OPT2025: 17th Annual Workshop on Optimization for Machine Learning

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

**Fivos Kalogiannis** FKALOGIANNIS@UCSD.EDU
*UC San Diego, CSE*
**Gabriele Farina** GFARINA@MIT.EDU
*MIT, EECS*

## Abstract

Multi-agent reinforcement learning (MARL) has long been seen as inseparable from Markov games [30]. Yet, the most remarkable achievements of practical MARL have arguably been in extensive-form games (EFGs)—spanning games like Poker, Stratego, and Hanabi. At the same time, little is known about provable equilibrium convergence for MARL algorithms applied to EFGs as they stumble upon the inherent nonconvexity of the optimization landscape and the failure of the value-iteration subroutine in EFGs. To this goal, we utilize contemporary advances in nonconvex optimization theory to prove that regularized alternating policy gradient with (i) *direct policy parametrization*, (ii) *softmax policy parametrization*, and (iii) *softmax policy parametrization with natural policy gradient updates* converge to an approximate Nash equilibrium (NE) in the *last-iterate* in imperfect-information perfect-recall zero-sum EFGs. Namely, we observe that since the individual utilities are concave with respect to the sequence-form strategy, they satisfy gradient dominance w.r.t. the behavioral strategy—or, *policy*, in reinforcement learning terms. We exploit this structure to further prove that the regularized utility satisfies the much stronger Polyak-Łojasiewicz condition. In turn, we show that the different flavors of alternating policy gradient methods converge to an $\epsilon$-approximate NE with a number of iterations and trajectory samples that are polynomial in $1/\epsilon$ and the natural parameters of the game. Our work is a preliminary—yet principled—attempt in bridging the conceptual gap between the theory of Markov and imperfect-information EFGs while it aspires to stimulate a deeper dialogue between them.

## 1. Introduction

Beyond its classical role in economics, game-theoretic environments now benchmark and improve AI planning: self-play language games can enhance LLM reasoning [7], and cross-domain benchmarks probe strategic reasoning [8]. Despite extensive work on policy optimization for imperfect-information games [19, 27, 34, 41, 46], polynomial-time convergence guarantees remain elusive, motivating the question:

*Do policy gradient methods provably converge to an equilibrium in imperfect-information EFGs using a polynomial number of iterations/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 gradient updates*—and a highly nonconvex policy optimization landscape—*which we prove that is benign*.

### 1.1. Contributions

We answer (★) in the affirmative by developing a policy gradient method (`Alt-EntRegPG`). Our algorithmic approach leads to last-iterate convergence to a regularized NE of the EFG. We, namely, contribute, (1) a novel *exploration scheme* that makes convergence with a polynomial number of samples possible (2) proof of the *gradient domination property* for the utilities in imperfect-information EFGs (3) a demonstration that REINFORCE is an unbiased gradient estimator of utilities in EFGs (4) regularized alternating *policy gradient* with *direct softmax parametrization* and alternating *natural policy gradient* steps converge in to a Nash equilibrium in imperfect-information EFGs.

### 1.2. Overview of Techniques

Our algorithmic guarantees are based on one idea: cast equilibrium computation as constrained two-sided PŁ optimization over policies. With suitable regularization the utility becomes hidden concave, each utility satisfies a proximal PŁ (pPŁ, a form of strong gradient-domination) inequality, and alternating gradient descent/ascent converges.

**Hidden concavity $\Rightarrow$ PŁ.** Utilities in EFGs are concave in sequence-form; with a suitable regularizer they become strongly concave. Enforcing a positive lower bound on reach probabilities makes the map from sequence-form strategies to behavioral policies a uniform-Lipschitz bijection. Writing the loss as $f(x) = -H(c(x))$ with $u = c(x)$ and $H$ $\mu$-strongly convex in $u$ (hidden strong convexity), we have $H(u) - H^\star \leq \frac{1}{2\mu}\|\nabla_u H(u)\|^2$. If $c^{-1}$ has Lipschitz modulus $L_{c^{-1}}$, the chain rule gives the PŁ inequality $f(x) - f^\star \leq \frac{L_{c^{-1}}^2}{2\mu}\|\nabla f(x)\|^2$; analogous arguments yields the pPŁ.

**Convergence** To show convergence of regularized alternating gradient descent ascent we utilize a Lyapunov function argument similar to that of [51].

## 2. Preliminaries

**Definition 1 (IIEFG)** *A two player zero-sum imperfect information extensive-form game, $\Gamma$, 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. We will note $S := \max\{|\mathcal{S}_1|, |\mathcal{S}_2|\}$.*
- *$\mathcal{A} := \{\mathcal{A}_s\}_{s\in\mathcal{S}_1}, \mathcal{B} := \{\mathcal{B}_s\}_{s\in\mathcal{S}_2}$ 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 $A_s := |\mathcal{A}_s|$, $A := \max_s A_s$ and $B_s := |\mathcal{B}_s|$, $B := \max_s B_s$.*
- *$r : \mathcal{H} \to [0, 1]$ is a payoff function mapping leaves of $\mathcal{T}$ to a payoff for player 1; player 2 gets the opposite payoff.*

**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:

$$\mu_1^{\pi_1}(s,a) := \mu_1^{\pi_1}(\sigma_1(s))\pi_1(a|s) \quad \forall s \in \mathcal{S}_1, \forall a \in \mathcal{A}_s,$$

where $\sigma_1(s), \sigma_2(s)$ denote the last infoset-action pair $(s', a'), s \in \mathcal{S}_1$ and $(s', b'), s \in \mathcal{S}_2$ player 1 and 2 encountered when descending from the game tree's root to history $h$. The set of sequence-form strategies, $\mathcal{M}_1, \mathcal{M}_2$ are convex polytopes and for player 2, 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 $r$.

**Definition 2 ($\epsilon$-NE)** *For an $\epsilon > 0$, a policy profile $\pi_1^\star, \pi_2^\star$ is an $\epsilon$-approximate Nash equilibrium of $\Gamma$, if $\forall \pi_1$ and $\pi_2$,*

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

**The bidilated regularizer.** Introduced in [32], the unweighted *bidilated regularizer* is defined using the entropic regularizer $\psi(\cdot)$ multiplied by the total reach probability of each insfoset. Since it depends on both players' policies we write $\mathcal{R}(\pi_1, \pi_2), \mathcal{R}(\pi_2, \pi_2)$ which are defined as $\mathcal{R}_1(\pi_1, \pi_2) := \mathbb{E}_{h\sim\pi}\left[\sum_h \psi(\pi_1(\cdot|h))\right]$ and $\mathcal{R}_2(\pi_1, \pi_2) := \mathbb{E}_{h\sim\pi}\left[\sum_h \psi(\pi_2(\cdot|h))\right]$.

**Policy parametrization** We consider direct policy parametrization and softmax policy parametrization. The parameters of directly parametrized policies will be denoted as $x \in \mathbb{R}^A, A = \sum_s A_s$ and $y \in \mathbb{R}^B, B = \sum_s B_s$ accordingly. While, parameters of softmax policies will be denoted $\chi, \theta$ with $\chi \in \mathbb{R}^A, A = \sum_s A_s$ and $\theta \in \mathbb{R}^B, B = \sum_s B_s$. 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).

For technical reasons related to the variance of stochastic gradient estimation and controlling moduli of relative smoothness, we make the following assumption.

**Assumption 1** *For an $\varepsilon > 0$, both players' policies, for every infoset and action, satisfy*

$$\pi_1(a|s) \geq \varepsilon, \ \forall s \in \mathcal{S}_1, \forall a \in \mathcal{A}_s \quad \pi_2(b|s) \geq \varepsilon, \ \forall s \in \mathcal{S}_2, \forall b \in \mathcal{B}_s. \qquad (\varepsilon\text{-trunc.})$$

## 3. Main Results

### 3.1. Efficient Exploration Strategy

We propose a novel approach to exploration. Each player is expected to reach every subsequence with probability $\frac{\gamma}{|\mathcal{S}_i|}$. The rule is simple:

**Assumption 2 (Efficient Exploration)** *Both players follow the following exploration strategy:*

- *At the start of each game, the player flips a biased coin that shows "heads" with probability $\gamma$.*
- *If the coin shows "heads", the player selects a sequence uniformly at random and then executes it.*
- *After this sequence, or if the coin shows "tails", the player resumes play according to their policy.*

**Remark 3** *It is noteworthy that using this exploration strategy, one can exercise direct control over the modulus of gradient domination. Whereas, policy gradient literature [1, 9, 35, 52] needs to make an assumption on the boundedness of the distribution mismatch coefficient.*

### 3.2. Gradient Domination Properties

We demonstrate that even though the utility of each player is highly nonconvex w.r.t. the policy, it satisfies the following gradient domination property.

**Definition 4 (pPŁ-condition [25])** *Assume $F : \mathbb{R}^d \to \mathbb{R}$ defined as $F(x) := f(x) + g(x)$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be an $\ell$-smooth function and $g : \mathbb{R}^d \to \mathbb{R}$ 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 pPŁ condition with modulus $\alpha > 0$ if, for every $x \in \mathcal{X}$,*

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

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

**Lemma 5 (pPŁ for EFG; restated from Thms. 42 to 44)** *Let an imperfect-information EFG, $\Gamma$, perturbed with the pair of bidilated regularizers $(\mathcal{R}_1, \mathcal{R}_2)$ with a coefficient $\tau > 0$. Then, each player's utility satisfies the pPŁ condition with a modulus*

$$\alpha = \tau \times \mathsf{poly}\left(\gamma, \tfrac{1}{A}, \tfrac{1}{B}, \frac{1}{2^{D(\mathcal{T})}}, \tfrac{1}{|\mathcal{S}_1|}, \tfrac{1}{|\mathcal{S}_2|}, \tfrac{1}{|\mathcal{H}|}\right).$$

### 3.3. Convergence of Alternating Regularized Policy Gradient

Having established the required background and notation, we are ready to present our main results. In Theorem 8 we show the convergence of simple alternating regularized policy gradient to an approximate NE in the last iterate. Moving to Theorem 10, 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 11).

Throughout, $\eta_x, \eta_y$ denote the stepsizes and $\hat{\nabla}_\cdot^\tau$ denotes the (REINFORCE) gradient estimate of the utility w.r.t. to a player's parameters accounting for both regularization terms. In order to estimate the gradient of the regularized game, the players need to exchange information about their regularizers. During each rollout, the trajectory's reward is increased by the sum of the opponent's local regularizer at every information set encountered from the root to the terminal node (chance nodes excluded).

**Definition 6** (REINFORCE) *Let $\xi$ denote a trajectory of infoset and actions sampled by implementing policies $\pi_1, \pi_2$, $\xi := (s_{(1)}, a_{(k)}, \dots)$. We define* REINFORCE, $\left(\widehat{\nabla}_x, \widehat{\nabla}_y\right)$, *to be the stochastic gradient estimators:*

$$\widehat{\nabla}_x = r_\xi \sum_{k=1}^{K_\xi} \nabla_x \log \pi_x\big(a_{(k)}|s_{(k)}\big) \quad \textit{and} \quad \widehat{\nabla}_y = r_\xi \sum_{k=1}^{K_\xi} \nabla_x \log \pi_y\big(b_{(k)}|s_{(k)}\big). \qquad \text{(REINFORCE)}$$

**Lemma 7 (Appendix F.1)** *Stochastic gradient estimator* REINFORCE *is unbiased and of bounded variance for both direct and softmax parametrized policies.*

**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,

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

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

**Theorem 8 (Informal; restated from Thm. 49)** *With direct policy parametrization and the Euclidean bidilated regularizer, alternating policy-gradient algorithm attains a last-iterate $\epsilon$-Nash equilibrium in*

$$T = \mathsf{poly}\left(\frac{1}{\epsilon}, \frac{1}{\gamma}, A, B, 2^{D(\mathcal{T})}, |\mathcal{S}_1|, |\mathcal{S}_2|, |\mathcal{H}|\right) \text{ iterations},$$

*using batches of* $\mathsf{poly}\left(\frac{1}{\epsilon}, \frac{1}{\gamma}, A, B, 2^{D(\mathcal{T})}, |\mathcal{S}_1|, |\mathcal{S}_2|, |\mathcal{H}|\right)$ *trajectory samples at each step.*

**Remark 9** *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.*

**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 $X_R, \Theta_R$ guarantees that Equation ($\varepsilon$-trunc.) is satisfied,

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

**Theorem 10 (Informal; restated from Thm. 50)** *Alternating policy-gradient algorithm with softmax policy parametrization and the entropic bidilated regularizer, converges in expectation in the last-iterate to an $\epsilon$-Nash equilibrium after a number of iterations $T$, that is*

$$T = \mathsf{poly}\left(\frac{1}{\epsilon}, \frac{1}{\gamma}, A, B, 2^{D(\mathcal{T})}, |\mathcal{S}_1|, |\mathcal{S}_2|, |\mathcal{H}|\right) \text{ iterations},$$

*using batches of* $\mathsf{poly}\left(\frac{1}{\epsilon}, \frac{1}{\gamma}, A, B, 2^{D(\mathcal{T})}, |\mathcal{S}_1|, |\mathcal{S}_2|, |\mathcal{H}|\right)$ *trajectory samples at each step.*

**Natural Policy Gradient** Finally, we consider the natural policy gradient algorithm [23] which is an adaptation of natural gradient [2]. This algorithm is of particular interest due to its intimate connection to the TRPO, PPO [44, 45] policy optimization algorithms. Natural policy gradient uses a *Fisher information matrix* induced by the policy as a preconditioner for policy gradient updates:

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

We cast *natural policy gradient* steps as *mirror descent steps* with a Mahalanobis norm induced by the Fisher information matrix (for a more nuanced discussion on this connection see [39]).

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

The update scheme can be equivalently written as:

$$\begin{aligned}
\chi_{t+1} &= \arg\min_{\chi \in X_R} \left\| \chi_t - \eta_x \mathbf{F}^\dagger_\chi(\chi_t,\theta_t) \nabla_\chi V(\chi_t,\theta_t) - \chi \right\|^2_{\mathbf{F}_\chi(\chi_t,\theta_t)} \\
\theta_{t+1} &= \arg\min_{\theta \in \theta_R} \left\| \theta_t + \eta_y \mathbf{F}^\dagger_\theta(\chi_{t+1},\theta_t) \nabla_\theta V(\chi_{t+1},\theta_t) - \theta \right\|^2_{\mathbf{F}_\theta(\chi_{t+1},\theta_t)}
\end{aligned} \qquad \texttt{(Alt-EntRegNPG)}$$

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

$$\begin{aligned}
\overline{\pi}_{1,t+1}(\cdot|s) &\propto \pi_{1,t}(\cdot|s)^{1-\eta_x \tau} \odot \exp\left(\eta_x Q^{\pi_t}_\tau(s,\cdot)\right) \\
\pi_{1,t+1}(\cdot|s) &\approx \arg\min_{\pi \in \Pi^R_1} \mathrm{KL}\Big(\pi(\cdot|s) \Big\| \overline{\pi}_{1,t+1}(\cdot|s)\Big)
\end{aligned}$$

To see why the second approximate equality holds, we note that the Mahalanobis distance over the parameters induced by the Fisher information matrix of the softmax policy, is a second-order approximation of policy KL. The derivation and an extensive discussion are deferred to Appendix G.3 (also, see [21]).

