EXPERT-REGULARIZED EQUILIBRIUM COMPUTATION IN TWO-PLAYER ZERO-SUM CONVEX MARKOV GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate the optimization landscape of Markov game equilibria that are anchored on (expert) reference policies. The technique of anchoring, coupled with self-play training, has achieved super-human performance in several games. Despite its strong empirical success, its formal analysis has lagged behind and has so far been restricted to stateless matrix games and/or direct policy parametrizations. We show that any two-player zero-sum reference-regularized Markov game can naturally be placed within the framework of a convex Markov game (cMG). This introduces a novel instance of "imitation" within the cMG framework—complementing previously recognized aspects such as behavioral imitation, creativity, safety, and equilibrium fairness. Framing the setting in this way, we establish two complementary convergence results. First, in the unconstrained case, we prove that simultaneous gradient descent-ascent converges locally to a (regularized) Nash equilibrium, offering theoretical support for the simultaneous-update schemes that dominate practice. Such guarantees, however, cannot in general be promoted to global ones: simultaneous GDA is known to fail globally even under a two-sided Polyak-Łojasiewicz condition. To obtain global convergence, we turn to softmax policies under alternating updates. We prove that saddle-point computation satisfies a two-sided (proximal) PŁ condition for softmax parametrizations and design an alternating policy gradient scheme that converges to a (regularized) Nash equilibrium with an explicit polynomial rate and sample complexity. In doing so, we obtain the first convergence proof of softmax policy gradient to a Nash equilibrium in two-player zero-sum convex Markov games, offering a crisp theoretical justification for the use of softmax policy gradient updates in numerous practical applications.

1 INTRODUCTION

Self-play training (Samuel, 1959; Tesauro et al., 1995) has produced some of the most striking successes in artificial intelligence, achieving human or super-human performance in complex strategic environments such as Chess, Go, Stratego, and Poker (Silver et al., 2018; 2017; Perolat et al., 2022; Brown & Sandholm, 2019). Its applications also extend to social and cooperative settings such as Diplomacy and Hanabi (FAIR et al., 2022; Bard et al., 2020). In self-play, an agent learns by facing versions of itself, continually adapting to the strategies generated during training. Yet the resulting conventions can be difficult for people to interpret or coordinate with (Lerer & Peysakhovich, 2019; Jacob et al., 2022). This is not merely an interpretability concern: when an agent must collaborate with human players, success depends on compatibility with human behavior as well as strategic strength.

Expert anchoring addresses this tension by coupling strategic improvement with a reference policy learned from expert demonstrations (Bakhtin et al., 2022; Jacob et al., 2022). Rather than learning solely against other self-play agents, each player optimizes its performance while a regularizer discourages departure from the expert-trained reference. This recipe contributed to human-compatible play in Diplomacy and Hanabi, connecting competitive learning with the demands of collaboration. The same principle appears in language-model post-training: policy regularization limits drift from a base model while learning from verifiable rewards or human preferences, including in multi-turn settings (Rafailov et al., 2023; Munos et al., 2024; Wu et al., 2025). Preferences need not admit a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

single scalar ranking; when they are intransitive, a game-theoretic formulation captures comparisons that reward maximization alone cannot express (Munos et al., 2024). These applications motivate understanding the dynamics of reference-regularized learning.

These successes share a common learning rule: each agent repeatedly takes gradient steps toward a *regularized best response* to its opponent's current policy, while remaining anchored to its reference. Two questions require separate answers. *First, do these dynamics converge at all?* The policy-space objective is nonconvex-nonconcave, and first-order game dynamics can cycle or diverge without additional structure (Mertikopoulos et al., 2018; Hsieh et al., 2021). *Second, if they converge, do the limit policies have game-theoretic meaning?* In two-player zero-sum games, a regularized Nash equilibrium certifies that neither player can improve its regularized objective by deviating alone. In cooperative or general-sum settings, convergence need not imply joint optimality, and the appropriate solution concept requires further care. We focus on the zero-sum case as a foundation for understanding the broader applications.

Scope of the guarantees. Our global guarantee concerns finite, two-player zero-sum cMGs with smooth occupancy utilities, full-support references, and constrained alternating softmax updates. The local simultaneous-GDA theorem assumes smoothness, two-sided PŁ, and bounded marginal optimizers. Neither result covers arbitrary neural policies, LLM systems, or general-sum collaboration.

Beyond matrix games and direct policies (Jacob et al., 2022; Bakhtin et al., 2022), we ask: *what geometry makes expert-regularized learning converge in stateful games, and to which equilibrium?*

1.1 Contributions

We make contributions to both min–max optimization and (convex) Markov games:

- *Local simultaneous GDA.* For any smooth min-max objective satisfying local two-sided PŁ and the stated boundedness condition, simultaneous GDA converges locally and linearly to a saddle. This addresses the simultaneous-update schemes used in practice, while a recent counterexample (Mourrat, 2026) shows why two-sided PŁ alone does not support a global claim.
- *Expert anchoring as a cMG.* We place reference-KL-regularized Markov games within the cMG framework (Gemp et al., 2024). This identification is useful because discounted entropy gives curvature in each player's occupancy measure, whereas the regularizer in some prior cMG analyses does not give the same symmetric geometry (Kalogiannis et al., 2025). It also connects standard RL regularization to bidilated regularization in extensive-form games (Perolat et al., 2021; Kalogiannis & Farina, 2025).
- *Global softmax convergence.* We establish a global last-iterate guarantee for alternating softmax policy gradient in cMGs, directly in policy parameters and without support or isolation assumptions on the target equilibrium (Zeng et al., 2022; Zhang et al., 2022). The proof controls the rank-deficient softmax Jacobian on a constrained logit set and combines this geometry with sampled policy gradients. For fixed τ, the method uses $\widetilde{O}(\tau^{-3}\epsilon^{-6})$ iterations and $\widetilde{O}(\tau^{-6}\epsilon^{-13})$ trajectories.

New technical ingredients. Our central contribution is a quantitative link between expert-KL regularization and the geometry of softmax policy parameters. A Pinsker argument and the occupancy–policy map give a uniform curvature modulus for discounted negative entropy. Reference KL adds a linear occupancy term, preserving this curvature. Crucially, we show that the *opponent's* KL term is also linear in one's own occupancy: regularizing both players therefore preserves the two-sided geometry needed for equilibrium computation, rather than establishing curvature for only one player's problem.

The second step transfers this occupancy-space curvature to the nonconvex logit objective through the occupancy inverse and a nonsmooth chain rule for feasible directions. Softmax has an all-ones null direction; we control its remaining singular values uniformly on the constrained logit set to obtain an explicit two-sided pPŁ modulus and polynomial rates. Our local simultaneous-GDA result instead uses a stability argument near a saddle. We obtain the global guarantee using the alternating-pPŁ theorem of Kalogiannis et al. (2025) and standard finite-horizon REINFORCE bias and variance bounds.

Reference	Setting	Algorithm	Policy Param.	Assumption	Convergence
Daskalakis et al. (2020)	NFG, MG	Simultaneous PG	Direct	No	best-iterate
Cen et al. (2021; 2022)	NFG, MG	Simultaneous PG	Direct	No	last-iterate
Jacob et al. (2022)	Anchored NFG	Regularized PG	Direct	No	last-iterate
Zeng et al. (2022)	Regularized MG	Alternating PG	Softmax	Full-support NE	last-iterate
Perolat et al. (2021; 2022)	EFG	FTRL dynamics	Direct	Monotonicity	asymptotic
Sokota et al. (2022)	EFG	MMD	Direct	Monotonicity	last-iterate
Liu et al. (2024)	EFG	Policy MD	Direct	EFG structure	best-iterate
Abe et al. (2023; 2024)	Monotone games	MD variant	Direct	Monotonicity	last-iterate
Meng et al. (2025)	EFG	RM/CFR+ variant	Direct	EFG structure	last-iterate
Ours	**cMG**	**Alt./nested regularized PG**	**Softmax**	**No target-NE support/isolation assumption**	**polynomial last-iterate**

NFG = normal-form game; MG = Markov game; cMG = convex Markov game; EFG = extensive-form game; PG = policy gradient; MD = mirror descent; MMD = magnetic mirror descent.

Table 1: Comparison with related convergence guarantees. EFG results are included to clarify scope: our guarantees are for softmax policy-gradient methods in convex Markov games, not CFR/RM algorithms for EFGs. For the monotone-games / EFG MMD row, last-iterate rates are for monotone games; the EFG guarantee is not a Markov-game result.

Nonlinear occupancy utilities break Bellman consistency (Zhang et al., 2020); pPŁ instead links stationarity to global optimality in each player's constrained problem (Karimi et al., 2016; Fatkhullin & He, 2024).

1.2 Related Work

Nonconvex min-max guarantees commonly rely on gradient domination or Minty-type conditions (Nouiehed et al., 2019; Daskalakis et al., 2020; Mertikopoulos et al., 2018; Diakonikolas et al., 2021); two-sided PŁ alone does not ensure global simultaneous-GDA convergence (Mourrat, 2026).

General-utility RL covers imitation, exploration, safety, and fairness through occupancy objectives (Hazan et al., 2019; Zhang et al., 2020; Gemp et al., 2024). Policy-gradient analyses mainly address cMDPs (Zhang et al., 2021; Barakat et al., 2023), with some softmax rates depending on nonuniform constants (Mei et al., 2020). Game results use linear rewards and exact gradients (Zeng et al., 2022), or purpose-built regularizers and direct/ϵ-greedy policies (Kalogiannis et al., 2025). Our analysis supplies quantitative two-sided geometry for expert-KL regularization under softmax parametrization. Table 1 compares the closest guarantees.

2 Preliminaries

Convex Markov games generalize cMDPs to multiple agents (Zahavy et al., 2021; Gemp et al., 2024). We introduce the policy parametrization, gradient oracle, and regularity conditions used below.

Notation. $\mu, \nu, d, \chi, \psi, r$ are vectors of appropriate dimension while $\alpha, \delta, \epsilon, L, \ell, L_F, \ell_F, L_U, \ell_U, L_c, \ell_c$, are scalars. Matrices are denoted using boldface capital letters. In general, $\|\cdot\|$ will be the ℓ_2-vector norm for a vector and spectral norm for a matrix.

2.1 (Convex) Markov Games

Definition 1 (Two-player zero-sum cMG)**.** An infinite-horizon zero-sum convex Markov game is a tuple $\Gamma = (\mathcal{S}, \mathcal{A}, \mathcal{B}, \mathbb{P}, F, \gamma, \varrho)$. Player 1 is the minimizer, with policy π_1 over the finite action space $\mathcal{A}$; player 2 is the maximizer, with policy π_2 over $\mathcal{B}$. The game has a finite state space $\mathcal{S}$, initial distribution $\varrho \in \Delta(\mathcal{S})$, transition kernel $\mathbb{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$, and discount factor $\gamma \in [0, 1)$. Its continuous occupancy objective $F : \Delta(\mathcal{S} \times \mathcal{A}) \times \Delta(\mathcal{S} \times \mathcal{B}) \to \mathbb{R}$ is convex in player 1's occupancy μ and concave in player 2's occupancy ν. The corresponding policy

objective is $U(\pi_1, \pi_2) := F(\mu(\pi_1, \pi_2), \nu(\pi_1, \pi_2))$, where the occupancies are defined below. Thus the equilibrium problem is $\min_{\pi_1} \max_{\pi_2} U(\pi_1, \pi_2)$.

