
Policy Gradient Methods Converge Globally in Imperfect-Information Extensive-Form Games

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Multi-agent reinforcement learning (MARL) has long been seen as inseparable
2 from Markov games (Littman, 1994). Yet, the most remarkable achievements of
3 practical MARL have arguably been in extensive-form games (EFGs)—spanning
4 games like Poker, Stratego, and Hanabi. At the same time, little is known about
5 provable equilibrium convergence for MARL algorithms applied to EFGs as they
6 stumble upon the inherent nonconvexity of the optimization landscape and the
7 failure of the value-iteration subroutine in EFGs. To this goal, we utilize con-
8 temporary advances in nonconvex optimization theory to prove that regularized
9 alternating policy gradient with (i) *direct policy parametrization*, (ii) *softmax policy*
10 *parametrization*, and (iii) *softmax policy parametrization with natural policy gra-*
11 *dient updates* converge to an approximate Nash equilibrium (NE) in the *last-iterate*
12 in imperfect-information perfect-recall zero-sum EFGs. Namely, we observe that
13 since the individual utilities are concave with respect to the sequence-form strat-
14 egy, they satisfy gradient dominance w.r.t. the behavioral strategy—or, *policy*, in
15 reinforcement learning terms. We exploit this structure to further prove that the
16 regularized utility satisfies the much stronger Polyak-Łojasiewicz condition. In
17 turn, we show that the different flavors of alternating policy gradient methods con-
18 verge to an ϵ -approximate NE with a number of iterations and trajectory samples
19 that are polynomial in $1/\epsilon$ and the natural parameters of the game. Our work is a
20 preliminary—yet principled—attempt in bridging the conceptual gap between the
21 theory of Markov and imperfect-information EFGs while it aspires to stimulate a
22 deeper dialogue between them.

23 1 Introduction

24 Reinforcement learning (RL) dominates contemporary applied and theoretical research. The flagship
25 of RL, *policy optimization methods*, appears to lend reasoning capabilities to language models (Shao
26 et al., 2024), defeats human Go world champions (Silver et al., 2016), and navigates real-world roads
27 safely (Lu et al., 2023). As is evident from even more examples (Vinyals et al., 2019; Schrittwieser
28 et al., 2020), machine gameplay has transformed by incorporating RL techniques in its algorithmic
29 arsenal. Although theoretical literature (Littman, 1994) posits that the canonical model of MARL
30 are Markov games (MGs), MARL has handled imperfect-information extensive-form games (EFGs)
31 with commendable success (Brown and Sandholm, 2019b; Bard et al., 2020; Perolat et al., 2022).

32 At first, the theory and practice of imperfect-information EFGs can seem saturated. Exhaustive
33 research in the properties of EFGs has exposed its convex structure using *sequence-form* strate-
34 gies (Koller et al., 1996; Von Stengel, 1996) and yielded the different counterfactual-regret minimiza-
35 tion algorithms (CFR) (Zinkevich et al., 2007; Tammelin, 2014; Brown and Sandholm, 2019a) that can
36 solve games using tabular policies with unmatched computational efficiency. Notwithstanding, these

techniques seem to hit a wall when faced with large-scale games whose size makes the use of tabular policies infeasible and calls for a *policy network* (or, more generally, policy function approximation). The picture is even more grave when CFR needs to be combined with model-free counterfactual value estimation. Its call for importance sampling yields a feedback of prohibitively high variance. Further, CFR’s average-iterate convergence makes the task of extracting a policy parametrization highly nontrivial. Since practitioners have extensively studied policy optimization for imperfect-information games (Lanctot et al., 2017; Srinivasan et al., 2018; Lockhart et al., 2019; Hennes et al., 2020; Rudolph et al., 2025) without offering guarantees of polynomial time convergence, we are naturally lead to the question:

*Do policy gradients methods provably converge to an equilibrium in
imperfect-information EFGs using a polynomial number of iterations and samples?* (♥)

To answer, we need to face the two obstacles that imperfect-information games raise against optimization, the failure of value iteration—which we sidestep by solely using policy gradients—and a highly nonconvex policy optimization landscape—which we prove that is benign.

Failure of value iteration In MARL for MGs, the overwhelming majority (Shapley, 1953; Wei et al., 2021; Zhao et al., 2022; Alacaoglu et al., 2022; Zhang et al., 2019) of existing algorithmic solutions for equilibrium learning or computation makes use of a *value iteration* subroutine or a *value critic*—which is in essence a backwards induction of the estimated value of the game. Whereas, solving imperfect-information games requires leveraging the opponent’s uncertainty about the underlying state. In other words, one needs to tradeoff exploiting private information and the benefit of keeping it secret. This precludes solving subtree-by-subtree conditioned on private information and leads to the emergence of behaviors such as bluffing at optimality.

Gradient Domination in Nonconvex Problems. Contemporary machine learning is arguably propelled by large-scale optimization of systems of astounding size to perform increasingly elaborate tasks. The corresponding objective functions are by no means convex in terms of parameters, which precludes theoretical guarantees of even reaching local optimum in a reasonable number of iterations (Murty and Kabadi, 1985). Yet, practice indicates a different reality and theory is gradually catching up. It has painstakingly been demonstrated that the nonconvexity of various ML optimization problems is seriously benign—more often than not, *stationarity implies global optimality*. Cases in point, gradient domination is exhibited for *the loss functions of overparametrized neural networks* (Liu et al., 2022a; Scaman et al., 2022), *the linear quadratic regulator* (Fazel et al., 2018), *value functions of Markov decision processes (MDPs)* (Agarwal et al., 2021; Bhandari and Russo, 2024), *matrix completion* (Ge et al., 2016), *dictionary learning* (Sun et al., 2015), and more. For a thorough discussion of gradient domination (Karimi et al., 2016; Li and Pong, 2018; Drusvyatskiy and Paquette, 2019; Drusvyatskiy and Lewis, 2018; Liao et al., 2024; Rebjock and Boumal, 2024). With the latter in mind, one could make the case that when game theory researchers pursuit equilibrium computation in general nonconvex games (Cai et al., 2024a; Angelopoulos et al., 2025) they set the bar too high. Still, the study of benign nonconvexity seems of great importance and rather underexplored (Yang et al., 2020; Mulvaney-Kemp et al., 2021; Vlatakis-Gkaragkounis et al., 2021; Sakos et al., 2023).

1.1 Contributions

We answer (♥) in the affirmative by developing three independent policy gradient methods (Theorems 3.1 to 3.3). All three algorithmic approaches lead to last-iterate convergence to a regularized NE of the EFG without requiring that the agents to share any information about their strategies by framing the resulting error as inexact gradient feedback (Devolder et al., 2014). The only “communication” that takes place is, taking turns in updating the policies and agreeing to pick different stepsizes. We, namely, contribute,

1. alternating policy gradient with *direct parametrization* and ℓ_2 -norm regularization
2. alternating policy gradient with *softmax parametrization* and *entropy regularization*
3. alternating *natural policy gradient* with *softmax parametrization* and *entropy regularization*

and prove their last-iterate convergence. On a sidenote, we offer a sharper dependence of the PŁ modulus to the hidden convexity modulus than the one suggested by (Karimi et al., 2016, Appendix G) for constrained optimization.

1.2 Overview of Techniques

The theoretical guarantees for our three algorithmic solutions are pinpointed by a simple unifying conceptual principle. That is, *the nonconvex optimization problem* of computing an equilibrium by directly optimizing the behavioral strategies (or, policies) *is a constrained two-sided PL optimization problem* where alternating gradient descent ascent is known to converge. Namely, we show that the optimization landscape viewed in terms of *policies* is nonconvex in a rather benign way; the utility is *hidden concave*. In particular, after appropriate regularization, each utility function satisfies a strong gradient domination property, *i.e.*, the proximal Polyak-Łojasiewicz.

Hidden concavity. Going into more detail, utilities in EFGs are concave in terms of *sequence-form* strategies. An appropriate *regularizer* enhances convexity to strong concavity. Moreover, enforcing a positive lower bound on the probability of reaching every information set yields a uniform Lipschitz constant for the bijection that maps sequence-form strategies to behavioral policies. Taken together, these two observations imply a strong gradient-domination condition for each player’s policy.

PL condition. For the sake of offering an intuitive exposition, we forego the nuances of constrained optimization to explain how the PL condition is proven to hold. We say that an optimization problem $\min_x f(x)$ exhibits *hidden strong convexity* when there exists an invertible mapping $u = c(x)$ and a function $H(u)$ that is μ -strongly convex in u and $f(x) = H(c(x))$. Strong convexity implies that $f(x) - f^* \equiv H(u) - H^* \leq \frac{1}{2\mu} \|\nabla_u H(u)\|^2$. Now, a bounded Lipschitz modulus $L_{c^{-1}} > 0$ of the inverse transform, $c^{-1}(u) = x$, leads to the PL inequality $f(x) - f^* \leq \frac{1}{2\mu L_{c^{-1}}^2} \|\nabla f(x)\|^2$ by merely applying the chain rule of differentiation. Similar arguments work for the proximal-PL.

Convergence. Then, *alternating gradient descent ascent* on $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$,

$$x_{t+1} \leftarrow \text{Proj}_{\mathcal{X}} [x_t - \eta_x \nabla_x f(x_t, y_t)]; \quad y_{t+1} \leftarrow \text{Proj}_{\mathcal{Y}} [y_t + \eta_y \nabla_y f(x_{t+1}, y_t)],$$

is proven to converge to a saddle-point point using a typical Lyapunov function argument. We tune the stepsizes η_x, η_y in such a way that one player learns faster than the other. Since the function is PL, this means that after each update the optimizer is significantly approximated. Intuitively, after enough iterations, the update scheme can be viewed as optimizing for $\Phi(x) := \max_{y \in \mathcal{Y}} f(x, y)$ as $x_{t+1} \approx \text{Proj}_{\mathcal{X}} (x_t - \eta_x \nabla_x \Phi(x_t))$. Crucially, our convergence analysis sets aside the usual regret minimization arguments that are used to either prove *average-iterate* or *best-iterate* convergence (*e.g.*, Anagnostides et al. (2022); Liu et al. (2024)).

1.3 Comparison to Related Work

We point out two particular results (Sokota et al., 2022; Liu et al., 2024) directly related to our endeavor of policy gradient/optimization methods for incomplete-information EFGs. Although the magnetic mirror descent method proposed in (Sokota et al., 2022) does not come with guarantees in EFGs, it exhibits impressive empirical performance. (Liu et al., 2024) lays the foundation of our approach as it introduces the *bidilated regularizer* although it does not offer a convergence guarantee that is polynomial in the parameters of the game and $1/\epsilon$.

	Altern./Simult. Updates	Provable Convergence	Regularization	Feedback
(Liu et al., 2024)	simultaneous	yes, best-iterate*	bidilated	CFR, Q, \bar{Q}
(Sokota et al., 2022)	simultaneous	no	policy entropy	Q
Ours	alternating	yes, last-iterate, polynomial time	bidilated	$\nabla_{\theta} V, Q$

Table 1: Comparison of policy gradient/optimization methods.

CFR, $Q, \bar{Q}, \nabla_{\theta} V$ stand for counterfactual value, action-value, traject. action-value, and policy gradient.

* Guarantees are pseudo-polynomial in the game-size.

Our work follows arguments utilized in the context of policy gradient methods for Markov decision processes (MDPs) and MGs. Namely, we use arguments from (Mei et al., 2020; Cen et al., 2022a) as

the entropic bidilated regularizer is almost identical to discounted entropy. Further, we use arguments from (Zhang et al., 2021) to show that the mapping from sequence-form strategies to policies is Lipschitz continuous. Further, our analysis bears similarity to that of (Kalogiannis et al., 2024).

1.4 A comparison of Markov and imperfect-information extensive-form games

Imperfect-information extensive-form games (EFGs) and Markov games (MGs) both model multi-stage strategic interaction. They differ sharply in what each player can observe while they maintain marked similarities in the way strategies are represented (*behavioral strategies* and *policies*), the *hidden concave* representation of utilities (concavity w.r.t. *sequence-form strategies* and *occupancy measures*), and regularization choices for optimization. The table and discussion below summarize this comparison along the axes of observability, strategy space, utility convex reformulation, regularization and optimization landscape.

	Game State	Observable State	Control Variables	Utility Concave In
EFG	History $h \in \mathcal{T}$	Info set $s \in \mathcal{S}$	Behavioral Strategy $\pi(\cdot s)$	Sequence-form Strategy μ^π
	<i>each a node of game tree graph \mathcal{T}</i>	<i>each a disjoint set of multiple histories h</i>	<i>distribution over actions at info set s</i>	<i>independent of opponents' strategies</i>
MG	State s		Markovian Policy $\pi(\cdot s)$	State-action Occupancy measure λ^π
	<i>fully observable by all players potentially recurring in the finite or infinite horizon of the game</i>		<i>distribution over actions at state s</i>	<i>depends on opponents' policies</i>

Table 2: Imperfect-information extensive-form games (EFG) vs. Markov games (MG).

Clearly, an *info set* (information set) in an imperfect-information EFG is to a behavioral strategy what a state is to a policy in an MG. However, imperfect information (or *partial observability*) leads to a discrepancy between the expected return of an info set in an EFG and the expected return state in an MG as highlighted in (Nayyar et al., 2013; Sokota et al., 2023). Interestingly, the concave reparametrization of EFG utilities exhibits a structure more favorable than the corresponding one in MGs. In particular, the utility is concave in sequence-form strategies of EFGs and the latter depend solely on a player’s own behavioral strategy. This comes in stark contrast to the state-action occupancy measure of MGs which need to satisfy *Bellman flow* constraints conditioned on opponents’ strategies.

Finally, similarities of the regularization techniques in EFGs and MGs are cornerstone to our work. The EFG entropic *bidilated regularizer* (Liu et al., 2024), \mathcal{R} , and the very commonly used MDP discounted entropy (Mei et al., 2020; Cen et al., 2022a,b), \mathcal{E} , are virtually identical. We note that, in EFGs a regularizer is mostly used in context of directly optimizing in the sequence-form space. They induce a distance generating function of mirror descent instantiations. Some more recent works have used it to make the game strongly-monotone and guarantee convergence of gradient descent methods (Liu et al., 2022b). Liu et al. (2024), in the context of policy optimization, define the bidilated regularizer whose policy gradients can be estimated without importance sampling. Illustratively, the two regularizers read side-by-side (γ is a discount factor of MDPs):

$$\mathcal{R}(\pi) := \mathbb{E}_{\tau \sim \pi} \left[\sum_{s_{(k)} \in \tau} \psi(\pi(\cdot | s_{(k)})) \right] \quad \Bigg| \quad \mathcal{E}(\pi) := \mathbb{E}_{\tau \sim \pi} \left[\sum_k^H \gamma^{k-1} \psi(\pi(\cdot | s_{(k)})) \right].$$

2 Preliminaries

In this section we introduce the key ingredients required for our analysis. For EFGs, we highlight how the utility is expressed as a concave function of the sequence-form strategies. We also review the—Euclidean or entropic—bidilated regularizer whose strong convexity underpins our gradient-domination arguments. With regards to RL theory, we recall the definition of the value and action-value functions and show that trajectory samples, or *roll-outs*, give unbiased Monte-Carlo

estimates of both the utility and the bidilated regularizer via the (REINFORCE) estimator (Williams, 1992; Sutton et al., 1999). Finally, we review the optimization notions of hidden concavity and gradient dominance, used to prove convergence in of our algorithmic solutions.

2.1 Extensive-Form Games

We briefly go over the definition of an EFG and move on to the sequence-form strategies and the corresponding regularizers.

Definition 1 (EFG). *A two player zero-sum extensive-form game, Γ , is defined by the tuple $(\mathcal{T}, \mathcal{H}, \mathcal{S}, \mathcal{A}, \mathcal{B}, r)$. A special chance player, c , models uncontrollable randomness while,*

- \mathcal{T} is a rooted game tree of height $D(\mathcal{T})$,
- $\mathcal{H} := \mathcal{H}_1 \cup \mathcal{H}_2 \cup \mathcal{H}_c$ is the set of \mathcal{T} 's nodes, referred to as histories. Each history, h , belongs to exactly one of the sets $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_c$ depending on the player responsible for taking action at h .
- $\mathcal{S} := \mathcal{S}_1 \cup \mathcal{S}_2$ is a finite set of information sets (infosets). The infosets partition histories, \mathcal{H}_i , of the acting player i into sets of nodes that are indistinguishable.
- $\mathcal{A} := \{\mathcal{A}_s\}_{s \in \mathcal{S}_1}, \mathcal{B} := \{\mathcal{B}_s\}_{s \in \mathcal{S}_2}$ are the action sets of player 1 and 2, respectively. Each infoset $s \in \mathcal{S}$ has a corresponding set of actions \mathcal{A}_s , and respectively \mathcal{B}_s . Further, we will denote $A_s := |\mathcal{A}_s|$, $A_{\max} := \max_s A_s$ and $B_s := |\mathcal{B}_s|$, $B_{\max} := \max_s B_s$.
- $r : \mathcal{H} \rightarrow [-1, 1]$ is a payoff function mapping leaves of \mathcal{T} to a payoff for player 1; player 2 gets the opposite payoff.

A perfect recall assumption is made, ensuring that players remember their past observations and actions. This implies that nodes in the same infoset have the same past observation sequence. We will use $\sigma_1(s), \sigma_2(s)$ to denote the last parent infoset-action pair (s', a') , $s \in \mathcal{S}_1$ and (s', b') , $s \in \mathcal{S}_2$ encountered when descending from the game tree's root to history h . $\sigma_1(\cdot), \sigma_2(\cdot)$ are either unique for non-root nodes or the null set for the root. We will overload notation $\sigma_1(h)$ to mean $\sigma_1(s)$ for the infoset s where h belongs (resp. for $\sigma_2(h)$).

Sequence-Form Strategies A player's behavioral strategy is a probability distribution over actions at each of their infosets. In *sequence-form*, the strategy of player 1 is defined as:

$$\mu_1^{\pi_1}(s, a) := \mu_1^{\pi_1}(\sigma(s))\pi_1(a|s) \quad \forall s \in \mathcal{S}_1, \forall a \in \mathcal{A}_s.$$

The sequence-form strategy of player 2 is defined in a symmetric fashion. Introduced in (Romanovskii, 1962; Von Stengel, 1996; Koller et al., 1996), sequence-form strategies are generalizations of simplices and express the sequential structure of an EFG. The set of sequence-form strategies, $\mathcal{M}_1, \mathcal{M}_2$ are convex polytopes as they are defined only by linear equalities and non-negativity constraints. The chance player's contribution to the probability of reaching history h is given by $\mu_c(h)$ and it is assumed to be strictly positive for reachable nodes. For player 1, the expected utility is given by the bilinear form:

$$V^{\pi_1, \pi_2} := (\mu_1^{\pi_1})^\top \mathbf{R} \mu_2^{\pi_2},$$

where \mathbf{R} is the matrix representation of payoff function r .

Forward, we will refer to behavioral strategies as policies which will be denoted as π_1, π_2 . The solution concept we are after is an ϵ -approximate Nash equilibrium.

Definition 2 (ϵ -NE). *A policy profile π_1^*, π_2^* is an ϵ -approximate Nash equilibrium of an EFG Γ , if, for any policies π_1 and π_2 it holds true that,*

$$V^{\pi_1, \pi_2^*} - \epsilon \leq V^{\pi_1^*, \pi_2^*} \leq V^{\pi_1, \pi_2^*} + \epsilon.$$

The bidilated regularizer. Introduced in (Liu et al., 2024), the unweighted *bidilated regularizer* is defined using a strongly-convex regularizer $\psi(\cdot)$ and summed multiplied by the total reach probability of each infoset. Since it depends on both players' policies we write $\mathcal{R}(\pi_1, \pi_2), \mathcal{R}(\pi_2, \pi_2)$, with

$$\mathcal{R}_1(\pi_1, \pi_2) := \mathbb{E}_{h \sim \pi} \left[\sum_h \psi(\pi_1(\cdot|h)) \right] \quad \text{and} \quad \mathcal{R}_2(\pi_1, \pi_2) := \mathbb{E}_{h \sim \pi} \left[\sum_h \psi(\pi_2(\cdot|h)) \right].$$

2.2 RL fundamentals

Moving on, we define the value, action-value, and advantage functions in the context of EFGs. Inspired by the occupancy measure of MGs, we define the history occupancy measure d^π for a given policy profile $\pi := (\pi_1, \pi_2)$ which simply is the reach probability of each history and comes in handy as a shorthand notation in the description of the algorithms and their analysis. Moreover, we recall the definitions of direct and softmax policy parametrization. Last but not least, we demonstrate how the (REINFORCE) gradient estimator computes policy gradients for EFGs for both the unregularized and regularized utility.

Value, action-value, and advantage functions. Without loss of generality, we assume that players get a payoff only on a terminal history \bar{h} . This way we can define the *value function* of an infoset s , as the expected utility if the game were to start at a history h_0 belonging to s ,

$$V^\pi(s) := \mathbb{E}_{\bar{h} \sim \pi} [r(\bar{h}) | h_0 \in s]$$

In a similar vein, we define the action-value function, or Q , as the expected utility if the game started at a history h_0 belonging in s and after the player had taken action a_0 , (or, resp. b_0),

$$Q_1^\pi(s, a) := \mathbb{E}_{\bar{h} \sim \pi} [r(\bar{h}) | h_0 \in s, a_0 = a] \quad \text{and} \quad Q_2^\pi(s, b) := \mathbb{E}_{\bar{h} \sim \pi} [r(\bar{h}) | h_0 \in s, b_0 = b].$$

Finally, the advantage function is defined for each player as the difference between an action-value and the infoset's value $A_1^\pi(s, a) := V^\pi(s) - Q_1^\pi(s, a)$ and $A_2^\pi(s, b) := V^\pi(s) - Q_2^\pi(s, b)$. Similar to the state occupancy measure of an MG, we can define the history occupancy measure $d^\pi : \mathcal{H} \rightarrow [0, 1]$ which is defined as, $d^\pi(h) := \mathbb{E}_{h' \sim \pi} [\mathbb{1}\{h' = h\}]$. Overloading notation, for an infoset $s \in \mathcal{S}$ $d^\pi(s) := \sum_{h \in s} d^\pi(h)$.

