Efficient Decentralized Learning of Generalized Quantal Response Equilibrium

Zehao Zhao[1] **Apurv Shukla**[2] **Rahul Jain**[1] **Vijay Subramanian**[3]

[1]Department of Electrical and Computer Engineering, University of Southern California, (zhaozeha, rahul.jain)@usc.edu
[2]Sustainability Center, University of Michigan, Dearborn, apurvshu@umich.edu
[3]EECS Department, University of Michigan, Ann Arbor, vsgsubram@umich.edu

Abstract

We study a solution concept for bounded rational agents in finite normal-form general-sum games called Generalized Quantal Response Equilibrium (GQRE) which generalizes Quantal Response Equilibrium (McKelvey and Palfrey, 1995). In our setup, each player can individually maximize a smooth, regularized expected utility of the mixed profiles used, reflecting both bounded rationality that subsumes stochastic choice and also individual choice of behaviors. After establishing existence under mild conditions, we present a computationally efficient no-regret decentralized learning algorithm via a smoothened version of the Frank–Wolfe algorithm. Our algorithm uses noisy gradient estimates via bandit-feedback from a simulation oracle that reports on repeated plays of the game. We analyze finite-time convergence properties of our algorithm under assumptions that ensure uniqueness of equilibrium, using a novel class of gap functions that generalize the Nash gap. We end by demonstrating the effectiveness of our method on a set of complex general-sum games such as high-rank two-player games, large action two-player games, and known examples of difficult multi-player games.

1 INTRODUCTION

AI agents are increasingly prevalent in many settings that involve strategic decision-making, such as in algorithmic trading or autonomous driving. Often, such systems include a variety of agents, some being rational, some sophisticated and advanced AI agents, some more limited algorithmic agents, and some being humans. A key question then is whether we can predict the outcomes of such interactions. Classical solution concepts in game theory often rely on the assumptions that all agents are fully rational, know the game parameters (or distributions), and always play best responses (or equilibrium strategies). While this assumption yields elegant theoretical constructs such as Nash Equilibrium (NE), there is plenty of evidence in behavioral game theory (see (Haile et al., 2008; Selfridge, 1989; Goeree et al., 2003)) that shows that this the assumption doesn't hold, particularly when players are faced with uncertainty, limited information, or cognitive constraints. As a result, alternative models have been developed to better align equilibrium predictions with observed behavior in strategic environments. Further, learning to play NE of general sum games is known to be hard in-general—impossible with deterministic dynamics for some games (Milionis et al., 2023), and also using just gradient feedback (Jordan, 1993; Hart and Mas-Colell, 2003) for others. This has been another motivation to study alternate behavioral models.

In this paper, we first introduce the Generalized Quantal Response Equilibrium (GQRE), a generalization of several existing notions of QRE (McKelvey and Palfrey, 1995). This is done by introducing a general agent-specific divergence function which can be a ϕ-divergence, Wasserstein distance or Renyi divergence, leading to different notions of QRE. Thereafter, each agent chooses its response to the strategies of other players by minimizing the cost of the strategy (e.g., using an agent-specific divergence function) while preserving some minimum (expected) utility. An interpretation of this generalization is that it allows for individual-specific payoff perturbations that correspond to different risk sensitivities. Importantly, these perturbations are assumed to be chosen by the agents on the basis of the type of bounded rationality they would like to exhibit. Empirical evidence supporting QRE and similar payoff perturbations can be found in (Goeree et al., 2003; Selfridge, 1989). Under (similar) minimal assumptions, we show existence of GQRE (Theorem 2.1) by showing equivalence to a perturbed game's NE. This leads to the characterization of GQRE via a variational inequality (Lemma 2.1), and a polynomial-time verification algorithm for GQRE and approximate GQRE (Lemma 2.2).

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:8182–8208.

Whereas the GQRE is subsumed in the generalized equilibrium concept proposed by (Allen and Rehbeck, 2021), our main reason to study the GQRE is to understand how agents can interact with others based on limited knowledge of the game including their payoffs, and also the behavioral choices made by other agents, but still reach some notion of equilibrium through repeated interactions. Hence, our main contribution is the analysis of a computationally efficient no-regret learning algorithm based on a smoothened Franke-Wolfe (FW) algorithm for computing the GQRE for general-sum games in the independent learning setting; the algorithm is an example of the family proposed in (Pena, 2019), where asymptotic optimality was shown with noiseless gradients. In particular, our main result establishes asymptotic and non-asymptotic performance guarantees for this algorithm (Theorem 3.1) when using a simulator oracle that provides noisy bandit-feedback to the agents similar to (Kocák et al., 2014; Heliou et al., 2017), which results in the noisy estimates having unbounded noise variance; we also note (Cui et al., 2022) that uses a regularized FW scheme with bandit feedback for a potential game. Our analysis establishes negative drift for a novel Lyapunov (gap) function; our gap function is not used by (Pena, 2019), and so it can be used to study the broader family of algorithms in the reference. The analysis of the associated variational inequality holds when the strategy space is a polytope and could be of independent interest, and it also holds for a slightly relaxed set of assumptions than in antecedent work (see Section 3).

We validate all our results on a suite of numerical experiments including zero-sum, rank-k and strongly-monotone games for strategy spaces as large as 1000 actions—see Section 4. For all our instances we show that our smoothened Frank-Wolfe based independent learning algorithm outperforms existing state-of-the-art in terms of iteration complexity for a given Nash-gap, policy verification and regret. It also exhibits more stable performance relative to other schemes. Thus, in matrix games, relative to NE, GQRE incorporates features of human decision-making and is more amenable to decentralized computation.

In summary, our contributions are four-fold: (i) we study a solution concept that generalizes several existing QRE notions and allows us to study different behavioral choices of agents; (ii) we establish existence under minimal conditions and present a computationally tractable decentralized algorithm that uses bandit-feedback of payoff gradients from a simulator oracle; (iii) we prove non-asymptotic performance guarantees for the learning algorithm; and (iv) we provide numerical evidence of the algorithm's efficacy relative to other algorithms in the literature.

Related Work. As mentioned earlier, whereas our equilibrium concept is subsumed in the concept studied in (Allen and Rehbeck, 2021), their complete information setting doesn't apply to cases where agents choose their own behaviors with few assumptions about their opponents as it needs information about the strategies of the opponents, and furthermore, they don't study computational aspects of their equilibrium. Another work that studies an equilibrium concept related to ours is (Melo, 2022), but the focus is on QRE and to find simple conditions that guarantee uniqueness of the equilibrium. From the computational perspective, closest to our work is the work by (Gui and Taşkesen, 2025)—they propose related solution concepts but only ones in which the utilities are perturbed randomly via methods characterized by (Natarajan et al., 2009); the learning algorithm presented is exactly the dual averaging scheme of (Mertikopoulos and Zhou, 2019). The work in (Mazumdar et al., 2024) introduces risk-averse quantal response wherein perturbations can depend on both players; their equilibrium concept is also subsumed by (Allen and Rehbeck, 2021). Under an assumption of joint convexity of the risk-measure and with an additional QR related perturbation, (Mazumdar et al., 2024) show that the coarse-correlated equilibrium (which are computable through no-regret schemes like follow-the-perturbed-leader) of a different game yields risk-averse quantal equilibrium; however, they don't discuss or analyze specific learning schemes. In contrast, we propose an efficient no-regret learning algorithm for independent learning and establish non-asymptotic (Theorem 3.1) guarantees on its performance. From a computational regret characterization viewpoint, also related are the dual-averaging (Mertikopoulos and Zhou, 2019; Hsieh et al., 2021), optimistic gradient (Golowich et al., 2020), mirror-prox Nemirovski et al. (2009); Juditsky et al. (2011) and proximal gradient Jordan et al. (2024) algorithms. We compare the performance of our algorithm against these in Section 4 and demonstrate its performance advantages.

2 GENERALIZED QUANTAL RESPONSE EQUILIBRIUM

Notation. Consider a finite normal form game $\mathcal{G} \triangleq \{N, A, \{u_i\}_{i\in N}\}$—$N$ is the set of players, each player $i \in N$ has a finite strategy set A_i and a base utility matrix [1] $u_i : \prod_{j\in[N]} A_j \mapsto [0,1]$. We will denote the mixed strategy profile $(\pi_1, \ldots, \pi_N)$ with $\pi_i \in \Delta(A_i)$ for all $i \in [N]$ as $\boldsymbol{\pi}$ and also as (π_i, π_{-i}) where $-i$ stands for $[N] \setminus \{i\}$. For agent $i \in [N]$, when convenient, we take $a_i \in A_i$ to denote a pure action and also probability distribution δ_{a_i} on A_i with mass only on a_i.

Generalized quantal choice. Random utility models in stochastic choice theory (Anderson et al., 1992; Hofbauer and Sandholm, 2002) study the random payoff model: agent i selecting action a_i obtains a random additional payoff ε_i,

[1] For a given set of utility matrices across all agents, without loss of generality, we can subtract the minimum utility entry, and then divide by the largest utility to reduce values to $[0,1]$ to get a strategically equivalent game.

where $\varepsilon_i \sim f_i$ (a strictly positive density function independent of the base payoffs). Upon realization of the payoffs, the agent chooses the action with the highest realized payoff, and therefore, the probability of selecting action $a_i \in A_i$ with other players choosing a_{-i} is:

$$\pi_i(a_i) = p_i(a_i; u, a_{-i}) \triangleq \mathbb{P}\left(a_i \in \arg\max_{\tilde{a}_i \in A_i} u_i(\tilde{a}_i, a_{-i}) + \varepsilon_i\right).$$

In our work, we will, however, pivot to the perspective originally proposed by (Hofbauer and Sandholm, 2002)—the goal of agent i is to select a mixed strategy $\pi_i \in \Delta(A_i)$ with an expected payoff $\sum_{a_i \in A_i} \pi_i(a_i)\mathbb{E}_{\pi_{-i}}[u(a_i; A_{-i})]$ when the opponent(s) plays mixed strategy π_{-i}, but must pay a cost $f_i(\pi_i)/\lambda_i$. Therefore, the agent's choice given other players strategy π_{-i} arises from the Fenchel conjugate computation for f_i, namely,

$$\xi_i^{\lambda_i}(\pi_{-i}; f_i) \triangleq \arg\max_{\pi_i \in \Delta(A_i)} \lambda u_i(\pi_i, \pi_{-i}) - f_i(\pi_i), \quad (1)$$

where $u_i(\pi_i, \pi_{-i})$ is shorthand for the multi-linear function $\mathbb{E}_{\prod_{i\in[N]}\pi_i}[u_i(A_1, \ldots, A_N)]$; linearity in π_i results in the Fenchel conjugate definition.

Before considering generalized response functions, we illustrate the logit Quantal Response (McKelvey and Palfrey, 1995). The standard logit QR given by:

$$\xi_i(\pi_{-i}) \triangleq \left(\frac{\exp\left(\lambda_i \sum_{a_{-i}\in A_{-i}} \pi_{-i}(a_{-i})u_i(a_i, a'_{-i})\right)}{\sum_{a'_i \in A_i} \exp\left(\lambda_i \sum_{a_{-i}\in A_{-i}} \pi_{-i}(a'_{-i})u_i(a_i, a'_{-i})\right)}\right)_{a_i \in A_i} \quad (2)$$

can be obtained by solving (1) with the specific strictly convex function $f_i(\pi_i) = \sum_{a_i \in A_i} \pi_i(a_i)\log(\pi_i(a_i))$.

