theta_{t+1})$ the KL-terms cancels out, thus we have $f_{t+1}(\theta_{t+1}) - f_t(\theta_{t+1}) = \mathcal{P}(\pi_{t+1} \succ \pi_{t+1}) - \mathcal{P}(\pi_t \succ \pi_{t+1})$, which we denote as $\mathcal{P}(\pi_{t+1} - \pi_t \succ \pi_{t+1})$ using the linearity of expectation. For the final term, we use a connection between $f_t^\star \equiv V_\beta^\star(\pi_t)$ and a convex conjugate of entropy, thus $f_{t+1}^\star - f_t^\star \ge \mathcal{P}(\pi_{t+1} - \pi_t \succ \nu_{t+1})$, where ν_{t+1} is a best response against π_{t+1}. As a result, using bilinearity and bi-Lipschitzness of preferences

$$\mathcal{R}_t \le \mathcal{P}(\pi_{t+1} - \pi_t \succ \pi_{t+1} - \nu_{t+1}) \le \tfrac{1}{2}\|\pi_{t+1} - \pi_t\|_{1,\rho}\|\nu_{t+1}^\star - \pi_{t+1}\|_{1,\rho}\,.$$

Finally, using Assumption **(A1)**, Pinsker's inequality, and the connection $\beta\,\mathrm{KL}_\rho(\pi_{t+1}\|\nu_{t+1}^\star) = S_{t+1}^2$, we have $\mathcal{R}_t \le (\gamma_t/2)\sqrt{G^2\|g_t\|_2^2\cdot S_{t+1}^2/\beta}$. Plugging this inequality into the inequality connecting S_{t+1}^2 and S_t^2 and solving the resulting quadratic inequality, we conclude the statement. □

Why growing batches? Unlike standard analysis of SGD under the PL-assumption on a fixed objective, the exploitability $\mathrm{SubOpt}_\beta(\pi_t) = V_\beta(\pi_t;\pi_t) - V_\beta^\star(\pi_t)$ uses a *moving* best-response baseline. Controlling its drift yields an intrinsic noise term of order $\gamma_t\|\xi_t\|_2^2$, which motivates the need of a batch size B_t increasing with t. We also note that an implicit assumption about growing batch sizes appears in the existing analysis of the extragradient method (Zhou et al., 2025), where the noise term is scaled by σ^2/β^2, with σ^2 denoting the gradient variance. As remarked by Madden et al. (2024), similar issues arise in the analysis of projected gradient descent in the non-convex setting, and are typically resolved by a similar growing mini-batch condition (Ghadimi et al., 2016) or by variance-reduction techniques (J. Reddi et al., 2016).

Why $\beta m \geq G^2$? This condition ensures the best-response map is well-contractive and one-step improvement is achievable; similar "sufficient regularization" assumptions also appear in Mirror-Prox-style analyses, see Nemirovski (2004) for a discussion. We note that in the convex case (Sokota et al., 2023), it was shown that a similar algorithm does not require such assumption; however, in a parametrized non-convex setting, it is unclear whether this is an artifact of the analysis or a necessity for the method's stability. We note that empirically the method may not achieves monotonic improvements in suboptimality over the course of training in the case of small β, making the last-iterate convergence non-achievable; see Appendix E.1 for numerical experiments. We leave the complete identification of the gap between these two settings as a direction for future work.

For the tabular softmax and Fisher-compatible parameterizations, the following corollary summarizes the resulting iteration and sample complexity. The rates depend on the regularity of the reference policy under ρ: it controls the tails of the stochastic gradient estimator and hence the clipping level M. This yields two regimes: the context-free and the contextual settings.

Corollary 1 (SPG iteration and sample complexity, informal)**.** *Fix confidence level $\delta \in (0,1)$ and target accuracy $\varepsilon \in (0,1)$. Assume $\mathbb{E}_{x\sim\rho}\big[\log^2\big(1/\pi^{\mathrm{ref}}_{\min}(x)\big)\big] < \infty$, where $\pi^{\mathrm{ref}}_{\min}(x) \triangleq \min_{y\in\mathcal{Y}} \pi^{\mathrm{ref}}(y \mid x)$. Then SPG returns an $\mathcal{O}(\varepsilon + \varepsilon_{\mathrm{PL}})$-VNW for the β-regularized game in the following settings:*

(i) Context-free, tabular softmax *(Mei et al., 2020). With clipping $M = \Theta(\log(1/\varepsilon))$ and $\beta \gtrsim \beta_{\min}$ (where $\beta_{\min}$ depends on the minimum reference mass),*

$$N_{\mathrm{iter}}(\varepsilon) = T = \widetilde{\mathcal{O}}\big(\varepsilon^{-1}\big), \qquad N_{\mathrm{sample}}(\varepsilon) = \widetilde{\mathcal{O}}\big(\varepsilon^{-2}\big).$$

(ii) Contextual, Fisher non-degenerate compatible parameterization *(Yuan et al., 2022). With clipping $M = \Theta(1/\sqrt{\varepsilon})$ and $\beta \gtrsim \beta_{\min}$ (where $\beta_{\min}$ depends on the Fisher conditioning),*

$$N_{\mathrm{iter}}(\varepsilon) = T = \widetilde{\mathcal{O}}\big(\varepsilon^{-2}\big), \qquad N_{\mathrm{sample}}(\varepsilon) = \widetilde{\mathcal{O}}\big(\varepsilon^{-4}\big).$$

We refer to Corollary 2 and 4 in Appendix C for proofs and complete statements. We also notice that under a stronger condition $\mathbb{E}_{x\sim\rho}[1/\pi^{\mathrm{ref}}_{\min}(x)] < \infty$, one can use a milder clipping level and recover the same improved rates in both the contextual and context-free cases. For general context spaces, however, $\mathbb{E}[1/\pi^{\mathrm{ref}}_{\min}(x)] < \infty$ can be substantially more restrictive than the log-moment condition.

5 Proximal Point Method with Self-Play Policy Gradients

In this section, we propose a way to overcome a theoretical limitation of the self-play policy-gradient (SPG) method by embedding it in the Proximal Point (PP) method.

5.1 (Approximate) Proximal Point Method

A well-known method for computing a NE is the Proximal Point (PP) method (Martinet, 1970; Rockafellar, 1976). At each iteration, the PP method computes the NE of an auxiliary game that is additionally regularized toward the previous iterate. To formally define the iterates $(\pi_k)_{k\geq 0}$, we initialize $\pi_0 \equiv \pi^{\mathrm{ref}}$, and for a PP step size $\eta > 0$ we define π_{k+1} as a solution to the following game:

$$\max_{\pi\in\Pi} \min_{\mu\in\Pi} \Big\{ \mathcal{P}_\beta(\pi \succ \mu) - (\beta/\eta)\,\mathrm{KL}_\rho(\pi\|\pi_k) + (\beta/\eta)\,\mathrm{KL}_\rho(\mu\|\pi_k) \Big\}, \tag{7}$$

which is equivalent to the original game, up to additional regularization toward the previous iterate π_k. Alternatively, we can view π_{k+1} as a fixed point of the best-response operator:

$$\pi_{k+1} \in \arg\min_{\pi\in\Pi} V_k(\pi;\pi_{k+1}), \qquad V_k(\pi;\mu) \triangleq V_\beta(\pi;\mu) + (\beta/\eta)\,\mathrm{KL}_\rho(\pi\|\pi_k). \tag{8}$$

This approach often yields accelerated convergence rates for solving games, but it is primarily a *conceptual* method rather than a *practical* one, since exactly computing the NE of the regularized game can be just as challenging as solving the original game. The key observation is that the PP subproblems in (8) induce stronger effective regularization than the original game (when η is small), which allows us to address the theoretical limitation of self-play policy gradients. This motivation is similar to the motivation behind the Mirror-Prox algorithm (Nemirovski, 2004), which efficiently approximates a PP iteration by performing two best-response steps for a well-chosen value of η.

PP residual and convergence guarantees. As discussed earlier, computing (8) exactly is challenging, especially in the function-approximation setting. The exact iterates can equivalently be characterized by $\nabla_\pi V_k(\pi_{k+1}, \pi_{k+1}) \propto \mathbf{1}$, i.e., the gradient vector becomes constant. This motivates the following definition: we call π_{k+1} an $\varepsilon_{\mathrm{PP}}$-approximate PP update if

$$\|\nabla_\pi V_k(\pi_{k+1}, \pi_{k+1})\|_{\mathrm{sp},\rho}^2 \le \varepsilon_{\mathrm{PP}}\,, \tag{9}$$

where $\|\cdot\|_{\mathrm{sp},\rho}^2$ is an expected (over contexts) span seminorm of a vector; by definition, it equals zero on constant vectors. Under this definition, it is possible to derive the following guarantees; for the full statement and proof, see Appendix B.

Proposition 1 (Approximate PP convergence)**.** Assume that (9) holds for all $k \ge 0$. Then the iterates of approximate PP converge toward $\pi_\beta^\star$ up to a residual floor:

$$\mathrm{KL}_\rho(\pi_\beta^\star \| \pi_k) \le (1+\eta/2)^{-k}\ \mathrm{KL}_\rho(\pi_\beta^\star \| \pi_0) + 2\varepsilon_{\mathrm{PP}}/\beta^2\,, \tag{10}$$

and similarly $\mathrm{SubOpt}_\beta(\pi_k)$ and $\|\log \pi_k - \log \pi_\beta^\star\|_{\mathrm{sp},\rho}$ decays at the rate $(1+\eta/2)^{-k}$ with an additive $\mathcal{O}(\varepsilon_{\mathrm{PP}})$ term (see Proposition 2 in the Appendix B for explicit constants).

Choice of η. Larger η improves the outer contraction factor in (10), but it also weakens the proximal regularizer β/η and can make the inner games harder to solve. This is the same trade-off that motivates the choice $\eta = \mathcal{O}(\beta)$ in Mirror-Prox-style algorithms for this problem (Cen et al., 2024), which resolves the inner problem using two iterations of the best-response operator.

5.2 PP–SPG Method

We now describe how we implement the inexact PP step (7) using SPG, and summarize the resulting convergence guarantees.

Outer–inner structure. We set $\pi_0 = \pi^{\mathrm{ref}}$ and run K outer PP steps. At outer iteration k, we fix π_k and (approximately) solve the proximal subgame (7) by running SPG for T_k inner steps. Denote

$$\beta_{\mathrm{target}} \triangleq \beta/\eta, \qquad \lambda \triangleq \beta + \beta_{\mathrm{target}} = \beta\big(1+\tfrac{1}{\eta}\big), \qquad \tilde{\pi}_k \propto [\pi^{\mathrm{ref}}]^{\beta/\lambda} \times [\pi_k]^{\beta_{\mathrm{target}}/\lambda}.$$

The inner problem is exactly a λ-regularized preference game anchored at $\tilde{\pi}_k$, i.e., it is of the same form as the game analyzed for SPG in Section 4.

Inner objective for SPG. Let π_θ be a parametrized policy. The min-player objective against policy π at the outer step k is

$$J_k(\theta; \pi) \triangleq \mathcal{P}(\pi \succ \pi_\theta) + \beta\ \mathrm{KL}_\rho(\pi_\theta \| \pi^{\mathrm{ref}}) + \beta_{\mathrm{target}}\ \mathrm{KL}_\rho(\pi_\theta \| \pi_k). \tag{11}$$

Self-play corresponds to using $\pi = \pi_\theta$ in J_k (as in Section 4), and updating parameters with stochastic gradients and an appropriately chosen improvement operator $\mathcal{T}_k : \Theta \to \Theta$ to ensure regularity:

$$\theta_{k,t+1} = \mathcal{T}_k(\theta_{k,t} - \gamma_{k,t}\, g_{k,t}), \qquad \pi_{k,t+1} = \pi_{\theta_{k,t+1}}. \tag{12}$$

At the end of the inner loop, we set $\pi_{k+1} \triangleq \pi_{k,T_k}$. A concrete example of $g_{k,t}$ is the clipped pairwise REINFORCE estimator as in Section 4, see Algorithm 1 in Appendix D for a formal definition and the pseudo-code.

Connection to COMAL (Liu et al., 2024b). Our method is closely related in spirit to COMAL, which is motivated as a practical implementation of the conceptual prox method for *unregularized* preference games. COMAL proceeds by repeatedly solving a KL-regularized subgame centered at a reference policy, and then updating the reference to the newly computed iterate. In contrast, our goal is to compute the Nash equilibrium of the *fixed* β-regularized game (anchored at π_{ref}), and we apply an (inexact) proximal-point outer loop *on top of* this baseline regularization. Concretely, each PP outer step adds an *additional* KL term toward the previous iterate, which strengthens the effective regularization of the inner problem and yields an improved contraction guarantee for the outer loop. This additional proximal regularizer allows us to overcome the limitation of plain self-play policy gradients in our setting and derive end-to-end convergence rates.

5.3 Theoretical Guarantees

Approximate PP requires a bound on the residual (9), which is a gradient in *policy space*. Our inner solver operates in *parameter space*. We bridge this gap via a gradient-compatibility condition (Assumption 11 in Appendix D.2) showing that, up to constants,

$$\|\nabla_\pi V_k(\pi_\theta, \pi_\theta)\|_{\mathrm{sp},\rho}^2 \lesssim \|\nabla_\theta J_k(\theta; \pi_\theta)\|_2^2 .$$

Notably, this assumption holds under both tabular softmax and Fisher-compatible parameterizations. Combined with the smoothness of J_k, this yields the key implication (Lemma 24): if the inner loop returns π_{k+1} with $\mathrm{SubOpt}_\lambda^{\widetilde{\pi}_k}(\pi_{k+1}) \le \varepsilon_{\mathrm{in}}$, then

$$\|\nabla_\pi V_k(\pi_{k+1}, \pi_{k+1})\|_{\mathrm{sp},\rho}^2 \le \varepsilon_{\mathrm{PP}} \qquad \text{with} \qquad \varepsilon_{\mathrm{PP}} = \mathcal{O}(\varepsilon_{\mathrm{in}}) \ \text{(explicit in Appendix D)}. \tag{13}$$

Therefore, controlling the *inner* SPG suboptimality suffices to control the *outer* PP residual, which allows us to provide end-to-end guarantees for the combined method. The main caveat in the analysis is that, because our anchor sequence $(\widetilde{\pi}_k)_{k\ge 0}$ depends on the previous iterate π_k, we cannot uniformly control its behavior over all contexts, but we still can control the second moment thanks to established convergence in $\|\log \pi_k - \log \pi_\beta^\star\|_{\mathrm{sp},\rho}$.

Theorem 2 (PP–SPG sample complexity, informal)**.** *Fix arbitrary $\beta \le 1$, confidence $\delta \in (0,1)$, and target accuracy $\varepsilon \in (0,1)$. Run PP–SPG for $K = \widetilde{\mathcal{O}}(1)$ outer steps, and set the inner target accuracy $\varepsilon_{\mathrm{in}} = \Theta(\varepsilon)$. Then, with probability at least $1-\delta$, the output π_K satisfies $\mathrm{SubOpt}_\beta(\pi_K) \le \mathcal{O}(\varepsilon + \varepsilon_{\mathrm{PL}})$. Moreover, the combined method achieves the same order of total iteration and sample complexity as in Corollary 1.*

See Corollaries 7 and 8 in Appendix D.2 for full statements and proofs. The contextual rate improves to $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ under stronger assumptions on $\mathcal{T}$ and π^{ref}. With PP alone, we control only a second log-moment, so gradients may be heavy-tailed. SVRG-style variance reduction could further reduce sampling complexity (J. Reddi et al., 2016); we leave this direction to future work.

5.4 Nash Prox: Practical Deep Learning Implementation

Next, we provide a practical implementation, which we call `Nash Prox`. We consider the following policies: an online policy π_t with parameters θ_t, a *target policy* π_t^{target} with parameters $\theta_t^{\mathrm{target}}$, and a fixed reference policy π^{ref}. Next, we define the following IPO-style loss function (cf. (6))

$$\mathcal{L}_{\texttt{Prox}}(\theta; \pi^{\mathrm{target}}, \pi^{\mathrm{ref}}) \triangleq \mathbb{E}_{x\sim\rho,\, y,y'\sim \mathtt{sg}(\pi_\theta(\cdot|x))}\left[\left(\ell_\theta(x,y) - \ell_\theta(x,y') - \tfrac{p-1/2}{\beta+\beta_{\mathrm{target}}}\right)^2\right], \tag{14}$$

where $\ell_\theta(x,y) \triangleq (\beta/\lambda)\cdot \log(\pi_\theta(y|x)/\pi^{\mathrm{ref}}(y|x)) + (\beta_{\mathrm{target}}/\lambda)\cdot \log(\pi_\theta(y|x)/\pi^{\mathrm{target}}(y|x))$ for $\lambda \triangleq \beta + \beta_{\mathrm{target}}$, and p is an estimator of $\mathcal{P}(y \succ y' \mid x)$. As explained before, the stochastic gradient of this loss function corresponds to our method, but it is easier to implement in practice. To implement the algorithm exactly as we discussed in Section 5.2, we need to update the target policy with the weights of the online policy every T_k steps, where T_k defines the update schedule. However, following a very common strategy in deep reinforcement learning (Mnih et al., 2015; Lillicrap et al., 2016), we employ a more practical and elegant approach and update the target with an exponential moving average instead, resulting in the following updates:

$$\theta_{t+1} = \theta_t - \alpha_t \cdot \nabla_\theta \mathcal{L}_{\texttt{Prox}}(\theta_t; \pi_t^{\mathrm{target}}, \pi^{\mathrm{ref}}), \qquad \theta_{t+1}^{\mathrm{target}} = (1-\kappa)\cdot \theta_t^{\mathrm{target}} + \kappa \cdot \theta_{t+1},$$

where α_t is a learning rate and the parameter $\kappa \in [0,1]$ implicitly controls the number of steps for one proximal update. Thus, we approximate the solution of one proximal subproblem with $T_k \approx 1/\kappa$ gradient steps. Also, we found it empirically beneficial to anneal the value of κ as $\kappa_t = 1/(c\cdot t + 1)$ for some value of c, where $K \approx 1/c$ corresponds to the approximate number of outer PP iterates.

6 Experiments

We use the simple contextual dueling bandit problem as an initial experiment to study the deep learning implementation of `Nash Prox`. We refer to Appendix E.1 for an additional small-scale experiment on the Rock-Paper-Scissors game.

Game definition. Let us fix a number of actions $Y \geq 2$ and a positive integer $r \geq 1$. We consider a dueling bandit game with a context space $\mathcal{X} = \mathbb{R}^{r \times r}$ and an action space $\mathcal{Y} = \{1, \dots, Y\}$. Preference probabilities are defined as follows

$$\mathcal{P}(y \succ y'|x) \triangleq \sigma(A_{y,y'} - A_{y',y}) \,,$$

for $A \triangleq U\Theta_x V^\top$, where $U \in \mathbb{R}^{Y \times r}$ and $V \in \mathbb{R}^{Y \times r}$ are fixed matrices, $\Theta_x \in \mathbb{R}^{r \times r}$ is a corresponding context matrix, and $\sigma(\cdot)$ is a sigmoid function. This type of dueling bandit instance is a generalization of a low-rank linear bandit problem. Notice that for any $r \geq 2$ this problem does not admit a Bradley-Terry model. The distribution over contexts ρ is assumed to be a standard Gaussian random matrix (i.e., elements of Θ_x are i.i.d. with distribution $\mathcal{N}(0, 1)$). We aim to find a policy $\pi \colon \mathcal{X} \to \Delta_{\mathcal{Y}}$ that approximates a β-regularized VNW. We refer to Appendix E for more details on the setup.

Results. We compare our method with the following baselines: Online IPO (Calandriello et al., 2024), Online DPO (Guo et al., 2024), EGPO (Zhou et al., 2025), Nash MD (Munos et al., 2023), and `Nash Prox`. The results are presented in Figure 1. We can observe that `Nash Prox` with an adaptive $\kappa_t = 1/(0.3 \cdot t + 1)$ and $\beta_{\text{target}} = 10 \times \beta$ outperforms all the baselines. We also observe that the two-step stabilization procedure of EGPO is insufficient to stabilize training in the function approximation setting, while Online DPO diverges because it is not designed to solve the underlying preference game in the first place. Additionally, we observe non-monotonous behavior of suboptimality for all the baselines, including online IPO. *This empirical fact is a direct demonstration of a necessity of the stability condition on β for last-iterate convergence guarantees, which is typically relies on contraction principle.* In Appendix E, we provide an additional ablation study on the choice of κ on the optimization procedure as well as a comparison between hard and soft updates.

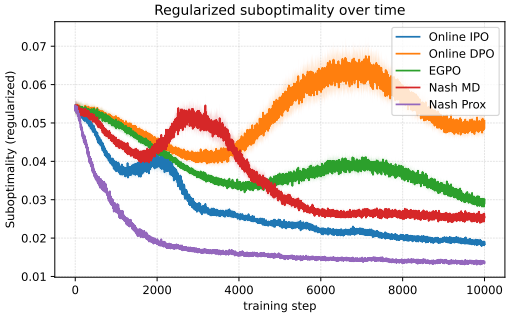


Figure 1: Comparison of `Nash Prox` (in violet) with baseline methods. Suboptimality is averaged over 25 random seeds; shaded regions indicate standard error.

6.1 LLM Alignment

Experiment setup. For our LLM-based experiments, we use the Gemma-3-4B (Team et al., 2025) pretrained model checkpoints and train on the RLHFlow (Dong et al., 2024) datasets for all the analysis. In particular, we first perform SFT on the RLHFlow SFT dataset (RLHFlow Team, 2024c) and then all our NLHF experiments use LoRA adapters (Hu et al., 2022) of rank 16 with $\alpha = 32$ on the resulting checkpoint, using a subset of RLHFlow Prompt collection (RLHFlow Team, 2024b). The pairwise judge model is Gemma-2-2B trained via the Robust Reward Models (Liu et al., 2025) method. All the experiments were performed using the TRL library (von Werra et al., 2020).

Baselines. We compare the practical version of `Nash Prox` against Online DPO (Guo et al., 2024) and Online IPO (Calandriello et al., 2024). We also considered `NashMD` (Munos et al., 2023) and EGPO (Zhou et al., 2025), but did not include them in our experiments because their current implementations are prohibitively slow in our setting (we estimate on the order of 10^3 GPU-hours per configuration). For `NashMD`, the main bottleneck is the need to sample from geometric mixtures, while for EGPO the reference implementation incurs substantial overhead due to repeated optimizer off-loading and reloading. Additional details on baselines and hyperparameters are provided in Appendix E.

Results. We report in Table 1 the pairwise win-rates between the different methods on a held-out test set of prompts for the separate Gemma3-27B-IT judge model. We observe that `Nash Prox` outperforms all the baselines. From the implementation side, the main difference between `Nash Prox` and Online IPO is the use of additional regularization with respect to a target model, and our results show that this regularization can be valuable.

Table 1: Pairwise win rates (mean $\pm$ 99.9%-confidence intervals). Statistically significant wins are in **bold**. Confidence intervals are in a smaller font size. Row/column for `Nash Prox` is highlighted.

Win rate	SFT	Online DPO	Online IPO	`Nash Prox`
SFT	–	$0.0674_{\pm 0.011}$	$0.0636_{\pm 0.011}$	$0.0567_{\pm 0.010}$
Online DPO	$\mathbf{0.9326}_{\pm 0.011}$	–	$\mathbf{0.5283}_{\pm 0.017}$	$0.4656_{\pm 0.017}$
Online IPO	$\mathbf{0.9364}_{\pm 0.011}$	$0.4717_{\pm 0.017}$	–	$0.4373_{\pm 0.017}$
`Nash Prox`	$\mathbf{0.9433}_{\pm 0.010}$	$\mathbf{0.5344}_{\pm 0.017}$	$\mathbf{0.5627}_{\pm 0.017}$	–

7 Conclusion

In this work, we study the problem of computing Nash equilibria in KL-regularized preference games that arise in NLHF, where the regularizer enforces proximity to a reference policy. We first analyze a simple self-play policy gradient method under general policy parameterizations and establish last-iterate, high-probability convergence guarantees. The analysis also highlights a structural difficulty of self-play, namely that the best-response baseline changes across iterations and may force a stringent effective-regularization condition. To remove this restriction, we embed self-play within a proximal point scheme, whose proximal subgames induce stronger regularization and yield last-iterate convergence without requiring a lower bound on the original regularization strength. Finally, we instantiate this framework with practical stochastic policy-gradient updates and EMA-based stabilization mechanisms, resulting in a concrete algorithm, `Nash Prox`, that matches state-of-the-art empirical performance while retaining guarantees in the parameterized setting.

Impact Statement

This paper aims to advance methods for aligning learning systems with human feedback without assuming transitive preferences. Potential societal impacts include improved safety and robustness of interactive AI systems, and risks include mis-specifying human preferences or reinforcing annotator biases.

Acknowledgements

The work of D.Tiapkin has been supported by the Paris Île-de-France Région in the framework of DIM AI4IDF. The work of E. Moulines has been funded by the European Union (ERC-2022-SYG-OCEAN-101071601). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. This project was provided with computing and storage resources by GENCI at IDRIS thanks to the grant 2025-AD011016276 on the supercomputer Jean Zay's A100 and H100 partitions.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98): 1–76, 2021.

Ahmadian, A., Cremer, C., Gallé, M., Fadaee, M., Kreutzer, J., Pietquin, O., Üstün, A., and Hooker, S. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T. B., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862, 2022. URL https://api.semanticscholar.org/CorpusID:248118878.

Bengs, V., Busa-Fekete, R., El Mesaoudi-Paul, A., and Hüllermeier, E. Preference-based online learning with dueling bandits: a survey. *J. Mach. Learn. Res.*, 22(1), jan 2021. ISSN 1532-4435.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Calandriello, D., Guo, Z. D., Munos, R., Rowland, M., Tang, Y., Avila Pires, B., Richemond, P. H., Le Lan, C., Valko, M., Liu, T., Joshi, R., Zheng, Z., and Piot, B. Human alignment of large language models through online preference optimisation. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5409–5435. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/calandriello24a.html.

Cen, S., Wei, Y., and Chi, Y. Fast policy extragradient methods for competitive games with entropy regularization. *Journal of machine learning Research*, 25(4):1–48, 2024.

Choi, E., Ahmadian, A., Geist, M., Pietquin, O., and Gheshlaghi Azar, M. Self-improving robust preference optimization. In Yue, Y., Garg, A., Peng, N., Sha, F., and Yu, R. (eds.), *International Conference on Learning Representations*, volume 2025, pp. 4047–4068, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/0bc795afae289ed465a65a3b4b1f4eb7-Paper-Conference.pdf.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Dao, T. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

Ding, Y., Zhang, J., and Lavaei, J. On the global optimum convergence of momentum-based policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pp. 1910–1934. PMLR, 2022.

Dong, H., Xiong, W., Pang, B., Wang, H., Zhao, H., Zhou, Y., Jiang, N., Sahoo, D., Xiong, C., and Zhang, T. RLHF workflow: From reward modeling to online RLHF. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a13aYUU9eU.

Dudík, M., Hofmann, K., Schapire, R. E., Slivkins, A., and Zoghi, M. Contextual dueling bandits. In *Conference on Learning Theory*, pp. 563–587. PMLR, 2015.

Fatkhullin, I., Barakat, A., Kireeva, A., and He, N. Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9827–9869. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/fatkhullin23a.html.

Gardner, M. Mathematical games: The paradox of the nontransitive dice and the elusive principle of indifference. *Scientific American*, 223(12):110–114, 1970.

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305, 2016.

Gheshlaghi Azar, M., Daniel Guo, Z., Piot, B., Munos, R., Rowland, M., Valko, M., and Calandriello, D. A general theoretical paradigm to understand learning from human preferences. In Dasgupta, S., Mandt, S., and Li, Y. (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Guo, S., Zhang, B., Liu, T., Liu, T., Khalman, M., Llinares, F., Rame, A., Mesnard, T., Zhao, Y., Piot, B., et al. Direct language model alignment from online AI feedback. *arXiv preprint arXiv:2402.04792*, 2024.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Huang, F., Gao, S., Pei, J., and Huang, H. Momentum-based policy gradient methods. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4422–4433. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/huang20a.html.

J. Reddi, S., Sra, S., Poczos, B., and Smola, A. J. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/291597a100aadd814d197af4f4bab3a7-Paper.pdf.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

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

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *ICLR (Poster)*, 2015. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2015.html#KingmaB14.

Klimenko, A. Y. Intransitivity in theory and in the real world. *Entropy*, 17(6):4364–4412, 2015. ISSN 1099-4300. doi: 10.3390/e17064364. URL https://www.mdpi.com/1099-4300/17/6/4364.

Kool, W., van Hoof, H., and Welling, M. Buy 4 reinforce samples, get a baseline for free! 2019.

Kreweras, G. Aggregation of preference orderings. *Mathematics and Social Sciences I: Proceedings of the seminars of Menthon-Saint-Bernard, France (1–27 July 1960) and of Gösing, Austria (3–27 July 1962)*, pp. 73–79, 1965.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Labbi, S., Mangold, P., Tiapkin, D., and Moulines, E. On global convergence rates for federated policy gradient under heterogeneous environment. *arXiv preprint arXiv:2505.23459*, 2025.

Leahy, J.-M., Kerimkulov, B., Siska, D., and Szpruch, L. Convergence of policy gradient for entropy regularized mdps with neural network approximation in the mean-field regime. In *International Conference on Machine Learning*, pp. 12222–12252. PMLR, 2022.

Li, S., Zhao, Y., Varma, R., Salpekar, O., Noordhuis, P., Li, T., Paszke, A., Smith, J., Vaughan, B., Damania, P., et al. Pytorch distributed: Experiences on accelerating data parallel training. *arXiv preprint arXiv:2006.15704*, 2020.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1509.02971.

Liu, J., Li, W., and Wei, K. Elementary analysis of policy gradient methods. *arXiv preprint arXiv:2404.03372*, 2024a.

Liu, T., Xiong, W., Ren, J., Chen, L., Wu, J., Joshi, R., Gao, Y., Shen, J., Qin, Z., Yu, T., Sohn, D., Makarova, A., Liu, J. Z., Liu, Y., Piot, B., Ittycheriah, A., Kumar, A., and Saleh, M. RRM: Robust reward model training mitigates reward hacking. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=88AS5MQnmC.

Liu, Y., Zhang, K., Basar, T., and Yin, W. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33: 7624–7636, 2020.

Liu, Y., Oikonomou, A., Zheng, W., Cai, Y., and Cohan, A. Comal: A convergent meta-algorithm for aligning llms with general preferences, 2024b. URL https://arxiv.org/abs/2410.23223.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Lu, M., Aghaei, M., Raj, A., and Vaswani, S. Towards principled, practical policy gradient for bandits and tabular mdps. *arXiv preprint arXiv:2405.13136*, 2024.

Madden, L., Dall'Anese, E., and Becker, S. High probability convergence bounds for non-convex stochastic gradient descent with sub-weibull noise. *Journal of Machine Learning Research*, 25 (241):1–36, 2024.

Martinet, B. Brève communication. Régularisation d'inéquations variationnelles par approximations successives. *Revue française d'informatique et de recherche opérationnelle. Série rouge*, 4(R3): 154–158, 1970. URL https://www.numdam.org/item/M2AN_1970__4_3_154_0/.

May, K. O. Intransitivity, utility, and the aggregation of preference patterns. *Econometrica*, 22:1, 1954. URL https://api.semanticscholar.org/CorpusID:156169619.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. On the global convergence rates of softmax policy gradient methods. In *International conference on machine learning*, pp. 6820–6829. PMLR, 2020.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015. ISSN 00280836. URL http://dx.doi.org/10.1038/nature14236.

Munos, R., Valko, M., Calandriello, D., Azar, M. G., Rowland, M., Guo, D., Tang, Y., Geist, M., Mésnard, T., Michi, A., Selvi, M., Girgin, S., Momchev, N., Bachem, O., Mankowitz, D. J., Precup, D., and Piot, B. Nash learning from human feedback, 2023.

Nash Jr, J. F. Equilibrium points in N-person games. *Proceedings of the National Academy of Sciences of the United States of America*, 36(1):48–49, 1950.

Nemirovski, A. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Nesterov, Y. et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Papini, M., Binaghi, D., Canonaco, G., Pirotta, M., and Restelli, M. Stochastic variance-reduced policy gradient. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4026–4035. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/papini18a.html.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alch'e-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Pásztor, B., Buening, T. K., and Krause, A. Stackelberg learning from human feedback: Preference optimization as a sequential game. In *Eighteenth European Workshop on Reinforcement Learning*, 2025. URL https://openreview.net/forum?id=If5eE5hCB5.

Pinelis, I. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pp. 1679–1706, 1994.

Rivière, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., et al. Gemma 2: Improving open language models at a practical size. *CoRR*, 2024.

RLHFlow Team. Rlhflow pairwise preference dataset, 2024a. URL https://huggingface.co/datasets/RLHFlow/pair_preference_model_dataset.

RLHFlow Team. Rlhflow prompt collection, 2024b. URL https://huggingface.co/datasets/RLHFlow/prompt-collection-v0.1.

RLHFlow Team. Rlhflow-sft-dataset, 2024c. URL https://huggingface.co/datasets/RLHFlow/RLHFlow-SFT-Dataset-ver2.

Rockafellar, R. T. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

Schulman, J. and Lab, T. M. Lora without regret. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20250929. https://thinkingmachines.ai/blog/lora/.

Shani, L., Rosenberg, A., Cassel, A., Lang, O., Calandriello, D., Zipori, A., Noga, H., Keller, O., Piot, B., Szpektor, I., et al. Multi-turn reinforcement learning with preference human feedback. *Advances in Neural Information Processing Systems*, 37:118953–118993, 2025.

Sokota, S., D'Orazio, R., Kolter, J. Z., Loizou, N., Lanctot, M., Mitliagkas, I., Brown, N., and Kroer, C. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DpE5UYUQzZH.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Swamy, G., Dann, C., Kidambi, R., Wu, S., and Agarwal, A. A minimaximalist approach to reinforcement learning from human feedback. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47345–47377. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/swamy24a.html.

Tang, X., Yoon, S., Son, S., Yuan, H., Gu, Q., and Bogunovic, I. Rspo: Regularized self-play alignment of large language models. *arXiv preprint arXiv:2503.00030*, 2025.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Tversky, A. Intransitivity of preferences. *Psychological Review*, 76:31–48, 1969. URL https://api.semanticscholar.org/CorpusID:144609998.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

von Neumann, J. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928. ISSN 0025-5831; 1432-1807/e.

von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., Huang, S., Rasul, K., and Gallouédec, Q. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

Wang, M., Ma, C., Chen, Q., Meng, L., Han, Y., Xiao, J., Zhang, Z., Huo, J., Su, W. J., and Yang, Y. Magnetic preference optimization: Achieving last-iterate convergence for language model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PDnEDS244P.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

Wu, Y., Sun, Z., Yuan, H., Ji, K., Yang, Y., and Gu, Q. Self-play preference optimization for language model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=a3PmRgAB5T.

Wu, Y., Viano, L., Chen, Y., Zhu, Z., Antonakopoulos, K., Gu, Q., and Cevher, V. Multi-step alignment as markov games: An optimistic online gradient descent approach with convergence guarantees. *arXiv preprint arXiv:2502.12678*, 2025b.

Yuan, R., Gower, R. M., and Lazaric, A. A general sample complexity analysis of vanilla policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pp. 3332–3380. PMLR, 2022.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. The k-armed dueling bandits problem. *J. Comput. Syst. Sci.*, 78(5):1538–1556, sep 2012. ISSN 0022-0000. doi: 10.1016/j.jcss.2011.12.028. URL https://doi.org/10.1016/j.jcss.2011.12.028.

Zeng, S., Doan, T., and Romberg, J. Regularized gradient descent ascent for two-player zero-sum markov games. *Advances in Neural Information Processing Systems*, 35:34546–34558, 2022.

Zermelo, E. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29:436–460, 1929. URL http://eudml.org/doc/168081.

Zhang, J., Kim, J., O'Donoghue, B., and Boyd, S. Sample efficient reinforcement learning with reinforce. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10887–10895, May 2021. doi: 10.1609/aaai.v35i12.17300. URL https://ojs.aaai.org/index.php/AAAI/article/view/17300.

Zhang, Y., Yu, D., Ge, T., Song, L., Zeng, Z., Mi, H., Jiang, N., and Yu, D. Improving llm general preference alignment via optimistic online mirror descent. *arXiv preprint arXiv:2502.16852*, 2025a.

Zhang, Y., Yu, D., Peng, B., Song, L., Tian, Y., Huo, M., Jiang, N., Mi, H., and Yu, D. Iterative nash policy optimization: Aligning LLMs with general preferences via no-regret learning. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=Pujt3ADZgI.

Zhou, R., Fazel, M., and Du, S. S. Extragradient preference optimization (egpo): Beyond last-iterate convergence for nash learning from human feedback. *arXiv preprint arXiv:2503.08942*, 2025.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *ArXiv*, abs/1909.08593, 2019. URL https://api.semanticscholar.org/CorpusID:202660943.

Zoghi, M., Whiteson, S., Munos, R., and Rijke, M. Relative upper confidence bound for the k-armed dueling bandit problem. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 10–18, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/zoghi14.html.

A Notation

Symbol	Meaning
$\mathcal{X}, \mathcal{Y}$	Context space and finite action set; $\lvert\mathcal{Y}\rvert = Y$.
ρ	Distribution over contexts $x \in \mathcal{X}$.
π, μ	Policies $\mathcal{X} \to \Delta_{\mathcal{Y}}$; $\pi(y\lvert x)$ is action-probability.
$\mathcal{P}(y \succ y' \mid x)$	Preference probability; satisfies $\mathcal{P}(y \succ y'\lvert x) + \mathcal{P}(y' \succ y\lvert x) = 1$.
$\mathbf{P}_x$	Preference matrix at context x: $(\mathbf{P}_x)_{y,y'} = \mathcal{P}(y \succ y'\lvert x)$.
$\mathcal{P}(\pi \succ \mu \mid x)$	Contextual preference: $\pi(x)^\top \mathbf{P}_x \mu(x)$.
$\mathcal{P}(\pi \succ \mu)$	Overall preference: $\mathbb{E}_{x\sim\rho}[\pi(x)^\top \mathbf{P}_x \mu(x)]$.
$\langle f, g\rangle_\rho$	Inner product: $\mathbb{E}_{x\sim\rho}[\langle f(x), g(x)\rangle]$.
$\lVert\cdot\rVert_{1,\rho},\ \lVert\cdot\rVert_{\infty,\rho}$	$\sqrt{\mathbb{E}_{x\sim\rho}[\lVert\cdot\rVert_1^2]}$ and $\sqrt{\mathbb{E}_{x\sim\rho}[\lVert\cdot\rVert_\infty^2]}$.
$\lVert\cdot\rVert_{\mathrm{sp},\rho}$	Span seminorm in context: $\sqrt{\mathbb{E}_{x\sim\rho}[\lVert\cdot\rVert_{\mathrm{sp}}^2]}$.
$\mathrm{KL}_\rho(\pi\Vert\mu)$	Contextual KL: $\mathbb{E}_{x\sim\rho}[\mathrm{KL}(\pi(x)\Vert\mu(x))]$.
$\widetilde{\pi}$	Anchor/reference policy (generic).
$\mathcal{P}_\lambda^{\widetilde{\pi}}(\pi \succ \mu)$	λ-regularized payoff: $\mathcal{P}(\pi \succ \mu) - \lambda\,\mathrm{KL}_\rho(\pi\Vert\widetilde{\pi}) + \lambda\,\mathrm{KL}_\rho(\mu\Vert\widetilde{\pi})$.
$\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi, \mu)$	Regularized suboptimality: $\frac{1}{2} - \mathcal{P}_\lambda^{\widetilde{\pi}}(\pi \succ \mu)$.
$V_\lambda^{\widetilde{\pi}}(\pi, \mu)$	Regularized value: $\mathcal{P}(\mu \succ \pi) + \lambda\,\mathrm{KL}_\rho(\pi\Vert\widetilde{\pi})$.
$\nu_\lambda^{\widetilde{\pi},\star}(\mu)$	Best response: $\arg\min_\pi V_\lambda^{\widetilde{\pi}}(\pi, \mu)$.
$V_\lambda^{\widetilde{\pi},\star}(\mu)$	Best-response value: $\min_\pi V_\lambda^{\widetilde{\pi}}(\pi, \mu)$.
β	Regularization strength of the "outer/original" preference game.
η	Proximal-point (PP) step size; $\beta_{\mathrm{target}} = \beta/\eta$ is the proximal penalty.
π_t	Outer PP iterate (policy at PP step t).
$V_t(\pi, \mu)$	PP inner value: $V_\beta(\pi, \mu) + (\beta/\eta)\,\mathrm{KL}_\rho(\pi\Vert\pi_t)$.
ζ_{t+1}	PP residual: $\nabla_\pi V_t(\pi_{t+1}, \pi_{t+1})$.
ε	PP inner-solve accuracy target: $\lVert\zeta_{t+1}\rVert_{\mathrm{sp},\rho}^2 \le \varepsilon$.
θ, π_θ	Parametric policy: parameter $\theta \in \mathbb{R}^d$ with induced policy π_θ.
$J^{\widetilde{\pi}}(\theta; \pi)$	Best-response objective: $\mathcal{P}(\pi \succ \pi_\theta) + \lambda\,\mathrm{KL}_\rho(\pi_\theta\Vert\widetilde{\pi})$.
g_t, ξ_t, b_t, B_t	Stochastic gradient, noise, bias, and mini-batch size in SPG.
$m_{\widetilde{\pi},\lambda}, \varepsilon_{\mathrm{PL}}$	(Approx.) PL constant and additive PL slack in Assumption 4.

Table 2: Notation used in Appendix B–D.