Policies. Policies are precisely parametrized behavioral strategies. We will consider two parametrizations of policies, (i) *direct parametrization*, and (ii) *softmax parametrization*. For directly parametrized policies, we denote the parameters as x, y which are $x \in \times_{s \in \mathcal{S}_1} \Delta(\mathcal{A}_s), y \in \times_{s \in \mathcal{S}_2} \Delta(\mathcal{B}_s)$.

The parameters of softmax policies will be denoted χ, θ with $\chi \in \mathbb{R}^A, A = \sum_s A_s$ and $\theta \in \mathbb{R}^B, B = \sum_s B_s$. As a reminder, the policy is generated through the softmax transform,

$$\pi_\chi(a|s) = \frac{\exp(\chi_{s,a})}{\sum_{a'} \exp(\chi_{s,a'})} \quad \text{and} \quad \pi_\theta(b|s) = \frac{\exp(\theta_{s,b})}{\sum_{b'} \exp(\theta_{s,b'})}.$$

Gradient estimation with REINFORCE. The ability to estimate a gradient of the value function using trajectory samples, or *roll-outs*, has endowed the theory and practice of RL with the rich toolbox of gradient-based optimization. In fact, the (REINFORCE) gradient estimator (Williams, 1992; Sutton et al., 1999) is also an unbiased estimator of the policy gradient in the EFG setting, and thus provides a sound foundation for our analysis.

Definition 3 (REINFORCE). Let τ denote a trajectory of infosets and actions sampled by implementing policies π_1, π_2 , $\tau := (s_{(1)}, a_{(k)}, \dots)$. We define REINFORCE, $(\hat{\nabla}_x, \hat{\nabla}_y)$, to be the stochastic gradient estimators:

$$\hat{\nabla}_x = r_\tau \sum_{k=1}^{K_\tau} \nabla_x \log \pi_x(a_{(k)} | s_{(k)}) \quad \text{and} \quad \hat{\nabla}_y = r_\tau \sum_{k=1}^{K_\tau} \nabla_y \log \pi_y(b_{(k)} | s_{(k)}). \quad (\text{REINFORCE})$$

The addition of regularization, leads to the definition a regularized value function, V_τ ,

$$V_\tau^\pi(s) := \mathbb{E}_{\bar{h} \sim \pi} [r(\bar{h}) + \tau \sum_h [\psi(\pi_1(\cdot | h)) + \psi(\pi_2(\cdot | h))] | h_0 \in s].$$

Regularized Q -value and advantage functions Q_τ^π, A_τ^π are defined accordingly (see Appendix C.2). Furthermore, (REINFORCE) can be minimally modified to estimate the policy gradient of the regularized value function without importance sampling (discussed in detail in Appendix F.1).

230 **Sufficient Exploration.** Following the policy parameterization rule (γ -trunc) defined in (Liu et al.,
 231 2024), every info set is visited with probability at least $\gamma > 0$. Besides ensuring the usual *sufficient*
 232 *exploration* condition for MDPs (Assumption 1), this lower bound also serves a purpose similar to a
 233 bounded mismatch coefficient in the sense of (Agarwal et al., 2021; Daskalakis et al., 2020).

234 **Assumption 1** (Sufficient exploration). *Both players, every action a_i in every info set is played with*
 235 *a probability s_i is played by at least some probability $\gamma_{s,a}, \gamma_{s,b} > 0$,*

$$\pi_1(a|s) \geq \gamma_{s,a}, \forall s \in \mathcal{S}_i, a \in \mathcal{A}_s \quad \text{and} \quad \pi_2(b|s) \geq \gamma_{s,b}, \forall s \in \mathcal{S}_i, b \in \mathcal{B}_s. \quad (\gamma\text{-trunc})$$

236 We denote by $\gamma > 0$ the lower bound of reaching any info set of the EFG.

237 Guaranteeing that Assumption 1 holds is straightforward for directly parametrized policies. The
 238 players need to pick policies x, y , from the cartesian product of appropriately truncated simplices, to
 239 be denoted $\mathcal{X}^\gamma, \mathcal{Y}^\gamma$ respectively.

240 As for softmax parametrized policies, (γ -trunc) is achieved when both players' parameters are
 241 restricted to the polytopes X_R, Θ_R . To demonstrate, X_R is defined in the following manner, $X_R :=$
 242 $\left\{ \chi \in \mathbb{R}^A, A = \sum_s A_s : \chi_s^\top \mathbf{1} = 0, \forall s \in \mathcal{S}_1, |\chi_{s,i} - \chi_{s,j}| \leq 2R, \forall i, j \in [A_s] \right\}$, and the definition
 243 of Θ_R follows suit. We highlight that the images of X_R, Ψ_R under the softmax map are convex sets
 244 (Lemma D.4) and we will denote the resulting truncated policy sets as Π_1^R, Π_2^R .

245 2.3 Hidden concavity and gradient domination

246 In this subsection, we define the two key backbone concepts of hidden concavity and gradient
 247 domination. Gradient domination of a weak or strong form has been extensively investigated in
 248 the theory of RL and MARL (Bhandari and Russo, 2024; Agarwal et al., 2021; Mei et al., 2020;
 249 Zhang et al., 2019; Daskalakis et al., 2020). Simply put, the nonconvex value function satisfies a
 250 gradient-domination property and any stationary point is globally optimal. Thus, any guarantee of
 251 convergence to a stationary point is elevated to a guarantee of convergence to global optimality.

252 **Definition 4** (Hidden convexity). *A nonconvex function $f : \mathcal{X} \rightarrow \mathbb{R}$ defined over the set \mathcal{X} is*
 253 *said to be hidden (strongly) convex if there exists (i) a bijective mapping $c : \mathcal{X} \rightarrow \mathcal{U}$ for some*
 254 *convex set \mathcal{U} ; (ii) a function $H : \mathcal{U} \rightarrow \mathbb{R}$ that is strongly convex with modulus $\alpha_H \geq 0$; such that*
 255 *$f(x) = H(c(x)), \forall x \in \mathcal{X}$.*

256 When the Lipschitz continuity modulus of the inverse transform c^{-1} , is uniformly bounded it implies
 257 the gradient domination condition as shown in (Fatkhullin et al., 2023, Prop. 2) coupled with (Karimi
 258 et al., 2016, App. G).

Definition 5 (pPLcondition (Karimi et al., 2016)). *Assume $F : \mathbb{R}^d \rightarrow \mathbb{R}$ defined as $F(x) :=$*
 $f(x) + g(x)$. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be an ℓ -smooth function and $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex. Define

$$\mathcal{D}_g(x, \ell) := -2\ell \min_z \left\{ \langle \nabla f(x), z - x \rangle + \frac{\ell}{2} B(z|x) + g(z) - g(x) \right\}.$$

for a choice of Bregman divergence $B(\cdot|\cdot)$. We say that F satisfies the pPL condition with modulus
 $\alpha > 0$ if, for every $x \in \mathcal{X}$,

$$\frac{1}{2} \mathcal{D}_g(x, \ell) \geq \alpha (F(x) - F^*),$$

259 where $F^* = \min_x F(x)$. When g is the indicator function of a set \mathcal{X} we write $\mathcal{D}_{\mathcal{X}}(x, \ell)$.

260 For our first contribution, we establish that the utility of an imperfect-information EFG under different
 261 policy parametrizations is pPL with regards to the policy.

262 **Lemma 2.1** (pPL for EFG; restated from Lemmata E.1 to E.3). *Let an imperfect-information EFG,*
 263 *Γ , perturbed with the pair of bidilated regularizers $(\mathcal{R}_1, \mathcal{R}_2)$ with a coefficient $\tau > 0$. Then, each*
 264 *player's utility satisfies the pPL condition with a modulus $\alpha = \tau \times \text{poly}\left(\frac{1}{\gamma}, A_{\max}, B_{\max}, \frac{1}{2D(\tau)}\right)$.*

265 3 Convergence of Alternating Regularized Policy Gradient

266 Having established the required background and notation, we are ready to present our main results.
 267 In Theorem 3.1 we show the convergence of simple alternating regularized policy gradient to an

approximate NE in the last iterate. Moving to Theorem 3.2, we prove a similar result for softmax-parametrized policies. Finally, we analyze *alternating regularized natural policy gradient* through a mirror-descent lens, demonstrate its relationship to multiplicative weight updates of the policies, and prove its convergence to an approximate NE in the last iterate (Theorem 3.3).

Throughout, η_x, η_y denote the stepsizes and $\hat{\nabla}^\tau$ denotes the (REINFORCE) gradient estimate of the utility w.r.t. to a player’s parameters accounting only for their own regularization term.

3.1 Direct Policy Parametrization

The first result we present is the a simple policy gradient scheme with alternating updates and a Euclidean regularizer. The parameter updates of alternating regularized policy gradient takes the following form,

$$\begin{aligned} x_{t+1} &= \text{Proj}_{\mathcal{X}^\gamma} \left[x_t - \eta_x \hat{\nabla}_x^\tau(x_t, y_t) \right] \\ y_{t+1} &= \text{Proj}_{\mathcal{Y}^\gamma} \left[y_t + \eta_y \hat{\nabla}_y^\tau(x_{t+1}, y_t) \right]. \end{aligned} \quad (\text{Alt-RegPG})$$

where $\text{Proj}_{\mathcal{X}^\gamma}, \text{Proj}_{\mathcal{Y}^\gamma}$ denote the Euclidean projection of the parameters to the truncated simplices dictated by (γ -trunc). We state our first convergence theorem which settles question (♥) and defer its formal statement to the Appendix G.1.

Theorem 3.1 (Informal; restated from Thm. G.1). *With direct policy parametrization and the Euclidean bidilated regulariser, alternating policy-gradient algorithm attains a last-iterate ϵ -Nash equilibrium in*

$$T = \text{poly} \left(\frac{1}{\epsilon}, \frac{1}{\gamma}, A_{\max}, B_{\max}, 2^{D(\mathcal{T})} \right) \text{ iterations,}$$

using batches of $\text{poly} \left(\frac{1}{\epsilon}, \frac{1}{\gamma}, A_{\max}, B_{\max}, 2^{D(\mathcal{T})} \right)$ trajectory samples at each step.

Remark 1. *We note that the exponential dependence on $D(\mathcal{T})$ is still polynomial in the game size as the height has itself logarithmic dependence in size of the game.*

3.2 Softmax Policy Parametrization

We move on to convergence under softmax parametrization and entropic regularization. This choice of parametrization is an important step towards getting provable guarantees for policy gradient methods in imperfect-information EFGs using function approximation (e.g. neural networks). The projection to X_R, Θ_R guarantees that Equation (γ -trunc) is satisfied,

$$\begin{aligned} \chi_{t+1} &= \text{Proj}_{X_R} \left[\chi_t - \eta_x \hat{\nabla}_\chi^\tau(\chi_t, \theta_t) \right] \\ \theta_{t+1} &= \text{Proj}_{\Theta_R} \left[\theta_t + \eta_y \hat{\nabla}_\theta^\tau(\chi_{t+1}, \theta_t) \right] \end{aligned} \quad (\text{Alt-EntRegPG})$$

Theorem 3.2 (Informal; restated from Thm. G.2). *Alternating policy-gradient algorithm with softmax policy parametrization and the entropic bidilated regulariser, converges in expectation in the last-iterate to an ϵ -Nash equilibrium after a number of iterations T , that is*

$$T = \text{poly} \left(\frac{1}{\epsilon}, \frac{1}{\gamma}, A_{\max}, B_{\max}, 2^{D(\mathcal{T})} \right) \text{ iterations,}$$

using batches of $\text{poly} \left(\frac{1}{\epsilon}, \frac{1}{\gamma}, A_{\max}, B_{\max}, 2^{D(\mathcal{T})} \right)$ trajectory samples at each step.

3.3 Natural Policy Gradient

Finally, we consider the natural policy gradient algorithm (Kakade, 2001) which is an adaptation of natural gradient (Amari, 1998). This algorithm is of particular interest due to its intimate connection to the TRPO, PPO (Schulman et al., 2015, 2017) policy optimization algorithms. Natural policy

300 gradient uses a *Fisher information matrix* induced by the policy as a preconditioner for policy gradient
 301 updates:

$$\mathbf{F}_\chi(\chi, \theta) := \sum_s d^{\chi, \theta}(s) \sum_a \pi_\chi(a|s) \nabla \log \pi_\chi(a|s) [\nabla \log \pi_\chi(a|s)]^\top$$

302 We cast *natural policy gradient* steps as *mirror descent steps* with a Mahalanobis norm induced by
 303 the Fisher information matrix (for a more nuanced discussion on this connection see (Raskutti and
 304 Mukherjee, 2015)).

$$\begin{aligned} \chi_{t+1} &= \arg \min_{\chi \in X_R} \langle \nabla_\chi V(\chi_t, \theta_t), \chi - \chi_t \rangle + \frac{1}{2\eta_x} \|\chi - \chi_t\|_{\mathbf{F}_\chi(\chi_t, \theta_t)}^2 \\ \theta_{t+1} &= \arg \min_{\theta \in \Theta_R} \langle \nabla_\theta V(\chi_{t+1}, \theta_t), \theta - \theta_t \rangle + \frac{1}{2\eta_y} \|\theta - \theta_t\|_{\mathbf{F}_\theta(\chi_{t+1}, \theta_t)}^2 \end{aligned}$$

305 The update scheme can be equivalently written as:

$$\begin{aligned} \chi_{t+1} &= \arg \min_{\chi \in X_R} \left\| \chi_t - \eta_x \mathbf{F}_\chi^\dagger(\chi_t, \theta_t) \nabla_\chi V(\chi_t, \theta_t) - \chi \right\|_{\mathbf{F}_\chi(\chi_t, \theta_t)}^2 \\ \theta_{t+1} &= \arg \min_{\theta \in \Theta_R} \left\| \theta_t + \eta_y \mathbf{F}_\theta^\dagger(\chi_{t+1}, \theta_t) \nabla_\theta V(\chi_{t+1}, \theta_t) - \theta \right\|_{\mathbf{F}_\theta(\chi_{t+1}, \theta_t)}^2 \end{aligned} \quad (\text{Alt-RegNPG})$$

306 More importantly, we note that in policy space, the update scheme of natural policy gradient takes a
 307 very simple form which, as expected, reads, for player 1 (\odot is element-wise multiplication):

$$\begin{aligned} \bar{\pi}_{1,t+1}(\cdot|s) &\propto \pi_{1,t}(\cdot|s)^{1-\eta_x \tau} \odot \exp(\eta_x Q_\tau^{\pi_t}(s, \cdot)) \\ \pi_{1,t+1}(\cdot|s) &\approx \arg \min_{\pi \in \Pi_1^R} \text{KL}(\pi(\cdot|s) \| \bar{\pi}_{1,t+1}(\cdot|s)) \end{aligned}$$

308 To see why the second approximate equality holds, we note that the Mahalanobis distance over
 309 the parameters induced by the Fisher information matrix of the softmax policy, is a second-order
 310 approximation of policy KL. The derivation and an extensive discussion are deferred to Appendix G.3.

311 **Theorem 3.3** (Informal; restated from Thm. G.3). *For an appropriate tuning of $\eta_x, \eta_y > 0$, the*
 312 *last-iterate of alternating regularized natural policy gradient (Alt-RegNPG) converges in expectation*
 313 *to an ϵ -approximate Nash equilibrium in a number of iterations T that is:*

$$T = \text{poly} \left(\frac{1}{\epsilon}, \frac{1}{\gamma}, A_{\max}, B_{\max}, 2^{D(\mathcal{T})} \right).$$

314

315 4 Conclusion

316 We studied three different policy gradient methods for imperfect-information perfect-recall zero-sum
 317 EFGs under a unifying optimization principle. We managed to provide the first global last-iterate
 318 convergence guarantees of independent policy gradient methods to an ϵ -approximate Nash equilibrium.
 319 Furthermore, our analysis requires a number of iterations and samples that is polynomial in $1/\epsilon$ and
 320 the parameters of the game. To do so, we exploited the favorable properties (PL-condition) of the
 321 otherwise nonconvex optimization landscape. We departed from the usual route of regret analysis in
 322 EFGs and opted for more conventional convergence analysis arguments. We hope to motivate further
 323 exchange between theoretical MARL research and the theory of EFGs as we strongly believe in the
 324 potential this communication fosters.

325 **Future directions.** On another note, as our main concerns was proving polynomial time con-
 326 vergence of policy gradient in EFGs, our analysis is at places loose. We firmly believe that the
 327 convergence rates and constant dependencies can be improved, *e.g.*, by using the machinery of
 328 treeplex norms (Fan et al., 2024), relatively-smooth optimization (Lu et al., 2018; Mei et al., 2021;
 329 Fatkhullin and He, 2024), and other policy optimization arguments (Zhan et al., 2023; Cen et al.,
 330 2022b). Being more particular, we would like to see guarantees that do not call for mini-batching and
 331 possibly use variance reduction techniques. Moreover, fundamental questions about the limit points
 332 of policy gradient methods in EFGs —similar to those of (Giannou et al., 2022) for MGs—are open.
 333 More broadly, do forms of benign nonconvexity (like hidden convexity) refine the results of (Cai
 334 et al., 2024b; Angelopoulos et al., 2025)?

References

- Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.
- Ahmet Alacaoglu, Luca Viano, Niao He, and Volkan Cevher. A natural actor-critic framework for zero-sum Markov games. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 307–366. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/alacaoglu22a.html>.
- Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.
- Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. On last-iterate convergence beyond zero-sum games. In *International Conference on Machine Learning*, pages 536–581. PMLR, 2022.
- Anastasios N Angelopoulos, Michael I Jordan, and Ryan J Tibshirani. Gradient equilibrium in online learning: Theory and applications. *arXiv preprint arXiv:2501.08330*, 2025.
- 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.
- Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *Operations Research*, 72(5):1906–1927, 2024.
- Noam Brown and Tuomas Sandholm. Solving imperfect-information games via discounted regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1829–1836, 2019a.
- Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019b.
- Yang Cai, Constantinos Daskalakis, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. On tractable phi-equilibria in non-concave games. *arXiv preprint arXiv:2403.08171*, 2024a.
- Yang Cai, Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Weiqiang Zheng. Fast last-iterate convergence of learning in games requires forgetful algorithms. *arXiv preprint arXiv:2406.10631*, 2024b.
- Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578, 2022a.
- 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*, 2022b.
- 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.
- Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $\mathcal{O}(\frac{1}{k})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018.
- Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146:37–75, 2014.
- Dmitriy Drusvyatskiy and Adrian S Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.
- Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178:503–558, 2019.

- 382 Zhiyuan Fan, Christian Kroer, and Gabriele Farina. On the optimality of dilated entropy and lower
383 bounds for online learning in extensive-form games. *arXiv preprint arXiv:2410.23398*, 2024.
- 384 Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Optimistic regret minimization for extensive-
385 form games via dilated distance-generating functions. *Advances in neural information processing*
386 *systems*, 32, 2019.
- 387 Ilyas Fatkhullin and Niao He. Taming nonconvex stochastic mirror descent with general bregman
388 divergence. In *International Conference on Artificial Intelligence and Statistics*, pages 3493–3501.
389 PMLR, 2024.
- 390 Ilyas Fatkhullin, Niao He, and Yifan Hu. Stochastic optimization under hidden convexity. *arXiv*
391 *preprint arXiv:2401.00108*, 2023.
- 392 Maryam Fazel, Rong Ge, Sham M. Kakade, and Mehran Mesbahi. Global convergence of policy
393 gradient methods for the linear quadratic regulator. In *International Conference on Machine*
394 *Learning*, 2018. URL <https://api.semanticscholar.org/CorpusID:51881649>.
- 395 Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *Advances*
396 *in neural information processing systems*, 29, 2016.
- 397 Angeliki Giannou, Kyriakos Lotidis, Panayotis Mertikopoulos, and Emmanouil-Vasileios Vlastakis-
398 Gkaragkounis. On the convergence of policy gradient methods to nash equilibria in general
399 stochastic games. *Advances in Neural Information Processing Systems*, 35:7128–7141, 2022.
- 400 Daniel Hennes, Dustin Morrill, Shayegan Omidshafiei, Rémi Munos, Julien Perolat, Marc Lanctot,
401 Audrunas Gruslys, Jean-Baptiste Lespiau, Paavo Parmas, Edgar Duñez Guzmán, and Karl Tuyls.
402 Neural replicator dynamics: Multiagent learning via hedging policy gradients. In *Proceedings*
403 *of the 19th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS*
404 *’20*, page 492–501, Richland, SC, 2020. International Foundation for Autonomous Agents and
405 Multiagent Systems. ISBN 9781450375184.
- 406 Samid Hoda, Andrew Gilpin, Javier Pena, and Tuomas Sandholm. Smoothing techniques for
407 computing nash equilibria of sequential games. *Mathematics of Operations Research*, 35(2):
408 494–512, 2010.
- 409 Sashank J Reddi, Suvrit Sra, Barnabas Póczos, and Alexander J Smola. Proximal stochastic methods
410 for nonsmooth nonconvex finite-sum optimization. *Advances in neural information processing*
411 *systems*, 29, 2016.
- 412 Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14,
413 2001.
- 414 Fivos Kalogiannis, Jingming Yan, and Ioannis Panageas. Learning equilibria in adversarial team
415 markov games: A nonconvex-hidden-concave min-max optimization problem. *arXiv preprint*
416 *arXiv:2410.05673*, 2024.
- 417 Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-
418 gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on*
419 *machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.
- 420 Daphne Koller, Nimrod Megiddo, and Bernhard Von Stengel. Efficient computation of equilibria for
421 extensive two-person games. *Games and economic behavior*, 14(2):247–259, 1996.
- 422 Christian Kroer, Kevin Waugh, Fatma Kılınc-Karzan, and Tuomas Sandholm. Faster algorithms for
423 extensive-form game solving via improved smoothing functions. *Mathematical Programming*, 179
424 (1):385–417, 2020.
- 425 Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat,
426 David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement
427 learning. *Advances in neural information processing systems*, 30, 2017.
- 428 Guoyin Li and Ting Kei Pong. Calculus of the exponent of kurdyka-łojasiewicz inequality and
429 its applications to linear convergence of first-order methods. *Foundations of computational*
430 *mathematics*, 18(5):1199–1232, 2018.