Occupancy Measures. We define the state-action occupancy measures:

- For player 1, let $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$ s.t. $\mu_{s,a} := (1-\gamma)\mathbb{E}_{\pi_1,\pi_2}\left[\sum_h^\infty \gamma^h \mathbb{1}\{s^{(h)} = s, a^{(h)} = a\} \,|\, \varrho\right]$
- For player 2, let $\nu \in \Delta(\mathcal{S} \times \mathcal{B})$ s.t. $\nu_{s,b} := (1-\gamma)\mathbb{E}_{\pi_1,\pi_2}\left[\sum_h^\infty \gamma^h \mathbb{1}\{s^{(h)} = s, b^{(h)} = b\} \mid \varrho\right]$

The state-occupancy measure $d \in \Delta(\mathcal{S})$ is the common marginal $d_s := \sum_a \mu_{s,a} = \sum_b \nu_{s,b}$ for all $s \in \mathcal{S}$.

A *convex Markov decision process* (cMDP) is the single-agent specialization of a cMG. Two important special cases are worth noting: a two-player, zero-sum *Markov game* (MG) corresponds to the utility $F(\mu, \nu) := \sum_s \frac{1}{d_s}\langle \mu_s, \mathbf{R}_s \nu_s \rangle$ for some reward matrix $\mathbf{R}$ of appropriate dimension, while a *Markov decision process* (MDP) corresponds to the linear utility $F(\mu) := \langle \mu, r \rangle$ for some reward vector r. The inclusion of MGs is strict: even in one state, KL-regularized objectives have curvature that no fixed reward matrix can represent. Crucially, Gemp et al. (2024) prove that under reasonable and mild conditions NEs exist in cMGs.

We now state two standard assumptions in the cMDP/cMG literature (Zhang et al., 2020; Kalogiannis et al., 2025): the initial state distribution has full support, and the utility function is Lipschitz continuous and smooth with respect to the occupancy measure.

Assumption 1. The initial state distribution satisfies $\varrho(s) \geq \varrho_{\min}$ with $\varrho_{\min} > 0$ for all $s \in \mathcal{S}$.

Assumption 2 (Regularity of Utility)**.** The function $F(\lambda)$ is L_F-Lipschitz continuous and ℓ_F-smooth.

Assumption 1 allows us to conclude the continuity and invertibility of the mapping from policies to occupancy measures: $\pi(a|s) = \mu_{s,a} / \sum_{a'} \mu_{s,a'}$.

Lemma 1 (Continuity of Occupancy Measures)**.** *In a convex Markov game, let $\mu(\pi_1, \pi_2) \in \Delta(\mathcal{S} \times \mathcal{A})$ and $\nu(\pi_1, \pi_2) \in \Delta(\mathcal{S} \times \mathcal{B})$ be the occupancy measures for players 1 and 2 with policies (π_1, π_2), correspondingly. Let μ^{-1} and ν^{-1} be the corresponding occupancy-to-policy mappings such that $\mu^{-1}(\mu(\pi_1, \pi_2)) = \pi_1$ and $\nu^{-1}(\nu(\pi_1, \pi_2)) = \pi_2$. Then:*

- *μ, ν are L_c-Lipschitz continuous with ℓ_c-Lipschitz continuous gradients w.r.t. $(\pi_1, \pi_2) \in \Pi_1 \times \Pi_2$, where $L_c := \frac{\sqrt{|\mathcal{S}|}(|\mathcal{A}|+|\mathcal{B}|)}{(1-\gamma)^2}$ and $\ell_c := \frac{2\gamma\sqrt{|\mathcal{S}|}(|\mathcal{A}|+|\mathcal{B}|)^{3/2}}{(1-\gamma)^3}$.*
- *The inverse mappings μ^{-1}, ν^{-1} are $L_{c^{-1}}$-Lipschitz continuous with $L_{c^{-1}} := \frac{2}{(1-\gamma)\min_s \varrho(s)}$.*

Policy Parametrization. In this work, we consider (constrained) softmax parametrization which takes the form:

$$\pi_1(a|s) = \pi_\chi(a|s) \propto \exp(\chi_{s,a}) \quad \text{and} \quad \pi_2(b|s) = \pi_\psi(b|s) \propto \exp(\psi_{s,b})$$

where $\chi \in X_R$ and $\psi \in Y_R$ are (state, action)-dependent softmax policy parameters (i.e., logits) with

$$X_R := \left\{ \chi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \;\middle|\; \chi_s^\top \mathbf{1} = 0,\ \forall s, |\chi_{s,i} - \chi_{s,j}| \leq 2R,\ \forall s, i, j \right\},$$

where $R > 0$ and Ψ_R is defined analogously over $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{B}|}$. The sets X_R, Y_R are convex and their image, under the softmax map, is convex with lower bounded minimum probabilities. Note the constraint $\chi_s^\top \mathbf{1} = 0$ is without loss of generality as changes to the sum of χ_s do not change the softmax probabilities.

By Assumption 2, as shown in (Zhang et al., 2021), it follows that the utility is Lipshcitz-continous and smooth w.r.t. the parameters of the policies:

Lemma 2. *For a cMG where Assumption 2 holds, the utility function is L_U-Lipschitz continuous and ℓ_U with respect to the parameters of the players' softmax policies, (χ, ψ):*

$$|U(\chi, \psi) - U(\chi', \psi')| \leq L_U \, \|(\chi, \psi) - (\chi', \psi')\| \,,$$
$$\|\nabla U(\chi, \psi) - \nabla U(\chi', \psi')\| \leq \ell_U \, \|(\chi, \psi) - (\chi', \psi')\| \,,$$

for any $(\chi, \psi), (\chi', \psi') \in X_R \times \Psi_R$ with $L_U := L_F L_c, \ell_U := \ell_F L c^2 + L_F \ell_c$.

Solution Concept. The solution concepts we consider are (i) Nash equilibrium, (ii) regularized Nash equilibrium, and (iii) quantal response equilibrium (a special case of the regularized equilibrium). Recall that in a Nash equilibrium, no player has an incentive to unilaterally deviate:

Definition 2 (ϵ-NE)**.** A policy profile parametrized by $(\chi^\star, \psi^\star) \in X \times \Psi$ is an ϵ-approximate Nash equilibrium (ϵ-NE), if for any $\chi, \psi \in X_R \times \Psi_R$ it holds that:

$$U(\chi^\star, \psi) - \epsilon \leq U(\chi^\star, \psi^\star) \leq U(\chi, \psi^\star) + \epsilon. \tag{ϵ-NE}$$

The regularized Nash equilibrium is a generalization of the quantal response equilibrium (McKelvey & Palfrey, 1995). Its definition stipulates a reference policy for each agent and effectively captures the Nash equilibria of expert-anchored games.

Definition 3 (ϵ-RNE & ϵ-QRE)**.** Let $(\pi_{1,\text{ref}}, \pi_{2,\text{ref}})$ be a pair of reference policies and $\tau > 0$. A profile $(\chi^\star, \psi^\star) \in X_R \times \Psi_R$ is an ϵ-regularized Nash equilibrium if Equation (ϵ-NE) holds with

$$U^\tau(\chi, \psi) = U(\chi, \psi) + \tau \mathrm{KL}^\gamma(\pi_1^\chi \| \pi_{1,\text{ref}}) - \tau \mathrm{KL}^\gamma(\pi_2^\psi \| \pi_{2,\text{ref}}).$$

The positive KL term penalizes deviations by minimizing player 1, while the negative KL term penalizes deviations by maximizing player 2. In the case that $(\pi_{1,\text{ref}}, \pi_{2,\text{ref}})$ are the uniform policies, then $(\chi^\star, \psi^\star)$ is called an ϵ-approximate quantal response equilibrium.

2.2 Policy Gradient Estimators

The cornerstone upon which policy gradient methods for MDPs are built is the policy gradient theorem (Williams, 1992). A simple application of the chain rule on top of it yields the policy gradient theorem for cMDPs, $\nabla_\chi U(\chi, \psi) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \frac{\partial F}{\partial \mu_{s,a}(\chi, \psi)} \nabla_\chi \mu_{s,a}(\chi, \psi)$.

To estimate the gradient from trajectories, we treat the utility's derivative w.r.t. the occupancy measure as a pseudo-reward and proceed as if one were to compute the policy gradient of a regular MDP.

Definition 4 (Gradient Estimator)**.**
Let $\xi = (s^{(0)}, a^{(0)}, s^{(1)}, \ldots, s^{(H-1)}, a^{(H-1)})$ be a trajectory of length H sampled under policy parameter profile (χ, ψ) and initial distribution ϱ. Given a pseudo-reward vector $z = \nabla_\mu F_1(\mu(\chi, \psi); y)$, the *gradient estimator* $\hat{g}_\chi(\xi \mid x, z) := \sum_{h=0}^{H-1} \gamma^h \, z(s^{(h)}, a^{(h)}) \left(\sum_{h'=0}^{h} \nabla_\chi \log \pi^\chi(a^{(h')} \mid s^{(h')}) \right)$.
For a batch of M independent trajectories $\{\xi_m\}_{m=1}^M$, the *batch gradient estimator* is given by $\hat{\nabla}_\chi U(\chi, \psi) := \frac{1}{M} \sum_{m=1}^M \hat{g}_\chi(\xi_m \mid x, z)$. The definition for $\hat{\nabla}_\psi U(\chi, \psi)$ is analogous.

It remains to characterize the variance and bias of the REINFORCE gradient estimator which is known to have a bounded variance $\sigma^2 > 0$ and bias $\delta > 0$ Lemma 20.

2.3 Discounted Entropy and Kullback-Leibler Divergence

We define *discounted entropy*, $\mathcal{H}(\cdot)$, as $\mathcal{H}(\pi) := (1-\gamma)\mathbb{E}_{(a_h, s_h) \sim \pi} \left[\sum_h \gamma^{h-1} \log \pi(a_h | s_h) \right]$. While, for a given reference policy π_{ref}, we define the *discounted Kullback-Leibler divergence*: $\mathrm{KL}(\pi \| \pi_{1,\text{ref}}) := (1-\gamma)\mathbb{E}_{(a_h, s_h) \sim \pi} \left[\sum_h \gamma^{h-1} \big(\log \pi(a_h | s_h) - \log \pi_{1,\text{ref}}(a_h | s_h) \big) \right]$.