The space of discrete probability measure of dimension d is denoted by $\Delta_d = \{p \in \mathbb{R}^d : p_i \ge 0, \sum_{i=1}^d p_i = 1\}$. In the same manner, we associate for any set $\mathcal{Y}$ of size $Y = \lvert\mathcal{Y}\rvert$ $\Delta_{\mathcal{Y}}$ with Δ_Y. Given two discrete probability measure $p, q \in \Delta_d$, let us define the KL-divergence between them as $\mathrm{KL}(p\Vert q) \triangleq \sum_{i=1}^d p_i \log(p_i/q_i)$. For two vectors $x, y \in \mathbb{R}^d$, we define an inner product as $\langle x, y\rangle \triangleq \sum_{i=1} x_i y_i$.

B Analysis of Proximal Point Method

In this section, we analyze the exact and approximate proximal point methods.

B.1 Preference Properties

We work in the contextual dueling bandit setting $(\mathcal{X}, \mathcal{Y}, \mathcal{P}, \rho)$ described in Section 3. The action space $\mathcal{Y}$ is finite with $\lvert\mathcal{Y}\rvert = Y \ge 2$, and in each round, a context $x \in \mathcal{X}$ is drawn from ρ. We assume that the context space is finite or countable for simplicity. We define a policy space Π as a functional space with a domain $\mathcal{X}$ and a codomain $\Delta_{\mathcal{Y}}$.

For every context $x \in \mathcal{X}$ and actions $y, y' \in \mathcal{Y}$, we are given preference probabilities $\mathcal{P}(y \succ y' \mid x) \in [0, 1]$ that satisfy the symmetry assumption

$$\mathcal{P}(y \succ y' \mid x) + \mathcal{P}(y' \succ y \mid x) = 1 \qquad \forall x \in \mathcal{X},\ y, y' \in \mathcal{Y}.$$

For each context x, we define the preference matrix $\mathbf{P}_x \in [0,1]^{|\mathcal{Y}|\times|\mathcal{Y}|}$ with entries $(\mathbf{P}_x)_{y,y'} \triangleq \mathcal{P}(y \succ y' \mid x)$. For policies $\pi, \mu \colon \mathcal{X} \to \Delta_{\mathcal{Y}}$, we define the contextual preference

$$\mathcal{P}(\pi \succ \mu \mid x) \triangleq \mathbb{E}_{y\sim\pi(x),\, y'\sim\mu(x)}\big[\mathcal{P}(y \succ y' \mid x)\big] = \pi(x)^\intercal \mathbf{P}_x \mu(x)\,,$$

and the overall preference

$$\mathcal{P}(\pi \succ \mu) \triangleq \mathbb{E}_{x\sim\rho}\big[\mathcal{P}(\pi \succ \mu \mid x)\big] = \mathbb{E}_{x\sim\rho}\big[\pi(x)^\intercal \mathbf{P}_x \mu(x)\big]\,.$$

More generally, for measurable functions $v, u \colon \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$ (not necessarily probability distributions), we extend the notation as

$$\mathcal{P}(v \succ u \mid x) \triangleq v(x)^\intercal \mathbf{P}_x u(x), \qquad \mathcal{P}(v \succ u) \triangleq \mathbb{E}_{x\sim\rho}\big[v(x)^\intercal \mathbf{P}_x u(x)\big]. \tag{15}$$

On such functions, we use the following inner product and norms:

$$\langle f, g\rangle_\rho \triangleq \mathbb{E}_{x\sim\rho}[\langle f(x), g(x)\rangle]\,,\ \|f\|_{1,\rho} \triangleq \sqrt{\mathbb{E}_{x\sim\rho}[\|f(x)\|_1^2]}\,,\ \|f\|_{\infty,\rho} \triangleq \sqrt{\mathbb{E}_{x\sim\rho}[\|f(x)\|_\infty^2]}\,.$$

For a function $g \colon \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$, we extend the span seminorm by

$$\|g\|_{\mathrm{sp},\rho}^2 \triangleq \mathbb{E}_{x\sim\rho}\big[\|g(x)\|_{\mathrm{sp}}^2\big],$$

where $\|\cdot\|_{\mathrm{sp}}$ on $\mathbb{R}^{|\mathcal{Y}|}$ is the span seminorm introduced in Section 3.

For any functional $\mathcal{F}$ defined on a policy space Π, we define its derivative at π as the function $\nabla\mathcal{F}(\pi) \colon \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$ satisfying the directional derivative identity

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(\pi + th)\Big|_{t=0} = \langle \nabla\mathcal{F}(\pi), h\rangle_\rho \qquad \forall h \colon \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}, \langle h(x), \mathbf{1}\rangle = 0 \text{ for all } x \in \mathcal{X}\,.$$

In particular, it can be characterized as $(\nabla\mathcal{F}(\pi))(x,y) = \frac{\partial}{\partial\pi(y|x)}\mathcal{F}(\pi)$ for ρ-almost all x and all $y \in \mathcal{Y}$.

Finally, for policies π, μ we set

$$\mathrm{KL}_\rho(\pi\|\mu) \triangleq \mathbb{E}_{x\sim\rho}[\mathrm{KL}(\pi(x)\|\mu(x))],$$

and we recall Pinsker's inequality applied pointwise in x:

$$\|\pi(x) - \mu(x)\|_1^2 \le 2\ \mathrm{KL}(\pi(x)\|\mu(x)),$$

which implies

$$\|\pi - \mu\|_{1,\rho}^2 \le 2\ \mathrm{KL}_\rho(\pi\|\mu). \tag{16}$$

We assume throughout that all policies have full support on $\mathcal{Y}$ for ρ-almost all x, so that the expressions involving $\log\pi(y|x)$ are well-defined.

Lemma 2 (Lipschitz property of preferences)**.** *Let $u, v \colon \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$ be measurable functions such that $\langle v(x), \mathbf{1}\rangle = 0$ for all $x \in \mathcal{X}$. Then*

$$|\mathcal{P}(u \succ v)| \le \frac{1}{2}\,\|u\|_{1,\rho}\,\|v\|_{1,\rho}, \qquad |\mathcal{P}(v \succ u)| \le \frac{1}{2}\,\|u\|_{1,\rho}\,\|v\|_{1,\rho}.$$

Proof. Fix any $x \in \mathcal{X}$. Using $\sum_{y'} v(x,y') = 0$, for any $y \in \mathcal{Y}$ we can write

$$(\mathbf{P}_x v(x))_y = \sum_{y'} \mathcal{P}(y \succ y' \mid x) v(x,y') = \sum_{y'} (\mathcal{P}(y \succ y' \mid x) - 1/2) v(x,y').$$

Hence

$$|(\mathbf{P}_x v(x))_y| \le \max_{y'} |\mathcal{P}(y \succ y' \mid x) - 1/2| \sum_{y'} |v(x,y')| \le \frac{1}{2}\|v(x)\|_1,$$

which implies $\|\mathbf{P}_x v(x)\|_\infty \le \frac{1}{2}\|v(x)\|_1$. Therefore,

$$|\mathcal{P}(u \succ v \mid x)| = |u(x)^\intercal \mathbf{P}_x v(x)| \le \|u(x)\|_1 \|\mathbf{P}_x v(x)\|_\infty \le \frac{1}{2}\|u(x)\|_1 \|v(x)\|_1.$$

Taking expectation over $x \sim \rho$ and applying Cauchy–Schwarz,

$$|\mathcal{P}(u \succ v)| \le \frac{1}{2}\,\mathbb{E}_{x\sim\rho}[\|u(x)\|_1 \|v(x)\|_1] \le \frac{1}{2}\,\|u\|_{1,\rho}\,\|v\|_{1,\rho}.$$

The bound for $|\mathcal{P}(v \succ u)|$ is obtained in the same way. □

B.2 Regularized Preference and Value

Fix a reference policy $\widetilde{\pi} \in \Pi$ and $\lambda > 0$. We consider the contextual λ-regularized preference game

$$\max_{\pi\in\Pi}\min_{\mu\in\Pi}\Big\{\mathcal{P}_\lambda^{\widetilde{\pi}}(\pi \succ \mu) \triangleq \mathcal{P}(\pi \succ \mu) - \lambda\,\mathrm{KL}_\rho(\pi\|\widetilde{\pi}) + \lambda\,\mathrm{KL}_\rho(\mu\|\widetilde{\pi})\Big\}. \tag{17}$$

The (contextual) regularized suboptimality is defined as

$$\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi,\mu) \triangleq \frac{1}{2} - \mathcal{P}_\lambda^{\widetilde{\pi}}(\pi \succ \mu), \qquad \mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi) \triangleq \max_{\mu\in\Pi}\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi,\mu).$$

We introduce the regularized value function

$$V_\lambda^{\widetilde{\pi}}(\pi,\mu) \triangleq \mathcal{P}(\mu \succ \pi) + \lambda\,\mathrm{KL}_\rho(\pi\|\widetilde{\pi})\,, \tag{18}$$

and denote

$$\nu_\lambda^{\widetilde{\pi},\star}(\mu) \in \arg\min_{\pi\in\Pi} V_\lambda^{\widetilde{\pi}}(\pi,\mu), \qquad V_\lambda^{\widetilde{\pi},\star}(\mu) \triangleq \min_{\pi\in\Pi} V_\lambda^{\widetilde{\pi}}(\pi,\mu)\,.$$

Note that $\nu_\lambda^{\widetilde{\pi},\star}(\mu)$ might be non-unique on ρ-zero measure sets, but $V_\lambda^{\widetilde{\pi},\star}(\mu)$ is always well-defined.

Lemma 3 (Exploitability gap via value function)**.** *For any policy $\pi \in \Pi$, it holds*

$$\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi) = V_\lambda^{\widetilde{\pi}}(\pi,\pi) - V_\lambda^{\widetilde{\pi},\star}(\pi)\,.$$

Proof. By definition of $\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi,\mu)$ and using $\mathcal{P}(\mu \succ \pi) = 1 - \mathcal{P}(\pi \succ \mu)$,

$$\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi,\mu) = \frac{1}{2} - \mathcal{P}(\pi \succ \mu) + \lambda\,\mathrm{KL}_\rho(\pi\|\widetilde{\pi}) - \lambda\,\mathrm{KL}_\rho(\mu\|\widetilde{\pi}) = V_\lambda^{\widetilde{\pi}}(\pi,\mu) - \tfrac{1}{2} - \lambda\,\mathrm{KL}_\rho(\mu\|\widetilde{\pi})\,.$$

Maximizing over μ yields

$$\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi) = \max_\mu \mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi,\mu) = \max_\mu\big(V_\lambda^{\widetilde{\pi}}(\pi,\mu) - \tfrac{1}{2} - \lambda\,\mathrm{KL}_\rho(\mu\|\widetilde{\pi})\big).$$

In particular, evaluating at $\mu = \pi$ gives

$$V_\lambda^{\widetilde{\pi}}(\pi,\pi) = \frac{1}{2} + \lambda\,\mathrm{KL}_\rho(\pi\|\widetilde{\pi}).$$

On the other hand,

$$V_\lambda^{\widetilde{\pi},\star}(\pi) = \min_{\nu\in\Pi}\big(\mathcal{P}(\pi \succ \nu) + \lambda\,\mathrm{KL}_\rho(\nu\|\widetilde{\pi})\big),$$

so

$$V_\lambda^{\widetilde{\pi}}(\pi,\pi) - V_\lambda^{\widetilde{\pi},\star}(\pi) = \max_{\nu\in\Pi}\big(\tfrac{1}{2} + \lambda\,\mathrm{KL}_\rho(\pi\|\widetilde{\pi}) - \mathcal{P}(\pi \succ \nu) - \lambda\,\mathrm{KL}_\rho(\nu\|\widetilde{\pi})\big),$$

where the right-hand side exactly matches the definition of $\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi)$. □

Lemma 4 (Strong convexity in π with respect to KL_ρ)**.** *Fix $\mu, \widetilde{\pi} \in \Pi$. Then, for any $\pi, \pi' \in \Pi$,*

$$V_\lambda^{\widetilde{\pi}}(\pi,\mu) - V_\lambda^{\widetilde{\pi}}(\pi',\mu) = \langle \nabla_\pi V_\lambda^{\widetilde{\pi}}(\pi',\mu), \pi - \pi'\rangle_\rho + \lambda\,\mathrm{KL}_\rho(\pi\|\pi')\,, \tag{19}$$

where $\nabla_\pi V_\lambda^{\widetilde{\pi}}(\pi',\mu)$, defined as a Fréchet derivative of $V_\lambda^{\widetilde{\pi}}(\pi',\mu)$ with respect to the first argument, is given element-wise by

$$\big(\nabla_\pi V_\lambda^{\widetilde{\pi}}(\pi',\mu)\big)(x,y) \triangleq \big(\mathbf{P}_x^\intercal \mu(x)\big)_y + \lambda\Big(1 + \log\frac{\pi'(y|x)}{\widetilde{\pi}(y|x)}\Big),$$

In particular, if $\pi' = \nu_\lambda^{\widetilde{\pi},\star}(\mu) \in \arg\min_\pi V_\lambda^{\widetilde{\pi}}(\pi,\mu)$, then

$$V_\lambda^{\widetilde{\pi}}(\pi,\mu) - V_\lambda^{\widetilde{\pi},\star}(\mu) \geq \lambda\ \mathrm{KL}_\rho(\pi\|\nu_\lambda^{\widetilde{\pi},\star}(\mu)).$$

Proof. Write

$$V_\lambda^{\widetilde{\pi}}(\pi,\mu)=\mathbb{E}_{x\sim\rho}\Big[\mu(x)^\intercal \mathbf{P}_x\pi(x)+\lambda f_x(\pi(x))\Big],\qquad f_x(p)\triangleq\sum_{y\in\mathcal{Y}}p_y\log\frac{p_y}{\widetilde{\pi}(y|x)}.$$

Fix $\mu,\widetilde{\pi}$ and π'. Let $h\colon\mathcal{X}\to\mathbb{R}^{|\mathcal{Y}|}$ be any direction and consider

$$\Phi(t)\triangleq V_\lambda^{\widetilde{\pi}}(\pi'+th,\mu).$$

For each x and y, the derivative of $p\mapsto p\log\frac{p}{\tau}$ at $p=p'$ is

$$\frac{\mathrm{d}}{\mathrm{d}p}\Big(p\log\frac{p}{\widetilde{\pi}(y|x)}\Big)\Big|_{p=p'}=\log\frac{p'}{\widetilde{\pi}(y|x)}+1.$$

Thus, pointwise in x,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big[\mu(x)^\intercal\mathbf{P}_x(\pi'(x)+th(x))\Big]_{t=0}=\mu(x)^\intercal\mathbf{P}_xh(x),$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\Big[\lambda f_x(\pi'(x)+th(x))\Big]_{t=0}=\lambda\sum_y h(x,y)\Big(\log\frac{\pi'(y|x)}{\widetilde{\pi}(y|x)}+1\Big).$$

Therefore

$$\Phi'(0)=\mathbb{E}_{x\sim\rho}\Big[\sum_y\Big((\mathbf{P}_x^\intercal\mu(x))_y+\lambda\big(1+\log\tfrac{\pi'(y|x)}{\widetilde{\pi}(y|x)}\big)\Big)h(x,y)\Big]=\langle\nabla_\pi V_\lambda^{\widetilde{\pi}}(\pi',\mu),h\rangle_\rho,$$

with

$$\big(\nabla_\pi V_\lambda^{\widetilde{\pi}}(\pi',\mu)\big)(x,y)=(\mathbf{P}_x^\intercal\mu(x))_y+\lambda\Big(1+\log\tfrac{\pi'(y|x)}{\widetilde{\pi}(y|x)}\Big).$$

Now consider the difference

$$V_\lambda^{\widetilde{\pi}}(\pi,\mu)-V_\lambda^{\widetilde{\pi}}(\pi',\mu)=\mathbb{E}_{x\sim\rho}\Big[\mu(x)^\intercal\mathbf{P}_x\big(\pi(x)-\pi'(x)\big)+\lambda\big(f_x(\pi(x))-f_x(\pi'(x))\big)\Big].$$

Similarly,

$$\langle\nabla_\pi V_\lambda^{\widetilde{\pi}}(\pi',\mu),\pi-\pi'\rangle_\rho=\mathbb{E}_{x\sim\rho}\Big[\mu(x)^\intercal\mathbf{P}_x\big(\pi(x)-\pi'(x)\big)+\lambda\sum_y\big(1+\log\tfrac{\pi'(y|x)}{\widetilde{\pi}(y|x)}\big)\big(\pi(y|x)-\pi'(y|x)\big)\Big].$$

Subtracting, we isolate the Bregman term:

$$V_\lambda^{\widetilde{\pi}}(\pi,\mu)-V_\lambda^{\widetilde{\pi}}(\pi',\mu)-\langle\nabla_\pi V_\lambda^{\widetilde{\pi}}(\pi',\mu),\pi-\pi'\rangle_\rho$$
$$=\lambda\,\mathbb{E}_{x\sim\rho}\Big[f_x(\pi(x))-f_x(\pi'(x))-\sum_y\big(1+\log\tfrac{\pi'(y|x)}{\widetilde{\pi}(y|x)}\big)\big(\pi(y|x)-\pi'(y|x)\big)\Big].$$

Fix x and abbreviate $p_y=\pi(y|x)$, $p'_y=\pi'(y|x)$, $\tau_y=\widetilde{\pi}(y|x)$. Then

$$f_x(p)=\sum_y p_y\log\frac{p_y}{\tau_y},\qquad\frac{\partial f_x}{\partial p_y}(p')=\log\frac{p'_y}{\tau_y}+1.$$

The inner expression is the Bregman divergence of f_x at $p\|p'$:

$$f_x(p)-f_x(p')-\sum_y\frac{\partial f_x}{\partial p_y}(p')(p_y-p'_y).$$

A standard calculation gives

$$f_x(p)-f_x(p')-\sum_y\frac{\partial f_x}{\partial p_y}(p')(p_y-p'_y)=\sum_y p_y\log\frac{p_y}{p'_y}=\mathrm{KL}(p\|p').$$

Thus

$$V_\lambda^{\widetilde{\pi}}(\pi,\mu)-V_\lambda^{\widetilde{\pi}}(\pi',\mu)-\langle\nabla_\pi V_\lambda^{\widetilde{\pi}}(\pi',\mu),\pi-\pi'\rangle_\rho=\lambda\,\mathbb{E}_{x\sim\rho}[\mathrm{KL}(\pi(x)\|\pi'(x))]=\lambda\,\mathrm{KL}_\rho(\pi\|\pi'),$$

which gives (19).

For the final claim, let $\pi' = \nu_\lambda^{\widetilde{\pi},\star}(\mu)$ be a minimizer of $V_\lambda^{\widetilde{\pi}}(\cdot,\mu)$. First-order optimality implies

$$\langle \nabla_\pi V_\lambda^{\widetilde{\pi}}(\pi',\mu), \pi - \pi' \rangle_\rho \geq 0 \qquad \forall \pi \in \Pi.$$

Plugging π' into (19) and using $V_\lambda^{\widetilde{\pi}}(\pi',\mu) = V_\lambda^{\widetilde{\pi},\star}(\mu)$ yields

$$V_\lambda^{\widetilde{\pi}}(\pi,\mu) - V_\lambda^{\widetilde{\pi},\star}(\mu) \geq \lambda \, \mathrm{KL}_\rho(\pi \| \pi') = \lambda \, \mathrm{KL}_\rho(\pi \| \nu_\lambda^{\widetilde{\pi},\star}(\mu)) \,,$$

as claimed. □

Lemma 5 (Smoothness of the best-response value)**.** *Fix a reference policy $\widetilde{\pi} \in \Pi$ and $\lambda > 0$. Then, for any policies $\mu, \mu' \in \Pi$,*

$$\mathcal{P}(\mu - \mu' \succ \nu_\lambda^{\widetilde{\pi},\star}(\mu)) \leq V_\lambda^{\widetilde{\pi},\star}(\mu) - V_\lambda^{\widetilde{\pi},\star}(\mu') \leq \mathcal{P}(\mu - \mu' \succ \nu_\lambda^{\widetilde{\pi},\star}(\mu')) \,.$$

Proof. Note that the lower bound follows automatically from the upper bound by multiplying both sides by -1 and renaming μ and μ':

$$V_\lambda^{\widetilde{\pi},\star}(\mu) - V_\lambda^{\widetilde{\pi},\star}(\mu') \leq \mathcal{P}(\mu - \mu' \succ \nu_\lambda^{\widetilde{\pi},\star}(\mu')) \iff V_\lambda^{\widetilde{\pi},\star}(\mu') - V_\lambda^{\widetilde{\pi},\star}(\mu) \geq \mathcal{P}(\mu' - \mu \succ \nu_\lambda^{\widetilde{\pi},\star}(\mu'))$$

Thus, it is enough to prove only an upper bound. Fix an arbitrary context $x \in \mathrm{supp}(\rho)$ and consider the regularizer $q \in \Delta_\mathcal{Y}$ as

$$\varphi_x(q) \triangleq \lambda \, \mathrm{KL}(q \| \widetilde{\pi}(x)) \,,$$

and we define its convex conjugate φ_x^* on $g \in \mathbb{R}^\mathcal{Y}$ as

$$\varphi_x^*(g) \triangleq \sup_{q \in \Delta_\mathcal{Y}} \big\{ \langle g, q \rangle - \varphi_x(q) \big\} = \sup_{q \in \Delta_\mathcal{Y}} \big\{ \langle g, q \rangle - \lambda \, \mathrm{KL}(q \| \widetilde{\pi}(x)) \big\}.$$

We note that the convex conjugate is always a convex function. As a result, we have

$$\forall g, h \in \mathbb{R}^Y : \varphi_x^*(g) - \varphi_x^*(h) - \langle \nabla \varphi_x^*(h), g - h \rangle \geq 0 \,, \tag{20}$$

where $\nabla \varphi_x^*(g)$ is the unique maximizer in the definition of $\varphi_x^*(g)$.

Next, we relate $V_\lambda^{\widetilde{\pi},\star}(\mu)$ to φ_x^*. For any μ, π,

$$V_\lambda^{\widetilde{\pi}}(\pi,\mu) = \mathcal{P}(\mu \succ \pi) + \lambda \, \mathrm{KL}_\rho(\pi \| \widetilde{\pi}) = \mathbb{E}_{x \sim \rho}[\langle \mathbf{P}_x^\intercal \mu(x), \pi(x) \rangle + \lambda \, \mathrm{KL}(\pi(x) \| \widetilde{\pi}(x))] \,,$$

Thus, using the fact that optimization is performed over all functions from $\mathcal{X}$ to $\Delta_\mathcal{Y}$,

$$\begin{aligned} V_\lambda^{\widetilde{\pi},\star}(\mu) &= \min_{\pi \in \Pi} V_\lambda^{\widetilde{\pi}}(\pi,\mu) = - \sup_{\pi \in \Pi} \mathbb{E}_{x \sim \rho}[\langle -\mathbf{P}_x^\intercal \mu(x), \pi(x) \rangle - \lambda \, \mathrm{KL}(\pi(x) \| \widetilde{\pi}(x))] \\ &= -\mathbb{E}_{x \sim \rho} \left[\sup_{q \in \Delta_\mathcal{Y}} \langle -\mathbf{P}_x^\intercal \mu(x), q \rangle - \lambda \, \mathrm{KL}(q \| \widetilde{\pi}(x)) \right] = -\mathbb{E}_{x \sim \rho}[\varphi_x^*(-\mathbf{P}_x^\intercal \mu(x))] \,. \end{aligned}$$

Also, for any $x \in \mathrm{supp}(\rho)$, the unique minimizer of $V_\lambda^{\widetilde{\pi}}(\cdot,\mu)$ at context x is given by

$$[\nu_\lambda^{\widetilde{\pi},\star}(\mu)](x) = \arg\min_{q \in \Delta_\mathcal{Y}} \langle \mathbf{P}_x^\intercal \mu(x), q \rangle + \lambda \, \mathrm{KL}(q \| \widetilde{\pi}(x)) = \nabla \varphi_x^*(-\mathbf{P}_x^\intercal \mu(x)) \,.$$

Using these relations, we can now prove the smoothness property of $V_\lambda^{\widetilde{\pi},\star}$. Applying (20) pointwise at each $x \in \mathrm{supp}(\rho)$ with $g = -\mathbf{P}_x^\intercal \mu(x)$, $h = -\mathbf{P}_x^\intercal \mu'(x)$, we get

$$\varphi_x^*(-\mathbf{P}_x^\intercal \mu(x)) - \varphi_x^*(-\mathbf{P}_x^\intercal \mu'(x)) - \langle \nabla \varphi_x^*(-\mathbf{P}_x^\intercal \mu'(x)), -\mathbf{P}_x^\intercal (\mu(x) - \mu'(x)) \rangle \geq 0 \,.$$

Taking expectation over $x \sim \rho$ and multiplying by -1, we obtain

$$\mathbb{E}_{x \sim \rho}[-\varphi_x^*(-\mathbf{P}_x^\intercal \mu(x))] - \mathbb{E}_{x \sim \rho}[-\varphi_x^*(-\mathbf{P}_x^\intercal \mu'(x))] - \mathbb{E}_{x \sim \rho}[\langle \nabla \varphi_x^*(-\mathbf{P}_x^\intercal \mu'(x)), \mathbf{P}_x^\intercal (\mu(x) - \mu'(x)) \rangle] \leq 0 \,.$$

Rewriting using the relations to $V_\lambda^{\widetilde{\pi},\star}$ and $\nu_\lambda^{\widetilde{\pi},\star}$, we get

$$V_\lambda^{\widetilde{\pi},\star}(\mu) - V_\lambda^{\widetilde{\pi},\star}(\mu') - \mathcal{P}(\mu - \mu' \succ \nu_\lambda^{\widetilde{\pi},\star}(\mu')) \leq 0 \,.$$

□

Lemma 6 (Suboptimality decomposition)**.** *Let $\pi_\lambda^{\widetilde{\pi},\star}$ be a solution to (17). Then, for any policies $\pi, \mu \in \Pi$, it holds*

$$\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi, \mu) \le \mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi, \pi_\lambda^{\widetilde{\pi},\star}) + \frac{1}{2}\|\pi - \pi_\lambda^{\widetilde{\pi},\star}\|_{1,\rho} \cdot \|\mu - \pi_\lambda^{\widetilde{\pi},\star}\|_{1,\rho}\,.$$

Proof. By definition,

$$\begin{aligned}
\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi, \mu) &= \frac{1}{2} - \mathcal{P}(\pi \succ \mu) + \lambda\,\mathrm{KL}_\rho(\pi\|\widetilde{\pi}) - \lambda\,\mathrm{KL}_\rho(\mu\|\widetilde{\pi}) \\
&= \underbrace{\frac{1}{2} - \mathcal{P}(\pi \succ \pi_\lambda^{\widetilde{\pi},\star}) + \lambda\,\mathrm{KL}_\rho(\pi\|\widetilde{\pi}) - \lambda\,\mathrm{KL}_\rho(\pi_\lambda^{\widetilde{\pi},\star}\|\widetilde{\pi})}_{\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi, \pi_\lambda^{\widetilde{\pi},\star})} \\
&\quad + \underbrace{\frac{1}{2} - \mathcal{P}(\pi_\lambda^{\widetilde{\pi},\star} \succ \mu) + \lambda\,\mathrm{KL}_\rho(\pi_\lambda^{\widetilde{\pi},\star}\|\widetilde{\pi}) - \lambda\,\mathrm{KL}_\rho(\mu\|\widetilde{\pi})}_{\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_\lambda^{\widetilde{\pi},\star}, \mu)} \\
&\quad + \mathcal{P}(\pi - \pi_\lambda^{\widetilde{\pi},\star} \succ \pi_\lambda^{\widetilde{\pi},\star} - \mu)\,,
\end{aligned}$$

where, by bilinearity (15) and symmetry, we used

$$\mathcal{P}(\pi - \pi_\lambda^{\widetilde{\pi},\star} \succ \pi_\lambda^{\widetilde{\pi},\star} - \mu) = \mathcal{P}(\pi \succ \pi_\lambda^{\widetilde{\pi},\star}) - \mathcal{P}(\pi \succ \mu) - \mathcal{P}(\pi_\lambda^{\widetilde{\pi},\star} \succ \pi_\lambda^{\widetilde{\pi},\star}) + \mathcal{P}(\pi_\lambda^{\widetilde{\pi},\star} \succ \mu),$$

and $\mathcal{P}(\pi_\lambda^{\widetilde{\pi},\star} \succ \pi_\lambda^{\widetilde{\pi},\star}) = 1/2$.

Since $\pi_\lambda^{\widetilde{\pi},\star}$ is a Nash equilibrium of the regularized game, we have

$$\mathcal{P}_\lambda^{\widetilde{\pi}}(\pi_\lambda^{\widetilde{\pi},\star} \succ \mu) \ge \frac{1}{2}$$

for all μ, hence $\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_\lambda^{\widetilde{\pi},\star}, \mu) \le 0$. It remains to bound the last term. For each x, both $\pi_\lambda^{\widetilde{\pi},\star}(x)$ and $\mu(x)$ are distributions, so $\sum_y(\pi_\lambda^{\widetilde{\pi},\star}(y|x) - \mu(y|x)) = 0$. Therefore, by Lemma 2 applied to $u(x) = \pi(x) - \pi_\lambda^{\widetilde{\pi},\star}(x)$, $v(x) = \pi_\lambda^{\widetilde{\pi},\star}(x) - \mu(x)$,

$$\left|\mathcal{P}(\pi - \pi_\lambda^{\widetilde{\pi},\star} \succ \pi_\lambda^{\widetilde{\pi},\star} - \mu)\right| \le \frac{1}{2}\|\pi - \pi_\lambda^{\widetilde{\pi},\star}\|_{1,\rho} \cdot \|\pi_\lambda^{\widetilde{\pi},\star} - \mu\|_{1,\rho}\,.$$

Combining the above inequalities yields

$$\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi, \mu) \le \mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi, \pi_\lambda^{\widetilde{\pi},\star}) + \frac{1}{2}\|\pi - \pi_\lambda^{\widetilde{\pi},\star}\|_{1,\rho} \cdot \|\mu - \pi_\lambda^{\widetilde{\pi},\star}\|_{1,\rho}\,.$$

□

B.3 Proximal Point Method

We now specialize to the β-regularized contextual game with respect to $\pi^{\mathrm{ref}} \in \Pi$:

$$\max_{\pi\in\Pi}\min_{\mu\in\Pi}\left\{\mathcal{P}_\beta(\pi \succ \mu) \triangleq \mathcal{P}(\pi \succ \mu) - \beta\,\mathrm{KL}_\rho(\pi\|\pi^{\mathrm{ref}}) + \beta\,\mathrm{KL}_\rho(\mu\|\pi^{\mathrm{ref}})\right\}.$$

We define

$$V_\beta(\pi, \mu) \triangleq \mathcal{P}(\mu \succ \pi) + \beta\,\mathrm{KL}_\rho(\pi\|\pi^{\mathrm{ref}})\,, \qquad V_\beta^\star(\mu) \triangleq \min_{\pi\in\Pi} V_\beta(\pi, \mu)\,,$$

and let $\pi_\beta^\star$ denote the β-regularized von Neumann winner in this contextual game. The suboptimality is

$$\mathrm{SubOpt}_\beta(\pi, \mu) \triangleq \tfrac{1}{2} - \mathcal{P}_\beta(\pi \succ \mu), \qquad \mathrm{SubOpt}_\beta(\pi) \triangleq \max_\mu \mathrm{SubOpt}_\beta(\pi, \mu).$$

Method description. The iterates of the approximate proximal point (PP) method can be written as follows:

$$\pi_{t+1} \approx \arg\max_{\pi\in\Pi} \min_{\mu\in\Pi} \left\{ \mathcal{P}_\beta(\pi \succ \mu) - \frac{\beta}{\eta}\,\mathrm{KL}_\rho(\pi\|\pi_t) + \frac{\beta}{\eta}\,\mathrm{KL}_\rho(\mu\|\pi_t) \right\}, \tag{21}$$

where $\eta > 0$ is a PP learning rate. To define the success criteria for approximation, we introduce the regularized value

$$V_t(\pi,\mu) \triangleq V_\beta(\pi,\mu) + \frac{\beta}{\eta}\,\mathrm{KL}_\rho(\pi\|\pi_t)\,, \qquad V_t^\star(\mu) \triangleq \min_{\pi\in\Pi} V_t(\pi,\mu)\,.$$

For the exact solution to (21), the updated policy π_{t+1} is its own best response, i.e., $\nabla_\pi V_t(\pi_{t+1},\pi_{t+1})(x) \propto \mathbf{1}$ for all $x \in \mathcal{X}$. In the contextual setting, we measure the approximation error by

$$\left\|\nabla_\pi V_t(\pi_{t+1},\pi_{t+1})\right\|^2_{\mathrm{sp},\rho} \triangleq \mathbb{E}_{x\sim\rho}\left[\left\|\nabla_\pi V_t(\pi_{t+1},\pi_{t+1})(x)\right\|^2_{\mathrm{sp}}\right].$$

We say that π_{t+1} is an ε-approximation of (21) if

$$\left\|\nabla_\pi V_t(\pi_{t+1},\pi_{t+1})\right\|^2_{\mathrm{sp},\rho} \le \varepsilon.$$

For brevity, we denote $\zeta_{t+1} \triangleq \nabla_\pi V_t(\pi_{t+1},\pi_{t+1})$.

Equivalence with a mixed-reference value. For each $x \in \mathcal{X}$, consider the regularization term

$$\beta\,\mathrm{KL}(\pi(x)\|\pi^{\mathrm{ref}}(x)) + \frac{\beta}{\eta}\,\mathrm{KL}(\pi(x)\|\pi_t(x)).$$

A direct calculation shows that there exists a policy $\tilde\pi_t$ such that

$$\tilde\pi_t(y \mid x) \propto [\pi^{\mathrm{ref}}(y \mid x)]^{\eta/(1+\eta)}[\pi_t(y \mid x)]^{1/(1+\eta)},$$

and

$$\beta\,\mathrm{KL}(\pi(x)\|\pi^{\mathrm{ref}}(x)) + \frac{\beta}{\eta}\,\mathrm{KL}(\pi(x)\|\pi_t(x)) = \beta\Big(1+\frac{1}{\eta}\Big)\,\mathrm{KL}(\pi(x)\|\tilde\pi_t(x)) + C_t(x),$$

where $C_t(x)$ is independent of $\pi(x)$. Averaging over $x \sim \rho$, we obtain

$$V_t(\pi,\mu) = \mathcal{P}(\mu \succ \pi) + \beta\Big(1+\frac{1}{\eta}\Big)\,\mathrm{KL}_\rho(\pi\|\tilde\pi_t) + \mathrm{const}(\pi^{\mathrm{ref}},\pi_t).$$

Hence, by Lemma 4 with $\lambda = \beta(1+1/\eta)$ and reference $\tilde\pi_t$, we have

$$V_t(\pi,\mu) - V_t(\pi',\mu) = \langle \nabla_\pi V_t(\pi',\mu), \pi - \pi'\rangle_\rho + \beta\Big(1+\frac{1}{\eta}\Big)\,\mathrm{KL}_\rho(\pi\|\pi') \tag{22}$$

for all $\pi,\pi',\mu \in \Pi$.

Proposition 2 (Convergence of approximate PP method)**.** Assume that each iterate π_{t+1} is an ε-approximation in the sense that $\left\|\nabla_\pi V_t(\pi_{t+1},\pi_{t+1})\right\|^2_{\mathrm{sp},\rho} \le \varepsilon$ for all $t \ge 0$. Then, for any $t \in \mathbb{N}$, it holds

$$\mathrm{KL}_\rho(\pi^\star_\beta\|\pi_t) \le (1+\eta/2)^{-t}\cdot \mathrm{KL}_\rho(\pi^\star_\beta\|\pi_0) + \frac{2\varepsilon}{\beta^2}\,,$$

and

$$\mathrm{SubOpt}_\beta(\pi_t) \le (1+\eta/2)^{-t}\cdot\left(\frac{1}{\beta}+\frac{2\beta}{\eta}+3\right)\cdot \mathrm{KL}_\rho(\pi^\star_\beta\|\pi_0) + \left(2+\frac{2}{\beta^3}+\frac{4}{\eta\beta}+\frac{6}{\beta^2}\right)\varepsilon\,,$$

and, moreover,

$$\left\|\log\pi_t - \log\pi^\star_\beta\right\|^2_{\mathrm{sp},\rho} \le (1+\eta)^{-t}\left\|\log\pi_0 - \log\pi^\star_\beta\right\|^2_{\mathrm{sp},\rho} + \frac{2}{\beta^2}\,\mathrm{KL}_\rho(\pi^\star_\beta\|\pi_0)\cdot(1+\eta/2)^{-t} + \frac{2(1+\beta^2)\cdot\varepsilon}{\beta^4}\,.$$

Remark 1 (On the choice of η)**.** The present bound suggests that taking $\eta \to +\infty$ is the best solution since it will drive convergence in just one iterate. However, the actual tradeoff is implicitly hidden in the value of ε: the speed of convergence of the internal subproblems depends on η and a smaller η makes these subproblems better-conditioned thanks to stronger regularization. In particular, the convergence of Mirror Prox-style methods requires $\eta = \mathcal{O}(\beta)$ if transferred to our setting (Cen et al., 2024).

Proof. We split the proof into three parts: convergence in KL_ρ, convergence in suboptimality, and convergence in the span seminorm of log-probabilities.