431 Feng-Yi Liao, Lijun Ding, and Yang Zheng. Error bounds, pl condition, and quadratic growth for
432 weakly convex functions, and linear convergences of proximal point methods. In *6th Annual*
433 *Learning for Dynamics & Control Conference*, pages 993–1005. PMLR, 2024.

434 Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In
435 *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

436 Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized
437 non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:
438 85–116, 2022a.

439 Mingyang Liu, Asuman Ozdaglar, Tiancheng Yu, and Kaiqing Zhang. The power of regularization in
440 solving extensive-form games. *arXiv preprint arXiv:2206.09495*, 2022b.

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

443 Weiming Liu, Huacong Jiang, Bin Li, and Houqiang Li. Equivalence analysis between counterfactual
444 regret minimization and online mirror descent. In *International Conference on Machine Learning*,
445 pages 13717–13745. PMLR, 2022c.

446 Edward Lockhart, Marc Lanctot, Julien Pérolat, Jean-Baptiste Lespiau, Dustin Morrill, Finbarr
447 Timbers, and Karl Tuyls. Computing approximate equilibria in sequential adversarial games by
448 exploitability descent. *arXiv preprint arXiv:1903.05614*, 2019.

449 Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by
450 first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

451 Yiren Lu, Justin Fu, George Tucker, Xinlei Pan, Eli Bronstein, Rebecca Roelofs, Benjamin Sapp,
452 Brandyn White, Aleksandra Faust, Shimon Whiteson, et al. Imitation is not enough: Robustifying
453 imitation with reinforcement learning for challenging driving scenarios. In *2023 IEEE/RSJ*
454 *International Conference on Intelligent Robots and Systems (IROS)*, pages 7553–7560. IEEE,
455 2023.

456 Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence
457 rates of softmax policy gradient methods. In *International conference on machine learning*, pages
458 6820–6829. PMLR, 2020.

459 Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity
460 in first-order non-convex optimization. In *International Conference on Machine Learning*, pages
461 7555–7564. PMLR, 2021.

462 Julie Mulvaney-Kemp, SangWoo Park, Ming Jin, and Javad Lavaei. Dynamic regret bounds for
463 online nonconvex optimization. 2021.

464 Katta G Murty and Santosh N Kabadi. Some np-complete problems in quadratic and nonlinear
465 programming. Technical report, 1985.

466 Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis. Decentralized stochastic control
467 with partial history sharing: A common information approach. *IEEE Transactions on Automatic*
468 *Control*, 58(7):1644–1658, 2013.

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

472 Garvesh Raskutti and Sayan Mukherjee. The information geometry of mirror descent. *IEEE*
473 *Transactions on Information Theory*, 61(3):1451–1457, 2015.

474 Quentin Rebjock and Nicolas Boumal. Fast convergence to non-isolated minima: four equivalent
475 conditions for c^2 functions. *Mathematical Programming*, pages 1–49, 2024.

476 I Romanovskii. Reduction of a game with complete memory to a matrix game. *Soviet Mathematics*,
477 3:678–681, 1962.

478 Max Rudolph, Nathan Lichtle, Sobhan Mohammadpour, Alexandre Bayen, J Zico Kolter, Amy Zhang,
479 Gabriele Farina, Eugene Vinitsky, and Samuel Sokota. Reevaluating policy gradient methods for
480 imperfect-information games. *arXiv preprint arXiv:2502.08938*, 2025.

481 Iosif Sakos, Emmanouil-Vasileios Vlatakis-Gkaragkounis, Panayotis Mertikopoulos, and Georgios
482 Piliouras. Exploiting hidden structures in non-convex games for convergence to nash equilibrium.
483 *Advances in Neural Information Processing Systems*, 36:66979–67006, 2023.

484 Kevin Scaman, Cedric Malherbe, and Ludovic Dos Santos. Convergence rates of non-convex
485 stochastic gradient descent under a generic lojasiewicz condition and local smoothness. In
486 *International conference on machine learning*, pages 19310–19327. PMLR, 2022.

487 Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon
488 Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari,
489 go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

490 John Schulman, Sergey Levine, P. Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region
491 policy optimization. *ArXiv*, abs/1502.05477, 2015. URL <https://api.semanticscholar.org/CorpusID:16046818>.
492

493 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy
494 optimization algorithms. *ArXiv*, abs/1707.06347, 2017. URL <https://api.semanticscholar.org/CorpusID:28695052>.
495

496 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
497 Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical
498 reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

499 Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):
500 1095–1100, 1953.

501 David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche,
502 Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering
503 the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

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

507 Samuel Sokota, Ryan D’Orazio, Chun Kai Ling, David J Wu, J Zico Kolter, and Noam Brown.
508 Abstracting imperfect information away from two-player zero-sum games. In *International*
509 *Conference on Machine Learning*, pages 32169–32193. PMLR, 2023.

510 Sriram Srinivasan, Marc Lanctot, Vinicius Zambaldi, Julien Pérolat, Karl Tuyls, Rémi Munos, and
511 Michael Bowling. Actor-critic policy optimization in partially observable multiagent environments.
512 *Advances in neural information processing systems*, 31, 2018.

513 Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere. In *2015*
514 *International Conference on Sampling Theory and Applications (SampTA)*, pages 407–410. IEEE,
515 2015.

516 Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods
517 for reinforcement learning with function approximation. *Advances in neural information processing*
518 *systems*, 12, 1999.

519 Oskari Tammelin. Solving large imperfect information games using cfr+. *arXiv preprint*
520 *arXiv:1407.5042*, 2014.

521 Oriol Vinyals, Igor Babuschkin, Junyoung Chung, Michael Mathieu, Max Jaderberg, Wojciech M
522 Czarnecki, Andrew Dudzik, Aja Huang, Petko Georgiev, Richard Powell, et al. Alphastar:
523 Mastering the real-time strategy game starcraft ii. *DeepMind blog*, 2:20, 2019.

524 Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, and Georgios Piliouras. Solving
525 min-max optimization with hidden structure via gradient descent ascent. *Advances in Neural*
526 *Information Processing Systems*, 34:2373–2386, 2021.

527 Bernhard Von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*,
528 14(2):220–246, 1996.

529 Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of
530 decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In
531 *Conference on learning theory*, pages 4259–4299. PMLR, 2021.

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

534 Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance reduction for a class of
535 nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*,
536 33:1153–1165, 2020.

537 Sihan Zeng, Thinh Doan, and Justin Romberg. Regularized gradient descent ascent for two-player
538 zero-sum markov games. *Advances in Neural Information Processing Systems*, 35:34546–34558,
539 2022.

540 Wenhao Zhan, Shicong Cen, Baihe Huang, Yuxin Chen, Jason D Lee, and Yuejie Chi. Policy mirror
541 descent for regularized reinforcement learning: A generalized framework with linear convergence.
542 *SIAM Journal on Optimization*, 33(2):1061–1091, 2023.

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

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

549 Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to nash
550 equilibria in zero-sum linear quadratic games. *Advances in Neural Information Processing Systems*,
551 32, 2019.

552 Yulai Zhao, Yuandong Tian, Jason Lee, and Simon Du. Provably efficient policy optimization for
553 two-player zero-sum markov games. In *International Conference on Artificial Intelligence and
554 Statistics*, pages 2736–2761. PMLR, 2022.

555 Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization
556 in games with incomplete information. *Advances in neural information processing systems*, 20,
557 2007.

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: Proofs of all claims are provided in the appendix

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discussed them in the conclusion section

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Yes found in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: experiments are small scale. code will be uploaded

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: code and proofs are in the supplemental material

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: code is shared

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: small scale experiments, confidence intervals included

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer “Yes” if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: description of laptop

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: we follow the NeurIPS code of ethics

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: work is theoretical. probably unlikely that it will have direct societal impacts

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: we cite previous work

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: no new assets released

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: no crowdsourcing

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: theoretical research

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines: only editing grammar

- 871 • The answer NA means that the core method development in this research does not
872 involve LLMs as any important, original, or non-standard components.
- 873 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
874 for what should or should not be described.

Appendix

875		
876		
877	A Further Related Work	22
878	B Optimization Lemmata	24
879	B.1 A variation of the descent lemma	26
880	B.2 Min-max Optimization	27
881	B.3 Regarding the Mahalanobis distance	29
882	B.4 Rewrite proof for function smooth relative to the Mahalanobis distance	29
883	C Further Preliminaries on EFGs	35
884	C.1 The behavioral and sequence-form strategies	35
885	C.2 Value, Action-Value, and Advantage Functions	36
886	C.3 Properties of the bidilated regularizer	37
887	D Regarding the Policy Parametrization	39
888	D.1 Definitions	39
889	E Gradient Domination	42
890	E.1 Direct Policy Patametrization	42
891	E.2 Softmax policy parametrization	43
892	E.3 Mahalanobis-PL	43
893	F Gradient Estimators	43
894	F.1 A Policy Gradient Theorem	43
895	F.2 Gradient Inexactness Bound	46
896	G Convergence Analysis	47
897	G.1 Direct Policy Parametrization	47
898	G.2 Softmax Policy Parametrization	47
899	G.3 Natural Policy Gradient	47
900	H Efficient Exploration	48
901		

A Further Related Work

903 **Relevant MARL for MG works** In MDP and MG literature, policy optimization seems to come in
904 two flavors—an *online learning()* approach and a *stochastic optimization* one. In the current work,
905 we opt for the second approach.

906 The work of (Zeng et al., 2022) is particularly close to ours. Yet, we highlight that they make a
907 rather strong assumption; they assume that the probability of playing each action in the support of

908 the regularized Nash equilibrium is lower-bounded by a constant independent of the regularization
 909 coefficient τ . In turn, we circumvent such an assumption by exercising direct control over the
 910 minimum probability of playing any action by projecting the parameters of the softmax parameters
 911 onto a convex polytope.

912 **Theory of Policy Gradient Methods** The policy gradient method (Williams, 1992; Sutton et al.,
 913 1999)

- 914 • (Agarwal et al., 2021) prove the convergence of directly parametrized policy gradient. They
 915 use the convergence result of gradient descent for smooth nonconvex function along a
 916 gradient domination lemma to demonstrate a $O(1/\epsilon^2)$ convergence rate to optimality. Later,
 917 (Zhang et al., 2020, 2021) use the *hidden concave* structure of the problem to improve the
 918 convergence rate to $O(1/\epsilon)$.
- 919 • (Mei et al., 2020) provide the first non-asymptotic convergence rate result for the policy
 920 gradient method using discounted entropy regularization (the analogue of bidilated entropy
 921 regularization). The proof of convergence uses a novel nonuniform PL condition.
- 922 • (Cen et al., 2022a) analyze natural policy gradient (NPG) with discounted entropy regular-
 923 ization. Natural policy gradient can be seen as a form of *preconditioned* gradient descent.
 924 Natural policy gradient effectively boils down to policy multiplicative weight updates using
 925 the Q -functions as feedback. The analysis of convergence uses a linear dynamical system.

B Optimization Lemmata

Definition 6 (Stationarity Proxies). Assume a function $F : f + I_{\mathcal{X}}(\cdot)$ such that $f : \mathcal{X} \rightarrow \mathbb{R}$ is ℓ -smooth relative to $\|\cdot\|_{\mathbf{M}}$ and $I_{\mathcal{X}}(\cdot)$ is the indicator function of the set \mathcal{X} . We define the following stationarity proxies,

- gradient of the Mahalanobis proximal mapping (MPM),

$$\|G_{\rho}(x)\| := \rho^2 \left\| x - \text{prox}_{F/\rho}(x) \right\|_{\mathbf{M}_t}^2$$

$$\text{with } \text{prox}_{F/\rho}(\cdot) := \arg \min_{x'} \{F(x') + \frac{\rho}{2} \|\cdot - x'\|_{\mathbf{M}}^2\}.$$

- Mahalanobis gradient mapping (MGM),

$$\|G_{\rho}^+(x)\| := \rho^2 \|x - x^+\|,$$

$$\text{where } x^+ := \arg \min_{x \in \mathcal{X}} \|x - \rho \mathbf{M}^{-1} \nabla f(x)\|_{\mathbf{M}}^2,$$

- Mahalanobis forward-backward mapping (MFBM),

$$\mathcal{D}(x, \rho) := -2\rho \min_{x'} \{\langle \nabla f(x), x' - x \rangle + \frac{\rho}{2} \|x - x'\|_{\mathbf{M}}^2 + I_{\mathcal{X}}(x') - I_{\mathcal{X}}(x)\},$$

Lemma B.1. The following properties hold true for the proximal point and the Mahalanobis Moreau envelope,

- $\nabla F_{\rho}(x) = \frac{1}{\rho}(x - \hat{x})$
- $\text{dist}(0, \partial F(\hat{x})) \leq \|\nabla F_{\rho}(x)\|_{\mathbf{M}^{-1}}$
- $F(\hat{x}) \leq F_{\rho}(\hat{x}) \leq F(x)$

Proof. The first and last items follow easily from the definition and standard arguments (Davis and Drusvyatskiy, 2018). The middle one uses the optimality condition of $\hat{x} := \text{prox}_{\rho F}(x)$,

$$0 \in \partial \left(F(\hat{x}) + \frac{1}{\rho} \mathbf{M}(\hat{x} - x) \right),$$

from which we conclude,

$$\frac{1}{\rho} \mathbf{M}(x - \hat{x}) \in \partial F(\hat{x}).$$

Finally, we conclude that $\min_{s_{\hat{x}} \in \partial F(\hat{x})} \|s_{\hat{x}}\|_{\mathbf{M}^{-1}}^2 \leq \frac{1}{\rho^2} \|x - \hat{x}\|_{\mathbf{M}}^2$. \square

Definition 7 (pPL, KL). Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be an L -Lipschitz continuous function with ℓ -Lipschitz continuous gradient. Then,

- Proximal Polyak-Łojasiewicz (pPL): f is said to satisfy the proximal Polyak-Łojasiewicz condition if $\exists \alpha > 0$ s.t.

$$\frac{1}{2} \mathcal{D}_{\mathcal{X}}(x, \ell) \geq \alpha [f(x) - f(x^*)]$$

- Kurdyka-Łojasiewicz (KL): f is said to satisfy if $\exists \bar{\alpha}$ s.t.

$$\min_{s_x \in \partial(f + I_{\mathcal{X}})(x)} \|s_x\|^2 \geq 2\bar{\alpha} [f(x) - f(x^*)], \quad \forall x \in \mathcal{X}.$$

The definitions for the Mahalanobis analogues of pPL and KL follow straight-forward extension.

Lemma B.2. Let f be an ℓ -smooth function relative to $\|\cdot\|_{\mathbf{M}}$ defined over the convex set \mathcal{X} . If f satisfies the (Mahalanobis) KL condition with modulus μ_{kl} , it also satisfies the pPL condition with a modulus of $\mu_{\text{ppl}} = \frac{\mu_{\text{kl}}}{202}$.

953 *Proof.* First, we define $F(x) := f(x) + I_{\mathcal{X}}(x)$, with $I_{\mathcal{X}}(\cdot)$ being the indicator function. We highlight
 954 that since $I_{\mathcal{X}}(\cdot)$ is convex and f is ℓ -smooth (relative to $\|\cdot\|_{\mathbf{M}}^2$), then F is ℓ -weakly convex (relative
 955 to $\|\cdot\|_{\mathbf{M}}^2$). This means that the proximal point of the function F/ρ is well defined for any $\rho > \ell$.
 956 Now, assume a point $x \in \mathcal{X}$ and $\hat{x} := \text{prox}_{F/\rho}(x)$. By assumption, for any $\hat{x} \in \mathcal{X}$, it holds true that,

$$\frac{1}{2} \|s_{\hat{x}}\|^2 \geq \alpha [f(\hat{x}) - f^*]$$

957 where $s_{\hat{x}} \in \partial F(\hat{x})$. The latter implies that for the gradient of the Mahalanobis-Moreau envelope of
 958 F , it holds that,

$$\begin{aligned} \frac{1}{2} \|\nabla F_{1/\rho}(x)\|_{\mathbf{M}}^2 &\geq \alpha [f(\hat{x}) - f^*] \\ &= \alpha + \alpha [f(\hat{x}) - f(x)] \\ &\geq \alpha [f(x) - f^*] - \alpha \left(\frac{1}{2\rho} \mathcal{D}(x, \rho) + \frac{\ell + \rho}{2} \|x - \hat{x}\|_{\mathbf{M}}^2 \right) \end{aligned} \quad (1)$$

959 where (1) follows from the fact that F is an ℓ -weakly convex function, and for every $v \in \partial F(x)$. To
 960 see this, we write that due to weak convexity (relative to $\|\cdot\|_{\mathbf{M}}^2$),

$$\begin{aligned} F(\hat{x}) &\geq F(x) + \langle v, \hat{x} - x \rangle - \frac{\ell}{2} \|x - \hat{x}\|_{\mathbf{M}}^2 \\ &= F(x) + \langle v, \hat{x} - x \rangle + \frac{\rho}{2} \|x - \hat{x}\|_{\mathbf{M}}^2 - \frac{\ell + \rho}{2} \|x - \hat{x}\|_{\mathbf{M}}^2 \\ &\geq F(x) + \min_{y \in \mathcal{Y}} \left\{ \langle \nabla f(x), y - x \rangle + \frac{\rho}{2} \|x - y\|_{\mathbf{M}}^2 \right\} - \frac{\ell + \rho}{2} \|x - \hat{x}\|_{\mathbf{M}}^2 \\ &= F(x) - \frac{1}{2\rho} \mathcal{D}(x, \rho) - \frac{\ell + \rho}{2} \|x - \hat{x}\|_{\mathbf{M}}^2 \end{aligned}$$

961 Collecting the terms,

$$\left(\frac{1}{2} + \alpha \frac{\ell + \rho}{2\rho^2} \right) \|\nabla F_{\rho}(x)\|_{\mathbf{M}}^2 + \frac{\alpha}{2\rho} \mathcal{D}(x, \rho) \geq \alpha [f(x) - f^*].$$

962 A direct generalization of (Karimi et al., 2016, Lemma 1), implies that for the MFBM and a choice of
 963 $\rho_1, \rho_2 > 0$ such that $\rho_1 > \rho_2$, then $\mathcal{D}(x, \rho_1) \geq \mathcal{D}(x, \rho_2)$. As such, we write,

$$\left(\frac{1}{2} + \alpha \frac{\ell + \rho}{2\rho^2} \right) \|\nabla F_{1/\rho}(x)\|_{\mathbf{M}}^2 + \frac{\alpha}{2\rho} \mathcal{D}(x, 2\rho) \geq \alpha [f(x) - f^*].$$

964 We can pick $\rho = 4\ell$ which then yields,

$$\left(\frac{1}{2} + \frac{12\alpha}{\ell} \right) \|\nabla F_{1/(4\ell)}(x)\|_{\mathbf{M}}^2 + \frac{\alpha}{8\ell} \mathcal{D}(x, 4\ell) \geq \alpha [f(x) - f^*].$$

965 Observing that $\alpha \leq \ell$ in general, we re-write:

$$\frac{25}{2} \|\nabla F_{1/(4\ell)}(x)\|_{\mathbf{M}}^2 + \frac{1}{8} \mathcal{D}(x, 4\ell) \geq \alpha [f(x) - f^*].$$

966 Now, from (Fatkhullin and He, 2024, Lemmata 4.1 & 4.2), we know that,

$$16\mathcal{D}(x, 4\ell) \geq \|\nabla F_{1/\rho}(\hat{x})\|_{\mathbf{M}}^2$$

967 which we plugin in the former inequality to finally conclude that,

$$\frac{1}{2} \mathcal{D}(x, 4\ell) \geq \frac{\mu}{202} [f(x) - f^*].$$

968 □

969 **Remark 2.** The latter lemma provides a bound that is significantly tighter than the one implied
 970 by the analysis found (Karimi et al., 2016, Appendix G) which connects the moduli of the KL and
 971 pPL conditions.

972 **B.1 A variation of the descent lemma**

973 The following lemma is a consequence of the three-point identity of the Mahalanobis norm and the
 974 smoothness of f .