**Theorem 11 (Informal; restated from Thm. 51)** *For an appropriate tuning of $\eta_x, \eta_y > 0$, the last-iterate of alternating regularized natural policy gradient* (Alt-EntRegNPG) *converges in expectation to an $\epsilon$-approximate Nash equilibrium in a number of iterations $T$ that is:*

$$T = \mathsf{poly}\left(\frac{1}{\epsilon}, \frac{1}{\gamma}, A, B, 2^{D(\mathcal{T})}, |\mathcal{S}_1|, |\mathcal{S}_2|, |\mathcal{H}|\right).$$

## 4. Conclusion

We studied three different policy gradient methods for imperfect-information perfect-recall zero-sum EFGs under a unifying optimization principle. We managed to provide the first global last-iterate convergence guarantees of independent policy gradient methods to an $\epsilon$-approximate Nash

equilibrium. Furthermore, our analysis requires a number of iterations and samples that is polynomial in $1/\epsilon$ and the parameters of the game. To do so, we exploited the favorable properties (PŁ-condition) of the otherwise nonconvex optimization landscape. We departed from the usual route of regret analysis in EFGs and opted for more conventional convergence analysis arguments. We hope to motivate further exchange between theoretical MARL research and the theory of EFGs as we strongly believe in the potential this communication fosters.

## References


[1] 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.

[2] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2): 251–276, 1998.

[3] Anastasios N Angelopoulos, Michael I Jordan, and Ryan J Tibshirani. Gradient equilibrium in online learning: Theory and applications. *arXiv preprint arXiv:2501.08330*, 2025.

[4] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *Operations Research*, 72(5):1906–1927, 2024.

[5] 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*, 2024.

[6] Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4): 2563–2578, 2022.

[7] Pengyu Cheng, Yong Dai, Tianhao Hu, Han Xu, Zhisong Zhang, Lei Han, Nan Du, and Xiaolong Li. Self-playing adversarial language game enhances llm reasoning. *Advances in Neural Information Processing Systems*, 37:126515–126543, 2024.

[8] Anthony Costarelli, Mat Allen, Roman Hauksson, Grace Sodunke, Suhas Hariharan, Carlson Cheng, Wenjie Li, Joshua Clymer, and Arjun Yadav. Gamebench: Evaluating strategic reasoning abilities of llm agents. *arXiv preprint arXiv:2406.06613*, 2024.

[9] 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.

[10] Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate o (k-1/4) on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018.

[11] 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.

[12] Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178:503–558, 2019.

[13] Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Optimistic regret minimization for extensive-form games via dilated distance-generating functions. *Advances in neural information processing systems*, 32, 2019.

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

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

[16] Maryam Fazel, Rong Ge, Sham M. Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, 2018. URL https://api.semanticscholar.org/CorpusID:51881649.

[17] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *Advances in neural information processing systems*, 29, 2016.

[18] Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[19] Daniel Hennes, Dustin Morrill, Shayegan Omidshafiei, Rémi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, Paavo Parmas, Edgar Duèñez Guzmán, and Karl Tuyls. Neural replicator dynamics: Multiagent learning via hedging policy gradients. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, page 492–501, Richland, SC, 2020. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450375184.

[20] Samid Hoda, Andrew Gilpin, Javier Pena, and Tuomas Sandholm. Smoothing techniques for computing nash equilibria of sequential games. *Mathematics of Operations Research*, 35(2): 494–512, 2010.

[21] Chloe Ching-Yun Hsu, Celestine Mendler-Dünner, and Moritz Hardt. Revisiting design choices in proximal policy optimization. *arXiv preprint arXiv:2009.10897*, 2020.

[22] Sashank J Reddi, Suvrit Sra, Barnabas Poczos, and Alexander J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. *Advances in neural information processing systems*, 29, 2016.

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

[24] Fivos Kalogiannis, Emmanouil-Vasileios Vlatakis-Gkaragkounis, Ian Gemp, and Georgios Piliouras. Solving zero-sum convex markov games. In *Forty-second International Conference on Machine Learning*, 2025.

[25] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.

[26] Christian Kroer, Kevin Waugh, Fatma Kılınç-Karzan, and Tuomas Sandholm. Faster algorithms for extensive-form game solving via improved smoothing functions. *Mathematical Programming*, 179(1):385–417, 2020.

[27] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

[28] Guoyin Li and Ting Kei Pong. Calculus of the exponent of kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, 18(5):1199–1232, 2018.

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

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

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

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

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

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

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

[36] Julie Mulvaney-Kemp, SangWoo Park, Ming Jin, and Javad Lavaei. Dynamic regret bounds for constrained online nonconvex optimization based on polyak–lojasiewicz regions. *IEEE Transactions on Control of Network Systems*, 10(2):599–611, 2023. doi: 10.1109/TCNS.2022.3203798.

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

[38] Konstantinos Oikonomidis, Emanuel Laude, and Panagiotis Patrinos. Forward-backward splitting under the light of generalized convexity. *arXiv preprint arXiv:2503.18098*, 2025.

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

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

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

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

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

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

[45] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017. URL https://api.semanticscholar.org/CorpusID:28695052.

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

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

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

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

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

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

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

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

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

## Appendix A. Further Related Work

**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 [37]. 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—significantly often, *stationarity implies global optimality*. Cases in point, gradient domination is exhibited for *the loss functions of overparametrized neural networks* [31, 43], *the linear quadratic regulator* [16], *value functions of Markov decision processes (MDPs)* [1, 4], *matrix completion* [17], *dictionary learning* [47], and more. For a thorough discussion of gradient domination and other regularity conditions we refer the reader to [11, 12, 25, 28, 29, 38, 40] and references therein. With the latter in mind, one could make the case that when game theory researchers seek equilibrium computation in general nonconvex games [3, 5] they set the bar too high. Still, the study of benign nonconvexity seems of great importance and rather underexplored [36, 42, 49, 51].

**Relevant MARL for MG works** In MDP and MG literature, policy optimization seems to come in two flavors—an *online learning* [18] approach and a *stochastic optimization* one. In the current work, we opt for the second approach.

The approach of [52] is particularly similar to ours. Yet, we highlight that they make a rather strong assumption; they assume that the probability of playing each action in the support of the regularized Nash equilibrium is lower-bounded by a constant independent of the regularization coefficient $\tau$. In turn, we circumvent such an assumption by exercising direct control over the minimum probability of playing any action by projecting the parameters of the softmax parameters onto a convex polytope. Also, we provide guarantees for natural policy gradient iterations.

**Theory of Policy Gradient Methods** The policy gradient method [48, 50]

- [1] prove the convergence of directly parametrized policy gradient. They use the convergence result of gradient descent for smooth nonconvex function along a gradient domination lemma to demonstrate a $O(1/\epsilon^2)$ convergence rate to optimality. Later, [53, 54] use the *hidden concave* structure of the problem to improve the convergence rate to $O(1/\epsilon)$.

- [35] provide the first non-asymptotic convergence rate result for the policy gradient method using discounted entropy regularization (the analogue of bidilated entropy regularization). The proof of convergence uses a novel nonuniform PŁ condition.

- [6] analyze natural policy gradient (NPG) with discounted entropy regularization. Natural policy gradient can be seen as a form of *preconditioned* gradient descent. Natural policy gradient effectively boils down to policy multiplicative weight updates using the $Q$-functions as feedback. The analysis of convergence uses a linear dynamical system.

## Appendix B. Optimization Lemmata

**Definition 12 (Stationarity Proxies)** *Assume a function $F : f + I_{\mathcal{X}}(\cdot)$ such that $f : \mathcal{X} \to \mathbb{R}$ is $\ell$-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)*,*

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

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

- *Mahalanobis gradient mapping* (MGM)*,*

$$\Delta_\rho^+(x) := \rho^2 \left\| x - x^+ \right\|_{\mathbf{M}_t}^2,$$

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

- *Mahalanobis forward-backward mapping* (MFBM)*,*

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

**Lemma 13** *The following properties hold true for the proximal point and the Mahalanobis Moreau envelope,*

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

**Proof** The first and last items follow easily from the definition and standard arguments [10]. The middle one uses the optimality condition of $\hat{x} := \operatorname{prox}_{\rho F}(x)$,

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

from which we conclude,

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

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

**Definition 14 (pPŁ, KŁ)** *Let $f : \mathcal{X} \to \mathbb{R}$ be an $L$-Lipschitz continuous function with $\ell$-Lipschitz continuous gradient. Then,*

- Proximal Polyak-Łojasiewicz (pPŁ): *$f$ is said to satisfy the proximal Polyak-Łojasiewicz condition if* $\exists \alpha > 0$ *s.t.*
$$\frac{1}{2}\mathcal{D}_{\mathcal{X}}(x, \ell) \geq \alpha \left[f(x) - f(x^\star)\right]$$
- Kurdyka-Łojasiewicz (KŁ): *$f$ is said to satisfy if* $\exists \overline{\alpha}$ *s.t.*
$$\min_{s_x \in \partial(f + I_{\mathcal{X}})(x)} \|s_x\|^2 \geq 2\overline{\alpha}\left[f(x) - f(x^\star)\right], \quad \forall x \in \mathcal{X}.$$

The definitions for the Mahalanobis analogues of pPŁ and KŁ follow straightforward extension.

**Lemma 15** *Let $f$ be an $\ell$-smooth function relative to $\|\cdot\|^2_{\mathbf{M}}$ defined over the convex set $\mathcal{X}$. If $f$ satisfies the* (Mahalanobis) KŁ condition *with modulus $\mu_{\mathrm{kl}}$, it also satisfies the* (Mahalanobis) pPŁ *condition with a modulus of $\mu_{\mathrm{ppl}} = \frac{\mu_{\mathrm{kl}}}{202}$.*

**Proof** First, we define $F(x) := f(x) + I_{\mathcal{X}}(x)$, with $I_{\mathcal{X}}(\cdot)$ being the indicator function. We highlight that since $I_{\mathcal{X}}(\cdot)$ is convex and $f$ is $\ell$-smooth (relative to $\|\cdot\|^2_{\mathbf{M}}$), then $F$ is $\ell$-weakly convex (relative to $\|\cdot\|^2_{\mathbf{M}}$). This means that the proximal point of the function $F/\rho$ is well defined for any $\rho > \ell$.

Now, assume a point $x \in \mathcal{X}$ and $\hat{x} := \mathrm{prox}_{F/\rho}(x)$. By assumption, for any $\hat{x} \in \mathcal{X}$, it holds true that,

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

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

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

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

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

Collecting the terms,

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

A direct generalization of [25, Lemma 1], implies that for the MFBM and a choice of $\rho_1, \rho_2 > 0$ such that$\rho_1 > \rho_2$, then $\mathcal{D}(x, \rho_1) \geq \mathcal{D}(x, \rho_1)$. As such, we write,

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

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

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

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

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

Now, from [14, Lemmata 4.1 & 4.2], we know that,

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

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

$$\frac{1}{2}\mathcal{D}(x, 4\ell) \geq \frac{\mu}{101}[f(x) - f^\star].$$

■

**Remark 16** *The latter lemma provides a bound that is significantly tighter than the one implied by the analysis found [25, Appendix G] which connects the moduli of the KŁand pPŁconditions.*

### B.1. A Variation of the Descent Lemma

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

**Lemma 17 ([22, Lemma 1])** *Let $f : \mathcal{X} \to \mathbb{R}$ be an $\ell$-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*

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

*Then, the following inequality holds true:*

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

**Lemma 18** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a closed convex set, and let $f : \mathcal{X} \to \mathbb{R}$ be an $\ell$-smooth function 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 $v \in \mathbb{R}^d$, define $x^+ = \mathrm{Proj}_{\mathcal{X},\mathbf{M}_t}(x - \eta v)$ . Then the following inequality holds:*

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

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

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

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

- Invoking Theorem 17 with $x = x$, $y = \overline{x}_+$, $z = x$, $v = \nabla f(x)$

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

- Again, invoking Theorem 17 but with $x = x$, $y = x_+$, $z = \overline{x}_+$, $v$,

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

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

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

Adding (4),

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

- (5) follows from the application of Young's inequality on

$$\left\langle \nabla f(x) - v, x^+ - \overline{x}^+ \right\rangle = \left\langle \mathbf{M}_t{}^{-1/2}\nabla f(x) - v, \mathbf{M}_t{}^{1/2}x^+ - \overline{x}^+ \right\rangle;$$

- (6) follows by dropping the non-positive terms; non-positivity follows from the choice of the step-size, $\eta \le \frac{1}{5\ell}$.

■

### B.2. Min-Max Optimization

**Lemma 19** *Let $f : \mathcal{X} \times \mathcal{Y}$ be an $\ell$-smooth function, $\rho > 0$, two points $y, y' \in \mathcal{Y}$, and a point $x \in \mathcal{X}$. Then, the following inequality holds:*

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

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

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

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

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

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

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

We note that:

- The first inequality follows from the triangle inequality.
- In the second inequality, we applied the reverse triangle inequality.
- The third uses the Cauchy-Schwarz inequality.
- Finally, the second to last uses Theorem 25 while, the last one, invokes the $\ell$-Lipschitz continuity of the gradient.

■

**Lemma 20** *Let $f : \mathcal{X} \times \mathcal{Y}$ be an $\ell$-smooth function such that for any $x \in \mathcal{X}$, $f(x, \cdot)$ satisfies the proximal-PŁ condition with modulus $\alpha > 0$. Then, the function $\Phi(x) := \arg\max_{y\in\mathcal{Y}} f(x, y)$ is $\ell_\star$-smooth, with*

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

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

$$\frac{\alpha}{2}\left\|y^\star(x') - y^\star(x)\right\|^2 \leq f(x, y^\star(x)) - f(x, y^\star(x')). \tag{7}$$

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-PŁ condition, we write,

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

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

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

Hence,

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

where the last line follows from a slight sharpening of the proof of Theorem 19 (for the function $h(y,x) = -f(x,y)$ and $\mathbf{M} = \mathbf{I}$). Finally, piecing inequalities (7), (8), and (9) together,

$$\left\|y^\star(x) - y^\star(x')\right\| \le \frac{\ell}{\alpha}\left\|x - x'\right\|. \tag{10}$$

What is left to do is to observe the following, due to Danskin's theorem and $\ell$-smoothness of $f$,

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

The latter inequality follows from (10) and completes the proof. ■

**Lemma 21 ([24, Lemma D.3])** *Let $f : \mathcal{X}\times\mathcal{Y}$ be an $\ell$-smooth function. Additionally, assume that $f(\cdot, y)$ is $\mu_x$-pPŁ for all $y \in \mathcal{Y}$ and $f(x,\cdot)$ is $\mu_y$-pPŁ for all $x \in \mathcal{X}$. Then, it holds true that:*

$$\Phi^\star := \min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x,y) = \max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} f(x,y).$$

.