Convergence in KL_ρ. Define $A_t \triangleq \sqrt{\mathrm{KL}_\rho(\pi^\star_\beta \| \pi_t)}$. Applying (22) with $\pi = \pi^\star_\beta$, $\pi' = \pi_{t+1}$, $\mu = \pi_{t+1}$, we obtain

$$\begin{aligned}\beta\Big(1+\frac{1}{\eta}\Big)\,\mathrm{KL}_\rho(\pi^\star_\beta\|\pi_{t+1}) &= V_t(\pi^\star_\beta,\pi_{t+1}) - V_t(\pi_{t+1},\pi_{t+1}) + \langle \nabla_\pi V_t(\pi_{t+1},\pi_{t+1}), \pi_{t+1}-\pi^\star_\beta\rangle_\rho \\ &= V_\beta(\pi^\star_\beta,\pi_{t+1}) + \frac{\beta}{\eta}\,\mathrm{KL}_\rho(\pi^\star_\beta\|\pi_t) - V_\beta(\pi_{t+1},\pi_{t+1}) - \frac{\beta}{\eta}\,\mathrm{KL}_\rho(\pi_{t+1}\|\pi_t) \\ &\quad + \langle \zeta_{t+1}, \pi_{t+1}-\pi^\star_\beta\rangle_\rho.\end{aligned}$$

Using the definition of V_β and $\mathcal{P}_\beta$, we have

$$V_\beta(\pi^\star_\beta,\pi_{t+1}) - V_\beta(\pi_{t+1},\pi_{t+1}) = \mathcal{P}_\beta(\pi_{t+1}\succ\pi^\star_\beta) - \frac{1}{2} = -\,\mathrm{SubOpt}_\beta(\pi_{t+1},\pi^\star_\beta) \le 0.$$

Dropping the non-negative term $\frac{\beta}{\eta}\,\mathrm{KL}_\rho(\pi_{t+1}\|\pi_t)$, we get

$$\beta\Big(1+\frac{1}{\eta}\Big)A_{t+1}^2 \le \frac{\beta}{\eta}A_t^2 + \langle \zeta_{t+1}, \pi_{t+1}-\pi^\star_\beta\rangle_\rho - \mathrm{SubOpt}_\beta(\pi_{t+1},\pi^\star_\beta)\,, \tag{23}$$

where we can drop the last term too since it is non-positive. For each x, since $\sum_y(\pi_{t+1}(y|x) - \pi^\star_\beta(y|x)) = 0$, we have for any scalar function $c(x)$

$$\langle \zeta_{t+1}(x), \pi_{t+1}(x)-\pi^\star_\beta(x)\rangle = \langle \zeta_{t+1}(x) - c(x)\cdot\mathbf{1}, \pi_{t+1}(x)-\pi^\star_\beta(x)\rangle.$$

Thus

$$\langle \zeta_{t+1}, \pi_{t+1}-\pi^\star_\beta\rangle_\rho \le \mathbb{E}_{x\sim\rho}\big[\|\zeta_{t+1}(x) - c(x)\cdot\mathbf{1}\|_\infty \|\pi_{t+1}(x)-\pi^\star_\beta(x)\|_1\big].$$

Using Cauchy–Schwarz,

$$\begin{aligned}\langle \zeta_{t+1}, \pi_{t+1}-\pi^\star_\beta\rangle_\rho &\le \mathbb{E}_{x\sim\rho}\big[\|\zeta_{t+1}(x) - c(x)\cdot\mathbf{1}\|_\infty \|\pi_{t+1}(x)-\pi^\star_\beta(x)\|_1\big] \\ &\le \sqrt{\mathbb{E}_{x\sim\rho}[\|\zeta_{t+1}(x)-c(x)\cdot\mathbf{1}\|_\infty^2]\cdot\mathbb{E}_{x\sim\rho}\Big[\|\pi_{t+1}(x)-\pi^\star_\beta(x)\|_1^2\Big]}\,.\end{aligned}$$

Finally, minimizing over $c(x)$, we have

$$\langle \zeta_{t+1}, \pi_{t+1}-\pi^\star_\beta\rangle_\rho \le \|\zeta_{t+1}\|_{\mathrm{sp},\rho}\cdot\|\pi_{t+1}-\pi^\star_\beta\|_{1,\rho}\,.$$

By Pinsker's inequality (16),

$$\|\pi_{t+1}-\pi^\star_\beta\|_{1,\rho} \le \sqrt{2\,\mathrm{KL}_\rho(\pi^\star_\beta\|\pi_{t+1})} = \sqrt{2}\,A_{t+1},$$

hence

$$\langle \zeta_{t+1}, \pi_{t+1}-\pi^\star_\beta\rangle_\rho \le \sqrt{2}\,\|\zeta_{t+1}\|_{\mathrm{sp},\rho}A_{t+1}.$$

Substituting into (23), we obtain

$$\beta\Big(1+\frac{1}{\eta}\Big)A_{t+1}^2 \le \frac{\beta}{\eta}A_t^2 + \sqrt{2}\,\|\zeta_{t+1}\|_{\mathrm{sp},\rho}A_{t+1}.$$

Dividing by $\beta(1+1/\eta)$, we have

$$A_{t+1}^2 \le \frac{1}{1+\eta}A_t^2 + 2\cdot\frac{\eta\sqrt{2\|\zeta_{t+1}\|_{\mathrm{sp},\rho}^2}}{2\beta\cdot(1+\eta)}\cdot A_{t+1}\,.$$

Solving this quadratic inequality in A_{t+1}

$$A_{t+1} \le \frac{\eta\sqrt{2\|\zeta_{t+1}\|_{\mathrm{sp},\rho}^2}}{2\beta\cdot(1+\eta)} + \sqrt{\frac{1}{1+\eta}A_t^2 + \frac{\eta^2\|\zeta_{t+1}\|_{\mathrm{sp},\rho}^2}{2\beta^2\cdot(1+\eta)^2}}\,.$$

Taking the square of both sides and using an inequality $(a+b)^2 \le (1+\alpha)a^2 + (1+1/\alpha)b^2$ for any $\alpha > 0$

$$A_{t+1}^2 \le \frac{1+\alpha}{1+\eta}A_t^2 + (2+\alpha+1/\alpha)\frac{\eta^2\|\zeta_{t+1}\|_{\mathrm{sp},\rho}^2}{2\beta^2\cdot(1+\eta)^2}\,.$$

Taking α as a solution to $(1+\alpha)/(1+\eta) = 1/(1+\eta/2)$, we have $2+\alpha+1/\alpha = (1+\eta)^2/(\eta/2 \cdot (1+\eta/2))$, thus we achieve

$$A_{t+1}^2 \le \frac{1}{1+\eta/2} A_t^2 + \frac{\eta \|\zeta_{t+1}\|_{\mathrm{sp},\rho}^2}{\beta^2 \cdot (1+\eta/2)} \le \frac{1}{1+\eta/2} A_t^2 + \frac{\eta \cdot \varepsilon}{\beta^2 \cdot (1+\eta/2)},$$

where the condition on a gradient error $\|\zeta_{t+1}\|_{\mathrm{sp},\rho}^2 \le \varepsilon$ is applied. Unrolling this inequality, we achieve for any $T \in \mathbb{N}$

$$A_T^2 \le \frac{A_0^2}{(1+\eta/2)^T} + \frac{\eta \cdot \varepsilon}{\beta^2(1+\eta/2)} \cdot \frac{1}{1-1/(1+\eta/2)} \le \frac{A_0^2}{(1+\eta/2)^T} + \frac{2\varepsilon}{\beta^2}.$$

Convergence in suboptimality. Rearranging (23), we have

$$\mathrm{SubOpt}_\beta(\pi_{t+1}, \pi_\beta^\star) \le (\beta/\eta) A_t^2 + \sqrt{2} \|\zeta_{t+1}\|_{\mathrm{sp},\rho} \cdot A_{t+1}.$$

Let $\mu_{t+1}^\star \in \arg\max_\mu \mathrm{SubOpt}_\beta(\pi_{t+1}, \mu)$ be a worst-case opponent so that $\mathrm{SubOpt}_\beta(\pi_{t+1}) = \mathrm{SubOpt}_\beta(\pi_{t+1}, \mu_{t+1}^\star)$. Applying Lemma 6 with $\pi_\lambda^{\pi^{\mathrm{ref}},\star} = \pi_\beta^\star$ and $\lambda = \beta, \widetilde{\pi} = \pi^{\mathrm{ref}}$ and triangle inequality, we get

$$\begin{aligned}
\mathrm{SubOpt}_\beta(\pi_{t+1}) &\le (\beta/\eta) A_t^2 + \sqrt{2} \|\zeta_{t+1}\|_{\mathrm{sp},\rho} \cdot A_{t+1} + \frac{1}{2} \|\pi_{t+1} - \pi_\beta^\star\|_{1,\rho} \cdot \|\mu_{t+1}^\star - \pi_\beta^\star\|_{1,\rho} \\
&\le (\beta/\eta) A_t^2 + \sqrt{2} \|\zeta_{t+1}\|_{\mathrm{sp},\rho} \cdot A_{t+1} + \frac{1}{2} \|\pi_{t+1} - \pi_\beta^\star\|_{1,\rho} \cdot \|\mu_{t+1}^\star - \pi_{t+1}\|_{1,\rho} \\
&\quad + \frac{1}{2} \|\pi_{t+1} - \pi_\beta^\star\|_{1,\rho}^2.
\end{aligned}$$

By Pinsker's inequality (16),

$$\|\pi_{t+1} - \pi_\beta^\star\|_{1,\rho} \le \sqrt{2 \,\mathrm{KL}_\rho(\pi_\beta^\star \| \pi_{t+1})} = \sqrt{2} \cdot A_{t+1}.$$

Next, we show that

$$\mathrm{KL}_\rho(\pi_{t+1} \| \mu_{t+1}^\star) \le \frac{1}{\beta} \mathrm{SubOpt}_\beta(\pi_{t+1}).$$

For any π, μ, a direct calculation gives

$V_\beta(\pi, \pi) - V_\beta(\mu, \pi) = \mathcal{P}(\pi \succ \pi) + \beta \,\mathrm{KL}_\rho(\pi \| \pi^{\mathrm{ref}}) - \mathcal{P}(\pi \succ \mu) - \beta \,\mathrm{KL}_\rho(\mu \| \pi^{\mathrm{ref}}) = \mathrm{SubOpt}_\beta(\pi, \mu)$.

Therefore, for fixed π,

$$\mathrm{SubOpt}_\beta(\pi) = \max_\mu \mathrm{SubOpt}_\beta(\pi, \mu) = V_\beta(\pi, \pi) - \min_\mu V_\beta(\mu, \pi).$$

This implies that $\mu_{t+1}^\star$ is a minimizer of $V_\beta(\cdot, \pi_{t+1})$ in its first argument. Applying Lemma 4 with $\lambda = \beta$, $\mu = \pi_{t+1}$, $\pi = \pi_{t+1}$, and $\pi' = \mu_{t+1}^\star$, we obtain

$$V_\beta(\pi_{t+1}, \pi_{t+1}) - V_\beta(\mu_{t+1}^\star, \pi_{t+1}) \ge \beta \,\mathrm{KL}_\rho(\pi_{t+1} \| \mu_{t+1}^\star).$$

The left-hand side is precisely $\mathrm{SubOpt}_\beta(\pi_{t+1})$, hence

$$\beta \,\mathrm{KL}_\rho(\pi_{t+1} \| \mu_{t+1}^\star) \le \mathrm{SubOpt}_\beta(\pi_{t+1}).$$

Using Pinsker's inequality once more,

$$\|\pi_{t+1} - \mu_{t+1}^\star\|_{1,\rho} \le \sqrt{2 \,\mathrm{KL}_\rho(\pi_{t+1} \| \mu_{t+1}^\star)} \le \sqrt{\frac{2}{\beta} \mathrm{SubOpt}_\beta(\pi_{t+1})}.$$

Combining the two bounds, we get

$$\frac{1}{2} \|\pi_{t+1} - \pi_\beta^\star\|_{1,\rho} \cdot \|\pi_{t+1} - \mu_{t+1}^\star\|_{1,\rho} \le \sqrt{1/\beta}\, A_{t+1} \sqrt{\mathrm{SubOpt}_\beta(\pi_{t+1})}.$$

Denoting $B_{t+1} = \sqrt{\mathrm{SubOpt}_\beta(\pi_{t+1})}$, we have

$$B_{t+1}^2 \le A_{t+1}^2 + (\beta/\eta) A_t^2 + \sqrt{2} \|\zeta_{t+1}\|_{\mathrm{sp},\rho} A_{t+1} + 2\sqrt{1/(4\beta)} A_{t+1} B_{t+1}.$$

Solving this quadratic, we have

$$B_{t+1} \le \sqrt{1/(4\beta)} \cdot A_{t+1} + \sqrt{(1 + 1/(4\beta)) \cdot A_{t+1}^2 + (\beta/\eta) A_t^2 + \sqrt{2 \cdot \|\zeta_{t+1}\|_{\mathrm{sp},\rho}^2} \cdot A_{t+1}},$$

and, after taking square and using an inequality $(a+b)^2 \le 2a^2 + 2b^2$ and $2ab \le a^2 + b^2$

$$B_{t+1}^2 \le (2 + 1/\beta) A_{t+1}^2 + 2(\beta/\eta) A_t^2 + 2\sqrt{2\|\zeta_{t+1}\|_{\mathrm{sp},\rho}^2} \cdot A_{t+1} \le (1/\beta + 3) A_{t+1}^2 + 2(\beta/\eta) A_t^2 + 2\|\zeta_{t+1}\|_{\mathrm{sp},\rho}^2.$$

After plugging in the bound on A_{t+1}^2, A_t^2, and $\|\zeta_{t+1}\|_{\mathrm{sp},\rho}^2$ we conclude the statement.

Convergence in log-probabilities. Finally, we establish convergence in the span seminorm of log-probabilities. Define $C_t \triangleq \|\log \pi_t - \log \pi^\star_\beta\|_{\mathrm{sp},\rho}$. A first-order optimality analysis of the regularized best response in V_t shows that, for each context x, the best response to π_{t+1} (denoted ν_{t+1}) satisfies

$$\beta(1+1/\eta)\log \nu_{t+1}(y|x) = \beta \log \pi^{\mathrm{ref}}(y|x) + (\beta/\eta)\log \pi_t(y|x) - \mathcal{P}(\pi_{t+1}(x) \succ y \mid x) + c_{t+1}(x)\,,$$

where $c_{t+1}(x)$ is a normalizing constant. Similarly, the regularized VNW $\pi^\star_\beta$ satisfies

$$\beta(1+1/\eta)\log \pi^\star_\beta(y|x) = \beta \log \pi^{\mathrm{ref}}(y|x) + (\beta/\eta)\log \pi^\star_\beta(y|x) - \mathcal{P}(\pi^\star_\beta(x) \succ y \mid x) + c^\star(x)\,.$$

Subtracting these two expressions and computing sp, ρ-norm yields

$$\beta(1+1/\eta)\|\log \nu_{t+1} - \log \pi^\star_\beta\|_{\mathrm{sp},\rho} \le (\beta/\eta)\|\log \pi_t - \log \pi^\star_\beta\|_{\mathrm{sp},\rho} + \|\mathcal{P}(\pi^\star_\beta - \pi_{t+1} \succ \cdot)\|_{\mathrm{sp},\rho}.$$

Using Lemma 2 and Pinsker's inequality, one can bound

$$\|\mathcal{P}(\pi^\star_\beta - \pi_{t+1} \succ \cdot)\|^2_{\mathrm{sp},\rho} \le \frac{1}{2}\,\mathrm{KL}_\rho(\pi^\star_\beta \| \pi_{t+1}) = \frac{1}{2}A^2_{t+1}.$$

Moreover, comparing the gradient expression

$$\zeta_{t+1}(x,y) = \nabla_\pi V_t(\pi_{t+1},\pi_{t+1})(x,y) = \mathcal{P}(\pi_{t+1}(x) \succ y \mid x) + \beta \log \frac{\pi_{t+1}(y|x)}{\pi^{\mathrm{ref}}(y|x)} + \beta/\eta \log \frac{\pi_{t+1}(y|x)}{\pi_t(y|x)} + d_{t+1}(x)\mathbf{1}\,,$$

with the best-response equation for ν_{t+1} shows that

$$\|\zeta_{t+1}\|_{\mathrm{sp},\rho} = \beta(1+1/\eta)\|\log \nu_{t+1} - \log \pi_{t+1}\|_{\mathrm{sp},\rho}\,.$$

Thus, we have

$$\beta(1+1/\eta)\|\log \pi_{t+1} - \log \pi^\star_\beta\|_{\mathrm{sp},\rho} \le (\beta/\eta)\|\log \pi_t - \log \pi^\star_\beta\|_{\mathrm{sp},\rho} + A_{t+1}\cdot\sqrt{1/2} + \|\zeta_{t+1}\|_{\mathrm{sp},\rho}\,.$$

or, after rearranging and dividing by $\beta(1+1/\eta)$,

$$C_{t+1} \le \frac{1}{1+\eta}C_t + \frac{A_{t+1}}{\beta(1+1/\eta)\cdot\sqrt{2}} + \frac{\|\zeta_{t+1}\|_{\mathrm{sp},\rho}}{\beta(1+1/\eta)}\,.$$

Taking the square and applying the inequality $(a+b)^2 \le (1+\alpha)a^2 + (1+1/\alpha)b^2$ twice: first time with $\alpha = \eta$ and the second one with $\alpha = 1$ implies

$$C^2_{t+1} \le \frac{1}{1+\eta}C^2_t + \frac{1}{\beta^2\cdot(1+1/\eta)}\left(\frac{A_{t+1}}{\sqrt{2}} + \|\zeta_{t+1}\|_{\mathrm{sp},\rho}\right)^2 \le \frac{1}{1+\eta}C^2_t + \frac{\eta\cdot A^2_{t+1}}{\beta^2(1+\eta)} + \frac{2\eta\cdot\|\zeta_{t+1}\|^2_{\mathrm{sp},\rho}}{\beta^2(1+\eta)}\,.$$

Plugging in the bound on A^2_{t+1} and $\|\zeta_{t+1}\|^2_{\mathrm{sp},\rho}$ implies

$$C^2_{t+1} \le \frac{1}{1+\eta}C^2_t + \frac{\eta\cdot\mathrm{KL}_\rho(\pi^\star_\beta\|\pi_0)}{\beta^2(1+\eta)}(1+\eta/2)^{-(t+1)} + \frac{2\eta(1+\beta^2)\cdot\varepsilon}{\beta^4(1+\eta)}$$

Unrolling this recursion, we achieve for any $t \in \mathbb{N}$

$$C^2_t \le (1+\eta)^{-t}C^2_0 + \frac{2\,\mathrm{KL}_\rho(\pi^\star_\beta\|\pi_0)}{\beta^2}\cdot(1+\eta/2)^{-t} + \frac{2\eta(1+\beta^2)\cdot\varepsilon}{\beta^4(1+\eta)}\cdot\frac{1}{1-1/(1+\eta)}\,,$$

and simplifying the last term we conclude the statement. □

C Self-Play Policy Gradients

In this section, we study a symmetric preference game with KL regularization that appears as a single inner problem of the proximal point method:

$$\max_{\pi\in\Pi}\min_{\pi'\in\Pi}\left\{\mathcal{P}^{\widetilde{\pi}}_\lambda(\pi \succ \pi') \triangleq \mathcal{P}(\pi \succ \pi') - \lambda\,\mathrm{KL}_\rho(\pi\|\widetilde{\pi}) + \lambda\,\mathrm{KL}_\rho(\pi'\|\widetilde{\pi})\right\}. \tag{24}$$

We solve this sub-problem using a self-play policy gradient method.

Policy parametrization. We consider a general policy parameterization: for a parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ let $\theta \mapsto \pi_\theta \in \Pi$ be a corresponding policy. A canonical example is the standard softmax parametrization with $\Theta = \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ and $\pi_\theta(y|x) \propto \mathrm{e}^{\theta_{x,y}}$.

Best-response objective. For an arbitrary competitor policy $\pi \in \Pi$ and a parameter $\theta \in \mathbb{R}^d$, define

$$J^{\widetilde{\pi}}(\theta;\pi) \triangleq \mathcal{P}(\pi \succ \pi_\theta) + \lambda\,\mathrm{KL}_\rho(\pi_\theta\|\widetilde{\pi})\,.$$

We will repeatedly relate $J^{\widetilde{\pi}}$ to the value function $V_\lambda^{\widetilde{\pi}}$ and its best-response value $V_\lambda^{\widetilde{\pi},\star}$ (see Appendix B).

The corresponding parametrized best-response operator and value are defined as

$$\theta_\pi^\star \triangleq \arg\min_{\theta\in\Theta} J^{\widetilde{\pi}}(\theta;\pi), \qquad J^{\widetilde{\pi},\star}(\pi) \triangleq \min_{\theta\in\Theta} J^{\widetilde{\pi}}(\theta;\pi)\,. \tag{25}$$

Our objective instead is to minimize the exploitability gap (or suboptimality) $\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_\theta) = \max_{\pi'\in\Pi}\{\frac{1}{2} - \mathcal{P}_\lambda^{\widetilde{\pi}}(\pi_\theta \succ \pi')\}$. Mirror Prox approximates the solution to (24) by two applications of the best-response operator. However, it requires solving (25) exactly, which is infeasible beyond the tabular setting.

Best-response approximation via gradient step. Rather than solving (25) exactly, we approximate it using a single gradient step:

$$\theta_{t+1} = \mathcal{T}(\theta_t - \gamma_t \cdot g_t)\,, \qquad \pi_{t+1} = \pi_{\theta_{t+1}}\,, \tag{26}$$

where $(\gamma_t)_{t\geq 0}$ is a sequence of step-sizes, $\mathcal{T}$ is a projection-like operator that guarantees policy improvement, and g_t is a stochastic estimator of $\nabla J^{\widetilde{\pi}}(\theta_t;\pi_t)$ computed from a mini-batch of size $B_t \geq 1$. For an example of g_t, we refer to Appendix C.2.

Parametrization assumptions. Let $\Theta_{\mathcal{T}} \triangleq \{\mathcal{T}(\theta) \mid \theta \in \Theta\}$ denote the image of the improvement operator $\mathcal{T}$, and assume $\theta_0 \in \Theta_{\mathcal{T}}$. Following (Yuan et al., 2022), we provide a set of assumptions on the parametrization $\theta \mapsto \pi_\theta$.

Assumption 2 (Lipschitzness of parametrization)**.** For all $\theta, \theta' \in \Theta$

$$\|\pi_\theta - \pi_{\theta'}\|_{1,\rho} \leq G\|\theta - \theta'\|_2$$

Assumption 3 (Smoothness)**.** For any $\pi \in \Pi$, the function $J^{\widetilde{\pi}}(\theta;\pi)$ is $L_{\widetilde{\pi},\lambda}$-smooth, i.e., for all $\theta, \theta' \in \Theta$:

$$-\frac{L_{\widetilde{\pi},\lambda}}{2}\|\theta-\theta'\|_2^2 \leq J^{\widetilde{\pi}}(\theta';\pi) - J^{\widetilde{\pi}}(\theta;\pi) - \langle\nabla J^{\widetilde{\pi}}(\theta;\pi), \theta'-\theta\rangle \leq \frac{L_{\widetilde{\pi},\lambda}}{2}\|\theta-\theta'\|_2^2\,,$$

We notice that smoothness implies the following useful inequality (see, e.g., Nesterov et al. 2018, Theorem 2.1.5)

$$\|\nabla J^{\widetilde{\pi}}(\theta;\pi_\theta)\|_2^2 \leq 2L_{\widetilde{\pi},\lambda}(J^{\widetilde{\pi}}(\theta;\pi_\theta) - J^{\widetilde{\pi},\star}(\pi_\theta)) \leq 2L_{\widetilde{\pi},\lambda}\cdot\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_\theta)\,.$$

Assumption 4 (Approximate Polyak–Łojasiewicz inequality)**.** The function $J^{\widetilde{\pi}}(\theta;\pi)$ satisfies the approximate version of Polyak–Łojasiewicz (PL) inequality: there exists a constant $m_{\widetilde{\pi},\lambda} \in (0, L_{\widetilde{\pi},\lambda}]$ such that for any $\theta \in \Theta_{\mathcal{T}}$

$$\varepsilon_{\mathrm{PL}} + \|\nabla J^{\widetilde{\pi}}(\theta;\pi_\theta)\|_2^2 \geq 2m_{\widetilde{\pi},\lambda}\cdot\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_\theta)\,.$$

In particular, additive error $\varepsilon_{\mathrm{PL}}$ naturally appears in this bound since, in general, $V_\lambda^{\widetilde{\pi},\star}(\pi_\theta) \neq J^{\widetilde{\pi},\star}(\pi_\theta)$ and, thus, we can guarantee only

$$\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_\theta) = V_\lambda^{\widetilde{\pi}}(\pi_\theta;\pi_\theta) - V_\lambda^{\widetilde{\pi},\star}(\pi_\theta) \geq J^{\widetilde{\pi}}(\theta;\pi_\theta) - J^{\widetilde{\pi},\star}(\pi_\theta)\,,$$

since the result of minimization of value over $\Pi_\Theta \triangleq \{\pi_\theta\}_{\theta\in\Theta}$ and the full policy class Π might be different.

Assumption 5 (Improvement operator)**.** For any $\theta \in \Theta$, the operator $\mathcal{T}$ does not increase exploitability:

$$\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_{\mathcal{T}(\theta)}) \le \mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_\theta)\,.$$

This assumption is always satisfied by $\mathcal{T} = \mathrm{Id}$; however, for some particular cases, it is necessary to consider non-trivial $\mathcal{T}$ to satisfy other assumptions (primarily Assumption 4). We notice that $\mathcal{T}$ is *not* a projection in the usual sense: since our objective $\theta \mapsto \mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_\theta)$ is non-convex, it means that a usual (Euclidean) projection may be detrimental for the current optimization progress.

Assumption 6 (Mini-batch gradient noise)**.** Let $(\mathcal{F}_t)_{t\ge0}$ be the filtration induced by the algorithm iterates: $\mathcal{F}_t \triangleq \sigma(\{\theta_j, \pi_j\}_{j\le t})$. For each t, let $B_t \ge 1$ denote the mini-batch size used to construct g_t, and define the gradient bias and noise

$$\xi_t \triangleq g_t - \mathbb{E}[g_t|\mathcal{F}_t]\,, \qquad b_t \triangleq \mathbb{E}[g_t|\mathcal{F}_t] - \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t)\,.$$

Then, for any $t \in \mathbb{N}$, conditionally on $\mathcal{F}_t$, the following holds:

(i) Bounded bias:

$$\|b_t\|_2 \le \varepsilon_{\mathrm{grad}}\,.$$

(ii) Subgaussian tails with variance proxy scaling as $1/B_t$: there exists a constant $\sigma^2_{\widetilde{\pi},\lambda} > 0$ such that for all $u \in \mathbb{R}^d$ and $s \in \mathbb{R}$,

$$\log \mathbb{E}\big[\exp\big(s\cdot\langle u, \xi_t\rangle\big)\big|\mathcal{F}_t\big] \le \frac{s^2\sigma^2_{\widetilde{\pi},\lambda}\cdot\|u\|_2^2}{2B_t}\,.$$

(iii) Subexponential tails for the squared norm with variance proxy scaling as $1/B_t$: there exists a constant $v^2_{\widetilde{\pi},\lambda} > 0$ such that for all $s \in [0, B_t/v^2_{\widetilde{\pi},\lambda}]$,

$$\log \mathbb{E}\big[\exp\big(s\|\xi_t\|_2^2\big)\big|\,\mathcal{F}_t\big] \le s\cdot\frac{v^2_{\widetilde{\pi},\lambda}}{B_t}\,.$$

Remark 2. Assumption 6 captures the standard $1/B_t$ variance reduction from averaging a mini-batch of B_t independent samples. For example, suppose $g_t = \frac{1}{B_t}\sum_{i=1}^{B_t}\hat{g}(\theta_t, \pi_{\theta_t}; Z_{t,i})$, where conditionally on $\mathcal{F}_t$ the random variables $\{Z_{t,i}\}_{i=1}^{B_t}$ are independent and the single-sample noises $\hat{g}(\theta_t, \pi_{\theta_t}; Z_{t,i}) - \mathbb{E}_{Z_{t,i}}[\hat{g}(\theta_t, \pi_{\theta_t}; Z_{t,i})]$ are subgaussian with variance proxy $\bar{\sigma}^2$ (possibly depending on the dimension). Then the averaged noise ξ_t is subgaussian with variance proxy $\bar{\sigma}^2/B_t$, and $\|\xi_t\|^2$ is also subexponential of the form $\bar{v}^2/B_t$ for some $\bar{v}^2 > 0$ (see, e.g., Jin et al. 2019). This covers, in particular, the cases of bounded per-sample gradients and purely Gaussian noise.

In Appendix C.3, we explicitly verify these assumptions of a simple softmax parametrization, and in Appendix C.4, we verify them for a general family of policies under compatible Fisher-nondegenerate parameterization.

C.1 Convergence Guarantees

In the following, we analyze two cases separately: the deterministic and the stochastic. Before that, we prove the following important result, which we split into two parts for the sake of better result exposition.

Lemma 7 (Descent Lemma I)**.** *Assume Assumptions 2-3-5. For the iterates of policy-gradient self-play (26), for any sequence of learning rates $(\gamma_t)_{t\ge0}$ and for any $t \ge 0$, it holds that*

$$\sqrt{\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_{t+1})} \le \sqrt{\frac{\gamma_t^2 G^2}{8\lambda}\|g_t\|_2^2} + \sqrt{\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) - \gamma_t\big\langle\nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), g_t\big\rangle + \Big(L_{\widetilde{\pi},\lambda} + \frac{G^2}{4\lambda}\Big)\frac{\gamma_t^2}{2}\|g_t\|_2^2}\,.$$

Remark 3 (On the additional term.)**.** Compared to standard descent-lemma arguments for a fixed smooth objective, Lemma 7 contains an extra term. This term reflects that the best-response objective is *dynamic*: after each update, the competitor policy π_t changes, so the objective function $\theta \mapsto J^{\widetilde{\pi}}(\theta;\pi_t)$ shifts across iterations.

Proof. We decompose one iteration of (26) as

$$\theta_t^+ \triangleq \theta_t - \gamma_t g_t, \qquad \theta_{t+1} = \mathcal{T}(\theta_t^+), \qquad \pi_t^+ \triangleq \pi_{\theta_t^+}.$$

Step 1: Smoothness in θ. By Assumption 3, for any $t \in \mathbb{N}$,

$$J^{\widetilde{\pi}}(\theta_t^+;\pi_t) \le J^{\widetilde{\pi}}(\theta_t;\pi_t) + \left\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), \theta_t^+ - \theta_t \right\rangle + \frac{L_{\widetilde{\pi},\lambda}}{2}\|\theta_t^+ - \theta_t\|_2^2 .$$

Using $\theta_t^+ - \theta_t = -\gamma_t g_t$, we obtain

$$J^{\widetilde{\pi}}(\theta_t^+;\pi_t) \le J^{\widetilde{\pi}}(\theta_t;\pi_t) - \gamma_t \left\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), g_t \right\rangle + \frac{L_{\widetilde{\pi},\lambda}\gamma_t^2}{2}\|g_t\|_2^2 . \tag{27}$$

Subtracting $V_\lambda^{\widetilde{\pi},\star}(\pi_t)$ from both sides and using Lemma 3, which gives

$$J^{\widetilde{\pi}}(\theta_t;\pi_t) - V_\lambda^{\widetilde{\pi},\star}(\pi_t) = V_\lambda^{\widetilde{\pi}}(\pi_t;\pi_t) - V_\lambda^{\widetilde{\pi},\star}(\pi_t) = \mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t),$$

we obtain

$$V_\lambda^{\widetilde{\pi}}(\pi_t^+;\pi_t) - V_\lambda^{\widetilde{\pi},\star}(\pi_t) \le \mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) - \gamma_t \left\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), g_t \right\rangle + \frac{L_{\widetilde{\pi},\lambda}\gamma_t^2}{2}\|g_t\|_2^2 . \tag{28}$$

Step 2: Relating $J^{\widetilde{\pi}}$ to $\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t^+)$. Using the identity $\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_\theta) = V_\lambda^{\widetilde{\pi}}(\pi_\theta;\pi_\theta) - V_\lambda^{\widetilde{\pi},\star}(\pi_\theta)$, we have

$$V_\lambda^{\widetilde{\pi}}(\pi_t^+;\pi_t) - V_\lambda^{\widetilde{\pi},\star}(\pi_t) = \underbrace{V_\lambda^{\widetilde{\pi}}(\pi_t^+;\pi_t^+) - V_\lambda^{\widetilde{\pi},\star}(\pi_t^+)}_{\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t^+)} + \left(V_\lambda^{\widetilde{\pi}}(\pi_t^+;\pi_t) - V_\lambda^{\widetilde{\pi}}(\pi_t^+;\pi_t^+)\right) + \left(V_\lambda^{\widetilde{\pi},\star}(\pi_t^+) - V_\lambda^{\widetilde{\pi},\star}(\pi_t)\right).$$

Substituting this identity into (28) yields

$$\begin{aligned}\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t^+) \le\ & \mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) - \gamma_t \left\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), g_t \right\rangle + \frac{L_{\widetilde{\pi},\lambda}\gamma_t^2}{2}\|g_t\|_2^2 \\ & + \left(V_\lambda^{\widetilde{\pi}}(\pi_t^+;\pi_t^+) - V_\lambda^{\widetilde{\pi}}(\pi_t^+;\pi_t)\right) + \left(V_\lambda^{\widetilde{\pi},\star}(\pi_t) - V_\lambda^{\widetilde{\pi},\star}(\pi_t^+)\right).\end{aligned} \tag{29}$$

By definition of $V_\lambda^{\widetilde{\pi}}$,

$$V_\lambda^{\widetilde{\pi}}(\pi_t^+;\pi_t^+) - V_\lambda^{\widetilde{\pi}}(\pi_t^+;\pi_t) = \mathcal{P}(\pi_t^+ - \pi_t \succ \pi_t^+).$$

Moreover, the lower bound of Lemma 5 yields

$$V_\lambda^{\widetilde{\pi},\star}(\pi_t^+) - V_\lambda^{\widetilde{\pi},\star}(\pi_t) \ge \mathcal{P}(\pi_t^+ - \pi_t \succ \nu_t^+),$$

where ν_t^+ is a best response to π_t^+. Thus

$$V_\lambda^{\widetilde{\pi},\star}(\pi_t) - V_\lambda^{\widetilde{\pi},\star}(\pi_t^+) \le -\mathcal{P}(\pi_t^+ - \pi_t \succ \nu_t^+).$$

Substituting these into (29) and using bilinearity of $\mathcal{P}$, we obtain

$$\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t^+) \le \mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) - \gamma_t \left\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), g_t \right\rangle + \frac{L_{\widetilde{\pi},\lambda}\gamma_t^2}{2}\|g_t\|_2^2 + \mathcal{P}(\pi_t^+ - \pi_t \succ \pi_t^+ - \nu_t^+). \tag{30}$$

Step 3: Bounding the preference term. Lemma 2 implies

$$\mathcal{P}(\pi_t^+ - \pi_t \succ \pi_t^+ - \nu_t^+) \le \frac{1}{2}\|\pi_t^+ - \pi_t\|_{1,\rho} \cdot \|\nu_t^+ - \pi_t^+\|_{1,\rho}.$$

Assumption 2 gives $\|\pi_t^+ - \pi_t\|_{1,\rho} \le G\|\theta_t^+ - \theta_t\|_2 = G\gamma_t\|g_t\|_2$. By Pinsker's inequality and Lemma 4,

$$\|\nu_t^+ - \pi_t^+\|_{1,\rho}^2 \le 2\,\mathrm{KL}_\rho(\pi_t^+\|\nu_t^+) \le \frac{2}{\lambda}\cdot\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t^+), \quad \text{so} \quad \|\nu_t^+ - \pi_t^+\|_{1,\rho} \le \sqrt{\frac{2}{\lambda}\,\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t^+)}.$$

Consequently,

$$\mathcal{P}(\pi_t^+ - \pi_t \succ \pi_t^+ - \nu_t^+) \le \frac{G\gamma_t}{2}\|g_t\|_2 \cdot \sqrt{\frac{2}{\lambda}\operatorname{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t^+)}\,.$$

As a result, (30) becomes

$$\begin{aligned}\operatorname{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t^+) \le\ & \operatorname{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) - \gamma_t\big\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), g_t\big\rangle + \frac{\gamma_t^2 \cdot L_{\widetilde{\pi},\lambda}}{2}\|g_t\|_2^2 \\ & + \frac{G\gamma_t}{2}\|g_t\|_2 \cdot \sqrt{\frac{2}{\lambda}\operatorname{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t^+)}\,.\end{aligned} \tag{31}$$

Let $S \triangleq \sqrt{\operatorname{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t^+)}$. Then (31) can be written as

$$S^2 - 2S\sqrt{\frac{G^2\gamma_t^2}{8\lambda}\|g_t\|_2^2} \le \operatorname{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) - \gamma_t\big\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), g_t\big\rangle + \frac{\gamma_t^2 \cdot L_{\widetilde{\pi},\lambda}}{2}\|g_t\|_2^2\,.$$

Adding $\frac{G^2\gamma_t^2}{8\lambda}\|g_t\|_2^2$ to both sides and completing the square yields

$$\left(S - \sqrt{\frac{G^2\gamma_t^2}{8\lambda}\|g_t\|_2^2}\right)^2 \le \operatorname{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) - \gamma_t\big\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), g_t\big\rangle + \left(L_{\widetilde{\pi},\lambda} + \frac{G^2}{4\lambda}\right)\frac{\gamma_t^2}{2}\|g_t\|_2^2\,.$$

Taking square roots, we get

$$\begin{aligned}\sqrt{\operatorname{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t^+)} \le\ & \sqrt{\frac{\gamma_t^2 G^2}{8\lambda}\|g_t\|_2^2} \\ & + \sqrt{\operatorname{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) - \gamma_t\big\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), g_t\big\rangle + \left(L_{\widetilde{\pi},\lambda} + \frac{G^2}{4\lambda}\right)\frac{\gamma_t^2}{2}\|g_t\|_2^2}\,.\end{aligned}$$

By Assumption 5, $\operatorname{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_{t+1}) \le \operatorname{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t^+)$, so

$$\begin{aligned}\sqrt{\operatorname{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_{t+1})} \le\ & \sqrt{\frac{\gamma_t^2 G^2}{8\lambda}\|g_t\|_2^2} \\ & + \sqrt{\operatorname{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) - \gamma_t\big\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), g_t\big\rangle + \left(L_{\widetilde{\pi},\lambda} + \frac{G^2}{4\lambda}\right)\frac{\gamma_t^2}{2}\|g_t\|_2^2}\,.\end{aligned}$$

□

Lemma 8 (Descent Lemma II)**.** *Assume Assumptions 2-3-4-5 and $\lambda m_{\widetilde{\pi},\lambda} \ge G^2$. For the iterates of policy-gradient self-play (26), for any sequence of learning rates $(\gamma_t)_{t\ge 0}$ satisfying $\gamma_t \le 1/(2L_{\widetilde{\pi},\lambda})$, and for any $t \ge 0$, it holds that*

$$\begin{aligned}\operatorname{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_{t+1}) \le\ & \left(1 - \frac{\gamma_t m_{\widetilde{\pi},\lambda}}{2}\right)\operatorname{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) - \alpha_t\left\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), \xi_t\right\rangle \\ & + \frac{3L_{\widetilde{\pi},\lambda}}{2}\cdot\gamma_t^2(\|\xi_t\|_2^2 + \|b_t\|_2^2) + \frac{3G^2}{4\lambda\, m_{\widetilde{\pi},\lambda}}\cdot\gamma_t(\|\xi_t\|_2^2 + \|b_t\|_2^2) \\ & + \frac{3}{4}\cdot\gamma_t(\varepsilon_{\mathrm{PL}} + \|b_t\|_2^2)\,,\end{aligned}$$

where $\xi_t \triangleq g_t - \mathbb{E}[g_t|\mathcal{F}_t]$ and $b_t \triangleq \mathbb{E}[g_t|\mathcal{F}_t] - \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t)$ are the gradient noise and bias respectively, $m_{\widetilde{\pi},\lambda}$ is defined in Assumption 4, and

$$\alpha_t \triangleq \frac{1 - \gamma_t m_{\widetilde{\pi},\lambda}/2}{1 - \gamma_t m_{\widetilde{\pi},\lambda}}\cdot\gamma_t\cdot\left(1 - \gamma_t L_{\widetilde{\pi},\lambda} - \frac{G^2}{2\lambda m_{\widetilde{\pi},\lambda}}\right).$$

Remark 4. In this Lemma, we have introduced an additional condition on λ to be a large enough constant. Although hypothetically it might be possible that $m_{\widetilde{\pi},\lambda}$ decreasing with a value of λ, in any known examples $m_{\widetilde{\pi},\lambda}$ non-decreases with an increase of λ, making an inequality $\lambda m_{\widetilde{\pi},\lambda} \ge 0$ solvable. We refer to Proposition 5 and Proposition 6 for examples.

Proof. Starting from Lemma 7, one can take squares and apply an inequality $(a+b)^2 \le (1+1/A_t)a^2 + (1+A_t)b^2$ for any $A_t > 0$:

$$\begin{aligned}\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_{t+1}) &\le \frac{1+A_t}{A_t}\cdot\frac{\gamma_t^2 G^2}{8\lambda}\|g_t\|_2^2 \\ &\quad + (1+A_t)\left(\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) - \gamma_t\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), g_t\rangle + \left(L_{\widetilde{\pi},\lambda} + \frac{G^2}{4\lambda}\right)\frac{\gamma_t^2}{2}\|g_t\|_2^2\right) \\ &= (1+A_t)\left(\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) - \gamma_t\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), g_t\rangle + \left(L_{\widetilde{\pi},\lambda} + \frac{G^2}{4\lambda}\left(1+\frac{1}{A_t}\right)\right)\frac{\gamma_t^2}{2}\|g_t\|_2^2\right).\end{aligned} \tag{32}$$

To simplify expressions, we also upper bound $1+1/A_t$ by $2+1/A_t$ in the last term. The reasons for this loose bound will be more evident lately.

Define the gradient noise $\xi_t \triangleq g_t - \mathbb{E}[g_t|\mathcal{F}_t]$ and bias $b_t \triangleq \mathbb{E}[g_t|\mathcal{F}_t] - \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t)$. Then $\|g_t\|_2^2 = \|\nabla J^{\widetilde{\pi}}(\theta_t;\pi_t) + b_t + \xi_t\|_2^2 \le \|\nabla J^{\widetilde{\pi}}(\theta_t;\pi_t)\|_2^2 + 2\|\xi_t\|_2^2 + 2\|b_t\|_2^2 + 2\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), \xi_t + b_t\rangle$. Substituting this into (32) yields

$$\begin{aligned}\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_{t+1}) \le (1+A_t)\Big(&\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) - \gamma_t\big(1 - \tfrac{\gamma_t}{2}(L_{\widetilde{\pi},\lambda} + \tfrac{G^2}{4\lambda}\cdot\tfrac{2A_t+1}{A_t})\big)\big\|\nabla J^{\widetilde{\pi}}(\theta_t;\pi_t)\big\|_2^2 \\ &- \gamma_t\big(1 - \gamma_t(L_{\widetilde{\pi},\lambda} + \tfrac{G^2}{4\lambda}\cdot\tfrac{2A_t+1}{A_t})\big\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), b_t\big\rangle \\ &- \gamma_t\big(1 - \gamma_t(L_{\widetilde{\pi},\lambda} + \tfrac{G^2}{4\lambda}\cdot\tfrac{2A_t+1}{A_t})\big\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), \xi_t\big\rangle \\ &+ \Big(L_{\widetilde{\pi},\lambda} + \tfrac{G^2}{4\lambda}\cdot\tfrac{2A_t+1}{A_t}\Big)\gamma_t^2(\|\xi_t\|_2^2 + \|b_t\|_2^2)\Big).\end{aligned}$$

Now, we set

$$A_t = \frac{\gamma_t m_{\widetilde{\pi},\lambda}}{2(1-\gamma_t m_{\widetilde{\pi},\lambda})},$$

where $m_{\widetilde{\pi},\lambda}$ is defined in Assumption 4. Note that A_t is positive since $\gamma_t \le 1/(2L_{\widetilde{\pi},\lambda}) \le 1/(2m_{\widetilde{\pi},\lambda})$ thanks to Assumption 3, Assumption 4. One checks that

$$1 + A_t = \frac{1-\gamma_t m_{\widetilde{\pi},\lambda}/2}{1-\gamma_t m_{\widetilde{\pi},\lambda}}, \qquad \frac{1+2A_t}{4\lambda A_t} = \frac{1-\gamma_t m_{\widetilde{\pi},\lambda}}{2\lambda\gamma_t m_{\widetilde{\pi},\lambda}} + \frac{1}{2\lambda} = \frac{1}{2\lambda\gamma_t m_{\widetilde{\pi},\lambda}}.$$

Simplifying the constants, we obtain

$$\begin{aligned}\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_{t+1}) \le \frac{1-\gamma_t m_{\widetilde{\pi},\lambda}/2}{1-\gamma_t m_{\widetilde{\pi},\lambda}}\Big(&\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) \\ &- \gamma_t\Big(1 - \frac{\gamma_t L_{\widetilde{\pi},\lambda}}{2} - \frac{G^2}{4\lambda m_{\widetilde{\pi},\lambda}}\Big)\big\|\nabla J^{\widetilde{\pi}}(\theta_t;\pi_t)\big\|_2^2 && \triangleq (\mathbf{A}) \\ &- \gamma_t\Big(1 - \gamma_t L_{\widetilde{\pi},\lambda} - \frac{G^2}{2\lambda m_{\widetilde{\pi},\lambda}}\Big)\big\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), b_t\big\rangle\Big) \\ - \frac{1-\gamma_t m_{\widetilde{\pi},\lambda}/2}{1-\gamma_t m_{\widetilde{\pi},\lambda}}&\gamma_t\Big(1 - \gamma_t L_{\widetilde{\pi},\lambda} - \frac{G^2}{2\lambda m_{\widetilde{\pi},\lambda}}\Big)\big\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), \xi_t\big\rangle && \triangleq (\mathbf{B}) \\ + \frac{1-\gamma_t m_{\widetilde{\pi},\lambda}/2}{1-\gamma_t m_{\widetilde{\pi},\lambda}}&\Big(L_{\widetilde{\pi},\lambda} + \frac{G^2}{2\lambda\gamma_t m_{\widetilde{\pi},\lambda}}\Big)\frac{\gamma_t^2}{2}\|\xi_t + b_t\|_2^2. && \triangleq (\mathbf{C})\end{aligned}$$

Next, we verify that under our assumptions, the first (deterministic) term contracts. First, we verify that $1 - \gamma_t \cdot L_{\widetilde{\pi},\lambda} - G^2/(2\lambda m_{\widetilde{\pi},\lambda}) \ge 0$. Indeed, since $\gamma_t \le 1/(2L_{\widetilde{\pi},\lambda})$ and $\lambda m_{\widetilde{\pi},\lambda} \ge G^2$, we have

$$1 - \gamma_t \cdot L_{\widetilde{\pi},\lambda} - \frac{G^2}{2\lambda m_{\widetilde{\pi},\lambda}} \ge 1 - \frac{1}{2} - \frac{1}{2} \ge 0.$$

Thus, the coefficient in front of $\big\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), b_t\big\rangle$ is non-negative. By Cauchy–Schwarz and an inequality $ab \le a^2/2 + b^2/2$, we have

$$\begin{aligned}-\gamma_t\Big(1 - \gamma_t\cdot L_{\widetilde{\pi},\lambda} - \tfrac{G^2}{2\lambda m_{\widetilde{\pi},\lambda}}\Big)\big\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), b_t\big\rangle \le \gamma_t\Big(\tfrac{1}{2} - \tfrac{\gamma_t\cdot L_{\widetilde{\pi},\lambda}}{2} - \tfrac{G^2}{4\lambda m_{\widetilde{\pi},\lambda}}\Big)\big\|\nabla J^{\widetilde{\pi}}(\theta_t;\pi_t)\big\|_2^2 \\ + \frac{\gamma_t}{2}\|b_t\|_2^2,\end{aligned}$$

and substituting this into $(\mathbf{A})$ yields

$$(\mathbf{A}) \leq \frac{1-\gamma_t m_{\widetilde{\pi},\lambda}/2}{1-\gamma_t m_{\widetilde{\pi},\lambda}}\Big(\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) - \frac{\gamma_t}{2}\big\|\nabla J^{\widetilde{\pi}}(\theta_t;\pi_t)\big\|_2^2 + \frac{\gamma_t}{2}\|b_t\|_2^2\Big)\,.$$

Thus, the coefficient in front of $\big\|\nabla J^{\widetilde{\pi}}(\theta_t;\pi_t)\big\|_2^2$ is positive. By Assumption 4,

$$\big\|\nabla J^{\widetilde{\pi}}(\theta_t;\pi_t)\big\|_2^2 \geq 2m_{\widetilde{\pi},\lambda}\ \mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) - \varepsilon_{\mathrm{PL}}\,,$$

therefore $(\mathbf{A})$ can be bounded as

$$(\mathbf{A}) \leq (1-\gamma_t m_{\widetilde{\pi},\lambda}/2)\,\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) + \frac{1-\gamma_t m_{\widetilde{\pi},\lambda}/2}{1-\gamma_t m_{\widetilde{\pi},\lambda}}\cdot\Big(\frac{\gamma_t}{2}\varepsilon_{\mathrm{PL}} + \frac{\gamma_t}{2}\|b_t\|_2^2\Big)\,.$$

To bound the last term, we use the fact that since $L_{\widetilde{\pi},\lambda} \geq m_{\widetilde{\pi},\lambda}$, we have $\gamma_t m_{\widetilde{\pi},\lambda} \leq 1/2$ and thus $(1-\gamma_t m_{\widetilde{\pi},\lambda}/2) \leq 3/2\cdot(1-\gamma_t m_{\widetilde{\pi},\lambda})$.