Lemma B.3 ((J Reddi et al., 2016, Lemma 1)). *Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be an ℓ -smooth function relative to $\|\cdot\|_{\mathbf{M}_t}$ and a point $x \in \mathcal{X} \subseteq \mathbb{R}^d$. Also, define the vector $v \in \mathbb{R}^d$ and $y \in \mathcal{X}$ to be*

$$y := \text{Proj}_{\mathcal{X}, \mathbf{M}_t} (x - \eta \mathbf{M}_t^{-1} v).$$

975 Then, the following inequality holds true:

$$\begin{aligned} f(y) &\leq f(z) + \langle \nabla f(x) - v, y - z \rangle \\ &\quad + \left(\frac{\ell}{2} - \frac{1}{2\eta} \right) \|y - x\|_{\mathbf{M}_t}^2 + \left(\frac{\ell}{2} + \frac{1}{2\eta} \right) \|z - x\|_{\mathbf{M}_t}^2 - \frac{1}{2} \|y - z\|_{\mathbf{M}_t}^2. \end{aligned}$$

976

977 **Lemma B.4.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a closed convex set, and let $f : \mathcal{X} \rightarrow \mathbb{R}$ be an ℓ -smooth function
 978 relative to $\|\cdot\|_{\mathbf{M}_t}$ for some $\ell > 0$. Suppose $\eta > 0$ with $\eta \leq \frac{1}{5\ell}$. For any $x \in \mathcal{X}$ and any vector
 979 $v \in \mathbb{R}^d$, define $x^+ = \text{Proj}_{\mathcal{X}, \mathbf{M}_t} (x - \eta v)$. Then the following inequality holds:*

$$f(x^+) \leq f(x) - \frac{\eta}{6} \mathcal{D}_{\mathcal{X}}(x, 1/\eta) + \frac{\eta}{2} \|\nabla f(x) - v\|_{\mathbf{M}_t^{-1}}^2.$$

980

981 *Proof.* First, we define $\bar{x}^+ := \text{Proj}_{\mathcal{X}, \mathbf{M}_t} \left(x - \frac{1}{\rho} \mathbf{M}_t^{-1} \nabla f(x) \right)$.

- 982 • Invoking ℓ -smoothness relative to $\|\cdot\|_{\mathbf{M}_t}$ of f for x, \bar{x}_+ and assuming $\rho > 0$ with $\rho \geq \ell$,

$$\begin{aligned} f(\bar{x}_+) &\leq f(x) + \langle \nabla f(x), \bar{x}_+ - x \rangle + \frac{\ell}{2} \|x_+ - x\|_{\mathbf{M}_t}^2 \\ &\leq f(x) + \langle \nabla f(x), \bar{x}_+ - x \rangle + \frac{\rho}{2} \|x_+ - x\|_{\mathbf{M}_t}^2 \\ &= f(x) - \left(\langle \nabla f(x), x - \bar{x}_+ \rangle - \frac{\rho}{2} \|x_+ - x\|_{\mathbf{M}_t}^2 \right) \\ &= f(x) - \frac{1}{2\rho} \mathcal{D}_{\mathbf{M}_t}(x, \rho). \end{aligned} \tag{2}$$

- 983 • Invoking Lemma B.3 with $x = x, y = \bar{x}_+, z = x, v = \nabla f(x)$

$$f(\bar{x}_+) \leq f(x) + \left(\frac{\ell}{2} - \frac{1}{\rho} \right) \|\bar{x}_+ - x\|_{\mathbf{M}_t}^2. \tag{3}$$

- 984 • Again, invoking Lemma B.3 but with $x = x, y = x_+, z = \bar{x}_+, v$,

$$\begin{aligned} f(x_+) &\leq f(\bar{x}_+) + \langle \nabla f(x) - v, x_+ - \bar{x}_+ \rangle \\ &\quad + \left(\frac{\ell}{2} - \frac{1}{2\eta} \right) \|x_+ - x\|_{\mathbf{M}_t}^2 + \left(\frac{\ell}{2} + \frac{1}{2\eta} \right) \|\bar{x}_+ - x\|_{\mathbf{M}_t}^2 - \frac{1}{2\eta} \|x_+ - \bar{x}_+\|_{\mathbf{M}_t}^2. \end{aligned} \tag{4}$$

985 Combining the previous inequalities as $1/3 \times (2)$ and $2/3 \times (3)$, and letting $1/\rho = \eta \leq \frac{1}{\ell}$
 986 yields,

$$f(\bar{x}_+) \leq f(x) - \frac{1}{6\eta} \mathcal{D}_{\mathcal{X}}(x, 1/\eta) + \left(\frac{\ell}{3} - \frac{2}{3\eta} \right) \|\bar{x}_+ - x\|_{\mathbf{M}_t}^2$$

Adding (4),

$$\begin{aligned}
f(x_+) &\leq f(x) - \frac{\eta}{6} \mathcal{D}_{\mathcal{X}}(x, 1/\eta) + \left(\frac{\ell}{3} - \frac{2}{3\eta} \right) \|\bar{x}_+ - x\|_{\mathbf{M}_t}^2 \\
&\quad + \langle \nabla f(x) - v, x_+ - \bar{x}_+ \rangle \\
&\quad + \left(\frac{\ell}{2} - \frac{1}{2\eta} \right) \|x_+ - x\|_{\mathbf{M}_t}^2 + \left(\frac{\ell}{2} + \frac{1}{2\eta} \right) \|\bar{x}_+ - x\|_{\mathbf{M}_t}^2 - \frac{1}{2\eta} \|x_+ - \bar{x}_+\|_{\mathbf{M}_t}^2 \\
&\leq f(x) - \frac{\eta}{6} \mathcal{D}_{\mathcal{X}}(x, 1/\eta) + \left(\frac{5\ell}{6} - \frac{1}{6\eta} \right) \|\bar{x}_+ - x\|_{\mathbf{M}_t}^2 \\
&\quad + \frac{\rho}{2} \|\nabla f(x) - v\|_{\mathbf{M}_t^{-1}}^2 + \frac{1}{2\rho} \|x_+ - \bar{x}_+\|_{\mathbf{M}_t}^2 \\
&\quad + \left(\frac{\ell}{2} - \frac{1}{2\eta} \right) \|x_+ - x\|_{\mathbf{M}_t}^2 - \frac{1}{2\eta} \|x_+ - \bar{x}_+\|_{\mathbf{M}_t}^2 \tag{5} \\
&= f(x) - \frac{\eta}{6} \mathcal{D}_{\mathcal{X}}(x, 1/\eta) + \left(\frac{5\ell}{6} - \frac{1}{6\eta} \right) \|\bar{x}_+ - x\|_{\mathbf{M}_t}^2 \\
&\quad + \frac{\eta}{2} \|\nabla f(x) - v\|_{\mathbf{M}_t^{-1}}^2 \\
&\quad + \left(\frac{\ell}{2} - \frac{1}{2\eta} \right) \|x_+ - x\|_{\mathbf{M}_t}^2 \\
&\leq f(x) - \frac{\eta}{6} \mathcal{D}_{\mathcal{X}}(x, 1/\eta) + \frac{\eta}{2} \|\nabla f(x) - v\|_{\mathbf{M}_t^{-1}}^2 \tag{6}
\end{aligned}$$

- (5) follows from the application of Young's inequality on $\langle \nabla f(x) - v, x^+ \bar{x}^+ \rangle = \langle \mathbf{M}_t^{-1/2} \nabla f(x) - v, \mathbf{M}_t^{1/2} x^+ \bar{x}^+ \rangle$;
- (6) follows by dropping the non-positive terms; non-positivity follows from the choice of the step-size, $\eta \leq \frac{1}{5\ell}$.

992

□

993 B.2 Min-max Optimization

Lemma B.5. *Let $f : \mathcal{X} \times \mathcal{Y}$ be an ℓ -smooth function, $\rho > 0$, two points $y, y' \in \mathcal{Y}$, and a point $x \in \mathcal{X}$. Then, the following inequality holds:*

$$|\mathcal{D}_{\mathcal{X}}(x, \rho; y) - \mathcal{D}_{\mathcal{X}}(x, \rho; y')| \leq 3\lambda_{\max}(\mathbf{M}_t^{-1})\ell^2 \|y - y'\|^2.$$

994

995 *Proof.* We define $\bar{x}, \bar{x}' \in \mathcal{X}$ to be:

$$\begin{aligned}
\bar{x} &:= \text{Proj}_{\mathcal{X}, \mathbf{M}_t} \left(x - \frac{1}{\rho} \mathbf{M}_t^{-1} \nabla_x f(x, y) \right); \\
\bar{x}' &:= \text{Proj}_{\mathcal{X}, \mathbf{M}_t} \left(x - \frac{1}{\rho} \mathbf{M}_t^{-1} \nabla_x f(x, y') \right).
\end{aligned}$$

996 By the definition of $\mathcal{D}_{\mathcal{X}}(x, \rho; y)$ we write:

$$\begin{cases} \frac{1}{2\rho} \mathcal{D}_{\mathcal{X}}(x, \rho; y) = \langle \nabla f(x, y), x - \bar{x} \rangle - \frac{\rho}{2} \|x - \bar{x}\|_{\mathbf{M}_t}^2; \\ \frac{1}{2\rho} \mathcal{D}_{\mathcal{X}}(x, \rho; y') = \langle \nabla f(x, y'), x - \bar{x}' \rangle - \frac{\rho}{2} \|x - \bar{x}'\|_{\mathbf{M}_t}^2. \end{cases}$$

997 Considering the difference $\mathcal{D}_{\mathcal{X}}(x, \rho; y) - \mathcal{D}_{\mathcal{X}}(x, \rho; y')$ we see that:

$$\begin{aligned}
& \frac{1}{2\rho} |\mathcal{D}_{\mathcal{X}}(x, \rho; y) - \mathcal{D}_{\mathcal{X}}(x, \rho; y')| \\
&= \left| \langle \nabla_x f(x, y) - \nabla_x f(x, y'), \bar{x}' - \bar{x} \rangle - \frac{\rho}{2} \left(\|x - \bar{x}\|_{\mathbf{M}_t}^2 - \|x - \bar{x}'\|_{\mathbf{M}_t}^2 \right) \right| \\
&\leq |\langle \nabla_x f(x, y) - \nabla_x f(x, y'), \bar{x}' - \bar{x} \rangle| + \frac{\rho}{2} \left| \left(\|x - \bar{x}\|_{\mathbf{M}_t}^2 - \|x - \bar{x}'\|_{\mathbf{M}_t}^2 \right) \right| \\
&\leq |\langle \nabla_x f(x, y) - \nabla_x f(x, y'), \bar{x}' - \bar{x} \rangle| + \frac{\rho}{2} \|\bar{x} - \bar{x}'\|_{\mathbf{M}_t}^2 \\
&\leq \|\nabla_x f(x, y) - \nabla_x f(x, y')\|_{\mathbf{M}_t^{-1}} \|\bar{x}' - \bar{x}\|_{\mathbf{M}_t} + \frac{\rho}{2} \|\bar{x} - \bar{x}'\|_{\mathbf{M}_t}^2 \\
&\leq \frac{1}{\rho} \|\nabla_x f(x, y) - \nabla_x f(x, y')\|_{\mathbf{M}_t^{-1}}^2 + \frac{1}{2\rho} \|\nabla_x f(x, y) - \nabla_x f(x, y')\|_{\mathbf{M}_t^{-1}}^2 \\
&\leq \frac{\lambda_{\max}(\mathbf{M}_t^{-1})}{\rho} \|\nabla_x f(x, y) - \nabla_x f(x, y')\|^2 + \frac{\lambda_{\max}(\mathbf{M}_t^{-1})}{2\rho} \|\nabla_x f(x, y) - \nabla_x f(x, y')\|^2 \\
&\leq \frac{3\lambda_{\max}(\mathbf{M}_t^{-1})\ell^2}{2\rho} \|y - y'\|^2.
\end{aligned}$$

998 We note that:

- 999 • The first inequality follows from the triangle inequality.
- 1000 • In the second inequality, we applied the reverse triangle inequality.
- 1001 • The third uses the Cauchy-Schwarz inequality.
- 1002 • Finally, the second to last uses Lemma B.7 while, the last one, invokes the ℓ -Lipschitz
- 1003 continuity of the gradient.

1004

□

1005 **Lemma B.6.** Let $f : \mathcal{X} \times \mathcal{Y}$ be an ℓ -smooth function such that for any $x \in \mathcal{X}$, $f(x, \cdot)$ satisfies the
1006 proximal-PŁ condition with modulus $\alpha > 0$. Then, the function $\Phi(x) := \arg \max_{y \in \mathcal{Y}} f(x, y)$ is
1007 ℓ_* -smooth, with

$$\ell_* := \ell \left(1 + \frac{\ell}{\alpha} \right).$$

1008 *Proof.* We effectively need to show Lipschitz continuity of the maximizers $y^*(\cdot) := \arg \max_x$ and
1009 the proof will follow from Danskin's lemma and f 's own ℓ -smoothness. So, we write by the quadratic
1010 growth condition,

$$\frac{\alpha}{2} \|y^*(x') - y^*(x)\|^2 \leq f(x, y^*(x)) - f(x, y^*(x')). \quad (7)$$

1011 We denote $\mathcal{D}_{\mathcal{Y}}(\cdot, \rho; x) := -2\rho \arg \min_{z \in \mathcal{Y}} \{ \langle -\nabla f(x, y), z - y \rangle + \frac{\rho}{2} \|y - z\|^2 \}$ and by the proximal-
1012 PŁ condition, we write,

$$f(x, y^*(x)) - f(x, y^*(x')) \leq \frac{1}{2\alpha} \mathcal{D}_{\mathcal{Y}}(y, \ell; x). \quad (8)$$

Now, we aim to bound $\mathcal{D}_{\mathcal{Y}}(y, \ell; x)$ by $\|y^*(x) - y^*(x')\|^2$. We observe that,

$$\mathcal{D}_{\mathcal{Y}}(y^*(x), \ell; x) = 0.$$

1013 Hence,

$$\begin{aligned}
\mathcal{D}_{\mathcal{Y}}(y^*(x'), \ell; x) &= \mathcal{D}_{\mathcal{Y}}(y^*(x'), \ell; x) - \mathcal{D}_{\mathcal{Y}}(y^*(x), \ell; x) \\
&\leq 2\ell^2 \|x - x'\|^2
\end{aligned} \quad (9)$$

1014 where the last line follows from a slight sharpening of the proof of Lemma B.5 (for the function
1015 $h(y, x) = -f(x, y)$ and $\mathbf{M} = \mathbf{I}$). Finally, piecing inequalities (7), (8), and (9) together,

$$\|y^*(x) - y^*(x')\| \leq \frac{\ell}{\alpha} \|x - x'\|. \quad (10)$$

1016 What is left to do is to observe the following, due to Danskin's theorem and ℓ -smoothness of f ,

$$\begin{aligned}\|\nabla_x \Phi(x) - \nabla_x \Phi(x')\| &= \|\nabla_x f(x, y^*(x)) - \nabla_x f(x', y^*(x'))\| \\ &\leq \ell \|(x, y^*(x)) - (x', y^*(x'))\| \\ &\leq \ell \|x - x'\| + \frac{\ell^2}{\alpha} \|x - x'\|.\end{aligned}$$

1017 The latter inequality follows from (10) and completes the proof. \square

1018 B.3 Regarding the Mahalanobis distance

1019 Throughout, we will refer to a positive-semidefinite matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$ and its Moore-Penrose
1020 pseudo-inverse $\mathbf{M}^\dagger \in \mathbb{R}^{d \times d}$. Although in general a PSD matrix cannot define a distance, restricting
1021 $x, y \in \mathbb{R}^d$ such that $(x - y) \in \ker(\mathbf{M})^\perp$, then $\|x - y\|_{\mathbf{M}}^2 := \sqrt{(x - y)^\top \mathbf{M} (x - y)}$ satisfies all
1022 properties of a metric. As we shall see, this seemingly arbitrary assumption is satisfied for every
1023 pair of consecutive updates of natural policy gradient steps. The matrix rank-deficient matrix we are
1024 interested in is policy gradient Fisher information matrix, and for softmax policy parametrization, it
1025 is rank deficient in the direction $\mathbf{1} \in \mathbb{R}^d$. Further, the gradient $\nabla f(x)$ as

1026 **Proposition 1.** Assume that $\theta_0 = \mathbf{0}$. Also, let $v_t^\top \mathbf{1} = 0, \forall t \in \{1, 2, 3, \dots\}$. Then, setting
1027 $\theta_{t+1} = \theta_t - \eta \mathbf{M}^\dagger v_t$ guarantees that,

$$(\theta_{t+1} - \theta_t)^\top \mathbf{1} \quad \text{and} \quad \theta_t^\top \mathbf{1} = 0, \forall t.$$

1028 *Proof.* Since, $\theta_{t+1} = \theta_t - \eta \mathbf{M}^\dagger v_t$, we see that $\theta_{t+1}^\top \mathbf{1} = (\theta_t - \eta \mathbf{M}^\dagger v_t)^\top \mathbf{1} = 0$ and $(\theta_{t+1} - \theta_t)^\top \mathbf{1} =$
1029 0 . \square

1030 **Proposition 2.** Let $\Theta \subseteq \mathbb{R}^d$ be a convex compact set. Assume that $\theta_0 = \mathbf{0}$. Also, let $v_t^\top \mathbf{1} = 0, \forall t \in$
1031 $\{1, 2, 3, \dots\}$. Then, the following minimization problem has a unique solution,

$$\min_{\theta \in \Theta, \text{s.t. } (\theta - \theta_t)^\top \mathbf{1} = 0} \|(\theta_t - \eta \mathbf{M}^\dagger v_t) - \theta\|_{\mathbf{M}}^2.$$

1032 Further, it is equivalent to the minimization problem,

$$\min_{\theta \in \Theta, \text{s.t. } (\theta - \theta_t)^\top \mathbf{1} = 0} \left\{ \langle v_t, \theta - \theta_t \rangle + \frac{1}{2\eta} \|\theta - \theta_t\|_{\mathbf{M}}^2 \right\}.$$

1033 *Proof.* It is clear that, for $\theta, \chi \in \Theta, \theta^\top \mathbf{1} = \chi^\top \mathbf{1} = 0$ the function $\|\theta\|_{\mathbf{M}}^2, \|\theta - \chi\|_{\mathbf{M}}^2$ is strongly
1034 convex in θ . Hence, both problems attain a unique minimum.

1035 For the first problem, the first-order optimality conditions for the write,

$$\langle \theta^+ - (\theta - \eta \mathbf{M}^\dagger v_t), \theta - \theta^+ \rangle \geq 0, \quad \forall \theta \in \Theta, \theta^\top \mathbf{1} = 0.$$

1036 Noting that, $(\theta^+ - (\theta - \eta \mathbf{M}^\dagger v_t))^\top \mathbf{1} = 0$ and $(\theta - \theta^+)^\top \mathbf{1} = 0$,

$$\langle \mathbf{M}\theta^+ - \mathbf{M}\theta + \eta v_t, \mathbf{M}^\dagger(\theta - \theta^+) \rangle \geq 0, \quad \forall \theta \in \Theta, \theta^\top \mathbf{1} = 0$$

1037 But, since the matrix \mathbf{M} is PSD and the last inequality is a condition on the sign of the inner-product,
1038 it can be written equivalently as,

$$\langle \mathbf{M}\theta^+ - \mathbf{M}\theta + \eta v_t, (\theta - \theta^+) \rangle \geq 0, \quad \forall \theta \in \Theta, \theta^\top \mathbf{1} = 0.$$

1039 The final inequality, is exactly the first-order optimality condition for the second minimization
1040 problem. \square

1041 B.4 Rewrite proof for function smooth relative to the Mahalanobis distance

1042 B.4.1 Supporting Lemmata

1043 **Lemma B.7.** Let v_1, v_2 be vectors in \mathbb{R}^d and $\mathcal{X} \subseteq \mathbb{R}^d$ be a compact convex set and a scalar $\eta > 0$.
1044 Also, let points $x_1^+, x_2^+ \in \mathcal{X}$ such that:

$$x_1^+ := \text{Proj}_{\mathcal{X}, \mathbf{M}_t} (x - \eta \mathbf{M}_t^{-1} v_1);$$

$$x_2^+ := \text{Proj}_{\mathcal{X}, \mathbf{M}_t} (x - \eta \mathbf{M}_t^{-1} v_2).$$

1045 Then, it holds true that:

$$\|x_1^+ - x_2^+\|_{\mathbf{M}_t} \leq \eta \|v_1 - v_2\|_{\mathbf{M}_t^{-1}}.$$

1046 .