Via a Lagrangian relaxation, we note that (1) can also be interpreted as

$$\xi_i^{\lambda}(u, \pi_{-i}) = \arg\min_{\pi_i \in \Delta(A_i)} f_i(\pi_i) \quad \text{s.t. } \mathbb{E}_{\pi_i \times \prod_{j\in -i}\pi_j}[u_i(A_i, A_{-i})] \geq \gamma_i, \quad (3)$$

where there is an equivalence[2] between γ_i and λ_i in (1) since unconstrained choice of strategy corresponding to $\gamma_i = \min_{\pi_i \in \Delta(A_i)} \mathbb{E}_{\pi_i \times \pi_{-i}}[u_i(A_i, A_{-i})]$ and $\lambda_i = 0$, and utility maximization corresponding to $\gamma_i = \max_{\pi_i \in \Delta(A_i)} \mathbb{E}_{\pi_i \times \pi_{-i}}[u_i(A_i, A_{-i})]$ and $\lambda_i \uparrow +\infty$. We interpret the above as follows: agent i chooses a mixed strategy by minimizing an agent-specific cost function $f_i(\cdot)$ subject to the (mixed) strategies preserving some expected utility based on a value given by γ_i (equivalently, λ_i) that they choose, leading to a generalized quantal response.

[2]For all $\gamma_i \in \left[\min_{\pi_i \in \Delta(A_i)} \mathbb{E}_{\pi_i \times \prod_{j\in -i}\pi_j}[u_i(A_i, A_{-i})], \mathbb{E}_{\pi_i(f_i) \times \prod_{j\in -i}\pi_j}[u_i(A_i, A_{-i})]\right]$, we have $\lambda_i = 0$ due to the form in (1) using the distribution $\pi_i(f_i)$ that minimizes f_i over $\Delta(A_i)$, but the discussion is still valid.

Definition 2.1 (Generalized Quantal Response (GQR))**.** *Given a normal form game $\mathcal{G}$ and a strictly convex function f_i, the generalized quantal response of agent i, $\xi^i(\pi_{-i}; f_i) \in \Delta(A_i)$, is given by:*

$$\xi_i^{\lambda_i}(\pi_{-i}; f_i) \triangleq \arg\max_{\pi_i \in \Delta(A_i)} \lambda_i u_i(\pi_i, \pi_{-i}) - f_i(\pi_i). \quad (4)$$

When the function f_i measures the divergence between a base distribution, $\pi_i(f_i)$ over actions (such as uniform over all actions), the generalized quantal choice $\xi^i()$ is the distribution that minimizes the penalty paid by the player for deviating from base behavior, i.e.,

$$\xi_i^{\lambda_i}(\pi_{-i}; f_i) \triangleq \arg\max_{\pi_i \in \Delta(A_i)} \lambda_i u_i(\pi_i, \pi_{-i}) - d_{f_i}(\pi_i, \pi_i(f_i)), \quad (5)$$

where $d_{f_i}(\cdot, \cdot)$ is the corresponding divergence measure. Depending on the divergence measure, this generalization can incorporate different choice models such as fair response (see Example 2.2) that transcends the original motivations of best-responding with choice smoothing.

Example 2.1 (Divergence Measures)**.** *Below we list a few divergence measures that could be used for the f_i functions in a GQR—see details in Appendix D. In all cases we will be considering discrete distributions, say P and Q, over set $\{1, 2, \ldots, N\}$; we will also be taking Q to be a reference distribution, say the uniform distribution $U_{\{1,2,\ldots,N\}}$. Then, we have:*

1. (ϕ-divergence): Consider a convex function ϕ such that $\phi(1) = 0$. The ϕ-divergence between two distributions P and Q such that P is absolutely continuous with respect to Q is defined as $\sum_{j=1}^{n} Q(j)\phi\left(\frac{P(j)}{Q(j)}\right)$ (where we set $0/0 = 1$). Note that the KL-divergence can be obtained with $\phi(x) = x\log x$, the total variation distance can be obtained using $\phi(x) = \frac{1}{2}|x-1|$, the α-divergences using $\phi(x) = \frac{x^\alpha - 1}{\alpha(\alpha-1)}$ for $\alpha \in (0,1) \cup (1, \infty)$ (with continuity used to define values of $\alpha \in \{0, 1\}$), and for alpha ≥ 1, and the χ^α-divergences using $\phi(x) = 0.5|x-1|^\alpha$.

2. (Wasserstein Distance): The Wasserstein distance of order p between two distribution P and Q is given by $W_p(P, Q) = \inf_{\gamma \in \Gamma(P,Q)}(\mathbb{E}_{(X,Y)\sim\gamma}[d(X, Y)^p])^{1/p}$, where Γ is the set of all joint distributions with marginals P and Q, and for all $x, y \in \{1, 2, \ldots, N\}$, $d(x, y) = |x - y|$.

3. (Rényi Divergence): When distribution P is absolutely continuous with respect to distribution Q, for $\alpha \in (0, 1) \cup (1, +\infty)$ define $d_\alpha(P, Q) \triangleq \frac{1}{1-\alpha}\log\sum_{i=1}^{N} P(i)^\alpha Q^{1-\alpha}(i)$. The value at $\alpha = 1$ is the KL-divergence (defined via continuity). Similarly, (via continuity), the value at $\alpha = 0$ is given by $\sum_{i=1}^{N} Q(i)1_{\{P(i)>0\}}$, and at $\alpha = \infty$ is given by $\log(\sup_{i\in\{1,2,\ldots,N\}} P(i)/Q(i)$. Only for $\alpha \in [0, 1]$ does convexity in P hold.

4. (Hellinger Distance): The Hellinger distance between two distribution P and Q is given by $1/\sqrt{2}\sqrt{\sum_{i=1}^{N}(\sqrt{P(i)}-\sqrt{Q(i)})^2}$.

Convex program (4) yields useful properties for a GQR.

Proposition 2.1 (GQR Properties)**.** *Given a normal form game $\mathcal{G}$, for all $\lambda > 0$ we have: the GQR in (4) exists if f_i is convex, and is unique and continuous in π_{-i} if f_i is strictly convex*

Example 2.2. *Consider the GQR for a divergence measure from Example 2.1—see details in Appendix D.*
(Fair Quantal Response) Using the total variation distance—set $f(t) = \frac{1}{2}|t-1|$)—in (4) leads to a sparse response function where actions with low utility may get probability zero, in contrast to the KL-divergence regularization in (2). (Also, see (Gui and Taşkesen, 2025, Example 4.10).)

Remark 2.1. *We emphasize the following: (i) Classical models of fictitious play Fudenberg and Levine (1998); Hofbauer and Sandholm (2002) and games with penalized cost van Damme (1982), consider penalty functions which are infinitely steep at the boundary of the simplex with positive-definite Hessian. Here, the functions f_i need only be strictly convex (or even just convex for existence).*
(ii) The GQR in (4) is not a regular quantal response as per (Goeree et al., 2005) since our choice functions allow for zero mass over some actions (see Example 2.2), and thus, do not satisfy their interior criteria.

Let $\mathcal{P}(\mathcal{S})$ denote the power set given the set $\mathcal{S}$. Then, we have the following definition.

Definition 2.2 ($(\boldsymbol{\lambda}, \boldsymbol{f})$-Generalized Quantal Response Equilibrium (GQRE))**.** *Define the correspondence $\Gamma_{GQRE}^{(\boldsymbol{\lambda},\boldsymbol{f})} : \prod_{i\in[N]}\Delta(A_i) \mapsto \prod_{i\in[N]}\mathcal{P}(\Delta(A_i))$ as:*

$$\Gamma_{GQRE}^{(\boldsymbol{\lambda},\boldsymbol{f})}(\boldsymbol{\pi}) \triangleq \left\{(\tilde{\pi}_i,\ldots,\tilde{\pi}_N) : \forall \tilde{\pi}_i \in \xi_i^{\lambda_i}(\pi_{-i}; f_i)\ \forall i \in [N]\right\}.$$

Then, a GQRE is given by an N-tuple of policies $\boldsymbol{\pi}^* = (\pi_1^*,\ldots,\pi_N^*)$ that satisfy

$$\boldsymbol{\pi}^* \in \Gamma_{GQRE}^{(\boldsymbol{\lambda},\boldsymbol{f})}(\boldsymbol{\pi}^*),\ \textit{i.e.,}\ \pi_i^* \in \xi_i^{\lambda_i}(\pi_{-i}^*; f_i),\ \forall i \in [N].$$

We will also consider the notion of an approximate equilibrium concept, ϵ-GQRE.

Definition 2.3 ($(\boldsymbol{\lambda}, \boldsymbol{f})$ ϵ-GQRE)**.** *For $\epsilon \geq 0$, a strategy profile $\boldsymbol{\pi}^*$ is a $(\boldsymbol{\lambda}, \boldsymbol{f})$ ϵ-GQRE if*

$$\forall i \in [N],\ \lambda_i u_i(\pi_i^*, \pi_{-i}^*) - f_i(\pi_i^*) \geq$$
$$\lambda_i u_i\left(\xi_i^{\lambda_i}(\pi_{-i}^*; f_i), \pi_{-i}^*\right) - f_i\left(\xi_i^{\lambda_i}(\pi_{-i}^*; f_i)\right) - \epsilon,$$

i.e., the utility achieved is within ϵ of the best possible utility for every agent given the strategy profile.

Definition 2.2 defines a GQRE as a fixed point a'la Nash Equilibrium (NE). We now show that a GQRE for a game $\mathcal{G}$ is an NE for a perturbed game $\mathcal{G}_f$. Given game $\mathcal{G} = \{[N], \{A_i\}, \{u_i\}_{i\in N}\}$ and functions $\{f_i\}_{i\in[N]}$, for a mixed strategy profile $\boldsymbol{\pi} = (\pi_1, \pi_2)$, define the perturbed utilities as

$$u_i^{f_i}(\pi_i, \pi_{-i}) \triangleq \lambda_i u_i(\pi_i, \pi_{-i}) - f_i(\pi_i),$$

and the perturbed game $\mathcal{G}_f$ as $\mathcal{G}_f = \{[N], \{\Delta(A_i)\}_{i\in N}, \{u_i^{f_i}\}_{i\in[N]}\}$.

Assumption 2.1. *We assume the following for the perturbed utilities: (i) The gradients of the perturbed utilities are Lipschitz-continuous, and (ii) The gradients of the perturbed utilities have a Lipschitz-continuous Jacobian matrix H with each element defined $\forall i, j \in [N]$ as*

$$H_{(i,a_i),(j,a_j)}(\boldsymbol{\pi}) = \frac{\partial^2 u_i^{f_i}(\boldsymbol{\pi})}{\partial \pi_{i,a_i}\partial \pi_{j,a_j}},\ \forall a_i \in A_i, a_j \in A_j.$$

Assumption 2.2 ((Strict) Diagonal Dominance)**.** *The game $\mathcal{G}_f$ satisfies strict diagonal dominance if the matrix $H(\boldsymbol{\pi}) + H^T(\boldsymbol{\pi})$ is negative definite for all $\boldsymbol{\pi} \in \prod_{i\in[N]}\Delta(A_i)$.*

Since $u_i^{f_i}(\boldsymbol{\pi}) = \lambda_i u_i(\pi_i, \pi_{-i}) - f_i(\pi_i)$ for each $i \in [N]$, H has a block-form where the $H_{(i,\cdot),(j,\cdot)}$ block depends only on f_i, if $i = j$ (which, in fact, is the Hessian Δf_i) and u_i the utility matrix of agent i. We will use this important property in our analysis. With $\{f_i\}_{i\in[N]}$ being convex, we have a concave game, and by (Rosen, 1965) an NE always exists. Also, Assumption 2.2 is exactly the condition for a unique NE from (Rosen, 1965). We will also assume that the unique equilibrium is fully mixed. We summarize this next.