For $(\mathbf{B})$, since the sign of $\langle\nabla J^{\widetilde{\pi}}(\theta_t;\pi_t),\xi_t\rangle$ can be arbitrary, we cannot provide a clear upper bound on this term and we only define a coefficient in front of it as α_t:

$$(\mathbf{B}) = -\alpha_t\cdot\langle\nabla J^{\widetilde{\pi}}(\theta_t;\pi_t),\xi_t\rangle$$

For $(\mathbf{C})$, we also use the bound $(1-\gamma_t m_{\widetilde{\pi},\lambda}/2) \leq 3/2\cdot(1-\gamma_t m_{\widetilde{\pi},\lambda})$, thus

$$(\mathbf{C}) \leq \frac{3L_{\widetilde{\pi},\lambda}}{2}\cdot\gamma_t^2(\|\xi_t\|_2^2+\|b_t\|_2^2) + \frac{3G^2}{4\lambda m_{\widetilde{\pi},\lambda}}\cdot\gamma_t(\|\xi_t\|_2^2+\|b_t\|_2^2)\,.$$

Combining all the bounds, we conclude the statement. □

Remark 5 (On the necessity of the $\gamma_t\|\xi_t\|_2^2$ term)**.** Lemma 8 contains a noise contribution of order $\gamma_t\|\xi_t\|_2^2$, whereas in the usual PL analysis of SGD on a *fixed* smooth objective one typically pays only $\mathcal{O}(\gamma_t^2\|\xi_t\|_2^2)$. This linear-in-$\gamma_t$ variance penalty is intrinsic to our setting.

The key difference is that our progress measure

$$\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) \;=\; V_\lambda^{\widetilde{\pi}}(\pi_t;\pi_t) - V_\lambda^{\widetilde{\pi},\star}(\pi_t)$$

uses the best-response value $V_\lambda^{\widetilde{\pi},\star}(\pi_t)$ as a *moving baseline*. After each update, the opponent changes from π_t to π_{t+1}, so $V_\lambda^{\widetilde{\pi},\star}(\pi_t)$ drifts. Controlling this drift is exactly what produces the extra square-root term in Lemma 7; see (31), which contains (up to constants)

$$\gamma_t\,\|g_t\|_2\,\sqrt{\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t^+)}.$$

With stochastic gradients $g_t = \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t)+\xi_t$, this term injects noise at the level $s_t \triangleq \sqrt{\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t)}$ as an additive perturbation $\Theta(\gamma_t\|\xi_t\|_2)$. Squaring to return to $\mathrm{SubOpt}_t = s_t^2$ produces a cross term of size $\Theta(\gamma_t\, s_t\|\xi_t\|_2)$. To close a recursion in SubOpt_t, this cross term must be absorbed into the deterministic PL decrease, whose scale is only $\Theta(\gamma_t m_{\widetilde{\pi},\lambda})\, s_t^2$. Applying Young's inequality therefore leaves an unavoidable remainder of order $\Theta(\gamma_t/m_{\widetilde{\pi},\lambda})\|\xi_t\|_2^2$ (with constants also involving $1/\lambda$ through the KL-based control of the baseline drift), which explains the $\gamma_t\|\xi_t\|_2^2$ term in Lemma 8.

In particular, this term vanishes when $\xi_t \equiv 0$, but in the stochastic case it becomes the dominant noise contribution unless $\mathbb{E}\|\xi_t\|_2^2$ decays (e.g., via growing mini-batches as in Assumption 6 or via variance reduction).

Proposition 3 (Deterministic rates)**.** Assume Assumptions 2-3-4-5 and $\lambda m_{\widetilde{\pi},\lambda} \geq G^2$. For the iterates of policy-gradient self-play (26) with exact gradients (i.e., $g_t = \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t)$), let a sequence $(\gamma_t)_{t\geq 0}$ be a constant such that $\gamma_t \equiv \gamma \leq 1/(2L_{\widetilde{\pi},\lambda})$ and for any $t \geq 0$, it holds that

$$\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_t) \leq \Big(1-\frac{\gamma m_{\widetilde{\pi},\lambda}}{2}\Big)^t\,\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_0) + \frac{3\cdot\varepsilon_{\mathrm{PL}}}{2m_{\widetilde{\pi},\lambda}}\,,$$

where $m_{\widetilde{\pi},\lambda}$ is defined in Assumption 4.

Proof. Directly follows from Lemma 8 since in this case $\xi_t \equiv b_t \equiv 0$, combined with the fact that $L_{\widetilde{\pi},\lambda} \geq m_{\widetilde{\pi},\lambda}$ and, consequently, $\gamma m_{\widetilde{\pi},\lambda}/2 < 1$ due to the choice of a step-size. □

Proposition 4 (Stochastic rates)**.** Assume Assumptions 2-3-4-5, Assumption 6 and $\lambda m_{\widetilde{\pi},\lambda} \geq G^2$. Define $\kappa_{\widetilde{\pi},\lambda} = \frac{L_{\widetilde{\pi},\lambda}}{m_{\widetilde{\pi},\lambda}} \geq 1$. Define sequences $(\gamma_t)_{t\geq 0}$ and $(B_t)_{t\geq 0}$ as follows

$$\gamma_t = \frac{4t + 32\kappa_{\widetilde{\pi},\lambda} - 2}{m_{\widetilde{\pi},\lambda}(t + 8\kappa_{\widetilde{\pi},\lambda})^2} = \Theta\left(\frac{1}{m_{\widetilde{\pi},\lambda} t}\right), \qquad B_t = \left\lceil \frac{t + 8\kappa_{\widetilde{\pi},\lambda}}{m_{\widetilde{\pi},\lambda}} \right\rceil = \Theta\left(\frac{t}{m_{\widetilde{\pi},\lambda}}\right).$$

Let $\delta \in (0,1)$. Then, for the iterates of policy-gradient self-play (26), for any $t \geq 0$, with probability at least $1-\delta$ it holds that

$$\begin{aligned}
\mathrm{SubOpt}^{\widetilde{\pi}}_{\lambda}(\pi_t) \leq{}& \frac{64 \cdot \kappa^2_{\widetilde{\pi},\lambda} \log(\mathrm{e}/\delta)}{(t + 8\kappa_{\widetilde{\pi},\lambda} - 1)^2} \mathrm{SubOpt}^{\widetilde{\pi}}_{\lambda}(\pi_0) \\
&+ \frac{24 \cdot \kappa_{\widetilde{\pi},\lambda} \log(\mathrm{e}/\delta) \cdot \log(1 + t/(2\kappa_{\widetilde{\pi},\lambda}))}{(t + 8\kappa_{\widetilde{\pi},\lambda} - 1)^2} \left(\frac{675 \cdot \sigma^2_{\widetilde{\pi},\lambda}}{49} + 2v^2_{\widetilde{\pi},\lambda} \right) \\
&+ \frac{6G^2 \cdot v^2_{\widetilde{\pi},\lambda} \log(\mathrm{e}/\delta)}{\lambda m_{\widetilde{\pi},\lambda} \cdot (t + 8\kappa_{\widetilde{\pi},\lambda} - 1)} + \frac{6\log(\mathrm{e}/\delta)}{m_{\widetilde{\pi},\lambda}} \cdot \left(\varepsilon_{\mathrm{PL}} + 15/7 \cdot \varepsilon^2_{\mathrm{grad}}\right).
\end{aligned}$$

Remark 6 (On the growing mini-batch condition)**.** Compared with a standard high-probability analysis of stochastic gradient descent under a PL condition (Madden et al., 2024), our bound requires a growing mini-batch size $B_t \asymp t$. This is not merely a technical artifact: it compensates for a new source of noise compared to the classical PL-SGD setting, as discussed in Remark 5. In particular, the linear-in-γ_t noise term in Lemma 8 would lead to a noise floor of the iterates if a constant mini-batch size were used even with decaying learning rates.

As remarked in (Madden et al., 2024, Appendix D), a similar issue appears in the analysis of projected non-convex SGD, and can be addressed either via a comparable growing mini-batch condition (Ghadimi et al., 2016) or via variance-reduction techniques (J. Reddi et al., 2016), which we do not employ in our setting.

Proof. First, we notice that a proposed sequence of learning rates satisfies the assumption of Lemma 8. Indeed,

$$\gamma_t = \frac{4t + 32\kappa_{\widetilde{\pi},\lambda} - 2}{m_{\widetilde{\pi},\lambda}(t + 8\kappa_{\widetilde{\pi},\lambda})^2} \leq \gamma_0 = \frac{32\kappa_{\widetilde{\pi},\lambda} - 2}{m_{\widetilde{\pi},\lambda} \cdot 64 \cdot \kappa^2_{\widetilde{\pi},\lambda}} \leq \frac{1}{2m_{\widetilde{\pi},\lambda} \cdot \kappa_{\widetilde{\pi},\lambda}} = \frac{1}{2L_{\widetilde{\pi},\lambda}}.$$

Thus, Lemma 8 implies

$$\begin{aligned}
\mathrm{SubOpt}^{\widetilde{\pi}}_{\lambda}(\pi_{t+1}) \leq{}& \left(1 - \frac{\gamma_t m_{\widetilde{\pi},\lambda}}{2}\right) \mathrm{SubOpt}^{\widetilde{\pi}}_{\lambda}(\pi_t) - \alpha_t \left\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), \xi_t \right\rangle \\
&+ \frac{3}{2} L_{\widetilde{\pi},\lambda} \cdot \gamma_t^2 (\|\xi_t\|_2^2 + \|b_t\|_2^2) + \frac{3G^2}{4\lambda\, m_{\widetilde{\pi},\lambda}} \cdot \gamma_t (\|\xi_t\|_2^2 + \|b_t\|_2^2) \\
&+ \frac{3}{4} \gamma_t (\varepsilon_{\mathrm{PL}} + \|b_t\|_2^2).
\end{aligned} \tag{33}$$

We notice that under our choice of $(\gamma_t)_{t\geq 0}$ it holds

$$1 - \frac{\gamma_t m_{\widetilde{\pi},\lambda}}{2} = \frac{(t + 8\kappa_{\widetilde{\pi},\lambda})^2 - 2(t + 8\kappa_{\widetilde{\pi},\lambda}) + 1}{(t + 8\kappa_{\widetilde{\pi},\lambda})^2} = \frac{(t - 1 + 8\kappa_{\widetilde{\pi},\lambda})^2}{(t + 8\kappa_{\widetilde{\pi},\lambda})^2}.$$

Thus, we can multiply both sides of (33) by $(t + 8\kappa_{\widetilde{\pi},\lambda})^2$ and denote $X_t = (t + 8\kappa_{\widetilde{\pi},\lambda} - 1)^2 \cdot \mathrm{SubOpt}^{\widetilde{\pi}}_{\lambda}(\pi_t)$

$$\begin{aligned}
X_{t+1} \leq{}& X_t - \alpha_t (t + 8\kappa_{\widetilde{\pi},\lambda})^2 \left\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), \xi_t \right\rangle \\
&+ (t + 8\kappa_{\widetilde{\pi},\lambda})^2 \left(\frac{3}{2} L_{\widetilde{\pi},\lambda} \gamma_t^2 + \frac{3\gamma_t G^2}{4\lambda m_{\widetilde{\pi},\lambda}} \right) (\|\xi_t\|_2^2 + \|b_t\|_2^2) + \frac{3\gamma_t}{4} (\varepsilon_{\mathrm{PL}} + \|b_t\|_2^2).
\end{aligned}$$

Next, we denote $Y_t = -\alpha_t(t+8\kappa_{\widetilde{\pi},\lambda})^2\langle \nabla J^{\widetilde{\pi}}(\theta_t;\pi_t), \xi_t\rangle$ and for this bound we have

$$
\begin{aligned}
|-\alpha_t(t+8\kappa_{\widetilde{\pi},\lambda})^2| &= \frac{1-\gamma_t m_{\widetilde{\pi},\lambda}/2}{1-\gamma_t m_{\widetilde{\pi},\lambda}} \cdot \gamma_t \cdot \left(1-\gamma_t L_{\widetilde{\pi},\lambda} - \frac{G^2}{2\lambda m_{\widetilde{\pi},\lambda}}\right) \cdot (t+8\kappa_{\widetilde{\pi},\lambda})^2 \\
&\leq \frac{3}{2}\gamma_t (t+8\kappa_{\widetilde{\pi},\lambda})^2 \leq 12 \cdot \frac{t+8\kappa_{\widetilde{\pi},\lambda}-1/2}{m_{\widetilde{\pi},\lambda}} \leq \frac{90\cdot(t+8\kappa_{\widetilde{\pi},\lambda}-1)}{7 m_{\widetilde{\pi},\lambda}}\,,
\end{aligned}
$$

thus, Assumption 6 and Assumption 4 imply for any $s \in \mathbb{R}$

$$
\begin{aligned}
\log \mathbb{E}[\exp(sY_t)|\mathcal{F}_t] &\leq \frac{s^2}{2} \cdot (\alpha_t(t+8\kappa_{\widetilde{\pi},\lambda})^2)^2 \frac{\sigma^2_{\widetilde{\pi},\lambda}}{B_t} \cdot \|\nabla J^{\widetilde{\pi}}(\theta_t;\pi_t)\|_2^2 \\
&\leq \frac{s^2}{2} \cdot \underbrace{\frac{2\cdot 8100 \cdot L_{\widetilde{\pi},\lambda}}{49\cdot m_{\widetilde{\pi},\lambda}\cdot(t+8\kappa_{\widetilde{\pi},\lambda})} \cdot \sigma^2_{\widetilde{\pi},\lambda}}_{\tilde{B}_t^2} \underbrace{(t+8\kappa_{\widetilde{\pi},\lambda}-1)^2\,\mathrm{SubOpt}^{\widetilde{\pi}}_{\lambda}(\pi_t)}_{X_t}\,,
\end{aligned}
$$

where for the last inequality we applied Assumption 3. Finally, we define

$$
\begin{aligned}
Z_t \triangleq\ & (t+8\kappa_{\widetilde{\pi},\lambda})^2\left(\frac{3}{2}L_{\widetilde{\pi},\lambda}\gamma_t^2 + \frac{3\gamma_t G^2}{4\lambda\, m_{\widetilde{\pi},\lambda}}\right)\|\xi_t\|_2^2 && \triangleq Z_{t,1} \\
&+ (t+8\kappa_{\widetilde{\pi},\lambda})^2\left(\frac{3}{2}L_{\widetilde{\pi},\lambda}\gamma_t^2 + \frac{3\gamma_t G^2}{4\lambda\, m_{\widetilde{\pi},\lambda}}\right)\|b_t\|_2^2 && \triangleq Z_{t,2} \\
&+ \frac{3}{4}(t+8\kappa_{\widetilde{\pi},\lambda})^2\gamma_t\cdot(\varepsilon_{\mathrm{PL}} + \|b_t\|_2^2) && \triangleq Z_{t,3}
\end{aligned}
$$

To evaluate this tail behavior, we first bound the first term

$$
Z_{t,1} = (t+8\kappa_{\widetilde{\pi},\lambda})^2\left(\frac{3}{2}L_{\widetilde{\pi},\lambda}\gamma_t^2 + \frac{3\gamma_t G^2}{4\lambda m_{\widetilde{\pi},\lambda}}\right)\|\xi_t\|_2^2 \leq \left(\frac{24 L_{\widetilde{\pi},\lambda}}{(m_{\widetilde{\pi},\lambda})^2} + \frac{3(t+8\kappa_{\widetilde{\pi},\lambda})G^2}{\lambda (m_{\widetilde{\pi},\lambda})^2}\right)\|\xi_t\|_2^2\,,
$$

and the similar bound of $Z_{t,2}$ holds, where we replace $\|\xi_t\|_2^2$ by $\|b_t\|_2^2$, however, we perform one additional simplification that comes from a bound $G^2 \leq \lambda m_{\widetilde{\pi},\lambda}$, thus

$$
Z_{t,2} \leq \left(\frac{24\kappa_{\widetilde{\pi},\lambda}}{m_{\widetilde{\pi},\lambda}} + \frac{3(t+8\kappa_{\widetilde{\pi},\lambda})}{m_{\widetilde{\pi},\lambda}}\right)\|b_t\|_2^2\,.
$$

For the last term, we have

$$
Z_{t,3} = \frac{3}{4}(t+8\kappa_{\widetilde{\pi},\lambda})^2\gamma_t(\varepsilon_{\mathrm{PL}} + \|b_t\|_2^2) \leq \frac{3(t+8\kappa_{\widetilde{\pi},\lambda})}{m_{\widetilde{\pi},\lambda}}(\varepsilon_{\mathrm{PL}} + \|b_t\|_2^2)\,.
$$

thus we can apply Assumption 6 and achieve

$$
\begin{aligned}
\log\, \mathbb{E}[\exp(sZ_t)|\mathcal{F}_t] &\leq s\cdot\left(\frac{24 L_{\widetilde{\pi},\lambda}}{(m_{\widetilde{\pi},\lambda})^2} + \frac{3(t+8\kappa_{\widetilde{\pi},\lambda})G^2}{\lambda(m_{\widetilde{\pi},\lambda})^2}\right)\frac{v^2_{\widetilde{\pi},\lambda}}{B_t} \\
&\quad + s\cdot\left(\frac{24\kappa_{\widetilde{\pi},\lambda}}{m_{\widetilde{\pi},\lambda}} + \frac{3(t+8\kappa_{\widetilde{\pi},\lambda})}{m_{\widetilde{\pi},\lambda}}\right)\varepsilon^2_{\mathrm{grad}} + s\cdot\frac{3(t+8\kappa_{\widetilde{\pi},\lambda})}{m_{\widetilde{\pi},\lambda}}(\varepsilon_{\mathrm{PL}} + \varepsilon^2_{\mathrm{grad}}) \\
&\leq s\left[\left(\frac{24\kappa_{\widetilde{\pi},\lambda}}{t+8\kappa_{\widetilde{\pi},\lambda}} + \frac{3G^2}{\lambda m_{\widetilde{\pi},\lambda}}\right)v^2_{\widetilde{\pi},\lambda} + \frac{24\kappa_{\widetilde{\pi},\lambda}\cdot\varepsilon^2_{\mathrm{grad}}}{m_{\widetilde{\pi},\lambda}} + \frac{3\cdot(t+8\kappa_{\widetilde{\pi},\lambda})}{m_{\widetilde{\pi},\lambda}}(\varepsilon_{\mathrm{PL}} + 2\varepsilon^2_{\mathrm{grad}})\right] \triangleq s\cdot\tilde{C}_t
\end{aligned}
$$

for a range of $s \in [0, 1/\tilde{C}_t)$. Thus, Madden et al. (2024, Theorem 9) implies that for our system of inequalities $X_{t+1} \leq X_t + Y_t + Z_t$, any $t \in \mathbb{N}$ with probability at least $1-\delta$ it holds

$$
X_t \leq K_t \log(\mathrm{e}/\delta)\,,
$$

where K_t is a sequence that satisfies the following equations: $K_{t+1}^2 \geq (K_t + 2\tilde{C}_t)K_{t+1} + \tilde{B}_t^2 K_t$ and $K_0 \geq X_0$, which is satisfied for a sequence $K_{t+1} = K_t + 2\tilde{C}_t + \tilde{B}_t^2$ and $K_0 =$

$64\kappa^2_{\widetilde{\pi},\lambda}\,\mathrm{SubOpt}^{\widetilde{\pi}}_\lambda(\pi_0) \geq X_0$. In particular, we have

$$
\begin{aligned}
K_t &= 64\kappa^2_{\widetilde{\pi},\lambda}\,\mathrm{SubOpt}^{\widetilde{\pi}}_\lambda(\pi_0) + \frac{16200\cdot\kappa_{\widetilde{\pi},\lambda}}{49}\cdot\sigma^2_{\widetilde{\pi},\lambda}\cdot\sum_{s=0}^{t-1}\frac{1}{s+8\kappa_{\widetilde{\pi},\lambda}} + 48\kappa_{\widetilde{\pi},\lambda}\cdot v^2_{\widetilde{\pi},\lambda}\cdot\sum_{s=0}^{t-1}\frac{1}{s+8\kappa_{\widetilde{\pi},\lambda}} \\
&\quad + 6t\cdot\left(\frac{v^2_{\widetilde{\pi},\lambda}\cdot G^2}{\lambda\cdot m_{\widetilde{\pi},\lambda}} + \frac{4\kappa_{\widetilde{\pi},\lambda}\cdot\varepsilon^2_{\mathrm{grad}}}{m_{\widetilde{\pi},\lambda}}\right) + \frac{6}{m_{\widetilde{\pi},\lambda}}\left(\varepsilon_{\mathrm{PL}} + 2\varepsilon^2_{\mathrm{grad}}\right)\cdot\sum_{s=0}^{t-1}(s+8\kappa_{\widetilde{\pi},\lambda}) \\
&\leq 64\kappa^2_{\widetilde{\pi},\lambda}\,\mathrm{SubOpt}^{\widetilde{\pi}}_\lambda(\pi_0) + 24\kappa_{\widetilde{\pi},\lambda}\log(1+t/(2\kappa_{\widetilde{\pi},\lambda}))\cdot\left(\frac{675\sigma^2_{\widetilde{\pi},\lambda}}{49} + 2v^2_{\widetilde{\pi},\lambda}\right) \\
&\quad + 6t\cdot\left(\frac{v^2_{\widetilde{\pi},\lambda}\cdot G^2}{\lambda\cdot m_{\widetilde{\pi},\lambda}} + \frac{4\kappa_{\widetilde{\pi},\lambda}\cdot\varepsilon^2_{\mathrm{grad}}}{m_{\widetilde{\pi},\lambda}}\right) + \frac{6}{m_{\widetilde{\pi},\lambda}}\left(\varepsilon_{\mathrm{PL}} + 2\varepsilon^2_{\mathrm{grad}}\right)\cdot\frac{t(t+16\kappa_{\widetilde{\pi},\lambda}-1)}{2}\,.
\end{aligned}
$$

Before proceeding further, we notice that since $\kappa_{\widetilde{\pi},\lambda}\geq 1$, we can use the following bounds for any $t\geq 0$:

$$
\frac{t}{(t+8\kappa_{\widetilde{\pi},\lambda}-1)^2} \leq \frac{1}{t+8\kappa_{\widetilde{\pi},\lambda}-1}\,, \qquad \frac{4t\kappa_{\widetilde{\pi},\lambda}}{(t+8\kappa_{\widetilde{\pi},\lambda}-1)^2} \leq \frac{1}{7}\,, \qquad \frac{t(t+16\kappa_{\widetilde{\pi},\lambda}-1)}{2(t+8\kappa_{\widetilde{\pi},\lambda}-1)^2} \leq 1\,.
$$

Finally, we divide both bounds on X_t and K_t by $(t+8\kappa_{\widetilde{\pi},\lambda}-1)^2$ and thus we have the following bound for exploitability that holds for any $t\in\mathbb{N}$ with probability at least $1-\delta$

$$
\begin{aligned}
\mathrm{SubOpt}^{\widetilde{\pi}}_\lambda(\pi_t) &\leq \frac{64\cdot\kappa^2_{\widetilde{\pi},\lambda}\log(\mathrm{e}/\delta)}{(t+8\kappa_{\widetilde{\pi},\lambda}-1)^2}\,\mathrm{SubOpt}^{\widetilde{\pi}}_\lambda(\pi_0) \\
&\quad + \frac{24\cdot\kappa_{\widetilde{\pi},\lambda}\log(\mathrm{e}/\delta)\cdot\log(1+t/(2\kappa_{\widetilde{\pi},\lambda}))}{(t+8\kappa_{\widetilde{\pi},\lambda}-1)^2}\left(\frac{675\cdot\sigma^2_{\widetilde{\pi},\lambda}}{49} + 2v^2_{\widetilde{\pi},\lambda}\right) \\
&\quad + \frac{6G^2\cdot v^2_{\widetilde{\pi},\lambda}\log(\mathrm{e}/\delta)}{\lambda m_{\widetilde{\pi},\lambda}\cdot(t+8\kappa_{\widetilde{\pi},\lambda}-1)} + \frac{6(\varepsilon_{\mathrm{PL}}+15/7\cdot\varepsilon^2_{\mathrm{grad}})\log(\mathrm{e}/\delta)}{m_{\widetilde{\pi},\lambda}}\,.
\end{aligned}
$$

□

C.2 Gradient Estimator

As a particular instance, we employ the clipped gradient estimator that allows to control the variance of the estimator via assumptions on the parameterization only.

Given $\theta\in\Theta$, a context $x\in\mathcal{X}$, two actions $y,y'\in\mathcal{Y}$, and an unbiased estimate $p\in[0,1]$ of $\mathcal{P}(y\succ y'|x)$, define the advantage estimator

$$
A^{\pi_\theta}(x,y,y',p) = 1/2 - p + \lambda\log\frac{\pi_\theta(y|x)}{\widetilde{\pi}(y|x)} - \lambda\log\frac{\pi_\theta(y'|x)}{\widetilde{\pi}(y'|x)}\,.
$$

Then, for a threshold parameter $M>0$, we define the clipped advantage estimator

$$
\tilde{A}^{\pi_\theta}_M(x,y,y',p) = \mathrm{clip}_{-M,M}(A^{\pi_\theta}(x,y,y',p)) = \begin{cases} -M & \text{if } A^{\pi_\theta}(x,y,y',p) < -M\,, \\ A^{\pi_\theta}(x,y,y',p) & \text{if } A^{\pi_\theta}(x,y,y',p)\in[-M,M]\,, \\ M & \text{if } A^{\pi_\theta}(x,y,y',p) > M\,. \end{cases}
$$

Then we define two versions of the stochastic gradient estimator:

$$
G(\theta|x,y,y',p) \triangleq \frac{1}{2}\cdot(\nabla\log\pi_\theta(y|x) - \nabla\log\pi_\theta(y'|x))A^{\pi_\theta}(x,y,y',p)\,, \tag{34}
$$

$$
G_M(\theta|x,y,y',p) \triangleq \frac{1}{2}\cdot(\nabla\log\pi_\theta(y|x) - \nabla\log\pi_\theta(y'|x))\tilde{A}^{\pi_\theta}_M(x,y,y',p)\,. \tag{35}
$$

Then we have the following properties of this estimator.

Lemma 9. *Assume $\|\nabla\log\pi_\theta(y|x)\|_2^2 \leq M_g^2$ for any $x\in\mathrm{supp}(\rho), y\in\mathcal{Y}, \theta\in\Theta_{\mathcal{T}}$. Then, the stochastic gradient estimator defined in (34) satisfies the following properties:*

- *A gradient estimator $G(\theta|x,y,y',p)$ is unbiased, i.e., for any θ it holds $\mathbb{E}[G(\theta|x,y,y',p)] = \nabla J^{\widetilde{\pi}}(\theta;\pi_\theta)$.*
- *The clipped gradient estimator $G_M(\theta|x,y,y',p)$ is biased, and its bias satisfies*

$$\|\mathbb{E}[G_M(\theta|x,y,y',p)] - \nabla J^{\widetilde{\pi}}(\theta;\pi_\theta)\|_2 \le 2M_g\lambda(\mathbb{E}_{x\sim\rho}\left[\left(\log\frac{1}{\widetilde{\pi}_{\min}(x)} - \frac{M-1/2}{2\lambda}\right)_+\right] + 2M_g\lambda \mathrm{e}^{-(M-1/2)/(2\lambda)}\,.$$

- *For any θ, it holds almost surely that*

$$\|G_M(\theta|x,y,y',p) - \mathbb{E}[G_M(\theta|x,y,y',p)]\|_2 \le 2M_g \cdot M\,.$$

Proof. For the first part of the statement, we notice that by linearity of expectation, zero expectation of the score function $\mathbb{E}[\nabla \log \pi_\theta(y|x)] = 0$, and symmetry of preferences, we have

$$\mathbb{E}[G(\theta|x,y,y',p)] = \mathbb{E}\left[\nabla \log \pi_\theta(y|x)\left(\mathcal{P}(\pi \succ y \mid x) + \lambda \log\frac{\pi_\theta(y|x)}{\widetilde{\pi}(y|x)}\right)\right],$$

where $x \sim \rho, y, y' \sim \pi_\theta(\cdot|x)$, and $\pi = \pi_\theta$. The last expression is a standard REINFORCE estimator (Williams, 1992) for a regularized RL problem $J^{\widetilde{\pi}}(\theta;\pi_\theta)$, and it is known that the last expression is equal to $\nabla J^{\widetilde{\pi}}(\theta;\pi_\theta)$.

For the second part of the statement, we use the unbiasedness of $G(\theta|x,y,y',p)$ and the definition of the clipped estimator to write

$$\begin{aligned}
\|\mathbb{E}[G_M(\theta|x,y,y',p)] - \nabla J^{\widetilde{\pi}}(\theta;\pi_\theta)\|_2 &= \|\mathbb{E}[G_M(\theta|x,y,y',p)] - \mathbb{E}[G(\theta|x,y,y',p)]\|_2 \\
&= \left\|\mathbb{E}\left[\frac{1}{2}\cdot(\nabla \log \pi_\theta(y|x) - \nabla \log \pi_\theta(y'|x))\left(\tilde{A}^{\pi_\theta}_M(x,y,y',p) - A^{\pi_\theta}(x,y,y',p)\right)\right]\right\|_2 \\
&\le \frac{1}{2}\mathbb{E}\left[\|\nabla \log \pi_\theta(y|x) - \nabla \log \pi_\theta(y'|x)\|_2 \cdot |\tilde{A}^{\pi_\theta}_M(x,y,y',p) - A^{\pi_\theta}(x,y,y',p)|\right] \\
&\le M_g \cdot \mathbb{E}\left[|\tilde{A}^{\pi_\theta}_M(x,y,y',p) - A^{\pi_\theta}(x,y,y',p)|\right].
\end{aligned}$$

Next, we notice that by definition of the clipping operator it holds

$$|\tilde{A}^{\pi_\theta}_M(x,y,y',p) - A^{\pi_\theta}(x,y,y',p)| = (|A^{\pi_\theta}(x,y,y',p)| - M)_+\,,$$

where $(x)_+ = \max\{0,x\}$. Next, define $\ell_\theta(x,y) = \log\frac{\pi_\theta(y|x)}{\widetilde{\pi}(y|x)}$ as the log-ratio random variable. Then, we can rewrite the advantage as

$$A^{\pi_\theta}(x,y,y',p) = 1/2 - p + \lambda(\ell_\theta(x,y) - \ell_\theta(x,y'))\,.$$

Assuming $M \ge 1$, we can use the triangle inequality to write

$$\begin{aligned}
\mathbb{E}[(|A^{\pi_\theta}(x,y,y',p)| - M)_+] &= \mathbb{E}[(|1/2 - p + \lambda(\ell_\theta(x,y) - \ell_\theta(x,y'))| - M)_+] \\
&\le \mathbb{E}[(|1/2 - p| + \lambda|\ell_\theta(x,y)| + \lambda|\ell_\theta(x,y')| - M)_+] \\
&\le \mathbb{E}[(1/2 + \lambda|\ell_\theta(x,y)| + \lambda|\ell_\theta(x,y')| - M)_+] \\
&\le 2\lambda \cdot \mathbb{E}\left[\left(|\ell_\theta(x,y)| - \frac{M-1/2}{2\lambda}\right)_+\right],
\end{aligned}$$

where in the last inequality we used the fact that $(a+b-c)_+ \le (a-c/2)_+ + (b-c/2)_+$ for any $a,b,c \ge 0$, and the fact that y and y' are i.i.d. samples from $\pi_\theta(\cdot|x)$. Next, defining $M' = (M-1/2)/(2\lambda)$, we can further bound

$$\mathbb{E}\left[(|\ell_\theta(x,y)| - M')_+\right] \le \mathbb{E}\left[(\ell_\theta(x,y) - M')_+\right] + \mathbb{E}\left[(-\ell_\theta(x,y) - M')_+\right].$$

For the first term, we apply the following inequality; $\ell_\theta(x,y) = \log\frac{\pi_\theta(y|x)}{\widetilde{\pi}(y|x)} \le \log\frac{1}{\widetilde{\pi}(y|x)} \le \log\frac{1}{\widetilde{\pi}_{\min}(x)}$ for any $x \in \mathcal{X}, y \in \mathcal{Y}$ since $\pi_\theta(y|x) \le 1$. Thus, we have

$$\mathbb{E}\left[(\ell_\theta(x,y) - M')_+\right] \le \mathbb{E}_{x\sim\rho}\left[\left(\log\frac{1}{\widetilde{\pi}_{\min}(x)} - M'\right)_+\right].$$

Next, we study the second term. Let us fix a context $x \in \mathcal{X}$ and define the set $\mathcal{Y}_x^- = \{y \in \mathcal{Y} : \pi_\theta(y|x)/\widetilde{\pi}(y|x) \le \mathrm{e}^{-M'}\}$. Then, we have

$$\begin{aligned}\mathbb{E}\big[(-\ell_\theta(x,y) - M')_+\big] &= \mathbb{E}_{x\sim\rho}\left[\sum_{y\in\mathcal{Y}_x^-} \pi_\theta(y|x)\left(\log\frac{\widetilde{\pi}(y|x)}{\pi_\theta(y|x)} - M'\right)\right] \\ &= \mathbb{E}_{x\sim\rho}\left[\sum_{y\in\mathcal{Y}_x^-} \pi_\theta(y|x)\log\frac{\widetilde{\pi}(y|x)}{\mathrm{e}^{M'}\cdot\pi_\theta(y|x)}\right],\end{aligned}$$

and, applying an inequality $\log(x) \le x - 1$ for any $x \ge 0$, we achieve

$$\mathbb{E}\big[(-\ell_\theta(x,y) - M')_+\big] \le \mathbb{E}_{x\sim\rho}\left[\sum_{y\in\mathcal{Y}_x^-} \pi_\theta(y|x)\left(\frac{\widetilde{\pi}(y|x)}{\mathrm{e}^{M'}\cdot\pi_\theta(y|x)} - 1\right)\right] \le \mathrm{e}^{-M'}.$$

Finally, the last statement of the lemma follows directly from the definition of the clipped estimator. □

Before proceeding further, we notice that the bias of the clipped estimator can be made arbitrarily small by increasing the clipping threshold M; however, its specific range depends on the behavior of the reference policy $\widetilde{\pi}$. In particular, if $\widetilde{\pi}$ has a small minimum probability $\widetilde{\pi}_{\min}(x)$ for some contexts x, then the bias can be large for moderate values of M. Thus, we need to introduce additional assumptions on the reference policy to control this bias effectively.

Lemma 10. *Assume $\|\nabla \log \pi_\theta(y|x)\|_2^2 \le M_g^2$ for any $x \in \mathrm{supp}(\rho), y \in \mathcal{Y}, \theta \in \Theta_{\mathcal{T}}$. Then, for any desired bias level $\varepsilon_{\mathrm{grad}} > 0$, the clipped gradient estimator defined in (35) satisfies the following properties.*

- *If the reference policy satisfies $\mathbb{E}_{x\sim\rho}[\log^2(1/\widetilde{\pi}_{\min}(x))] \le V_{\widetilde{\pi}}$ for some constant $V_{\widetilde{\pi}} > 0$, then under the choice $M = M_2(\varepsilon_{\mathrm{grad}}) \triangleq 1/2 + 4\lambda^2 M_g(V_{\widetilde{\pi}} + 1)/\varepsilon_{\mathrm{grad}}$ the bias is bounded as $\varepsilon_{\mathrm{grad}}$ and it holds*
$$\|G_M(\theta|x,y,y',p) - \mathbb{E}[G_M(\theta|x,y,y',p)]\|_2 \le D_2(\varepsilon_{\mathrm{grad}}),$$
where $D_2(\varepsilon_{\mathrm{grad}}) \triangleq M_g \cdot \left(1 + \frac{8\lambda^2 M_g(V_{\widetilde{\pi}}+1)}{\varepsilon_{\mathrm{grad}}}\right)$.
- *If the reference policy satisfies $\mathbb{E}_{x\sim\rho}[1/\widetilde{\pi}_{\min}(x)] \le D_{\widetilde{\pi}}$ for some constant $D_{\widetilde{\pi}} > 0$, then under the choice $M = M_\infty(\varepsilon_{\mathrm{grad}}) \triangleq 1/2 + 2\lambda\log\left(\frac{2M_g\lambda(D_{\widetilde{\pi}}+1)}{\varepsilon_{\mathrm{grad}}}\right)$ the bias is bounded as $\varepsilon_{\mathrm{grad}}$ and it holds*
$$\|G_M(\theta|x,y,y',p) - \mathbb{E}[G_M(\theta|x,y,y',p)]\|_2 \le D_\infty(\varepsilon_{\mathrm{grad}}),$$
where $D_\infty(\varepsilon_{\mathrm{grad}}) \triangleq M_g \cdot \left(1 + 4\lambda\log\left(\frac{2M_g\lambda(D_{\widetilde{\pi}}+1)}{\varepsilon_{\mathrm{grad}}}\right)\right)$.