Smoothness

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle - \frac{\ell}{2} \|x - y\|^2 \\ &\geq f(x) + \langle \nabla f(x), y - x \rangle - \frac{\ell}{2\lambda_{\min}(\mathbf{M}_t)} \|x - y\|_{\mathbf{M}_t}^2 \end{aligned}$$

$$\begin{aligned} f(y) &\leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\ell}{2} \|x - y\|^2 \\ &\leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\ell}{2\lambda_{\min}(\mathbf{M}_t)} \|x - y\|_{\mathbf{M}_t}^2 \end{aligned}$$

$$\begin{aligned} \Phi(x') &\leq \Phi(x) + \langle \nabla \Phi(x), x' - x \rangle + \frac{\ell_\phi}{2} \|x' - x\|^2 \\ &\leq \Phi(x) + \langle \nabla \Phi(x), x' - x \rangle + \frac{\ell_\phi}{2\lambda_{\min}(\mathbf{M}_t)} \|x' - x\|_{\mathbf{M}_t}^2 \end{aligned}$$

$$\begin{aligned} \|\nabla_x \Phi(x) - \nabla_x f(x, y)\|_{\mathbf{M}_t^{-1}}^2 &\leq \lambda_{\max}(\mathbf{M}_t^{-1}) \|\nabla_x \Phi(x) - \nabla_x f(x, y)\|^2 \\ &\leq \lambda_{\max}(\mathbf{M}_t^{-1}) \ell^2 \|y^*(x) - y\|^2 \end{aligned}$$

PL and QG

$$\frac{\mu_{\text{qg}}}{2} \|y^*(x) - y\|^2 \leq \Phi(x) - f(x, y)$$

$$\begin{aligned} f(x, y) - \min_{x'} f(x', y) &\leq \frac{1}{2\mu_x} \mathcal{D}_{\mathcal{X}}(x, \alpha; y) \\ \max_y' f(x, y') - f(x, y) &\leq \frac{1}{2\mu_y} \mathcal{D}_{\mathcal{Y}}(y, \alpha; x) \end{aligned}$$

Controlling δ

$$\begin{aligned} \|\nabla f(x) - g\|_{\mathbf{M}_t^{-1}}^2 &= \|\mathbf{M}_t^{-1} (\nabla f(x) - g)\|_{\mathbf{M}_t}^2 \\ &\leq \lambda_{\max}(\mathbf{M}_t) \|A - \tilde{A}\|^2 \end{aligned}$$

1047 B.4.2 Convergence of alternating gradient descent ascent

1048 We consider the iteration scheme,

$$\begin{aligned} x_{t+1} &= \arg \min_{x \in \mathcal{X}} \langle \nabla f(x), x - x_t \rangle + \frac{1}{2\eta_x} \|x - x_t\|_{\mathbf{M}_{x,t}}^2 \\ y_{t+1} &= \arg \min_{y \in \mathcal{Y}} \langle -\nabla f(x_{t+1}, y_t), y - y_t \rangle + \frac{1}{2\eta_y} \|y - y_t\|_{\mathbf{M}_{y,t}}^2 \end{aligned} \tag{11}$$

1049 **Assumption 2** (Unbiased Gradient Estimators and Bounded Second Moments). *For all iterations t ,*
 1050 *the gradient estimators $\hat{g}_x(x^t, y^t)$ and $\hat{g}_y(x^t, y^t)$ satisfy*

$$\mathbb{E} [\hat{g}_x(x^t, y^t)] = \nabla_x f(x^t, y^t),$$

1051

$$\mathbb{E} [\hat{g}_y(x^t, y^t)] = \nabla_y g(x^t, y^t),$$

1052 *and*

$$\mathbb{E} [\|\hat{g}_x(x^t, y^t)\|^2] \leq \sigma_x^2,$$

1053

$$\mathbb{E} [\|\hat{g}_y(x^t, y^t)\|^2] \leq \sigma_y^2.$$

1054 *In turn, $\|g_x(x^t, y^t) - \nabla_x f(x^t, y^t)\| \leq \delta_x$, $\|g_y(x^t, y^t) - \nabla_y f(x^t, y^t)\| \leq \delta_y$.*

1055 **Theorem B.1.** *Let $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ be a bounded in the interval Δ_f , ℓ -smooth function and*
 1056 *\mathcal{X}, \mathcal{Y} be two convex sets with Euclidean diameters. Moreover, assume that f satisfies a two-sided*
 1057 *pPL condition with moduli μ_x for all $y \in \mathcal{Y}$ and μ_y for any $x \in \mathcal{X}$. Additionally, let (\hat{g}_x, \hat{g}_y) be an*
 1058 *inexact stochastic gradient oracle satisfying Assumption 2.*

1059 • *When $\mathbf{M}_t = \mathbf{I}$, after T iterations of (11) with a choice of stepsizes $\eta_x = \frac{\mu_y^2}{960\ell^3}$ and $\eta_y = \frac{1}{5\ell}$,*
 1060 *it holds true that:*

$$\begin{aligned} & \mathbb{E}\Phi(x_T) - \Phi^* + \frac{1}{10} (\mathbb{E}\Phi(x_T) - \mathbb{E}f(x_T, y_T)) \\ & \leq \exp\left(-\frac{\mu_x \mu_y^2}{960\ell^3} T\right) \Delta_f + \frac{c_1 \sigma_x^2}{\mu_x} + \frac{c_1 \delta_x^2}{\mu_x} + \frac{c_2 \ell^2 \sigma_y^2}{\mu_x \mu_y^2} + \frac{c_2 \ell^2 \delta_y^2}{\mu_x \mu_y^2}, \end{aligned}$$

1061 *where, $c_1, c_2 \in O(1)$.*

1062 • *For a general choice of $\mathbf{M}_t = \mathbf{I}$, after T iterations of (11) with a choice of stepsizes*
 1063 *$\eta_x = \frac{\mu_y^3}{960\lambda_{\max}\ell^3\lambda_{\max}}$ and $\eta_y = \frac{1}{5\ell}$, it holds true that:*

$$\begin{aligned} & \mathbb{E}\Phi(x_T) - \Phi^* + \frac{1}{10} (\mathbb{E}\Phi(x_T) - \mathbb{E}f(x_T, y_T)) \\ & \leq \exp\left(-\frac{\mu_x \mu_y^2}{960\ell^3} T\right) \Delta_f + \frac{c_1 \sigma_x^2}{\mu_x} + \frac{c_1 \delta_x^2}{\mu_x} + \frac{c_2 \ell^2 \sigma_y^2}{\mu_x \mu_y^2} + \frac{c_2 \ell^2 \delta_y^2}{\mu_x \mu_y^2}, \end{aligned}$$

1064 *where, $c_1, c_2 \in O(1)$.*

1065 **Descent on Φ** In order to guarantee descent, we pick $\eta_x \leq \frac{1}{5\ell\lambda_{\max}(\mathbf{M}_t^{-1})}$. By Lemma B.4 we write,

$$\begin{aligned} \mathbb{E}\Phi(x_{t+1}) & \leq \mathbb{E}\Phi(x_t) - \frac{\eta_x}{6} \mathbb{E}\mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x) + \eta_x \mathbb{E} \|\nabla_x \Phi(x_t) - \nabla_x f(x_t, y_t)\|_{\mathbf{M}_t^{-1}}^2 \\ & \quad + 2\eta_x \sigma_x^2 + 2\eta_x \delta_x^2. \end{aligned}$$

1066 Equivalently, subtracting Φ^* from both sides yields,

$$\begin{aligned} \mathbb{E}\Phi(x_{t+1}) - \Phi^* & \leq \mathbb{E}\Phi(x_t) - \Phi^* - \frac{\eta_x}{6} \mathbb{E}\mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x) + \eta_x \mathbb{E} \|\nabla_x \Phi(x_t) - \nabla_x f(x_t, y_t)\|_{\mathbf{M}_t^{-1}}^2 \\ & \quad + 2\eta_x \sigma_x^2 + 2\eta_x \delta_x^2. \end{aligned}$$

1067 Further, a simple re-arrangement reads,

$$\begin{aligned} \mathbb{E}\Phi(x_{t+1}) - \mathbb{E}\Phi(x_t) & \leq -\frac{\eta_x}{6} \mathbb{E}\mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x) + \eta_x \mathbb{E} \|\nabla_x \Phi(x_t) - \nabla_x f(x_t, y_t)\|_{\mathbf{M}_t^{-1}}^2 \\ & \quad + 2\eta_x \sigma_x^2 + 2\eta_x \delta_x^2. \end{aligned}$$

1068 **Ascent on $f(x, \cdot)$** Requiring that $\eta_y \leq \frac{1}{5\ell\mathbf{M}_t}$,

$$\mathbb{E}f(x_{t+1}, y_{t+1}) \geq \mathbb{E}f(x_{t+1}, y_t) + \frac{\eta_y}{6} \mathbb{E}\mathcal{D}_{\mathcal{Y}}(y_t, 1/\eta_y; x_{t+1}) - \eta_y \delta^2 - \eta_y \sigma_y^2$$

1069 Multiplying by -1 and adding $\Phi(x_{t+1})$ will yield,

$$\begin{aligned} \mathbb{E} [\Phi(x_{t+1}) - f(x_{t+1}, y_{t+1})] & \leq \left(1 - \frac{\mu_y \eta_y}{6}\right) \mathbb{E} [\Phi(x_{t+1}) - f(x_{t+1}, y_t)] + \eta_y \delta^2 + \eta_y \sigma_y^2 \\ & = \left(1 - \frac{\mu_y \eta_y}{6}\right) \mathbb{E} [\Phi(x_t) - f(x_t, y_t) + f(x_t, y_t) - f(x_{t+1}, y_t) + \Phi(x_{t+1}) - \Phi(x_t)] \\ & \quad + \eta_y \delta^2 + \eta_y \sigma_y^2. \end{aligned}$$

Descent Upper bound on $f(\cdot, y)$

$$\begin{aligned}\mathbb{E}f(x_{t+1}, y_t) &\geq \mathbb{E}f(x_t, y_t) - \frac{3\eta_x}{2} \mathbb{E} \|G_{1/\eta_x}(x_t)\|_{\mathbf{M}_t}^2 - \frac{9\eta_x\sigma_x^2}{2} - \frac{7\eta_x\delta_x^2}{2} \\ &\geq \mathbb{E}f(x_t, y_t) - \frac{3\eta_x}{2} \mathbb{E}\mathcal{D}_{\mathcal{X}}(x_t, 1/\eta_x; y_t) - \frac{9\eta_x\sigma_x^2}{2} - \frac{7\eta_x\delta_x^2}{2}\end{aligned}$$

1070 Re-arranging,

$$\mathbb{E}f(x_t, y_t) - \mathbb{E}f(x_{t+1}, y_t) \leq \frac{3\eta_x}{2} \mathbb{E}\mathcal{D}_{\mathcal{X}}(x_t, 1/\eta_x; y_t) + \frac{9\eta_x\sigma_x^2}{2} + \frac{7\eta_x\delta_x^2}{2}$$

$$\begin{aligned}&\mathbb{E} [\Phi(x_{t+1}) - f(x_{t+1}, y_{t+1})] \\ &\leq \left(1 - \frac{\mu_y\eta_y}{6}\right) \mathbb{E} [\Phi(x_t) - f(x_t, y_t)] \\ &+ \left(1 - \frac{\mu_y\eta_y}{6}\right) \mathbb{E} \left[-\frac{\eta_x}{6} \mathbb{E}\mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x) + \eta_x \mathbb{E} \|\nabla_x \Phi(x_t) - \nabla_x f(x_t, y_t)\|_{\mathbf{M}_t^{-1}}^2 \right] \\ &+ \left(1 - \frac{\mu_y\eta_y}{6}\right) \mathbb{E} \left[\frac{3\eta_x}{2} \mathbb{E}\mathcal{D}_{\mathcal{X}}(x_t, 1/\eta_x; y_t) \right] \\ &+ \eta_y\delta_y^2 + \eta_y\sigma_y^2 + \eta_x \left(1 - \frac{\mu_y\eta_y}{6}\right) \left(\frac{13}{2}\sigma_x^2 + \frac{11}{2}\delta_x^2\right)\end{aligned}$$

$$\begin{aligned}&U_{t+1} + \alpha W_{t+1} \\ &\leq U_t - \frac{\eta_x}{6} \mathbb{E}\mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x) + \eta_x \mathbb{E} \|\nabla_x \Phi(x_t) - \nabla_x f(x_t, y_t)\|_{\mathbf{M}_t^{-1}}^2 \\ &\quad + \alpha \left(1 - \frac{\mu_y\eta_y}{6}\right) \mathbb{E} W_t \\ &\quad + \alpha \left(1 - \frac{\mu_y\eta_y}{6}\right) \mathbb{E} \left[-\frac{\eta_x}{6} \mathbb{E}\mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x) + \eta_x \mathbb{E} \|\nabla_x \Phi(x_t) - \nabla_x f(x_t, y_t)\|_{\mathbf{M}_t^{-1}}^2 \right] \\ &\quad + \alpha \left(1 - \frac{\mu_y\eta_y}{6}\right) \frac{3\eta_x}{2} \mathbb{E} [\mathcal{D}_{\mathcal{X}}(x_t, 1/\eta_x; y_t)] \\ &\quad + \alpha\eta_y\delta_y^2 + \alpha\eta_y\sigma_y^2 + \alpha\eta_x \left(1 - \frac{\mu_y\eta_y}{6}\right) \left(\frac{13}{2}\sigma_x^2 + \frac{11}{2}\delta_x^2\right) + 2\eta_x\sigma_x^2 + 2\eta_x\delta_x^2 \\ &\leq U_t - \frac{\eta_x}{6} \mathbb{E}\mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x) + \eta_x \mathbb{E} \|\nabla_x \Phi(x_t) - \nabla_x f(x_t, y_t)\|_{\mathbf{M}_t^{-1}}^2 \\ &\quad + \alpha \left(1 - \frac{\mu_y\eta_y}{6}\right) \mathbb{E} W_t \\ &\quad + \alpha \left(1 - \frac{\mu_y\eta_y}{6}\right) \mathbb{E} \left[-\frac{\eta_x}{6} \mathbb{E}\mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x) + \eta_x \mathbb{E} \|\nabla_x \Phi(x_t) - \nabla_x f(x_t, y_t)\|_{\mathbf{M}_t^{-1}}^2 \right] \\ &\quad + \alpha \left(1 - \frac{\mu_y\eta_y}{6}\right) \frac{3\eta_x}{2} \mathbb{E} [|\mathcal{D}_{\mathcal{X}}(x_t, 1/\eta_x; y_t) - \mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x)| + \mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x)] \quad (12) \\ &\quad + \alpha\eta_y\delta_y^2 + \alpha\eta_y\sigma_y^2 + \alpha\eta_x \left(1 - \frac{\mu_y\eta_y}{6}\right) \left(\frac{13}{2}\sigma_x^2 + \frac{11}{2}\delta_x^2\right) + 2\eta_x\sigma_x^2 + 2\eta_x\delta_x^2 \\ &\leq U_t - \frac{\eta_x}{6} \mathbb{E}\mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x) + \eta_x \mathbb{E} \|\nabla_x \Phi(x_t) - \nabla_x f(x_t, y_t)\|_{\mathbf{M}_t^{-1}}^2 \\ &\quad + \alpha \left(1 - \frac{\mu_y\eta_y}{6}\right) \mathbb{E} W_t \\ &\quad + \alpha \left(1 - \frac{\mu_y\eta_y}{6}\right) \mathbb{E} \left[-\frac{\eta_x}{6} \mathbb{E}\mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x) + \eta_x \mathbb{E} \|\nabla_x \Phi(x_t) - \nabla_x f(x_t, y_t)\|_{\mathbf{M}_t^{-1}}^2 \right] \\ &\quad + \alpha \left(1 - \frac{\mu_y\eta_y}{6}\right) \frac{3\eta_x}{2} \mathbb{E} \left[3\lambda_{\max}(\mathbf{M}_t^{-1})\ell^2 \|y_t - y^*(x_t)\|^2 + \mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x) \right] \quad (13) \\ &\quad + \alpha\eta_y\delta_y^2 + \alpha\eta_y\sigma_y^2 + \alpha\eta_x \left(1 - \frac{\mu_y\eta_y}{6}\right) \left(\frac{13}{2}\sigma_x^2 + \frac{11}{2}\delta_x^2\right) + 2\eta_x\sigma_x^2 + 2\eta_x\delta_x^2 \\ &\leq U_t - \frac{\eta_x}{6} \mathbb{E}\mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x) + \eta_x\lambda_{\max}(\mathbf{M}_t^{-1})\ell^2 \mathbb{E} \|y^*(x_t) - y_t\|^2 \\ &\quad + \alpha \left(1 - \frac{\mu_y\eta_y}{6}\right) \mathbb{E} W_t \\ &\quad + \alpha \left(1 - \frac{\mu_y\eta_y}{6}\right) \mathbb{E} \left[-\frac{\eta_x}{6} \mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x) + \eta_x\lambda_{\max}(\mathbf{M}_t^{-1})\ell^2 \|y^*(x_t) - y_t\|^2 \right]\end{aligned}$$

$$\begin{aligned}
& + \alpha \left(1 - \frac{\mu_y \eta_y}{6}\right) \frac{3\eta_x}{2} \mathbb{E} \left[3\lambda_{\max}(\mathbf{M}_t^{-1}) \ell^2 \|y_t - y^*(x_t)\|^2 + \mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x) \right] \\
& + \alpha \eta_y \delta_y^2 + \alpha \eta_y \sigma_y^2 + \alpha \eta_x \left(1 - \frac{\mu_y \eta_y}{6}\right) \left(\frac{13}{2} \sigma_x^2 + \frac{11}{2} \delta_x^2\right) + 2\eta_x \sigma_x^2 + 2\eta_x \delta_x^2
\end{aligned}$$

- 1071 • (12) uses the fact that $a \leq |a - b| + b$ for $a = \mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x; y_t)$, $b = \mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x)$
1072 • (13) uses Lemma B.5 and Daskin's theorem.

$$\begin{aligned}
U_{t+1} + \alpha W_{t+1} & \leq U_t - \frac{\eta_x}{6} \mathbb{E} \mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x) + \frac{2\eta_x \lambda_{\max}(\mathbf{M}_t^{-1}) \ell^2}{\mu_{\text{qg}}} W_t \\
& + \alpha \left(1 - \frac{\mu_y \eta_y}{6}\right) \mathbb{E} W_t \\
& + \alpha \left(1 - \frac{\mu_y \eta_y}{6}\right) \mathbb{E} \left[-\frac{\eta_x}{6} \mathbb{E} \mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x) + \frac{2\eta_x \lambda_{\max}(\mathbf{M}_t^{-1}) \ell^2}{\mu_{\text{qg}}} W_t \right] \\
& + \alpha \left(1 - \frac{\mu_y \eta_y}{6}\right) \frac{3\eta_x}{2} \mathbb{E} \left[\frac{6\lambda_{\max}(\mathbf{M}_t^{-1}) \ell^2}{\mu_{\text{qg}}} W_t \right] \\
& + \alpha \eta_y \delta_y^2 + \alpha \eta_y \sigma_y^2 + \alpha \eta_x \left(1 - \frac{\mu_y \eta_y}{6}\right) \left(\frac{13}{2} \sigma_x^2 + \frac{11}{2} \delta_x^2\right) + 2\eta_x \sigma_x^2 + 2\eta_x \delta_x^2 \\
& \leq \varpi_1 U_t + \alpha \varpi_2 W_t \\
& + \alpha \eta_y \delta_y^2 + \alpha \eta_y \sigma_y^2 + \alpha \eta_x \left(1 - \frac{\mu_y \eta_y}{3}\right) \left(\frac{13}{2} \sigma_x^2 + \frac{11}{2} \delta_x^2\right) + 2\eta_x \sigma_x^2 + 2\eta_x \delta_x^2
\end{aligned}$$

$$\begin{aligned}
\varpi_1 & := 1 - \mu_x \eta_x \left(\frac{1}{3} - \alpha \left(1 - \frac{\mu_y \eta_y}{6}\right) \frac{1}{3} + \alpha \left(1 - \frac{\mu_y \eta_y}{6}\right) 3 \right); \\
\varpi_2 & := 1 + \frac{2\eta_x \lambda_{\max}(\mathbf{M}_t^{-1}) \ell^2}{\alpha \mu_{\text{qg}}} - \frac{\mu_y \eta_y}{6} + \left(1 - \frac{\mu_y \eta_y}{6}\right) \frac{11\eta_x \lambda_{\max}(\mathbf{M}_t^{-1}) \ell^2}{\mu_{\text{qg}}}.
\end{aligned}$$

1073 For ϖ_1 , letting $\alpha = 1/10$

$$\begin{aligned}
\varpi_1 & = 1 - \mu_x \eta_x \left(\frac{1}{3} - \frac{1}{10} \left(1 - \frac{\mu_y \eta_y}{6}\right) \frac{1}{3} + \frac{1}{10} \left(1 - \frac{\mu_y \eta_y}{6}\right) 3 \right) \\
& = 1 - \mu_x \eta_x \frac{1}{3} - \mu_x \eta_x \frac{8}{30} \left(1 - \frac{\mu_y \eta_y}{6}\right) \leq 1 - \frac{\mu_x \eta_x}{3}.
\end{aligned}$$

1074 For ϖ_2 , we distinguish two cases relevant to our algorithms, $\mathbf{M}_t = \mathbf{I}$ and a general choice of \mathbf{M}_t .

- 1075 • For, $\mathbf{M}_t = \mathbf{I}$, it holds that $\lambda_{\max}(\mathbf{M}_t^{-1}) = 1$, and $\mu_{\text{qg}} = \mu_y$. So we write,

$$\begin{aligned}
\varpi_2 & = 1 + \frac{20\eta_x \ell^2}{\mu_y} - \frac{\mu_y \eta_y}{6} + \left(1 - \frac{\mu_y \eta_y}{6}\right) \frac{11\eta_x \ell^2}{\mu_y} \\
& = 1 - \frac{\eta_x \ell^2}{\mu_y} \left(-20 + \frac{\mu_y^2 \eta_y}{6\eta_x \ell^2} - 11 \left(1 - \frac{\mu_y \eta_y}{6}\right) \right) \\
& \leq 1 - \frac{\eta_x \ell^2}{\mu_y} (-20 + 32 - 11)
\end{aligned}$$

1076 Let $\frac{\mu_y^2 \eta_y}{\eta_x \ell^2} = 192$. Then, choosing $\eta_y = \frac{1}{5\ell}$ yields $\eta_x = \frac{\mu_y^3}{960\ell^3}$.