Theorem 2.1 (GQRE of $\mathcal{G}$ is NE of $\mathcal{G}_f$)**.** *A pair of policies (π_1^*, π_2^*) is the GQRE for $\mathcal{G}$ if and only if they are the Nash Equilibria for $\mathcal{G}_f$. Since $\mathcal{G}_f$ is a concave game when $\{f_i\}_{i\in[N]}$ are convex functions, a GQRE always exists. Further, by Assumption 2.2 of strict diagonal dominance, the GQRE is unique.*

Under a bounded influence condition, Melo (2022) studied Theorem 2.1 for QRE and developed simple conditions for uniqueness. Gui and Taşkesen (2025); Mertikopoulos and Zhou (2019) use (Rosen, 1965) as well; again, the equilibrium concept in Allen and Rehbeck (2021) also fits.

Remark 2.2. *In our definition of GQR and GQRE, we assumed that each agent i chose its (mixed) strategy from $\Delta(A_i)$. We can also assume that each agent i chooses its strategy from the ϵ_i-exploration simplex $\Delta(A_i, \epsilon_i)$ where the probability of choosing each action is at least ϵ_i for $\epsilon_i \in [0, \frac{1}{|A_i|}]$. The existence and uniqueness of an $\boldsymbol{\epsilon}$-exploration GQRE follow by the same logic as for a GQRE. We note that learning such an equilibrium will be covered naturally by our algorithm.*

An immediate consequence of Theorem 2.1 is a variational inequality (Facchinei and Pang, 2003) characterization for a GQRE as an alternative to the fixed-point characterization in Definition 2.2.

Lemma 2.1 (VI characterization)**.** *GQREs of the game $\mathcal{G}$ are solutions to variational inequality below: π^* such that*

$$\langle \pi - \pi_i^*, \nabla_\pi u_i^{f_i}(\boldsymbol{\pi}^*) \rangle \leq 0, \ \forall\, \pi \in \Delta(A_i), \forall i \in [N]. \quad (6)$$

Further, (6) is monotone since $u_i^{f_i}$ is concave in π_i, and strongly monotone if Assumption 2.2 holds.

The variational inequality characterization helps in designing efficient computational methods for computing the generalized quantal response equilibria by borrowing from the rich literature on optimization and variational inequalities. It also leads to a verification procedure for checking whether a given strategy profile is a GQRE or an ϵ-GQRE.

GQRE Verification. Having established a notion of GQRE, we propose an algorithmic verification mechanism to check whether a given strategy (π_i, π_{-i}) is a GQRE. For this, we use Theorem 2.1, i.e., the equivalence between GQRE of $\mathcal{G}$ and NE of $\mathcal{G}_f$. Checking whether a strategy is an NE for $\mathcal{G}_f$ is equivalent to checking if π_i is in the best-response set of player i. Since the utility functions are concave, this will be a convex optimization problem. Consequently, performing this check for all players can be done in polynomial-time.

Lemma 2.2 (Verification of GQRE and ϵ-GQRE)**.** *The following hold:*
1. A strategy profile $\boldsymbol{\pi}^*$ is a GQRE if and only if

$$\max_{a_i \in A_i} \langle \delta_{a_i} - \pi_i^*, \nabla_{\pi_i} u_i^{f_i}(\pi^*) \rangle \leq 0, \ \forall i \in [N].$$

This can be verified in $O(|N||A|)$ time.
2. A given strategy (π_i, π_{-i}) is an ϵ-GQRE for

$$\epsilon = \max_{i \in [N]} \epsilon_i, \ \textit{where}$$
$$\epsilon_i = u_i^{f_i}(\xi_i^{\lambda_i}(\pi_{-i}; f_i), \pi_{-i}) - u_i^{f_i}(\pi_i, \pi_{-i}), \ \forall i \in [N],$$

and determining the Nash gap ϵ requires solving N convex programs.

3 COMPUTING A GQRE

Having discussed the concept of a GQRE and its existence, we now discuss a means to compute it. From Lemma 2.1 it is clear that gradient schemes can be an option if the exact gradient or a noisy version of it is available. This has been a fruitful avenue in the optimization literature—using the exact gradient via the prox method as in Nemirovski (2004), or using a noisy gradient given by a martingale difference sequence (MDS) as in Nemirovski et al. (2009); Juditsky et al. (2011); Jordan et al. (2024); Mertikopoulos and Zhou (2019); Heliou et al. (2017); Golowich et al. (2020); Hsieh et al. (2021). Compared to the bulk of the literature that uses bounded conditional variance of the gradient estimates, we will use bandit feedback from a simulator oracle as in Heliou et al. (2017); Kocák et al. (2014), for which we neither have bounded variance nor independence across agents.

Gradient access via bandit feedback. Our simulator based oracle takes (mixed) strategies from all the agents, and plays the underlying finite game $M \geq 1$ times (independently if mixed strategies are used). The oracle then sends agent i a table listing M_i, the number of times action $a_i \in A_i$ was played, and the cumulative payoff $U_i(a_i)$ for the action over the M plays. Then, given strategy profile $\boldsymbol{\pi}$ and M plays $\{\boldsymbol{A}(m) = (A_1(m), \ldots, A_N(m))\}_{m=1}^M$ *i.i.d.* with distribution $\prod_{i \in [N]} \pi_i$, by taking expectations (given $\boldsymbol{\pi}$)

$$\begin{aligned} U_i(a_i) &= \sum_{m=1}^{M} u_i(\boldsymbol{A}(m)) 1_{\{A_i(m) = a_i\}}, \text{ and} \\ \mathbb{E}[U_i(a_i)] &= M \pi_i(a_i) u_i(a_i, \pi_{-i}). \end{aligned} \quad (7)$$

Hence, assuming that $\pi_i(a_i) > 0$, an unbiased estimate of the gradient of agent i's perturbed utility from just the finite game is given $\forall a_i \in A_i$ by

$$\begin{aligned} \hat{F}_i(a_i) &= \frac{\lambda_i}{\pi_i(a_i)} \frac{U(a_i)}{M} \text{ with} \\ \mathrm{Var}(\hat{F}_i(a_i)) &= \frac{\lambda_i^2 \widetilde{\mathrm{Var}}(a_i, \pi_i, \pi_{-i}; u_i)}{M \pi_i(a_i)}, \end{aligned} \quad (8)$$

where we have $\forall a_i \in A_i$,

$$\widetilde{\mathrm{Var}}(a_i, \pi_i, \pi_{-i}; u_i) \triangleq \sum_{\substack{\boldsymbol{a}_{-i} \in \\ \prod_{j \in -i} A_j}} \pi_{-i}(\boldsymbol{a}_{-i}) u_i^2(a_i, \boldsymbol{a}_{-i}) - \pi_i(a_i) u_i^2(a_i, \pi_{-i}). \quad (9)$$

Note that $\forall a_i \in A_i$, $\mathrm{Var}(\hat{F}_i(a_i))$ is unbounded over the range $\pi_i(a_i) \in (0, 1]$; this complicates algorithm design and analysis. Note that due to independent plays, the correlations of the gradient estimates across each agent i's actions are easily computed. Finally, the estimates being obtained from a common dataset are dependent across agents. Since each agent i knows its π_i and f_i, it can then form an unbiased estimate $F_i = \hat{F}_i - \nabla_{\pi_i} f_i(\pi_i)$ of $\nabla_{\pi_i} u_i^{f_i}(\boldsymbol{\pi})$. The variance and distributional characterizations still hold unchanged.

Algorithm Design. Another departure from existing work is in the choice of algorithmic schemes used by us to compute GQRE. Most existing work focuses on schemes such as dual-averaging, projected gradient descent, optimistic gradient descent, or extra-gradient methods to compute the solution to carefully chosen optimization problem(s). We instead focus on a projection-free method—the Frank-Wolfe algorithm. However, the Frank-Wolfe scheme is hard to analyze as the choice of directions suggested can be set-valued—assuming current utility gradient $\nabla_{\pi_i} u_i^{f_i}(\boldsymbol{\pi})$ for

agent $i \in [N]$, in the Frank-Wolfe scheme, the agent chooses probability vector s_i^* such that

$$s_i^* \in \Gamma_i^{\mathrm{FW}}(\boldsymbol{\pi}) \triangleq \arg\max_{s \in \Delta(A_i)} \langle s, \nabla_{\pi_i} u_i^{f_i}(\boldsymbol{\pi}) \rangle,$$

and then updates its strategy to $\pi_i + \gamma(s_i^* - \pi_i)$ with step-size $\gamma \in (0,1)$. Whereas the VI characterization shows that $\pi_i^* \in \Gamma^{\mathrm{FW}}(\boldsymbol{\pi}^*)$ for all $i \in [N]$ holds if and only if $\boldsymbol{\pi}^*$ is GQRE, and the performance of this algorithm is good (see Section 4.1), it is hard to analyze. The main reason for this is that the gap function [3] for this algorithm (a potential Lyapunov function) is Lipschitz but not everywhere differentiable. Due to this issue, we smoothen the Frank-Wolfe algorithm vis a "proximal" perturbation: $\forall i \in [N]$,

$$s_i^*(\boldsymbol{\pi}) \triangleq \arg\max_{s \in \Delta(A_i)} \langle s, \nabla_{\pi_i} u_i^{f_i}(\boldsymbol{\pi}) \rangle - \eta \mathrm{KL}(s, \pi_i),$$

$$\forall a_i \in A_i,\ s_{i,a_i}^*(\boldsymbol{\pi}) = \frac{\pi_{i,a_i} \exp\left(\frac{1}{\eta} \nabla_{\pi_i} u_{i,a_i}^{f_i}(\boldsymbol{\pi})\right)}{\left\langle \pi_i, \exp\left(\frac{1}{\eta} \nabla_{\pi_i} u_i^{f_i}(\boldsymbol{\pi})\right)\right\rangle}, \tag{10}$$

and thereafter, we define $\forall i \in [N]$,

$$\begin{aligned} V_i(\boldsymbol{\pi}) &\triangleq \max_{s \in \Delta(A_i)} \langle s - \pi_i, \nabla_{\pi_i} u_i^{f_i}(\boldsymbol{\pi}) \rangle - \eta \mathrm{KL}(s, \pi_i) \\ &= \eta \log\left(\left\langle \pi_i, \exp\left(\frac{1}{\eta} \nabla_{\pi_i} u_i^{f_i}(\boldsymbol{\pi})\right)\right\rangle\right) \\ &\quad - \eta \left\langle \pi_i, \frac{1}{\eta} \nabla_{\pi_i} u_i^{f_i}(\boldsymbol{\pi}) \right\rangle, \end{aligned} \tag{11}$$