**Lemma 22 ([24, Lemma D.4])** *Let $f : \mathcal{X}\times\mathcal{Y}$ be an $\ell$-smooth function. Additionally, assume that $f(\cdot, y)$ is $\mu_x$-pPŁ for all $y \in \mathcal{Y}$ and $f(x,\cdot)$ is $\mu_y$-pPŁ for all $x \in \mathcal{X}$. Then, the function $\Phi(x) := \max_{y\in\mathcal{Y}} f(x,y)$ is $\mu_x$-pPŁ.*

### B.3. Regarding the Mahalanobis Distance

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

**Proposition 23** *Assume that $\theta_0 = \mathbf{0}$. Also, let $v_t^\top \mathbf{1} = 0,\ \forall t \in \{1, 2, 3, \dots\}$. Then, setting $\theta_{t+1} = \theta_t - \eta \mathbf{M}^\dagger v_t$ guarantees that,*

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

**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 $(\theta_{t+1} - \theta_t)^\top \mathbf{1} = 0$. ■

**Proposition 24** *Let $\Theta \subseteq \mathbb{R}^d$ be a convex compact set. Assume that $\theta_0 = \mathbf{0}$. Also, let $v_t^\top \mathbf{1} = 0,\ \forall t \in \{1, 2, 3, \dots\}$. Then, the following minimization problem has a unique solution,*

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

*Further, it is equivalent to the minimization problem,*

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

**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 convex in $\theta$. Hence, both problems attain a unique minimum.

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

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

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

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

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

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

The final inequality, is exactly the first-order optimality condition for the second minimization problem. ■

### B.4. Alternating Mirror Descent using a Changing Mahalanobis DGF

#### B.4.1. SUPPORTING LEMMATA

**Lemma 25 ()** *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 $\eta > 0$. Also, let points $x_1^+, x_2^+ \in \mathcal{X}$ such that:*

$$x_1^+ := \operatorname*{Proj}_{\mathcal{X}, \mathbf{M}_t} \left( x - \eta {\mathbf{M}_t}^{-1} v_1 \right);$$
$$x_2^+ := \operatorname*{Proj}_{\mathcal{X}, \mathbf{M}_t} \left( x - \eta {\mathbf{M}_t}^{-1} v_2 \right).$$

*Then, it holds true that:*

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

.

**Smoothness Relative to the Mahalanobis Distance**

**Proposition 26** *Let $f$ be a function $\ell$-smooth relative to the $\ell_2$-distance. Then, it is $\frac{\ell}{\lambda_{\min}(\mathbf{M}_t)}$-smooth relative to the Mahalanobis distance induced by a positive definite matrix $\mathbf{M}_t$.*

**Proof** We will merely demonstrate that if $f$ is $\ell$-smooth (relative to $\ell_2$-distance) it is also the case that:

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

For one direction we use vector norm equivalence to write:

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

Correspondingly for the opposite direction:

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

■

### B.4.2. CONVERGENCE OF ALTERNATING DESCENT-ASCENT

Through, we consider this section, we consider the iteration following scheme,

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

We make a standard assumption on the gradient estimators and their second moments.

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

$$\mathbb{E}\left[\hat{g}_x(x_t, y_t)\right] = g_x(x_t, y_t),$$

$$\mathbb{E}\left[\hat{g}_y(x_t, y_t)\right] = g_y(x_t, y_t),$$

*and*

$$\mathbb{E}\left[\|\hat{g}_x(x_t, y_t)\|^2\right] \leq \sigma_x^2,$$

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

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

**Theorem 27** *Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ an $\ell$-smooth function and bounded in the interval $\Delta_f$. Further, assume $\mathcal{X}, \mathcal{Y}$ to be two convex sets with Euclidean diameters, $\operatorname{diam}(\mathcal{X}), \operatorname{diam}(\mathcal{Y})$. Moreover, assume that $f$ satisfies a two-sided* pPŁ condition *with moduli $\mu_x$ for all $y \in \mathcal{Y}$ and $\mu_y$ for any $x \in \mathcal{X}$. Additionally, let $(\hat{g}_x, \hat{g}_y)$ be an inexact stochastic gradient oracle satisfying Assumption 3.*

- *When $\mathbf{M}_{\cdot t} = \mathbf{I}$, after $T$ iterations of* (Alt-GDA) *with a choice of stepsizes $\eta_x = \frac{\mu_y^2}{960\ell^3}$ and $\eta_y = \frac{1}{5\ell}$, it holds true that:*

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

  *where, $\Delta_f := \max_{x\in\mathcal{X}, y\in\mathcal{Y}} f(x,y) - \min_{x\in\mathcal{X}, y\in\mathcal{Y}} f(x,y)$ and $c_1, c_2 \in O(1)$.*

- *For a general positive definite choice of $\mathbf{M}_{\cdot t}$ (Mahalanobis metric), after $T$ iterations of* (Alt-GDA) *with a choice of stepsizes $\eta_x = \frac{\mu_y^2}{960\ell^3\lambda_{\max}^2}$ and $\eta_y = \frac{1}{5\ell\lambda_{\max}}$, it holds true that:*

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

  *where, $\Delta_f := \max_{x\in\mathcal{X}, y\in\mathcal{Y}} f(x,y) - \min_{x\in\mathcal{X}, y\in\mathcal{Y}} f(x,y)$, $\lambda_{\max} := \max_t \lambda_{\max}(\mathbf{M}_{\cdot,t}^{-1})$ and $c_1, c_2 \in O(1)$.*

**Proof** To prove convergence we will use the Lyapunov function $L(x,y) := U(x,y) + cW(x,y)$ with $U(x,y) := \mathbb{E}\left[\Phi(x) - \Phi^\star\right]$, $W(x,y) := \mathbb{E}\left[\Phi(x) - f(x,y)\right]$ and $c > 0$. Intuitively, $U(x,y)$ measures $x$'s success in achieving the unique minmax value $\Phi^\star$, while $W(x,y)$ measures $y$'s success in achieving to be a best-response to its corresponding $x$. We begin with some preliminary work to ultimately setup a recursion on $L$.

**Descent on $\Phi$** In order to guarantee descent, by Theorem 20, Theorem 26, and Theorem 18, it suffices to pick $\eta_x \leq \frac{1}{5\ell\lambda_{\max}(\mathbf{M}_{x,t})}$. Then, we can write,

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

Equivalently, subtracting $\Phi^\star$ from both sides yields,

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

Further, a simple re-arrangement reads,

$$\mathbb{E}\Phi(x_{t+1}) - \mathbb{E}\Phi(x_t) \leq -\frac{\eta_x}{6}\mathbb{E}\mathcal{D}_{\mathcal{X}}^{\Phi}(x_t, 1/\eta_x) + \eta_x\mathbb{E}\left\|\nabla_x\Phi(x_t) - \nabla_x f(x_t, y_t)\right\|_{\mathbf{M}_{x,t}^{-1}}^2$$

$$+ 2\eta_x\sigma_x^2 + 2\eta_x\delta_x^2.$$

Requiring that $\eta_y \leq \frac{1}{5\ell\lambda_{\max}(\mathbf{M}_{y,t})}$, (Theorem 26 and Theorem 18), we write:

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

Invoking Theorem 22, multiplying by $-1$, and adding $\Phi(x_{t+1})$ will yield,

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

As a reminder, $\Phi$ is a pPL function relative to the Mahalanobis distance induced by $\mathbf{M}_t$ by Theorem 22.

**Upper bound on the descent of $f(\cdot, y)$** From the smoothnes of $f$:

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

Re-arranging to isolate $f(x_t, y_t) - f(x_{t+1}, y_t)$,

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

Putting the pieces together for $\Phi(x_t) - f(x_t, y_t)$, we get:

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

**Decrease in the Lyapunov function** We consider the Lyapunov function $L(x, y) := U(x, y) + cW(x, y)$ with $U(x, y) := \mathbb{E}\left[\Phi(x) - \Phi^\star\right]$, $W(x, y) := \mathbb{E}\left[\Phi(x) - f(x, y)\right]$ and shorthand notation $U_t = U(x_t, y_t), W_t = W(x_t, y_t)$. Here $U_t$ measures primal suboptimality via the PL condition on $\Phi$, while $W_t$ captures the dual gap $\Phi(x_t) - f(x_t, y_t)$.

$$\begin{aligned}
&U_{t+1} + cW_{t+1} \\
&\quad\leq U_t - \frac{\eta_x}{6}\mathbb{E}\mathcal{D}^{\Phi}_{\mathcal{X}}(x_t, 1/\eta_x) + \eta_x\mathbb{E}\left\|\nabla_x\Phi(x_t) - \nabla_x f(x_t, y_t)\right\|^2_{\mathbf{M}^{-1}_{x,t}} \\
&\qquad + c\left(1 - \frac{\mu_y\eta_y}{6}\right)\mathbb{E}W_t
\end{aligned}$$

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

- (11) uses the fact that $a \le |a-b|+b$ for $a=\mathcal{D}_{\mathcal{X}}(x_t,1/\eta_x;y_t), b=\mathcal{D}_{\mathcal{X}}^{\Phi}(x_t,1/\eta_x)$. This decomposition isolates the term $|\mathcal{D}_{\mathcal{X}}-\mathcal{D}_{\mathcal{X}}^{\Phi}|$, which can then be controlled using the Mahalanobis continuity lemma in $y$.
- (12) uses Theorem 19 and Danskin's theorem; this yields a bound $|\mathcal{D}_{\mathcal{X}}-\mathcal{D}_{\mathcal{X}}^{\Phi}| \le 3\lambda_{\max}(\mathbf{M}_{x,t}^{-1})\ell^2\|y_t-y^\star(x_t)\|^2$.

$$
\begin{aligned}
U_{t+1}+cW_{t+1} \le\ & U_t-\frac{\eta_x}{6}\mathbb{E}\mathcal{D}_{\mathcal{X}}^{\Phi}(x_t,1/\eta_x)+\frac{2\eta_x\lambda_{\max}(\mathbf{M}_{x,t}^{-1})\ell^2}{\mu_{\mathrm{qg}}}W_t\\
&+ c\left(1-\frac{\mu_y\eta_y}{6}\right)\mathbb{E}W_t\\
&+ c\left(1-\frac{\mu_y\eta_y}{6}\right)\mathbb{E}\left[-\frac{\eta_x}{6}\mathbb{E}\mathcal{D}_{\mathcal{X}}^{\Phi}(x_t,1/\eta_x)+\frac{2\eta_x\lambda_{\max}(\mathbf{M}_{x,t}^{-1})\ell^2}{\mu_{\mathrm{qg}}}W_t\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ c\left(1-\frac{\mu_y\eta_y}{6}\right)\frac{3\eta_x}{2}\mathbb{E}\left[\frac{6\lambda_{\max}(\mathbf{M}_{x,t}^{-1})\ell^2}{\mu_{\mathrm{qg}}}W_t\right]\\
&+ c\eta_y\delta_y^2 + c\eta_y\sigma_y^2 + c\eta_x\left(1-\tfrac{\mu_y\eta_y}{6}\right)\left(\tfrac{13}{2}\,\sigma_x^2+\tfrac{11}{2}\,\delta_x^2\right)+2\eta_x\sigma_x^2+2\eta_x\delta_x^2\\
\le\ &\varpi_1 U_t + c\varpi_2 W_t\\
&+ c\eta_y\delta_y^2 + c\eta_y\sigma_y^2 + c\eta_x\left(1-\tfrac{\mu_y\eta_y}{3}\right)\left(\tfrac{13}{2}\,\sigma_x^2+\tfrac{11}{2}\,\delta_x^2\right)+2\eta_x\sigma_x^2+2\eta_x\delta_x^2
\end{aligned}
$$

We then collect the coefficients in front of $U_t$ and $W_t$ in the previous inequality into $\varpi_1$ and $\varpi_2$, respectively, so that the Lyapunov recursion can be written compactly as $U_{t+1}+cW_{t+1}\le\varpi_1 U_t + c\varpi_2 W_t+\text{noise}$. *I.e.*,

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

For $\varpi_1$, letting $c=1/10$

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

For $\varpi_2$, we distinguish two cases relevant to our algorithms, $\mathbf{M}_t=\mathbf{I}$ and a general choice of $\mathbf{M}_t$.

- For $\mathbf{M}_t=\mathbf{I}$, it holds that $\lambda_{\max}\left(\mathbf{M}_{\cdot,t}^{-1}\right)=1$, and $\mu_{\mathrm{qg}}=\mu_y$. So we write

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

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

- For a general choice of $\mathbf{M}_t$, let $\lambda_{\max}:=\max\{\lambda_{\max}(\mathbf{M}_{x,t}^{-1}),\lambda_{\max}(\mathbf{M}_{y,t}^{-1})\}$ and $\overline{\mu_y}\leftarrow\min\{\mu_{\mathrm{qg}},\mu_y\}$,

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

  Similarly, we need to set

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

  This in turn yields $\eta_y=\frac{1}{5\lambda_{\max}\ell}$ and $\eta_x=\frac{\overline{\mu_y}^2}{960\ell^3\lambda_{\max}^2}$.