2.4 Continuity of the Regularized Utility

It is crucial for the convergence of the softmax policy gradient method that the function is smooth. Indeed, we know from the literature that this is the case:

Lemma 3 (Smoothness (Mei et al., 2020, Lemma 14))**.** *The entropic regularizer $\mathcal{H}$ is $\frac{4+\log|\mathcal{A}|}{(1-\gamma)^3}$-smooth with respect to the parameters of the softmax policies.*

Lemma 4 ((Zhang et al., 2021, Lemma 5.3))**.** *Given Assumption 2, the regularized utility function of the cMG is ℓ_U-Lipschitz continuous for softmax policies with $\ell_U := \frac{2\ell_F + 4L_F}{(1-\gamma)^2}$.*

2.5 Gradient Domination Conditions

Here we formally define the proximal Polyak-Łojasiewciz inequality (pPŁ) which is the backbone of our theoretical convergence analysis. In short, it is a variant of the Polyak-Łojasiewicz condition for

functions with constrained domains. In a nutshell, it dictates that stationarity implies optimality, but in a stronger sense than the general weak gradient domination condition of MDPs (Agarwal et al., 2020). Further, it unifies the optimization landscape of the linear quadratic games (Zhang et al., 2019) and regularized (convex) Markov games.

The pPŁ condition for an ℓ-smooth function $f : X \to \mathbb{R}$ is defined using gradient stationarity, $\mathcal{D}_X(\cdot, \ell)$, which in the unconstrained case retrieves the squared norm of the gradient $\|\nabla_\chi f(\chi)\|^2$. This is an intuitive way of understanding why this is a stronger notion of gradient domination.

Definition 5 (pPŁ)**.** Let f be an ℓ-smooth function, $f : X \to \mathbb{R}$ defined over the convex and compact set $X \subseteq \mathbb{R}^d$. Let the forward-backward envelope, $\mathcal{D}_X(\cdot, \ell)$, be defined as: $\mathcal{D}_X(\chi, \ell) := -2\ell \min_{\chi' \in X} \left\{ \langle \nabla f(\chi), \chi' - \chi \rangle + \frac{\ell}{2} \|\chi - \chi'\|^2 \right\}$. Then, f is said to satisfy the proximal Polyak-Łojasiewicz condition with modulus $\alpha > 0$ if for all $x \in X$, it holds true that:

$$\frac{1}{2\alpha} \mathcal{D}_X(\chi, \ell) \geq f(\chi) - \min_{\chi^\star \in X} f(\chi^\star).$$

When, $X = \mathbb{R}^n$ the pPŁ reduces to the PŁ condition.

3 Insights on Discounted Entropy Regularization for MGs

In this section, we demonstrate favorable properties of discounted entropy regularization for MGs. Later, we will see that the properties of discounted entropy render it a natural choice for regularization in MGs. Much like a Bregman divergence regularization term can turn a convex-concave game into a strongly-convex-strongly-concave one, adding discounted entropy transforms the weak gradient domination of the cMG's utility—that is nonconvex in policy—into a strong gradient domination condition. Crucially, this regularization strengthens the gradient domination condition for each agent while leaving that of its opponent intact.

3.1 Properties of Discounted Entropy

Our first observation comes by picking up the thread of the properties of the discounted entropy from where Neu et al. (2017) left it. In particular, we study its curvature on the feasible occupancy image obtained while holding the opponent fixed and find a uniform lower bound. This strong-convexity modulus is used to extract a uniform lower bound on the modulus of the pPŁ inequality. This is a departure from the previous analysis of (Mei et al., 2020) that considers a non-uniform PŁ condition. The rates we derive in terms of $\frac{1}{\epsilon}$ are slower compared to (Mei et al., 2020; 2021), yet they are expressed only in terms of the explicit parameters of the (convex) MDP.

Lemma 5. *For either player with the opponent's policy fixed, discounted negative entropy is uniformly $\alpha_{\mathcal{H}}$-strongly convex on that player's feasible normalized occupancy image, where* $\alpha_{\mathcal{H}} = \frac{(1-\gamma)\varrho_{\min}}{L_c^2} = \frac{\varrho_{\min}(1-\gamma)^5}{|\mathcal{S}|(|\mathcal{A}|+|\mathcal{B}|)^2}$.

Next, we can prove that the utility satisfies the pPŁ condition w.r.t. the policies' parameters $\chi \in X_R$ for a finite $R > 0$. This argument has been observed most recently in (Fatkhullin et al., 2023) for function over constrained domains.

Now, the proof for discounted KL follows by the same arguments as the ones for discounted entropy after a reduction. For a reference policy π_{ref} and a utility function $F(\mu)$, we can define a new utility function

$$\tilde{F}(\mu) = F(\mu) - \tau \sum_{s,a} \mu_{s,a} \log \pi_{\text{ref}}(a|s).$$

The addition of the linear term, $-\tau \sum_{s,a} \mu_{s,a} \log \pi_{\text{ref}}(a|s)$, does not alter the strong convexity modulus of the function and hence all our previous arguments go through.

4 Local Convergence of Simultaneous GDA

The single-agent pPŁ condition translates to the two-player setting: under entropy or KL regularization, the cMG utility is two-sided pPŁ. This is enough to tame the simultaneous-update dynamics that

dominate self-play training in practice, where both players take a gradient step against the opponent's current iterate within the same round. Although simultaneous GDA is known to oscillate or diverge on general saddle-point problems, under two-sided PŁ it converges locally at an exponential rate. We state the result abstractly; specialization to regularized cMGs is immediate.

Theorem 6. *Let f be C^2 in a neighborhood of a saddle point, with locally Lipschitz gradient, and satisfy a two-sided PŁ condition there with moduli $\alpha_1, \alpha_2 > 0$. Also assume that, for every $y \in \mathbb{R}^n$, $\arg\min_{x\in\mathbb{R}^m} f(\cdot, y)$ is bounded, and symmetrically for $\arg\max_{y\in\mathbb{R}^n} f(x, \cdot)$. Then gradient descent-ascent with a sufficiently small stepsize, initialized sufficiently close to $(x^\star, y^\star)$, converges to it at an exponential rate.*

Remark 1 (Boundedness of marginal optimizers). The boundedness assumption on $\arg\min_x f(\cdot, y)$ and $\arg\max_y f(x, \cdot)$ is mild for the regularized cMGs of interest. Discounted-KL (or discounted-entropy) regularization with strength $\tau > 0$ enforces full support of every regularized best-response policy. Under softmax parametrization, this translates into bounded logits.

A recent construction of Mourrat (2026) exhibits a two-sided PŁ function with a unique saddle-point on which GDA cycles from far-away initializations, ruling out global convergence in this regime. We therefore turn to alternating GDA, which does admit global guarantees.

5 Computing an Expert-Regularized Equilibrium

Having investigated the regularization properties of discounted entropy and KL, and having established a local convergence guarantee for simultaneous GDA, we now extend our analysis to alternating updates training and obtain global convergence guarantees. We focus on an alternating-update variant of piKL (Jacob et al., 2022; Bakhtin et al., 2022), in which each player updates sequentially using the other's most recent iterate:

$$\chi_{t+1} \leftarrow \Pi_X \left[\chi_t - \eta_1 \hat{\nabla}_\chi U^\tau(\chi_t, \psi_t)\right]; \quad \psi_{t+1} \leftarrow \Pi_\Psi \left[\psi_t + \eta_2 \hat{\nabla}_\psi U^\tau(\chi_{t+1}, \psi_t)\right] \qquad \text{(Alt-piKL)}$$

Convergence of (Alt-piKL) requires the regularized utility U^τ to satisfy a two-sided pPŁ condition—per (Kalogiannis et al., 2025, Theorem 3.4)). When only one player regularizes, the condition holds with respect to that player's policy immediately. The nontrivial step is showing that pPŁ is preserved on both sides when both players regularize simultaneously. This follows from Lemma 5: the opponent's discounted entropy is linear in a player's own occupancy measure, and therefore contributes no additional curvature. Concretely:

Lemma 7. *In a two-player zero-sum MG or cMG, the maximizer's discounted entropy regularization is concave in the maximizer's occupancy measure and convex in the minimizer's occupancy measure; symmetrically, the minimizer's discounted negative entropy regularization is convex in its own occupancy measure and concave in the maximizer's.*

This lemma immediately implies a two-sided pPŁ condition for the utility of the regularized game. Beyond this, it unifies two lines of work: discounted entropy regularization in Markov games and bidilated regularization in imperfect-information extensive-form games (Liu et al., 2025). Our analysis makes it clear that these settings are equivalent under an optimization lens—as imperfect-information extensive-form games were recently shown to exhibit a two-sided pPŁ condition (Kalogiannis & Farina, 2025).

Further, Lemma 7 yields the following corollary:

Corollary 8 (Reference-regularized MGs are cMGs)**.** *Let Γ be a two-player zero-sum Markov game with reward matrix, and let $(\pi_{1,ref}, \pi_{2,ref})$ be a pair of reference policies with $\pi_{1,ref}(a \mid s) > 0$ and $\pi_{2,ref}(b \mid s) > 0$ for all s, a, b. For any $\tau > 0$, the reference-KL-regularized utility defines a two-player zero-sum cMG.*

We note that, alternating regularized softmax policy gradient descent-ascent (Alt-Reg-PG) is an instance of Alt-piKL where $\pi_{1,\text{ref}}, \pi_{2,\text{ref}}$ are the uniform policies, $\pi_{1,\text{ref}}(s, \cdot) = \frac{\mathbf{1}}{|\mathcal{A}|}$ and $\pi_{2,\text{ref}}(s, \cdot) = \frac{\mathbf{1}}{|\mathcal{B}|}$ for all $s \in \mathcal{S}$.

Full-support references keep the KL terms finite; the constants worsen as $\min_a \pi_{i,\text{ref}}(a \mid s)$ approaches zero for either player i at some state s. Constraining the logits gives a uniform lower bound

on the nonzero spectrum of the softmax Jacobian and thus on the pPŁ modulus. The full-simplex approximation below quantifies the effect of this constraint.

Theorem 9. *Assume a two-player zero-sum cMG satisfying assumption 2, full-support reference policies $\pi_{1,ref}, \pi_{2,ref}$, and constrained logits $X_R \times \Psi_R$. For $\tau > 0$,* Alt-piKL *with $\eta_1 = \Theta(\alpha_\tau^2/\ell_U^3)$ and $\eta_2 = \Theta(1/\ell_U)$ reaches an ϵ_{opt}-RNE against deviations in $X_R \times \Psi_R$ using*

$$T = O\left(\frac{\ell_U^3}{\alpha_\tau^3}\log\frac{1}{\epsilon_{\mathrm{opt}}}\right), \qquad M = O\left(\frac{\ell_U^2}{\alpha_\tau^3\epsilon_{\mathrm{opt}}}\right),$$

trajectories per iteration, each of length

$$H = \Theta\left(\frac{1}{1-\gamma}\log\frac{L_F\ell_F\ell_U}{(1-\gamma)\alpha_\tau\epsilon_{\mathrm{opt}}}\right),$$

where $\alpha_\tau = \frac{\tau\alpha_{\mathcal{H}}}{202(1+(\max\{|\mathcal{A}|,|\mathcal{B}|\}-1)e^{2R})^2(L_{c^{-1}})^2}$ and $\alpha_{\mathcal{H}} = (1-\gamma)\varrho_{\min}/L_c^2$.