Proof. We start from the bias bound provided in Lemma 9:

$$\begin{aligned}\|\mathbb{E}[G_M(\theta|x,y,y',p)] - \nabla J^{\widetilde{\pi}}(\theta;\pi_\theta)\|_2 \le 2M_g\lambda\mathbb{E}_{x\sim\rho}&\left[\left(\log\frac{1}{\widetilde{\pi}_{\min}(x)} - \frac{M-1/2}{2\lambda}\right)_+\right] \\ &+ 2M_g\lambda\mathrm{e}^{-(M-1/2)/(2\lambda)}.\end{aligned}$$

Let us define $M' = (M-1/2)/(2\lambda)$ and then study the first term inside the parentheses and represent it as

$$\mathbb{E}_{x\sim\rho}\left[\left(\log\frac{1}{\widetilde{\pi}_{\min}(x)} - M'\right)_+\right] \le \int_{M'}^\infty \mathbb{P}\left(\log\frac{1}{\widetilde{\pi}_{\min}(x)} \ge u\right)\mathrm{d}u.$$

For the first part of the statement, we apply Markov's inequality to write

$$\mathbb{P}\left(\log\frac{1}{\widetilde{\pi}_{\min}(x)} \ge u\right) \le \frac{\mathbb{E}_{x\sim\rho}[\log^2(1/\widetilde{\pi}_{\min}(x))]}{u^2} \le \frac{V_{\widetilde{\pi}}}{u^2},$$

and thus

$$\mathbb{E}_{x\sim\rho}\left[\left(\log\frac{1}{\widetilde{\pi}_{\min}(x)}-M'\right)_+\right]\le\int_{M'}^{\infty}\frac{V_{\widetilde{\pi}}}{u^2}\mathrm{d}u=\frac{V_{\widetilde{\pi}}}{M'}\,.$$

To control the bias at level $\varepsilon_{\mathrm{grad}}$, we set M' such that $2M_g\lambda(V_{\widetilde{\pi}}/M'+\mathrm{e}^{-M'})\le\varepsilon_{\mathrm{grad}}$. Since $\mathrm{e}^{-M'}\le 1/M'$, it is sufficient to set $M'=2M_g\lambda(V_{\widetilde{\pi}}+1)/\varepsilon_{\mathrm{grad}}$, which leads to the choice of $M_2(\varepsilon_{\mathrm{grad}})$ in the statement of the lemma.

For the second part of the statement, we again start from the integral representation and apply Markov's inequality to write

$$\mathbb{P}\left(\log\frac{1}{\widetilde{\pi}_{\min}(x)}\ge u\right)=\mathbb{P}\left(\frac{1}{\widetilde{\pi}_{\min}(x)}\ge\mathrm{e}^{u}\right)\le\frac{\mathbb{E}_{x\sim\rho}[1/\widetilde{\pi}_{\min}(x)]}{\mathrm{e}^{u}}\le\frac{D_{\widetilde{\pi}}}{\mathrm{e}^{u}}\,,$$

and thus

$$\mathbb{E}_{x\sim\rho}\left[\left(\log\frac{1}{\widetilde{\pi}_{\min}(x)}-M'\right)_+\right]\le\int_{M'}^{\infty}\frac{D_{\widetilde{\pi}}}{\mathrm{e}^{u}}\mathrm{d}u=D_{\widetilde{\pi}}\cdot\mathrm{e}^{-M'}\,.$$

As a result, we have

$$\|\mathbb{E}[G_M(\theta|x,y,y',p)]-\nabla J^{\widetilde{\pi}}(\theta;\pi_\theta)\|_2\le 2M_g\lambda\left(D_{\widetilde{\pi}}\cdot\mathrm{e}^{-M'}+\mathrm{e}^{-M'}\right)=2M_g\lambda(D_{\widetilde{\pi}}+1)\cdot\mathrm{e}^{-M'}\,.$$

To control the bias at level $\varepsilon_{\mathrm{grad}}$, we set M' such that $2M_g\lambda(D_{\widetilde{\pi}}+1)\cdot\mathrm{e}^{-M'}\le\varepsilon_{\mathrm{grad}}$, which leads to the choice of $M_\infty(\varepsilon_{\mathrm{grad}})$ in the statement of the lemma.

□

Next, we provide the properties of the corresponding mini-batch stochastic gradient estimator. At iteration t we draw an i.i.d. mini-batch $(x_j,y_j,y'_j,p_j)_{j\in[B_t]}$ with $x_j\sim\rho$, $y_j,y'_j\sim\pi_{\theta_t}(\cdot|x_j)$ and p_j an unbiased estimator of $\mathcal{P}(y_j\succ y'_j|x_j)$, and set

$$g_t=\frac{1}{B_t}\sum_{j=1}^{B_t}G_{M_k(\varepsilon_{\mathrm{grad}})}(\theta_t|x_j,y_j,y'_j,p_j)\,,\tag{36}$$

where $\varepsilon_{\mathrm{grad}}$ is a desired bias level and $k\in\{2,\infty\}$ indicates which choice of the clipping threshold from Lemma 10 is used. Let $(\mathcal{F}_t)_{t\ge0}$ be the filtration generated by the iterates: $\mathcal{F}_t=\sigma(\{\theta_k,\pi_k\}_{k\le t})$.

Lemma 11. *Assume that the conditions of Lemma 10 hold true for some desired bias level $\varepsilon_{\mathrm{grad}}>0$ and $k\in\{2,\infty\}$. Then, the stochastic gradient estimator defined in (36) satisfies for any $\mathcal{F}_t$-measurable vector $u\in\mathbb{R}^d$*

$$\forall s\in\mathbb{R}:\log\mathbb{E}[\exp(s\langle u,\xi_t\rangle)|\mathcal{F}_t]\le\frac{s^2D_k^2(\varepsilon_{\mathrm{grad}})}{2B_t}\|u\|_2^2\,,$$
$$\forall s\in[0,B_t/(6D_k^2(\varepsilon_{\mathrm{grad}}))]:\log\mathbb{E}\left[\exp(s\|\xi_t\|_2^2)|\mathcal{F}_t\right]\le 6s\cdot D_k^2(\varepsilon_{\mathrm{grad}})/B_t\,.$$

In particular, it satisfies Assumption 6 with a bias $\varepsilon_{\mathrm{grad}}$, subgaussian constant $\sigma^2_{\widetilde{\pi},\lambda}=D_k^2(\varepsilon_{\mathrm{grad}})$ and subexponential constant $v^2_{\widetilde{\pi},\lambda}=6D_k^2(\varepsilon_{\mathrm{grad}})$.

Proof. We have to compute an exact subgaussianity constant for a random vector ξ_t bounded almost surely by $D_k(\varepsilon_{\mathrm{grad}})$, which is standard (see, e.g., Vershynin (2018)) and we write the proof for the sake of completeness. Let us fix a vector $u\in\mathbb{R}^d$ and define $\xi_t^1,\dots,\xi_t^{B_t}$ as the corresponding components of the noise vector $\xi_t^j=G_{M_k}(\theta_t|x_j,y_j,y'_j,p_j)-\mathbb{E}[G_{M_k}(\theta_t|x_j,y_j,y'_j,p_j)]$. Then, we define $X_i=\langle u,\xi_t^i\rangle$ as a centered bounded random variable with $|X_i|\le\|u\|_2D_k(\varepsilon_{\mathrm{grad}})$, and thus (conditional) Hoeffding's lemma (Hoeffding, 1963) implies

$$\mathbb{E}[\mathrm{e}^{sX_i}|\mathcal{F}_t]\le\exp\left(\frac{s^2\cdot\|u\|_2^2D_k^2(\varepsilon_{\mathrm{grad}})}{2}\right)\qquad\forall s\in\mathbb{R}\,,$$

and, using independence, we have $\mathbb{E}[\mathrm{e}^{s\langle u,\xi_t\rangle}|\mathcal{F}_t]\le\mathrm{e}^{s^2\|u\|_2^2D_k^2(\varepsilon_{\mathrm{grad}})/(2B_t)}$. To control the norm of ξ_t, we apply Pinelis (1994, Theorem 3.5):

$$\mathbb{P}[\|\xi_t\|_2\ge u|\mathcal{F}_t]\le 2\exp\left(-\frac{u^2\cdot B_t}{2D_k^2(\varepsilon_{\mathrm{grad}})}\right),$$

thus, defining $v^2 = D_k^2(\varepsilon_{\text{grad}})/B_t$, for a $s > 0$ the following holds

$$\mathbb{E}[\exp(s\|\xi_t\|_2^2)|\mathcal{F}_t] = \int_0^\infty \mathbb{P}[\exp(s\|\xi_t\|_2^2) > u|\mathcal{F}_t]\mathrm{d}u$$
$$\leq 1 + \int_1^\infty \mathbb{P}\left[\|\xi_t\|_2 > \sqrt{\frac{1}{s}\log u}|\mathcal{F}_t\right] \leq 1 + 2\int_1^\infty u^{-\frac{1}{2v^2 s}}\mathrm{d}u\,.$$

The right-hand side is finite as $s < 1/(4v^2) < 1/(2v^2)$, thus we have

$$\mathbb{E}[\exp(s\|\xi_t\|_2^2)] \leq 1 + 2\frac{1}{\frac{1}{2v^2 s} - 1} = \frac{1 + 2v^2 s}{1 - 2v^2 s} \leq \exp(6v^2 s)\,,$$

where in the last inequality we used an inequality $(1+x)/(1-x) \leq \exp(3x)$ for $x \leq [0, 1/2]$. □

C.3 Verification for the Softmax Parametrization

In this section, we verify Assumptions 2–6 for the standard softmax parametrization with an appropriate improvement operator and stochastic gradient estimator. For simplicity, we assume a context-free setting, although a generalization to contextual setting is straightforward due to separated optimization over any $x \in \mathcal{X}$.

Parametrization. We fix $\Theta = \mathbb{R}^{\mathcal{Y}}$ and define

$$\pi_\theta = \mathrm{softmax}(\theta), \qquad \pi_\theta(y) = \frac{\exp(\theta_y)}{\sum_{y'}\exp(\theta_{y'})}\,. \tag{37}$$

We also fix a reference policy $\nu \in \Delta_{\mathcal{Y}}$ with full support, and define an improvement operator $\mathcal{T} := \mathcal{T}_\tau^\nu$ in Section C.3.2 below, where $\tau \in (0, \tau_0]$ is chosen as in Lemma 16. We will show that this choice of $\mathcal{T}$ satisfies Assumption 5 and yields a uniform PL constant for Assumption 4.

Proposition 5 (Softmax parametrization satisfies Assumptions 2–6)**.** Fix any full-support reference policy $\nu \in \Delta_{\mathcal{Y}}$, a constant $\varepsilon_{\text{grad}} > 0$, and let $\tau = \tau_0$ be as in Lemma 16. Consider the softmax parametrization (37) with improvement operator $\mathcal{T} := \mathcal{T}_\tau^\nu$ and stochastic gradient estimator g_t defined in (36). Then, Assumptions 2–6 hold with constants

$$G = 1, \quad L_{\widetilde{\pi},\lambda} = \tfrac{5}{2}\big(1 + \lambda\log(1/\widetilde{\pi}_{\min})\big) + \lambda(4 + \log|\mathcal{Y}|), \quad m_{\widetilde{\pi},\lambda} = \lambda \mathrm{e}^{-2/\lambda}\cdot c_\nu^2, \quad \varepsilon_{\text{PL}} = 0\,,$$
$$\varepsilon_{\text{grad}} = \varepsilon_{\text{grad}}, \qquad \sigma^2_{\widetilde{\pi},\lambda} = D^2_{\widetilde{\pi},\lambda}(\varepsilon_{\text{grad}}), \qquad v^2_{\widetilde{\pi},\lambda} = 6D^2_{\widetilde{\pi},\lambda}(\varepsilon_{\text{grad}})\,,$$

where

$$c_\nu \triangleq \exp(\min\{-2\|\log\widetilde{\pi} - \log\nu\|_{\text{sp}}, \log(\nu_{\min}/(1+\nu_{\min}))\})\cdot\nu_{\min}\,,$$
$$D^2_{\widetilde{\pi},\lambda}(\varepsilon_{\text{grad}}) \triangleq 2\left(1 + 4\lambda\log\left(\frac{2\sqrt{2}\cdot\lambda(1+\widetilde{\pi}_{\min})}{\widetilde{\pi}_{\min}}\right) + 4\lambda\log(1/\varepsilon_{\text{grad}})\right)^2.$$

and $\nu_{\min} = \min_y \nu(y)$, $\widetilde{\pi}_{\min} = \min_y \widetilde{\pi}(y)$.

Proof. Assumption 2 follows from Lemma 12. Assumption 3 follows from Lemma 13. Assumption 5 and the uniform PL condition of Assumption 4 follow from Lemmas 16 and Corollary 3. Finally, Assumption 6 follows from Lemma 17. □

Corollary 2 (SPG iteration and sample complexity: tabular softmax)**.** *Fix $\delta \in (0,1)$ and $\varepsilon \in (0,1)$. Consider the context-free tabular softmax parametrization from Appendix C.3 with improvement operator $\mathcal{T} = \mathcal{T}_{\tau_0}^\nu$ from Lemma 16, and run SPG (26) with (γ_t, B_t) as in Proposition 4. Assume $\widetilde{\pi}_{\min} = \min_{y\in\mathcal{Y}}\widetilde{\pi}(y) > 0$ and let $m_{\widetilde{\pi},\lambda} = \lambda\mathrm{e}^{-2/\lambda}c_\nu^2$ be as in Corollary 3 (so $\lambda m_{\widetilde{\pi},\lambda} = \lambda^2\mathrm{e}^{-2/\lambda}c_\nu^2$). Assume $\lambda m_{\widetilde{\pi},\lambda} \geq 1$ (e.g., it suffices that $\lambda \geq \max\{2, \mathrm{e}^{1/2}/c_\nu\}$).*

Choose the clipped pairwise estimator from Lemma 17 with bias level

$$\varepsilon^2_{\text{grad}} = \frac{7\, m_{\widetilde{\pi},\lambda}\,\varepsilon}{180\log(\mathrm{e}/\delta)} \qquad \textit{(so in particular } M = \tfrac{1}{2} + 2\lambda\log\big(\tfrac{2\sqrt{2}\,\lambda(1+\widetilde{\pi}_{\min})}{\widetilde{\pi}_{\min}\varepsilon_{\text{grad}}}\big) = \Theta(\log(1/\varepsilon))\textit{)}.$$

Then with probability at least $1-\delta$ *the iterate* π_T *satisfies* $\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_T) \le \varepsilon$ *(equivalently,* π_T *is an* ε*-VNW for the* λ*-regularized game) once*

$$N_{\mathrm{iter}}(\varepsilon,\delta) = T = \widetilde{\mathcal{O}}\left(\frac{v^2_{\widetilde{\pi},\lambda}}{\lambda m_{\widetilde{\pi},\lambda}} \cdot \frac{\log(\mathrm{e}/\delta)}{\varepsilon}\right), \qquad N_{\mathrm{sample}}(\varepsilon,\delta) = \sum_{t=0}^{T-1} B_t = \widetilde{\mathcal{O}}\left(\frac{T^2}{m_{\widetilde{\pi},\lambda}}\right).$$

In particular (since here $v^2_{\widetilde{\pi},\lambda} = 6D^2_{\widetilde{\pi},\lambda}(\varepsilon_{\mathrm{grad}}) = \widetilde{\mathcal{O}}(\log^2(1/\varepsilon))$*,* $N_{\mathrm{iter}}(\varepsilon) = \widetilde{\mathcal{O}}(\varepsilon^{-1})$ *and* $N_{\mathrm{sample}}(\varepsilon) = \widetilde{\mathcal{O}}(\varepsilon^{-2})$ *up to polylog factors.*

Proof. By Proposition 5 and Lemma 17, Assumptions 2–6 hold with $G = 1$, $\varepsilon_{\mathrm{PL}} = 0$, and the stated $m_{\widetilde{\pi},\lambda}$ and $v^2_{\widetilde{\pi},\lambda}$. The condition $\lambda m_{\widetilde{\pi},\lambda} \ge G^2$ ensures Proposition 4 applies. With the chosen $\varepsilon_{\mathrm{grad}}$, the bias term in Proposition 4 satisfies

$$\frac{6\log(\mathrm{e}/\delta)}{m_{\widetilde{\pi},\lambda}} \cdot \frac{15}{7}\varepsilon^2_{\mathrm{grad}} \le \varepsilon/2.$$

Taking $T = \widetilde{\mathcal{O}}\big(\frac{v^2_{\widetilde{\pi},\lambda}}{\lambda m_{\widetilde{\pi},\lambda}} \cdot \frac{\log(\mathrm{e}/\delta)}{\varepsilon}\big)$ makes the remaining (decaying) terms in Proposition 4 at most $\varepsilon/2$, hence $\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_T) \le \varepsilon$. Finally, $B_t = \lceil (t + 8\kappa_{\widetilde{\pi},\lambda})/m_{\widetilde{\pi},\lambda} \rceil$ implies $\sum_{t<T} B_t = \widetilde{\mathcal{O}}(T^2/m_{\widetilde{\pi},\lambda})$. □

C.3.1 Lipschitzness and smoothness

Lemma 12 (Lipschitz parametrization, Lemma 24 of Mei et al. 2020)**.** *For any* $\theta, \theta' \in \Theta$ *it holds*

$$\|\pi_\theta - \pi_{\theta'}\|_1 \le \|\theta - \theta'\|_{\mathrm{span}} \le \|\theta - \theta'\|_2.$$

In particular, Assumption 2 holds with $G = 1$.

Lemma 13 (Smoothness, cf. Lemmas 2 and 14 of Mei et al. 2020)**.** *The softmax parametrization (37) satisfies Assumption 3 with*

$$L_{\widetilde{\pi},\lambda} = \tfrac{5}{2}\big(1 + \lambda\log(1/\widetilde{\pi}_{\min})\big) + \lambda(4 + \log|\mathcal{Y}|)\,,$$

where $\widetilde{\pi}_{\min} = \min_y \widetilde{\pi}(y)$.

Proof. The result follows from Mei et al. (2020) by (i) considering a one-state bandit MDP, (ii) absorbing the cross-entropy part of the KL term into the reward and rescaling, and (iii) taking the discount factor $\gamma = 0$. □

Lemma 14 (Non-Uniform Polyak–Łojasiewicz inequality, Lemma 15 of Mei et al. 2020)**.** *A function* $J^{\widetilde{\pi}}(\theta;\pi)$ *satisfies non-uniform Polyak–Łojasiewicz inequality with a constant* $c_\lambda(\theta) = \lambda \min_y \pi_\theta^2(y)$*, i.e., for any* $\theta \in \Theta$ *and* $\pi \in \Pi$

$$\|\nabla J^{\widetilde{\pi}}(\theta;\pi)\|_2^2 \ge 2c_\lambda(\theta)(J^{\widetilde{\pi}}(\theta;\pi) - J^{\widetilde{\pi},\star}(\pi))\,.$$

Proof. Follows from Lemma 15 of Mei et al. (2020) taking $S = 1, \gamma = 0$. □

We note that, according to Mei et al. (2020), this parameterization, by default, violates the usual PL-condition (Assumption 4) because its coefficient depends on the minimal probability of the current policy, which can become arbitrarily small during the algorithm iterates. To address this, we introduce an improvement operator that enables control over the minimum probability.

C.3.2 Improvement Operator and Uniform PL

We use a minimal-probability truncation-style argument, similar to that of Zhang et al. (2021) and Labbi et al. (2025). Let us fix a policy $\nu \in \Pi$ with full support and a constant $\tau > 0$. For any policy π define a ratio w.r.t. ν as $r_\nu(y) \triangleq \pi(y)/\nu(y)$, the low-ratio set $\mathcal{Y}_\tau(\pi) \triangleq \{y \in \mathcal{Y} \mid r_\nu(y) < \tau\}$, and the max-ratio action $y_{\max} = \arg\max_{y\in\mathcal{Y}} r_\nu(y)$, where ties are resolved arbitrarily.

Next, we define the improved ratios r'_ν by

$$r'_\nu(y) \triangleq \begin{cases} \tau & y \in \mathcal{Y}_\tau(\pi)\,, \\ r_\nu(y) - \frac{1}{\nu(y_{\max})} \sum_{z \in \mathcal{Y}_\tau(\pi)} \nu(z)(\tau - r_\nu(z)) & y = y_{\max}\,, \\ r_\nu(y) & \text{otherwise}\,. \end{cases} \tag{38}$$

and the corresponding policy update $\pi^+ = \mathcal{U}^\nu_\tau(\pi)$ via $\pi^+(y) \triangleq \nu(y) \cdot r'_\nu(y)$. It is easy to verify that π^+ defines a correct probability distribution.

Lemma 15. *Let*

$$\tau_0 \;\triangleq\; \min\Big\{\exp\big(-\tfrac{1}{\lambda} - 2\|\log\widetilde{\pi} - \log\nu\|_{\mathrm{sp}}\big),\; (1 + 1/\nu_{\min})^{-1}\Big\},$$

where $\nu_{\min} = \min_{y\in\mathcal{Y}} \nu(y)$. *Then for any* $\tau \in (0, \tau_0]$ *and for* $\pi^+ = \mathcal{U}^\nu_\tau(\pi)$ *it holds*

$$\mathrm{SubOpt}^{\widetilde{\pi}}_\lambda(\pi^+) \le \mathrm{SubOpt}^{\widetilde{\pi}}_\lambda(\pi)\,, \qquad \forall y \in \mathcal{Y} : \pi^+(y) \ge \tau \cdot \nu(y)\,.$$

Proof. Without loss of generality, we can assume $|\mathcal{Y}_\tau(\pi)| > 0$ since otherwise $\pi^+ = \pi$ and the second condition is trivially satisfied. Let us define $\delta_\tau(\pi) = \sum_{y\in\mathcal{Y}_\tau(\pi)} \nu(y) \cdot (\tau - r_\nu(y)) \ge 0$. In terms of this accumulated ratio, we have

$$\|\pi - \pi^+\|_1 = \sum_{y\in\mathcal{Y}} \nu(y)|r_\nu(y) - r'_\nu(y)| = \sum_{y\in\mathcal{Y}_\tau(\pi)} \nu(y)\cdot(r'_\nu(y) - r_\nu(y)) + \nu(y_{\max})(r_\nu(y) - r'_\nu(y)) = 2\delta_\tau(\pi)\,. \tag{39}$$

Then we notice that $\mathrm{SubOpt}^{\widetilde{\pi}}_\lambda(\pi^+) \le \mathrm{SubOpt}^{\widetilde{\pi}}_\lambda(\pi)$ follows from the following inequality for any competitor policy $\pi^c \in \Pi$

$$\Delta(\pi^c) \triangleq \mathcal{P}^{\widetilde{\pi}}_\lambda(\pi^+ \succ \pi^c) - \mathcal{P}^{\widetilde{\pi}}_\lambda(\pi \succ \pi^c) \ge 0\,.$$

To show that, we apply the following decomposition

$$\Delta(\pi^c) = \mathcal{P}^{\widetilde{\pi}}_\lambda(\pi^+ \succ \pi^c) - \mathcal{P}^{\widetilde{\pi}}_\lambda(\pi \succ \pi^c) = \underbrace{\sum_{y\in\mathcal{Y}} (\pi^+(y) - \pi(y))\mathcal{P}(y \succ \pi^c)}_{T_1} + \lambda \underbrace{\big[\mathrm{KL}(\pi\|\widetilde{\pi}) - \mathrm{KL}(\pi^+\|\widetilde{\pi})\big]}_{T_2}\,.$$

For the first term, we have the following lower bound

$$T_1 = \sum_{y\in\mathcal{Y}} (\pi^+(y) - \pi(y)) \cdot \mathcal{P}(y \succ \pi^c) \ge -\|\pi^+ - \pi\|_1 \|\mathcal{P}(y \succ \pi^c)\|_{\mathrm{sp}} \ge -\delta_\tau(\pi)\,,$$

since $\pi^+ - \pi$ is orthogonal to $\mathbf{1}$, $\|\mathcal{P}(y \succ \pi^c)\|_{\mathrm{sp}} \le 1/2$, and (39) provides an expression for a ℓ_1-norm.

For the second term, we first notice that for any policy π' it holds

$$\mathrm{KL}(\pi'\|\widetilde{\pi}) = \sum_{y\in\mathcal{Y}} \pi'(y) \log \frac{\pi'(y)}{\nu(y)} + \sum_{y\in\mathcal{Y}} \pi'(y) \log \frac{\nu(y)}{\widetilde{\pi}(y)} = \mathrm{KL}(\pi'\|\nu) + \sum_{y\in\mathcal{Y}} \pi'(y) \log \frac{\nu(y)}{\widetilde{\pi}(y)}\,,$$

thus

$$T_2 = \underbrace{\mathrm{KL}(\pi\|\nu) - \mathrm{KL}(\pi^+\|\nu)}_{T_3} + \sum_{y\in\mathcal{Y}} (\pi(y) - \pi^+(y)) \log \frac{\nu(y)}{\widetilde{\pi}(y)} \ge T_3 - 2\delta_\tau(\pi)\|\log\nu - \log\widetilde{\pi}\|_{\mathrm{sp}}\,.$$

Finally, to analyze T_3, we reformulate KL-divergence as an f-divergence for $f(x) \triangleq x\log x - (x-1)$ and have

$$T_3 = \sum_{y\in\mathcal{Y}} \nu(y)[f(r_\nu(y)) - f(r'_\nu(y))] = \underbrace{\sum_{y\in\mathcal{Y}_\tau(\pi)} \nu(y)[f(r_\nu(y)) - f(r'_\nu(y))]}_{T_{3,1}} + \underbrace{\nu(y_{\max})(f(r_\nu(y_{\max})) - f(r'_\nu(y_{\max})))}_{T_{3,2}}\,.$$

Due to convexity of f, we have for any $u, v \in \mathbb{R}$: $f(u) - f(v) \geq f'(v)(u - v) = \log v \cdot (u - v)$, and thus

$$T_{3,1} = \sum_{y \in \mathcal{Y}_\tau(\pi)} \nu(y)[f(r_\nu(y)) - f(r'_\nu(y))] \geq \sum_{y \in \mathcal{Y}_\tau(\pi)} \nu(y) f'(r'_\nu(y))(r_\nu(y) - r'_\nu(y)) \,.$$

For all $y \in \mathcal{Y}_\tau(\pi)$ it holds $r_\nu(y) < r'_\nu(y)$ and $r'_\nu(y) \geq \tau$ and, as a result, $f'(r'_\nu(y)) \geq f'(\tau) = \log(\tau)$. Thus, for $\tau \leq 1$,

$$T_{3,1} \geq \log(1/\tau) \cdot \sum_{y \in \mathcal{Y}_\tau(\pi)} \nu(y)(r'_\nu(y) - r_\nu(y)) = \log(1/\tau) \cdot \delta_\tau(\pi) \,.$$

To analyze the second term, let us define $\nu(\mathcal{Y}_\tau(\pi)) = \sum_{y \in \mathcal{Y}_\tau(\pi)} \nu(y)$. Then we have

$$\begin{aligned} 1 = \sum_{y \in \mathcal{Y}} \nu(y) r_\nu(y) &= \sum_{y \in \mathcal{Y}_\tau(\pi)} \nu(y) \underbrace{r_\nu(y)}_{\leq \tau} + \sum_{y \notin \mathcal{Y}_\tau(\pi)} \nu(y) \underbrace{r_\nu(y)}_{\leq r_\nu(y_{\max})} \\ &\leq \nu(\mathcal{Y}_\tau(\pi)) \cdot \tau + (1 - \nu(\mathcal{Y}_\tau(\pi))) \cdot r_\nu(y_{\max}) \,. \end{aligned}$$

As a result, we have

$$r_\nu(y_{\max}) \geq \frac{1 - \nu(\mathcal{Y}_\tau(\pi))\tau}{1 - \nu(\mathcal{Y}_\tau(\pi))} = 1 + \frac{\nu(\mathcal{Y}_\tau(\pi))}{1 - \nu(\mathcal{Y}_\tau(\pi))} \cdot (1 - \tau) \geq 1 + \nu(\mathcal{Y}_\tau(\pi))(1 - \tau) \,.$$

At the same time, we have $\delta_\tau(\pi) \leq \tau \nu(\mathcal{Y}_\tau(\pi))$, and thus

$$r'_\nu(y_{\max}) \geq 1 + \nu(\mathcal{Y}_\tau(\pi)) \cdot (1 - \tau) - \frac{\delta_\tau(\pi)}{\nu(y_{\max})} \geq 1 + \nu(\mathcal{Y}_\tau(\pi)) \cdot \left(1 - \tau\left(1 + \frac{1}{\nu_{\min}}\right)\right),$$

Taking $\tau \leq (1 + 1/\nu_{\min})^{-1}$, we have $r'_\nu(y_{\max}) \geq 1$, and, thus, by the same convexity argument

$$T_{3,2} \geq \nu(y_{\max}) \underbrace{\log(r'_\nu(y_{\max}))}_{\geq 0} \underbrace{(r_\nu(y_{\max}) - r'_\nu(y_{\max}))}_{\geq 0} \geq 0 \,.$$

Combining all bounds together, we have

$$\begin{aligned} \Delta(\pi^c) &\geq -(1 + 2\lambda \|\log \nu - \log \widetilde{\pi}\|_{\mathrm{sp}}) \delta_\tau(\pi) + \lambda \log(1/\tau) \cdot \delta_\tau(\pi) \\ &\geq \delta_\tau(\pi) \cdot (\lambda \log(1/\tau) - 1 - 2\lambda \|\log \nu - \log \widetilde{\pi}\|_{\mathrm{sp}}) \,, \end{aligned}$$

which is non-negative for all

$$0 < \tau \leq \exp\left\{-\frac{1}{\lambda} - 2\|\log \widetilde{\pi} - \log \nu\|_{\mathrm{sp}}\right\}.$$

The second part of the statement for $y \in \mathcal{Y}_\tau(\pi)$ follows automatically, whereas for $y_{\max}$ it follows from a bound $r'_\nu(y_{\max}) \geq 1$. □

We then lift this operator to parameter space as an operator $\mathcal{T}^\nu_\tau \colon \mathbb{R}^{|\mathcal{Y}|} \to \mathbb{R}^{|\mathcal{Y}|}$, defined as

$$\mathcal{T}^\nu_\tau(\theta) = \log \mathcal{U}^\nu_\tau(\pi_\theta) \,.$$

Alternatively, we can define it (equivalently) as follows

$$\mathcal{T}^\nu_\tau(\theta)(y) \triangleq \theta(y) + \begin{cases} \log(\tau \cdot \nu(y)/\pi_\theta(y)) \,, & \pi_\theta(y)/\nu(y) < \tau \,, \\ \log\Big(1 - \sum_{y'} [\nu(y') \cdot \tau - \pi_\theta(y')]_+ / \pi_\theta(y)\Big) & y = \arg\max_{y' \in \mathcal{Y}} \frac{\pi_\theta(y')}{\nu(y')} \,, \\ 0 \,, & \text{otherwise} \,. \end{cases} \tag{40}$$

where $[x]_+ = \max\{x, 0\}$. For this operator, the following lemma is a straightforward consequence.

Lemma 16. *Let*

$$\tau_0 \;\triangleq\; \min\Big\{\exp\big(-\tfrac{1}{\lambda} - 2\|\log \widetilde{\pi} - \log \nu\|_{\mathrm{sp}}\big),\; (1 + 1/\nu_{\min})^{-1}\Big\},$$

where $\nu_{\min} = \min_{y \in \mathcal{Y}} \nu(y)$. Then for any $\tau \in (0, \tau_0]$ and for $\theta^+ = \mathcal{T}^\nu_\tau(\theta)$ it holds

$$\mathrm{SubOpt}^{\widetilde{\pi}}_\lambda(\pi_{\theta^+}) \leq \mathrm{SubOpt}^{\widetilde{\pi}}_\lambda(\pi_\theta) \,, \qquad \forall y \in \mathcal{Y} : \pi_{\theta^+}(y) \geq \tau \cdot \nu(y) \,.$$

In particular, Assumption 5 holds for $\mathcal{T} = \mathcal{T}^\nu_{\tau_0}$ for any full-support distribution ν.

Proof. Follows directly from a fact that $\pi_{\theta^+} = \mathcal{U}^{\nu}_{\tau}(\pi_\theta)$ and Lemma 15. □

Corollary 3 (Uniform Polyak–Łojasiewicz inequality)**.** *Let $\mathcal{T} = \mathcal{T}^{\nu}_{\tau}$ be an improvement operator defined in Lemma 16. For any $\theta \in \Theta_{\mathcal{T}}$ the following inequality holds almost surely*

$$\|\nabla J^{\widetilde{\pi}}(\theta; \pi_\theta)\|_2^2 \geq 2c^{\nu}_{\lambda,\tau} \cdot \mathrm{SubOpt}^{\widetilde{\pi}}_{\lambda}(\pi_\theta)\,,$$

where $c^{\nu}_{\lambda,\tau} = \lambda\tau^2\nu^2_{\min}$. In particular, for a choice $\tau = \tau_0$, we have the following simplified factor

$$m_{\widetilde{\pi},\lambda} = \lambda \mathrm{e}^{-2/\lambda} \cdot c_\nu^2, \qquad c_\nu = \exp(\min\{-2\|\log\widetilde{\pi} - \log\nu\|_{\mathrm{sp}}, \log(\nu_{\min}/(1+\nu_{\min}))\}) \cdot \nu_{\min}\,,$$

so Assumption 4 holds with $m_{\widetilde{\pi},\lambda}$ and $\varepsilon_{\mathrm{PL}} = 0$.

Proof. For any $\theta \in \Theta_{\mathcal{T}}$, Lemma 16 implies

$$\pi_\theta(y) \geq \tau\nu(y), \qquad \min_{y\in\mathcal{Y}} \pi_\theta(y) \geq \tau \cdot \nu_{\min}\,.$$

Thus, by Lemma 14, we have

$$\|\nabla J^{\widetilde{\pi}}(\theta; \pi_\theta)\|_2^2 \geq 2\lambda \min_y \pi^2_\theta(y) \cdot \mathrm{SubOpt}^{\widetilde{\pi}}_{\lambda}(\pi_\theta) \geq 2\lambda\tau^2\nu^2_{\min} \cdot \mathrm{SubOpt}^{\widetilde{\pi}}_{\lambda}(\pi_\theta)\,.$$

After plugging-in the $\tau = \tau_0$ we have

$$c^{\nu}_{\lambda,\tau_0} = \lambda \min\{\exp(-2/\lambda) \cdot \exp(-4\|\log\widetilde{\pi} - \log\nu\|_{\mathrm{sp}}), [\nu_{\min}/(1+\nu_{\min})]^2\}\nu^2_{\min}\,.$$

Since $\exp(-2/\lambda) \leq 1$, we can simplify the expression and achieve the following bound

$$c^{\nu}_{\lambda,\tau_0} \geq m_{\widetilde{\pi},\lambda} = \lambda \mathrm{e}^{-2/\lambda} \cdot \exp(\min\{-4\|\log\widetilde{\pi} - \log\nu\|_{\mathrm{sp}}, 2\log(\nu_{\min}/(1+\nu_{\min}))\}) \cdot \nu^2_{\min}$$

□

C.3.3 Gradient Estimation and Noise Bounds

For softmax parametrization, we use the truncated pairwise policy gradient estimator, discussed in Appendix C.2 and in this section we verify the conditions of Lemma 11 for this parameterization.

Lemma 17. *Let us consider the context-free softmax parametrization of policies π_θ. Then, the mini-batch estimator g_t in (36) with the choice $k = \infty$ satisfies Assumption 6 with bias $\varepsilon_{\mathrm{grad}}$, subgaussian constant $\sigma^2_{\widetilde{\pi},\lambda} = D^2_{\widetilde{\pi},\lambda}(\varepsilon_{\mathrm{grad}})$, and subexponential constant $v^2_{\widetilde{\pi},\lambda} = 6D^2_{\widetilde{\pi},\lambda}(\varepsilon_{\mathrm{grad}})$, where*

$$\begin{aligned} D^2_{\widetilde{\pi},\lambda}(\varepsilon_{\mathrm{grad}}) &\triangleq 2\Big(1 + 4\lambda\log\Big(\frac{2\sqrt{2}\,\lambda(1+\widetilde{\pi}_{\min})}{\widetilde{\pi}_{\min}}\Big) + 4\lambda\log(1/\varepsilon_{\mathrm{grad}})\Big)^2 \\ &= 2\Big(1 + 4\lambda\log\Big(\frac{2\sqrt{2}\,\lambda(1+\widetilde{\pi}_{\min})}{\widetilde{\pi}_{\min}\cdot\varepsilon_{\mathrm{grad}}}\Big)\Big)^2, \end{aligned}$$

and $\widetilde{\pi}_{\min} = \min_{y\in\mathcal{Y}}\widetilde{\pi}(y)$.

Proof. For this lemma, we verify the conditions of Lemma 11. First of all, we notice that $\|\nabla\log\pi_\theta(y|x)\|_2^2 \leq 2$ since for softmax parameterization we have the following identity $[\nabla\log\pi_\theta(y)]_{y'} = \mathbb{1}\{y = y'\} - \pi_\theta(y')$. Next, we see that since $\widetilde{\pi}(y) > 0$ for any $y \in \mathcal{Y}$, then the second part of Lemma 11 holds with $D_{\widetilde{\pi}} = 1/\widetilde{\pi}_{\min}$.

Thus, for any $\varepsilon_{\mathrm{grad}} > 0$, Assumption 6 is satisfied with a bias bound $\varepsilon_{\mathrm{grad}}$, subgaussian constant $\sigma^2_{\widetilde{\pi},\lambda} = D^2_{\widetilde{\pi},\lambda}$ and subexponential constant $v^2_{\widetilde{\pi},\lambda} = 6D^2_{\widetilde{\pi},\lambda}$, where

$$D^2_{\widetilde{\pi},\lambda} \triangleq 2\left(1 + 4\lambda\log\left(\frac{2\sqrt{2}\cdot\lambda(1+\widetilde{\pi}_{\min})}{\widetilde{\pi}_{\min}}\right) + 4\lambda\log(1/\varepsilon_{\mathrm{grad}})\right)^2.$$

□

C.4 Verification under Compatible Fisher Non-Degenerate parameterization

In this section, we state another set of assumptions under which Assumptions 2–6 hold. In the following, we assume $\Theta = \mathbb{R}^d$ for some $d > 0$. We start from the general assumption on the policy $\widetilde{\pi}$ which depends only on this policy and does not depend on the parametrization.

Assumption 7 (Reference-policy regularity)**.** Define the per-context minimum mass of the reference policy

$$\widetilde{\pi}_{\min}(x) \triangleq \min_{y \in \mathcal{Y}} \widetilde{\pi}(y|x).$$

Assume that $\widetilde{\pi}_{\min}(x) > 0$ for ρ-almost every x, and that one of the following moment conditions holds:

(i) Log-moment condition:

$$\mathbb{E}_{x\sim\rho}\left[\log^2\left(\frac{1}{\widetilde{\pi}_{\min}(x)}\right)\right] \le V_{\widetilde{\pi}} \quad \text{for some } V_{\widetilde{\pi}} > 0.$$

(ii) Inverse-mass condition:

$$\mathbb{E}_{x\sim\rho}\left[\frac{1}{\widetilde{\pi}_{\min}(x)}\right] \le D_{\widetilde{\pi}} \quad \text{for some } D_{\widetilde{\pi}} > 0.$$

This assumption controls how often the reference policy assigns extremely small probability to some action. It is used to bound moments (and clipping bias) of log-ratio terms such as $\log(\pi_\theta(y|x)/\widetilde{\pi}(y|x))$ that appear in the KL-regularized policy-gradient estimator.

Condition (ii) is strictly stronger and implies (i). Indeed, for all $u \ge 1$, $(\log u)^2 \le \frac{4}{\mathrm{e}^2}\, u$, so taking $u = 1/\widetilde{\pi}_{\min}(x)$ yields

$$\mathbb{E}_{x\sim\rho}\left[\log^2\left(\frac{1}{\widetilde{\pi}_{\min}(x)}\right)\right] \le \frac{4}{\mathrm{e}^2}\,\mathbb{E}_{x\sim\rho}\left[\frac{1}{\widetilde{\pi}_{\min}(x)}\right] \le \frac{4}{\mathrm{e}^2} D_{\widetilde{\pi}}.$$

We state both variants because (i) is sufficient for some bounds (e.g., smoothness), whereas (ii) can yield slightly cleaner bias/threshold choices for clipped estimators.

Assumption 8 (Bounded score function)**.** For all $\theta \in \Theta$, the score and the log-policy Hessian are bounded:

$$\max_{x\in\mathrm{supp}(\rho), y\in\mathcal{Y}} \|\nabla_\theta \log \pi_\theta(y|x)\|_2^2 \le M_g^2, \qquad \mathbb{E}_{x\sim\rho, y\sim\pi_\theta(\cdot|x)}\left[\|\nabla_\theta^2 \log \pi_\theta(y|x)\|_{\mathrm{op}}^2\right] \le M_h^2.$$

This assumption is standard in literature of convergence of policy gradient methods (see, e.g., Papini et al. (2018); Huang et al. (2020); Ding et al. (2022); Yuan et al. (2022); Fatkhullin et al. (2023)).

Assumption 9 (Fisher Non-Degeneracy)**.** The Fisher information matrix is non-degenerate for every $\theta \in \Theta_{\mathcal{T}}$:

$$F_\rho(\theta) \triangleq \mathbb{E}_{x\sim\rho, y\sim\pi_\theta(\cdot|x)}[\nabla \log \pi_\theta(y|x)[\nabla \log \pi_\theta(y|x)]^\intercal] \succeq m_F I.$$

This assumption is natural for any natural policy gradient analysis, and it holds for certain exponential family parameterized policies (Ding et al., 2022, Section 8) and for certain neural policy families. We refer to a discussion in Liu et al. (2020, Section B.2).

Assumption 10 (Compatible parametrization)**.** There exists $\varepsilon_{\mathrm{bias}} \ge 0$ such that for $u_\star(\theta) = F_\rho(\theta)^{-1}\nabla J^{\widetilde{\pi}}(\theta; \pi_\theta)$ it holds for any $\theta \in \Theta_{\mathcal{T}}$

$$\mathbb{E}_{x\sim\rho}\left[\left\|\mathcal{P}(\pi_\theta \succ \cdot) + \lambda \log\frac{\pi_\theta(\cdot|x)}{\widetilde{\pi}(\cdot|x)} - u_\star(\theta)^\intercal \nabla \log \pi_\theta(\cdot|x)\right\|_{\mathrm{sp}}^2\right] \le \varepsilon_{\mathrm{bias}}\,.$$

We note that this assumption corresponds to an Assumption 4.6 of Ding et al. (2022) for $\varepsilon_{\mathrm{bias}} = 0$ and corresponds to the classical assumption of Sutton et al. (1999) as well as actively used for analysis of natural policy gradient methods Kakade (2001); Agarwal et al. (2021). We note that in our case, the norm is more restrictive than $\pi_\theta(\cdot|x)$-weighted ℓ_2 norm, however we argue that even a stronger assumption is possible to satisfy in practice by using an expressive enough parametrization.