$$\text{and set } V(\boldsymbol{\pi}) \triangleq \sum_{i \in [N]} \bar{V}_i(\boldsymbol{\pi}), \tag{12}$$

where we use the Donsker-Varadhan variational formula to solve for the optimizer and obtain the optimum value. Since choosing $\tilde{\pi} = \pi_i$ makes the objective in the optimization defining V_i to take value 0, the value of $V_i(\boldsymbol{\pi})$ is non-negative for each $\pi_i \in \Delta(A_i)$; this also holds using Jensen's inequality. Further, by the VI characterization for a GQRE $\boldsymbol{\pi}^*$, each $V_i(\boldsymbol{\pi}^*) = 0$ since at a non-GQRE $\boldsymbol{s}$,the linear term is non-positive by the VI, and the KL-divergence term is negative. Further, for a non-GQRE $\boldsymbol{\pi}$, starting from $\tilde{\pi} = \pi_i$, we can make a small enough perturbation in the right direction so that objective is positive; for this we use the fact that gradient of $\mathrm{KL}(\tilde{\pi}|\pi_i)$ is 0 at $\tilde{\pi} = \pi_i$. This proves that $\bar{V}_i(\boldsymbol{\pi})$ is a valid smooth gap function that we can use to study convergence to GQRE—see Appendix B for details. However, the variance of the gradient estimate being unbounded discussed earlier coupled with the unbounded gradient of the KL divergence in the second term, forces us to solve for the feasible direction within the ϵ-exploration simplex $\Delta(A_i;\epsilon)$ for $\epsilon \in (0, \frac{1}{|A_i|}]$, i.e., solving for

$$\forall i \in [N],\ s_i^*(\boldsymbol{\pi};\epsilon) \triangleq \arg\max_{s \in \Delta(A_i;\epsilon)} \frac{1}{2}\|s - s_i^*(\boldsymbol{\pi})\|_2^2. \tag{13}$$

[3] Gap function, $\sum_{i \in [N]} \max_{s \in \Delta(A_i)} \langle s - \pi_i, \nabla_{\pi_i} u_i^{f_i}(\boldsymbol{\pi}) \rangle$, is differentiable only if the maximizer is unique.

Using the unique $\lambda^* \in [\epsilon - \max_{a_i \in A_i} s_{i,a_i}^*(\boldsymbol{\pi}), 0]$ that solves (using a line search)

$$\sum_{a_i \in A_i} \max\left(s_{i,a_i}^*(\boldsymbol{\pi}) - \epsilon + \lambda, 0\right) = 1 - |A_i|\epsilon, \tag{14}$$

we set $\forall a_i \in A_i$,

$$\begin{aligned} s_{i,a_i}^*(\boldsymbol{\pi};\epsilon) &= \epsilon + \max\left(s_{i,a_i}^*(\boldsymbol{\pi}) - \epsilon + \lambda^*, 0\right) \\ &= \max(s_{i,a_i}^*(\boldsymbol{\pi}) + \lambda^*, \epsilon). \end{aligned} \tag{15}$$

Thus, the algorithm has all agents independently updating their strategy via their part of $\boldsymbol{\pi} + \gamma(\boldsymbol{s}^*(\boldsymbol{\pi};\epsilon) - \boldsymbol{\pi})$. This, then, leads to the independent and decentralized learning scheme outlined in Algorithm 1. As mentioned earlier, the basic algorithm (without adjusting for unbounded variance) is within the family of algorithms in (Pena, 2019), and method used to counter the unbounded variance is different from (Heliou et al., 2017) (convex combination with the uniform distribution) and (Kocák et al., 2014) (adding a constant to the denominator); also see (Cui et al., 2022) where a positive definite matrix is used for a Frank-Wolfe scheme. We note that our main result Theorem 3.1 also applies to the schemes used by (Kocák et al., 2014; Heliou et al., 2017), but from results in Section 4 we will also see that our choice provides a distinct advantage.

Algorithm 1 Smoothened Frank-Wolfe for decentralized learning with bandit feedback

Initialise: all players $i \in [N]$ choose a policy $\pi_{i,0}$ and parameter $\eta > 0$; choose sequences $\{(\gamma_t, \epsilon_t)\}_{t=1}^T$ such that, for all $t \in [T]$, $\gamma_t \in (0,1)$, $\epsilon_t \in (0, \frac{1}{N}]$, and $\epsilon_{t+1} \leq \epsilon_t$.
for $t = 1 : T$ **do**
  All players $i \in [N]$ submit their policy $\pi_{i,t-1}$ to the oracle
  Oracle plays the game M times with policy profile $\boldsymbol{\pi}_{t-1} = (\pi_{1,t-1}, \ldots, \pi_{N,t-1})$: $\{\boldsymbol{A}_t(m) = (A_{1,t}(m), \ldots, A_{N,t}(m))\}_{m=1}^M$ are *i.i.d.* with distribution $\prod_{i \in [N]} \pi_{i,t-1}$.
  Oracle returns outcomes: for $i \in [N]$ and all $a_i \in A_i$, $M_t(a_i) = \sum_{m=1}^M 1_{\{A_{i,t}(m) = a_i\}}$ and $U_t(a_i) = \sum_{m=1}^M u_i(\boldsymbol{A}_t(m)) 1_{\{A_{i,t}(m) = a_i\}}$.
  for $i \in [N]$ **do**
    Generate payoff gradient from game play: for all $a_i \in A_i$, $\hat{F}_{i,t}(a_i) = \frac{\lambda_i}{\pi_{i,t-1}(a_i)} \frac{U_t(a_i)}{M}$
    Generate noisy payoff gradient at $\boldsymbol{\pi}_{t-1}$: $F_{i,t} = \hat{F}_{i,t} - \nabla_\pi f_i(\pi)|_{\pi = \pi_{i,t-1}}$
    $\tilde{s}_{i,t} \leftarrow \arg\max_{s \in \Delta(A_i)} \langle s, F_{i,t} \rangle - \eta \mathrm{KL}(s|\pi_{i,t-1})$
    $s_{i,t} \leftarrow \arg\max_{s \in \Delta(A_i;\epsilon_t)} \frac{1}{2}\|s - \tilde{s}_{i,t}\|_2^2$
    $\pi_{i,t} \leftarrow \pi_{i,t-1} + \gamma_t (s_{i,t} - \pi_{i,t-1})$
  end for
end for

Smoothened Gap. Since we have a gap function $V(\boldsymbol{\pi})$ which is non-negative and 0 exactly for a GQRE, we assess the performance of Algorithm 1 using the expected gap value at the T^{th} timestep given by $R(T) = \mathbb{E}[V(\boldsymbol{\pi}_T)]$, where the randomness is induced by bandit feedback based gradient estimates.

Theorem 3.1 (No-regret guarantee)**.** *Under Assumptions 2.1 and 2.2, and at time-step t step-size $\gamma_t = \frac{1}{t+1}$, $\epsilon_t = \frac{1}{(t+1)\max_i |A_i|}$ and the number of samples M from the simulator chosen as $M_t = \lceil \frac{1}{\epsilon_t \gamma_t^2} \rceil$, the expected gap value of Algorithm 1 at time $T \geq 1$ satisfies:*

$$R(T) \leq \frac{C_7}{T} + \frac{C_6 \log(T)}{T},$$

and further, $\lim_{T\to\infty} V(\boldsymbol{\pi}(T)) = 0$ in probability.

Proof. The proof in Appendix A follows by showing that $V(\boldsymbol{\pi})$ is a valid Lyapunov function by proving that it has a negative drift. We carefully account for the noise due to the simulator, and here the choice of the number of samples is crucial. The fact that for all $i \in [N]$, $\pi_{i,t} \in \Delta(A_i, \epsilon_t)$ with $\epsilon_t > 0$, so that $\pi_{i,t}(a_i) \geq \epsilon_t > 0$ for $a_i \in A_i$, is also crucial along with $\epsilon_{t+1} \leq \epsilon_t$; see line 1 of Algorithm 1. The Fenchel conjugate representation in the algorithm and Fenchel duality are also critical. □

The proof ideas also extend the result to the schemes in (Kocák et al., 2014; Heliou et al., 2017)—see Appendix A.

4 NUMERICAL RESULTS

Games: We evaluate **Algorithm 1** on two families of two-player games: (1) *Strongly Monotone Games* of size $n \times n$ with $n \in \{10, 20, 100, 200\}$; and (2) *Rank-$k$ General-Sum Games* of size $m \times m$ with $m = 5$ and rank parameter $k \in \{1, 2, 3, 4\}$. All experiments use the same hyperparameters: $\mu = 1.0$, skew $= 0.3$ (used only in the strongly monotone setting), $T = 1000$ iterations, $M = 100$ samples per gradient estimate, step size $\eta = 1.0$, $\lambda_i = 1.0$, and $f_i(\pi_i) = \sum_{a_i \in A_i} \pi_{i,a_i} \log p_{i,a_i}$ for mixed strategy π_i; we solve for logit QRE; results for other penalty functions are presented in the Appendix. Each result is averaged over 20 independent runs. At each iteration, we draw M simulator samples to form stochastic gradient estimates and evaluate convergence by tracking the GQRE gap over time.

Strongly Monotone Games: For $n \in \{10, 20, 100, 200\}$, we generate $S = \mu I_n$, $M \sim \mathcal{N}(0,1)^{n\times n}$, $K = \frac{M - M^\top}{2}$, $K \leftarrow \frac{\text{skew}}{\|K\|_2} K$, and $A = S + K$, $B = S - K$, with $\mu = 1.0$, skew $= 0.3$. This construction ensures strong monotonicity by making the sum $A + B = 2\mu I_n$ positive definite. The symmetric component $S = \mu I_n$ guarantees strong convexity–concavity, while the antisymmetric matrix K introduces structured interaction between the players without violating monotonicity. The resulting game models have a unique and well-behaved NE.

Rank-k General-Sum Games: For $m \times m$ games with $m = 5$ and each $k \in \{1, 2, 3, 4\}$, we sample

$$U, V \sim \mathcal{N}(0,1)^{5\times k},\ M = U V^\top,$$
$$K \sim \mathcal{N}(0,1)^{5\times 5},\ S = \tfrac{1}{2}(K - K^\top),$$

and set $(A, B) = (M + S,\ M - S)$ so that $\text{rank}(A+B) = k$. Here, M is a low-rank matrix of rank at most k that controls the cooperative component of the game, while S introduces an antisymmetric term that governs the competitive interaction. By construction, the sum $A + B = 2M$ has rank k, ensuring that the underlying game lies on a low-dimensional manifold. These settings are useful for evaluating algorithm performance in structured general-sum environments with varying complexity Kannan and Theobald (2010).

The Baseline Algorithms: We consider the following state-of-the-art algorithms as baselines for comparison: (i) *Uniform-Mix Smoothened Frank-Wolfe:* Algorithm 1 with the (Heliou et al., 2017) unbounded variance combating scheme instead of Step 10; (ii) *Bandit Frank–Wolfe (Hard FW)* (Frank and Wolfe, 1956): replaces the smoothed best response with a hard $\arg\max$ on the stochastic gradient estimate within the Frank–Wolfe update, isolating the effect of smoothing; (iii) *Extragradient (EG)* (Chen et al., 2022): performs a Mirror–Prox update by taking a provisional gradient step followed by a corrective gradient step, yielding robust convergence in saddle-point problems; (iv) *Optimistic Gradient Descent (OGD)* (Jordan et al., 2024): uses the update $\log \pi \leftarrow \log \pi + \gamma_t(2G_t - G_{t-1})$ to add a "predictor–corrector" term that accelerates convergence under predictable gradient sequences; and (v) *Adaptive OGD–PGD (PGD)* (Jordan et al., 2024): applies an exponentiated gradient step with $\gamma_t = 1/\sqrt{t}$ and then projects onto an ε-simplex to adaptively balance exploration and exploitation under stochastic noise. We do not present dual-averaging schemes from (Mertikopoulos and Zhou, 2019; Hsieh et al., 2021) as their performance was significantly worse.