Remark 2. After choosing $R = \Theta(\log(1/\epsilon))$ to control full-simplex truncation and $\epsilon_{\mathrm{opt}} = \Theta(\epsilon)$, the iteration count scales as $\widetilde{O}(\tau^{-3}\epsilon^{-6})$. The exponent 6 arises because the pPŁ modulus is controlled through the nonzero spectrum of the constrained softmax map; it is an iteration bound, not a total trajectory bound.

From constrained logits to the full policy simplices. Let player i have d_i actions. Given any full-simplex deviation π_i, define statewise

$$\pi_i^{(R)} = (1-\beta_i)\pi_i + \beta_i\,\mathrm{Unif}_i, \qquad \beta_i = \frac{d_i}{e^{2R}+d_i-1}, \qquad t_i = \frac{d_i-1}{e^{2R}+d_i-1}.$$

Then $\pi_i^{(R)}$ is representable by logits of radius R and $\mathrm{TV}(\pi_i^{(R)}, \pi_i) \le t_i$ at every state. Writing $b_i = \max_s \log(1/\min_a \pi_{i,\mathrm{ref}}(a \mid s))$ and letting h_2 denote binary entropy, the induced change in regularized utility is at most

$$\Delta_i(R) = 2L_F L_c|\mathcal{S}|t_i + \tau\left[h_2(t_i) + t_i\log(d_i-1) + \frac{2t_i(b_i+\gamma b_{-i})}{1-\gamma}\right].$$

Consequently, an ϵ_{opt}-RNE over $X_R \times \Psi_R$ is an $(\epsilon_{\mathrm{opt}} + \max_i \Delta_i(R))$-RNE over the full policy simplices. Since $\Delta_i(R) = O(Re^{-2R})$ up to fixed game/reference constants, $R = \Theta(\log(1/\epsilon))$ makes the truncation error $O(\epsilon)$ and yields $\alpha_\tau = \widetilde{\Omega}(\tau\epsilon^2)$. For fixed τ and $\epsilon_{\mathrm{opt}} = \Theta(\epsilon)$, the resulting costs are $\widetilde{O}(\tau^{-3}\epsilon^{-6})$ iterations, $\widetilde{O}(\tau^{-6}\epsilon^{-13})$ trajectories, and $\widetilde{O}((1-\gamma)^{-1}\tau^{-6}\epsilon^{-13})$ sampled transitions. These trajectory bounds assume oracle access to $z = \nabla_\mu F$ as a pseudo-reward and do not include the cost of producing z; H controls truncation bias and M controls sampling error.

Corollary 10. *Let $B_{\mathrm{ref}} = b_1 + b_2$. Choosing $\epsilon_{\mathrm{opt}} = \epsilon/3$, R so that $\max_i \Delta_i(R) \le \epsilon/3$, and $\tau = \epsilon/(3B_{\mathrm{ref}})$ makes* Alt-Reg-PG *return an ϵ-NE over the full policy simplices. Suppressing logarithms and fixed game/reference constants, this requires $\widetilde{O}(\epsilon^{-9})$ iterations, $\widetilde{O}(\epsilon^{-19})$ trajectories, and $\widetilde{O}((1-\gamma)^{-1}\epsilon^{-19})$ sampled transitions.*

Recovering the single-agent guarantee. Retaining the minimizing player in the cMDP specialization, the alternating scheme reduces to projected stochastic gradient descent on U^τ:

$$\chi_{t+1} \leftarrow \Pi_{X_R}\left[\chi_t - \eta\hat{\nabla}_\chi U^\tau(\chi_t)\right]. \tag{Reg-PG}$$

Corollary 11. *Assume a cMDP such that assumption 2 holds and the agent employs constrained softmax policy parametrization. Then* (Reg-PG) *converges to an ϵ-approximate global minimum of the objective in $T = \tilde{O}\left(\frac{\ell_U}{L_{c^{-1}}^2\epsilon^2\alpha_\tau} + \frac{\sigma^2}{L_{c^{-1}}^4\epsilon^5\alpha_\tau^2}\right)$ iterations, with batch size $M = 1$ and horizon $H = O\left(\log\frac{\ell_U}{\epsilon\tau}\right)$.*

This corollary recovers and sharpens the cMDP analysis of Zhang et al. (2020)—which does not address softmax parametrization explicitly—and complements the entropy-regularized analysis of Mei et al. (2020), whose rate depends on an MDP-dependent constant whose scaling with explicit problem parameters is left unquantified.

6 Empirical Validation

We compare simultaneous piKL and Alt-piKL in three two-player OpenSpiel games (Lanctot et al., 2019): Pig (score 10, three die outcomes), Kuhn poker, and Oshi Zumo (coins 10, size 2, minimum bid 1, horizon 10). All use exact gradients and best responses, unconstrained softmax, two-sided discounted KL to uniform references, $\gamma = 0.95$, and $\tau = 0.01$. OpenSpiel player 0 is our minimizer and player 1 our maximizer. Pig's 1,134-state approximation omits the 1,000-decision timeout (discounted tail below 1.1×10^{-21}); Kuhn ties logits across 12 information sets; and Oshi uses 936 time-augmented states. Simultaneous rates for Pig/Kuhn/Oshi are $1000/1/10$ (also 0.3 for Kuhn); Alt-piKL uses $(100, 1000)/(0.3, 1)/(1, 10)$, updating the minimizer first at the slower rate. Ten paired seeds (0–9) use $\mathcal{N}(0, 0.2^2)$ logits without clipping or projection; budgets are 3,000 iterations for Pig and 10,000 for Kuhn/Oshi.

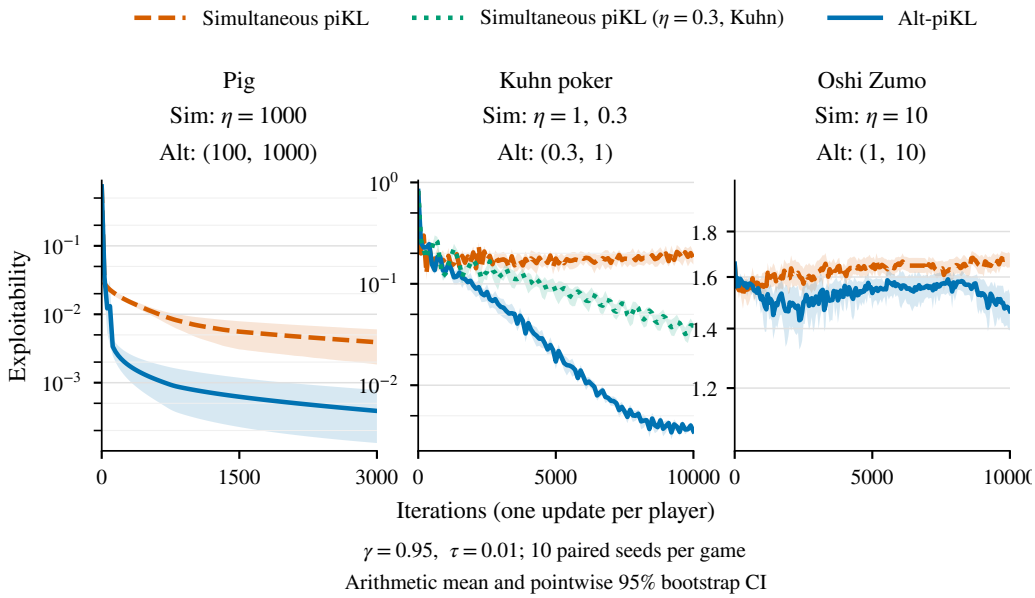


Figure 1: Mean unregularized best-response gap on logarithmic scales (each panel uses its own range) and pointwise 95% bootstrap confidence intervals over ten seeds. One iteration updates both players once; Alt-piKL uses two joint gradient evaluations, versus one for simultaneous piKL.

Exploitability is the sum of unregularized discounted deviation gains. Endpoint means are $0.00390/0.000387$ for Pig (Sim/Alt), $0.180/0.0388/0.00351$ for Kuhn (Sim-1/Sim-0.3/Alt), and $1.670/1.465$ for Oshi (Sim/Alt; Figure 1); Oshi's respective 95% intervals are $[1.641, 1.701]$ and $[1.394, 1.530]$. Alt has the lowest finite-budget gap in all three configurations, although Oshi remains highly exploitable and nonmonotone. Because update order and the first player's rate both change, this is neither a tuned ranking nor a test of the constrained-logit guarantee; imperfect-information Kuhn lies outside our Markov-game theorem.

7 Conclusion

We analyze expert-anchored self-play through Alt-piKL with softmax policies (Jacob et al., 2022; Bakhtin et al., 2022). Casting these games as cMGs, we establish occupancy-space curvature and transfer it to a two-sided proximal Polyak-Łojasiewicz condition. This yields local convergence of simultaneous GDA under the stated boundedness assumptions and global convergence of constrained alternating policy gradient in two-player zero-sum cMGs, with polynomial iteration and trajectory bounds. Three discounted-game experiments illustrate game- and rate-dependent behavior with unconstrained logits. Together, our results provide a foundation for expert-anchored self-play under tabular softmax parametrization; neural function approximation remains open.

REFERENCES

Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Atsushi Iwasaki. Adaptively perturbed mirror descent for learning in games. *arXiv preprint arXiv:2305.16610*, 2023.

Kenshi Abe, Mitsuki Sakamoto, Kaito Ariu, and Atsushi Iwasaki. Boosting perturbed gradient ascent for last-iterate convergence in games. *arXiv preprint arXiv:2410.02388*, 2024.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020.

Anton Bakhtin, David J Wu, Adam Lerer, Jonathan Gray, Athul Paul Jacob, Gabriele Farina, Alexander H Miller, and Noam Brown. Mastering the game of no-press diplomacy via human-regularized reinforcement learning and planning. *arXiv preprint arXiv:2210.05492*, 2022.

Anas Barakat, Ilyas Fatkhullin, and Niao He. Reinforcement learning with general utilities: Simpler variance reduction and large state-action space. In *International Conference on Machine Learning*, pp. 1753–1800. PMLR, 2023.

Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456): 885–890, 2019.

Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34:27952–27964, 2021.

Shicong Cen, Yuejie Chi, Simon S Du, and Lin Xiao. Faster last-iterate convergence of policy optimization in zero-sum markov games. *arXiv preprint arXiv:2210.01050*, 2022.

Christian Clason and Tuomo Valkonen. Introduction to nonsmooth analysis and optimization. *arXiv preprint arXiv:2001.00216*, 2020.

Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33: 5527–5540, 2020.

Jelena Diakonikolas, Constantinos Daskalakis, and Michael I Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2746–2754. PMLR, 2021.