Under these assumptions, we can prove the following result.

Proposition 6 (Compatible Fisher non-degenerate parameterization satisfies Assumptions 2–6)**.** Consider a general parametrization $\theta \mapsto \pi_\theta$ and an improvement operator $\mathcal{T}$ which satisfy Assumptions 8, Assumption 9, and Assumption 10, and assume that a reference policy $\widetilde{\pi}$ satisfies Assumption 7-(i). Then Assumption 2–5 hold with

$$G = M_g, \quad L_{\widetilde{\pi},\lambda} = \left(\tfrac{1}{2} + \lambda\sqrt{2+V_{\widetilde{\pi}}}\right)(M_g^2+M_h)+\lambda M_g^2, \quad m_{\widetilde{\pi},\lambda} = \frac{\lambda m_F^2}{2M_g^2}, \quad \varepsilon_{\mathrm{PL}} = \frac{\varepsilon_{\mathrm{bias}} \cdot m_F^2}{M_g^2}.$$

Additionally, for any $\varepsilon_{\mathrm{grad}} > 0$, a truncated pairwise estimator (36) with $k = 2$ satisfies Assumption 6 with a bias of level $\varepsilon_{\mathrm{grad}}$ and a subgaussian and subexponential parameters

$$\sigma^2_{\widetilde{\pi},\lambda} = M_g^2\left(1 + \frac{8\lambda^2 M_g(V_{\widetilde{\pi}}+1)}{\varepsilon_{\mathrm{grad}}}\right)^2, \qquad v^2_{\widetilde{\pi},\lambda} = 6M_g^2\left(1 + \frac{8\lambda^2 M_g(V_{\widetilde{\pi}}+1)}{\varepsilon_{\mathrm{grad}}}\right)^2.$$

If additionally a reference policy $\widetilde{\pi}$ satisfies Assumption 7-(ii), then a truncated pairwise estimator (36) with $k = +\infty$ satisfies Assumption 6 with a bias level $\varepsilon_{\mathrm{grad}} > 0$ and other factors

$$\sigma^2_{\widetilde{\pi},\lambda} = M_g^2\left(1 + 4\lambda \log\left(\frac{2\lambda \cdot M_g(D_{\widetilde{\pi}}+1)}{\varepsilon_{\mathrm{grad}}}\right)\right)^2, \qquad v^2_{\widetilde{\pi},\lambda} = 6\sigma^2_{\widetilde{\pi},\lambda}.$$

Proof. Follows from Lemma 18, Lemma 19, Lemma 21, and Lemma 22. □

Corollary 4 (SPG iteration and sample complexity: compatible Fisher parameterization)**.** *Fix $\delta \in (0,1)$ and $\varepsilon \in (0,1)$. Assume the log-moment condition of Assumption 7(i) with constant $V_{\widetilde{\pi}}$, and assume Assumptions 8–10 hold with constants M_g, M_h, m_F and $\varepsilon_{\mathrm{bias}}$. Let $m_{\widetilde{\pi},\lambda} = \lambda m_F^2/(2M_g^2)$ and $\varepsilon_{\mathrm{PL}} = \varepsilon_{\mathrm{bias}} m_F^2/M_g^2$ as in Proposition 6. Assume $\lambda m_{\widetilde{\pi},\lambda} \geq G^2$ (here $G = M_g$), i.e.,*

$$\lambda \geq \lambda^{\mathrm{F}}_{\min} \triangleq \sqrt{2}\,\frac{M_g^2}{m_F}.$$

Run SPG (26) with (γ_t, B_t) as in Proposition 4, and use the clipped estimator from Proposition 6 with $k = 2$ at bias level

$$\varepsilon^2_{\mathrm{grad}} = \frac{7\, m_{\widetilde{\pi},\lambda} \cdot \varepsilon}{180 \log(\mathrm{e}/\delta)} \qquad (\text{so } M = \tfrac{1}{2} + \tfrac{4\lambda^2 M_g(V_{\widetilde{\pi}}+1)}{\varepsilon_{\mathrm{grad}}} = \Theta(1/\sqrt{\varepsilon})).$$

Then with probability at least $1-\delta$, once

$$N_{\mathrm{iter}}(\varepsilon,\delta) = T = \widetilde{\mathcal{O}}\left(\frac{G^2 v^2_{\widetilde{\pi},\lambda}}{\lambda m_{\widetilde{\pi},\lambda}} \cdot \frac{\log(\mathrm{e}/\delta)}{\varepsilon}\right) = \widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon^2}\right),$$
$$N_{\mathrm{sample}}(\varepsilon,\delta) = \sum_{t=0}^{T-1} B_t = \widetilde{\mathcal{O}}(T^2/m_{\widetilde{\pi},\lambda}) = \widetilde{\mathcal{O}}(\varepsilon^{-4}),$$

the final iterate π_T satisfies the approximate VNW guarantee

$$\mathrm{SubOpt}^{\widetilde{\pi}}_{\lambda}(\pi_T) \leq \varepsilon + \frac{6\log(\mathrm{e}/\delta)}{m_{\widetilde{\pi},\lambda}}\varepsilon_{\mathrm{PL}} = \varepsilon + \frac{12\log(\mathrm{e}/\delta)}{\lambda}\varepsilon_{\mathrm{bias}}.$$

Proof. Proposition 6 provides $G = M_g$, $m_{\widetilde{\pi},\lambda}$, $\varepsilon_{\mathrm{PL}}$, and shows Assumption 6 holds with $v^2_{\widetilde{\pi},\lambda} = 6M_g^2\big(1+8\lambda^2 M_g(V_{\widetilde{\pi}}+1)/\varepsilon_{\mathrm{grad}}\big)^2$. The condition $\lambda m_{\widetilde{\pi},\lambda} \geq G^2$ allows applying Proposition 4. With the chosen $\varepsilon_{\mathrm{grad}}$, $\frac{6\log(\mathrm{e}/\delta)}{m_{\widetilde{\pi},\lambda}} \cdot \frac{15}{7}\varepsilon^2_{\mathrm{grad}} \leq \varepsilon/2$. Also, since $\varepsilon^2_{\mathrm{grad}} = \Theta(\varepsilon)$, we have $v^2_{\widetilde{\pi},\lambda} = \Theta(1/\varepsilon)$, so taking $T = \widetilde{\mathcal{O}}\big(\frac{G^2 v^2_{\widetilde{\pi},\lambda}}{\lambda m_{\widetilde{\pi},\lambda}} \cdot \frac{\log(\mathrm{e}/\delta)}{\varepsilon}\big) = \widetilde{\mathcal{O}}(\log(\mathrm{e}/\delta)/\varepsilon^2)$ makes the remaining terms in Proposition 4 at most $\varepsilon/2$, yielding $\mathrm{SubOpt}^{\widetilde{\pi}}_{\lambda}(\pi_T) \leq \varepsilon + \frac{6\log(\mathrm{e}/\delta)}{m_{\widetilde{\pi},\lambda}}\varepsilon_{\mathrm{PL}}$. Finally, summing $B_t = \lceil (t+8\kappa_{\widetilde{\pi},\lambda})/m_{\widetilde{\pi},\lambda} \rceil$ gives $\sum_{t<T} B_t = \widetilde{\mathcal{O}}(T^2/m_{\widetilde{\pi},\lambda})$. □

C.4.1 Lipschitzness and smoothness

We note that this type of results for general policy parameterizations are typically derived for unregularized reinforcement learning problems (Papini et al., 2018; Yuan et al., 2022). As a result, it is necessary to prove modified versions.

Lemma 18 (Lipschitzness). *Assume Assumption 8. For any $\theta, \theta' \in \Theta$ it holds*

$$\|\pi_\theta - \pi_{\theta'}\|_{1,\rho} \le M_g \|\theta - \theta'\|_2.$$

In particular, Assumption 2 holds with a constant $G = M_g$.

Proof. From Lemma 24 of Mei et al. (2020) we have for any $x \in \mathcal{X}$

$$\|\pi_\theta(\cdot|x) - \pi_{\theta'}(\cdot|x)\|_1 \le \|\log \pi_\theta(\cdot|x) - \log \pi_{\theta'}(\cdot|x)\|_{\mathrm{sp}}\,.$$

Next, by Taylor expansion we have

$$\log \pi_\theta(\cdot|x) - \log \pi_{\theta'}(\cdot|x) = \nabla \log \pi_{\tilde\theta}(\cdot|x)^\intercal(\theta - \theta')\,,$$

where $\tilde\theta$ is some point on the line segment between θ and θ'. By our assumption on the parameter set Θ we have $\tilde\theta \in \Theta$, thus, applying Assumption 8,

$$\begin{aligned}\|\pi_\theta - \pi_{\theta'}\|_{1,\rho}^2 &= \mathbb{E}_{x\sim\rho}[\|\pi_\theta(\cdot|x) - \pi_{\theta'}(\cdot|x)\|_1^2] \le \mathbb{E}_{x\sim\rho}[\|\nabla \log \pi_{\tilde\theta}(\cdot|x)^\intercal(\theta - \theta')\|_\infty^2] \\ &\le \mathbb{E}_{x\sim\rho}\left[\max_{y\in\mathcal{Y}} \|\nabla \log \pi_{\tilde\theta}(y|x)\|_2^2\right] \cdot \|\theta - \theta'\|_2^2 \le M_g^2 \|\theta - \theta'\|_2^2\,.\end{aligned}$$

□

Lemma 19 (Smoothness). *Assume Assumption 8 and Assumption 7-(i). Then $\theta \mapsto J^{\widetilde\pi}(\theta;\pi)$ is $L_{\widetilde\pi,\lambda}$-smooth on Θ with*

$$L_{\widetilde\pi,\lambda} = \left(\tfrac{1}{2} + \lambda\sqrt{2 + V_{\widetilde\pi}}\right)(M_g^2 + M_h) + \lambda M_g^2\,,$$

i.e., satisfies Assumption 3 with a constant $L_{\widetilde\pi,\lambda}$.

Proof. To simplify the notation, define

$$s_\theta(x,y) \triangleq \nabla_\theta \log \pi_\theta(y|x), \qquad H_\theta(x,y) \triangleq \nabla_\theta^2 \log \pi_\theta(y|x).$$

Then the standard identities give

$$\nabla_\theta \pi_\theta(y|x) = \pi_\theta(y|x)\, s_\theta(x,y), \qquad \nabla_\theta^2 \pi_\theta(y|x) = \pi_\theta(y|x)\big(s_\theta(x,y)s_\theta(x,y)^\intercal + H_\theta(x,y)\big).$$

Moreover, since $\sum_y \pi_\theta(y|x) = 1$, differentiating twice yields the "Bartlett identity"

$$\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\big[s_\theta(x,y)s_\theta(x,y)^\intercal + H_\theta(x,y)\big] = 0, \qquad \forall x \in \mathcal{X}, \theta \in \Theta. \tag{41}$$

Next, define the per-sample regularized reward

$$r_\theta(x,y) \triangleq \mathcal{P}(\pi \succ y \mid x) + \lambda \log \frac{\pi_\theta(y|x)}{\widetilde\pi(y|x)}\,.$$

Using $\nabla_\theta \ell_\theta(x,y) = s_\theta(x,y)$ and $\mathbb{E}_{y\sim\pi_\theta}[s_\theta] = 0$, one obtains the usual policy-gradient form

$$\nabla_\theta J^{\widetilde\pi}(\theta;\pi) = \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\Big[s_\theta(x,y)\, r_\theta(x,y)\Big]\,.$$

Differentiating once more gives

$$\nabla_\theta^2 J^{\widetilde\pi}(\theta;\pi) = \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\Big[r_\theta(x,y)\big(s_\theta s_\theta^\intercal + H_\theta\big)\Big] \;+\; \lambda\, \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\Big[s_\theta s_\theta^\intercal\Big],$$

where we used $\nabla_\theta r_\theta(x,y) = \lambda s_\theta(x,y)$.

Fix $x \in \mathcal{X}$. Because of (41), for any scalar $c(x)$ we have

$$\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\Big[c(x)\,(s_\theta s_\theta^\intercal + H_\theta)\Big] = 0.$$

Therefore we may replace $r_\theta(x,y)$ by a centered version without changing the expectation: pick any $a(x), b_\theta(x) \in \mathbb{R}$ and write

$$r_\theta(x,y) = \big(u_\pi(x,y) - a(x)\big) + \lambda\big(\ell_\theta(x,y) - b_\theta(x)\big) \;+\; \big(a(x) + \lambda b_\theta(x)\big),$$

and the last constant term vanishes when multiplied by $(s_\theta s_\theta^\intercal + H_\theta)$ and averaged over y. Let $S_\theta(x,y) \triangleq s_\theta s_\theta^\intercal + H_\theta$. Now use Cauchy–Schwarz to bound

$$\begin{aligned}\Big\|\mathbb{E}_{x,y}\big[r_\theta(x,y)S_\theta(x,y)\big]\Big\|_{\mathrm{op}} &= \Big\|\mathbb{E}_{x,y}\big[\big(r_\theta(x,y) - a(x) - \lambda b_\theta(x)\big)\, S_\theta(x,y)\big]\Big\|_{\mathrm{op}} \\ &\le \mathbb{E}_{x,y}\big[|r_\theta(x,y) - a(x) - \lambda b_\theta(x)|\, \|S_\theta(x,y)\|_{\mathrm{op}}\big] \\ &\le \sqrt{\mathbb{E}_{x,y}[(r_\theta(x,y) - a(x) - \lambda b_\theta(x))^2] \cdot \mathbb{E}_{x,y}[\|S_\theta(x,y)\|_{\mathrm{op}}^2]}\,.\end{aligned}$$

For the Hessian term, we have by Minkowski inequality since $\|S_\theta\|_{\mathrm{op}} \le \|s_\theta\|_2^2 + \|H_\theta\|_{\mathrm{op}}$

$$\sqrt{\mathbb{E}_{x,y}[\|S_\theta(x,y)\|_{\mathrm{op}}^2]} \le \sqrt{\mathbb{E}_{x,y}[\|s_\theta(x,y)\|_2^4]} + \sqrt{\mathbb{E}_{x,y}[\|H_\theta(x,y)\|_{\mathrm{op}}^2]} \le M_g^2 + M_h\,.$$

For the reward variance term, we first apply Minkowski inequality again:

$$\begin{aligned}\sqrt{\mathbb{E}_{x,y}[(r_\theta(x,y) - a(x) - \lambda b_\theta(x))^2]} &\le \sqrt{\mathbb{E}_{x,y}[(u_\pi(x,y) - a(x))^2]} \\ &\quad + \lambda\sqrt{\mathbb{E}_{x,y}[(\ell_\theta(x,y) - b_\theta(x))^2]}\,.\end{aligned}$$

For the first term, since $u_\pi(x,y) \in [0,1]$, we can pick $a(x) = 1/2$ to get

$$\sqrt{\mathbb{E}_{x,y}[(u_\pi(x,y) - a(x))^2]} \le \frac{1}{2}.$$

For the second term, we take $b_\theta(x) = 0$ and decompose as follows:

$$\begin{aligned}\mathbb{E}_{x,y}\big[(\ell_\theta(x,y) - b_\theta(x))^2\big] &= \mathbb{E}_x\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[\left(\log\frac{\pi_\theta(y|x)}{\widetilde{\pi}(y|x)}\right)_+^2\right] \\ &\quad + \mathbb{E}_x\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[\left(-\log\frac{\pi_\theta(y|x)}{\widetilde{\pi}(y|x)}\right)_+^2\right].\end{aligned}$$

For the positive part, we use a bound $\pi_\theta(y|x) \le 1$ to get

$$\mathbb{E}_x\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[\left(\log\frac{\pi_\theta(y|x)}{\widetilde{\pi}(y|x)}\right)_+^2\right] \le \mathbb{E}_x\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[\log^2\frac{1}{\widetilde{\pi}(y|x)}\right] \le V_{\widetilde{\pi}},$$

where the last inequality follows from Assumption 7-(i). For the negative part, we use a bound $\log(x) \le x - 1$ for any $x > 0$ and get the following

$$\begin{aligned}\mathbb{E}_x\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[\left(-\log\frac{\pi_\theta(y|x)}{\widetilde{\pi}(y|x)}\right)_+^2\right] &= \mathbb{E}_x\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[\left(\log\frac{\widetilde{\pi}(y|x)}{\pi_\theta(y|x)}\right)_+^2\right] \\ &= 2\int_0^\infty u\mathbb{P}_{x\sim\rho,y\sim\pi_\theta(\cdot|x)}\left[\log\frac{\widetilde{\pi}(y|x)}{\pi_\theta(y|x)} \ge u\right]\mathrm{d}u\,.\end{aligned}$$

To bound the last probability, we notice that for $Z = \log\frac{\widetilde{\pi}(y|x)}{\pi_\theta(y|x)}$ we have $\mathbb{E}[\mathrm{e}^Z] = 1$, thus, by Markov's inequality

$$\mathbb{P}_{x\sim\rho,y\sim\pi_\theta(\cdot|x)}\left[\log\frac{\widetilde{\pi}(y|x)}{\pi_\theta(y|x)} \ge u\right] \le \mathrm{e}^{-u}\,,$$

thus

$$\mathbb{E}_x\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[\left(-\log\frac{\pi_\theta(y|x)}{\widetilde{\pi}(y|x)}\right)_+^2\right] \le 2\int_0^\infty u\mathrm{e}^{-u}\mathrm{d}u = 2\,.$$

For the remaining term,

$$\Big\|\lambda\,\mathbb{E}_{x,y}[s_\theta s_\theta^\intercal]\Big\|_{\mathrm{op}} \le \lambda\,\mathbb{E}_{x,y}\|s_\theta s_\theta^\intercal\|_{\mathrm{op}} = \lambda\,\mathbb{E}_{x,y}\|s_\theta\|_2^2 \le \lambda M_g^2.$$

Combining all the bounds yields the uniform Hessian bound

$$\|\nabla_\theta^2 J^{\widetilde{\pi}}(\theta;\pi)\|_{\mathrm{op}} \le \Big(\tfrac{1}{2} + \lambda\sqrt{2 + V_{\widetilde{\pi}}}\Big)(M_g^2 + M_h) + \lambda M_g^2\,.$$

A uniform bound on the Hessian implies $L_{\widetilde{\pi},\lambda}$-smoothness (e.g. by the standard second-order Taylor remainder bound), proving the lemma. □

C.4.2 Polyak–Łojasiewicz inequality

Lemma 20 (Gradient comparison)**.** *Under Assumptions 8-9-10, for any $\theta \in \Theta$ the following inequality holds*

$$\|\nabla V_\lambda^{\widetilde{\pi}}(\pi_\theta, \pi_\theta)\|_{\mathrm{sp},\rho}^2 \le 2\varepsilon_{\mathrm{bias}} + \frac{2M_g^2}{m_F^2}\|\nabla J^{\widetilde{\pi}}(\theta; \pi_\theta)\|_2^2 .$$

Proof. First, we notice that $[\nabla V_\lambda^{\widetilde{\pi}}(\pi_\theta, \pi_\theta)](x, y) = \mathcal{P}(\pi_\theta \succ y \mid x) + \lambda(1 + \log \frac{\pi_\theta(y|x)}{\widetilde{\pi}(y|x)})$, thus, Assumption 10 implies $\|\nabla V_\lambda^{\widetilde{\pi}}(\pi_\theta, \pi_\theta) - (u_\star)^\intercal \nabla \log \pi_\theta(y|x)\|_{\mathrm{sp},\rho}^2 \le \varepsilon_{\mathrm{bias}}$, thus, applying triangle inequality and an inequality $(a+b)^2 \le 2a^2 + 2b^2$,

$$\begin{aligned}
\|\nabla V_\lambda^{\widetilde{\pi}}(\pi_\theta, \pi_\theta)\|_{\mathrm{sp},\rho}^2 &\le 2\varepsilon_{\mathrm{bias}} + 2\|(u_\star)^\intercal \nabla \log \pi_\theta(y|x)\|_{\mathrm{sp},\rho}^2 \\
&\le 2\varepsilon_{\mathrm{bias}} + 2\|(u_\star)^\intercal \nabla \log \pi_\theta(y|x)\|_{\infty,\rho}^2 \\
&\le 2\varepsilon_{\mathrm{bias}} + 2\max_{y \in \mathcal{Y}} \|F_\rho(\theta)^{-1}\nabla J^{\widetilde{\pi}}(\theta; \pi_\theta)\|_2^2 \cdot \|\nabla \log \pi_\theta(y|x)\|_{2,\rho}^2 \\
&\le 2\varepsilon_{\mathrm{bias}} + \frac{2M_g^2}{m_F^2}\|\nabla J^{\widetilde{\pi}}(\theta; \pi_\theta)\|_2^2 .
\end{aligned}$$

□

Lemma 21 (PL inequality)**.** *Assume Assumptions 8-9-10. Then for any $\theta \in \Theta_{\mathcal{T}}$ the following holds*

$$\lambda m_F^2 / M_g^2 \cdot \mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_\theta) \le \|\nabla J^{\widetilde{\pi}}(\theta; \pi_\theta)\|_2^2 + \frac{\varepsilon_{\mathrm{bias}} \cdot m_F^2}{M_g^2} .$$

In particular, Assumption 4 is satisfied with $m_{\widetilde{\pi},\lambda} = 0.5 \cdot \lambda (m_F/M_g)^2$ and $\varepsilon_{\mathrm{PL}} = \varepsilon_{\mathrm{bias}} \cdot (m_F/M_g)^2$.

Proof. First, we follow a standard proof for PL-inequality in the strongly convex case, initially context-wise.

Let us fix a context $x \in \mathrm{supp}(\rho)$ and a competitor policy $\mu \in \Pi$, and then we can define a context-wise value as $V_\lambda^{\widetilde{\pi}}(\pi, \mu|x) = \mathcal{P}(\mu \succ \pi \mid x) + \lambda \mathrm{KL}(\pi(x)\|\widetilde{\pi}(x))$, such that $V_\lambda^{\widetilde{\pi}}(\pi, \mu) = \mathbb{E}_{x\sim\rho}\big[V_\lambda^{\widetilde{\pi}}(\pi, \mu|x)\big]$.

Let us denote $\pi_\mu^\star(\cdot|x) = \arg\min_\pi V_\lambda^{\widetilde{\pi}}(\pi, \mu|x)$ for all $x \in \mathrm{supp}(\rho)$ simultaneously. We notice that this policy is the same as $\pi_\mu^\star = \arg\min_\pi V_\lambda^{\widetilde{\pi}}(\pi, \mu)$ due to the additive structure of subproblems with different contexts x. Since $V_\lambda^{\widetilde{\pi}}(\pi, \mu|x)$ is λ-strongly convex in $\pi(x)$ with respect to ℓ_1-norm, thus, for any π

$$V_\lambda^{\widetilde{\pi},\star}(\mu|x) \ge V_\lambda^{\widetilde{\pi}}(\pi, \mu|x) + \langle [\nabla_\pi V_\lambda^{\widetilde{\pi}}(\pi, \mu)](x), \pi_\mu^\star(x) - \pi(x)\rangle + (\lambda/2)\|\pi(x) - \pi_\mu^\star(x)\|_1^2 .$$

After rearranging the terms, noticing that $\langle \mathbf{1}, \pi_\mu^\star(x) - \pi(x)\rangle = 0$ and using a bound $ab \le a^2/(2\lambda) + \lambda b^2/2$

$$\begin{aligned}
V_\lambda^{\widetilde{\pi}}(\pi, \mu|x) - V_\lambda^{\widetilde{\pi},\star}(\mu|x) &\le \big\langle [\nabla_\pi V_\lambda^{\widetilde{\pi}}(\pi, \mu)](x) + c(x)\mathbf{1}, \pi_\mu^\star(x) - \pi(x)\big\rangle - (\lambda/2)\|\pi(x) - \pi_\mu^\star(x)\|_1^2 \\
&\le \frac{1}{2\lambda}\|[\nabla_\pi V_\lambda^{\widetilde{\pi}}(\pi, \mu)](x) + c(x)\mathbf{1}\|_\infty^2 ,
\end{aligned}$$

where $c\colon \mathcal{X} \to \mathbb{R}$ is an arbitrary baseline function. Minimizing the right-hand side over $c(x)$ and taking the expectation over $x \sim \rho$ we have

$$V_\lambda^{\widetilde{\pi}}(\pi, \mu) - V_\lambda^{\widetilde{\pi},\star}(\mu) \le \frac{1}{2\lambda}\|\nabla_\pi V_\lambda^{\widetilde{\pi}}(\pi, \mu)\|_{\mathrm{sp},\rho}^2 ,$$

and, taking $\pi = \mu = \pi_\theta$, we get

$$\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_\theta) \le \frac{1}{2\lambda}\|\nabla V_\lambda^{\widetilde{\pi}}(\pi_\theta, \pi_\theta)\|_{\mathrm{sp},\rho}^2 .$$

Finally, applying Lemma 20 to derive

$$\mathrm{SubOpt}_\lambda^{\widetilde{\pi}}(\pi_\theta) \le \frac{M_g^2}{\lambda m_F^2}\|\nabla J^{\widetilde{\pi}}(\theta; \pi_\theta)\|_2^2 + \frac{\varepsilon_{\mathrm{bias}}}{\lambda} .$$

By rearranging the terms, we conclude the proof. □

C.4.3 Gradient Estimation and Noise Bounds

For general parametrization, we also use the truncated pairwise policy gradient estimator, discussed in Appendix C.2. Next, we verify the conditions of Lemma 11 for this parameterization.

Lemma 22. *Assume Assumptions 8 for parameterization and Assumption 7-(i) for a reference policy. Then, for any $\varepsilon_{\mathrm{grad}} > 0$, a mini-batch gradient estimator g_t defined in (36) with $k = 2$ satisfies Assumption 6 with a bias level $\varepsilon_{\mathrm{grad}}$, subgaussian constant $\sigma^2_{\widetilde{\pi},\lambda} = D_2^2(\varepsilon_{\mathrm{grad}})$, $v^2_{\widetilde{\pi},\lambda} = 6D_2^2(\varepsilon_{\mathrm{grad}})$, where*

$$D_2^2(\varepsilon_{\mathrm{grad}}) = M_g^2\left(1 + \frac{8\lambda^2 M_g(V_{\widetilde{\pi}} + 1)}{\varepsilon_{\mathrm{grad}}}\right)^2.$$

If, additionally, Assumption 7-(ii) holds, then a mini-batch gradient estimator g_t defined in (36) with $k = +\infty$ also satisfies Assumption 6 with a bias level $\varepsilon_{\mathrm{grad}}$, subgaussian constant $\sigma^2_{\widetilde{\pi},\lambda} = D_\infty^2(\varepsilon_{\mathrm{grad}})$, $v^2_{\widetilde{\pi},\lambda} = 6D_\infty^2(\varepsilon_{\mathrm{grad}})$, where

$$D_\infty^2(\varepsilon_{\mathrm{grad}}) = M_g^2\left(1 + 4\lambda \log\left(\frac{2\lambda \cdot M_g(D_{\widetilde{\pi}} + 1)}{\varepsilon_{\mathrm{grad}}}\right)\right)^2.$$

Proof. Directly follows from Lemma 11. □

D Proximal Point Method with Self-Play Policy Gradients

We now instantiate the proximal point (PP) method from Section B using the self-play policy gradient (SPG) procedure from Section C as an inexact inner solver.

D.1 Algorithm Description

Setup. Let $\pi^{\mathrm{ref}} \in \Pi$ be a reference policy and let $\theta_0 \in \mathbb{R}^d$ be initial parameters such that $\pi_0 = \pi_{\theta_0} = \pi^{\mathrm{ref}}$. We fix: (i) the regularization strength $\beta > 0$ of the original game, (ii) the proximal point step size $\eta > 0$, and (iii) a sequence of self-play learning rates $(\gamma_{k,t})_{k\geq 0,\, t\geq 0}$ and batch sizes $(B_{k,t})_{k\geq 0,\, t\geq 0}$. For convenience, we also denote $\beta_{\mathrm{target}} \triangleq \beta/\eta$.

The PP update at outer iteration k aims to compute

$$\pi_{k+1} \approx \arg\max_{\pi\in\Pi} \min_{\pi'\in\Pi} \Big\{ \mathcal{P}(\pi \succ \pi') - \beta\, \mathrm{KL}_\rho(\pi\|\pi^{\mathrm{ref}}) + \beta\, \mathrm{KL}_\rho(\pi'\|\pi^{\mathrm{ref}}) \\ - \beta_{\mathrm{target}}\, \mathrm{KL}_\rho(\pi\|\pi_k) + \beta_{\mathrm{target}}\, \mathrm{KL}_\rho(\pi'\|\pi_k) \Big\}.$$

We realize this update approximately by running self-play policy gradients (SPG) on the inner two-player game above, using a finite number T_k of self-play steps.

Inner objective. At outer iteration k, we keep π_k fixed and define for any $\theta \in \Theta$ and competitor policy $\pi \in \Pi$ the local objective

$$J_k(\theta;\pi) \triangleq \mathcal{P}(\pi \succ \pi_\theta) + \beta\, \mathrm{KL}_\rho(\pi_\theta\|\pi^{\mathrm{ref}}) + \beta_{\mathrm{target}}\, \mathrm{KL}_\rho(\pi_\theta\|\pi_k),$$

where π_θ is a parametrized policy satisfying Assumption 2-5. For a given π, minimizing $J_k(\theta;\pi)$ in θ corresponds to computing a regularized best response against π with an additional proximal term toward π_k.

Stochastic gradients. At inner iterate (k,t) we maintain parameters $\theta_{k,t}$ and the corresponding policy $\pi_{k,t} = \pi_{\theta_{k,t}}$. We assume access to a stochastic gradient estimator $g_{k,t}$ which satisfies Assumption 6.

Example: pairwise REINFORCE estimator. As a particular instance, we can use a truncated pairwise REINFORCE estimator, analogous to one defined in Appendix C.2. At inner iterate (k,t), given a batch size $B_{k,t}$, we sample contexts $x_{k,t,j} \sim \rho$, independent pairs $(y_{k,t,j}, y'_{k,t,j}) \sim \pi_{k,t}(\cdot \mid$

Algorithm 1 Proximal Point Method with Self-Play Policy Gradients (PP–SPG)

Require: Reference policy π^{ref}; regularization parameter $\beta > 0$; proximal step size $\eta > 0$; number of outer iterations K; inner iteration lengths $(T_k)_{k=0}^{K-1}$; learning rates $(\gamma_{k,t})_{k,t}$; batch sizes $(B_{k,t})_{k,t}$.
Ensure: Approximate von Neumann winner policy π_K.

Initialize parameters θ_0 such that $\pi_0 = \pi_{\theta_0} = \pi^{\mathrm{ref}}$.
$\beta_{\mathrm{target}} \leftarrow \beta/\eta$.
for $k = 0$ **to** $K-1$ **do**
  # Outer (proximal point) loop
  $\theta_{k,0} \leftarrow \theta_k$; $\pi_{k,0} \leftarrow \pi_k$.
  for $t = 0$ **to** $T_k - 1$ **do**
    # Inner self-play loop
    for $j = 1$ **to** $B_{k,t}$ **do**
      Sample context $x_{k,t,j} \sim \rho$.
      Sample $y_{k,t,j} \sim \pi_{k,t}$ and $y'_{k,t,j} \sim \pi_{k,t}$ independently.
      Obtain an unbiased estimate $p_{k,t,j}$ of $\mathcal{P}\big(y_{k,t,j} \succ y'_{k,t,j} \mid x_{k,t,j}\big)$.
      Set $G_k^{(j)} \leftarrow G_k\big(\theta_{k,t} \mid x_{k,t,j}, y_{k,t,j}, y'_{k,t,j}, p_{k,t,j}\big)$ using Eq. (42).
    end for
    $g_{k,t} \leftarrow \frac{1}{B_{k,t}} \sum_{j=1}^{B_{k,t}} G_k^{(j)}$.
    $\theta_{k,t+1} \leftarrow \mathcal{T}_k\big(\theta_{k,t} - \gamma_{k,t}\, g_{k,t}\big)$.
    $\pi_{k,t+1} \leftarrow \pi_{\theta_{k,t+1}}$.
  end for
  $\theta_{k+1} \leftarrow \theta_{k,T_k}$; $\pi_{k+1} \leftarrow \pi_{k,T_k}$.
end for
return π_K.

$x_{k,t,j}) \otimes \pi_{k,t}(\cdot \mid x_{k,t,j})$ and obtain unbiased estimates $p_{k,t,j}$ of $\mathcal{P}(y_{k,t,j} \succ y'_{k,t,j} \mid x_{k,t,j})$. Define the single-sample estimator

$$G_k(\theta \mid x, y, y', p) = \frac{1}{2}\big(\nabla_\theta \log \pi_\theta(y|x) - \nabla_\theta \log \pi_\theta(y'|x)\big) \times \tag{42}$$
$$\times\, \mathrm{clip}_{[-M_k, M_k]}\Big[\tfrac{1}{2} - p + \beta\big(\log \tfrac{\pi_\theta(y|x)}{\pi^{\mathrm{ref}}(y|x)} - \log \tfrac{\pi_\theta(y'|x)}{\pi^{\mathrm{ref}}(y'|x)}\big) + \beta_{\mathrm{target}}\big(\log \tfrac{\pi_\theta(y|x)}{\pi_k(y|x)} - \log \tfrac{\pi_\theta(y'|x)}{\pi_k(y'|x)}\big)\Big],$$

where $\beta_{\mathrm{target}} = \beta/\eta$ and M_k is a clipping threshold. The mini-batch estimator at step (k,t) is

$$g_{k,t} \triangleq \frac{1}{B_{k,t}} \sum_{j=1}^{B_{k,t}} G_k\big(\theta_{k,t} \mid x_{k,t,j}, y_{k,t,j}, y'_{k,t,j}, p_{k,t,j}\big),$$

which is a (biased) estimate of $\nabla J_k(\theta_{k,t}; \pi_{k,t})$. As shown in Appendix C.2, this choice satisfies Assumption 6 under softmax and compatible Fisher non-degenerate parameterization.

Self-play update and improvement. To ensure uniform exploration and obtain a uniform PL constant, we apply the improvement operator $\mathcal{T}_k$ for the outer step k satisfying Assumption 5 (e.g., $\mathcal{T}_k \equiv \mathcal{T}_\tau^\nu$ from Section C.3.2 in the softmax case). Given $g_{k,t}$, the inner update is

$$\theta_{k,t+1} \;=\; \mathcal{T}_k\big(\theta_{k,t} - \gamma_{k,t}\, g_{k,t}\big), \qquad \pi_{k,t+1} = \pi_{\theta_{k,t+1}}. \tag{43}$$

Outer update. At the end of the inner loop, after T_k self-play steps, we set

$$\theta_{k+1} \triangleq \theta_{k,T_k}, \qquad \pi_{k+1} \triangleq \pi_{k,T_k},$$

and proceed to the next PP iteration. After K outer iterations, the policy π_K is our approximation of the Nash equilibrium of the original β-regularized preference game. We summarize the procedure in Algorithm 1.

D.2 Analysis of PP–SPG

In this section, we combine the approximate proximal point analysis from Section B with the self-play policy-gradient results of Section C to obtain convergence guarantees for Algorithm 1.

D.2.1 Inner PP game as a regularized preference game

Recall that the outer PP iteration (indexed by k) aims to compute

$$\pi_{k+1} \approx \arg\max_{\pi\in\Pi} \min_{\pi'\in\Pi} \Big\{ \mathcal{P}(\pi \succ \pi') - \beta\,\mathrm{KL}_\rho(\pi\|\pi^{\mathrm{ref}}) + \beta\,\mathrm{KL}_\rho(\pi'\|\pi^{\mathrm{ref}}) \\ - \beta_{\mathrm{target}}\,\mathrm{KL}_\rho(\pi\|\pi_k) + \beta_{\mathrm{target}}\,\mathrm{KL}_\rho(\pi'\|\pi_k) \Big\},$$

where $\beta_{\mathrm{target}} = \beta/\eta$ and $\eta > 0$ is the PP step size. As in Section B, we focus on the value function of the min-player (the regularized best-response objective)

$$V_k(\pi,\mu) \triangleq \mathcal{P}(\mu \succ \pi) + \beta\,\mathrm{KL}_\rho(\pi\|\pi^{\mathrm{ref}}) + \beta_{\mathrm{target}}\,\mathrm{KL}_\rho(\pi\|\pi_k)\,.$$

The self-play inner loop at outer iterate k is run on the objective

$$J_k(\theta;\pi) \triangleq \mathcal{P}(\pi \succ \pi_\theta) + \beta\,\mathrm{KL}_\rho(\pi_\theta\|\pi^{\mathrm{ref}}) + \beta_{\mathrm{target}}\,\mathrm{KL}_\rho(\pi_\theta\|\pi_k),$$

where π plays the role of the competitor policy in the self-play game against π_θ. To analyze the inner loop using the results from Section C, we first show that the inner game at outer step k can be expressed as a usual λ-regularized preference game with an appropriate anchor policy.

Lemma 23 (Inner game as λ-regularized preference game)**.** *Let $\lambda \triangleq \beta + \beta_{\mathrm{target}} = \beta\Big(1 + \frac{1}{\eta}\Big)$ and define the geometric mixture*

$$\forall (x,y) \in \mathcal{X}\times\mathcal{Y} : \widetilde{\pi}_k(y|x) \propto [\pi^{\mathrm{ref}}(y|x)]^{\beta/\lambda}\,[\pi_k(y|x)]^{\beta_{\mathrm{target}}/\lambda} \\ = [\pi^{\mathrm{ref}}(y|x)]^{\eta/(1+\eta)}\,[\pi_k(y|x)]^{1/(1+\eta)}.$$

Then, up to additive constants independent of π and θ,

$$V_k(\pi,\mu) = \mathcal{P}(\mu \succ \pi) + \lambda\,\mathrm{KL}_\rho(\pi\|\widetilde{\pi}_k) + \textit{const} = V_\lambda^{\widetilde{\pi}_k}(\pi,\mu) + \textit{const}\,,$$
$$J_k(\theta;\pi) = \mathcal{P}(\pi \succ \pi_\theta) + \lambda\,\mathrm{KL}_\rho(\pi_\theta\|\widetilde{\pi}_k) + \textit{const} = J_\lambda^{\widetilde{\pi}_k}(\theta;\pi) + \textit{const}\,,$$

where $J_\lambda^{\widetilde{\pi}_k}$ is precisely the best-response objective from Section C with regularization parameter λ and anchor $\widetilde{\pi} = \widetilde{\pi}_k$.

Proof. For an arbitrary policy π, we use the identity for any context $x \in \mathcal{X}$

$$\beta\,\mathrm{KL}(\pi(x)\|\pi^{\mathrm{ref}}(x)) + \beta_{\mathrm{target}}\,\mathrm{KL}(\pi(x)\|\pi_k(x)) = \lambda\,\mathrm{KL}(\pi(x)\|\widetilde{\pi}_k(x)) + \mathrm{const}(\pi^{\mathrm{ref}},\pi_k,x)\,,$$

which follows by expanding the KL terms and grouping the π-dependent and constant parts. Taking the expectation with respect to ρ yields the expression for V_k. Substituting $\pi = \pi_\theta$ gives the expression for J_k. □

By Lemma 23, each inner problem at outer iterate k is exactly of the form studied in Section C, with regularization parameter $\lambda = \beta + \beta_{\mathrm{target}}$ and anchor $\widetilde{\pi}_k$. The only difference is that $\widetilde{\pi}_k$ now depends on π_k, but λ is fixed across outer iterations.

D.2.2 Inner accuracy and PP residual

The approximate PP analysis in Proposition 2 requires that the inner solver at outer step k returns a policy π_{k+1} such that

$$\big\|\nabla_\pi V_k(\pi_{k+1},\pi_{k+1})\big\|_{\mathrm{sp},\rho}^2 \le \varepsilon, \tag{44}$$

for some accuracy parameter $\varepsilon > 0$, uniformly in k. We note that in general, it is possible to construct a sequence of π_k such that for each step $\mathrm{SubOpt}_\lambda^{\widetilde{\pi}_k}(\pi_{k+1}) \to 0$, but $\|\nabla_\pi V_k(\pi_{k+1},\pi_{k+1})\|_{\mathrm{sp},\rho}^2 \not\to 0$. Thus, we need one additional assumption on the properties of parametrization to ensure that an approximate solution to the inner problem will yield small PP residual.

Assumption 11 (Gradient compatibility)**.** There exists $\varepsilon_{\mathrm{gc}} > 0$ and a constant $C_{\mathrm{gc}} > 0$ such that for any $k \in \mathbb{N}$, for any $\theta \in \Theta_{\mathcal{T}_k}$, the following inequality holds

$$\|\nabla_\pi V_k(\pi_\theta,\pi_\theta)\|_{\mathrm{sp},\rho}^2 \le C_{\mathrm{gc}} \cdot \|\nabla J_k(\theta;\pi_\theta)\|_2^2 + \varepsilon_{\mathrm{gc}}\,.$$

This assumption holds under both softmax and Fisher compatible parameterizations (see, e.g., Lemma 20). Next, we state the required regularity assumption which should be satisfied for each inner game to guarantee that the self-play policy gradient convergence is independent of step k.