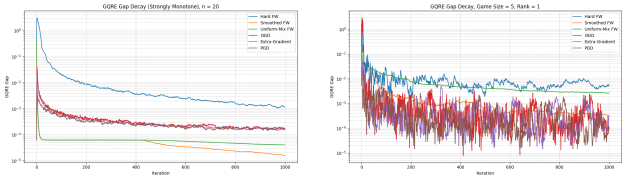
4.1 THE RESULTS

In Figure 1 we compare the five schemes in terms of the GQRE gap decay for two games from the game families we described above—the first, a strongly monotone game of size 20, and the second, a rank-1 game of size 5. Not only does Algorithm 1 perform better than existing schemes, but due to the projection step it also outperforms the Uniform-Mix Smoothened Frank-Wolfe algorithm.

Figure 2 shows GQRE gap for the strongly monotone games with game sizes of 100 and 200, Again, we see that Algorithm 1 outperforms all the other algorithms by a wide margin. Figure 3 shows GQRE gap for rank-k general-sum

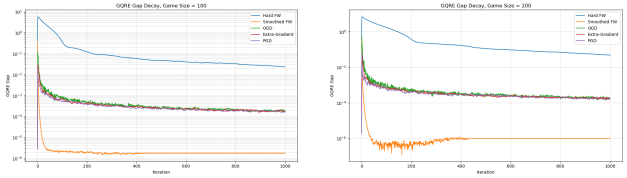
	action_0	action_1	action_2	action_3	action_4
Entropy–Entropy	0.183	0.184	0.149	0.302	0.182
Entropy–TV	0.189	0.192	0.171	0.262	0.186
L2–Entropy	0.169	0.171	0.130	0.363	0.167
L2–TV	0.179	0.185	0.155	0.307	0.174
Rényi–Rényi	0.129	0.130	0.129	0.482	0.129
TV–TV	0.200	0.200	0.200	0.200	0.200

Table 1: Representative Player 1 action distributions under different regularizer pairs (f_1, f_2) in the rank-2 game. Mixed regularizers (e.g., Entropy–TV, L2–Entropy, L2–TV) visibly shift mass allocation compared to homogeneous choices.



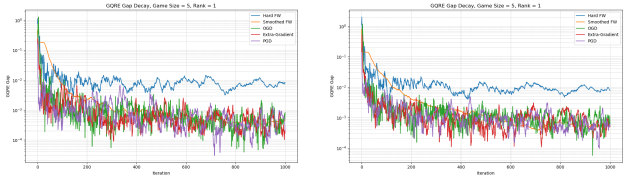
(a) Strongly monotone game ($n = 20$). (b) Rank-1 game of size 5.

Figure 1: GQRE gap decay in two classes of games.



(a) Strongly monotone game ($n = 100$). (b) Strongly monotone game ($n = 200$).

Figure 2: GQRE gap decay for strongly monotone games with sizes $n = 100, 200$.



(a) Rank-1 game. (b) Rank-3 game.

Figure 3: GQRE gap decay for rank-k general-sum games with size $m = 5$ and ranks $k = 1, 3$.

games for $k = 1, 3$. Here, we see that our algorithm performs similar to the other baselines, but is much more stable.

To understand the potential benefit of assigning different f_i functions to different agents, we ran simulations with heterogeneous regularizers with $\lambda = 1$ and Algorithm 1. Specifically, for the rank-k game of size 5, we considered all pairwise combinations of four divergences: Shannon entropy (KL), Rényi ($\alpha = 0.5$), $0.5\|\cdot\|_2^2$ relative to uniform, and the total variation (TV) distance to uniform.

Table 1 reports representative equilibrium distributions for Player 1 under each pair (f_1, f_2). Note that the choice of regularizer shapes the action probabilities: entropy and L2 spread mass more evenly across actions, Rényi accentuates a single action, and TV concentrates very close to uniform. More results are presented in Appendix C.

In Appendix C we present numerical results for more games—matching pennies, other strongly monotone games, and rank-k games (ranks ≤ 4). We also study Jordan's three-player two-action game from (Jordan, 1993), which has a unique NE, but which cannot be computed using just gradients (Hart and Mas-Colell, 2003); this issue shows up for Algorithm 1 via stability when Assumption 2.2 does not hold. Finally, impact of heterogeneous choice of f_is by agents is also studied.

Thus, the overall conclusion from this set of results is that the Smoothed FW algorithm we propose is quite effective at computing the GQRE equilibrium in decentralized setting wherein we have access to bandit-feedback of the gradients via a simulator oracle, and each agent adapts independently.

5 CONCLUSIONS

In this paper, we study the generalized quantal response equilibrium (GQRE), to predict outcomes of games with bounded-rational players learning independently, and provide sufficient conditions for existence of GQRE. We then introduced a decentralized Smoothened Frank–Wolfe algorithm—Algorithm 1—that can be used to compute the GQRE. Through an extensive set of numerical experiments, we showed that it is remarkably good at finding it.

References

Roy Allen and John Rehbeck. A generalization of quantal response equilibrium via perturbed utility. *Games*, 12(1): 20, 2021.

Simon P Anderson, Andre De Palma, and Jacques-Francois Thisse. *Discrete choice theory of product differentiation*. MIT press, 1992.

Ziyi Chen, Shaocong Ma, and Yi Zhou. Sample efficient stochastic policy extragradient algorithm for zero-sum Markov game. In *International Conference on Learning Representation*, 2022.

Qiwen Cui, Zhihan Xiong, Maryam Fazel, and Simon S Du. Learning in congestion games with bandit feedback. *Advances in Neural Information Processing Systems*, 35: 11009–11022, 2022.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Drew Fudenberg and David K Levine. *The theory of learning in games*, volume 2. MIT press, 1998.

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

Jacob K Goeree, Charles A Holt, and Thomas R Palfrey. Risk averse behavior in generalized matching pennies games. *Games and Economic Behavior*, 45(1):97–113, 2003.

Jacob K Goeree, Charles A Holt, and Thomas R Palfrey. Regular quantal response equilibrium. *Experimental economics*, 8:347–367, 2005.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. *Advances in Neural Information Processing Systems*, 33:20766–20778, 2020.

Yu Gui and Bahar Taşkesen. Statistical equilibrium of optimistic beliefs. *arXiv preprint arXiv:2502.09569*, 2025.

Philip A Haile, Ali Hortaçsu, and Grigory Kosenok. On the empirical content of quantal response equilibrium. *American Economic Review*, 98(1):180–200, 2008.

Sergiu Hart and Andreu Mas-Colell. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review*, 93(5):1830–1836, 2003.

Amélie Heliou, Johanne Cohen, and Panayotis Mertikopoulos. Learning with bandit feedback in potential games. *Advances in Neural Information Processing Systems*, 30, 2017.

Josef Hofbauer and William H Sandholm. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294, 2002.

Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos. Adaptive learning in continuous games: Optimal regret bounds and convergence to Nash equilibrium. In *Conference on Learning Theory*, pages 2388–2422. PMLR, 2021.

James S Jordan. Three problems in learning mixed-strategy Nash equilibria. *Games and Economic Behavior*, 5(3): 368–386, 1993.

Michael Jordan, Tianyi Lin, and Zhengyuan Zhou. Adaptive, doubly optimal no-regret learning in strongly monotone and exp-concave games with gradient feedback. *Operations Research*, 2024.

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Ravi Kannan and Thorsten Theobald. Games of fixed rank: A hierarchy of bimatrix games. *Economic Theory*, 42: 157–173, 2010.

Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. *Advances in Neural Information Processing Systems*, 27, 2014.

Eric Mazumdar, Kishan Panaganti, and Laixi Shi. Tractable equilibrium computation in Markov games through risk aversion. *arXiv preprint arXiv:2406.14156*, 2024.

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

Emerson Melo. On the uniqueness of quantal response equilibria and its application to network games. *Economic Theory*, 74(3):681–725, 2022.

Panayotis Mertikopoulos and Zhengyuan Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173:465–507, 2019.

Jason Milionis, Christos Papadimitriou, Georgios Piliouras, and Kelly Spendlove. An impossibility theorem in game dynamics. *Proceedings of the National Academy of Sciences*, 120(41):e2305349120, 2023.

Karthik Natarajan, Miao Song, and Chung-Piaw Teo. Persistency model and its applications in choice modeling. *Management Science*, 55(3):453–469, 2009.

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

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Javier Pena. Generalized conditional subgradient and generalized mirror descent: duality, convergence, and symmetry. *arXiv preprint arXiv:1903.00459*, 2019.

J Ben Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society*, pages 520–534, 1965.

Oliver G Selfridge. Adaptive strategies of learning a study of two-person zero-sum competition. In *Proceedings of the Sixth International Workshop on Machine Learning*, pages 412–415. Elsevier, 1989.

Eric Eleterius Coralie van Damme. *Refining the equilibrium concept for bimatrix games via control costs*. PhD thesis, Technische Hogeschool Eindhoven, 1982.

A PROOF OF THEOREM 3.1

We reiterate the statement of Theorem 3.1 below.

Theorem A.1 (No-regret guarantee). *Under Assumptions 2.1 and 2.2, and at time-step t step-size $\gamma_t = \frac{1}{t+1}$, $\epsilon_t = \frac{1}{(t+1)\max_i |A_i|}$ and the number of samples M from the simulator chosen as $M_t = \lceil \frac{1}{\epsilon_t \gamma_t^2} \rceil$, the expected gap value of Algorithm 1 at time $T \geq 1$ satisfies:*

$$R(T) \leq \frac{C_7}{T} + \frac{C_6 \log(T)}{T},$$

and further, $\lim_{T\to\infty} V(\boldsymbol{\pi}(T)) = 0$ in probability.

Reminder that

$$V_i(\boldsymbol{\pi}) = \eta \log\left(\langle \pi_i, \exp(F_i(\boldsymbol{\pi})/\eta)\rangle\right) - \langle \pi_i, F_i(\boldsymbol{\pi})\rangle, \text{and } V(\boldsymbol{\pi}) = \sum_{i\in[N]} V_i(\boldsymbol{\pi}).$$

We have for all $i \in [N]$,

$$\begin{aligned}
\nabla_{\pi_i} V_i(\boldsymbol{\pi}) &= \eta \frac{\exp(F_i(\boldsymbol{\pi})/\eta)}{\langle \pi_i, \exp(F_i(\boldsymbol{\pi})/\eta)\rangle} - F_i(\boldsymbol{\pi}) \\
&\quad + \nabla_{\pi_i} F_i(\boldsymbol{\pi})^T \left(s_i^*(\boldsymbol{\pi}, F_i(\boldsymbol{\pi})) - \pi_i\right) \\
\forall j \in -i,\ \nabla_{\pi_j} V_i(\boldsymbol{\pi}) &= \nabla_{\pi_j} F_i(\boldsymbol{\pi})^T \left(s_i^*(\boldsymbol{\pi}, F_i(\boldsymbol{\pi})) - \pi_i\right).
\end{aligned}$$

Our "RL" simulator provides noisy gradient $\boldsymbol{\delta}(\boldsymbol{\pi}) + F(\boldsymbol{\pi})$ (with $\boldsymbol{\delta}(\boldsymbol{\pi})$ being zero mean), which we use in the algorithm instead of $F(\boldsymbol{\pi})$. This gives $\boldsymbol{s}^*(\boldsymbol{\pi}, F(\boldsymbol{\pi}) + \boldsymbol{\delta}(\boldsymbol{\pi}))$ instead of $\boldsymbol{s}^*(\boldsymbol{\pi}, F(\boldsymbol{\pi}))$ in the noisy case. Further, the projection brings it to $\boldsymbol{s}_\epsilon^*(\boldsymbol{\pi}, F(\boldsymbol{\pi}) + \boldsymbol{\delta}(\boldsymbol{\pi}))$.