**Remark 28** *In fact,* $\mathbf{M}_t$ *is allowed to be positive semidefinite as long as the gradient throuhgout the iterations is in the kernel of* $\mathbf{M}_t$*.*

■

## Appendix C. Further Preliminaries on IIEFGs

### C.1. The Behavioral and Sequence-Form Strategies

**Lemma 29** *Under Assumption 2, the transforms $c_1^{-1} : \mathcal{M}_1 \to \mathcal{X}_\gamma, c_2^{-1} : \mathcal{M}_2 \to \mathcal{Y}_\gamma$ are Lipschitz continuous. I.e., for any $\mu_1, \mu_1'$, it holds true that,*

$$\left\|c_1^{-1}(\mu_1) - c_1^{-1}(\mu_1')\right\| \leq \frac{2|\mathcal{H}|\sqrt{A_{\max}}}{\gamma} \left\|\mu_1 - \mu_1'\right\|$$

*and for any $\mu_2, \mu_2'$,*

$$\left\|c_2^{-1}(\mu_2) - c_2^{-1}(\mu_2')\right\| \leq \frac{2|\mathcal{H}|\sqrt{B_{\max}}}{\gamma} \left\|\mu_2 - \mu_2'\right\|.$$

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

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

As a reminder, for all $s \in \mathcal{S}_1$ it holds that $\mu_1(s) \geq \frac{\gamma}{|\mathcal{H}|}$ by Assumption 2. Proceeding towards the desired inequality,

$$\begin{aligned}
&\left\|c_1^{-1}(\mu_1) - c_1^{-1}(\mu_1')\right\|^2 \\
&= \sum_{s\in\mathcal{S}_1}\sum_{a\in\mathcal{A}_s}\left[\left(\frac{\mu_1(s,a)}{\mu_1(s)} - \frac{\mu_1'(s,a)}{\mu_1(s)}\right) + \frac{\mu_1'(s) - \mu_1(s)}{\mu_1(s)\mu_1'(s)}\mu_1'(s,a)\right]^2 \\
&\leq 2\sum_{s\in\mathcal{S}_1}\sum_{a\in\mathcal{A}_s}\left(\frac{\mu_1(s,a)}{\mu_1(s)} - \frac{\mu_1'(s,a)}{\mu_1(s)}\right)^2 + 2\sum_{s\in\mathcal{S}_1}\sum_{a\in\mathcal{A}_s}\left(\frac{\mu_1'(s) - \mu_1(s)}{\mu_1(s)\mu_1'(s)}\right)^2 {\mu_1'}^2(s,a) \\
&\leq 2\sum_{s\in\mathcal{S}_1}\sum_{a\in\mathcal{A}_s}\left(\frac{\mu_1(s,a)}{\mu_1(s)} - \frac{\mu_1'(s,a)}{\mu_1(s)}\right)^2 + 2\sum_{s\in\mathcal{S}_1}\sum_{a\in\mathcal{A}_s}\left(\frac{\mu_1'(s) - \mu_1(s)}{\mu_1(s)\mu_1'(s)}\right)^2 {\mu_1'}^2(s) \\
&\leq \frac{2|\mathcal{H}|^2}{\gamma^2}\sum_{s\in\mathcal{S}_1}\sum_{a\in\mathcal{A}_s}\left(\mu_1(s,a) - \mu_1'(s,a)\right)^2 + \frac{2|\mathcal{H}|^2}{\gamma^2}\sum_{s\in\mathcal{S}_1}\sum_{a\in\mathcal{A}_s}\left(\mu_1'(s) - \mu_1(s)\right)^2 \\
&\leq \frac{2|\mathcal{H}|^2}{\gamma^2}\left\|\mu_1 - \mu_1'\right\|^2 + \frac{2A_{\max}|\mathcal{H}|^2}{\gamma^2}\sum_{s\in\mathcal{S}_1}\left(\mu_1'(s) - \mu_1(s)\right)^2 \\
&= \frac{2|\mathcal{H}|^2}{\gamma^2}\left\|\mu_1 - \mu_1'\right\|^2 + \frac{2A_{\max}|\mathcal{H}|^2}{\gamma^2}\sum_{s\in\mathcal{S}_1}\left(\sum_{a\in\mathcal{A}_s}\mu_1'(s,a) - \mu_1(s,a)\right)^2. \qquad (13)
\end{aligned}$$

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

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

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,

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

Plugging this inequality into (13) yields the desired bound. ■

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

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

- $\sigma_1(h), \sigma_2(h)$ returns the last history before $h$ where player 1 (player 2, resp.) took an action,
- $h \in s$ signifies that history $h$ belongs in the infoset $s$,
- $h' \succeq_{\mathcal{T}} h, h' \succeq_{\mathcal{T}} (h,a)$ signifies that $h'$ is a successor/child node of $h$, $(h,a)$;
- $h \in \xi, (h,a) \in \xi$ signifies that $h$, $h,a$ belongs in the game trajectory $\xi$ from the root to a terminal node.

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

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

The value function of each infoset is defined as,

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

Also, the action-value function reads:

$$Q_1^\pi(s,a) := \mathbb{E}_{\xi\sim\pi}\left[\sum_{h'\in\xi, h'\succeq_{\mathcal{T}}(h,a)} r(h') \Big| \exists h \in s : (h,a)\in\xi\right]$$
$$= \frac{1}{\sum_\xi \mathbb{P}^\pi(\xi)\mathbb{1}\{\exists h\in s:(h,a)\in\xi\}} \sum_\xi \mathbb{P}^\pi(\xi)\mathbb{1}\{\exists h\in s:(h,a)\in\xi\}\left[\sum_{\substack{h'\in\xi,\\ h'\succeq_{\mathcal{T}}(h,a)}} r(h')\right].$$

We define the advantage function to be:

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

Finally, let a policy pair $\pi_1, \pi_2$ and $\pi := (\pi_1, \pi_2)$. Let $\pi_1$ be parametrized by some vector $\theta$. We compute the policy gradient for $\theta$,

$$\begin{aligned}
\frac{\partial V_1^\pi}{\partial \theta_{s,a}} &= \frac{\partial}{\partial\theta_{s,a}} \sum_\xi r_1(\xi)\,\mathbb{P}^\pi(\xi) \\
&= \sum_\xi r_1(\xi)\,\mathbb{P}^\pi(\xi)\frac{\partial \log \mathbb{P}^\pi(\xi)}{\partial\theta_{s,a}} \\
&= \sum_\xi \sum_{a'} r_1(\xi)\,\mathbb{P}^\pi(\xi)\frac{\partial\log\pi_1(a'|s)}{\partial\theta_{s,a}}\mathbb{1}\{\exists h\in s:(h,a')\in\xi\} \\
&= \sum_\xi \sum_{a'} \left(r_1(\xi)\,\mathbb{P}^\pi(\xi)\frac{\mathbb{1}\{\exists h\in s:(h,a')\in\xi\}}{\pi_1(a'|s)}\right)\pi_1(a'|s)\frac{\partial\log\pi_1(a'|s)}{\partial\theta_{s,a}} \\
&= \sum_\xi \sum_{a'} \left(\left[\sum_{\substack{h'\in\xi,\\ h'\succeq_{\mathcal{T}}(h,a)}} r(h') + \sum_{\substack{h'\in\xi,\\ h'\prec_{\mathcal{T}}(h,a)}} r(h')\right]\mathbb{P}^\pi(\xi)\frac{\mathbb{1}\{\exists h\in s:(h,a')\in\xi\}}{\pi_1(a'|s)}\right)\pi_1(a'|s)\frac{\partial\log\pi_1(a'|s)}{\partial\theta_{s,a}} \\
&= \sum_\xi \sum_{a'} \left(\left[\sum_{\substack{h'\in\xi,\\ h'\prec_{\mathcal{T}}(h,a)}} r(h')\right]\mathbb{P}^\pi(\xi)\frac{\mathbb{1}\{\exists h\in s:(h,a')\in\xi\}}{\pi_1(a'|s)}\right)\pi_1(a'|s)\frac{\partial\log\pi_1(a'|s)}{\partial\theta_{s,a}} \\
&+ d^\pi(s)\sum_{a'}\pi_1(a'|s)\frac{\partial\log\pi_1(a'|s)}{\partial\theta_{s,a}}Q^\pi(s,a') \\
&= d^\pi(s)\sum_{a'}\pi_1(a'|s)\frac{\partial\log\pi_1(a'|s)}{\partial\theta_{s,a}}Q^\pi(s,a'). \qquad (14)
\end{aligned}$$

Where we have used the following fact,

$$\sum_\xi \sum_{a'} \left(\left[\sum_{\substack{h'\in\xi,\\ h'\prec_{\mathcal{T}}(h,a)}} r(h')\right]\frac{\mathbb{1}\{\exists h\in s:(h,a')\in\xi\}}{\pi_1(a'|s)}\right)\pi_1(a'|s)\frac{\partial\log\pi_1(a'|s)}{\partial\theta_{s,a}}$$

$$= \sum_{a'} \underbrace{\sum_{\xi} \left( \left[ \sum_{\substack{h' \in \xi, \\ h' \prec_{\mathcal{T}} (h,a)}} r(h') \right] \frac{\mathbb{1}\{\exists h \in s : (h, a') \in \xi\}}{\pi_1(a'|s)} \right)}_{=:C(s)} \pi_1(a'|s) \frac{\partial \log \pi_1(a'|s)}{\partial \theta_{s,a}}$$

$$= \sum_{a'} C(s) \pi_1(a'|s) \frac{\partial \log \pi_1(a'|s)}{\partial \theta_{s,a}}$$

$$= C(s) \sum_{a'} \pi_1(a'|s) \frac{\partial \log \pi_1(a'|s)}{\partial \theta_{s,a}}$$

$$= C(s) \frac{\partial}{\partial \theta_{s,a}} \sum_{a'} \pi_1(a'|s)$$

$$= C(s) \frac{\partial}{\partial \theta_{s,a}} 1 = 0.$$

Further, for direct policy parametrization, we get,

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

For the softmax policy parametrization, (14) yields,

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

### C.3. Properties of the Bidilated Regularizer

Introduced in [32], the bidilated regularizer offers an alternative to the commonly used dilated regularizer [20]. It can be seamlessly used along $Q$ feedback by dropping the need of importance sampling which would be necessary for the *dilated regularizer* when the gradient is estimated through trajectory roll-outs. The purpose of this refined regularizer was introducing a distance generating function in the sequence-form space that would not necessitate importance sampling.

#### C.3.1. STRONG CONVEXITY MODULUS

**Lemma 30** *For a choice of strongly convex function $\psi$, and a weighting scheme $\{w_{1,s}\}_{s \in \mathcal{S}_1}$, $\{w_{2,s}\}_{s \in \mathcal{S}_2}$ and let $\alpha_{\text{dil}} > 0$ be the modulus of the weighted dilated regularizer. Then, the corresponding bidilated regularizer is strongly convex,*

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

**Proof** These calculations were used in the proof of [32, Lemma D.1]; we repeat them for completeness. 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 regularizer is defined as,

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

We can slightly refine [32, Lemma C.1] in order to compute an explicit lower bound on the convexity modulus of different weighted bidilated regularizer depending on the choice of $\psi$. 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 conclude that,

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

By [33, Lemma D.2] we know that,

$$B_{\mathcal{R}_1^{\psi}}\left(\mu_1^{\pi_1'}\middle\|\mu_1^{\pi_1};\mu_2^{\pi_2}\right) \geq \frac{\gamma}{|\mathcal{H}|}\min_h \mu_c(h) B_{\psi}^{\mathrm{dil}}(\mu_1^{\pi_1'}\|\mu_1^{\pi_1}).$$

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

$$\alpha_{\mathrm{bi}} := \frac{\gamma}{|\mathcal{H}|}\min_h \mu_c(h)\,\alpha_{\mathrm{dil}}.$$

■

By [13, Corollary 1], we know that there exists a weighting scheme, such that the Euclidean dilated regularizer is 1-strongly convex w.r.t. the $\ell_2$-norm. The procedure assigns weights to nodes in a bottom-up fashion.

- At each leaf node $s$, the weights are set to
$$w_{1,s} = 1.$$
- For an internal node $s$, let $s_a, s_{a'}, \ldots$ denote its child nodes under actions $a, a', \ldots$. For each action $a$, compute
$$W_{1,s_a} = \sum_{s'\succeq_{\mathcal{T}}(s,a)} w_{1,s'}.$$
- The node's weights are then set to
$$w_{1,s} = 2\max_a W_{1,s_a}.$$

**Corollary 31 (Euclidean Regularizer)** *There exists a choice of weights, with* $\max_s w_{1,s}, \max_s w_{2,s} = \Theta(2^{D(\mathcal{T})})$*, and under the assumption that* $\min_s \mu_2(s) \geq \gamma$*, the bidilated Euclidean regularizer has a strong convexity modulus w.r.t. the* $\ell_2$*-norm,* $\alpha_{\text{bi}}$*,*

$$\alpha_{\text{bi}}^{\text{eucl}} := \frac{\gamma}{|\mathcal{H}|} \min_h \mu_c(h).$$

[26, Theorem 2] states that a recursion defines weights with $\max_s w_{1,s}, \max_s w_{2,s} = \Theta(2^{D(\mathcal{T})})$ such that the entropic dilated regularizer is strongly convex w.r.t. the $\ell_2$-norm.

**Corollary 32 (Entropic Regularizer)** *There exists a choice of weights, and under the assumption that* $\min_s \mu_2(s) \geq \gamma$*, the bidilated entropic regularizer has a strong convexity modulus w.r.t. the* $\ell_2$*-norm,* $\alpha_{\text{bi}}$*,*

$$\alpha_{\text{bi}}^{\text{ent}} := \frac{\gamma}{|\mathcal{H}|} \min_h \mu_c(h).$$

### C.3.2. LIPSCHITZ MODULI

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

**Euclidean regularizer**

**Lemma 33** *The weighted Euclidean bidilated regularizer is* $\ell$*-smooth with*

$$\ell := \Theta\left(2^{D(\mathcal{T})} D(\mathcal{T}) S\right).$$

**Proof** We write the bidilated regularizer as

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

For a fixed $\pi_2$, we have

$$\nabla_{\pi_1} \mathcal{R}_1^{\text{eucl}}(\pi_1, \pi_2) = \mathbf{J}_f(\pi_1, \pi_2)^\top g(\pi_1) + \mathbf{J}_g(\pi_1)^\top f(\pi_1, \pi_2),$$

where, $f(\pi_1, \pi_2), g(\pi_1) \in \mathbb{R}^{|\mathcal{H}|}$ with $f(\pi_1, \pi_2) = \sum_{h \in s} \mu_c(h) \mu_2^{\pi_2}(\sigma_2(h)) \mu_1^{\pi_1} \sigma_1((h))$ and $g_s(\pi_1) = w_{1,s} \left\|\pi_1(\cdot|s)\right\|^2$.

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

- 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 $\ell_2$-norm and the fact that $\pi_1(\cdot|s)$ lies in the simplex, *i.e.*, $\|\pi_1(\cdot|s)\|_2 \leq 1$. Also, the weight $w_{1,s}$ only scales the local quadratic term.
- 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 variables bounded by 1. Also, it holds that $\max_{\pi_1,\pi_2} \|f(\pi_1)\| \leq \sqrt{D(\mathcal{T})}$.

Concluding,

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

Symmetrically,

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

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

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

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

■

Symmetric arguments yield the bounds for $\nabla_\theta\mathcal{R}_1, \nabla_\theta\mathcal{R}_2$.