Meta Fundamental AI Research Diplomacy Team FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

Ilyas Fatkhullin and Niao He. Taming nonconvex stochastic mirror descent with general bregman divergence. In *International Conference on Artificial Intelligence and Statistics*, pp. 3493–3501. PMLR, 2024.

Ilyas Fatkhullin, Niao He, and Yifan Hu. Stochastic optimization under hidden convexity. *arXiv preprint arXiv:2401.00108*, 2023.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Ian Gemp, Andreas Haupt, Luke Marris, Siqi Liu, and Georgios Piliouras. Convex markov games: A framework for fairness, imitation, and creativity in multi-agent learning. *arXiv preprint arXiv:2410.16600*, 2024.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning (ICML)*, 2019.

Ya-Ping Hsieh, Panayotis Mertikopoulos, and Volkan Cevher. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In *International Conference on Machine Learning*, pp. 4337–4348. PMLR, 2021.

Athul Paul Jacob, David J Wu, Gabriele Farina, Adam Lerer, Hengyuan Hu, Anton Bakhtin, Jacob Andreas, and Noam Brown. Modeling strong and human-like gameplay with kl-regularized search. In *International Conference on Machine Learning*, pp. 9695–9728. PMLR, 2022.

Fivos Kalogiannis and Gabriele Farina. Policy gradient methods converge globally in imperfect-information extensive-form games. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Fivos Kalogiannis, Emmanouil-Vasileios Vlatakis-Gkaragkounis, Ian Gemp, and Georgios Piliouras. Solving zero-sum convex markov games. *arXiv preprint arXiv:2506.16120*, 2025.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pp. 795–811. Springer, 2016.

Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, et al. Openspiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

Adam Lerer and Alexander Peysakhovich. Learning existing social conventions via observationally augmented self-play. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 107–114, 2019.

Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. A policy-gradient approach to solving imperfect-information games with iterate convergence. *arXiv preprint arXiv:2408.00751*, 2024.

Mingyang Liu, Gabriele Farina, and Asuman E Ozdaglar. A policy-gradient approach to solving imperfect-information games with best-iterate convergence. In *The Thirteenth International Conference on Learning Representations*, 2025.

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

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

Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *International Conference on Machine Learning*, pp. 7555–7564. PMLR, 2021.

Linjian Meng, Tianpei Yang, Youzhi Zhang, Zhenxing Ge, Shangdong Yang, Tianyu Ding, Wenbin Li, Bo An, and Yang Gao. Efficient last-iterate convergence in solving extensive-form games. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629*, 2018.

Jean-Christophe Mourrat. Pl conditions do not guarantee convergence of gradient descent-ascent dynamics. *arXiv preprint arXiv:2602.16517*, 2026.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, et al. Nash learning from human feedback. In *Forty-first International Conference on Machine Learning*, 2024.

Aziz Ben Nejma. Polyak-łojasiewicz inequality is essentially no more general than strong convexity for c^2 functions. *arXiv preprint arXiv:2512.05285*, 2025.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, et al. From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *International Conference on Machine Learning*, pp. 8525–8535. PMLR, 2021.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T Connor, Neil Burch, Thomas Anthony, et al. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419): 1140–1144, 2018.

Samuel Sokota, Ryan D'Orazio, J Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. *arXiv preprint arXiv:2206.05825*, 2022.

Gerald Tesauro et al. Temporal difference learning and td-gammon. *Communications of the ACM*, 38 (3):58–68, 1995.

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

Yongtao Wu, Luca Viano, Yihang Chen, Zhenyu Zhu, Kimon Antonakopoulos, Quanquan Gu, and Volkan Cevher. Multi-step alignment as markov games: An optimistic online gradient descent approach with convergence guarantees. *arXiv preprint arXiv:2502.12678*, 2025.

Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 33:1153–1165, 2020.

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020.

Junyu Zhang, Chengzhuo Ni, Zheng Yu, Csaba Szepesvari, and Mengdi Wang. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021.

Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. *Advances in Neural Information Processing Systems*, 32, 2019.

Runyu Zhang, Jincheng Mei, Bo Dai, Dale Schuurmans, and Na Li. On the global convergence rates of decentralized softmax gradient play in markov potential games. *Advances in Neural Information Processing Systems*, 35:1923–1935, 2022.

A Properties of Discounted Entropy Regularization

We first examine discounted entropy as a function of a player's state-action occupancy with the opponent fixed. We show uniform strong convexity on this feasible occupancy image (Lemma 12) and then derive the pPŁ inequality for the regularized utility under constrained softmax policies (Lemma 14).

Lemma 12. *Fix either player's opponent policy. Discounted negative entropy is $\alpha_{\mathcal{H}}$-strongly convex on the resulting feasible set of normalized occupancies μ, where*

$$\alpha_{\mathcal{H}} := \frac{(1-\gamma)\varrho_{\min}}{L_c^2} = \frac{\varrho_{\min}(1-\gamma)^5}{|\mathcal{S}|(|\mathcal{A}|+|\mathcal{B}|)^2}.$$

Proof. Consider two occupancies μ, μ' induced by policies π, π' while the opponent policy is held fixed. Let $\mathcal{H}(\mu) := \sum_{s,a} \mu_{s,a} \log(\mu_{s,a}/d_s)$, where $d_s = \sum_a \mu_{s,a}$, and let $\pi(a \mid s) = \mu_{s,a}/d_s$. By Neu et al. (2017, Prop. 1),

$$D_{\mathcal{H}}(\mu', \mu) = \sum_s d'_s D_{\mathrm{KL}}(\pi'(\cdot \mid s) \| \pi(\cdot \mid s)) \,.$$

The normalization of the discounted occupancy gives $d'_s \geq (1-\gamma)\varrho(s) \geq (1-\gamma)\varrho_{\min}$. Applying Pinsker statewise and then summing globally yields

$$\begin{aligned} D_{\mathcal{H}}(\mu', \mu) &\geq \frac{(1-\gamma)\varrho_{\min}}{2} \sum_s \|\pi'(\cdot \mid s) - \pi(\cdot \mid s)\|_1^2 \\ &\geq \frac{(1-\gamma)\varrho_{\min}}{2} \|\pi' - \pi\|_2^2. \end{aligned}$$

With the opponent fixed, the global policy-to-occupancy Lipschitz bound in Lemma 1 gives $\|\mu' - \mu\|_2 \leq L_c \|\pi' - \pi\|_2$. Consequently,

$$D_{\mathcal{H}}(\mu', \mu) \geq \frac{(1-\gamma)\varrho_{\min}}{2L_c^2} \|\mu' - \mu\|_2^2 = \frac{\alpha_{\mathcal{H}}}{2} \|\mu' - \mu\|_2^2,$$

proving the claim. This argument uses the global Lipschitz inequality only after summing Pinsker's bounds; it makes no unsupported statewise inverse-Lipschitz step. □

Proposition 13. *With the opponent fixed, Kullback–Leibler regularization with respect to a reference policy π_{ref} is strongly convex on the player's feasible occupancy image.*

Proof. The proof follows from the simple observation that discounted KL is the summation of discounted entropy and a term linear in μ. A linear term does not change the curvature of a function.

$$\begin{aligned} F(\mu) &:= \sum_{s,a} \mu_{s,a} \log \frac{\pi(a|s)}{\pi_{\text{ref}}(a|s)} \\ &= \sum_{s,a} \mu_{s,a} \log \frac{\mu_{s,a}}{\sum_{a'} \mu_{s,a'}} - \sum_{s,a} \mu_{s,a} \log \pi_{\text{ref}}(a|s). \end{aligned}$$

□

Lemma 14. *A discounted entropy regularized cMG satisfies the pPŁ condition with modulus*

$$\alpha_\tau = \frac{\tau \alpha_{\mathcal{H}}}{202(1+(|\mathcal{A}|-1)e^{2R})^2 (L_{c^{-1}})^2},$$

i.e.:

$$\frac{1}{2\alpha_\tau}\mathcal{D}_X(\chi,\ell;\psi) \geq U^\tau(\chi,\psi) - \min_{\chi^\star \in X_R} U^\tau(\chi^\star,\psi);$$
$$\frac{1}{2\alpha_\tau}\mathcal{D}_\Psi(\psi,\ell;\chi) \geq \max_{\psi^\star \in \Psi_R} U^\tau(\chi,\psi^\star) - U^\tau(\chi,\psi).$$

Proof. We will prove the pPŁ condition for the minimizing player considering the policy maximizer fixed. The proof for the maximizer is symmetric. We start by considering the utility as a function of the occupancy measure establishing the Kurdyka-Łojasiewicz inequality from strong convexity. Then, we will use the chain rule to ultimately prove the Kurdyka-Łojasiewicz inequality for the policy and analogolously for the logits of the softmax policy.

Let us introduce some auxilliary notation from this proof:

- Π_{X_R} is a convex set of softmax policies parametrized by X_R;
- M_R denotes the convex set of state-action occupancy measures achievable from Π_{X_R};
- $I_{\mathcal{C}}$ will denote the indicator function of a convex set $\mathcal{C}$;
- $\mathbf{J}_f(x)$ denotes the Jacobian of vector-valued function f w.r.t. variable x;
- $\partial_x f$ will denote the Fréchet subdifferential of a function f w.r.t. variable x;
- For a fixed $\psi \in \Psi_R$, let $\tilde{F}$ and $\tilde{U}$ denote the utility of the minimizing player as a function of the occupancy measure (with the addition of the appropriate indicator function) and as a function of the policy and the policy parameters, respectively.

$$\tilde{F}(\mu) := U^\tau(\mu^{-1}(\mu), \pi^\psi) + I_{M_R}(\mu);$$
$$\tilde{U}(\pi^\chi) := U^\tau(\pi^\chi, \pi^\psi) + I_{\Pi_{X_R}}(\pi^\chi);$$
$$\tilde{U}(\chi) := U^\tau(\pi^\chi, \pi^\psi) + I_{X_R}(\chi).$$

First, from α-strong convexity of the regularized $\tilde{F}$, we conclude the Kurdyka-Łojasiewicz condition w.r.t. μ:

$$\tilde{F}(\mu) - \min_{\mu^\star \in M_R} \tilde{F}(\mu^\star) \leq \frac{1}{2\alpha}\|v_\mu\|^2, \quad \forall v_\mu \in \partial_\mu F(\mu;\psi). \tag{1}$$

For the next step, our goal is to connect $\|v_\mu\|$ where $v_\mu \in \partial_\mu \tilde{F}(\mu)$ to $\|v_\chi\|$ for $v_\chi \in \partial_\chi U(\chi)$. For that purpose, we use the chain rule (Clason & Valkonen, 2020, Theorem 13.23) to write,

$$\partial_\chi \tilde{U}(\pi^\chi) \subseteq \{\mathbf{J}^\top_{\pi^\chi}(\chi)\mathbf{J}^\top_\mu(\pi^\chi)v_\mu \mid v_\mu \in \partial_\mu \tilde{F}\}.$$