- 1077 • For a general choice of \mathbf{M}_t , let $\lambda_{\max} := \lambda_{\max}(\mathbf{M}_t^{-1})$ and $\overline{\mu_y} \leftarrow \min\{\mu_{\text{qg}}, \mu_y\}$,

$$\begin{aligned}
\varpi_2 & = 1 + \frac{20\eta_x \lambda_{\max} \ell^2}{\overline{\mu_y}} - \frac{\overline{\mu_y} \eta_y}{6} + \left(1 - \frac{\overline{\mu_y} \eta_y}{6}\right) \frac{11\eta_x \lambda_{\max} \ell^2}{\overline{\mu_y}} \\
& = 1 - \frac{\lambda_{\max} \eta_x \ell^2}{\overline{\mu_y}} \left(-20 + \frac{\overline{\mu_y}^2 \eta_y}{6\lambda_{\max} \eta_x \ell^2} - 11 \left(1 - \frac{\overline{\mu_y} \eta_y}{6}\right) \right)
\end{aligned}$$

1078

Similarly, we need to set,

$$\frac{\overline{\mu}_y^2 \eta_y}{\lambda_{\max} \eta_x \ell^2} = 192.$$

1079

which in turn yields, $\eta_x = \frac{\mu_y^3}{960 \ell^3 \lambda_{\max}}$.

1080 C Further Preliminaries on EFGs

1081 C.1 The behavioral and sequence-form strategies

1082 **Lemma C.1.** *Under γ -sufficient exploration, the transforms $c_1^{-1} : \mathcal{M}_1 \rightarrow \mathcal{X}_\gamma$, $c_2^{-1} : \mathcal{M}_2 \rightarrow \mathcal{Y}_\gamma$ are*
 1083 *Lipschitz continuous. I.e., for any μ_1, μ'_1 , it holds true that,*

$$\|c_1^{-1}(\mu_1) - c_1^{-1}(\mu'_1)\| \leq \frac{2\sqrt{A_{\max}}}{\gamma} \|\mu_1 - \mu'_1\|$$

1084 and for any μ_2, μ'_2 ,

$$\|c_2^{-1}(\mu_2) - c_2^{-1}(\mu'_2)\| \leq \frac{2\sqrt{B_{\max}}}{\gamma} \|\mu_2 - \mu'_2\|.$$

1085 *Proof.* We will first observe the difference in c_1^{-1} in the (s, a) -th entry of the the vector valued
 1086 mapping:

$$\begin{aligned} \frac{\mu_1(a|s)}{\mu_1(\sigma(s))} - \frac{\mu'_1(a|s)}{\mu'_1(\sigma(s))} &= \left(\frac{\mu_1(a|s)}{\mu_1(\sigma(s))} - \frac{\mu'_1(a|s)}{\mu_1(\sigma(s))} \right) + \left(\frac{\mu'_1(a|s)}{\mu_1(\sigma(s))} - \frac{\mu'_1(a|s)}{\mu'_1(\sigma(s))} \right) \\ &= \left(\frac{\mu_1(a|s)}{\mu_1(\sigma(s))} - \frac{\mu'_1(a|s)}{\mu_1(\sigma(s))} \right) + \left(\frac{1}{\mu_1(\sigma(s))} - \frac{1}{\mu'_1(\sigma(s))} \right) \mu'_1(a|s) \\ &= \left(\frac{\mu_1(a|s)}{\mu_1(\sigma(s))} - \frac{\mu'_1(a|s)}{\mu_1(\sigma(s))} \right) + \frac{\mu'_1(\sigma(s)) - \mu_1(\sigma(s))}{\mu_1(\sigma(s))\mu'_1(\sigma(s))} \mu'_1(a|s) \end{aligned}$$

1087 Proceeding towards the desired inequality,

$$\begin{aligned} &\|c^{-1}(\mu_1) - c^{-1}(\mu'_1)\|^2 \\ &= \sum_{s \in \mathcal{S}_1} \sum_{a \in \mathcal{A}_s} \left[\left(\frac{\mu_1(a|s)}{\mu_1(\sigma(s))} - \frac{\mu'_1(a|s)}{\mu_1(\sigma(s))} \right) + \frac{\mu'_1(\sigma(s)) - \mu_1(\sigma(s))}{\mu_1(\sigma(s))\mu'_1(\sigma(s))} \mu'_1(a|s) \right]^2 \\ &\leq 2 \sum_{s \in \mathcal{S}_1} \sum_{a \in \mathcal{A}_s} \left(\frac{\mu_1(a|s)}{\mu_1(\sigma(s))} - \frac{\mu'_1(a|s)}{\mu_1(\sigma(s))} \right)^2 + 2 \sum_{s \in \mathcal{S}_1} \sum_{a \in \mathcal{A}_s} \left(\frac{\mu'_1(\sigma(s)) - \mu_1(\sigma(s))}{\mu_1(\sigma(s))\mu'_1(\sigma(s))} \right)^2 \mu_1'^2(a|s) \\ &\leq 2 \sum_{s \in \mathcal{S}_1} \sum_{a \in \mathcal{A}_s} \left(\frac{\mu_1(a|s)}{\mu_1(\sigma(s))} - \frac{\mu'_1(a|s)}{\mu_1(\sigma(s))} \right)^2 + 2 \sum_{s \in \mathcal{S}_1} \sum_{a \in \mathcal{A}_s} \left(\frac{\mu'_1(\sigma(s)) - \mu_1(\sigma(s))}{\mu_1(\sigma(s))\mu'_1(\sigma(s))} \right)^2 \mu_1'^2(s) \\ &\leq \frac{2}{\gamma^2} \sum_{s \in \mathcal{S}_1} \sum_{a \in \mathcal{A}_s} (\mu_1(a|s) - \mu'_1(a|s))^2 + \frac{2}{\gamma^2} \sum_{s \in \mathcal{S}_1} \sum_{a \in \mathcal{A}_s} (\mu'_1(\sigma(s)) - \mu_1(\sigma(s)))^2 \\ &\leq \frac{2}{\gamma^2} \|\mu_1 - \mu'_1\|^2 + \frac{2A_{\max}}{\gamma^2} \sum_{s \in \mathcal{S}_1} (\mu'_1(\sigma(s)) - \mu_1(\sigma(s)))^2 \\ &= \frac{2}{\gamma^2} \|\mu_1 - \mu'_1\|^2 + \frac{2A_{\max}}{\gamma^2} \sum_{s \in \mathcal{S}_1} \left(\sum_{a \in \mathcal{A}_s} \mu'_1(a|s) - \mu_1(a|s) \right)^2. \end{aligned} \tag{14}$$

1088 We need to upper bound the second term by some quantity proportional to $\|\mu_1 - \mu'_1\|$. We first note
 1089 that by the triangular inequality,

$$\begin{aligned} \left| \sum_{a \in \mathcal{A}_s} \mu'_1(a|s) - \mu_1(a|s) \right| &\leq \sum_{a \in \mathcal{A}_s} |\mu'_1(a|s) - \mu_1(a|s)| \\ &\leq \sqrt{A_{\max}} \|\mu'_1(\cdot|s) - \mu_1(\cdot|s)\|. \end{aligned}$$

1090 where the last inequality is due to the fact that $\|x\|_1 \leq \sqrt{d} \|x\|$, $\forall x \in \mathbb{R}^d$. As such, we can note that,

$$\sum_{s \in \mathcal{S}_1} \left(\sum_{a \in \mathcal{A}_s} \mu'_1(a|s) - \mu_1(a|s) \right)^2 \leq \sum_{s \in \mathcal{S}_1} \left(\sqrt{A_{\max}} \|\mu'_1(\cdot|s) - \mu_1(\cdot|s)\| \right)^2$$

$$\begin{aligned}
&= A_{\max} \sum_{s \in \mathcal{S}_1} \sum_{a \in \mathcal{A}_s} (\mu'_1(a|s) - \mu_1(a|s))^2 \\
&= A_{\max} \|\mu'_1 - \mu_1\|^2.
\end{aligned}$$

1091 Plugging this inequality into (14) yields the desired bound. \square

1092 C.2 Value, Action-Value, and Advantage Functions

1093 **On notation.** In this subsection, we will use the following shorthand notations,

- 1094 • $\sigma_1(h), \sigma_2(h)$ returns the last history before h where player 1 (player 2, resp.) took an action,
- 1095 • $h \in s$ signifies that history h belongs in the info set s ,
- 1096 • $h' \succeq_{\mathcal{T}} h, h' \succeq_{\mathcal{T}} (h, a)$ signifies that h' is a successor node of $h, (h, a)$;
- 1097 • $h \in \tau, (h, a) \in \tau$ signifies that h, h, a belongs in the game trajectory τ from the root to a
- 1098 terminal node.

1099 **Occupancy measure** For a policy pair $\pi := (\pi_1, \pi_2)$, we define $d^\pi : \mathcal{S} \rightarrow [0, 1]$ to be a finite
1100 measure over all the info sets—summing over all info sets $s \in \mathcal{S}$ yields the depth of the game tree
1101 $D(\mathcal{T})$ —where for any info set $s \in \mathcal{S}$,

$$d^\pi(s) := \sum_{h \in s} \mu_c(h) \mu_1^{\pi_1}(\sigma_1(h)) \mu_2^{\pi_2}(\sigma_2(h)).$$

1102 The value function of each info set is defined as,

$$\begin{aligned}
V_1^\pi(s) &:= \mathbb{E}_\pi \left[\sum_{\tau} r_1(\tau) \mathbb{P}(\tau) \mid \exists h \in s : h \in \tau \right] \\
&= \frac{1}{\sum_{h \in s} \mu_c(h) \mu_1^{\pi_1}(\sigma_1(h)) \mu_2^{\pi_2}(\sigma_2(h))} \sum_{h' : \exists h \in s, h' \succeq_{\mathcal{T}} h} \mu_c(h') \mu_1^{\pi_1}(\sigma_1(h')) \mu_2^{\pi_2}(\sigma_2(h')) r_1(h')
\end{aligned}$$

1103 We define the advantage function,

$$A_1^\pi(s, a) := Q_1^\pi(s, a) - V_1(s).$$

1104 Finally, we compute the policy,

$$\begin{aligned}
\frac{\partial V_1^\pi}{\partial \theta_{s,a}} &= \frac{\partial}{\partial \theta_{s,a}} \sum_{\tau} r_1(\tau) \mathbb{P}(\tau) \\
&= \sum_{\tau} r_1(\tau) \mathbb{P}(\tau) \frac{\partial \log \mathbb{P}(\tau)}{\partial \theta_{s,a}} \\
&= \sum_{\tau} \sum_{a'} r_1(\tau) \mathbb{P}(\tau) \frac{\partial \log \pi_1(a'|s)}{\partial \theta_{s,a}} \mathbb{1}\{\exists h \in s : (h, a') \in \tau\} \\
&= \sum_{\tau} \sum_{a'} \left(r_1(\tau) \mathbb{P}(\tau) \frac{1}{\pi_1(a'|s)} \right) \pi_1(a'|s) \frac{\partial \log \pi_1(a'|s)}{\partial \theta_{s,a}} \mathbb{1}\{\exists h \in s : h \in \tau\} \\
&= \left(\sum_{h \in s} \sum_{a'} \mu_c(h) \mu_1^{\pi_1}(\sigma_1(h)) \mu_2^{\pi_2}(\sigma_2(h)) Q(s, a') \right) \pi_1(a'|s) \frac{\partial \log \pi_1(a'|s)}{\partial \theta_{s,a}} \\
&= d^\pi(s) \sum_{a'} \pi_1(a'|s) \frac{\partial \log \pi_1(a'|s)}{\partial \theta_{s,a}} Q(s, a'). \tag{15}
\end{aligned}$$

1105 For direct policy parametrization, we get,

$$\frac{\partial V_1^\pi}{\partial \pi_1(s, a)} = d^\pi(s) Q(s, a).$$

1106 For the softmax policy parametrization, (15) yields,

$$\begin{aligned}
\frac{\partial V_1^\pi}{\partial \theta_{s,a}} &= d^\pi(s) \sum_{a'} \pi_1(a'|s) \frac{\partial \log \pi_1(a'|s)}{\partial \theta_{s,a}} Q(s, a') \\
&= d^\pi(s) \sum_{a'} \pi_1(a'|s) [\mathbb{1}\{a' = a\} - \pi_1(a'|s)] Q(s, a') \\
&= d^\pi(s) \pi_1(a|s) [Q(s, a) - V(s)] \\
&= d^\pi(s) \pi_1(a|s) A(s, a).
\end{aligned}$$

1107 C.3 Properties of the bidilated regularizer

1108 Introduced in (Liu et al., 2024), the bidilated regularizer offers an alternative to the commonly used
1109 dilated regularizer (Hoda et al., 2010). It can be seamlessly used along Q feedback by dropping the
1110 need of importance sampling which would be necessary for the *dilated regularizer* when the gradient
1111 is estimated through trajectory roll-outs. The purpose of this refined regularizer was introducing
1112 a distance generating function in the sequence-form space that would not necessitate importance
1113 sampling.

1114 C.3.1 Strong Convexity Modulus

1115 **Lemma C.2.** *For a choice of strongly convex function ψ , and a weighting scheme $\{w_{1,s}\}_{s \in \mathcal{S}_1}$,*
1116 *$\{w_{2,s}\}_{s \in \mathcal{S}_2}$ and*

$$\alpha_{\text{bi}} := \gamma \min_h \mu_c(h) \alpha_{\text{dil}}$$

1117 *Proof.* For an appropriate choice of weights $\{w_{1,s}\}_{s \in \mathcal{S}_1}$, $\{w_{2,s}\}_{s \in \mathcal{S}_2}$, the *weighted* bidilated regular-
1118 izer is defined as,

$$\begin{aligned}
\mathcal{R}_1^\psi(\mu_1^{\pi_1}, \mu_2^{\pi_2}) &:= \sum_s \mu_1^{\pi_1}(\sigma_1(s)) \left(\sum_{h \in s} \mu_c(h) \mu_2^{\pi_2}(\sigma_2(h)) \right) w_{1,s} \psi(\pi_1(\cdot|s)) \\
\mathcal{R}_2^\psi(\mu_1^{\pi_1}, \mu_2^{\pi_2}) &:= \sum_s \mu_2^{\pi_2}(\sigma_2(s)) \left(\sum_{h \in s} \mu_c(h) \mu_2^{\pi_2}(\sigma_1(h)) \right) w_{2,s} \psi(\pi_2(\cdot|s)).
\end{aligned}$$

1119 We can slightly refine (Liu et al., 2024, Lemma C.1) in order to compute an explicit lower bound on
1120 the convexity modulus of different weighted bidilated regularizer depending on the choice of ψ .

1121 From the fact that $\mathcal{R}_1(\mu_1^{\pi_1}, \mu_2^{\pi_2})$ is linear in $\mu_2^{\pi_2}$ and the definition of the Bregman divergence, we
1122 conclude that,

$$\begin{aligned}
&\left\langle \nabla(\mathcal{R}_1 + \mathcal{R}_2)(\mu_1^{\pi_1}, \mu_2^{\pi_2}) - \nabla(\mathcal{R}_1 + \mathcal{R}_2)(\mu_1^{\pi'_1}, \mu_2^{\pi'_2}), (\mu_1^{\pi_1}, \mu_2^{\pi_2}) - (\mu_1^{\pi'_1}, \mu_2^{\pi'_2}) \right\rangle \\
&\geq B_{\mathcal{R}_1^\psi}(\mu_1^{\pi'_1} \| \mu_1^{\pi_1}; \mu_2^{\pi_2}) + B_{\mathcal{R}_1^\psi}(\mu_1^{\pi_1} \| \mu_1^{\pi'_1}; \mu_2^{\pi_2}) + B_{\mathcal{R}_2^\psi}(\mu_2^{\pi_2} \| \mu_2^{\pi'_2}; \mu_1^{\pi_1}) + B_{\mathcal{R}_2^\psi}(\mu_2^{\pi'_2} \| \mu_2^{\pi_2}; \mu_1^{\pi_1}).
\end{aligned}$$

1123 By (Liu et al., 2022c, Lemma D.2) we know that,

$$B_{\mathcal{R}_1^\psi}(\mu_1^{\pi'_1} \| \mu_1^{\pi_1}; \mu_2^{\pi_2}) \geq \gamma \min_h \mu_c(h) B_\psi^{\text{dil}}(\mu_1^{\pi'_1} \| \mu_1^{\pi_1}).$$

1124 As such, for the strong convexity modulus of the weighted \mathcal{R}_1^ψ relative to the choice of norm
1125 appropriate for ψ , we write,

$$\alpha_{\text{bi}} := \gamma \min_h \mu_c(h) \alpha_{\text{dil}}.$$

1126 □

1127 By (Farina et al., 2019, Corollary 1), we know that there exists a weighting scheme, such that the
1128 Euclidean dilated regularizer is 1-strongly convex w.r.t. the ℓ_2 -norm.

1129 **Corollary C.1** (Euclidean Regularizer). *There exists a choice of weights, with*
 1130 $\max_s w_{1,s}, \max_s w_{2,s} = \Theta(2^{D(\mathcal{T})})$, *and under the assumption that* $\min_s \mu_2(s) \geq \gamma$, *the*
 1131 *bidilated Euclidean regularizer has a strong convexity modulus w.r.t. the ℓ_2 -norm, α_{bi} ,*

$$\alpha_{\text{bi}} := \gamma \min_h \mu_c(h).$$

1132 (Kroer et al., 2020, Theorem 2) states that a recursion defines weights with $\max_s w_{1,s}, \max_s w_{2,s} =$
 1133 $\Theta(2^{D(\mathcal{T})})$ such that the entropic dilated regularizer is strongly convex w.r.t. the ℓ_2 -norm.

1134 **Corollary C.2** (Entropic Regularizer). *There exists a choice of weights, and under the assumption*
 1135 *that* $\min_s \mu_2(s) \geq \gamma$, *the bidilated entropic regularizer has a strong convexity modulus w.r.t. the*
 1136 *ℓ_2 -norm, α_{bi} ,*

$$\alpha_{\text{bi}} := \gamma \min_h \mu_c(h).$$

1137 C.3.2 Lipschitz Moduli

1138 Here, we concern ourselves with the Lipschitz continuity of the regularizers and that of their gradients.

1139 Euclidean regularizer

1140 **Lemma C.3.** *The Euclidean bidilated regularizer is $\Theta(2^{D(\mathcal{T})} D(\mathcal{T}) \sqrt{|\mathcal{S}|})$ -smooth.*

Proof.

$$\mathcal{R}_1^{\text{eucl}}(\pi_1, \pi_2) := \langle f(\pi_1), g(\pi_1) \rangle$$

1141 where, $f(\pi_1), g(\pi_1) \in \mathbb{R}^{|\mathcal{S}_1|}$ with $f_s(\pi_1) = \sum_{h \in \mathcal{S}} \mu_c(h) \mu_2^{\pi_2}(\sigma(h)) \mu_1^{\pi_1}(\sigma(h))$ and $g_s(\pi_1) =$
 1142 $w_s \|\pi_1(\cdot|s)\|^2$.

$$\begin{aligned} & \|\nabla_{\pi_1} \mathcal{R}_1^{\text{eucl}}(\pi_1, \pi_2) - \nabla_{\pi_1} \mathcal{R}_1^{\text{eucl}}(\pi'_1, \pi_2)\| \\ & \leq \|(\mathbf{J}_f(\pi_1) - \mathbf{J}_f(\pi'_1))\| \|g(\pi_1)\| + \|\mathbf{J}_f(\pi'_1)\| \|g(\pi_1) - g(\pi'_1)\| \\ & + \|\mathbf{J}_g(\pi_1) - \mathbf{J}_g(\pi'_1)\| \|f(\pi_1)\| + \|\mathbf{J}_g(\pi'_1)\| \|f(\pi_1) - f(\pi'_1)\| \\ & \leq (\ell_f \sqrt{S_1} + L_f L_g + \ell_g \max_{\pi_1} \|f(\pi_1)\| + L_g) \|\pi_1 - \pi'_1\| \end{aligned}$$

1143 • For g , we see that $L_g := 2 \max_s w_s$ and $\ell_g := 2 \max_s w_s$ by the properties of the weighted
 1144 ℓ_2 -norm and the fact that $\pi_1(\cdot|s)$ lies in the simplex.

1145 • For f , it is easy to see that $L_f, \ell_f \leq D(\mathcal{T})$ since f is a multilinear function of non-negative
 1146 variables bounded by 1. Also, it holds that $\max_{\pi_1} \|f(\pi_1)\| \leq \sqrt{D(\mathcal{T})}$.