Based on Assumption 2.1, the bound we will use is

$$\begin{aligned}
V(\boldsymbol{\pi}(t+1)) - V(\boldsymbol{\pi}(t)) &\leq \langle \nabla_{\boldsymbol{\pi}} V(\boldsymbol{\pi}(t)), \boldsymbol{\pi}(t+1)) - \boldsymbol{\pi}(t))\rangle + \frac{C}{2}\|\boldsymbol{\pi}(t+1)) - \boldsymbol{\pi}(t)\|_2^2 \qquad (16) \\
&= \gamma_t \langle \nabla_{\boldsymbol{\pi}} V(\boldsymbol{\pi}(t)), \boldsymbol{s}_\epsilon^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t)\rangle \\
&\quad + \frac{C\gamma_t^2}{2}\|\boldsymbol{s}_\epsilon^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t) + \boldsymbol{\delta}(\boldsymbol{\pi}(t)))) - \boldsymbol{\pi}(t)\|_2^2, \qquad (17)
\end{aligned}$$

where the second representation uses our update rule. We prove this bound later on in Lemma A.1.

We will focus on the first term on the RHS of (17) after dividing by the γ_t term, namely,

$$\begin{aligned}
&\langle \nabla_{\boldsymbol{\pi}} V(\boldsymbol{\pi}(t)), \boldsymbol{s}_\epsilon^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t)\rangle \\
&= \sum_{i\in[N]} \langle \nabla_{\boldsymbol{\pi}} V_i(\boldsymbol{\pi}(t)), \boldsymbol{s}_\epsilon^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t)\rangle \\
&= \sum_{i\in[N]} \langle \nabla_{\boldsymbol{\pi}} V_i(\boldsymbol{\pi}(t)), \boldsymbol{s}_\epsilon^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t)))\rangle \\
&\qquad + \langle \nabla_{\boldsymbol{\pi}} V_i(\boldsymbol{\pi}(t)), \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t)\rangle.
\end{aligned}$$

The first term is small here as it is dependent on $\boldsymbol{s}_\epsilon^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t)))$, and this is a function of ϵ_t via our projection; this will also holds for the schemes of (Kocák et al., 2014; Heliou et al., 2017). We will include it in our statement of the result in the end, so we will focus on the second term, namely,

$$\begin{aligned}
&\sum_{i\in[N]} \langle \nabla_{\boldsymbol{\pi}} V_i(\boldsymbol{\pi}(t)), \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t)\rangle \\
&= \sum_{i\in[N]} \left\langle \eta \frac{\exp(F_i(\boldsymbol{\pi}(t))/\eta)}{\langle \pi_i(t), \exp(F_i(\boldsymbol{\pi}(t))/\eta)\rangle} - F_i(\boldsymbol{\pi}(t)), s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \pi_i(t) \right\rangle
\end{aligned}$$

$$
\begin{aligned}
&+ (\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t))^T H(\boldsymbol{\pi}(t))(\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t)) \\
&= \sum_{i\in[N]} \left\langle \eta \frac{\exp(F_i(\boldsymbol{\pi}(t))/\eta)}{\langle \pi_i(t), \exp(F_i(\boldsymbol{\pi}(t))/\eta)\rangle} - F_i(\boldsymbol{\pi}(t)), s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) \right\rangle \\
&\quad + (\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t))^T H(\boldsymbol{\pi}(t))(\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t)) \\
&\quad + \sum_{i\in[N]} \left\langle \eta \frac{\exp(F_i(\boldsymbol{\pi}(t))/\eta)}{\langle \pi_i(t), \exp(F_i(\boldsymbol{\pi})/\eta)\rangle} - F_i(\boldsymbol{\pi}), s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \pi_i(t) \right\rangle
\end{aligned} \tag{18}
$$

By Legendre-Fenchel duality, we have

$$
\begin{aligned}
V_i(\boldsymbol{\pi}) + \langle \pi_i, F_i(\boldsymbol{\pi})\rangle &= \max_{s\in\Delta(A_i)} \langle s, F_i(\boldsymbol{\pi})\rangle - \eta \mathrm{KL}(s|\pi_i) \\
&= \langle s_i^*(\boldsymbol{\pi}, F(\boldsymbol{\pi}), F_i(\boldsymbol{\pi})\rangle - \eta \mathrm{KL}(s_i^*(\boldsymbol{\pi}, F(\boldsymbol{\pi})|\pi_i).
\end{aligned}
$$

Thus, the third term in (18) yields

$$
\begin{aligned}
&\sum_{i\in[N]} \left\langle \eta \frac{\exp(F_i(\boldsymbol{\pi}(t))/\eta)}{\langle \pi_i(t), \exp(F_i(\boldsymbol{\pi}(t))/\eta)\rangle} - F_i(\boldsymbol{\pi}(t)), s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \pi_i(t) \right\rangle \\
&= \sum_{i\in[N]} \left\langle \eta \frac{\exp(F_i(\boldsymbol{\pi}(t))/\eta)}{\langle \pi_i(t), \exp(F_i(\boldsymbol{\pi}(t))/\eta)\rangle}, s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \pi_i(t) \right\rangle \\
&\quad - \sum_{i\in[N]} \langle F_i(\boldsymbol{\pi}), s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \pi_i(t)\rangle \\
&= \sum_{i\in[N]} \left\langle \eta \frac{\exp(F_i(\boldsymbol{\pi}(t))/\eta)}{\langle \pi_i(t), \exp(F_i(\boldsymbol{\pi}(t))/\eta)\rangle}, s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \pi_i(t) \right\rangle \\
&\quad - \sum_{i\in[N]} (V_i(\boldsymbol{\pi}(t)) + \eta \mathrm{KL}(s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)))|\pi_i(t))) \\
&= -V(\boldsymbol{\pi}(t)) + \sum_{i\in[N]} \left\langle \eta \frac{\exp(F_i(\boldsymbol{\pi}(t))/\eta)}{\langle \pi_i(t), \exp(F_i(\boldsymbol{\pi}(t))/\eta)\rangle}, s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \pi_i(t) \right\rangle \\
&\quad - \eta \sum_{i\in[N]} \mathrm{KL}(s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)))|\pi_i(t))
\end{aligned}
$$

Inserting the revised form of the third term into (18) yields

$$
\begin{aligned}
&\sum_{i\in[N]} \langle \nabla_{\boldsymbol{\pi}} V_i(\boldsymbol{\pi}(t)), \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t)\rangle \\
&= -V(\boldsymbol{\pi}(t)) - \sum_{i\in[N]} \langle F_i(\boldsymbol{\pi}(t)), s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)))\rangle \\
&\quad + (\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t))^T H(\boldsymbol{\pi}(t))(\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t)) \\
&\quad + \eta \sum_{i\in[N]} \left\langle \frac{\exp(F_i(\boldsymbol{\pi})/\eta)}{\langle \pi_i(t), \exp(F_i(\boldsymbol{\pi})/\eta)\rangle}, s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \pi_i(t) \right\rangle \\
&\quad - \eta \sum_{i\in[N]} \mathrm{KL}(s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)))|\pi_i(t))
\end{aligned} \tag{19}
$$

$$
\begin{aligned}
&\leq -V(\boldsymbol{\pi}(t)) - \sum_{i\in[N]} \langle F_i(\boldsymbol{\pi}(t)), s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)))\rangle \\
&\quad + (\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t))^T H(\boldsymbol{\pi}(t))(\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t)) \\
&\quad + \eta \sum_{i\in[N]} \Big(\mathrm{KL}(\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)))|\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t)))) \\
&\qquad\qquad - 2\mathrm{KL}(s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)))|\pi_i(t)) \Big)
\end{aligned} \tag{20}
$$

Remark A.1. *From the bound in (20) it is clear that the exact (or noiseless) gradient setting has negative drift, since $F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t)) = F(\boldsymbol{\pi}(t))$ in this setting. Hence, this algorithm is a good computational option when the payoff matrix is known. Assumption 2.2 is used to show that the quadratic forms are negative.*

The inequality in (20) is obtained the using the fact $\mathrm{KL}(p|q)$ for two distributions p and q is convex in q, and hence, satisfies the gradient inequality: for three distributions p, q and q_1, we have

$$\mathrm{KL}(p|q_1) \geq \mathrm{KL}(p|q) + \nabla_q \mathrm{KL}(p|q)^T (q_1 - q), \text{and } (\nabla_q \mathrm{KL}(p|q))_k = \frac{p_k}{q_k}, \ \forall k \in [K],$$

where we assume that the distributions are in a K-valued discrete space (and, sufficient for our purposes, both q and q_1 have full support). Then, we set the following:

$$p = \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))), \ q = \pi_i(t), \text{ and } q_1 = \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}_e(t))).$$

Using the expression for $\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)))$, this yields

$$\begin{aligned}
&\mathrm{KL}(s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)))|s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t)))) \\
&\geq \mathrm{KL}(s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)))|\pi_i(t)) + \left\langle \frac{\exp(F_i(\boldsymbol{\pi})/\eta)}{\langle \pi_i(t), \exp(F_i(\boldsymbol{\pi})/\eta)\rangle}, s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \pi_i(t) \right\rangle
\end{aligned}$$