Now, since $\mathbf{J}_{\pi^\chi}(\chi)$ is symmetric and rank deficient in the direction the all-ones vector $\mathbf{1}$ we see that,

$$\{\mathbf{J}^\top_{\pi^\chi}(\chi)\mathbf{J}^\top_\mu(\pi^\chi)v_\mu \mid v_\mu \in \partial_\mu \tilde{F}\} = \{\mathbf{J}^\top_{\pi^\chi}(\chi)\mathbf{J}^\top_\mu(\pi^\chi)v_\mu \mid v'_\mu \in \partial_\mu \tilde{F} \wedge \mathbf{1}^\top v'_\mu\}.$$

Hence, the inclusion translates into the inequality:

$$\begin{aligned}
\min_{v_\chi \in \partial \tilde{U}(\chi)} \|v_\chi\| &\geq \min_{v_\mu, v_\mu^\top \mathbf{1}=0} \left\|\mathbf{J}^\top_{\pi^\chi}(\chi)\mathbf{J}^\top_\mu(\pi^\chi)v_\mu\right\| \\
&\geq \min_{\|w\|_2=1, w^\top \mathbf{1}=0} \sqrt{w^\top \left[\mathbf{J}_\mu(\pi^\chi)\mathbf{J}_{\pi^\chi}(\chi)\mathbf{J}^\top_{\pi^\chi}(\chi)\mathbf{J}^\top_\mu(\pi^\chi)\right] w} \min_{v_\mu, v_\mu^\top \mathbf{1}=0} \|v_\mu\| \\
&\geq \min_{\|w\|_2=1, w^\top 1=0} \sqrt{w^\top \left[\mathbf{J}_\mu(\pi^\chi)\mathbf{J}_{\pi^\chi}(\chi)\mathbf{J}^\top_{\pi^\chi}(\chi)\mathbf{J}^\top_\mu(\pi^\chi)\right] w} \min_{v_\mu, v_\mu^\top \mathbf{1}=0} \|v_\mu\|
\end{aligned}$$

$$
\begin{aligned}
&\geq \lambda_{\min}(\mathbf{J}_\mu(\pi^\chi)) \min_{\|w\|_2=1, w^\top \mathbf{1}=0} \sqrt{w^\top \left[\mathbf{J}_{\pi^\chi}(\chi)\mathbf{J}_{\pi^\chi}^\top(\chi)\right] w} \min_{v_\mu, v_\mu^\top \mathbf{1}=0} \|v_\mu\| \\
&\geq \frac{1}{L_{c^{-1}}} \min_{s,\chi \in X_R} \text{softmax}(\chi_s) \min_{v_\mu, v_\mu^\top \mathbf{1}=0} \|v_\mu\| \\
&\geq \frac{1}{(1+(d-1)e^{2R})\, L_{c^{-1}}} \|v_\mu\| . \qquad (2)
\end{aligned}
$$

where,

- in the second inequality we use the fact that $\|Ax\| \geq \min_{w=1} \sqrt{w^\top A^\top A w}\, \|x\|$,
- in the third, we use the fact that $\lambda_{\min}(ABB^\top A^\top) \geq \lambda_{\min}(AA^\top)\lambda_{\min}(BB^\top)$,
- in the penultimate, we use the fact that the mapping of policies to occupancies is "1–1" and Lipschitz continuous along Lemma 1,
- finally, we use Lemma 16 for the lower bound of the block-diagonal matrix $\mathbf{J}_{\pi^\chi}(\chi)$ as a function of the smallest entry of the probability vector.

Plugging (2) into (1), we get:

$$
\frac{(1+(|\mathcal{A}|-1)e^{2R})^2 (L_{c^{-1}})^2}{2\tau\alpha_{\mathcal{H}}} \|v_\chi\|^2 \geq U^\tau(\chi,\psi) - \min_{\chi^\star \in X_R} U^\tau(\chi^\star,\psi).
$$

The, we invoke Lemma 18 to retrieve the claim. □

We note that in order to guarantee a uniform bound on α_τ that also allows an ϵ-small optimality gap, we demand that $\min_{s\in\mathcal{S},\, a\in\mathcal{A}} \pi_\chi(a|s) \geq \epsilon$ and pick $R = \frac{1}{2}\log\left(\frac{1-\epsilon}{\epsilon(|\mathcal{A}|-1)}\right)$ which results in

$$
\alpha_\tau \geq \frac{\tau\alpha_{\mathcal{H}}\epsilon^2}{202(L_{c^{-1}})^2}.
$$

B Constrained Softmax Parametrization

In this section we go over the properties of the softmax parametrized policies under constraints and prove the pPŁ condition for the utility function of the utility function for the softmax policies.

B.1 Properties of Constrained Softmax Parametrization

The properties of the constrained softmax parametrization follow from well known facts and arguments in the theory of the softmax function (Gao & Pavel, 2017). Formal proofs can be found in (Kalogiannis & Farina, 2025, Appendix D.2). The first lemma bounds the spectrum of the Jacobian of softmax which is crucial for the lower bound on the modulus of the pPŁ condition.

Lemma 15 (Spectrum of the Softmax Jacobian)**.** *Let* $\sigma = \text{softmax}(\theta) \in \mathbb{R}^d$ *and let* $\mathbf{J} := \mathbf{J}_{\text{softmax}}(\theta) \in \mathbb{R}^{d\times d}$ *denote the Jacobian of the softmax map. Then:*

(i) $\mathbf{J} = \text{diag}(\sigma) - \sigma\sigma^\top$.

(ii) $\mathbf{J}$ *is positive semidefinite with* $\text{rank}(\mathbf{J}) = d-1$.

(iii) $\mathbf{J}\mathbf{1} = 0$*, i.e.,* $\mathbf{1}$ *is a null eigenvector.*

(iv) The remaining $d-1$ *eigenvalues satisfy*

$$
\lambda_i \in \left[\min_{j\in[d]} \sigma_j,\, \max_{j\in[d]} \sigma_j\right] \subseteq (0,1).
$$

The constraints in the parameter space yield a maximum and minimum probability over all actions. Here, we quantify the bounds depending on the radius R.

Lemma 16 (Softmax Probability Bounds with Parameter Constraints)**.** *Define the set*

$$\Theta_R := \{\theta \in \mathbb{R}^d : \theta^\top \mathbf{1} = 0 \text{ and } \max_{i,j\in[d]} |\theta_i - \theta_j| \le 2R\}.$$

For any $\theta \in \Theta_R$, the softmax probabilities satisfy

$$\frac{1}{1+(d-1)e^{2R}} \le \min_{i\in[d]} \mathrm{softmax}_i(\theta) \le \max_{i\in[d]} \mathrm{softmax}_i(\theta) \le \frac{1}{1+(d-1)e^{-2R}}.$$

Finally, and most crucially, the image of the convex parameter set (X_R and Ψ_R) yields a convex set of policies, even though softmax is not a convex mapping.

Lemma 17 (Convexity of the Softmax Image)**.** *For $R > 0$, define the set*

$$\Theta_R := \{\theta \in \mathbb{R}^d : \theta^\top \mathbf{1} = 0 \text{ and } \max_{i,j\in[d]} |\theta_i - \theta_j| \le 2R\}$$

Then the image $\mathrm{softmax}(\Theta_R) \subseteq \Delta^{d-1}$ *is convex.*

B.2 Gradient Domination Conditions

Definition 6. Let f be a function $f\colon \mathcal{X} \to \mathbb{R}$. Further, define $F(\cdot) := f(\cdot) + I_{\mathcal{X}}(\cdot)$ where $I_{\mathcal{X}}(\cdot)$ is the indicator function of $\mathcal{X}$. The function is said to satisfy the Kurdyka-Łojasiewicz condition with a modulus $\mu_{\mathrm{kl}} > 0$, if for all x:

$$\frac{1}{2\mu_{\mathrm{kl}}} \|v_x\|^2 \ge f(x) - \min_{x^\star} f(x^\star), \quad \forall v_x \in \partial_x F(x).$$

Lemma 18 (KŁ Implies Proximal-PŁ)**.** *Let $f\colon \mathcal{X} \to \mathbb{R}$ be an ℓ-smooth function on a convex set $\mathcal{X} \subseteq \mathbb{R}^d$. If f satisfies the* KŁ condition *with modulus $\mu_{kl} > 0$, i.e.,*

$$\frac{1}{2}\,\mathrm{dist}^2(0, \partial F(x)) \ge \mu_{kl}\big(f(x) - f^\star\big) \quad \forall x \in \mathcal{X},$$

where $F := f + I_{\mathcal{X}}$, then f also satisfies the proximal-PŁ condition *with modulus $\mu_{pl} = \frac{\mu_{kl}}{202}$.*

Proof. The equivalence was established in Karimi et al. (2016) and the following proof that improves the constant is due to Kalogiannis & Farina (2025) and we include it for completeness. Define $F := f + I_{\mathcal{X}}$, where $I_{\mathcal{X}}$ is the indicator function of $\mathcal{X}$. Since f is ℓ-smooth and $I_{\mathcal{X}}$ is convex, the function F is ℓ-weakly convex, *i.e.*,

$$F(y) \ge F(x) + \langle v, y - x\rangle - \frac{\ell}{2}\|x - y\|^2 \quad \forall x, y \in \mathbb{R}^d,\ v \in \partial F(x).$$

In particular, for any $\rho > \ell$, the proximal operator $\mathrm{prox}_{F/\rho}$ is well-defined. Fix $x \in \mathcal{X}$, let $\rho > \ell$, and define $\hat{x} := \mathrm{prox}_{F/\rho}(x)$. By the optimality condition for the proximal operator,

$$s_{\hat{x}} := \rho(x - \hat{x}) \in \partial F(\hat{x}),$$

and moreover, $\nabla F_{1/\rho}(x) = \rho(x - \hat{x}) = s_{\hat{x}}$, where $F_{1/\rho}$ denotes the Moreau envelope of F with parameter $1/\rho$.