1147 Concluding,

$$\|\nabla_{\pi_1} \mathcal{R}_1^{\text{eucl}}(\pi_1, \pi_2) - \nabla_{\pi_1} \mathcal{R}_1^{\text{eucl}}(\pi'_1, \pi_2)\| \leq 64 \max_s w_s D(\mathcal{T}) \sqrt{S_1} \|\pi_1 - \pi'_1\|.$$

1148 Symmetrically,

$$\|\nabla_{\pi_2} \mathcal{R}_2^{\text{eucl}}(\pi_1, \pi_2) - \nabla_{\pi_2} \mathcal{R}_2^{\text{eucl}}(\pi_1, \pi'_2)\| \leq 64 \max_s w'_s D(\mathcal{T}) \sqrt{S_2} \|\pi_2 - \pi'_2\|.$$

1149 Now, we need to bound the smoothness modulus of $\nabla_{\pi_1} \mathcal{R}_2^{\text{eucl}}(\pi_1, \pi_2)$. Similarly, we write,

$$\mathcal{R}_2^{\text{eucl}}(\pi_1, \pi_2) := \langle f(\pi_1), g \rangle.$$

1150

$$\begin{aligned} \|\nabla_{\pi_1} \mathcal{R}_2^{\text{eucl}}(\pi_1, \pi_2) - \nabla_{\pi_1} \mathcal{R}_2^{\text{eucl}}(\pi'_1, \pi_2)\| & \leq \|\mathbf{J}_f(\pi_1) - \mathbf{J}_f(\pi'_1)\| \|g\| \\ & \leq 2 \max_s w'_s D(\mathcal{T}) \|\pi_1 - \pi'_1\|. \end{aligned}$$

1151

□

1152 Entropic regularizer

Lemma C.4. *The weighted entropic bidilated regularizer is ℓ -smooth with*

$$\ell := 128 \max_{i \in \{1,2\}} \{ \max_s w_{1,s} \} (1 + \log A_{\max}) \sqrt{S_1} D(\mathcal{T})$$

1153 .

Proof.

$$\mathcal{R}_2(\pi_\chi, \pi_\theta) := \langle f(\pi_\chi), g \rangle.$$

1154

$$\begin{aligned} \|\nabla_\chi \mathcal{R}_2(\pi_\chi, \pi_\theta) - \nabla_\chi \mathcal{R}_2(\pi_{\chi'}, \pi_\theta)\| &\leq \|\mathbf{J}_\pi(\chi)^\top \mathbf{J}_f(\pi_\chi) - \mathbf{J}_\pi(\chi')^\top \mathbf{J}_f(\pi_{\chi'})\| \|g\| \\ &\leq (\|\mathbf{J}_\pi(\chi)^\top \mathbf{J}_f(\pi_\chi) - \mathbf{J}_\pi(\chi)^\top \mathbf{J}_f(\pi_{\chi'})\| + \|\mathbf{J}_\pi(\chi)^\top \mathbf{J}_f(\pi_{\chi'}) - \mathbf{J}_\pi(\chi')^\top \mathbf{J}_f(\pi_{\chi'})\|) \|g\| \\ &\leq (\|\mathbf{J}_\pi(\chi)\| \|\mathbf{J}_f(\pi_\chi) - \mathbf{J}_f(\pi_{\chi'})\| + \|\mathbf{J}_f(\pi_{\chi'})\| \|\mathbf{J}_\pi(\chi) - \mathbf{J}_\pi(\chi')\|) \|g\| \\ &\leq (D(\mathcal{T}) + \frac{3}{4} D(\mathcal{T}) (S_1 A_{\max})^{3/2}) \max_s w_{2,s} \sqrt{\log B_{\max}} \|\chi - \chi'\|. \end{aligned}$$

1155

$$\begin{aligned} \|\nabla_\chi \mathcal{R}_2(\pi_\chi, \pi_\theta) - \nabla_\chi \mathcal{R}_2(\pi_{\chi'}, \pi_\theta)\| &\leq \|(\mathbf{J}_f(\chi) - \mathbf{J}_f(\chi'))\| \|g(\chi)\| + \|\mathbf{J}_f(\chi')\| \|g(\chi) - g(\chi')\| \\ &\quad + \|\mathbf{J}_g(\chi) - \mathbf{J}_g(\chi')\| \|f(\chi)\| + \|\mathbf{J}_g(\chi')\| \|f(\chi) - f(\chi')\| \end{aligned}$$

1156 Now, we re-purpose the lengthy calculations found in the proof of (Mei et al., 2020, Lemma 14), we
1157 consider $\chi = \chi_0 + \alpha u$ for some $u, \chi \in \mathbb{R}^A, \alpha \in \mathbb{R}$,

$$\left\| \frac{dg(\chi + \alpha u)}{d\alpha} \right\|_\infty \leq \max_s w_{1,s} \log A_{\max} \|u\|_2;$$

1158 hence, (since $\|x\|_2 \leq \sqrt{S_1} \|x\|_\infty$),

$$\left\| \frac{d^2 g(\chi + \alpha u)}{d\alpha^2} \right\|_2 \leq \max_s w_{1,s} \log A_{\max} \sqrt{S_1} \|u\|_2,$$

1159 or, $L_f = \log A_{\max} \sqrt{S_1}$. Similarly,

$$\left\| \frac{d^2 g(\chi + \alpha u)}{d\alpha^2} \right\|_\infty \leq 3 \max_s w_{1,s} (1 + \log A_{\max}) \|u\|_2;$$

1160 and, as such,

$$\left\| \frac{d^2 g(\chi + \alpha u)}{d\alpha^2} \right\|_2 \leq 3 \max_s w_{1,s} (1 + \log A_{\max}) \sqrt{S_1} \|u\|_2,$$

1161 or, $\ell_f = 3 \max_s w_{1,s} (1 + \log A_{\max}) \sqrt{S_1}$. □

1162 D Regarding the Policy Parametrization

1163 D.1 Definitions

1164 **Direct policy parametrization.** Both players, parameterize their policies (or, behavioral strategies),
1165 $\pi_1 : \mathcal{S}_1 \rightarrow \mathcal{A}$ and $\pi_2 : \mathcal{S}_2 \rightarrow \mathcal{B}$, using a concatenation of $|\mathcal{S}_1|$ and $|\mathcal{S}_2|$ probability vectors over the
1166 (potentially truncated) probability simplex $\Delta(\mathcal{A}_s), \Delta(\mathcal{B}_s)$ for all s in \mathcal{S}_1 and \mathcal{S}_2 correspondingly.
1167 The parameter space of player 1 is denoted with $\mathcal{X} := \prod_{s \in \mathcal{S}_1} \Delta(\mathcal{A}_s)$, while the parameter space of
1168 player 2 with $\mathcal{Y} := \prod_{s \in \mathcal{S}_2} \Delta(\mathcal{B}_s)$.

1169 **Softmax policy parametrization.** Softmax parametrized policies have a well-known definition.
 1170 The parameters of the corresponding policies are denoted χ, θ with $\chi \in \mathbb{R}^A, A = \sum_s A_s$ and
 1171 $\theta \in \mathbb{R}^B, B = \sum_s B_s$. For each info set s , the policy is,

$$\pi_\chi(a|s) = \frac{\exp(\chi_{s,a})}{\sum_{a'} \exp(\chi_{s,a'})} \quad \text{or} \quad \pi_\theta(b|s) = \frac{\exp(\theta_{s,b})}{\sum_{b'} \exp(\theta_{s,b'})}.$$

1172 Now, since we want to have control over the minimum eigenvalue of the Jacobian of $\text{softmax}(\cdot)$, we
 1173 restrict the parameter space to the following convex polytopes,

$$X_R := \left\{ \chi \in \mathbb{R}^A, A = \sum_s A_s : \chi_s^\top \mathbf{1} = 0, \forall s \in \mathcal{S}_1, |\chi_{s,i} - \chi_{s,j}| \leq 2R, \forall i, j \in [A_s] \right\};$$

$$\Theta_R := \left\{ \theta \in \mathbb{R}^B, B = \sum_s B_s : \theta_s^\top \mathbf{1} = 0, \forall s \in \mathcal{S}_2, |\theta_{s,i} - \theta_{s,j}| \leq 2R, \forall i, j \in [B_s] \right\}.$$

1174 D.1.1 General properties under parameter constraints

1175 **Lemma D.1.** Let $\mathbf{J} := \mathbf{J}_{\text{softmax}}(\theta) \in \mathbb{R}^{d \times d}$ be the jacobian of the softmax map. Its matrix form is:

$$\mathbf{J} = \text{diag}(\text{softmax}(\theta)) - \text{softmax}(\theta) \text{softmax}(\theta)^\top.$$

Further, the vector $\mathbf{1}$ is an eigenvector of \mathbf{J} with a corresponding eigenvalue of 0. The rest of the eigenvalues are

$$\lambda_i \in \left[\min_{i \in [d]} \text{softmax}_i(\theta), \max_{i \in [d]} \text{softmax}_i(\theta) \right].$$

1176

1177 *Proof.* For brevity, define $\sigma := \text{softmax}(\theta)$, and let $\text{diag}(v)$ be the $d \times d$ diagonal matrix whose
 1178 diagonal is equal to the entries $v \in \mathbb{R}^d$,

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

1179 First, we observe that the all-ones vector $\mathbf{1} \in \mathbb{R}^d$ is an eigenvector of \mathbf{J} with a corresponding
 1180 eigenvalue of 0,

$$\begin{aligned} \mathbf{J} &= \text{diag}(\sigma) \mathbf{1} - \sigma \sigma^\top \mathbf{1} \\ &= \sigma - \sigma(\sigma^\top \mathbf{1}) \\ &= \sigma - \sigma = 0. \end{aligned}$$

1181 By Weyl's inequality for two Hermitian matrices, A, B , we know that their eigenvalues indexed in a
 1182 descending order $\lambda_1(A) \geq \dots \geq \lambda_d(A)$ satisfy,

$$\lambda_{i+j-1}(A+B) \leq \lambda_i(A) + \lambda_j(B) \leq \lambda_{i+j-n}(A+B).$$

1183 $\lambda_i(\text{diag}(\sigma)) = \sigma_i^\downarrow$ while $\lambda_d(-\sigma \sigma^\top) = -\|\sigma\|_2^2 \in \left[-1, -\frac{1}{\sqrt{d}}\right]$. Hence,

- 1184 • $\lambda_{\min}^+(\mathbf{J}) \geq \min_{i \in [d]} \sigma_i(\theta)$ — by taking $i = d$ and $j = d - 1$;
- 1185 • $\sigma_2^\downarrow \leq \lambda_{\max}(\mathbf{J}) \leq \max_{i \in [d]} \sigma_i(\theta)$ — by taking $i = 2, j = 1$ for the LHS and $i = 1, j = 1$
 1186 for the RHS.

1187

□

1188 **Lemma D.2.** The softmax map is $\frac{3}{4}d^{3/2}$ -smooth.

Proof.

$$\frac{\partial^2 \sigma_i(x)}{\partial x_j \partial x_k} = \mathbb{1}\{i = j\} J_{ik} - J_{ik} \sigma_j(x) - \sigma_i(x) J_{jk}(x)$$

1189 as such $\left| \frac{\partial^2 \sigma_i(x)}{\partial x_j \partial x_k} \right| \leq \frac{3}{4}$. Then, $\|\nabla_x^3 \text{softmax}(x)\|_{\text{op}} \leq \frac{3}{4}d^{3/2}$. □

1190 **Lemma D.3.** Assume $\theta \in \mathbb{R}^d$ with $\theta \in \Theta_R := \{\theta \in \mathbb{R}^d : \theta^\top \mathbf{1} = 0 \text{ and } |\theta_i - \theta_j| \leq R, \forall i, j \in [d]\}$.
 1191 Then, the following bounds hold true,

$$\begin{aligned} 1192 \quad & \bullet \min_{i \in [d]} \text{softmax}_i(\theta) \geq \frac{1}{1 + (d-1)e^{2R}}; \\ 1193 \quad & \bullet \max_{i \in [d]} \text{softmax}_i(\theta) \geq \frac{1}{1 + (d-1)e^{-2R}}. \end{aligned}$$

1194 *Proof.*

1195 **Minimum probability lower bound.** W.l.o.g. we minimize the first coordinate. We write,

$$\frac{e^{\theta_1}}{\sum_i e^{\theta_i}} = \frac{1}{1 + \sum_{i>1} e^{\theta_i - \theta_1}}.$$

1196 By observing that,

$$e^{\theta_i - \theta_1} \leq \max_j e^{\theta_j - \theta_1}$$

1197 We can lower bound the value as,

$$\frac{e^{\theta_1}}{\sum_i e^{\theta_i}} \geq \frac{1}{1 + (d-1) \max_j \{e^{\theta_j - \theta_1}\}}$$

1198 It suffices to maximize the quantity $\max_{j \neq 1, \theta \in \Theta_R} \{\theta_j - \theta_1\}$ as the RHS quantity is non-increasing in
 1199 $\max_{j \neq 1, \theta \in \Theta_R} \{\theta_j - \theta_1\}$. I.e., the largest difference between two coordinates of a vector in the sphere
 1200 is $2R$. The minimum is achieved when $\theta_j - \theta_1 = 2R$ and $\theta_j = \theta_k, \forall j, k \geq 2$.

1201

1202 **Maximum probability lower bound.** Similarly, w.l.o.g, it suffices to maximize $\text{softmax}_1(\theta)$ for
 1203 $\theta \in \Theta_R$.

$$\begin{aligned} \frac{e^{\theta_1}}{\sum_i e^{\theta_i}} &= \frac{e^{\theta_1}}{e^{\theta_1} + \sum_{i \neq 1} e^{\theta_i}} \\ &\leq \frac{e^{\theta_1}}{e^{\theta_1} + (d-1)e^{\sum_i \theta_i / (d-1)}} \end{aligned}$$

1204 where the inequality follows from the convexity of e^x . For any $\theta \in \Theta_R$ the point $(\bar{\theta}) =$
 1205 $(\theta_1, \dots, \frac{\theta_i}{d-1}, \dots)$ is also in Θ_R due to the convexity of the set (it is a linear polytope). We can
 1206 simply optimize the objective,

$$\begin{aligned} &\max_{a,b} \frac{1}{1 + (d-1)e^{b-a}} \\ &\text{s.t. } |a - b| \leq R. \end{aligned}$$

1207 Due to the objective function's monotonicity in $b - a$, the program can be simplified even more into,

$$\begin{aligned} &\min_{a,b} b - a \\ &\text{s.t. } |a - b| \leq R. \end{aligned}$$

1208 Finally, it is clear that the last objective is minimized for $a - b = -2R$.

1209 □

1210 **Proposition 3.** Let p be a probability vector in Δ^{d-1} and define $\theta(p)$ to be the set of θ such that,
 1211 $\text{softmax}(\theta) = p$. For any two $\theta, \theta' \in \theta(p)$, there exists a $c \in \mathbb{R}$, such that $\theta = \theta' + c\mathbf{1}$.

1212 *Proof.* By assumption, $\text{softmax}(\theta) = \text{softmax}(\theta') = p$. For every entry i ,

$$p_i = \frac{e^{\theta_i}}{\sum_i e^{\theta_i}} = \frac{e^{\theta'_i}}{\sum_i e^{\theta'_i}}.$$

1213 Letting $Z := \sum_i^d e^{\theta_i}$, $Z' := \sum_i^d$, we observe,

$$\frac{e^{\theta_i}}{e^{\theta'_i}} = \frac{Z'}{Z} \implies \theta_i = \theta'_i + \log \frac{Z'}{Z}, \forall i \in \{1, \dots, d\}.$$

1214 Hence, any two θ, θ' that map to the same probability vector are translations of each other in the
1215 direction of $\mathbf{1}$. \square

1216 **Proposition 4.** Let a probability vector $p \in \Delta^{d-1}$ and the set, $\theta(p)$, of vectors $\theta \in \mathbb{R}^d$ such that
1217 $\text{softmax}(\theta) = p$. For the vector $\theta^* := \arg \min_{\theta \in \theta(p)}$ it holds true that,

$$\mathbf{1}^\top \theta = 0.$$

1218

1219 *Proof.* The set $\theta(p)$ takes the form $\theta(p) := \{(\theta_i = \log p_i + c) \mid c \in \mathbb{R}\} = \{\theta_0 + c\mathbf{1} \mid c \in \mathbb{R}\}$.
1220 Picking an arbitrary $\theta_0 \in \theta(p)$ to use as a reference, we can write the problem of minimizing $\|\theta\|_2$ as,

$$\min_{\theta \in \theta(p)} \|\theta\|^2 \equiv \min_{c \in \mathbb{R}} \|\theta_0 + c\mathbf{1}\|_2^2 \equiv \min_{c \in \mathbb{R}} \|\theta_0\|^2 + \langle \theta_0, c\mathbf{1} \rangle + \|c\mathbf{1}\|^2.$$

1221 By the first-order optimality conditions, $c = -\frac{1}{d}\theta_0^\top \mathbf{1}$. Plugging back this for θ^* , we see $\theta^* =$
1222 $\theta_0 - \mathbf{1}(\theta_0^\top \mathbf{1})$. We see that, $\mathbf{1}^\top \theta^* = \mathbf{1}^\top \theta_0 - \frac{d}{d}\theta_0^\top \mathbf{1} = 0$. \square

1223 **Lemma D.4.** Assume a fixed $0 < R < \infty$ and define the set Θ_R to be $\Theta_R := \{\theta \in \mathbb{R}^d : \theta^\top \mathbf{1} =$
1224 $0 \text{ and } |\theta_i - \theta_j| \leq R, \forall i, j \in [d]\}$. Then, $\text{softmax}(\Theta_R)$ is a convex set.

1225 *Proof.* For any $p \in \Delta^{d-1}$ for which $e^{-2R} \leq \frac{p_i}{p_j} \leq e^{2R}, \forall i, j \in [d]$, there exists $\theta \in \Theta_R$ such that
1226 $\text{softmax}(\theta) = p$. To see this, we apply the logarithm on the inequalities,

$$-2R \leq \log p_i - \log p_j \leq 2R. \quad (16)$$

1227 A vector χ with entries $\chi_i := \log p_i$ clearly implements p . By (16) we see that subtracting $\kappa =$
1228 $\frac{\max_j \log p_j + \min_k \log p_k}{2}$ from all entries yields a softmax-equivalent vector $\chi'_i := \log p_i - \kappa$ with
1229 $-R \leq \chi'_i \leq R$. Conversely, for any $\theta \in \Theta_R$, $e^{-2R} \leq \frac{\text{softmax}_i(\theta)}{\text{softmax}_j(\theta)} \leq e^{2R}$.

1230 Now, the set defined by the inequalities $p \in \Delta^{d-1}, e^{-2R} \leq \frac{p_i}{p_j} \leq e^{2R}$, is clearly a linear polytope
1231 and as such, convex. \square

1232 E Gradient Domination

1233 E.1 Direct Policy Patametrization

1234 **Lemma E.1.** The utility satisfies the pPLcondition for directly parametrized policies,

$$\begin{aligned} \tau \min_h \mu_c(h) \gamma^2 [V(x, y) - V(x^*(y), y)] &\leq \mathcal{D}_{\mathcal{X}}(x, \ell; y); \\ \tau \min_h \mu_c(h) \gamma^2 [V(x, y^*(x)) - V(x, y)] &\leq \mathcal{D}_{\mathcal{Y}}(y, \ell; x). \end{aligned}$$

1235

1236 *Proof.* We write the utility function of the regularized game,

$$H_\tau^{\text{eucl}}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) := \langle \mu_1^{\pi_1}, \mathbf{R} \mu_2^{\pi_2} \rangle - \tau \mathcal{R}_1^{\text{eucl}}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) + \tau \mathcal{R}_1^{\text{eucl}}(\mu_1^{\pi_1}, \mu_2^{\pi_2})$$

1237 For player 1, we know that the functions is strongly convex with an appropriate weight scheme
1238 $\{w_1, s\}, \{w_2, s\}$,

$$H_\tau^{\text{eucl}}(\mu'_1, \mu_2) \geq H_\tau^{\text{eucl}}(\mu_1, \mu_2) + \langle \nabla_{\mu_1} H_\tau^{\text{eucl}}(\mu_1, \mu_2), \mu'_1 - \mu_1 \rangle + \frac{\alpha_{\text{bi}}}{2} \|\mu_1 - \mu_2\|_2^2$$

1239 Strong convexity implies the KL condition for μ_1 . In turn, using the bound on the Lipschitz continuity
1240 modulus of the map $\mu_1 \mapsto x$,

$$H_\tau^{\text{eucl}}(\mu_1) - H_\tau^{\star, \text{eucl}} \leq \frac{1}{2\tau\gamma^2} \|s_x\|_2^2$$

1241 \square

1242 E.2 Softmax policy parametrization

1243 **Lemma E.2.** *The utility of the game with softmax-parametrized policies satisfies the two-sided*
 1244 *pPL condition,*

$$\begin{aligned} 2\tau\alpha_c^2 [V(\chi, \theta) - V(\chi^*, \theta)] &\leq \frac{1}{2} \mathcal{D}_{X_R}(\chi, \ell; \theta) \\ 2\tau\alpha_c^2 [V(\chi, \theta^*) - V(\chi, \theta)] &\leq \frac{1}{2} \mathcal{D}_{\Theta_R}(\theta, \ell; \chi). \end{aligned}$$

1245

1246 *Proof.* Similar to the previous argument as such omitted. \square

1247 E.3 Mahalanobis-PL

1248 **Lemma E.3.** *The utility of the game with softmax-parametrized policies satisfies the two-sided*
 1249 *Mahalanobis pPL condition,*

$$\begin{aligned} 2\tau\alpha_c^2 \frac{1}{\lambda_{\max}(\mathbf{M}^{-1})} [V(\chi, \theta) - V(\chi^*, \theta)] &\leq \frac{1}{2} \mathcal{D}_{X_R}(\chi, \ell; \theta) \\ 2\tau\alpha_c^2 \frac{1}{\lambda_{\max}(\mathbf{M}^{-1})} [V(\chi, \theta^*) - V(\chi, \theta)] &\leq \frac{1}{2} \mathcal{D}_{\Theta_R}(\theta, \ell; \chi). \end{aligned}$$

1250

1251 *Proof.* The proof follows from naively getting the Mahalanobis-KL conditions from the ℓ_2 KL and
 1252 then upper bounding it by the MFBM. \square

1253 F Gradient Estimators

1254 F.1 A Policy Gradient Theorem

1255 We define a trajectory τ to be a sequence of consecutive history-action pairs, $\tau =$
 1256 $((h^{(1)}, a_{i(1)}^{(1)}), (h^{(2)}, a_{i(2)}^{(2)}), \dots)$. The length of trajectory τ is noted as K_τ and it is bounded by the
 1257 game-tree's height, $D(\mathcal{T})$. We define \mathcal{K} to be the set of all trajectories and note that it is finite. After
 1258 a policy profile, (π_1, π_2) , is fixed, the probability of each trajectory $\tau \in \mathcal{K}$ taking place is the product
 1259 of the probability of each consecutive action,

$$\mathbb{P}^{\pi_1, \pi_2}(\tau) := \prod_{k=1}^{K_\tau} \pi_{i(k)}(a_{i(k)}^{(k)} | h^{(k)}).$$

1260 where $i(k)$ denotes the player that takes an action at timestep k .