Consider (19) where the RHS is given by

$$\begin{aligned}
\mathrm{RHS} &= -V(\boldsymbol{\pi}(t)) - \langle F(\boldsymbol{\pi}(t)), \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)))\rangle \\
&\quad + (\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t))^T H(\boldsymbol{\pi}(t))(\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t)) \\
&\quad + \eta \sum_{i\in[N]} \left\langle \frac{\exp(F_i(\boldsymbol{\pi}(t))/\eta)}{\langle \pi_i(t), \exp(F_i(\boldsymbol{\pi}(t))/\eta)\rangle}, s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \pi_i(t) \right\rangle \\
&\quad - \eta \sum_{i\in[N]} \mathrm{KL}(s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)))|\pi_i(t)) \\
&= -V(\boldsymbol{\pi}(t)) + (\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t))^T H(\boldsymbol{\pi}(t))(\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t)) \\
&\quad + \eta \sum_{i\in[N]} \left\langle \frac{\exp(F_i(\boldsymbol{\pi}(t))/\eta)}{\langle \pi_i(t), \exp(F_i(\boldsymbol{\pi}(t))/\eta)\rangle}, s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \pi_i(t) \right\rangle \\
&\quad - \eta \sum_{i\in[N]} \mathrm{KL}(s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)))|\pi_i(t)) \\
&\quad + (\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t))^T H(\boldsymbol{\pi}(t))(\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)))) \\
&\quad - \langle F(\boldsymbol{\pi}(t)), \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)))\rangle \\
&\quad + \eta \sum_{i\in[N]} \left\langle \frac{\exp(F_i(\boldsymbol{\pi})/\eta)}{\langle \pi_i(t), \exp(F_i(\boldsymbol{\pi})/\eta)\rangle}, s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) \right\rangle \\
&= G(\boldsymbol{\pi}(t)) + \langle L(\boldsymbol{\pi}(t)), \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)))\rangle
\end{aligned}$$

where

$$\begin{aligned}
G(\boldsymbol{\pi}(t)) &= -V(\boldsymbol{\pi}(t)) + (\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t))^T H(\boldsymbol{\pi}(t))(\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t)) \\
&\quad + \eta \sum_{i\in[N]} \left\langle \frac{\exp(F_i(\boldsymbol{\pi}(t))/\eta)}{\langle \pi_i(t), \exp(F_i(\boldsymbol{\pi}(t))/\eta)\rangle}, s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \pi_i(t) \right\rangle \\
&\quad - \eta \sum_{i\in[N]} \mathrm{KL}(s_i^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)))|\pi_i(t)) \\
&\leq -V(\boldsymbol{\pi}(t)),
\end{aligned}$$

but the other negative terms will be useful to drive other quantities negative—some of them arise from Assumption 2.2 regarding the quadratic forms. Defining vector $\boldsymbol{\lambda}(\boldsymbol{\pi})$ by setting the i^{th} block to be

$$\lambda_i(\boldsymbol{\pi}) = \eta \frac{\exp(F_i(\boldsymbol{\pi})/\eta)}{\langle \pi_i, \exp(F_i(\boldsymbol{\pi})/\eta)\rangle},$$

we have

$$L(\boldsymbol{\pi}(t)) = -F(\boldsymbol{\pi}(t)) + H^T(\boldsymbol{\pi}(t))(\boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) - \boldsymbol{\pi}(t)) + \boldsymbol{\lambda}(\boldsymbol{\pi}(t))$$

Now collecting everything we can replace (17) by

$$\begin{aligned}
&V(\boldsymbol{\pi}(t+1)) - V(\boldsymbol{\pi}(t)) \\
&\leq \gamma_t G(\boldsymbol{\pi}(t)) + \gamma_t \langle L(\boldsymbol{\pi}(t)), \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) \rangle \\
&\quad + \gamma_t \sum_{i \in [N]} \langle \nabla_{\boldsymbol{\pi}} V_i(\boldsymbol{\pi}(t)), s^*_{i,\epsilon}(\boldsymbol{\pi}(t), F_i(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - s^*_i(\boldsymbol{\pi}(t), F_i(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) \rangle \\
&\quad + \frac{C\gamma_t^2}{2} \|\boldsymbol{s}^*_\epsilon(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t) + \boldsymbol{\delta}(\boldsymbol{\pi}(t)))) - \boldsymbol{\pi}(t)\|_2^2 \\
&\leq \gamma_t G(\boldsymbol{\pi}(t)) + \gamma_t \langle L(\boldsymbol{\pi}(t)), s^*_i(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - s^*_i(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) \rangle \\
&\quad + 2\gamma_t \epsilon_t C_1 + C_2 \gamma_t^2
\end{aligned}$$

Hence, the expected drift is given by

$$\begin{aligned}
&\mathbb{E}[V(\boldsymbol{\pi}(t+1)) - V(\boldsymbol{\pi}(t)) | \boldsymbol{\pi}(t)] \\
&\leq \gamma_t G(\boldsymbol{\pi}(t)) + \gamma_t \mathbb{E}[\langle L(\boldsymbol{\pi}(t)), s^*_i(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - s^*_i(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) \rangle | \boldsymbol{\pi}(t)] \\
&\quad + 2\gamma_t \epsilon_t C_1 + C_2 \gamma_t^2.
\end{aligned}$$

Applying the Cauchy-Schwartz inequality and then bounding, we get

$$\begin{aligned}
&\mathbb{E}[V(\boldsymbol{\pi}(t+1)) - V(\boldsymbol{\pi}(t)) | \boldsymbol{\pi}(t)] \\
&\leq \gamma_t G(\boldsymbol{\pi}(t)) + \gamma_t \mathbb{E}[\langle L(\boldsymbol{\pi}(t)), \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - \boldsymbol{s}^*(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) \rangle | \boldsymbol{\pi}(t)] \\
&\quad + 2\gamma_t \epsilon_t C_1 + C_2 \gamma_t^2 \\
&= \gamma_t G(\boldsymbol{\pi}(t)) + \gamma_t \sum_{i \in [N]} \mathbb{E}[\langle L_i(\boldsymbol{\pi}(t)), s^*_i(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - s^*_i(\boldsymbol{\pi}(t), F(\boldsymbol{\pi}(t))) \rangle | \boldsymbol{\pi}(t)] \\
&\quad + 2\gamma_t \epsilon_t C_1 + C_2 \gamma_t^2 \\
&\leq \gamma_t G(\boldsymbol{\pi}(t)) + \gamma_t C_2 \eta \sum_{i \in [N]} \sqrt{\mathbb{E}[\|s^*_i(\boldsymbol{\pi}(t), F_i(\boldsymbol{\pi}(t)) + \boldsymbol{\delta}(\boldsymbol{\pi}(t))) - s^*_i(\boldsymbol{\pi}(t), F_i(\boldsymbol{\pi}(t)))\|_2^2]} \\
&\quad + 2\gamma_t \epsilon_t C_1 + C_2 \gamma_t^2 \\
&\leq \gamma_t G(\boldsymbol{\pi}(t)) + \gamma_t C_2 \sum_{i \in [N]} \sqrt{\mathbb{E}[\|\delta_i(\boldsymbol{\pi}(t))\|_2^2]} + 2\gamma_t \epsilon_t C_1 + C_2 \gamma_t^2 \\
&\leq \gamma_t G(\boldsymbol{\pi}(t)) + \gamma_t C_2 \sum_{i \in [N]} \sqrt{\sum_{a_i \in A_i} \frac{\lambda_i^2 \widetilde{\mathrm{Var}}(a_i, \pi_i, \pi_{-i}; u_i)}{M \pi_i(a_i)}} + 2\gamma_t \epsilon_t C_1 + C_2 \gamma_t^2 \\
&\leq \gamma_t G(\boldsymbol{\pi}(t)) + \gamma_t C_3 \frac{1}{\sqrt{M \epsilon_t}} + 2\gamma_t \epsilon_t C_1 + C_2 \gamma_t^2,
\end{aligned}$$

where we have used the fact that the softmax function (Gao and Pavel, 2017, Proposition 4) is $1/\eta$ Lipschitz in the l_2 norm with respect to the input vector—our function $s^*_i(\boldsymbol{\pi}(t), F_i(\boldsymbol{\pi}(t)))$ is a softmax function where the second component is the input vector, and we consider the difference between the values at $F_i(\boldsymbol{\pi}(t)) + \delta_i(\boldsymbol{\pi}(t)))$ and $F_i(\boldsymbol{\pi}(t))$. We have also used the variance characterization in (8), and we remind the reader that from (9) we have $\forall a_i \in A_i$,

$$\begin{aligned}
\widetilde{\mathrm{Var}}(a_i, \pi_i, \pi_{-i}; u_i) &= \sum_{\boldsymbol{a}_{-i} \in \prod_{j \in -i} A_j} \pi_{-i}(\boldsymbol{a}_{-i}) u_i^2(a_i, \boldsymbol{a}_{-i}) - \pi_i(a_i) u_i^2(a_i, \pi_{-i}) \\
&\leq \sum_{\boldsymbol{a}_{-i} \in \prod_{j \in -i} A_j} \pi_{-i}(\boldsymbol{a}_{-i}) u_i^2(a_i, \boldsymbol{a}_{-i}) \leq 1,
\end{aligned}$$

where we used our assumed upper bound of 1 on the utilities.

For completeness, we derive the variance characterization in (8) by showing the following:

$$\begin{aligned}
\mathbb{E}[U_i^2(a_i)] &= \mathbb{E}\left[\sum_{m,m'=1}^{M} u_i(\boldsymbol{A}(m))u_i(\boldsymbol{A}(m'))1_{\{A_i(m)=A_i(m')=a_i\}}\right] \\
&= \sum_{m=1}^{M} \mathbb{E}\left[u_i^2(\boldsymbol{A}(m))1_{\{A_i(m)=a_i\}}\right] \\
&\quad + \sum_{m,m'=1,m\neq m'}^{M} \mathbb{E}\left[u_i(\boldsymbol{A}(m))u_i(\boldsymbol{A}(m'))1_{\{A_i(m)=A_i(m)=a_i\}}\right] \\
&= M\pi_i(a_i) \sum_{\boldsymbol{a}_{-i}\in\prod_{j\in-i}A_j} \pi_{-i}(\boldsymbol{a}_{-i})u_i^2(a_i,\boldsymbol{a}_{-i}) + M(M-1)\pi_i^2(a_i)u_i^2(a_i,\pi_{-i}) \\
&= M\left(\pi_i(a_i) \sum_{\boldsymbol{a}_{-i}\in\prod_{j\in-i}A_j} \pi_{-i}(\boldsymbol{a}_{-i})u_i^2(a_i,\boldsymbol{a}_{-i}) - \pi_i^2(a_i)u_i^2(a_i,\pi_{-i})\right) + \mathbb{E}^2[U_i(a_i)] \\
\Rightarrow \mathrm{Var}(U_i(a_i)) &= M\left(\pi_i(a_i) \sum_{\boldsymbol{a}_{-i}\in\prod_{j\in-i}A_j} \pi_{-i}(\boldsymbol{a}_{-i})u_i^2(a_i,\boldsymbol{a}_{-i}) - \pi_i^2(a_i)u_i^2(a_i,\pi_{-i})\right) \\
&= M\pi_i(a_i)\left(\sum_{\boldsymbol{a}_{-i}\in\prod_{j\in-i}A_j} \pi_{-i}(\boldsymbol{a}_{-i})u_i^2(a_i,\boldsymbol{a}_{-i}) - \pi_i(a_i)u_i^2(a_i,\pi_{-i})\right).
\end{aligned}$$

If we increase the number of samples M per iteration as $M_t = \lceil \frac{1}{\epsilon_t\gamma_t^2} \rceil$, then also using the fact that $G(\boldsymbol{\pi}) \leq -V(\boldsymbol{\pi})$, we have

$$\mathbb{E}[V(\boldsymbol{\pi}(t+1)) - V(\boldsymbol{\pi}(t))|\boldsymbol{\pi}(t)] \leq -\gamma_t V(\boldsymbol{\pi}(t)) + C_4\gamma_t^2 + C_5\gamma_t\epsilon_t.$$

Taking $\epsilon_t \propto \gamma_t$, and taking expectations, we get

$$\mathbb{E}[V(\boldsymbol{\pi}(t+1))] \leq (1-\gamma_t)\mathbb{E}[V(\boldsymbol{\pi}(t))] + C_6\gamma_t^2.$$

Then, by telescoping we have:

$$\mathbb{E}\left[V(\boldsymbol{\pi}(t))\right] \leq \Pi_{s=1}^{t-1}(1-\gamma_s)V(\boldsymbol{\pi}(0)) + C_6\sum_{s=1}^{t-1}\gamma_s^2\Pi_{i=s+1}^{t-1}(1-\gamma_s)$$

with $\gamma_s = \frac{1}{s+1}$, we get:

$$\Pi_{s=1}^{t-1}(1-\gamma_s) = \frac{1}{t}$$

and

$$\sum_{s=0}^{t-1}\gamma_s^2\Pi_{j=s+1}^{t-1}(1-\gamma_j) = \sum_{s=0}^{t-1}\frac{1}{(s+1)^2}\Pi_{j=s+1}^{t-1}\frac{j}{j+1} \leq \frac{\log(t)}{t}.$$

Therefore,

$$\mathbb{E}[V(\boldsymbol{\pi}(t))] \leq \frac{V(\boldsymbol{\pi}(0))}{(t+1)} + \frac{C_6\log(t)}{t}.$$

Using Markov's inequality for the non-negative function $V(\boldsymbol{\pi})$, we conclude that $\lim_{t\to\infty} V(\boldsymbol{\pi}(t)) = 0$ in probability.