Apply the KŁ condition at the proximal point. By the KŁ assumption applied at $\hat{x} \in \mathcal{X}$,

$$\frac{1}{2}\|\nabla F_{1/\rho}(x)\|^2 = \frac{1}{2}\|s_{\hat{x}}\|^2 \ge \mu_{\mathrm{kl}}\big(f(\hat{x}) - f^\star\big). \tag{3}$$

Relate $f(\hat{x})$ to $f(x)$ via weak convexity. Using weak convexity of F with any $v \in \partial F(x)$:

$$\begin{aligned} F(\hat{x}) &\ge F(x) + \langle v, \hat{x} - x\rangle - \frac{\ell}{2}\|x - \hat{x}\|^2 \\ &= F(x) + \langle v, \hat{x} - x\rangle + \frac{\rho}{2}\|x - \hat{x}\|^2 - \frac{\ell + \rho}{2}\|x - \hat{x}\|^2 \end{aligned}$$

$$\geq F(x) + \min_{y\in\mathcal{X}} \left\{ \langle \nabla f(x), y - x\rangle + \frac{\rho}{2}\|x-y\|^2 \right\} - \frac{\ell+\rho}{2}\|x-\hat{x}\|^2$$
$$= F(x) - \frac{1}{2\rho}\mathcal{D}(x,\rho) - \frac{\ell+\rho}{2}\|x-\hat{x}\|^2,$$

where $\mathcal{D}(x,\rho) := -2\rho \min_{y\in\mathcal{X}} \{\langle \nabla f(x), y-x\rangle + \frac{\rho}{2}\|x-y\|^2\}$ is the forward-backward measure. Since $F = f$ on $\mathcal{X}$, rearranging gives

$$f(\hat{x}) - f(x) \geq -\frac{1}{2\rho}\mathcal{D}(x,\rho) - \frac{\ell+\rho}{2}\|x-\hat{x}\|^2. \tag{4}$$

Combine the bounds. Substituting (4) into (3):

$$\frac{1}{2}\|\nabla F_{1/\rho}(x)\|^2 \geq \mu_{\text{kl}}\big(f(x) - f^\star\big) - \mu_{\text{kl}}\left(\frac{1}{2\rho}\mathcal{D}(x,\rho) + \frac{\ell+\rho}{2}\|x-\hat{x}\|^2\right).$$

Using $\|x-\hat{x}\|^2 = \frac{1}{\rho^2}\|\nabla F_{1/\rho}(x)\|^2$ and rearranging:

$$\left(\frac{1}{2} + \mu_{\text{kl}}\frac{\ell+\rho}{2\rho^2}\right)\|\nabla F_{1/\rho}(x)\|^2 + \frac{\mu_{\text{kl}}}{2\rho}\mathcal{D}(x,\rho) \geq \mu_{\text{kl}}\big(f(x)-f^\star\big).$$

Choose ρ and simplify. Since $\mathcal{D}(x,\rho)$ is nondecreasing in ρ (cf. (Karimi et al., 2016, Lemma 1)), we have $\mathcal{D}(x,2\rho) \geq \mathcal{D}(x,\rho)$, so

$$\left(\frac{1}{2} + \mu_{\text{kl}}\frac{\ell+\rho}{2\rho^2}\right)\|\nabla F_{1/\rho}(x)\|^2 + \frac{\mu_{\text{kl}}}{2\rho}\mathcal{D}(x,2\rho) \geq \mu_{\text{kl}}\big(f(x)-f^\star\big).$$

Setting $\rho = 4\ell$:

$$\left(\frac{1}{2} + \frac{5\mu_{\text{kl}}}{16\ell}\right)\|\nabla F_{1/(4\ell)}(x)\|^2 + \frac{\mu_{\text{kl}}}{8\ell}\mathcal{D}(x,8\ell) \geq \mu_{\text{kl}}\big(f(x)-f^\star\big).$$

Since $\mu_{\text{kl}} \leq \ell$ in general, we have $\frac{1}{2} + \frac{5\mu_{\text{kl}}}{16\ell} \leq \frac{1}{2} + \frac{5}{16} = \frac{13}{16}$, so

$$\frac{13}{16}\|\nabla F_{1/(4\ell)}(x)\|^2 + \frac{\mu_{\text{kl}}}{8\ell}\mathcal{D}(x,8\ell) \geq \mu_{\text{kl}}\big(f(x)-f^\star\big).$$

Convert to the proximal-PŁ condition. By (Fatkhullin & He, 2024, Lemmas 4.1 and 4.2), we have

$$\|\nabla F_{1/(4\ell)}(x)\|^2 \leq 16\ell \cdot \mathcal{D}(x,8\ell).$$

Substituting and simplifying:

$$\left(\frac{13}{16}\cdot 16\ell + \frac{\mu_{\text{kl}}}{8}\right)\frac{\mathcal{D}(x,8\ell)}{\ell} \geq \mu_{\text{kl}}\big(f(x)-f^\star\big).$$

Since $\frac{\mu_{\text{kl}}}{8\ell} \leq \frac{1}{8}$, the left-hand side is bounded by $(13+\frac{1}{8})\mathcal{D}(x,8\ell) \leq \frac{105}{8}\mathcal{D}(x,8\ell)$. Thus,

$$\frac{1}{2}\mathcal{D}(x,8\ell) \geq \frac{\mu_{\text{kl}}}{101}\big(f(x)-f^\star\big),$$

which establishes the proximal-PŁ condition with modulus $\mu_{\text{pl}} = \frac{\mu_{\text{kl}}}{101}$. □

C Convergence of Reg-PG in cMDPs—Proofs of Section 3

In this section, we prove Corollary 11 using the analysis of stochastic gradient in nonconvex functions under generalized pPŁ (see (Fatkhullin et al., 2023)).

Theorem 19. *Assume a cMDP such that Assumption 2 is true. Then, Equation (Reg-PG) with*

$$\eta = \Theta\left(\frac{\tau\epsilon^3}{(L_{c^{-1}})^2\ell_U}\right), \quad H = \Theta\left(\frac{1}{1-\gamma}\log\frac{L_F\ell_F\ell_U}{(1-\gamma)L_{c^{-1}}\epsilon}\right)$$

converges to an ϵ-approximate global optimum of the utility after at most:

$$T = \tilde{O}\left(\frac{(L_{c^{-1}})^4\ell_U}{\tau^2\epsilon^5}\right)$$

iterations.

Proof. We invoke the arguments used in (Fatkhullin & He, 2024, Theorem 4.7) along a biased stochastic gradient oracle (their arguments are meant for an unbiased one). In particular, for the minimization problem for a function f that is ℓ-smooth and α-pPŁ,

$$\min_{\chi\in X} f(\chi) \equiv \min_{\chi\in\mathbb{R}^d}\{F(\chi) := f(\chi) + I_X(\chi)\},$$

where $I_X(\chi) = \begin{cases} +\infty, \text{ if } \chi \notin X; \\ 0, \text{ else} \end{cases}$. In addition, we assume a biased stochastic gradient oracle with bounded second moment,

$$\|\mathbb{E}[\hat{g}_{\chi_t}] - \nabla f(\chi_t)\| \le \delta;$$
$$\mathbb{E}\left[\|\hat{g}_{\chi_t}\|^2\right] \le \sigma^2.$$

To show convergence of stochastic gradient descent, Fatkhullin & He (2024) consider the Lyapunov function

$$V(\chi_t) := F_{1/\rho}(\chi_t) - F^\star + \eta\rho(F(\chi_t - F^\star)).$$

To accommodate the biased stochastic estimate, an alternative to (Fatkhullin & He, 2024, Equation 16) for $e_t := \hat{g}_{\chi_t} - \nabla f(\chi_t)$ we write,

$$\begin{aligned}\mathbb{E}\left[\eta\left\langle e_t, \chi_t - \chi_{t+1}\right\rangle - (1-\eta\ell)\left\|x_t - x_{t+1}\right\|^2\right] &\le \frac{\eta^2}{2(1-\eta\ell)}\mathbb{E}\left[\|e_t\|^2\right] \\ &\le \frac{\eta^2}{2(1-\eta\ell)}\mathbb{E}\left[2\|\delta\|^2 + 2\|\sigma\|^2\right].\end{aligned}$$

□

From the latter, along $\eta \le \frac{1}{2\ell}$ and (Fatkhullin & He, 2024, Eq. 15), we conclude that,

$$\mathbb{E}V(\chi_{t+1}) \le (1-\eta\mu/3)\,\mathbb{E}V(\chi_t) + 4\ell\eta^2\sigma^2 + 4\ell\eta^2\delta^2.$$

Which implies,

$$\mathbb{E}V(\chi_T) \le V(\chi_0)\exp\left(-\eta\alpha T/3\right) + \frac{12\ell\eta\sigma^2}{\alpha} + \frac{12\ell\eta\delta^2}{\alpha}.$$

Now, we plug in δ the bias of the REINFORCE gradient estimator, the pPŁ modulus of the regularized utility α_τ for $\tau = \frac{\epsilon}{\log|\mathcal{A}|}$ and its smoothness ℓ_U to tune the step-size and get the upper bound on the number of iterations T and the needed horizon of the trajectory samples H.

D Convergence in cMGs—Proofs of Section 5

First, we note that the variance and bias of the REINFORCE gradient estimator are bounded.

Lemma 20 (Variance and Bias of REINFORCE)**.** *The batched policy gradient estimator* REINFORCE*, with* $H > \frac{1}{\ln 1/\sqrt{\gamma}}$ *, has variance* σ^2 *and bias* δ *satisfying* $\sigma^2 = O(1)$ *and* $\delta = O\Big(\frac{L_F^2\,\ell_F}{(1-\gamma)^6}\,\exp(-(1-\gamma)(H-1))\Big)$.

In this section we formally prove our claims for the multi-agent expert-anchored equilibrium computation. We use the fact that a discounted KL regularized utility function is two-sided pPŁ and invoke arguments from recent results.

D.1 Convergence of Alt-piKL

Convergence of (Alt-piKL) is analyzed through the general convergence results of projected gradient descent-ascent (Alt-AGDA) under a two-sided pPŁ condition. The updates read:

$$\begin{aligned} x_{t+1} &\leftarrow \Pi_{\mathcal{X}}\left[x_t - \eta_1\hat{\nabla}_x f(x_t, y_t)\right] \\ y_{t+1} &\leftarrow \Pi_{\mathcal{Y}}\left[y_t + \eta_2\hat{\nabla}_y f(x_{t+1}, y_t)\right] \end{aligned} \qquad \text{(Alt-AGDA)}$$

Before proceeding, to the theorem we just state a standard assumption on the stochastic gradient oracle.

Assumption 3 (Unbiased Gradient Estimators with Bounded Second Moments)**.** For all iterations t, the stochastic gradient estimators $\hat{g}_x(x_t, y_t)$ and $\hat{g}_y(x_t, y_t)$ are unbiased:

$$\mathbb{E}\big[\hat{g}_x(x_t, y_t)\big] = g_x(x_t, y_t),$$
$$\mathbb{E}\big[\hat{g}_y(x_t, y_t)\big] = g_y(x_t, y_t),$$

and have bounded second moments:

$$\mathbb{E}\big[\|\hat{g}_x(x_t, y_t)\|^2\big] \leq \sigma_x^2,$$
$$\mathbb{E}\big[\|\hat{g}_y(x_t, y_t)\|^2\big] \leq \sigma_y^2.$$

Furthermore, the oracle g approximates the true gradient with bounded error:

$$\|g_x(x_t, y_t) - \nabla_x f(x_t, y_t)\| \leq \delta_x,$$
$$\|g_y(x_t, y_t) - \nabla_y f(x_t, y_t)\| \leq \delta_y.$$

Now, we restate the convergence of theorem for min-max optimization of two-sided pPL functions from Kalogiannis et al. (2025):

Theorem 21 ((Kalogiannis et al., 2025, Theorem D.8))**.** *Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be an L-Lipschitz, ℓ-smooth function, where $\mathcal{X}$ and $\mathcal{Y}$ are compact convex sets with Euclidean diameters $\mathrm{D}_{\mathcal{X}}$ and $\mathrm{D}_{\mathcal{Y}}$, respectively. Suppose that:*

(i) $f(\cdot, y)$ satisfies the pPL condition with modulus α_x for all $y \in \mathcal{Y}$;

(ii) $f(x, \cdot)$ satisfies the pPL condition with modulus α_y for all $x \in \mathcal{X}$;

(iii) $(\hat{g}_\chi, \hat{g}_\psi)$ is an inexact stochastic gradient oracle satisfying Assumption 3.