1261 **Lemma F.1.** *Under the assumption of (I), it holds true that the gradient estimator (REINFORCE) is*
 1262 *unbiased,*

$$\mathbb{E}_{\tau \sim \pi_1, \pi_2} [\hat{\nabla}_x] = \nabla_x V(\pi_1, \pi_2), \quad \text{and} \quad \mathbb{E}_{\tau \sim \pi_1, \pi_2} [\hat{\nabla}_y] = \nabla_y V(\pi_1, \pi_2);$$

1263 and also, its variance is bounded:

$$\begin{aligned} \mathbb{E}_{\tau \sim \pi_1, \pi_2} \left[\left\| \hat{\nabla}_x - \nabla_x V(\pi_1, \pi_2) \right\|^2 \right] &\leq \frac{A_{\max}^2 D(\mathcal{T})^2}{\varepsilon}; \\ \mathbb{E}_{\tau \sim \pi_1, \pi_2} \left[\left\| \hat{\nabla}_y - \nabla_y V(\pi_1, \pi_2) \right\|^2 \right] &\leq \frac{B_{\max}^2 D(\mathcal{T})^2}{\varepsilon}. \end{aligned}$$

1264 where A, B denote the maximum available number of action in any info set for player 1 and 2
 1265 respectively.

1266 *Proof.* We first show that the gradient estimator is unbiased. Indeed,

$$\begin{aligned}
\nabla_x V(\pi_1, \pi_2) &= \nabla_x \left(\sum_{\tau \in \mathcal{K}} r_\tau \mathbb{P}(\tau)^{\pi_1, \pi_2} \right) \\
&= \sum_{\tau \in \mathcal{K}} r_\tau \nabla_x \mathbb{P}(\tau) \\
&= \sum_{\tau \in \mathcal{K}} r_\tau \mathbb{P}(\tau) \nabla_x \log \mathbb{P}(\tau) \\
&= \sum_{\tau \in \mathcal{K}} r_\tau \mathbb{P}(\tau) \sum_{k=1}^{K_\tau} \left(\nabla_x \log \pi_{i(k)}(a_{i(k)}^{(k)} | h^{(k)}) \right) \\
&= \mathbb{E}_{\tau \sim \pi_1, \pi_2} \left[r_\tau \sum_{k=1}^{K_\tau} \nabla_x \log \pi_{i(k)}(a_{i(k)}^{(k)} | h^{(k)}) \right] \\
&= \mathbb{E}_{\tau \sim \pi_1, \pi_2} \left[r_\tau \sum_{k=1}^{K_\tau} \nabla_x \log \pi_1(a^{(k)} | s^{(k)}) \right] \\
&= \mathbb{E}_{\tau \sim \pi_1, \pi_2} [\hat{\nabla}_x]
\end{aligned}$$

1267 The proof for $\hat{\nabla}_y$ uses an identical argument. We will now proceed to show that the variance of the
1268 (REINFORCE) gradient estimator is bounded:

$$\begin{aligned}
\mathbb{E}_\tau \left[\left\| \hat{\nabla}_x - \mathbb{E}[\hat{\nabla}_x] \right\|^2 \right] &\leq \mathbb{E}_\tau \left[\left\| \hat{\nabla}_x \right\|^2 \right] \\
&= \mathbb{E}_\tau \left[\left\| r_\tau \sum_{k=1}^{K_\tau} \nabla_x \log \pi_1(a^{(k)} | s^{(k)}) \right\|^2 \right] \\
&\leq \mathbb{E}_\tau \left[\left\| \sum_{k=1}^{K_\tau} \nabla_x \log \pi_1(a^{(k)} | s^{(k)}) \right\|^2 \right] \\
&\leq \mathbb{E}_\tau \left[K_\tau \sum_{k=1}^{K_\tau} \left\| \nabla_x \log \pi_1(a^{(k)} | s^{(k)}) \right\|^2 \right] \\
&\leq D(\mathcal{T}) \mathbb{E}_\tau \left[\sum_{k=1}^{K_\tau} \left\| \nabla_x \log \pi_1(a^{(k)} | s^{(k)}) \right\|^2 \right] \\
&= D(\mathcal{T}) \mathbb{E}_\tau \left[\sum_{k=1}^{K_\tau} \sum_{s,a} \mathbb{1}\{s = s^{(k)}, a = a^{(k)}\} \frac{1}{\pi_1^2(a | s^{(k)})} \right] \\
&= D(\mathcal{T}) \mathbb{E}_\tau \left[\sum_{k=1}^{K_\tau} \sum_{s,a} \mathbb{1}\{s = s^{(k)}\} \frac{1}{\pi_1(a | s^{(k)})} \right] \\
&\leq \frac{A}{\varepsilon} D(\mathcal{T}) \mathbb{E}_\tau \left[\sum_{k=1}^{K_\tau} \sum_{s,a} \mathbb{1}\{s = s^{(k)}\} \right] \\
&= \frac{A}{\varepsilon} D(\mathcal{T}) \sum_{\tau \in \mathcal{K}} \mathbb{P}(\tau) \sum_{k=1}^{K_\tau} \sum_{s,a} \mathbb{1}\{s = s^{(k)}\} \\
&\leq \frac{A^2 D(\mathcal{T})^2}{\varepsilon}.
\end{aligned}$$

1269

□

1270 **Lemma F.2.** The variance of (REINFORCE) for softmax policy parametrization is bounded as
1271 $\sigma_\theta^2, \sigma_\chi^2 \leq 2D(\mathcal{T})^2$.

1272 *Proof.* We see that $\nabla_{\theta} \log \pi_{\theta}(a|s) = e_{s,a} - \pi_{\theta}(\cdot|s)$. Whence, $\|\nabla_{\theta} \log \pi_{\theta}(a|s)\| \leq \sqrt{2}$. Then, the
 1273 proof follows arguments similar to the previous one. \square

1274 **Policy gradient of the bidilated regularizer** We define the policy gradient estimator of the bidilated
 1275 regularizer, $\hat{\nabla}_x \mathcal{R}_1$, as:

$$\hat{\nabla}_x \mathcal{R}_1 := \left(\sum_k^{K_{\tau}} \psi(\pi_1(s^{(k)})) \right) \sum_{k=1}^{K_{\tau}} \nabla_x \log \pi_1(a^{(k)}|s^{(k)}) + \sum_k^{K_{\tau}} \nabla_x \psi(\pi_1(s^{(k)})).$$

1276 We will demonstrate that this gradient estimator is, in fact, both unbiased and enjoys a variance that
 1277 is bounded. We start with a preliminary proposition about an alternative expression of the regularizer.

1278 **Proposition 5.** For a policy profile π_1, π_2 , the bidilated regularizer, \mathcal{R}_1 can be alternatively defined
 1279 as:

$$\mathcal{R}_1(\pi_1, \pi_2) = \sum_{\tau \in \mathcal{K}} \mathbb{P}^{\pi_1, \pi_2}(\tau) \left(\sum_k^{K_{\tau}} \psi(\pi_1(s^{(k)})) \right).$$

Proof.

$$\begin{aligned} \mathcal{R}_1(\pi_1, \pi_2) &= \sum_{s \in \mathcal{S}_1} \mu_1^{\pi_1}(\sigma(s)) \left(\sum_{h \in s} \mu_c(h) \mu_2^{\pi_2}(\sigma(h)) \right) \psi(\pi_1(s)) \\ &= \sum_{s \in \mathcal{S}_1} \mathbb{P}^{\pi_1, \pi_2}(s) \psi(\pi_1(s)) \\ &= \sum_{s \in \mathcal{S}_1} \mathbb{E}_{\tau} \left[\sum_k^{K_{\tau}} \mathbb{I}\{s = s^{(k)}\} \psi(\pi_1(s)) \right] \\ &= \mathbb{E}_{\tau} \left[\sum_{s \in \mathcal{S}_1} \sum_k^{K_{\tau}} \mathbb{I}\{s = s^{(k)}\} \psi(\pi_1(s)) \right] \\ &= \mathbb{E}_{\tau} \left[\sum_k^{K_{\tau}} \sum_{s \in \mathcal{S}_1} \mathbb{I}\{s = s^{(k)}\} \psi(\pi_1(s)) \right] \\ &= \mathbb{E}_{\tau} \left[\sum_k^{K_{\tau}} \psi(\pi_1(s^{(k)})) \right] \\ &= \sum_{\tau \in \mathcal{K}} \mathbb{P}^{\pi_1, \pi_2}(\tau) \left(\sum_k^{K_{\tau}} \psi(\pi_1(s^{(k)})) \right). \end{aligned}$$

1280 \square

1281 With the latter expression, proving the desired properties is easier.

$$\begin{aligned} \nabla_x \mathcal{R}_1(\pi_1, \pi_2) &= \nabla_x \sum_{\tau \in \mathcal{K}} \mathbb{P}^{\pi_1, \pi_2}(\tau) \left(\sum_k^{K_{\tau}} \psi(\pi_1(s^{(k)})) \right) \\ &= \sum_{\tau \in \mathcal{K}} (\nabla_x \mathbb{P}^{\pi_1, \pi_2}(\tau)) \left(\sum_k^{K_{\tau}} \psi(\pi_1(s^{(k)})) \right) + \sum_{\tau \in \mathcal{K}} \mathbb{P}^{\pi_1, \pi_2}(\tau) \left(\nabla_x \sum_k^{K_{\tau}} \psi(\pi_1(s^{(k)})) \right) \\ &= \underbrace{\sum_{\tau \in \mathcal{K}} (\mathbb{P}^{\pi_1, \pi_2}(\tau) \nabla_x \log \mathbb{P}^{\pi_1, \pi_2}(\tau))}_{\varpi_1} \left(\sum_k^{K_{\tau}} \psi(\pi_1(s^{(k)})) \right) \end{aligned}$$

$$+ \underbrace{\sum_{\tau \in \mathcal{K}} \mathbb{P}^{\pi_1, \pi_2}(\tau) \left(\sum_k^{K_\tau} \nabla_x \psi(\pi_1(s^{(k)})) \right)}_{\varpi_2}$$

1282 For ϖ_1 , let us denote $R_\tau = \sum_k^{K_\tau} \psi(\pi_1(s^{(k)}))$,

$$\begin{aligned} \varpi_1 &= R_\tau \sum_{\tau \in \mathcal{K}} \mathbb{P}^{\pi_1, \pi_2}(\tau) \nabla_x \log \mathbb{P}^{\pi_1, \pi_2}(\tau) \\ &= \sum_{\tau \in \mathcal{K}} R_\tau \mathbb{P}(\tau) \nabla_x \log \mathbb{P}(\tau) \\ &= \sum_{\tau \in \mathcal{K}} R_\tau \mathbb{P}(\tau) \sum_{k=1}^{K_\tau} \left(\nabla_x \log \pi_{i(k)}(a_{i(k)}^{(k)} | h^{(k)}) \right) \\ &= \mathbb{E}_{\tau \sim \pi_1, \pi_2} \left[R_\tau \sum_{k=1}^{K_\tau} \nabla_x \log \pi_{i(k)}(a_{i(k)}^{(k)} | h^{(k)}) \right] \\ &= \mathbb{E}_{\tau \sim \pi_1, \pi_2} \left[R_\tau \sum_{k=1}^{K_\tau} \nabla_x \log \pi_1(a^{(k)} | s^{(k)}) \right]. \end{aligned}$$

1283 For ϖ_2 , we write,

$$\begin{aligned} \varpi_2 &= \sum_{\tau \in \mathcal{K}} \mathbb{P}^{\pi_1, \pi_2}(\tau) \sum_k^{K_\tau} \nabla_x \psi(\pi_1(s^{(k)})) \\ &= \mathbb{E}_\tau \left[\sum_k^{K_\tau} \nabla_x \psi(\pi_1(s^{(k)})) \right] \end{aligned}$$

1284 We will use similar arguments for the variance in the case of the (REINFORCE) gradient estimator.

$$\begin{aligned} &\mathbb{E} \left[\left\| \hat{\nabla}_x \mathcal{R}_1 - \mathbb{E} \left[\hat{\nabla}_x \mathcal{R}_1 \right] \right\|^2 \right] \\ &\leq \mathbb{E} \left[\left\| \hat{\nabla}_x \mathcal{R}_1 \right\|^2 \right] \\ &\leq \mathbb{E} \left[2 \underbrace{\left\| \left(\sum_k^{K_\tau} \psi(\pi_1(s^{(k)})) \right) \sum_{k=1}^{K_\tau} \nabla_x \log \pi_1(a^{(k)} | s^{(k)}) \right\|^2}_{\vartheta_1} + 2 \underbrace{\left\| \sum_k^{K_\tau} \nabla_x \psi(\pi_1(s^{(k)})) \right\|^2}_{\vartheta_2} \right] \end{aligned}$$

1285 For ϑ_1 , similar to Lemma F.1, we see that

$$\mathbb{E}[\vartheta_1] \leq \frac{A^2 \psi_{\max}^2 D(\mathcal{T})^2}{\varepsilon}.$$

1286 Whereas, for ϑ_2 ,

$$\begin{aligned} \mathbb{E}[\vartheta_2] &\leq \mathbb{E} \left[K_\tau \sum_k^{K_\tau} \left\| \nabla_x \psi(\pi_1(s^{(k)})) \right\|^2 \right] \\ &\leq \mathbb{E} \left[K_\tau \sum_k^{K_\tau} L_\psi^2 \right] \\ &\leq D(\mathcal{T})^2 L_\psi^2. \end{aligned}$$

1287 F.2 Gradient Inexactness Bound

1288 Perturbing the game by adding a *bidilated regularizer* term for each player transforms it into a
1289 strongly-convex strongly-concave game in the sequence-form space. Each bidilated regularizer

interleaves information of all players' strategies. This inevitably leaves its trace on the gradient of each perturbed utility function; a part of the gradient carries information about the opponents' strategies that cannot be accessed by merely sampling trajectories (e.g. the entropy or the squared ℓ_2 -norm of the opponents' behavioral strategies).

G Convergence Analysis

G.1 Direct Policy Parametrization

Theorem G.1. *With direct policy parametrization and the Euclidean bidilated regularizer, alternating policy-gradient algorithm attains a last-iterate ϵ -Nash equilibrium in*

$$T = \frac{1}{\epsilon^6} \text{poly} \left(\frac{1}{\epsilon}, \frac{1}{\gamma}, A_{\max}, B_{\max}, 2^{D(\mathcal{T})}, \min_h \mu_c(h) \right) \text{ iterations,}$$

using batches of $\text{poly} \left(\frac{1}{\epsilon}, \frac{1}{\gamma}, A_{\max}, B_{\max}, 2^{D(\mathcal{T})}, \min_h \mu_c(h) \right)$ trajectory samples at each step.

Proof. The proof follows as an application of Theorem B.1. \square

G.2 Softmax Policy Parametrization

Theorem G.2. *Alternating policy-gradient algorithm with softmax policy parametrization and the entropic bidilated regulariser, converges in expectation in the last-iterate to an ϵ -Nash equilibrium after a number of iterations T , that is*

$$T = \frac{1}{\epsilon^{11}} \text{poly} \left(\frac{1}{\epsilon}, \frac{1}{\gamma}, A_{\max}, B_{\max}, 2^{D(\mathcal{T})}, \min_h \mu_c(h) \right) \text{ iterations,}$$

using batches of $\text{poly} \left(\frac{1}{\epsilon}, \frac{1}{\gamma}, A_{\max}, B_{\max}, 2^{D(\mathcal{T})}, \min_h \mu_c(h), \min_h \mu_c(h) \right)$ trajectory samples at each step.

Proof. The theorem follows a corollary of Theorem B.1. \square

G.3 Natural Policy Gradient

G.3.1 The Fisher Information Matrix

$$\mathbf{F}(\chi) = \mathbb{E}_{s \sim d^{\chi, \theta}} \mathbb{E}_{a \sim \pi_\chi(\cdot|s)} \left[\nabla \log_\chi \pi_\chi(a|s) [\nabla_\chi \log \pi_\chi(a|s)]^\top \right]$$

The matrix $F(\chi)$ is a block diagonal matrix with its (s, s) -block being the matrix:

$$\mathbf{F}_s(\chi) = d^{\chi, \theta}(s) \left(\text{diag}(\pi_\chi(s)) - \pi_\chi(s) \pi_\chi(s)^\top \right).$$

Its pseudo-inverse, \mathbf{F}^\dagger , is again a block-diagonal matrix, with an (s, s) -block,

$$\mathbf{F}_s^\dagger(\chi) = \frac{1}{d^{\chi, \theta}(s)} \left(\text{diag}(\pi_\chi(s)) - \pi_\chi(s) \pi_\chi(s)^\top \right)^\dagger.$$

Interestingly, the matrix $\mathbf{Z} := \mathbf{F}^\dagger \mathbf{J}_{\text{softmax}}(\chi)$ is a block-diagonal matrix with entries $\frac{1}{d^{\chi, \theta}(s)} \mathbf{I}_{|\mathcal{A}_s| \times |\mathcal{A}_s|}$ on diagonal (s, s) -block.

The spectrum of the Fisher Information Matrix With the same arguments used in Lemma D.1, we can conclude that,

- $\lambda_{\min}(\mathbf{F}(\chi)) = 0$;
- $\lambda_{\min}^+(\mathbf{F}(\chi)_s) \geq d(s) \min_a \pi_\chi(a|s)$;

$$1317 \quad \bullet \quad d^{X,\theta}(s) \max_a \pi_\chi(a|s) \leq \lambda_{\max}(\mathbf{F}(\chi)_s) \leq d^{X,\theta}(s) \max_a \pi_\chi(a|s) + 1.$$

1318 Hence,

$$1319 \quad \bullet \quad \lambda_{\min}^+(\mathbf{F}(\chi)) \geq \min_{s,a} d^{X,\theta}(s) \pi_\chi(a|s);$$

$$1320 \quad \bullet \quad \min_s \frac{1}{\sqrt{|\mathcal{S}_1| |\mathcal{A}_s|}} \leq \lambda_{\max}(\mathbf{F}(\chi)) \leq 1.$$

1321 **Theorem G.3.**

1322 *Proof.* Again, this theorem can be seen as a corollary of Theorem B.1. □

1323 H Efficient Exploration

1324 Throughout our proofs, we have kept our complexity results parametric w.r.t. $1/\gamma$. A naive exploration
1325 rule that would dictate that the player merely picks behavioral strategies over the truncated simplex
1326 will give a $\gamma = O(\epsilon^{D(\mathcal{T})})$.

1327 We propose a different approach to exploration, inspired by van Damme’s notion of “tremble” in
1328 quasi-perfect equilibrium. In words, every player is expected to reach every subsequence with a
1329 probability $\gamma \frac{1}{|\mathcal{S}_i|}$. The rule is simple,

- 1330 • at the beginning of each game, the player throws a biased coin which lands on “heads” with
1331 probability γ . If so happens, the player executes a sequence of actions with probability $\frac{1}{|\mathcal{S}_i|}$.
1332 Afterwards, the player continues to play according to their own behavioral strategy.
- 1333 • In the case that the coin lands on “tails”, the player simply plays according to their behavioral
1334 strategy.

1335 We observe that in sequence-form, this means that $\mu(s) \geq \frac{\gamma}{|\mathcal{S}_1|} + \frac{\gamma}{|\mathcal{S}_1|} \sum_{s' \in \mathcal{S}_1} \mathbb{1}\{s' \succeq_{\mathcal{T}} s\}$. In
1336 other words, the sequence-form strategies are truncated by a set of linear constraints and as long
1337 as $\gamma \leq \frac{1}{|\mathcal{S}_1|}$, there set of feasible sequence-form strategies is non-empty. We now observe that the
1338 mapping, from μ to the part component of the behavioral policy the agent can in fact control, is

$$\pi(a|s) = \frac{\mu(s, a) - \frac{\gamma}{|\mathcal{S}_1|} \sum_{s' \in \mathcal{S}_1} \mathbb{1}\{s' \succeq_{\mathcal{T}} (s, a)\}}{\mu(s) - \frac{\gamma}{|\mathcal{S}_1|} \sum_{s' \in \mathcal{S}_1} \mathbb{1}\{s' \succeq_{\mathcal{T}} s\}}.$$

1339 As such, the Lipschitz continuity of mapping $\mu \mapsto \pi$, is Lipschitz continuous with a modulus,

$$\frac{|\mathcal{S}_1| \sqrt{A_{\max}}}{\gamma},$$

1340 by following the same line of arguments as the ones in Lemma C.1.