Assumption 12 (Uniform inner-game regularity)**.** Fix $\beta > 0$ and $\eta > 0$, and let

$$\lambda \triangleq \beta + \beta_{\text{target}} = \beta\Big(1 + \frac{1}{\eta}\Big), \qquad \beta_{\text{target}} = \beta/\eta,$$

and let $\widetilde{\pi}_k$ be the mixed anchor from Lemma 23. There exist constants

$$G,\ L_{\beta,\eta},\ m_{\beta,\eta},\ \varepsilon_{\text{PL}},\ \varepsilon_{\text{grad}},\ \sigma^2_{\beta,\eta},\ v^2_{\beta,\eta}\ >\ 0\,,$$

independent of k, and a family of improvement operators $(\mathcal{T}_k)_{k\geq 0}$ such that for every outer iteration k the inner objective $J_k(\theta;\pi) = J_\lambda^{\widetilde{\pi}_k}(\theta;\pi)$ satisfies Assumptions 2, 3, 4, and 5 with constants bounded by $(G, L_{\beta,\eta}, m_{\beta,\eta}, \varepsilon_{\text{PL}})$, and the mini-batch gradient estimator $g_{k,t}$ satisfies Assumption 6 with bias $\varepsilon_{\text{grad}}$ and variance proxies $(\sigma^2_{\beta,\eta}, v^2_{\beta,\eta})$. Moreover, a PP step-size η satisfies PL-compatibility condition uniformly:

$$\beta(1 + 1/\eta) \cdot m_{\beta,\eta} \geq G^2\,.$$

We note that this assumption does not hold automatically for our previously chosen examples since many constants, such as $L_{\widetilde{\pi},\lambda}, m_{\widetilde{\pi},\lambda}$, in Assumption 3–6 depend on the reference policy $\widetilde{\pi}$, and our sequence of policies $\widetilde{\pi}_k$ might be ill-behaved. However, in Appendix D.3 we show that this assumption is in fact satisfied thanks to established convergence of the proximal point method in span-seminorm of log-probabilities.

Lemma 24 (From SPG progress to PP residual)**.** *Fix $k \geq 0$. Assume Assumption 11 with constants $(C_{\text{gc}}, \varepsilon_{\text{gc}})$. Assume additionally that $J_k(\cdot;\pi)$ is $L_{\beta,\eta}$-smooth (as in Assumption 12). Then for any $\theta \in \Theta_{\mathcal{T}_k}$,*

$$\big\|\nabla_\pi V_k(\pi_\theta, \pi_\theta)\big\|^2_{\text{sp},\rho}\ \leq\ 2C_{\text{gc}}\, L_{\beta,\eta} \cdot \text{SubOpt}^{\widetilde{\pi}_k}_\lambda(\pi_\theta)\ +\ \varepsilon_{\text{gc}}.$$

In particular, if the inner solver returns π_{k+1} such that $\text{SubOpt}^{\widetilde{\pi}_k}_\lambda(\pi_{k+1}) \leq \varepsilon_{\text{in}}$, then the PP residual satisfies

$$\big\|\nabla_\pi V_k(\pi_{k+1}, \pi_{k+1})\big\|^2_{\text{sp},\rho}\ \leq\ \varepsilon_{\text{PP}} \quad \textit{with} \quad \varepsilon_{\text{PP}} \triangleq 2C_{\text{gc}} L_{\beta,\eta}\, \varepsilon_{\text{in}} + \varepsilon_{\text{gc}}.$$

Proof. Assumption 11 gives

$$\|\nabla_\pi V_k(\pi_\theta, \pi_\theta)\|^2_{\text{sp},\rho} \leq C_{\text{gc}}\, \|\nabla_\theta J_k(\theta;\pi_\theta)\|_2^2 + \varepsilon_{\text{gc}}.$$

By $L_{\beta,\eta}$-smoothness of $\theta \mapsto J_k(\theta;\pi_\theta)$ (Assumption 3), we have

$$\|\nabla_\theta J_k(\theta;\pi_\theta)\|_2^2 \leq 2L_{\beta,\eta}\, \text{SubOpt}^{\widetilde{\pi}_k}_\lambda(\pi_\theta)\,,$$

which allows us to conclude the statement. □

Additionally, we provide some regularity conditions on the reference policy π^{ref} that serves as the initial policy for our algorithm.

Lemma 25 (Regularity of reference policy.)**.** *Let $\pi^{\text{ref}} \in \Pi$ be a full-support reference policy. Then it holds for any $\beta > 0$*

$$\text{KL}_\rho(\pi^\star_\beta \| \pi^{\text{ref}}) \leq \frac{1}{2\beta}, \qquad \text{SubOpt}_\beta(\pi^{\text{ref}}) \leq \frac{1}{2}, \qquad \|\log \pi^{\text{ref}} - \log \pi^\star_\beta\|^2_{\text{sp},\rho} \leq \frac{1}{4\beta^2}\,.$$

Proof. By optimality conditions, $\pi^\star_\beta$ can be characterized as a best response against itself, thus, for any $x \sim \text{supp}(\rho)$ and $y \in \mathcal{Y}$ we have

$$\log \pi^\star_\beta(y|x) = \frac{1}{\beta}\big(1/2 - \mathcal{P}(\pi^\star_\beta(x) \succ y|x)\big) + \log \pi^{\text{ref}}(y|x) + c^\star(x)\,,$$

where $c^\star(x)$ is a log-normalization constant, which is equal to

$$c^\star(x) = -\log \mathbb{E}_{y\sim\pi^{\text{ref}}(\cdot|x)}\left[\exp\left(\frac{1}{\beta}\big(1/2 - \mathcal{P}(\pi^\star_\beta(x) \succ y|x)\big)\right)\right].$$

In particular, this expression automatically implies the last bound since

$$\|\log \pi^{\mathrm{ref}}(\cdot|x) - \log \pi^\star_\beta(\cdot|x)\|_{\mathrm{sp}} \le \frac{1}{\beta}\|1/2 - \mathcal{P}(\pi^\star_\beta(x) \succ \cdot|x)\|_\infty \le \frac{1}{2\beta}\,.$$

For the expression for the KL divergence, we use the log-ratio expression to write

$$\mathrm{KL}_\rho(\pi^\star_\beta\|\pi^{\mathrm{ref}}) = \mathbb{E}_{x\sim\rho}\left[\frac{1}{\beta}\mathbb{E}_{y\sim\pi^\star_\beta(\cdot|x)}\left[1/2 - \mathcal{P}(\pi^\star_\beta(x) \succ y|x)\right] + c^\star(x)\right] = \mathbb{E}_{x\sim\rho}[c^\star(x)]\,.$$

We note that by Jensen's inequality, it holds

$$\begin{aligned}\exp(c^\star(x)) &= \frac{1}{\mathbb{E}_{y\sim\pi^{\mathrm{ref}}(\cdot|x)}\left[\exp\left(\frac{1}{\beta}\left(1/2 - \mathcal{P}(\pi^\star_\beta(x) \succ y|x)\right)\right)\right]} \\ &\le \frac{1}{\exp\left(\frac{1}{\beta}\left(\frac{1}{2} - \mathcal{P}(\pi^\star_\beta(x) \succ \pi^{\mathrm{ref}}(x)|x)\right)\right)} \le \exp\left(\frac{1}{2\beta}\right),\end{aligned}$$

thus $\mathrm{KL}_\rho(\pi^\star_\beta\|\pi^{\mathrm{ref}}) \le 1/(2\beta)$. Finally, for the suboptimality, we use the definition:

$$\mathrm{SubOpt}_\beta(\pi^{\mathrm{ref}}) = \max_\pi\left\{\frac{1}{2} - \mathcal{P}_\beta(\pi^{\mathrm{ref}} \succ \pi)\right\},$$

where

$$\mathcal{P}_\beta(\pi^{\mathrm{ref}} \succ \pi) = \mathcal{P}(\pi^{\mathrm{ref}} \succ \pi) - \beta\,\mathrm{KL}_\rho(\pi^{\mathrm{ref}}\|\pi^{\mathrm{ref}}) + \beta\,\mathrm{KL}_\rho(\pi\|\pi^{\mathrm{ref}})\,.$$

In particular, $\mathcal{P}_\beta(\pi^{\mathrm{ref}} \succ \pi) = \mathcal{P}(\pi^{\mathrm{ref}} \succ \pi) + \beta\,\mathrm{KL}_\rho(\pi\|\pi^{\mathrm{ref}}) \ge 0$. Thus, we have $\mathrm{SubOpt}_\beta(\pi^{\mathrm{ref}}) \le 1/2$. □

Lemma 26 (Regularity of initial policy)**.** *Let $\pi^{\mathrm{ref}} \in \Pi$ be a full-support reference policy and let π_k be the policy at outer iteration k of Algorithm 1. Then, for any $\beta > 0$ and $\eta > 0$, it holds that the mixed anchor policy $\widetilde{\pi}_k$ from Lemma 23 satisfies*

$$\mathrm{SubOpt}^{\widetilde{\pi}_k}_\lambda(\pi_k) \le \mathrm{SubOpt}_\beta(\pi_k)\,,$$

where $\lambda = \beta + \beta_{\mathrm{target}}$ with $\beta_{\mathrm{target}} = \beta/\eta$.

Proof. Using a property of $\widetilde{\pi}_k$ being a geometric mixture, we have

$$\begin{aligned}\mathcal{P}^{\widetilde{\pi}_k}_\lambda(\pi' \succ \pi) &= \mathcal{P}(\pi' \succ \pi) - \lambda\,\mathrm{KL}_\rho(\pi'\|\widetilde{\pi}_k) + \lambda\,\mathrm{KL}_\rho(\pi\|\widetilde{\pi}_k) \\ &= \mathcal{P}(\pi' \succ \pi) - \beta\,\mathrm{KL}_\rho(\pi'\|\pi^{\mathrm{ref}}) + \beta\,\mathrm{KL}_\rho(\pi\|\pi^{\mathrm{ref}}) \\ &\qquad - \beta_{\mathrm{target}}\,\mathrm{KL}_\rho(\pi'\|\pi_k) + \beta_{\mathrm{target}}\,\mathrm{KL}_\rho(\pi\|\pi_k)\,.\end{aligned}$$

Taking $\pi' = \pi_k$, the expression above gives

$$\begin{aligned}\mathcal{P}^{\widetilde{\pi}_k}_\lambda(\pi_k \succ \pi) &= \mathcal{P}(\pi_k \succ \pi) - \beta\,\mathrm{KL}_\rho(\pi_k\|\pi^{\mathrm{ref}}) + \beta\,\mathrm{KL}_\rho(\pi\|\pi^{\mathrm{ref}}) \\ &\qquad - \beta_{\mathrm{target}}\,\mathrm{KL}_\rho(\pi_k\|\pi_k) + \beta_{\mathrm{target}}\,\mathrm{KL}_\rho(\pi\|\pi_k) \\ &= \mathcal{P}_\beta(\pi_k \succ \pi) + \beta_{\mathrm{target}}\,\mathrm{KL}_\rho(\pi\|\pi_k) \ge \mathcal{P}_\beta(\pi_k \succ \pi)\,.\end{aligned}$$

Thus, we have

$$\mathrm{SubOpt}^{\widetilde{\pi}_k}_\lambda(\pi_k) \le \mathrm{SubOpt}_\beta(\pi_k)\,.$$

□

Finally, we establish a uniform bound on the suboptimality and log-probability span-norm of the iterates π_k assuming that each inner problem is solved up to a fixed accuracy. This result will be useful to select the number of inner iterations in a consistent manner.

Lemma 27 (Uniform regularity of PP iterates)**.** *Assume $\beta \le 1$ and fix $K \ge 0$. Assume that for all $k < K$ the k-th proximal subproblem is solved up to accuracy $\varepsilon_{\mathrm{PP}} > 0$:*

$$\left\|\nabla_\pi V_k(\pi_{k+1}, \pi_{k+1})\right\|^2_{\mathrm{sp},\rho} \le \varepsilon_{\mathrm{PP}}\,.$$

Then the iterate π_K satisfies

$$\|\log \pi_K - \log \pi^{\mathrm{ref}}\|_{\mathrm{sp},\rho} \le \frac{\frac{1}{2} + \sqrt{\varepsilon_{\mathrm{PP}}}}{\beta}, \qquad \mathrm{KL}_\rho(\pi_K \| \pi^{\mathrm{ref}}) \le \frac{1 + 2\sqrt{\varepsilon_{\mathrm{PP}}}}{\beta},$$

and consequently

$$\mathrm{SubOpt}_\beta(\pi_K) \le \frac{3}{2} + 2\sqrt{\varepsilon_{\mathrm{PP}}}, \qquad \left\|\log \pi_K - \log \pi^\star_\beta\right\|^2_{\mathrm{sp},\rho} \le \frac{(1 + \sqrt{\varepsilon_{\mathrm{PP}}})^2}{\beta^2}.$$

Proof. Let $\beta_{\mathrm{target}} \triangleq \beta/\eta$ and recall

$$V_k(\pi, \mu) = \mathcal{P}(\mu \succ \pi) + \beta\, \mathrm{KL}_\rho(\pi \| \pi^{\mathrm{ref}}) + \beta_{\mathrm{target}}\, \mathrm{KL}_\rho(\pi \| \pi_k).$$

Write $\zeta_{k+1} \triangleq \nabla_\pi V_k(\pi_{k+1}, \pi_{k+1})$. A direct differentiation (same as in Lemma 4) gives, for ρ-a.e. x and all y,

$$\zeta_{k+1}(x, y) = \mathcal{P}(\pi_{k+1}(x) \succ y \mid x) + \beta\Big(1 + \log \frac{\pi_{k+1}(y \mid x)}{\pi^{\mathrm{ref}}(y \mid x)}\Big) + \beta_{\mathrm{target}}\Big(1 + \log \frac{\pi_{k+1}(y \mid x)}{\pi_k(y \mid x)}\Big).$$

The additive "+1" terms are constant across y and therefore irrelevant for the span seminorm. Thus, for each x we may subtract a baseline $c_{k+1}(x)$ and rewrite as

$$\mathcal{P}(\pi_{k+1}(x) \succ y \mid x) + \beta \log \frac{\pi_{k+1}(y \mid x)}{\pi^{\mathrm{ref}}(y \mid x)} + \beta_{\mathrm{target}} \log \frac{\pi_{k+1}(y \mid x)}{\pi_k(y \mid x)} = e_{k+1}(x, y),$$

where $e_{k+1}(x) \triangleq \zeta_{k+1}(x) - c_{k+1}(x)\mathbf{1}$ satisfies $\|e_{k+1}\|_{\mathrm{sp},\rho} = \|\zeta_{k+1}\|_{\mathrm{sp},\rho} \le \sqrt{\varepsilon_{\mathrm{PP}}}$.

Rearranging and using $\beta + \beta_{\mathrm{target}} = \beta(1 + 1/\eta) = \beta(1+\eta)/\eta$, we obtain the (approximate) log equation

$$\log \pi_{k+1} = \frac{\eta}{1+\eta} \log \pi^{\mathrm{ref}} + \frac{1}{1+\eta} \log \pi_k - \frac{\eta}{\beta(1+\eta)} \mathcal{P}(\pi_{k+1} \succ \cdot) + \frac{\eta}{\beta(1+\eta)} e_{k+1} + (\text{const}). \tag{45}$$

Subtract $\log \pi^{\mathrm{ref}}$ and take $\|\cdot\|_{\mathrm{sp},\rho}$. Using Minkowski and that $\|e_{k+1}\|_{\mathrm{sp},\rho} \le \sqrt{\varepsilon_{\mathrm{PP}}}$, we get

$$\|\log \pi_{k+1} - \log \pi^{\mathrm{ref}}\|_{\mathrm{sp},\rho} \le \frac{1}{1+\eta} \|\log \pi_k - \log \pi^{\mathrm{ref}}\|_{\mathrm{sp},\rho} + \frac{\eta}{\beta(1+\eta)} \|\mathcal{P}(\pi_{k+1} \succ \cdot)\|_{\mathrm{sp},\rho} + \frac{\eta}{\beta(1+\eta)} \sqrt{\varepsilon_{\mathrm{PP}}}.$$

For each x, the vector $y \mapsto \mathcal{P}(\pi_{k+1}(x) \succ y \mid x)$ takes values in $[0,1]$, hence $\|\mathcal{P}(\pi_{k+1}(x) \succ \cdot \mid x)\|_{\mathrm{sp}} \le \frac{1}{2}$, and therefore $\|\mathcal{P}(\pi_{k+1} \succ \cdot)\|_{\mathrm{sp},\rho} \le \frac{1}{2}$. Thus, defining $D_k \triangleq \|\log \pi_k - \log \pi^{\mathrm{ref}}\|_{\mathrm{sp},\rho}$,

$$D_{k+1} \le \frac{1}{1+\eta} D_k + \frac{\eta}{\beta(1+\eta)} \Big(\frac{1}{2} + \sqrt{\varepsilon_{\mathrm{PP}}}\Big).$$

Since $D_0 = \|\log \pi_0 - \log \pi^{\mathrm{ref}}\|_{\mathrm{sp},\rho} = 0$, unrolling the recursion yields for all K:

$$D_K \le \frac{1}{\beta} \Big(\frac{1}{2} + \sqrt{\varepsilon_{\mathrm{PP}}}\Big).$$

This proves the first bound in the lemma.

Bounding $\mathrm{KL}_\rho(\pi_K \| \pi^{\mathrm{ref}})$ from span of log-ratios. Fix x and abbreviate $p = \pi_K(\cdot \mid x)$, $q = \pi^{\mathrm{ref}}(\cdot \mid x)$, and $r = \log(p/q)$. Because $\sum_y q_y e^{r_y} = 1$, we have $1 \le e^{\max_y r_y}$ and $1 \ge e^{\min_y r_y}$, so $\max r \ge 0 \ge \min r$ and therefore

$$\max_y r_y \le \max r - \min r = 2\|r\|_{\mathrm{sp}}.$$

Then

$$\mathrm{KL}(p \| q) = \sum_y p_y r_y \le \max_y r_y \le 2\|r\|_{\mathrm{sp}}.$$

Taking expectation over $x \sim \rho$ and using $\mathbb{E}\|r(x)\|_{\mathrm{sp}} \le \|r\|_{\mathrm{sp},\rho}$ gives

$$\mathrm{KL}_\rho(\pi_K \| \pi^{\mathrm{ref}}) \le 2\| \log \pi_K - \log \pi^{\mathrm{ref}} \|_{\mathrm{sp},\rho} \le \frac{1 + 2\sqrt{\varepsilon_{\mathrm{PP}}}}{\beta}.$$

Suboptimality bound. By definition,

$$\begin{aligned}\mathrm{SubOpt}_\beta(\pi_K) &= \max_\mu \Big(\tfrac{1}{2} - \mathcal{P}(\pi_K \succ \mu) + \beta\, \mathrm{KL}_\rho(\pi_K \| \pi^{\mathrm{ref}}) - \beta\, \mathrm{KL}_\rho(\mu \| \pi^{\mathrm{ref}}) \Big) \\ &\le \tfrac{1}{2} + \beta\, \mathrm{KL}_\rho(\pi_K \| \pi^{\mathrm{ref}}).\end{aligned}$$

Plugging the KL bound yields

$$\mathrm{SubOpt}_\beta(\pi_K) \le \tfrac{1}{2} + \big(1 + 2\sqrt{\varepsilon_{\mathrm{PP}}}\big) = \tfrac{3}{2} + 2\sqrt{\varepsilon_{\mathrm{PP}}}\,.$$

Span bound to $\pi^\star_\beta$. By the triangle inequality,

$$\| \log \pi_K - \log \pi^\star_\beta \|_{\mathrm{sp},\rho} \le \| \log \pi_K - \log \pi^{\mathrm{ref}} \|_{\mathrm{sp},\rho} + \| \log \pi^{\mathrm{ref}} - \log \pi^\star_\beta \|_{\mathrm{sp},\rho}.$$

By Lemma 25, $\| \log \pi^{\mathrm{ref}} - \log \pi^\star_\beta \|_{\mathrm{sp},\rho} \le \frac{1}{2\beta}$. Therefore

$$\| \log \pi_K - \log \pi^\star_\beta \|_{\mathrm{sp},\rho} \le \frac{\frac{1}{2} + \sqrt{\varepsilon_{\mathrm{PP}}}}{\beta} + \frac{1}{2\beta} = \frac{1 + \sqrt{\varepsilon_{\mathrm{PP}}}}{\beta},$$

and squaring gives the last statement of the lemma. □

D.2.3 Deterministic PP–SPG

For a warm-up, assume the inner-loop gradients are exact: $g_{k,t} = \nabla J_k(\theta_{k,t}; \pi_{k,t})$.

Corollary 5 (Deterministic PP–SPG). *Assume $\beta \le 1$ and let $\varepsilon_{\mathrm{in}} > 0$. Assume that parameterization and PP learning rate η satisfy Assumptions 11 and 12. Then, the deterministic version of Algorithm 1 using*

$$T_k \ge \frac{4L_{\beta,\eta}}{m_{\beta,\eta}} \log\left(\frac{3/2 + 2\sqrt{\varepsilon_{\mathrm{PP}}}}{\varepsilon_{\mathrm{in}}} \right)$$

iterations for each inner loop, where $\varepsilon_{\mathrm{PP}} \triangleq 2C_{\mathrm{gc}} L_{\beta,\eta} \left(\varepsilon_{\mathrm{in}} + \frac{3\varepsilon_{\mathrm{PL}}}{2m_{\beta,\eta}} \right) + \varepsilon_{\mathrm{gc}}$, with a learning rate $\gamma_{k,t} \equiv \gamma = 1/(2L_{\beta,\eta})$, for any $k \ge 0$ it holds

$$\mathrm{KL}_\rho(\pi^\star_\beta \| \pi_k) \le (1 + \eta/2)^{-k} \cdot \frac{1}{2\beta} + \frac{2\varepsilon_{\mathrm{PP}}}{\beta^2}\,,$$

and

$$\mathrm{SubOpt}_\beta(\pi_k) \le (1 + \eta/2)^{-k} \cdot \left(\frac{5}{2\beta^2} + \frac{1}{\eta} \right) + \left(\frac{5}{2\beta^2} + \frac{1}{\eta} \right) \frac{4\varepsilon_{\mathrm{PP}}}{\beta}\,,$$

and, moreover,

$$\left\| \log \pi_k - \log \pi^\star_\beta \right\|^2_{\mathrm{sp},\rho} \le \frac{1}{4\beta^2}(1 + \eta)^{-k} + \frac{1}{\beta^3} \cdot (1 + \eta/2)^{-k} + \frac{4\varepsilon_{\mathrm{PP}}}{\beta^4}\,.$$

Proof. We want to show that for all $k \ge 0$, the following bounds hold:

$$\left\| \nabla_\pi V_k(\pi_{k+1}, \pi_{k+1}) \right\|^2_{\mathrm{sp},\rho} \le \varepsilon_{\mathrm{PP}} \triangleq 2C_{\mathrm{gc}} L_{\beta,\eta} \left(\varepsilon_{\mathrm{in}} + \frac{3\varepsilon_{\mathrm{PL}}}{2m_{\beta,\eta}} \right) + \varepsilon_{\mathrm{gc}}\,.$$

Assume now $k \ge 0$ and that the statements hold for $k' < k$. We will show they hold for k. By Lemma 25 for $k = 0$ and by the induction hypothesis and Lemma 27 for $k > 0$, we have

$$\mathrm{SubOpt}_\beta(\pi_k) \le \frac{3}{2} + 2\sqrt{\varepsilon_{\mathrm{PP}}}\,.$$

Applying Proposition 3, Lemma 26, and a bound $1 + x \le \mathrm{e}^x$ to the inner loop at each k gives

$$\begin{aligned}\mathrm{SubOpt}_\lambda^{\widetilde{\pi}_k}(\pi_{k+1}) &\le \left(1 - \frac{m_{\beta,\eta}}{4L_{\beta,\eta}}\right)^{T_k} \mathrm{SubOpt}_\lambda^{\widetilde{\pi}_k}(\pi_k) + \frac{3\varepsilon_{\mathrm{PL}}}{2m_{\beta,\eta}} \\ &\le \left(1 - \frac{m_{\beta,\eta}}{4L_{\beta,\eta}}\right)^{T_k} \mathrm{SubOpt}_\beta(\pi_k) + \frac{3\varepsilon_{\mathrm{PL}}}{2m_{\beta,\eta}} \\ &\le \exp\left(-\frac{m_{\beta,\eta}}{4L_{\beta,\eta}} T_k\right)\left(\frac{3}{2} + 2\sqrt{\varepsilon_{\mathrm{PP}}}\right) + \frac{3\varepsilon_{\mathrm{PL}}}{2m_{\beta,\eta}} \le \varepsilon_{\mathrm{in}} + \frac{3\varepsilon_{\mathrm{PL}}}{2m_{\beta,\eta}}\,.\end{aligned}$$

By Lemma 24,

$$\left\|\nabla_\pi V_k(\pi_{k+1}, \pi_{k+1})\right\|_{\mathrm{sp},\rho}^2 \le 2C_{\mathrm{gc}} L_{\beta,\eta}\left(\varepsilon_{\mathrm{in}} + \frac{3\varepsilon_{\mathrm{PL}}}{2m_{\beta,\eta}}\right) + \varepsilon_{\mathrm{gc}} = \varepsilon_{\mathrm{PP}}\,.$$

Thus Proposition 2 applies to the outer PP loop with $\varepsilon = \varepsilon_{\mathrm{PP}}$, yielding the following corollary. □

D.2.4 Stochastic PP–SPG

We now analyze the stochastic version of Algorithm 1, where the inner-loop gradients are estimated by mini-batches and satisfy Assumption 6. The proof mirrors the deterministic case, but we additionally track the confidence parameter and apply a union bound over outer iterations.

Corollary 6 (Stochastic PP–SPG). *Assume $\beta \le 1$ and let $\delta \in (0,1)$. Assume Assumptions 11 and 12. Fix a target inner accuracy $\varepsilon_{\mathrm{in}} > 0$ and a target number of outer steps $K \in \mathbb{N}$. Define the per-outer failure probability as*

$$\delta_k \triangleq \frac{\delta}{K}, \qquad k \in \{0, \dots, K-1\}\,.$$

and a target PP-residual accuracy as

$$\varepsilon_{\mathrm{PP}} \triangleq 2C_{\mathrm{gc}} L_{\beta,\eta}\left(\varepsilon_{\mathrm{in}} + \frac{6\log(K\mathrm{e}/\delta)}{m_{\beta,\eta}} \cdot \left(\varepsilon_{\mathrm{PL}} + 15/7 \cdot \varepsilon_{\mathrm{grad}}^2\right)\right) + \varepsilon_{\mathrm{gc}}\,.$$

Define $\kappa_{\beta,\eta} = \frac{L_{\beta,\eta}}{m_{\beta,\eta}} \ge 1$ and sequences $(\gamma_t)_{t\ge 0}$ and $(B_t)_{t\ge 0}$ as follows

$$\gamma_t = \frac{4t + 32\kappa_{\beta,\eta} - 2}{m_{\beta,\eta}(t + 8\kappa_{\beta,\eta})^2} = \Theta\left(\frac{1}{m_{\beta,\eta} t}\right), \qquad B_t = \left\lceil \frac{t + 8\kappa_{\beta,\eta}}{m_{\beta,\eta}} \right\rceil = \Theta\left(\frac{t}{m_{\beta,\eta}}\right).$$

Then, Algorithm 1 with $\gamma_{k,t} \equiv \gamma_t$ and $B_{k,t} \equiv B_t$, and

$$T_k = \widetilde{\mathcal{O}}\left(\frac{v_{\beta,\eta}^2}{\varepsilon_{\mathrm{in}}} + \frac{\kappa_{\beta,\eta}}{\sqrt{\varepsilon_{\mathrm{in}}}} + \sqrt{\frac{\kappa_{\beta,\eta}(\sigma_{\beta,\eta}^2 + v_{\beta,\eta}^2)}{\varepsilon_{\mathrm{in}}}}\right)$$

outputs $\{\pi_k\}_{k\in[0,\dots,K]}$ such that with probability at least $1-\delta$ for all $k = 0, \dots, K$ the following holds:

$$\begin{aligned}\mathrm{KL}_\rho(\pi_\beta^\star \| \pi_k) &\le (1+\eta/2)^{-k} \cdot \frac{1}{2\beta} + \frac{2\varepsilon_{\mathrm{PP}}}{\beta^2}, \\ \mathrm{SubOpt}_\beta(\pi_k) &\le (1+\eta/2)^{-k} \cdot \left(\frac{5}{2\beta^2} + \frac{1}{\eta}\right) + \left(\frac{5}{2\beta^2} + \frac{1}{\eta}\right)\frac{4\varepsilon_{\mathrm{PP}}}{\beta}, \\ \left\|\log \pi_k - \log \pi_\beta^\star\right\|_{\mathrm{sp},\rho}^2 &\le \frac{1}{4\beta^2}(1+\eta)^{-k} + \frac{1}{\beta^3}(1+\eta/2)^{-k} + \frac{4\varepsilon_{\mathrm{PP}}}{\beta^4}\,.\end{aligned}$$

Proof. We prove that the PP residual condition holds uniformly for all outer steps with probability at least $1-\delta$, and then apply Proposition 2 exactly as in the deterministic case.

Fix an outer iteration $k \in \{0, \dots, K-1\}$. Condition on the history up to the beginning of the k-th inner loop (so π_k and hence $\widetilde{\pi}_k$ are fixed). By Assumption 12, the inner objective J_k satisfies the same regularity constants $(L_{\beta,\eta}, m_{\beta,\eta}, \varepsilon_{\mathrm{PL}})$ and the stochastic gradients satisfy the same noise constants $(\varepsilon_{\mathrm{grad}}, \sigma_{\beta,\eta}^2, v_{\beta,\eta}^2)$. Therefore Proposition 4 is applicable to the inner loop.

Moreover, as in the deterministic proof, for $k = 0$ Lemma 25 controls the initial outer iterate, and for $k > 0$ the induction hypothesis together with Lemma 27 gives a uniform bound on $\mathrm{SubOpt}_\beta(\pi_k)$ and hence via Lemma 26 a uniform bound on $\mathrm{SubOpt}_\lambda^{\widetilde{\pi}_k}(\pi_k)$ required to instantiate the SPG bound. Concretely, on the event that the residual bounds hold for all steps $k' < k$, we have

$$\mathrm{SubOpt}_\lambda^{\widetilde{\pi}_k}(\pi_k) \le \mathrm{SubOpt}_\beta(\pi_k) \le \frac{3}{2} + 2\sqrt{\varepsilon_{\mathrm{PP}}}.$$

Now apply Proposition 4 to the inner run at step k with confidence $\delta_k = \delta/K$ after T_k steps and required learning rate and batch size sequences (γ_t) and (B_t). It implies that with probability at least $1 - \delta_k$,

$$\begin{aligned}\mathrm{SubOpt}_\lambda^{\widetilde{\pi}_k}(\pi_{k+1}) \le{}& \frac{64 \cdot \kappa_{\beta,\eta}^2 \log(Ke/\delta)}{(T_k + 8\kappa_{\beta,\eta} - 1)^2}\left(\frac{3}{2} + 2\sqrt{\varepsilon_{\mathrm{PP}}}\right) \\ &+ \frac{24 \cdot \kappa_{\beta,\eta} \log(Ke/\delta) \cdot \log(1 + T_k/(2\kappa_{\beta,\eta}))}{(T_k + 8\kappa_{\beta,\eta} - 1)^2}\left(\frac{675 \cdot \sigma_{\beta,\eta}^2}{49} + 2v_{\beta,\eta}^2\right) \\ &+ \frac{6G^2 \cdot v_{\beta,\eta}^2 \log(Ke/\delta)}{\lambda m_{\beta,\eta} \cdot (T_k + 8\kappa_{\beta,\eta} - 1)} + \frac{6\log(Ke/\delta)}{m_{\beta,\eta}} \cdot \left(\varepsilon_{\mathrm{PL}} + 15/7 \cdot \varepsilon_{\mathrm{grad}}^2\right),\end{aligned}$$

where $\lambda = \beta(1 + 1/\eta)$. To guarantee that all terms with dependence on T_k are at most $\varepsilon_{\mathrm{in}}$, it suffices to choose T_k such that

$$\frac{64 \cdot \kappa_{\beta,\eta}^2 \log(Ke/\delta)}{(T_k + 8\kappa_{\beta,\eta} - 1)^2}\left(\frac{3}{2} + 2\sqrt{\varepsilon_{\mathrm{PP}}}\right) \le \frac{\varepsilon_{\mathrm{in}}}{3}, \tag{46}$$

$$\frac{24 \cdot \kappa_{\beta,\eta} \log(Ke/\delta) \cdot \log(1 + T_k/(2\kappa_{\beta,\eta}))}{(T_k + 8\kappa_{\beta,\eta} - 1)^2}\left(\frac{675 \cdot \sigma_{\beta,\eta}^2}{49} + 2v_{\beta,\eta}^2\right) \le \frac{\varepsilon_{\mathrm{in}}}{3}, \tag{47}$$

$$\frac{6G^2 \cdot v_{\beta,\eta}^2 \log(Ke/\delta)}{\lambda m_{\beta,\eta} \cdot (T_k + 8\kappa_{\beta,\eta} - 1)} \le \varepsilon_{\mathrm{in}}/3\,. \tag{48}$$

Our choice $\beta(1 + 1/\eta) \cdot m_{\beta,\eta} = G^2$ simplifies the last condition (48) to

$$T_k \ge \frac{18 \cdot v_{\beta,\eta}^2 \log(Ke/\delta)}{\varepsilon_{\mathrm{in}}} - 8\kappa_{\beta,\eta} + 1\,,$$

which asymptotically scales as $\mathcal{O}\left(v_{\beta,\eta}^2 \log(K/\delta)/\varepsilon_{\mathrm{in}}\right)$. For the first two conditions (46) and (47), it suffices to choose

$$\begin{aligned}T_k \ge \max\Bigg\{&\sqrt{\frac{192 \cdot \kappa_{\beta,\eta}^2 \log(Ke/\delta)\left(\frac{3}{2} + 2\sqrt{\varepsilon_{\mathrm{PP}}}\right)}{\varepsilon_{\mathrm{in}}}}, \\ &\sqrt{\frac{72 \cdot \kappa_{\beta,\eta} \log(Ke/\delta)\left(\frac{675 \cdot \sigma_{\beta,\eta}^2}{49} + 2v_{\beta,\eta}^2\right)\log(1 + T_k/(2\kappa_{\beta,\eta}))}{\varepsilon_{\mathrm{in}}}}\Bigg\}.\end{aligned}$$

In particular, the first term scales as $\mathcal{O}\left(\kappa_{\beta,\eta}\sqrt{\frac{\log(K/\delta)}{\varepsilon_{\mathrm{in}}}}\right)$ and the second term scales as $\tilde{\mathcal{O}}\left(\sqrt{\frac{\kappa_{\beta,\eta}(\sigma_{\beta,\eta}^2 + v_{\beta,\eta}^2)\log(K/\delta)}{\varepsilon_{\mathrm{in}}}}\right)$.

Under this choice, we have shown that with probability at least $1 - \delta_k$,

$$\mathrm{SubOpt}_\lambda^{\widetilde{\pi}_k}(\pi_{k+1}) \le \varepsilon_{\mathrm{in}} + \frac{6\log(Ke/\delta)}{m_{\beta,\eta}} \cdot \left(\varepsilon_{\mathrm{PL}} + 15/7 \cdot \varepsilon_{\mathrm{grad}}^2\right).$$

On the same event, Lemma 24 gives

$$\left\|\nabla_\pi V_k(\pi_{k+1}, \pi_{k+1})\right\|_{\mathrm{sp},\rho}^2 \le 2C_{\mathrm{gc}}L_{\beta,\eta}\left(\varepsilon_{\mathrm{in}} + \frac{6\log(Ke/\delta)}{m_{\beta,\eta}} \cdot \left(\varepsilon_{\mathrm{PL}} + 15/7 \cdot \varepsilon_{\mathrm{grad}}^2\right)\right) + \varepsilon_{\mathrm{gc}} = \varepsilon_{\mathrm{PP}}.$$

Define the event

$$\mathcal{E} \triangleq \bigcap_{k=0}^{K-1} \mathcal{E}_k\,, \qquad \mathcal{E}_k \triangleq \left\{ \left\| \nabla_\pi V_k(\pi_{k+1}, \pi_{k+1}) \right\|_{\mathrm{sp},\rho}^2 \le \varepsilon_{\mathrm{PP}} \right\}.$$

By the per-step success probability $\mathbb{P}(\mathcal{E}_k) \ge 1 - \delta_k$ and a union bound, $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$. On $\mathcal{E}$, the residual condition (44) holds uniformly for all outer iterations with $\varepsilon = \varepsilon_{\mathrm{PP}}$. Thus Proposition 2 applies, giving the three displayed bounds. Finally, since $\pi_0 = \pi^{\mathrm{ref}}$ we use Lemma 25 to bound initial quantities as in the deterministic case. □

D.3 Assumption Verification

First, we prove a technical lemma that will be useful to verify all assumptions connected to a non-degeneracy of reference policies.

Lemma 28. *Let $\pi, \mu \in \Pi$ be two full-support policies and define $\pi_{\min}(x) = \min_{y\in\mathcal{Y}} \pi(y|x)$ and $\mu_{\min}(x) = \min_{y\in\mathcal{Y}} \mu(y|x)$. Then, the following bound holds*

$$\forall x \in \mathcal{X} : \max_{y\in\mathcal{Y}} \left| \log \frac{\pi(y|x)}{\mu(y|x)} \right| \le 2\|\log \pi(x) - \log \mu(x)\|_{\mathrm{sp}}\,,$$

$$\mathbb{E}_{x\sim\rho}\left[\max_{y\in\mathcal{Y}} \log^2\left(\frac{\pi(y|x)}{\mu(y|x)} \right) \right] \le 4\|\log \pi - \log \mu\|_{\mathrm{sp},\rho}^2\,.$$

Additionally, the following bound holds

$$\mathbb{E}_{x\sim\rho}\left[\log^2\left(\frac{1}{\pi_{\min}(x)} \right) \right] \le 2\mathbb{E}_{x\sim\rho}\left[\log^2\left(\frac{1}{\mu_{\min}(x)} \right) \right] + 8\|\log \pi - \log \mu\|_{\mathrm{sp},\rho}^2\,.$$

Proof. First, we use the relation between the span seminorm and a range:

$$\|\log \pi(x) - \log \mu(x)\|_{\mathrm{sp}} = \frac{1}{2}\left(\max_{y\in\mathcal{Y}} \log \frac{\pi(y|x)}{\mu(y|x)} - \min_{y\in\mathcal{Y}} \log \frac{\pi(y|x)}{\mu(y|x)} \right).$$

We notice that for any pair of policies it holds $\max_{y\in\mathcal{Y}} \log \frac{\pi(y|x)}{\mu(y|x)} \ge 0$ and $\min_{y\in\mathcal{Y}} \log \frac{\pi(y|x)}{\mu(y|x)} \le 0$, thus we have

$$\max\left\{ \max_{y\in\mathcal{Y}} \log \frac{\pi(y|x)}{\mu(y|x)}, -\min_{y\in\mathcal{Y}} \log \frac{\pi(y|x)}{\mu(y|x)} \right\} \le 2\|\log \pi(x) - \log \mu(x)\|_{\mathrm{sp}}\,.$$

Finally, we note that the left-hand side is equal to $\max_{y\in\mathcal{Y}} |\log \frac{\pi(y|x)}{\mu(y|x)}|$. Taking a square and expectation over $x \sim \rho$ concludes the first part of the proof.

The second part of the proof follows from a bound

$$\|\log \pi(x)\|_\infty \le \|\log \pi(x) - \log \mu(x)\|_\infty + \|\log \mu(x)\|_\infty \le 2\|\log \pi(x) - \log \mu(x)\|_{\mathrm{sp}} + \|\log \mu(x)\|_\infty\,,$$

which follows from the triangle inequality and the first part of the lemma, squaring both sides, taking expectation over $x \sim \rho$, and using a bound $(a+b)^2 \le 2a^2 + 2b^2$. □

In particular, this lemma allows us to show that if a policy π_k generated by the PP–SPG algorithm satisfies $\|\log \pi_k - \log \pi_\beta^\star\|_{\mathrm{sp},\rho}^2 \le C$ for some constant $C > 0$, then the mixed anchor $\widetilde{\pi}_k$ for the next step satisfies Assumption 7. Next, we prove the result that connects the inexact solution to the inner problem to PP residuals needed to establish the convergence of the outer algorithm.

D.4 Verification for Softmax Parametrization

In this section, we show that context-free softmax parametrization satisfies Assumptions 11 and 12.

Lemma 29. *Consider a context-free setting and assume that $\|\nabla V_k(\pi_{k+1}, \pi_{k+1})\|_{\mathrm{sp}}^2 \le \varepsilon_{\mathrm{PP}}$ for all $k < K$. Then for the next step K it holds*

$$\log(1/\widetilde{\pi}_{K,\min}) \le \log(1/\pi_{\min}^{\mathrm{ref}}) + \frac{1 + 2\sqrt{\varepsilon_{\mathrm{PP}}}}{(1+\eta)\beta}\,,$$

or, equivalently,

$$\frac{1}{\widetilde{\pi}_{K,\min}} \le \frac{1}{\pi_{\min}^{\mathrm{ref}}} \cdot \exp\left(\frac{1 + 2\sqrt{\varepsilon_{\mathrm{PP}}}}{(1+\eta)\beta} \right).$$

Proof. We start from the first inequality of Lemma 28 with $\pi = \pi^{\text{ref}}$ and $\mu = \widetilde{\pi}_k$

$$\|\log \pi^{\text{ref}} - \log \widetilde{\pi}_k\|_\infty \le 2\|\log \pi^{\text{ref}} - \log \widetilde{\pi}_k\|_{\text{sp}} .$$

For the left-hand side, we apply a reverse triangle inequality and a fact $\|\log \widetilde{\pi}_k\|_\infty = \log(1/\widetilde{\pi}_{k,\min})$. For the right-hand side, we use a definition of a geometric mixture. Overall, we have

$$\log(1/\widetilde{\pi}_{k,\min}) \le \log(1/\pi^{\text{ref}}_{\min}) + \frac{2}{1+\eta}\|\log \pi_k - \log \pi^{\text{ref}}\|_{\text{sp}} .$$

Next, we use Lemma 27 to bound the second term. Thus, we have

$$\log(1/\widetilde{\pi}_{k,\min}) \le \log(1/\pi^{\text{ref}}_{\min}) + \frac{1 + 2\sqrt{\varepsilon_{\text{PP}}}}{(1+\eta)\beta} .$$

Exponentiating both sides concludes the proof. □

Also, we will need the following technical lemma.