**Entropic regularizer**

**Lemma 34** *The weighted entropic bidilated regularizer is $\ell$-smooth with*

$$\ell := \Theta\left(2^{D(\mathcal{T})}D(\mathcal{T})\max\{(1+\log A_{\max}),(1+\log B_{\max})\}S\right).$$

**Proof** We write $\mathcal{R}_2$ as the inner product of $f(\pi_\chi) := d^{\pi_\chi,\pi_\theta}$ and $g := [\pi_\theta(b|s)\log\pi_\theta(b|s)]_{s,b}$. For notational convenience, we suppress dependence of $f, g$ on $\pi_\theta$.

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

We now bound the Lipschitz modulus of the gradient using the chain rule:

$$\begin{aligned}&\left\|\nabla_\chi\mathcal{R}_2(\pi_\chi,\pi_\theta) - \nabla_\chi\mathcal{R}_2(\pi_{\chi'},\pi_\theta)\right\| \leq \left\|\mathbf{J}_\pi(\chi)^\top\mathbf{J}_f(\pi_\chi) - \mathbf{J}_\pi(\chi')^\top\mathbf{J}_f(\pi_{\chi'})\right\|\|g\| \\ &\leq \left(\left\|\mathbf{J}_\pi(\chi)^\top\mathbf{J}_f(\pi_\chi) - \mathbf{J}_\pi(\chi)^\top\mathbf{J}_f(\pi_{\chi'})\right\| + \left\|\mathbf{J}_\pi(\chi)^\top\mathbf{J}_f(\pi_{\chi'}) - \mathbf{J}_\pi(\chi')^\top\mathbf{J}_f(\pi_{\chi'})\right\|\right)\|g\| \\ &\leq \left(\|\mathbf{J}_\pi(\chi)\|\left\|\mathbf{J}_f(\pi_\chi) - \mathbf{J}_f(\pi_{\chi'})\right\| + \left\|\mathbf{J}_f(\pi_{\chi'})\right\|\left\|\mathbf{J}_\pi(\chi) - \mathbf{J}_\pi(\chi')\right\|\right)\|g\| \\ &\leq \left(D(\mathcal{T}) + \tfrac{3}{4}D(\mathcal{T})(S_1A_{\max})^{3/2}\right)\max_s w_{2,s}\sqrt{\log B_{\max}}\left\|\chi - \chi'\right\|.\end{aligned}$$

For the Lipschitz modulus of $\nabla_\chi \mathcal{R}_1(\pi_\chi, \pi_\theta)$, we re-purpose the lengthy calculations found in the proof of [35, Lemma 14], we consider $\chi = \chi_0 + \alpha u$ for some $u, \chi \in \mathbb{R}^A, \alpha \in \mathbb{R}$,

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

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

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

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

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

and, as such,

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

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

## Appendix D. Regarding the Policy Parametrization

### D.1. Definitions

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

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

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

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

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

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

### D.2. General Properties under Parameter Constraints

**Lemma 35** *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:*

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

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

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

**Proof** For brevity, define $\sigma := \operatorname{softmax}(\theta)$, and let $\operatorname{diag}(v)$ be the $d \times d$ diagonal matrix "whose diagonal entries are given by $v \in \mathbb{R}^d$,

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

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

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

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

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

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

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

■

**Lemma 36 ([54, Lemma 5.3])** *The softmax map is* $8$*-smooth.*

**Lemma 37** *The softmax map* $\operatorname{softmax} : \mathbb{R}^d \to \mathbb{R}^d$ *has an* $\frac{3}{\sqrt{2}}d^{3/2}$*-smooth gradient.*

**Proof** Again we use $\sigma := \operatorname{softmax}(\theta)$ for brevity. We compute the second order derivatives:

$$\begin{aligned}\frac{\partial^2}{\partial\theta_j\partial\theta_k}\sigma_i &= \frac{\partial}{\partial\theta_k}[\sigma_i(\delta_{ij} - \sigma_j)]\\ &= \sigma_i(\delta_{ik} - \sigma_k)(\delta_{ij} - \sigma_j) - \sigma_i\sigma_j(\delta_{jk} - \sigma_k).\end{aligned}$$

Every term is a function of $\theta$ and it is true in general that

$$\begin{aligned}&\left|f(\theta)\,g(\theta)\,h(\theta)\;-\;f(\theta')\,g(\theta')\,h(\theta')\right|\;\leq\\&|f(\theta)-f(\theta')|\,|g(\theta)|\,|h(\theta)|\;+\;|g(\theta)-g(\theta')|\,|f(\theta')|\,|h(\theta)|\;+\;|h(\theta)-h(\theta')|\,|f(\theta')|\,|g(\theta')|.\end{aligned}$$

As such, we can write,

$$\left|\frac{\partial^2}{\partial\theta_j\partial\theta_k}\sigma_i(\theta)-\frac{\partial^2}{\partial\theta_j\partial\theta_k}\sigma_i(\theta')\right|\leq 3\left\|\theta-\theta'\right\|$$

■

**Lemma 38** *Assume $\theta\in\mathbb{R}^d$ with $\theta\in\Theta_R:=\{\theta\in\mathbb{R}^d:\theta^\top\mathbf{1}=0$ and $|\theta_i-\theta_j|\leq 2R,\ \forall i,j\in[d]\}$. Then, the following bounds hold true,*

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

**Proof**

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

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

By observing that,

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

We can lower bound the value as,

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

It suffices to maximize the quantity $\max_{j\neq1,\theta\in\Theta_R}\{\theta_j-\theta_1\}$ as the RHS quantity is non-increasing in $\max_{j\neq1,\theta\in\Theta_R}\{\theta_j-\theta_1\}$. *I.e.,* the largest difference between two coordinates of a vector in the sphere is $2R$. The minimum is achieved when $\theta_j-\theta_1=2R$ and $\theta_j=\theta_k,\forall j,k\geq2$.

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

$$\begin{aligned}\frac{e^{\theta_1}}{\sum_i e^{\theta_i}}&=\frac{e^{\theta_1}}{e^{\theta_1}+\sum_{i\neq1}e^{\theta_i}}\\&\leq\frac{e^{\theta_1}}{e^{\theta_1}+(d-1)e^{\sum_i\theta_i/(d-1)}}\end{aligned}$$

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

$$\max_{a,b} \frac{1}{1+(d-1)e^{b-a}}$$
$$\text{s.t. } |a-b| \leq 2R.$$

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

$$\min_{a,b} b-a$$
$$\text{s.t. } |a-b| \leq 2R.$$

Finally, it is clear that the last objective is minimized for $a-b=-2R$. Letting $\varepsilon \leq (d-1)^{-2}$. ■

In this vein, if we want to bound the minimum probability of the softmax parametrized policy by $\varepsilon > 0$ for some $R > 0$, we need to set $R \leq 1/2 \log\left(\frac{1-\varepsilon}{\varepsilon(d-1)}\right)$. Then, it is also the case that $\max_{\theta \in \Theta_R, i} \text{softmax}_i(\theta) \geq \frac{1-\varepsilon}{1-\varepsilon+\varepsilon(d-1)^2} \geq 1-\varepsilon-\varepsilon(d-1)^2$.

**Proposition 39** *Let $p$ be a probability vector in $\Delta^{d-1}$ and define $\theta(p)$ to be the set of $\theta$ such that $\text{softmax}(\theta) = p$. For any two $\theta, \theta' \in \theta(p)$, there exists a $c \in \mathbb{R}$ such that $\theta = \theta' + c\mathbf{1}$.*

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

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

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

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

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

**Proposition 40** *Let $p \in \Delta^{d-1}$ be a probability vector and the set, $\theta(p)$ , of vectors $\theta \in \mathbb{R}^d$ such that $\text{softmax}(\theta) = p$. For the vector $\theta^\star := \arg\min_{\theta \in \Theta(p)} \|\theta\|^2$ it holds true that,*

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

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

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

By the first-order optimality conditions, $c = -\frac{1}{d}\theta_0^\top \mathbf{1}$. Plugging back this for $\theta^\star$, we see $\theta^\star = \theta_0 - \frac{1}{d}\mathbf{1}(\theta_0^\top \mathbf{1})$. We see that, $\mathbf{1}^\top \theta^\star = \mathbf{1}^\top \theta_0 - \frac{d}{d}\boldsymbol{\theta}_0^\top \mathbf{1} = 0$. ■

**Lemma 41** *Assume a fixed $0 < R < \infty$ and define the set $\Theta_R$ to be $\Theta_R := \{\theta \in \mathbb{R}^d : \theta^\top \mathbf{1} = 0$ and $|\theta_i - \theta_j| \leq 2R,\ \forall i, j \in [d]\}$. Then, $\mathrm{softmax}(\Theta_R)$ is a convex set.*

**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 $\mathrm{softmax}(\theta) = p$. To see this, we apply the logarithm on the inequalities,

$$-2R \leq \log p_i - \log p_j \leq 2R. \tag{15}$$

A vector $\chi$ with entries $\chi_i := \log p_i$ clearly implements $p$. By (15) we see that subtracting $\kappa = \frac{\max_j \log p_j + \min_k \log p_k}{2}$ from all entries yields a softmax-equivalent vector $\chi'_i := \log p_i - \kappa$ with $-R \leq \chi'_i \leq R$. Conversely, for any $\theta \in \Theta_R$, $e^{-2R} \leq \frac{\mathrm{softmax}_i(\theta)}{\mathrm{softmax}_j(\theta)} \leq e^{2R}$.

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 and as such, convex. ■

## Appendix E. Gradient Domination

In this section we prove the gradient domination properties of the utilities of the game with different policy parametrizations. Further, for clarity, in place of $V_\tau^{x,y}$ we will use $V_\tau(x, y)$; and in place of $V_\tau^{\pi_\chi, \pi_\theta}$ we will use $V_\tau(\chi, \theta)$.

### E.1. Direct Policy Parametrization pPŁ

**Lemma 42** *The utility of the game regularized with the weighted bidilated Euclidean regularizer with a weighting scheme defined in* **??** *C.3.1, satisfies the* pPŁ *condition for directly parametrized policies,*

$$\frac{\tau \min_h \mu_c(h)\gamma^3}{101|\mathcal{H}|^3}[V_\tau(x, y) - V_\tau(x^\star, y)] \leq \frac{1}{2}\mathcal{D}_\mathcal{X}(x, \ell; y);$$
$$\frac{\tau \min_h \mu_c(h)\gamma^3}{101|\mathcal{H}|^3}[V_\tau(x, y^\star) - V_\tau(x, y)] \leq \frac{1}{2}\mathcal{D}_\mathcal{Y}(y, \ell; x).$$

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

$$H_\tau^{\mathrm{eucl}}(\mu_1,\mu_2) := \langle \mu_1, \mathbf{R}\mu_2\rangle - \tau\mathcal{R}_1^{\mathrm{eucl}}(\mu_1,\mu_2) + \tau\mathcal{R}_2^{\mathrm{eucl}}(\mu_1,\mu_2).$$

For player 1, we know that the function $H_\tau^{\mathrm{eucl}}$ is strongly convex with an appropriate weighting scheme $\{w_{1,s}\}$, (correspondingly $\{w_{2,s}\}$ for player 2),

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

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

$$H_\tau^{\mathrm{eucl}}(\mu_1,\mu_2) - \min_{\mu_1^\star} H_\tau^{\star,\mathrm{eucl}}(\mu_1{}^\star,\mu_2) \leq \frac{1}{2\tau\alpha_{\mathrm{bi}}^{\mathrm{eucl}}\left(\frac{\gamma}{|\mathcal{H}|}\right)^2}\left\|s_x\right\|_2^2. \tag{16}$$

Now, we know that $\alpha_{\mathrm{bi}}^{\mathrm{eucl}} = \frac{\gamma \min_h \mu_c(h)}{|\mathcal{H}|}$ (Theorem 31). The conclusion follows from Theorem 15. ■

### E.2. Softmax Policy Parametrization pPŁ

**Lemma 43** *The utility of the game with softmax-parametrized policies satisfies the two-sided* pPŁ *condition,*

$$\frac{\tau \min_h \mu_c(h)\gamma^3}{101|\mathcal{H}|^3(1+(A-1)e^{2R})^2}\left[V_\tau(\chi,\theta) - V_\tau(\chi^\star,\theta)\right] \leq \frac{1}{2}\mathcal{D}_{X_R}(\chi,\ell;\theta)$$
$$\frac{\tau \min_h \mu_c(h)\gamma^3}{101|\mathcal{H}|^3(1+(B-1)e^{2R})^2}\left[V_\tau(\chi,\theta^\star) - V_\tau(\chi,\theta)\right] \leq \frac{1}{2}\mathcal{D}_{\Theta_R}(\theta,\ell;\chi),$$

*where $\ell$ is the smoothness constant of the softmax-parametrized utility function.*

**Proof** The main challenge in proving this lemma is the fact that the softmax mapping is not a bijection; this is manifested with a rank-deficient Jacobian of the mapping.

Concretely, from (16), we know that the KŁ-condition holds for the policies. What remains to show is that the KŁ-condition also holds for the parameters $\chi$ (and $\theta$).

For some $R > 0$, let $\mathcal{X}_R := \mathrm{softmax}(X_R)$ be the convex set of softmax-parametrized policies where $X_R := \left\{\theta \in \mathbb{R}^A, A = \sum_s A_s :\ \chi_s^\top \mathbf{1} = 0, \forall s \in \mathcal{S}_1, |\chi_{s,i} - \chi_{s,j}| \leq 2R,\ \forall i,j \in [A_s]\right\}$. By overloading notation, let $V(\pi_\chi,\pi_\theta)$ be the loss function of the minimizing player as a function of policies $\pi_\chi,\pi_\theta$ and $V(\chi,\theta)$ the utility as a function of parameters $\chi,\theta$.

Now, we note that the subgradient $s \in \partial_{\pi_\chi}\left(V(\pi_\chi,\pi_\theta) + I_{\mathcal{X}_R}(\pi_\chi)\right)$ that minimizes $\|s\|$ is such that $s^\top\mathbf{1} = 0$. So when picking a norm-minimizing $s$, it suffices to look at the set of subgradients that are perpendicular to $\mathbf{1}$. Further, the chain rule applied on $V(\pi_\chi,\pi_\theta) + I_{\mathcal{X}_R}(\pi_\chi)$ yields,

$$\partial_\chi\left(V(\pi_\chi,\pi_\theta) + I_{\mathcal{X}_R}(\pi_\chi)\right) \subseteq \mathbf{J}(\chi)\left(\nabla_\pi V(\pi_\chi,\pi_\theta) + \partial_\pi I_{\mathcal{X}_R}(\pi_\chi)\right). \tag{17}$$

Moreover, we note that by the symmetry of $\mathbf{J}(\chi)$,

$$\left\|\mathbf{J}(\chi)s\right\|^2 = s^\top\mathbf{J}(\chi)^\top\mathbf{J}(\chi)s$$

$$\geq \lambda^{+}_{\min}(\mathbf{J}(\chi)^{\top}\mathbf{J}(\chi))\,\|s\|^2$$
$$\geq \left(\lambda^{+}_{\min}(\mathbf{J}(\chi))\right)^2\|s\|^2. \tag{18}$$

From inclusion (17) we infer that:

$$\min_{w\in\partial_\chi\left(V(\pi_\chi,\pi_\theta)+I_{\mathcal{X}_R}(\pi_\chi)\right)}\|w\| \geq \min_{v\in\mathbf{J}(\chi)\left(\nabla_\pi V(\pi_\chi,\pi_\theta)+\partial_\pi I_{\mathcal{X}_R}(\pi_\chi)\right)}\|v\|.$$

Theorem 38 provides the bound $\lambda^{+}_{\min}(\mathbf{J}(\chi)) \geq \frac{1}{1+(B-1)e^{2R}}$ and the conclusion is proven.