Then, after T iterations of (Alt-AGDA) *with step-sizes $\eta_1 = \frac{\alpha_y^2}{160\ell^3}$ and $\eta_2 = \frac{1}{5\ell}$, we have:*

$$\mathbb{E}\Phi(x_T) - \Phi^\star + \tfrac{1}{10}\left(\mathbb{E}\Phi(x_T) - \mathbb{E}f(x_T, y_T)\right)$$
$$\leq \exp\left(-\frac{\alpha_x \alpha_y^2}{160\ell^3} T\right) L(\mathrm{D}_{\mathcal{X}} + \mathrm{D}_{\mathcal{Y}}) + \frac{c_1\sigma_x^2}{\alpha_x} + \frac{c_1\delta_x^2}{\alpha_x} + \frac{c_2\ell^2\sigma_y^2}{\alpha_x\alpha_y^2} + \frac{c_2\ell^2\delta_y^2}{\alpha_x\alpha_y^2},$$

where, $c_1, c_2 \in O(1)$.

We put the pieces together, *i.e.,* the two-sided pPŁ lemma for the regularized utility Lemma 7 and Theorem 21.

Theorem 22. *Assume a two-player zero-sum cMG such that assumption 2 holds. Also, let $(\pi_{1,ref}, \pi_{2,ref})$ be a pair of reference policies and $\tau > 0$. Then,* (Alt-piKL) *with step-sizes $\eta_1 = \Theta\left(\frac{\alpha_\tau^2}{\ell_U^3}\right)$ and $\eta_2 = \Theta\left(\frac{1}{\ell_U}\right)$ converges to an ϵ-RNE after iterations that are:*

$$T = O\left(\frac{\ell_U^3}{\alpha_\tau^3} \log \frac{1}{\epsilon}\right),$$

where

$$\alpha_\tau := \frac{\tau\alpha_{\mathcal{H}}}{202\left(1 + (\max\{|\mathcal{A}|, |\mathcal{B}|\} - 1)\, e^{2R}\right)^2 (L_{c^{-1}})^2}, \qquad \alpha_{\mathcal{H}} = \frac{(1-\gamma)\varrho_{\min}}{L_c^2}.$$

by picking a trajectory horizon

$$H = \Theta\left(\frac{1}{1-\gamma} \log \frac{L_F \ell_F\, \ell_U}{(1-\gamma)\alpha_\tau \epsilon}\right)$$

and a sufficient batch-size is

$$M = O\left(\frac{\ell_U^2}{\alpha_\tau^3 \epsilon}\right).$$

Proof. The proof of Theorem 22 follows from Theorem 21. □

E LOCAL CONVERGENCE UNDER TWO-SIDED PŁ

Theorem 23. *Assume that $f(x, y)$ is C^2 in a neighborhood of a saddle point, has locally Lipschitz gradient there, and satisfies a two-sided PŁ condition with moduli $\alpha_1, \alpha_2 > 0$. Also assume that for every $y \in \mathbb{R}^n$ the set $\arg\min_{x\in\mathbb{R}^m} f(\cdot, y)$ is bounded and, symmetrically, $\arg\max_{y\in\mathbb{R}^n} f(x, \cdot)$ is bounded for every x. Then gradient descent-ascent with a sufficiently small stepsize $\eta > 0$, initialized sufficiently close to $(x^\star, y^\star)$, converges to it at an exponential rate.*

Proof. We use the Lyapunov linearization theorem for the discrete-time dynamical system induced by the gradient descent-ascent dynamics. The Jacobian of the dynamics about the saddle-point $(x^\star, y^\star)$ reads:

$$\mathbf{J}_{\mathrm{GDA}}(x^\star, y^\star) := \mathbf{I} - \eta \underbrace{\begin{bmatrix} \nabla^2_{xx} f(x^\star, y^\star) & \nabla^2_{xy} f(x^\star, y^\star) \\ -(\nabla^2_{xy} f(x^\star, y^\star))^\top & -\nabla^2_{yy} f(x^\star, y^\star) \end{bmatrix}}_{=:\mathbf{M}}.$$

We need to make sure that all eigenvalues ρ_i of $\mathbf{J}_{\mathrm{GDA}}(x^\star, y^\star)$ satisfy $|\rho_i| < 1$. As $\rho_i = 1 - \eta\lambda_i$, it suffices to make sure that the eigenvalues λ_i of $\mathbf{M}$ have a positive real part. Indeed, we can write $\mathbf{M} = \mathbf{S} + \mathbf{T}$ with $\mathbf{S} = \frac{1}{2}\left(\mathbf{M} + \mathbf{M}^\top\right)$ and $\mathbf{T} = \frac{1}{2}(\mathbf{M} - \mathbf{M}^\top)$.

$$\mathbf{S} = \begin{bmatrix} \nabla^2_{xx} f(x^\star, y^\star) & 0 \\ 0 & -\nabla^2_{yy} f(x^\star, y^\star) \end{bmatrix}, \quad \mathbf{T} = \begin{bmatrix} 0 & \nabla^2_{xy} f(x^\star, y^\star) \\ -(\nabla^2_{xy} f(x^\star, y^\star))^\top & 0 \end{bmatrix}.$$

Clearly, since $x^\star$ is a minimum of $f(\cdot, y^\star)$, the matrix $\nabla^2_{xx} f(x^\star, y^\star)$ is positive semidefinite. Symmetrically, $\nabla^2_{yy} f(x^\star, y^\star)$ is negative semidefinite. From the assumption of boundedness of the set of optimizers and Lemma 24, we can further conclude that these matrices are positive definite and negative definite, respectively. Since $\mathbf{S} \succ 0$, every eigenvalue of $\mathbf{M} = \mathbf{S} + \mathbf{T}$ has positive real part. Indeed, let $\mathbf{M}z = \lambda z$ with $z \neq 0$. Then

$$\lambda\|z\|^2 = z^*\mathbf{M}z = z^*\mathbf{S}z + z^*\mathbf{T}z.$$

Because $\mathbf{S} \succ 0$, we have $z^*\mathbf{S}z > 0$. Because $\mathbf{T}^\top = -\mathbf{T}$, the scalar $z^*\mathbf{T}z$ is purely imaginary. Therefore

$$\mathrm{Re}(\lambda) = \frac{z^*\mathbf{S}z}{\|z\|^2} > 0.$$

Hence every eigenvalue of $\mathbf{M}$ can be written as $\lambda_i = a_i + ib_i$ with $a_i > 0$. The eigenvalues of $\mathbf{J}_{\mathrm{GDA}} = \mathbf{I} - \eta\mathbf{M}$ are $\rho_i = 1 - \eta\lambda_i$. Thus $|\rho_i|^2 = |1 - \eta(a_i + ib_i)|^2 = (1 - \eta a_i)^2 + \eta^2 b_i^2 = 1 - 2\eta a_i + \eta^2(a_i^2 + b_i^2)$. Therefore $|\rho_i| < 1$ whenever

$$0 < \eta < \frac{2a_i}{a_i^2 + b_i^2}.$$

Since there are finitely many eigenvalues, choosing

$$0 < \eta < \min_i \frac{2a_i}{a_i^2 + b_i^2}$$

gives $|\rho_i| < 1$ for all i. □

Lemma 24. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^2 in a neighborhood of an isolated minimum $x^\star$, with locally Lipschitz gradient, and satisfy the α-PŁ condition there. Then $\nabla^2_{xx} f(x^\star) \succeq \alpha I$.*

Proof. Let $F(x) := f(x) - f(x^\star)$. By PŁ and isolatedness of $x^\star$, f has quadratic growth near $x^\star$: there exists a neighborhood V of $x^\star$ such that

$$F(x) \geq \frac{\alpha}{2}\|x - x^\star\|^2, \qquad x \in V.$$

Fix a unit vector u. By the fundamental theorem of calculus,

$$f(x^\star + tu) - f(x^\star) = \int_0^t \langle \nabla f(x^\star + su), u\rangle \, ds,$$

and

$$\nabla f(x^\star + su) = \int_0^s \nabla^2 f(x^\star + ru)u\, dr.$$

Therefore

$$f(x^\star + tu) - f(x^\star) = \int_0^t \int_0^s u^\top \nabla^2 f(x^\star + ru)u\, dr\, ds.$$

Since $\nabla^2 f$ is continuous, for every $\varepsilon > 0$ there exists $\delta > 0$ such that, for all $|r| \le \delta$,

$$u^\top \nabla^2 f(x^\star + ru)u \le u^\top \nabla^2 f(x^\star)u + \varepsilon.$$

Hence, for $0 < |t| \le \delta$,

$$f(x^\star + tu) - f(x^\star) \le \frac{t^2}{2}\left(u^\top \nabla^2 f(x^\star)u + \varepsilon\right).$$

Combining this with quadratic growth gives

$$\frac{\alpha}{2}t^2 \le \frac{t^2}{2}\left(u^\top \nabla^2 f(x^\star)u + \varepsilon\right).$$

Dividing by $t^2/2$ yields

$$u^\top \nabla^2 f(x^\star)u \ge \alpha - \varepsilon.$$

Letting $\varepsilon \downarrow 0$ gives

$$u^\top \nabla^2 f(x^\star)u \ge \alpha.$$

Since u was arbitrary,

$$\nabla^2 f(x^\star) \succeq \alpha I.$$

□

Lemma 25. *Assume a function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ that is ℓ-smooth, globally PŁ, and the sets $\arg\min_x f(\cdot, y)$, $\arg\max_y f(x, \cdot)$ are bounded for all $y \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$ respectively. Then, f admits a unique saddle-point.*

Proof. By (Nejma, 2025, Theorem 2.1) we conclude that $\arg\min_x f(\cdot, y)$, $\arg\max_y f(x, \cdot)$ are singletons for all y and x in general. Then, aiming for contradiction, we assume that there exist two distinct saddle-points $(x^\star, y^\star)$ and (x', y'). By the standard fact that all saddle-point have the same value (see e.g., (Yang et al., 2020, Lemma 2.1)),

$$f(x^\star, y^\star) = f(x', y') = \min_x f(x, y^\star) = \max_y f(x^\star, y).$$

Therefore, $f(x^\star, y^\star) = f(x', y^\star)$ leading to a contradiction unless $(x^\star, y^\star) = (x', y')$. Therefore, the saddle-point is unique. □

AI USE STATEMENT

Generative AI tools were used for manuscript editing, theoretical proof assistance, and experiment execution. The authors verified the mathematical arguments, code, and reported results and take responsibility for the final work.