Lemma 30. *Let $x \in \mathbb{R}^d$, then for any $\pi \in \Delta_d$*

$$\|H(\pi)x\|_{\text{sp}} \ge \pi_{\min} \cdot \|x\|_{\text{sp}} ,$$

where $\pi_{\min} = \min_{i\in[d]} \pi(i)$ and $H(\pi) = \operatorname{diag}(\pi) - \pi\pi^\intercal$.

Proof. Without loss of generality, we assume that $\pi(i) > 0$ for any $i \in [d]$, otherwise the statement trivially holds.

First, we notice that $\|x\|_{\text{sp}} = 1/2 \cdot (\max_i x_i - \min_i x_i)$ and define $i_{\max} = \arg\max_i x_i$ and $i_{\min} = \arg\min_i x_i$.

Then, we notice that $H(\pi)x = \pi \odot (x - \mu\mathbf{1})$, where $\mu = \pi^\intercal x$ is a mean of x under π. Since π is a positive measure, we have $x_{i_{\max}} \ge \mu$ and $x_{i_{\min}} \le \mu$, therefore

$$\max_i \pi(i)(x_i - \mu) \ge \pi(i_{\max}) \cdot \underbrace{(x_{i_{\max}} - \mu)}_{\ge 0} \ge \pi_{\min} \cdot (x_{i_{\max}} - \mu)$$

$$\min_i \pi(i)(x_i - \mu) \le \pi(i_{\min}) \underbrace{(x_{i_{\min}} - \mu)}_{\le 0} \le \pi_{\min} \cdot (x_{i_{\min}} - \mu) ,$$

thus

$$\begin{aligned}\|H(\pi)x\|_{\text{sp}} &= (1/2) \cdot \Big(\max_i \pi(i)(x_i - \mu) - \min_i \pi(i)(x_i - \mu)\Big) \\ &\ge (1/2)\pi_{\min}(x_{i_{\max}} - \mu - x_{i_{\min}} + \mu) = \pi_{\min}\|x\|_{\text{sp}}.\end{aligned}$$

□

Corollary 7 (Convergence guarantees for Softmax parametrization)**.** *Let $\theta \mapsto \pi_\theta$ be a context-free softmax parametrization and $(\mathcal{T}_k)_{k\ge 0}$ a sequence of improvement operators $\mathcal{T}_k \equiv \mathcal{T}^{\pi^{\text{ref}}}_{\tau_{k,0}}$ for $\tau_{k,0}$ equal to τ_0 with $\widetilde{\pi} = \widetilde{\pi}_k$, as defined in Lemma 16 with $\nu = \pi^{\text{ref}}$ and $\widetilde{\pi} = \widetilde{\pi}_k$. Assume $\beta \le 1$. Fix $\varepsilon_{\text{in}} \in (0, (\tau_0\pi^{\text{ref}}_{\min})^2/(2L_{\beta,\eta}))$ as a desired accuracy for approximate solving of the inner problem, where τ_0 and $L_{\beta,\eta}$ are defined below.*

Then the following statements hold.

(i) Assumption 11 holds with $C_{\text{gc}} = 1/(\tau_0 \cdot \pi^{\text{ref}}_{\min})^2$ and $\varepsilon_{\text{gc}} = 0$, where

$$\tau_0 \triangleq \min\left\{\exp\left(-\frac{3+\eta}{\beta(1+\eta)}\right), (1 + 1/\pi^{\text{ref}}_{\min})^{-1}\right\};$$

(ii) There exists a truncation threshold $M_k > 0$ such that Assumption 12 holds with

$$G = 1, \quad L_{\beta,\eta} = \frac{5}{2}\big(1 + \beta(1+1/\eta)\log(1/\pi^{\text{ref}}_{\min}) + 3/\eta\big) + \beta(1+1/\eta)(4 + \log|\mathcal{Y}|) ,$$

$$m_{\beta,\eta} = \beta(1+1/\eta)\exp\left(-\frac{2}{\beta(1+1/\eta)}\right)\times$$

$$\times \exp\left(\min\left\{-\frac{6}{\beta(1+\eta)}, 2\log\left(\frac{\pi^{\text{ref}}_{\min}}{1+\pi^{\text{ref}}_{\min}}\right)\right\}\right) \cdot (\pi^{\text{ref}}_{\min})^2 ,$$

$$\varepsilon_{\text{PL}} = 0 , \qquad \varepsilon_{\text{grad}} = \sqrt{\varepsilon_{\text{in}}}, \qquad \sigma^2_{\beta,\eta} = D^2_{\beta,\eta}, \qquad v^2_{\beta,\eta} = 6D^2_{\beta,\eta} ,$$

where

$$D^2_{\beta,\eta} = 2\left(1 + 4\beta(1+1/\eta)\log\left(\frac{4\sqrt{2}\cdot\beta(1+1/\eta)}{\pi^{\mathrm{ref}}_{\min}}\right) + 12/\eta + 2\beta(1+1/\eta)\log(1/\varepsilon_{\mathrm{in}})\right)^2,$$

and η is chosen as solution to $\beta(1+1/\eta)\cdot m_{\beta,\eta} = 1$.

(iii) Algorithm 1 outputs $\mathcal{O}(\varepsilon_{\mathrm{in}})$-optimal regularized policy after $K = \widetilde{\mathcal{O}}(1)$ outer iterations, $T_k = \widetilde{\mathcal{O}}(1/\varepsilon_{\mathrm{in}})$ inner iterations, using in total $\widetilde{\mathcal{O}}(1/\varepsilon^2_{\mathrm{in}})$ samples, ignoring constants depending on $\pi^{\mathrm{ref}}_{\min}, \pi^\star_{\beta,\min}, \beta, \eta$ and logarithmic terms in $\varepsilon_{\mathrm{in}}$ and the confidence level δ.

Proof. To prove the statements, we also need to prove that for all k it holds $\|\nabla V_k(\pi_{k+1}, \pi_{k+1})\|^2_{\mathrm{sp},\rho} \le 1$. We prove this statement by induction over steps k. We assume that for all $k' < k$ it holds $\|\nabla V_{k'}(\pi_{k'+1}, \pi_{k'+1})\|^2_{\mathrm{sp},\rho} \le 1$ for all $k' < k$. This assumption allows us to use Lemma 29 to control the minimum probability of the mixed anchor policy $\widetilde{\pi}_k$.

Next, we notice that under our improvement operator choice, for any k and any $\theta \in \Theta_{\mathcal{T}_k}$ it holds $\pi_{\theta,\min} \ge \tau_{k,0}\cdot\pi^{\mathrm{ref}}_{\min}$, where $\tau_{k,0}$ is defined in Lemma 16 with $\nu = \pi^{\mathrm{ref}}$, $\widetilde{\pi} = \widetilde{\pi}_k$, and $\lambda = \beta(1+1/\eta)$. In particular, we have the following expression for $\tau_{k,0}$:

$$\tau_{k,0} = \min\left\{\exp\left(-\tfrac{1}{\lambda} - 2\|\log\widetilde{\pi}_k - \log\pi^{\mathrm{ref}}\|_{\mathrm{sp}}\right),\ (1 + 1/\pi^{\mathrm{ref}}_{\min})^{-1}\right\}.$$

To bound this constant away from zero uniformly over k, we need to show that $\|\log\widetilde{\pi}_k - \log\pi^{\mathrm{ref}}\|_{\mathrm{sp}}$ is uniformly bounded over k. For that, we start from a definition of geometric mixture and then apply Lemma 27:

$$\|\log\widetilde{\pi}_k - \log\pi^{\mathrm{ref}}\|_{\mathrm{sp}} = \frac{1}{1+\eta}\|\log\pi_k - \log\pi^{\mathrm{ref}}\|_{\mathrm{sp}} \le \frac{1/2 + \sqrt{\varepsilon_{\mathrm{PP}}}}{\beta(1+\eta)} \le \frac{3}{2\beta(1+\eta)}.$$

Thus, we have

$$\tau_{k,0} \ge \min\left\{\exp\left(-\frac{3+\eta}{\beta(1+\eta)}\right), (1+1/\pi^{\mathrm{ref}}_{\min})^{-1}\right\} \triangleq \tau_0 > 0,$$

uniformly over k.

Next, we can prove the first statement of the lemma. Notice that $\nabla J^{\widetilde{\pi}_k}(\theta;\pi) = H(\pi_\theta)\cdot\nabla V_k(\pi_\theta;\pi)$, where $H(\pi_\theta) = \mathrm{diag}(\pi_\theta) - \pi_\theta\pi_\theta^\intercal$ is the parametrization Jacobian matrix. Thus, we can apply Lemma 30 to obtain for any $\theta \in \Theta_{\mathcal{T}_k}$

$$\|\nabla J^{\widetilde{\pi}_k}(\theta;\pi)\|_2 \ge \|\nabla J^{\widetilde{\pi}_k}(\theta;\pi)\|_{\mathrm{sp}} \ge \pi_{\theta,\min}\cdot\|\nabla V_k(\pi_\theta;\pi)\|_{\mathrm{sp}} \ge \tau_0\cdot\pi^{\mathrm{ref}}_{\min}\cdot\|\nabla V_k(\pi_\theta;\pi)\|_{\mathrm{sp}},$$

thus concluding the proof of the first statement with $C_{\mathrm{gc}} = 1/(\tau_0\cdot\pi^{\mathrm{ref}}_{\min})^2$ and $\varepsilon_{\mathrm{gc}} = 0$.

The second statement follows from Proposition 5, a bound on $\log(1/\widetilde{\pi}_{k,\min})$ which follows from Lemma 29, and taking $\varepsilon_{\mathrm{grad}} = \sqrt{\varepsilon_{\mathrm{in}}}$. In fact, a constant $G = 1$ does not depend on k; the bound on a smoothness coefficient follows using $\lambda = \beta(1+1/\eta)$:

$$\begin{aligned}
L_{\widetilde{\pi},\lambda} &= \frac{5}{2}\left(1 + \lambda\log(1/\widetilde{\pi}_{\min})\right) + \lambda(4 + \log|\mathcal{Y}|) \\
&\le \frac{5}{2}\left(1 + \beta(1+1/\eta)\log(1/\pi^{\mathrm{ref}}_{\min}) + \frac{1 + 2\sqrt{\varepsilon_{\mathrm{PP}}}}{\eta}\right) + \beta(1+1/\eta)(4+\log|\mathcal{Y}|) \\
&\le \frac{5}{2}\left(1 + \beta(1+1/\eta)\log(1/\pi^{\mathrm{ref}}_{\min}) + 3/\eta\right) + \beta(1+1/\eta)(4+\log|\mathcal{Y}|) \triangleq L_{\beta,\eta}.
\end{aligned}$$

After that bound, we can prove the induction step to show that $\|\nabla V_k(\pi_{k+1},\pi_{k+1})\|^2_{\mathrm{sp},\rho} \le 1$ for all k. In fact, we have

$$\|\nabla V_k(\pi_{k+1},\pi_{k+1})\|^2_{\mathrm{sp},\rho} \le 2C_{\mathrm{gc}}L_{\beta,\eta}\cdot\varepsilon_{\mathrm{in}} \le 2\cdot\frac{1}{(\tau_0\cdot\pi^{\mathrm{ref}}_{\min})^2}\cdot L_{\beta,\eta}\cdot\varepsilon_{\mathrm{in}} = \varepsilon_{\mathrm{PP}} \le 1,$$

where the last inequality holds for our choice of $\varepsilon_{\mathrm{in}}$ small enough.

Next, for bound on $m_{\widetilde{\pi},\lambda}$ we recall that under our choice of $\nu = \pi^{\mathrm{ref}}$

$$m_{\widetilde{\pi},\lambda} = \lambda \mathrm{e}^{-2/\lambda} \cdot c_{\pi^{\mathrm{ref}}}^2 \,,$$

for $c_{\pi^{\mathrm{ref}}} \triangleq \exp\bigl(\min\bigl\{-2\|\log\widetilde{\pi}_k - \log\pi^{\mathrm{ref}}\|_{\mathrm{sp}}, \log(\pi^{\mathrm{ref}}_{\min}/(1+\pi^{\mathrm{ref}}_{\min}))\bigr\}\bigr) \cdot \pi^{\mathrm{ref}}_{\min}$, where we repeat our calculations for $\|\log\widetilde{\pi}_k - \log\pi^{\mathrm{ref}}\|_{\mathrm{sp}}$ to obtain

$$c_{\pi^{\mathrm{ref}}} \geq \exp\left(\min\left\{-\frac{3}{\beta(1+\eta)}, \log(\pi^{\mathrm{ref}}_{\min}/(1+\pi^{\mathrm{ref}}_{\min}))\right\}\right) \cdot \pi^{\mathrm{ref}}_{\min} > 0\,,$$

and $m_{\widetilde{\pi},\lambda} \geq m_{\beta,\eta} > 0$ uniformly over k. Next, the bound on $\sigma^2_{\widetilde{\pi},\lambda}$ and $v^2_{\widetilde{\pi},\lambda}$ again follows automatically from a bound on $\widetilde{\pi}_{k,\min}$, which depends on the following quantity

$$\begin{aligned}
D^2_{\widetilde{\pi},\lambda}(\sqrt{\varepsilon_{\mathrm{in}}}) &= 2\left(1 + 4\lambda\log\left(\frac{2\sqrt{2}\cdot\lambda(1+\widetilde{\pi}_{\min})}{\widetilde{\pi}_{\min}}\right) + 2\lambda\log(1/\varepsilon_{\mathrm{in}})\right)^2 \\
&\leq 2\left(1 + 4\beta(1+\eta^{-1})\log\left(\frac{4\sqrt{2}\cdot\beta(1+1/\eta)}{\pi^{\mathrm{ref}}_{\min}}\right) + \frac{12}{\eta} + 2\beta(1+\eta^{-1})\log\left(\frac{1}{\varepsilon_{\mathrm{in}}}\right)\right)^2,
\end{aligned}$$

implying that the indicated bound on $\sigma^2_{\widetilde{\pi},\lambda}$ and $v^2_{\widetilde{\pi},\lambda}$ hold uniformly over k. To conclude the second statement of the lemma, we need to choose $\eta > 0$ such that $\beta(1+1/\eta)\cdot m_{\beta,\eta} = 1$. In fact, such η exists since the function $\eta \mapsto \beta(1+1/\eta)\cdot m_{\beta,\eta}$ is continuous and decreasing on $\mathbb{R}_{>0}$, with limits $\lim_{\eta\to 0}\beta(1+1/\eta)\cdot m_{\beta,\eta} = +\infty$ and

$$\lim_{\eta\to+\infty}\beta(1+1/\eta)\cdot m_{\beta,\eta} = \beta\exp(-2/\beta)\cdot\exp\left(\min\left\{-6/\beta, 2\log\left(\frac{\pi^{\mathrm{ref}}_{\min}}{1+\pi^{\mathrm{ref}}_{\min}}\right)\right\}\right)\cdot(\pi^{\mathrm{ref}}_{\min})^2 < 1$$

under the choice $\beta \leq 1$.

Finally, the third statement follows from Corollary 6 noticing that our constants depend on $\varepsilon_{\mathrm{in}}$ only logarithmically and polynomially on $\pi^{\mathrm{ref}}_{\min}, \beta, \eta$. □

D.5 Verification for Compatible Fisher Non-Degenerate Parametrization

In this section, we show that along the iterations of Algorithm 1, Assumptions 7–10 hold uniformly over k. In particular, Assumptions 8, 9, and 10 do not depend on the target policy, thus they cannot be disturbed by iterates.

Lemma 31. *Assume that $\|\nabla V_k(\pi_{k+1},\pi_{k+1})\|^2_{\mathrm{sp},\rho} \leq \varepsilon_{\mathrm{PP}}$ for all $k < K$. Then Assumption 7-(i) holds for the next step K with*

$$V_{\widetilde{\pi}_K} \leq 2\mathbb{E}_{x\sim\rho}\bigl[\log^2(1/\pi^{\mathrm{ref}}_{\min}(x))\bigr] + \frac{2(1+2\sqrt{\varepsilon_{\mathrm{PP}}})^2}{\beta^2(1+\eta)^2}\,.$$

Proof. First, we apply the second inequality of Lemma 28 with $\pi = \widetilde{\pi}_k$ and $\mu = \pi^{\mathrm{ref}}$:

$$\mathbb{E}_{x\sim\rho}\left[\log^2\frac{1}{\widetilde{\pi}_{k,\min}(x)}\right] \leq 2\mathbb{E}_{x\sim\rho}\left[\log^2\frac{1}{\pi^{\mathrm{ref}}_{\min}(x)}\right] + 8\|\log\widetilde{\pi}_k - \log\pi^{\mathrm{ref}}\|^2_{\mathrm{sp},\rho}\,.$$

For the second term we use the property of geometric mixture policies

$$\begin{aligned}
\|\log\widetilde{\pi}_k - \log\pi^{\mathrm{ref}}\|_{\mathrm{sp},\rho} &= \|1/(1+\eta)\log\pi_k + \eta/(1+\eta)\log\pi^{\mathrm{ref}} - \log\pi^{\mathrm{ref}}\|_{\mathrm{sp},\rho} \\
&\leq \frac{1}{1+\eta}\|\log\pi_k - \log\pi^{\mathrm{ref}}\|_{\mathrm{sp},\rho}\,.
\end{aligned}$$

Next, we apply Lemma 27 and achieve

$$\|\log\widetilde{\pi}_k - \log\pi^{\mathrm{ref}}\|_{\mathrm{sp},\rho} \leq \frac{1}{1+\eta}\cdot\frac{1/2+\sqrt{\varepsilon_{\mathrm{PP}}}}{\beta} \leq \frac{1+2\sqrt{\varepsilon_{\mathrm{PP}}}}{2\beta(1+\eta)}\,.$$

Taking a square and plugging back into the previous inequality concludes the proof. □

In particular, a simple induction shows that Assumption 7-(i) is satisfied for all $k \in \mathbb{N}$ simultaneously. Thus, we have the following final corollary.

Corollary 8 (Convergence guarantees for Fisher compatible parameterization)**.** *Let $\theta \mapsto \pi_\theta$ be a parametrization and $(\mathcal{T}_k)_{k\geq 0}$ a sequence of improvement operators that satisfies Assumption 8–10 for all $k \in \mathbb{N}$, the reference policy satisfies $V_{\pi^{\mathrm{ref}}} \triangleq \mathbb{E}_{x\sim\rho}[\log^2(1/\pi^{\mathrm{ref}}_{\min}(x))] < \infty$ and let $\beta \leq 1$.*

Then, fix $\varepsilon_{\mathrm{in}} > 0$ as a desired accuracy for approximate solving of the inner problem. Then the following statements hold.

(i) Assumption 11 holds with $C_{\mathrm{gc}} = 2M_g^2/m_F^2$ and $\varepsilon_{\mathrm{gc}} = 2\varepsilon_{\mathrm{bias}}$;

(ii) There exists a truncation threshold $M_k > 0$ such that Assumption 12 holds with

$$
\begin{aligned}
G = M_g, \qquad m_{\beta,\eta} = \frac{\beta(1+1/\eta)m_F^2}{2M_g^2}, \qquad \varepsilon_{\mathrm{PL}} = \frac{\varepsilon_{\mathrm{bias}} \cdot m_F^2}{M_g^2}\,,\\
L_{\beta,\eta} = \left(\tfrac{1}{2} + \tfrac{1}{\eta} \cdot \sqrt{2\beta^2(1+\eta)^2 + 2\beta^2(1+\eta)^2 V_{\pi^{\mathrm{ref}}} + 2(1+2\sqrt{\varepsilon_{\mathrm{PP}}})^2}\right)(M_g^2 + M_h)\\
+ \beta(1+1/\eta)M_g^2\,,\\
\varepsilon_{\mathrm{grad}} = \sqrt{\varepsilon_{\mathrm{in}}}\,, \qquad \sigma^2_{\beta,\eta} = D^2_{\beta,\eta}, \qquad v^2_{\beta,\eta} = 6D^2_{\beta,\eta}\,,\\
D^2_{\beta,\eta} \triangleq 2M_g^2\left(1 + \frac{128M_g^2[\beta^2(1+\eta)^2(1+4V_{\pi^{\mathrm{ref}}}^2) + 4(1+2\sqrt{\varepsilon_{\mathrm{PP}}})^4]}{\eta^4 \cdot \varepsilon_{\mathrm{in}}}\right),
\end{aligned}
$$

where $\varepsilon_{\mathrm{PP}} = \mathcal{O}(M_g^2/m_F^2 \cdot \varepsilon_{\mathrm{in}} + \varepsilon_{\mathrm{bias}})$ and a PP learning rate $\eta = \beta \cdot \frac{m_F}{\sqrt{2\cdot M_g^2 - \beta m_F}}$.

(iii) Algorithm 1 outputs $\mathcal{O}(\varepsilon_{\mathrm{in}} + \varepsilon_{\mathrm{bias}})$-optimal regularized policy after $K = \widetilde{\mathcal{O}}(1)$ outer iterations, $T_k = \widetilde{\mathcal{O}}(1/\varepsilon^2_{\mathrm{in}})$ inner iterations, using in total $\widetilde{\mathcal{O}}(1/\varepsilon^4_{\mathrm{in}})$ samples, ignoring constants depending on $M_g, M_h, m_F, \beta, V_{\pi^{\mathrm{ref}}}$ and logarithmic terms in $\varepsilon_{\mathrm{in}}, \varepsilon_{\mathrm{bias}}$ and the confidence level δ.

Proof. The statement (i) automatically follows from Lemma 20. The statement (ii) follows from Lemma 31 combined with Proposition 6 applied for each $k \in \mathbb{N}$ as well as a choice of $\varepsilon_{\mathrm{grad}} = \sqrt{\varepsilon_{\mathrm{in}}}$ and manipulations to simplify the expression for $D^2_{\beta,\eta}$:

$$
\begin{aligned}
D^2_{\widetilde{\pi}_k,\lambda}(\sqrt{\varepsilon_{\mathrm{in}}}) &= M_g^2\left(1 + \frac{8\lambda^2 M_g(V_{\widetilde{\pi}_k} + 1)}{\sqrt{\varepsilon_{\mathrm{in}}}}\right)^2 \leq 2M_g^2\left(1 + \frac{(8\lambda^2 M_g(V_{\widetilde{\pi}_k} + 1))^2}{\varepsilon_{\mathrm{in}}}\right)\\
&\leq 2M_g^2\left(1 + 64\beta^4(1+\eta)^4 \frac{M_g^2(1 + 2V_{\pi^{\mathrm{ref}}} + 2(1+2\sqrt{\varepsilon_{\mathrm{PP}}})^2/(\beta^2(1+\eta)^2))^2}{\varepsilon_{\mathrm{in}} \cdot \eta^4}\right)\\
&\leq 2M_g^2\left(1 + \frac{128M_g^2(\beta^2(1+\eta)^2(1+4V_{\pi^{\mathrm{ref}}}^2) + 4(1+2\sqrt{\varepsilon_{\mathrm{PP}}})^4)}{\eta^4 \cdot \varepsilon_{\mathrm{in}}}\right) \triangleq D^2_{\beta,\eta}\,,
\end{aligned}
$$

where we used Lemma 31 and an inequality $(a+b)^2 \leq 2a^2 + 2b^2$. Finally, we choose $\eta = \beta \cdot \frac{m_F}{\sqrt{2\cdot M_g^2 - \beta m_F}}$ to guarantee that $\beta(1+1/\eta) \cdot m_{\beta,\eta} = 1$.

For the statement (iii), we apply Corollary 6 with the constants from (ii). In particular, we choose $\varepsilon_{\mathrm{PP}} = \mathcal{O}(M_g^2/m_F^2(\varepsilon_{\mathrm{in}} + \varepsilon_{\mathrm{bias}}))$ to guarantee that the final policy is $\mathcal{O}(\varepsilon_{\mathrm{in}} + \varepsilon_{\mathrm{bias}})$-VNW. The number of outer iterations K follows from Corollary 6 and scales as $\widetilde{\mathcal{O}}(1)$. The number of inner iterations T_k follows from Corollary 6 and scales as $\widetilde{\mathcal{O}}(1/\varepsilon^2_{\mathrm{in}})$, where additional $1/\varepsilon_{\mathrm{in}}$ factor comes from a bound on $v^2_{\beta,\eta} \asymp 1/\varepsilon_{\mathrm{in}}$. Finally, the total number of samples follows from $\sum_{k=1}^K \sum_{t=1}^{T_k} B_{k,t} = \widetilde{\mathcal{O}}(1/\varepsilon^4_{\mathrm{in}})$. □

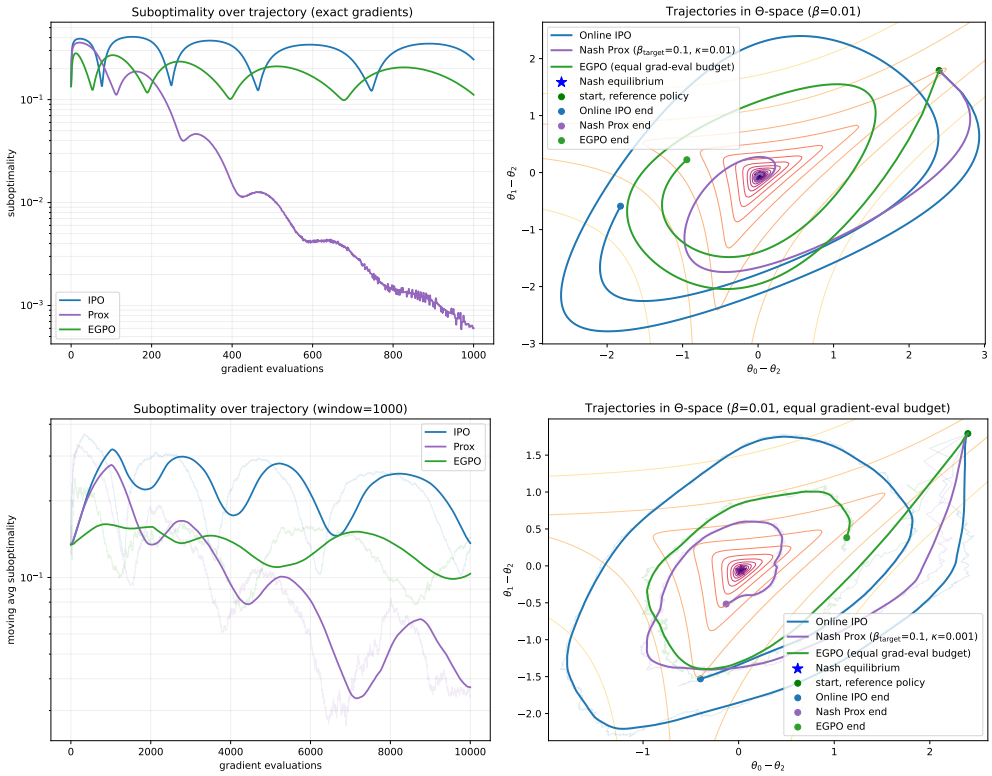


Figure 2: *Optimization trajectories for the Rock–Paper–Scissors example.* **Top row**: exact-gradient updates. **Bottom row**: stochastic-gradient (SGD) updates. **Left**: suboptimality versus gradient evaluations; for SGD (bottom) we plot an exponential moving average (EMA) to smooth single-trajectory noise. **Right**: trajectories in a 2D projection of the 3D parameter space (Θ-space).

E Detailed Experiment Description

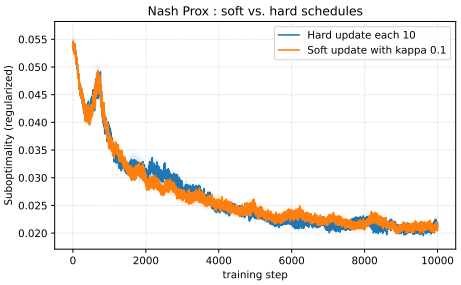
E.1 Rock-Paper-Scissors

In this section, we provide an additional experiment on a very simple 3-dimensional game of Rock-Paper-Scissors. It is a context-free game defined by the following matrix

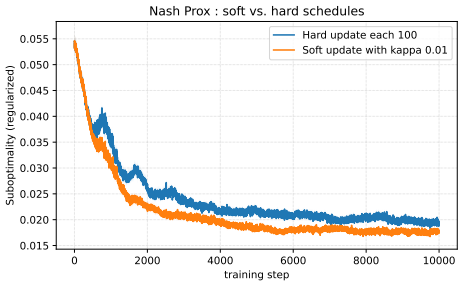
$$\mathbf{P} = \begin{bmatrix} 0.5 & 1 & 0 \\ 0 & 0.5 & 1 \\ 1 & 0 & 0.5 \end{bmatrix},$$

a reference policy $\pi^{\text{ref}} = (11/18, 1/3, 1/18)$, and $\beta = 0.01$. With this environment, we implemented in JAX (Bradbury et al., 2018) an exact and stochastic versions of Online IPO (Calandriello et al., 2024), EGPO (Zhou et al., 2025), and `Nash Prox`, using a learning rate $\alpha_t = t^{-1/2}$ for an exact version and $\alpha_t = 0.2 \times t^{-1/2}$ for a stochastic one, where t is an iteration number, additionally multiplied the IPO-style losses by an effective regularization to guarantee the same scaling of updates. The value of κ is normalized to make 10 total soft updates of the target policy at the training. The results are presented on Figure 2.

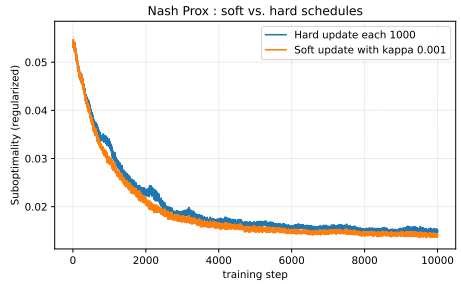
Overall, we observe that the trajectories of `Nash Prox` and EGPO are both present more stabilized behavior compared to Online IPO, but they are stabilized differently: two-step stabilization of EGPO turns out to be an effective measure, especially in the beginning of trajectory, but later the soft-anchoring of `Nash Prox` starts behaving better empirically. Additionally, we observe two facts about the geometry of the problem: (1) suboptimality landscape exhibits non-convex behavior, although it admits a gradient dominance geometry, and (2) all the methods exhibits non-monotonic behavior in suboptimality. In fact, (2) exactly prevents us to provide a *last-iterate* convergence theory for Online IPO for small values of β.



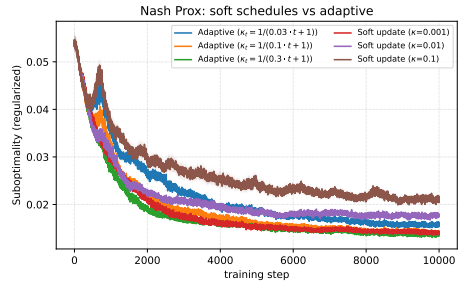
(a) Hard updates vs. soft updates, period 10;



(b) Hard updates vs. soft updates, period 100;



(c) Hard updates vs. soft updates, period 1000;



(d) Soft updates vs. adaptive schedule.

Figure 3: Comparison between different target policy update schedules for `Nash Prox`. Suboptimality is averaged over 25 random seeds; shaded regions indicate one standard deviation.

E.2 Matrix Games

Experiment setup. In our experiments, we fixed $r = 2, Y = 100, \beta = 0.01$ and a reference policy to be a uniform distribution $\pi^{\text{ref}}(y|x) = 1/Y$. The matrices U and V are generated as random Gaussian matrices. To parameterize the space of policies, we use a 3-layer MLP with a ReLU action function and 128 hidden units, taking a flattened matrix $\Theta_x \in \mathbb{R}^{2\times 2}$ as input and outputting logits over possible actions. We use the Adam optimizer (Kingma & Ba, 2015), and for all the baselines we perform a grid search over the learning rate using the grid $\{3\times10^{-3}, 10^{-3}, 3\times10^{-4}, 10^{-4}, 3\times10^{-5}\}$. For each batch, we sample 128 random games. For Nash Prox, we utilize $\beta_{\text{target}} = 10 \times \beta$ for all experiments. All experiments implemented using JAX (Bradbury et al., 2018), runtime for one configuration running in parallel for all 25 seeds is less than 5 minutes for all the methods.

Soft vs. hard updates. As a first additional experiment, we compare the influence of soft and hard updates for the target network. We fixed the learning rate to 3×10^{-4} and varied the update period before the update in hard and soft senses; see Figures 3a,3b,3c for the results. Overall, these results show the benefits of soft updates over hard ones in terms of suboptimality.

Adaptive choice of κ. Next, we verify the adaptive schedule $\kappa_t = 1/(0.3 \cdot t + 1)$, which changes over the optimization procedure; the results are presented in Figure 3d. In particular, we observe that the adaptive schedule performs on par with the best choice of standard soft updates; however, it is less sensitive to the choice of optimal value of the multiplier than the choice of optimal value of κ for usual soft updates.

E.3 LLM Alignment

In this section, we provide implementation and details as well as details on hyperparameter selection.

E.3.1 Loss implementation

We use a library TRL (von Werra et al., 2020) as a base for our implementation, and we use vLLM (Kwon et al., 2023) for efficient generation. Following the discussion in Section 5.4, we use

the following loss function to achieve the correct gradients using automatic differentiation in PyTorch (Paszke et al., 2019)

$$\widetilde{\mathcal{L}}_{\texttt{Nash Prox}}(\theta;\theta^{\text{target}}) \triangleq \frac{1}{B}\sum_{i=1}^{B}\left(\ell_\theta(x_i,y_i)-\ell_\theta(x_i,y_i')-\frac{\mathcal{P}(y_i\succ y_i'|x_i)-1/2}{\beta+\beta_{\text{target}}}\right)^2$$

$$\ell_\theta(x,y) \triangleq \frac{\beta}{\beta+\beta_{\text{target}}}\log\left(\frac{\pi_\theta(y|x)}{\pi^{\text{ref}}(y|x)}\right)+\frac{\beta_{\text{target}}}{\beta+\beta_{\text{target}}}\log\left(\frac{\pi_\theta(y|x)}{\pi_{\theta^{\text{target}}}(y|x)}\right),$$

where $\{x_i\}_{i\in[B]}$ are samples from the prompt dataset, $\{(y_i,y_i')\}_{i\in[B]}\}$ are generated from a policy π_{θ_t} using a temperature sampling with a temperature 1, and $\texttt{SG}$ is a stop-gradient operations. $\log\pi_\theta(y|x)$ is computed in an auto-regressive manner as $\log\pi_\theta(y|x)=\sum_{i=1}^{|y|}\log\pi_\theta(y_i|x,y_{<i})=\sum_{i=1}^{|y|}\texttt{logit}_\theta(y_i|x,y_{<i})-\texttt{LogSumExp}(\texttt{logit}_\theta(\cdot|x,y_{<i}))$, where $\texttt{logit}_\theta$ are raw logits outputted by the model with parameters θ.

E.3.2 Experiment description

We start our experiments from the Google Gemma-3-4B[1] (Team et al., 2025) pretrained checkpoint.

Supervised fine-tuning (SFT). For SFT, we use the `RLHFlow/RLHFlow-SFT-Dataset-ver2` dataset (RLHFlow Team, 2024c). This dataset, structured as conversations, is processed using a chat template following the Gemma3 format (`<bos><start_of_turn>role\ncontent<end_of_turn>\n...`), where the template maps the `assistant` role to `model`. System messages are dropped from the input, and training is performed on the train split. The dataset samples are tokenized, with a maximum sequence length of 8,192 tokens. We use sample packing to efficiently train on long sequences and pad sequences to the maximum length. Following standard SFT practice, the loss is computed only on the model's output tokens (the assistant's turns), not on the input prompts.

The model was fully fine-tuned (no LoRA PEFT adapter was used). Training was conducted for 2 epochs. Optimization was performed using a fused version of AdamW (Loshchilov & Hutter, 2019) 32-bit optimizer with a learning rate of 1.5×10^{-5}. A cosine learning rate schedule was applied with a warmup ratio of 0.05 of the total training steps. We used a micro batch size of 1 sequence per device and accumulated gradients over 16 steps, resulting in an effective batch size of 16 sequences per device. On the 8 A100 GPUs, we thus had an effective batch size of 128. Gradient clipping was applied with a maximum norm of 1.0, and no weight decay was used.

For improved memory efficiency and speed, we enabled gradient checkpointing and leveraged Flash Attention (Dao, 2024). Training utilized BFloat16 (BF16) and TF32 precision where supported.

Nash Learning from Human Feedback All subsequent NLHF experiments started from the SFT checkpoint described above. This SFT model also served as the initial policy and the reference policy (π^{ref}). During this phase, we used LoRA adapters Hu et al. (2022) with rank $r=16$ and $\alpha=32$ for all methods.

Datasets. For generating responses during NLHF training and for final evaluation, we used a subset of prompts from the `RLHFlow/prompt-collection-v0.1` dataset (RLHFlow Team, 2024b). We first filtered this dataset to have only prompt conversations of length less than 512 tokens, and used 5% of prompts as a separate validation set. Overall, we have $\approx$ 80,200 train prompts and $\approx$ 4,220 validation problems. For further details on the original data mixtures within `RLHFlow/prompt-collection-v0.1` and their licenses, we refer to (Dong et al., 2024).

Preference model. The pairwise preference model, used to provide comparison signals, was a Gemma2-2B model (Rivière et al., 2024). This model was trained on the `RLHFlow/pair_preference_model_dataset` dataset (RLHFlow Team, 2024a), with its training methodology detailed in (Liu et al., 2025). We employed a separate, more capable judge for the final evaluation of model performance: a Gemma3-27-IT model prompted to decide which completion better follows the instructions. On this stage, we perform two episodes of judgment per prompt in two different orders: (`prompt`, `completion A`, `completion B`) and (`prompt`, `completion B`,

[1] Published under Gemma license.

Table 3: Hyperparameter settings for the evaluated algorithms. The hyperparameters with the best performance in comparisons against the SFT checkpoint are presented.

Algorithm	Learning Rate	β	Algorithm-Specific Parameters
Online DPO	3×10^{-5}	1×10^{-2}	N/A
Online IPO	1×10^{-4}	1×10^{-2}	N/A
`Nash Prox`	3×10^{-5}	1×10^{-3}	$\kappa_t = 1/(0.1 \times t + 1)$, $\beta_{\text{target}} = 10 \times \beta$

`completion A`). If results were inconsistent (e.g., in both cases the judge selected different completions depending on their order), we consider such judgments as inconsistent and do not include them in the win-rate computation. On average, around 40% of the comparisons were inconsistent.

The primary evaluation metric was side-by-side pairwise win rate, and for hyperparameter selection in each algorithm, we used win rate against the SFT reference policy as a proxy metric. Confidence intervals were computed as 99.9%-confidence intervals using the first term of the empirical Bernstein inequality for Bernoulli random variables: for an estimate $\hat{p}$, we compute the intervals as $\sqrt{2\hat{p}(1-\hat{p}) \cdot \log(2/10^{-3}) \cdot 1/N}$, where N is a number of *consistent* judgments.

Training Configuration and Hyperparameter Tuning. For all NLHF experiments, we used the AdamW optimizer (Loshchilov & Hutter, 2019). The learning rate schedule featured a 0.1 warmup period (as a fraction of total training steps) followed by a linear decay. The effective global batch size was 32 prompts, following the practice of small batch sizes as indicated by Schulman & Lab (2025). We used per-device micro-batch size of 8 prompts, 2 GPUs A100, and 2 gradient accumulation steps (8 prompts/GPU $\times$ 2 GPUs $\times$ 2 grad_accum_steps). Training was conducted for 1 epoch over the $N_{\text{train}} = 80,200$ prompts, corresponding to approximately 2505 update steps. We employed gradient clipping with a maximum norm of 1.0.

For each algorithm, we perform a grid search over its key hyperparameters: learning rate over $\text{lr} \in \{3 \times 10^{-4}, 10^4, 3 \times 10^{-5}\}$, regularization parameter $\beta \in \{10^{-3}, 10^{-2}, 10^{-1}\}$, soft update schedule of form $\kappa_t = 1/(ct+1)$ for $c \in \{0.3, 0.1, 0.03\}$. For `Nash Prox`, we use always $\beta_{\text{target}} = 10 \times \beta$.

For each algorithm, we selected the best-performing hyperparameter configurations based on the win rate against the SFT reference policy, evaluated using the Gemma3-27B-IT judge. These selected checkpoints were then compared side-by-side in the final evaluation using the same judge model. The final reported results (see Table 3) represent the performance of the best configuration found through this process.

To manage memory and improve throughput during all NLHF training phases, we utilized BFloat16 (BF16) mixed-precision, gradient checkpointing, and PyTorch DDP (Li et al., 2020).

Computational Resources and Runtimes. All NLHF experiments were conducted on 2 NVIDIA A100 (80GB) GPUs, with one additional GPU used to host a pairwise reward model. A full training cycle for Nash Prox and baselines requires approximately 11 hours per algorithm.

Generation Parameters. During response generation, both for collecting experiences within NLHF algorithms (e.g., generating samples per prompt) and for final evaluation on the test set, we used temperature sampling with a temperature of $\tau = 1.0$. The maximum generation length was capped at 256 tokens.

Limitations. We notice that using the target network increases the memory footprint of the model, which is a limitation of our method. However, this increase is marginal since we do not need to backpropagate through the target model, and in the LoRA setting this memory footprint is negligible. Additionally, it does not significantly increase the running time of the algorithm and is straightforward to implement.