■

### E.3. Mahalanobis-pPŁ

**Lemma 44** *The utility of the game with softmax-parametrized policies satisfies the two-sided* Mahalanobis pPŁ *condition,*

$$\frac{\tau\min_h\mu_c(h)\gamma^3}{101\lambda_{\max}(\mathbf{M}^{-1})|\mathcal{H}|^3(1+(A-1)e^{2R})^2}\left[V_\tau(\chi,\theta)-V_\tau(\chi^\star,\theta)\right] \leq \frac{1}{2}\mathcal{D}_{X_R}(\chi,\ell;\theta)$$
$$\frac{\tau\min_h\mu_c(h)\gamma^3}{101\lambda_{\max}(\mathbf{M}^{-1})|\mathcal{H}|^3(1+(B-1)e^{2R})^2}\left[V_\tau(\chi,\theta^\star)-V_\tau(\chi,\theta)\right] \leq \frac{1}{2}\mathcal{D}_{\Theta_R}(\theta,\ell;\chi).$$

**Proof** We invoke (18) and the fact that $\|w\|^2_{\mathbf{M}^{-1}} \geq \lambda^{+}_{\min}\left(\mathbf{M}^{-1}\right)\|w\|^2$ for any $\langle w,v\rangle = 0,\ \forall v\in\ker\left(\mathbf{M}^{-1}\right)$. Also, we use Equation ($\varepsilon$-trunc.) and Assumption 2 to bound $\lambda^{+}_{\min}(\mathbf{M}^{-1})$. In detail, we know that,

$$\frac{|\mathcal{H}|^3(1+(A-1)e^{2R})^2}{\tau\min_h\mu_c(h)\gamma^3}\min_{w\in\partial_\chi\left(V(\pi_\chi,\pi_\theta)+I_{\mathcal{X}_R}(\pi_\chi)\right)}\|w\|^2 \geq V(\chi,\theta)-V(\chi^\star,\theta).$$

When $\mathbf{M} := \mathbf{F}(\chi,\theta)$, it is true that $\frac{\gamma^2\min_h\mu_c(h)}{|\mathcal{H}|^2}\varepsilon \leq \lambda_{\max}\left(\mathbf{F}(\chi,\theta)\right) \leq 1$.

■

**The spectrum of the Fisher Information Matrix** With the same arguments used in Theorem 35, we can conclude that,

- $\lambda_{\min}(\mathbf{F}(\chi,\theta)) = 0$;
- $\lambda^{+}_{\min}(\mathbf{F}(\chi,\theta)_s) \geq d(s)\min_a\pi_\chi(a|s)$;
- $d^{\chi,\theta}(s)\min_{s,a}\pi_\chi(a|s) \leq \lambda_{\max}(\mathbf{F}(\chi,\theta)_s) \leq d^{\chi,\theta}(s)\max_a\pi_\chi(a|s)+1$.

Hence,

- $\lambda^{+}_{\min}\left(\mathbf{F}(\chi,\theta)\right) \geq \min_{s,a}d^{\chi,\theta}(s)\pi_\chi(a|s)$;
- $\frac{\gamma^2\min_h\mu_c(h)}{|\mathcal{H}|^2}\varepsilon \leq \lambda_{\max}\left(\mathbf{F}(\chi,\theta)\right) \leq 1$.

Moreover, $d^{\chi,\theta}(s) \geq \frac{\gamma^2\min_h\mu_c(h)}{|\mathcal{H}|^2}$ by Assumption 2.

### E.4. Weak Gradient Domination

We now conclude this section with a proof of the weak gradient domination condition.

**Lemma 45 (Utility Weak Gradient Domination)** *Let $\Gamma$ be an IIEFG satisfying satisfying Assumption 2. Then, it holds true that,*

$$V^{\pi_1,\pi_2} - \min_{\pi_1'} V^{\pi_1',\pi_2} \leq \frac{1}{2\alpha_x} \max_{\pi_1'} \left\langle \nabla_{\pi_1} V^{\pi_1,\pi_2}, \pi_1 - \pi_1' \right\rangle;$$
$$\max_{\pi_2'} V^{\pi_1,\pi_2'} - V^{\pi_1,\pi_2} \leq \frac{1}{2\alpha_y} \max_{\pi_2'} \left\langle \nabla_{\pi_2} V^{\pi_1,\pi_2}, \pi_2' - \pi_2 \right\rangle,$$

*for $\alpha_x = \frac{\gamma}{\sqrt{2}|\mathcal{H}|^{\frac{3}{2}}A}$ and $\alpha_y = \frac{\gamma}{\sqrt{2}|\mathcal{H}|^{\frac{3}{2}}B}$.*

**Proof** We use [15, Prop. 2] by using the fact that the diameter of the treeplex is at most $\sqrt{2|\mathcal{H}||A|}$ and the fact that the Lipschitz of $\mu_1^{\pi_1} \to \pi_1$ is $\frac{|\mathcal{H}|\sqrt{A}}{\gamma}$. Then, we use the fact that $\max_{\|y-x\|\leq 1, y\in\mathcal{X}} \langle \nabla f(x), x - y\rangle = \min_{v\in\partial_x(f+I_{\mathcal{X}}(x))} \|v\|$. ■

## Appendix F. Gradient Estimators

In this section, we demonstrate that the well-known stochastic gradient estimator, REINFORCE, can be used yield an unbiased estimate of bounded variance of the gradients of the non-regularized and regularized imperfect-information game.

### F.1. A Policy Gradient Theorem

We define a trajectory $\xi$ to be a sequence of consecutive history-action pairs, $\xi = \left(\left(h^{(1)}, a^{(1)}_{i(1)}\right), \left(h^{(2)}, a^{(2)}_{i(2)}\right), \ldots\right)$. The length of trajectory $\xi$ is noted as $K_\xi$ and it is bounded by the 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 a policy profile, $(\pi_1, \pi_2)$, is fixed, the probability of each trajectory $\xi \in \mathcal{K}$ taking place is the product of the probability of each consecutive action,

$$\mathbb{P}^{\pi_1,\pi_2}(\xi) := \prod_{k=1}^{K_\xi} \pi_{i(k)}\left(a^{(k)}_{i(k)} | h^{(k)}\right).$$

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

**Lemma 46** *Under the assumption of ($\varepsilon$-trunc.), it holds true that the gradient estimator (REINFORCE) is unbiased,*

$$\mathbb{E}_{\xi\sim\pi_1,\pi_2}\left[\widehat{\nabla}_x\right] = \nabla_x V(\pi_1,\pi_2), \quad \textit{and} \quad \mathbb{E}_{\xi\sim\pi_1,\pi_2}\left[\widehat{\nabla}_y\right] = \nabla_y V(\pi_1,\pi_2);$$

*and also, its variance is bounded:*

$$\mathbb{E}_{\xi\sim\pi_1,\pi_2}\left[\left\|\widehat{\nabla}_x - \nabla_x V(\pi_1,\pi_2)\right\|^2\right] \leq \frac{A^2_{\max} D(\mathcal{T})^2}{\varepsilon};$$

$$\mathbb{E}_{\xi\sim\pi_1,\pi_2}\left[\left\|\widehat{\nabla}_y - \nabla_y V(\pi_1,\pi_2)\right\|^2\right] \leq \frac{B_{\max}^2 D(\mathcal{T})^2}{\varepsilon}.$$

*where $A, B$ denote the maximum available number of action in any infoset for player 1 and 2 respectively.*

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

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

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

$$\begin{aligned}
\mathbb{E}_\xi\left[\left\|\widehat{\nabla}_x - \mathbb{E}\left[\widehat{\nabla}_x\right]\right\|^2\right] &\leq \mathbb{E}_\xi\left[\left\|\widehat{\nabla}_x\right\|^2\right] \\
&= \mathbb{E}_\xi\left[\left\|r_\xi \sum_{k=1}^{K_\xi} \nabla_x \log \pi_1\left(a^{(k)}|s^{(k)}\right)\right\|^2\right] \\
&\leq \mathbb{E}_\xi\left[\left\|\sum_{k=1}^{K_\xi} \nabla_x \log \pi_1\left(a^{(k)}|s^{(k)}\right)\right\|^2\right] \\
&\leq \mathbb{E}_\xi\left[K_\xi \sum_{k=1}^{K_\xi} \left\|\nabla_x \log \pi_1\left(a^{(k)}|s^{(k)}\right)\right\|^2\right] \\
&\leq D(\mathcal{T})\mathbb{E}_\xi\left[\sum_{k=1}^{K_\xi} \left\|\nabla_x \log \pi_1\left(a^{(k)}|s^{(k)}\right)\right\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
&= D(\mathcal{T})\mathbb{E}_\xi\left[\sum_{k=1}^{K_\xi}\sum_{s,a}\mathbb{1}\{s=s^{(k)},a=a^{(k)}\}\frac{1}{\pi_1^2(a|s^{(k)})}\right]\\
&= D(\mathcal{T})\mathbb{E}_\xi\left[\sum_{k=1}^{K_\xi}\sum_{s,a}\mathbb{1}\{s=s^{(k)}\}\frac{1}{\pi_1(a|s^{(k)})}\right]\\
&\le \frac{A}{\varepsilon}D(\mathcal{T})\mathbb{E}_\xi\left[\sum_{k=1}^{K_\xi}\sum_{s,a}\mathbb{1}\{s=s^{(k)}\}\right]\\
&= \frac{A}{\varepsilon}D(\mathcal{T})\sum_{\xi\in\mathcal{K}}\mathbb{P}^{\pi_1,\pi_1}(\xi)\sum_{k=1}^{K_\xi}\sum_{s,a}\mathbb{1}\{s=s^{(k)}\}\\
&\le \frac{A^2D(\mathcal{T})^2}{\varepsilon}.
\end{aligned}
$$

■

**Lemma 47** *The variance of* (REINFORCE) *for softmax-parametrized policies is bounded as* $\sigma_\theta^2,\sigma_\chi^2\le 2D(\mathcal{T})^2$.

**Proof** We see that $\nabla_\theta\log\pi_\theta(a|s)=e_{s,a}-\pi_\theta(\cdot|s)$. From then on, $\|\nabla_\theta\log\pi_\theta(a|s)\|\le\sqrt{2}$ with probability 1. Then, the proof follows arguments similar to the previous one. ■

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

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

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

**Proposition 48** *For a policy profile* $\pi_1,\pi_2$*, the bidilated regularizer,* $\mathcal{R}_1$ *can be alternatively defined as:*

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

**Proof**

$$
\mathcal{R}_1(\pi_1,\pi_2)=\sum_{s\in\mathcal{S}_1}\mu_1^{\pi_1}(\sigma(s))\left(\sum_{h\in s}\mu_c(h)\mu_2^{\pi_2}(\sigma(h))\right)\psi(\pi_1(s))
$$

$$
\begin{aligned}
&= \sum_{s\in\mathcal{S}_1} \mathbb{P}^{\pi_1,\pi_2}(s)\psi(\pi_1(s)) \\
&= \sum_{s\in\mathcal{S}_1} \mathbb{E}_\xi \left[\sum_k^{K_\xi} \mathbb{1}\{s = s^{(k)}\}\psi(\pi_1(s))\right] \\
&= \mathbb{E}_\xi \left[\sum_{s\in\mathcal{S}_1}\sum_k^{K_\xi} \mathbb{1}\{s = s^{(k)}\}\psi(\pi_1(s))\right] \\
&= \mathbb{E}_\xi \left[\sum_k^{K_\xi}\sum_{s\in\mathcal{S}_1} \mathbb{1}\{s = s^{(k)}\}\psi(\pi_1(s))\right] \\
&= \mathbb{E}_\xi \left[\sum_k^{K_\xi} \psi\big(\pi_1\big(s^{(k)}\big)\big)\right] \\
&= \sum_{\xi\in\mathcal{K}} \mathbb{P}^{\pi_1,\pi_2}(\xi) \left(\sum_k^{K_\xi} \psi\big(\pi_1\big(s^{(k)}\big)\big)\right).
\end{aligned}
$$

■

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

$$
\begin{aligned}
&\nabla_x \mathcal{R}_1(\pi_1,\pi_2) \\
&\quad= \nabla_x \sum_{\xi\in\mathcal{K}} \mathbb{P}^{\pi_1,\pi_2}(\xi) \left(\sum_k^{K_\xi} \psi\big(\pi_1\big(s^{(k)}\big)\big)\right) \\
&\quad= \sum_{\xi\in\mathcal{K}} (\nabla_x \mathbb{P}^{\pi_1,\pi_2}(\xi)) \left(\sum_k^{K_\xi} \psi\big(\pi_1\big(s^{(k)}\big)\big)\right) + \sum_{\xi\in\mathcal{K}} \mathbb{P}^{\pi_1,\pi_2}(\xi) \left(\nabla_x \sum_k^{K_\xi} \psi\big(\pi_1\big(s^{(k)}\big)\big)\right) \\
&\quad= \underbrace{\sum_{\xi\in\mathcal{K}} (\mathbb{P}^{\pi_1,\pi_2}(\xi)\nabla_x \log \mathbb{P}^{\pi_1,\pi_2}(\xi)) \left(\sum_k^{K_\xi} \psi\big(\pi_1\big(s^{(k)}\big)\big)\right)}_{\varpi_1} \\
&\quad+ \underbrace{\sum_{\xi\in\mathcal{K}} \mathbb{P}^{\pi_1,\pi_2}(\xi) \left(\sum_k^{K_\xi} \nabla_x \psi\big(\pi_1\big(s^{(k)}\big)\big)\right)}_{\varpi_2}
\end{aligned}
$$

For $\varpi_1$, let us denote $r_\xi = \sum_k^{K_\xi} \psi\big(\pi_1\big(s^{(k)}\big)\big)$,

$$
\begin{aligned}
\varpi_1 &= r_\xi \sum_{\xi\in\mathcal{K}} \mathbb{P}^{\pi_1,\pi_2}(\xi)\nabla_x \log \mathbb{P}^{\pi_1,\pi_2}(\xi) \\
&= \sum_{\xi\in\mathcal{K}} r_\xi\, \mathbb{P}_\xi\, \nabla_x \log \mathbb{P}_\xi
\end{aligned}
$$

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

For $\varpi_2$, we write,

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

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

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

For $\vartheta_1$, similar to Theorem 46, we see that

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

Whereas, for $\vartheta_2$,

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

Finally, we note that when Assumption 2 is followed, then (REINFORCE) is also an unbiased estimator of bounded variance (same bounds as previously) of the perturbed version of the game. The reasoning is the same (when a player is exploring the gradient of the probability of an action is zero) and as such we omit it.

## Appendix G. Convergence Analysis

### G.1. Direct Policy Parametrization

**Theorem 49** *With direct policy parametrization and the Euclidean bidilated regularizer, alternating policy-gradient algorithm attains a last-iterate $\epsilon$-Nash equilibrium in*

$$T = \frac{1}{\epsilon^{12}}\mathsf{poly}\left(\tfrac{1}{\gamma}, |\mathcal{H}|, A, B, 2^{D(\mathcal{T})}, \tfrac{1}{\min_h \mu_c(h)}, |\mathcal{S}_1|, |\mathcal{S}_2|\right) \text{ iterations},$$

*using batches of* $\mathsf{poly}\left(\tfrac{1}{\epsilon}, \tfrac{1}{\gamma}, |\mathcal{H}|, A, B, 2^{D(\mathcal{T})}, \tfrac{1}{\min_h \mu_c(h)}, |\mathcal{S}_1|, |\mathcal{S}_2|\right)$ *trajectory samples at each step.*

**Proof** The proof follows as an application of Theorem 27. In a central role lies Theorem 42, which provides a two-sided pPŁ condition for the regularized game under direct policy parametrization, while in a supportive one the smoothness lemmata of the value function and the Euclidean bidilated regularizer when the policy is directly parametrized.

First, we relate equilibria of the regularized, truncated, exploration-perturbed game to equilibria of the original game. An $\epsilon$-NE of the regularized game is an $\epsilon'$-NE of the unregularized game where

$$\epsilon' = O\Big(\epsilon + \tau S 2^{D(\mathcal{T})} + \varepsilon S \max\{A, B\} + \gamma\Big).$$

The term contains the optimization error $\epsilon$, the regularization error (controlled by $\tau$), the truncation error (controlled by $\varepsilon$ through the minimum action probability), and the exploration-induced error (controlled by $\gamma$). To make each contribution $O(\epsilon)$ we choose

- $\gamma = \Theta(\epsilon)$,
- $\tau = \Theta\left(\frac{\epsilon}{\max_{i\in\{1,2\}} |\mathcal{S}_i| 2^{D(\mathcal{T})}}\right)$,
- $\varepsilon = \Theta\left(\frac{\epsilon}{\max_{i\in\{1,2\}} |\mathcal{S}_i| \max\{A,B\}}\right)$.

We now instantiate Theorem 27. By Theorem 42 the utility of the regularized game satisfies the two-sided pPŁ condition with moduli

$$\mu_x, \mu_y \;=\; \Theta\left(\frac{\tau \min_{h\in\mathcal{H}} \mu_c(h)\, \gamma^3}{|\mathcal{H}|^3}\right).$$

Combining the smoothness of the value function with that of the Euclidean bidilated regularizer (Theorem 33 ) yields an overall smoothness constant

$$\ell \;=\; \Theta\big(2^{D(\mathcal{T})} D(\mathcal{T}) \sqrt{|\mathcal{S}|}\big),$$

up to polynomial factors in $|\mathcal{S}_1|, |\mathcal{S}_2|, A, B$. The stochastic gradients used by `Alt-RegPG` are given by the REINFORCE estimator together with the gradient estimators for the bidilated regularizer; by Theorem 46 and the analysis of Appendix F.1 they are unbiased and have bounded per-trajectory variance

$$\mathbb{E}\big\|\widehat{\nabla}_x^{(1)} - \nabla_x V\big\|^2 \le \frac{A^2 D(\mathcal{T})^2}{\varepsilon}, \qquad \mathbb{E}\big\|\widehat{\nabla}_y^{(1)} - \nabla_y V\big\|^2 \le \frac{B^2 D(\mathcal{T})^2}{\varepsilon}.$$

If each update averages a mini-batch of $M$ i.i.d. trajectories, $\widehat{\nabla}_x = \frac{1}{M}\sum_{m=1}^{M}\widehat{\nabla}_x^{(m)}$ and $\widehat{\nabla}_y = \frac{1}{M}\sum_{m=1}^{M}\widehat{\nabla}_y^{(m)}$, then the averaged estimators have variances

$$\mathrm{Var}(\widehat{\nabla}_x) \;\le\; \frac{\sigma_x^2}{M}, \qquad \mathrm{Var}(\widehat{\nabla}_y) \;\le\; \frac{\sigma_y^2}{M},$$

with per-trajectory bounds $\sigma_x^2 \le A^2 D(\mathcal{T})^2/\varepsilon$ and $\sigma_y^2 \le B^2 D(\mathcal{T})^2/\varepsilon$. Substituting these into Theorem 27, the stochastic error terms are controlled (up to absolute constants) by $\sigma_x^2/(M\mu_x)$ and $\ell\,\sigma_y^2/(M\mu_x\mu_y^2)$. Requiring each to be at most $\epsilon$ leads to the condition

$$M \;\ge\; \max\left\{\frac{\sigma_x^2}{\epsilon\mu_x}, \frac{\ell\,\sigma_y^2}{\epsilon\mu_x\mu_y^2}\right\} = \max\left\{\frac{A^2 D(\mathcal{T})^2}{\epsilon\,\varepsilon\,\mu_x}, \frac{\ell\,B^2 D(\mathcal{T})^2}{\epsilon\,\varepsilon\,\mu_x\mu_y^2}\right\}.$$

Using the explicit forms of $\mu_x, \mu_y$ from Theorem 42 and the per-trajectory variance bounds from Theorem 46, this can be summarized (up to game-dependent constants) as choosing

$$M = \Theta\left(\max\left\{\frac{1}{\epsilon\,\varepsilon\,\tau\,\gamma^3}, \frac{\ell}{\epsilon\,\varepsilon\,\tau^3\gamma^9}\right\}\right).$$

Writing $S := \max\{|\mathcal{S}_1|, |\mathcal{S}_2|\}$ and using the tunings $\gamma = \Theta(\epsilon)$, $\tau = \Theta\big(\epsilon/(S2^{D(\mathcal{T})})\big)$, and $\varepsilon = \Theta\big(\epsilon/(SA_{\max})\big)$ from above, together with

$$\ell = \Theta\big(2^{D(\mathcal{T})}D(\mathcal{T})\sqrt{S}\big), \qquad \mu_x = \mu_y = \Theta\Big(\frac{\tau\gamma^3 \min_h \mu_c(h)}{|\mathcal{H}|^3}\Big) = \Theta\Big(\frac{\min_h \mu_c(h)}{S2^{D(\mathcal{T})}|\mathcal{H}|^3}\,\epsilon^4\Big),$$

a direct substitution yields the explicit bounds

$$M \ge \Theta\left(\frac{2^{D(\mathcal{T})}\,D(\mathcal{T})^2\,S^2\,A^3\,|\mathcal{H}|^3}{\min_h \mu_c(h)\,\epsilon^6}\right),$$
$$M \ge \Theta\left(\frac{2^{4D(\mathcal{T})}\,D(\mathcal{T})^3\,S^{9/2}\,A\,B^2\,|\mathcal{H}|^9}{\min_h \mu_c(h)^3\,\epsilon^{14}}\right).$$

For small $\epsilon$ the second constraint dominates, so it is sufficient to choose

$$M = \Theta\left(\frac{2^{4D(\mathcal{T})}\,D(\mathcal{T})^3\,S^{9/2}\,A\,B^2\,|\mathcal{H}|^9}{\min_h \mu_c(h)^3\,\epsilon^{14}}\right),$$

which spells out the precise dependence of the mini-batch size on $\epsilon$, $A$, $B$, $D(\mathcal{T})$, $|\mathcal{S}_1|$, $|\mathcal{S}_2|$, $|\mathcal{H}|$, and $\min_h \mu_c(h)$.

Under these conditions, Theorem 27 prescribes the concrete stepsizes

$$\eta_y = \frac{1}{5\ell}, \qquad\qquad \eta_x = \frac{\mu_y^2}{960\,\ell^3} = \frac{\tau^2\gamma^6(\min_{h\in\mathcal{H}}\mu_c(h))^2}{960\cdot 101^2\,|\mathcal{H}|^6\,\ell^3},$$

owing to the symmetric pPŁ moduli $\mu_x = \mu_y$ from Theorem 42. The resulting duality-gap decay is $\exp\Big(-\frac{\mu_x\mu_y^2}{960\ell^3}T\Big)$, so driving the deterministic term below $\epsilon$ requires

$$T = \frac{960\,\ell^3}{\mu_x\mu_y^2}\log\frac{\Delta_f}{\epsilon} = \frac{960\cdot 101^3\,|\mathcal{H}|^9\,\ell^3}{\tau^3\gamma^9\,(\min_{h\in\mathcal{H}}\mu_c(h))^3}\log\frac{\Delta_f}{\epsilon},$$

where $\Delta_f$ is the payoff range appearing in Theorem 27. Substituting the smoothness estimate from Theorems 31 and 33,

$$\ell = \Theta\Big(2^{D(\mathcal{T})} D(\mathcal{T}) \max_{i\in\{1,2\}} \sqrt{|\mathcal{S}_i|}\Big),$$

yields the following dependencies on the game parameters:

- $\eta_y = \Theta\left(\frac{1}{2^{D(\mathcal{T})} D(\mathcal{T}) \max_i \sqrt{|\mathcal{S}_i|}}\right)$;
- $\eta_x = \Theta\left(\frac{\tau^2\gamma^6(\min_{h\in\mathcal{H}} \mu_c(h))^2}{|\mathcal{H}|^6\left(2^{D(\mathcal{T})} D(\mathcal{T}) \max_i \sqrt{|\mathcal{S}_i|}\right)^3}\right)$;
- $T = \Theta\left(\frac{2^{3D(\mathcal{T})} D(\mathcal{T})^3 \left(\max_i \sqrt{|\mathcal{S}_i|}\right)^3 |\mathcal{H}|^9}{\tau^3\gamma^9(\min_{h\in\mathcal{H}} \mu_c(h))^3} \log\frac{1}{\epsilon}\right)$.

Finally, substituting the choices of $\gamma, \tau, \varepsilon$ from above into the expression for $T$ shows that $T$ scales as $\frac{1}{\epsilon^{12}}$ times a polynomial in the remaining game parameters, as claimed in the statement of the theorem. ■

### G.2. Softmax Policy Parametrization

**Theorem 50** *Alternating policy-gradient algorithm with softmax policy parametrization and the entropic bidilated regularizer converges in expectation in the last-iterate to an $\epsilon$-Nash equilibrium after a number of iterations $T$ given by*

$$T = \frac{1}{\epsilon^{18}} \text{ poly}\left(|\mathcal{H}|,\, A,\, B,\, 2^{D(\mathcal{T})},\, \frac{1}{\min_{h\in\mathcal{H}} \mu_c(h)},\, |\mathcal{S}_1|,\, |\mathcal{S}_2|\right),$$

*using batches of* $\text{poly}\left(\frac{1}{\epsilon}, \frac{1}{\gamma}, |\mathcal{H}|, A, B, 2^{D(\mathcal{T})}, \frac{1}{\min_h \mu_c(h)}, |\mathcal{S}_1|, |\mathcal{S}_2|\right)$ *trajectory samples at each step.*

**Proof** The theorem follows as a corollary of Theorem 27. By Theorem 43, the regularized game under softmax parametrization satisfies the two-sided pPŁ condition with moduli

$$\mu_x = \Theta\left(\frac{\tau \min_{h\in\mathcal{H}} \mu_c(h)\, \gamma^3}{|\mathcal{H}|^3\big(1 + (A-1)e^{2R}\big)^2}\right), \qquad \mu_y = \Theta\left(\frac{\tau \min_{h\in\mathcal{H}} \mu_c(h)\, \gamma^3}{|\mathcal{H}|^3\big(1 + (B-1)e^{2R}\big)^2}\right),$$

up to absolute constants. An $\epsilon$-NE for the regularized game is also an $\epsilon'$-NE for the unregularized game where

$$\epsilon' = O\Big(\epsilon + \gamma + \tau S 2^{D(\mathcal{T})} \max\{\log A, \log B\} + \varepsilon S(\max\{A, B\} - 1)^2\Big).$$

Then, we need to tune:

- $\gamma = \Theta(\epsilon)$;
- $\tau = \Theta\left(\frac{\epsilon}{\max_{i\in\{1,2\}} |\mathcal{S}_i| 2^{D(\mathcal{T})} \max\{\log A, \log B\}}\right)$;
- $\varepsilon = \Theta\left(\frac{\epsilon}{\max_{i\in\{1,2\}} |\mathcal{S}_i| (\max\{A,B\}-1)^2}\right)$.

We recall the smoothness parameter of the softmax-parametrized regularized utility function is

$$\ell = \Theta\Big(2^{D(\mathcal{T})} D(\mathcal{T}) \sqrt{|\mathcal{H}|} \;+\; 128 \max_{i\in\{1,2\}} \{\max_{s} w_{i,s}\} \,(1+\log A_{\max}) \sqrt{S} D(\mathcal{T})\Big),$$

by combining the Lipschitz bounds on the utility and the weighted entropic bidilated regularizer (Theorem 34). Then, from Theorem 27 we tune,

$$\eta_y = \Theta\left(\frac{1}{\ell}\right), \qquad \eta_x = \Theta\left(\frac{\alpha_y^2}{\ell^3}\right), \qquad T = \Theta\left(\frac{\ell^3}{\alpha_x \alpha_y^2} \log\frac{1}{\varepsilon}\right),$$

where $\ell$ is the smoothness constant for the utility and $\alpha_x, \alpha_y$ are the softmax pPŁ moduli of the two players. Invoking Theorem 43 for player 2 yields

$$\alpha_y = \Theta\left(\frac{\tau \min_{h\in\mathcal{H}} \mu_c(h)\, \gamma^3}{|\mathcal{H}|^3 \big(1+(B-1)e^{2R}\big)^2}\right),$$

and therefore, prior to relating $R$ to the truncation level $\varepsilon$,

$$\eta_x = \Theta\left(\frac{\tau^2 (\min_{h\in\mathcal{H}} \mu_c(h))^2\, \gamma^6}{|\mathcal{H}|^6 \big(1+(B-1)e^{2R}\big)^4\, \ell^3}\right).$$

Finally, using the explicit relationship between $R$ and the minimum action probability (so that $\big(1+(B-1)e^{2R}\big)^4$ can be expressed as a polynomial in $1/\varepsilon$) and simplifying constants leads to the following convenient. And, subsequently,

- $\eta_y = \Theta\left(\frac{1}{2^{D(\mathcal{T})}(1+\log B)\sqrt{|\mathcal{S}_2| D(\mathcal{T})}}\right),$
- $\eta_x = \Theta\Big(\frac{(\min_h \mu_c(h))^2\, \varepsilon^{12}}{|\mathcal{H}|^6 \max_{i\in\{1,2\}} |\mathcal{S}_i|^{14+3/2}\, D(\mathcal{T})^{2+3/2} \max\{\log A, \log B\}^5\, (B-1)^8\, 2^{3D(\mathcal{T})}}\Big).$

Finally, plugging the explicit expressions for $\alpha_x, \alpha_y$ from above into the generic bound $T = \Theta\Big(\frac{\ell^3}{\alpha_x \alpha_y^2} \log\frac{1}{\varepsilon}\Big)$ yields the precise parameter dependence

$$T = \Theta\left(\frac{\ell^3\, |\mathcal{H}|^9 \left(1+(A-1)e^{2R}\right)^2 \left(1+(B-1)e^{2R}\right)^4}{\tau^3 \left(\min_{h\in\mathcal{H}} \mu_c(h)\right)^3 \gamma^9} \log\frac{1}{\varepsilon}\right).$$

Using the relationship between $R$ and the minimum action probability to upper-bound $\big(1+(A-1)e^{2R}\big)$ and $\big(1+(B-1)e^{2R}\big)$ by polynomials in $1/\varepsilon$ and then substituting the tunings of $\gamma, \tau, \varepsilon$ we obtain an explicit dependence on the game parameters. Writing $S := \max\{|\mathcal{S}_1|, |\mathcal{S}_2|\}$ and using the smoothness estimate $\ell_{\text{softmax}}$ together with the truncation relation $\varepsilon$, a straightforward calculation yields

$$T = \Theta\left(\frac{2^{3D(\mathcal{T})}\, D(\mathcal{T})^3\, |\mathcal{H}|^9 S^{21/2} \left(\max\{A,B\}\right)^{12} \max\{\log A, \log B\}^6}{\epsilon^{18} (\min_{h\in\mathcal{H}} \mu_c(h))^3}\right).$$

As in the direct-parametrization case, we now quantify the effect of stochastic gradients. For softmax-parametrized policies, Theorem 47 shows that the REINFORCE estimator (combined with

the estimator for the entropic bidilated regularizer) is unbiased and has bounded variance per-trajector with $\sigma_\chi^2, \sigma_\theta^2 \leq \Theta\left(D(\mathcal{T})^2 + \tau 2^{D(\mathcal{T})}\right) = O\left(D(\mathcal{T})^2\right)$. We will control the stochastic error using mini-batches.

Substituting these into Theorem 27 with the softmax pPŁ moduli from Theorem 43,

$$\mu_x = \Theta\left(\frac{\tau \min_{h\in\mathcal{H}} \mu_c(h)\,\gamma^3}{|\mathcal{H}|^3\big(1+(A-1)e^{2R}\big)^2}\right), \qquad \mu_y = \Theta\left(\frac{\tau \min_{h\in\mathcal{H}} \mu_c(h)\,\gamma^3}{|\mathcal{H}|^3\big(1+(B-1)e^{2R}\big)^2}\right),$$

the stochastic error terms are controlled by

$$\frac{\sigma_x^2}{M\mu_x} \leq \Theta\left(\frac{2D(\mathcal{T})^2\,|\mathcal{H}|^3\big(1+(A-1)e^{2R}\big)^2}{M\,\tau \min_{h\in\mathcal{H}} \mu_c(h)\,\gamma^3}\right),$$
$$\frac{\ell\,\sigma_y^2}{M\mu_x\mu_y^2} \leq \Theta\left(\frac{2D(\mathcal{T})^2\,\ell\,|\mathcal{H}|^9\left(1+(A-1)e^{2R}\right)^2\left(1+(B-1)e^{2R}\right)^4}{M\,\tau^3\left(\min_{h\in\mathcal{H}} \mu_c(h)\right)^3\gamma^9}\right).$$

Requiring each to be at most $\epsilon$ gives the condition

$$M = \max\left\{\begin{array}{c}\Theta\left(\frac{D(\mathcal{T})^2\,|\mathcal{H}|^3\left(1+(A-1)e^{2R}\right)^2}{\epsilon\,\tau \min_{h\in\mathcal{H}} \mu_c(h)\,\gamma^3}\right), \\ \Theta\left(\frac{D(\mathcal{T})^2\,\ell\,|\mathcal{H}|^9\left(1+(A-1)e^{2R}\right)^2\left(1+(B-1)e^{2R}\right)^4}{\epsilon\,\tau^3\left(\min_{h\in\mathcal{H}} \mu_c(h)\right)^3\gamma^9}\right)\end{array}\right\}.$$

The second term dominates for small $\epsilon$, so it suffices to enforce

$$M \geq \Theta\left(\frac{D(\mathcal{T})^2\,\ell\,|\mathcal{H}|^9\left(1+(A-1)e^{2R}\right)^2\left(1+(B-1)e^{2R}\right)^4}{\epsilon\,\tau^3\left(\min_{h\in\mathcal{H}} \mu_c(h)\right)^3\gamma^9}\right).$$

To relate the dependence on $R$ to the truncation level, we use Theorem 38, which implies that if the minimum action probability under the softmax parametrization is at least $\varepsilon$, then $1+(A-1)e^{2R} \leq \frac{1}{\varepsilon}$, and $1+(B-1)e^{2R} \leq \frac{1}{\varepsilon}$, so

$$\left(1+(A-1)e^{2R}\right)^2\left(1+(B-1)e^{2R}\right)^4 \leq \frac{1}{\varepsilon^6}.$$

Combining this with $\ell$-smoothness from above yields the bound

$$M \geq \Theta\left(\frac{D(\mathcal{T})^2\,\ell\,|\mathcal{H}|^9}{\epsilon\,\tau^3\left(\min_{h\in\mathcal{H}} \mu_c(h)\right)^3\gamma^9\,\varepsilon^6}\right).$$

Finally, we denote $S := \max\{|\mathcal{S}_1|, |\mathcal{S}_2|\}$ and substitute the terms $\gamma, \tau, \varepsilon$, together with the definition of $\ell$, a direct calculation shows that it is sufficient to choose

$$M = \Theta\left(\frac{1}{\epsilon^{19}}\,\frac{2^{4D(\mathcal{T})}\,D(\mathcal{T})^3\,|\mathcal{H}|^9 S^{19/2}(\max\{A,B\})^1 2(\max\{\log A, \log B\})^4}{\left(\min_{h\in\mathcal{H}} \mu_c(h)\right)^3}\right).$$

■

## 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 $\mathbf{F}(\chi)$ is a blog 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 Theorem 35, we can conclude that,

- $\lambda_{\min}(\mathbf{F}(\chi,\theta)) = 0$;
- $\lambda^+_{\min}(\mathbf{F}_s(\chi,\theta)) \geq d^{\chi,\theta}(s)\min_a \pi_\chi(a|s)$;
- $\frac{\gamma^2 \min_h \mu_c(h)}{|\mathcal{H}|^2}\varepsilon \leq \lambda_{\max}\left(\mathbf{F}_s(\chi,\theta)\right) \leq 1$.

Hence,

- $\lambda^+_{\min}\left(\mathbf{F}(\chi,\theta)\right) \geq \min_{s,a} d^{\chi,\theta}(s)\pi_\chi(a|s)$;
- $\min_s \frac{1}{\sqrt{|\mathcal{H}||\mathcal{A}_s|}} \leq \lambda_{\max}\left(\mathbf{F}(\chi,\theta)\right) \leq 1$.

While, $d^{\chi,\theta}(s) \geq \frac{\gamma^2 \min_h \mu_c(h)}{|\mathcal{H}|^2}$ by Assumption 2.

**Theorem 51** *Alternating natural policy-gradient algorithm with softmax policy parametrization and the entropic bidilated regularizer converges in expectation in the last-iterate to an $\epsilon$-Nash equilibrium after a number of iterations $T$, that is*

$$T = \frac{1}{\epsilon^{36}}\,\text{poly}\left(\tfrac{1}{\gamma},\, |\mathcal{H}|,\, A,\, B,\, 2^{D(\mathcal{T})},\, \tfrac{1}{\min_{h\in\mathcal{H}}\mu_c(h)},\, |\mathcal{S}_1|,\, |\mathcal{S}_2|\right),$$

.

**Proof** This theorem is again an application of Theorem 27, now in its Mahalanobis form. For natural policy gradient, the updates are mirror-descent steps with a Mahalanobis metric induced by the Fisher information matrices, so we run Alt-GDA with $\mathbf{M}_{x,t} = \mathbf{F}_\chi(\chi_t,\theta_t)$ and $\mathbf{M}_{y,t} = \mathbf{F}_\theta(\chi_t,\theta_t)$.

By Theorem 44, for a general positive-semidefinite metric matrix $\mathbf{M}$ the game satisfies a two-sided Mahalanobis pPŁ condition with moduli

$$\tilde{\alpha}_x = \Theta\left(\frac{\tau \min_{h\in\mathcal{H}} \mu_c(h)\,\gamma^3}{\lambda_{\max}(\mathbf{M}^{-1})\,|\mathcal{H}|^3\big(1+(A-1)e^{2R}\big)^2}\right),$$
$$\tilde{\alpha}_y = \Theta\left(\frac{\tau \min_{h\in\mathcal{H}} \mu_c(h)\,\gamma^3}{\lambda_{\max}(\mathbf{M}^{-1})\,|\mathcal{H}|^3\big(1+(B-1)e^{2R}\big)^2}\right).$$

When we specialize $\mathbf{M}$ to the Fisher information matrices, the spectrum bounds in the previous subsection together with Assumption 2 and the truncation assumption imply

$$\lambda^+_{\min}\big(\mathbf{F}_\chi(\chi,\theta)\big) \gtrsim \frac{\gamma^2 \min_{h\in\mathcal{H}} \mu_c(h)\,\varepsilon}{|\mathcal{H}|^2}, \qquad \lambda^+_{\min}\big(\mathbf{F}_\theta(\chi,\theta)\big) \gtrsim \frac{\gamma^2 \min_{h\in\mathcal{H}} \mu_c(h)\,\varepsilon}{|\mathcal{H}|^2},$$

and hence, over the image of the Fisher matrices,

$$\lambda_{\max}\big(\mathbf{F}_\chi^{-1}(\chi,\theta)\big),\ \lambda_{\max}\big(\mathbf{F}_\theta^{-1}(\chi,\theta)\big) = O\left(\frac{|\mathcal{H}|^2}{\gamma^2 \min_{h\in\mathcal{H}} \mu_c(h)\,\varepsilon}\right).$$

Substituting these bounds for $\lambda_{\max}(\mathbf{M}^{-1})$ into the expressions above yields Mahalanobis pPŁ moduli

$$\tilde{\alpha}_x = \Theta\left(\frac{\tau(\min_{h\in\mathcal{H}} \mu_c(h))^2\,\gamma^5\,\varepsilon}{|\mathcal{H}|^5\big(1+(A-1)e^{2R}\big)^2}\right), \qquad \tilde{\alpha}_y = \Theta\left(\frac{\tau(\min_{h\in\mathcal{H}} \mu_c(h))^2\,\gamma^5\,\varepsilon}{|\mathcal{H}|^5\big(1+(B-1)e^{2R}\big)^2}\right).$$

The Mahalanobis version of Theorem 27 prescribes stepsizes (up to constants)

$$\eta_y = \Theta\left(\frac{1}{\ell'}\right), \qquad \eta_x = \Theta\left(\frac{\tilde{\alpha}_y^2}{\ell'^3 \lambda_{\max}^2}\right), \qquad T = \Theta\left(\frac{\ell'^3}{\tilde{\alpha}_x \tilde{\alpha}_y^2}\log\frac{1}{\varepsilon}\right),$$

where $\ell'$ is the smoothness constant of the objective under the Mahalanobis metric and $\lambda_{\max} := \max_t \lambda_{\max}(\mathbf{M}_{\cdot,t}^{-1})$. We use the same Euclidean smoothness constant $\ell_{\text{softmax}}$ as in the softmax-parametrized policy-gradient case. By the "Smoothness Relative to the Mahalanobis Distance" lemma,

$$\ell' = \frac{\ell_{\text{softmax}}}{\lambda^+_{\min}(\mathbf{M}_t)} \lesssim \frac{\ell_{\text{softmax}}|\mathcal{H}|^2}{\gamma^2 \min_{h\in\mathcal{H}} \mu_c(h)\,\varepsilon},$$

so overall $\ell'$ and $\lambda_{\max}$ contribute polynomial factors in $|\mathcal{H}|$, $1/\gamma$, $1/\min_h \mu_c(h)$, and $1/\varepsilon$.

As in the softmax-parametrized policy-gradient case, we relate equilibria of the truncated, regularized, exploration-perturbed game to equilibria of the original game. An $\epsilon$-NE of the perturbed game is an $\epsilon'$-NE of the unregularized game with

$$\epsilon' = O\Big(\epsilon + \gamma + \tau \max_{i\in\{1,2\}} |\mathcal{S}_i| 2^{D(\mathcal{T})} \max\{\log A, \log B\} + \varepsilon \max_{i\in\{1,2\}} |\mathcal{S}_i|(\max\{A,B\}-1)^2\Big),$$

so, as before, we choose

- $\gamma = \Theta(\epsilon)$;

- $\tau = \Theta\left(\frac{\epsilon}{\max_{i\in\{1,2\}} |\mathcal{S}_i| 2^{D(\mathcal{T})} \max\{\log A, \log B\}}\right)$;
- $\varepsilon = \Theta\left(\frac{\epsilon}{\max_{i\in\{1,2\}} |\mathcal{S}_i| (\max\{A,B\}-1)^2}\right)$.

Combining these tunings with the expressions for $\tilde{\alpha}_x, \tilde{\alpha}_y$, the effective smoothness $\ell'$, and the generic iteration bound $T = \Theta\big(\ell'^3/(\tilde{\alpha}_x \tilde{\alpha}_y^2) \log(1/\varepsilon)\big)$ yields

$$T = \Theta\left(\frac{2^{6D(\mathcal{T})}, D(\mathcal{T})^3 |\mathcal{H}|^{21} S_{\max}^{33/2} (\max\{A,B\})^{24} (\max\{\log A, \log B\})^6}{\epsilon^{36} \big(\min_{h\in\mathcal{H}} \mu_c(h)\big)^9}\right),$$

■