Note: With our choice at time-step t of $\gamma_t = \frac{1}{t+1}$, $\epsilon_t \propto \gamma_t$ and $M_t = \lceil \frac{1}{\epsilon_t\gamma_t^2} \rceil$, we note that the cumulative number of simulator uses until time T is polynomial in T as $\sum_{t=1}^{T} M_t = \Omega(T^4)$.

We now show that the gap function $V(\boldsymbol{\pi})$ is C-smooth. To show this, we show that the gradient of the gap function is Lipschitz continuous.

Lemma A.1. *$V(\boldsymbol{\pi})$ is C-smooth.*

Proof. The gradient of $V(\boldsymbol{\pi})$ is given by:

$$\nabla V(\boldsymbol{\pi}) = L(\boldsymbol{\pi}) = -F(\boldsymbol{\pi}) + H^T(\boldsymbol{\pi})(\boldsymbol{s}^*(\boldsymbol{\pi}, F(\boldsymbol{\pi})) - \boldsymbol{\pi}) + \boldsymbol{\lambda}(\boldsymbol{\pi}).$$

For any $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ we have

$$\begin{aligned}\nabla V(\boldsymbol{\pi}_1) - \nabla V(\boldsymbol{\pi}_2) = &-F(\boldsymbol{\pi}_1) + H^T(\boldsymbol{\pi}_1)(\boldsymbol{s}^*(\boldsymbol{\pi}_1, F(\boldsymbol{\pi}_1)) - \boldsymbol{\pi}_1) + \boldsymbol{\lambda}(\boldsymbol{\pi}_1) \\ &+ F(\boldsymbol{\pi}_2) - H^T(\boldsymbol{\pi}_2)(\boldsymbol{s}^*(\boldsymbol{\pi}_2, F(\boldsymbol{\pi}_2)) - \boldsymbol{\pi}_2) - \boldsymbol{\lambda}(\boldsymbol{\pi}_2)\end{aligned}$$

By assumption 2.1, we have that:

$$\begin{aligned}\|\nabla V(\boldsymbol{\pi}_1) - \nabla V(\boldsymbol{\pi}_2)\|_2 \leq &\|F(\boldsymbol{\pi}_2) - F(\boldsymbol{\pi}_1)\|_2 + \|H^T(\boldsymbol{\pi}_1)\|_2\|\boldsymbol{\pi}_2 - \boldsymbol{\pi}_1\|_2 \\ &+ \|H^T(\boldsymbol{\pi}_1)\|_2\|\boldsymbol{s}^*(\boldsymbol{\pi}_1, F(\boldsymbol{\pi}_1)) - \boldsymbol{s}^*(\boldsymbol{\pi}_2, F(\boldsymbol{\pi}_2))\|_2 \\ &+ \|H^T(\boldsymbol{\pi}_1) - H^T(\boldsymbol{\pi}_2)\|_2\|\boldsymbol{\pi}_2 - \boldsymbol{s}^*(\boldsymbol{\pi}_2, F(\boldsymbol{\pi}_2))\|_2 + \|\boldsymbol{\lambda}(\boldsymbol{\pi}_1) - \boldsymbol{\lambda}(\boldsymbol{\pi}_2)\|_2 \\ \leq &\alpha_V\|\pi_1 - \pi_2\| + \alpha_H(1+\alpha_s)\|\pi_1 - \pi_2\| + 2\alpha_H|A|\|\pi_1 - \pi_2\| \\ &+ c_0|A|\|\pi_1 - \pi_2\| \\ = &C\|\pi_1 - \pi_2\|\end{aligned}$$

where, $C = \alpha_V + \alpha_H(1+\alpha_S) + 2\alpha_H|A| + c_0$. It is easily shown by computing the gradient that $\boldsymbol{\lambda}(\boldsymbol{\pi})$ is Lipschitz. The C-smoothness of $V(\boldsymbol{\pi})$ allows us to claim the bound in (16). □

B GAP FUNCTION ANALYSIS

We start by reminding the reader of the gap function from (11) and (12):

$$\begin{aligned}V_i(\boldsymbol{\pi}) &\triangleq \max_{s\in\Delta(A_i)} \langle s - \pi_i, \nabla_{\pi_i} u_i^{f_i}(\boldsymbol{\pi})\rangle - \eta\mathrm{KL}(s, \pi_i) \\ &= \eta\log\left(\left\langle \pi_i, \exp\left(\frac{1}{\eta}\nabla_{\pi_i} u_i^{f_i}(\boldsymbol{\pi})\right)\right\rangle\right) - \eta\left\langle \pi_i, \frac{1}{\eta}\nabla_{\pi_i} u_i^{f_i}(\boldsymbol{\pi})\right\rangle, \\ \text{and } V(\boldsymbol{\pi}) &\triangleq \sum_{i\in[N]} \bar{V}_i(\boldsymbol{\pi}).\end{aligned}$$

We also remind the reader of the VI characterization of a GQRE from Lemma 2.2, namely, a strategy profile $\boldsymbol{\pi}^*$ is a GQRE if and only if

$$\max_{a_i\in A_i} \langle \delta_{a_i} - \pi_i^*, \nabla_{\pi_i} u_i^{f_i}(\pi^*)\rangle \leq 0, \ \forall i \in [N].$$

Whereas the relationship above is written for pure actions, by standard convex analysis, it extends to the following: a strategy profile $\boldsymbol{\pi}^*$ is a GQRE if and only if

$$\max_{s_i\in A_i} \langle s_i - \pi_i^*, \nabla_{\pi_i} u_i^{f_i}(\pi^*)\rangle \leq 0, \ \forall i \in [N]. \tag{21}$$

We start by proving that $V(\boldsymbol{\pi}^*) = 0$ for a GQRE $\boldsymbol{\pi}^*$. By the properties of the KL divergence function, we have

$$\langle \pi_i^* - \pi_i^*, \nabla_{\pi_i^*} u_i^{f_i}(\boldsymbol{\pi}^*)\rangle - \eta\mathrm{KL}(\pi^*, \pi_i^*) = 0$$

Then, by (21) and the properties of the KL divergence function, for any $\boldsymbol{s} \neq \boldsymbol{\pi}^*$ for all $i \in [N]$,

$$\langle s_i - \pi_i, \nabla_{\pi_i} u_i^{f_i}(\boldsymbol{\pi})\rangle \leq 0,$$

with the inequality strict for at least one $i \in [N]$. Further, $\mathrm{KL}(s_i, \pi_i^*) \geq 0$ for all $i \in [N]$ and strictly positive for at least one $i \in [N]$. Therefore, we have

$$\langle s_i - \pi_i^*, \nabla_{\pi_i} u_i^{f_i}(\boldsymbol{\pi}^*)\rangle - \eta\mathrm{KL}(s_i, \pi_i^*) \leq 0,$$

and hence, $V(\boldsymbol{\pi}^*) = 0$.

Next, let $\boldsymbol{\pi}$ which not a GQRE, be such that π_i has full support for all $i \in [N]$. Then, by (21), there is at least one $i \in [N]$, say 1 after relabeling, where there is an s_1 such that

$$\langle s_1 - \pi_1, \nabla_{\pi_1} u_1^{f_1}(\boldsymbol{\pi}) \rangle - \eta \mathrm{KL}(s_1, \pi_1) = \epsilon_1 > 0.$$

For $\delta \in (0, 1)$, set $\tilde{s}_1(\delta) = \pi_1 + \delta(s_1 - \pi_1)$, then for $\delta \ll 1$ we have

$$\begin{aligned} \mathrm{KL}(\tilde{s}_1, \pi_1) &= \sum_{a \in A_1} (\pi_{1,a} + \delta(s_{1,a} - \pi_{1,a})) \log\left(1 + \delta\left(\frac{s_{1,a}}{\pi_{1,a}} - 1\right)\right) \\ &\approx -\delta^2 \sum_{a \in A_1} \frac{(s_{1,a} - \pi_{1,a})^2}{\pi_{1,a}}, \end{aligned}$$

with error terms being higher order in δ, and hence, negligible. Then, we have

$$\begin{aligned} \langle \tilde{s}_1(\delta) - \pi_1, \nabla_{\pi_1} u_1^{f_1}(\boldsymbol{\pi}) \rangle - \eta \mathrm{KL}(\tilde{s}_1(\delta), \pi_1) &\approx \delta\epsilon - \delta^2 \eta \sum_{a \in A_1} \frac{(s_{1,a} - \pi_{1,a})^2}{\pi_{1,a}} \\ &= \delta\left(1 - \delta\eta \sum_{a \in A_1} \frac{(s_{1,a} - \pi_{1,a})^2}{\pi_{1,a}}\right). \end{aligned}$$

Then, we can make δ small enough that

$$\delta\eta \sum_{a \in A_1} \frac{(s_{1,a} - \pi_{1,a})^2}{\pi_{1,a}} < 1.$$

Hence, $V_1(\boldsymbol{\pi}) > 0$, and since $V_i(\boldsymbol{\pi}) \geq 0$ for all $i \in [N]$ for any $\boldsymbol{\pi}$ (take $\boldsymbol{s} = \boldsymbol{\pi}$ or $s_i = \boldsymbol{\pi}_i$ for all $i \neq 1$), it follows that $V(\boldsymbol{\pi}) > 0$.

Note that our gap function is 0 for all pure strategy profiles. However, in Algorithm 1 we always operate at fully mixed strategies, and there we have negative drift except for the unique fully mixed equilibrium, so the issue of the gap function being zero at pure strategy profiles doesn't cause trouble.

C ADDITIONAL NUMERICAL SIMULATIONS

Environment Setup: The experiments are done on a MacBook Air with an Apple M1 chip, 16 GB memory and 10 core CPU. All codes are written in Python3 using several open source packages. The running time for all experiments ranges from less than a minute to a few hours.

C.1 ADDITIONAL NUMERICAL RESULT FOR STRONGLY MONOTONE AND RANK-k GAMES

We start by presenting results for *Matching Pennies:* This is a class game with 2×2 payoff matrices: $A = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$, and $B = -A$. It is a zero-sum game (so rank is 0) in which each player tries to match/mismatch the other's action with a unique mixed-strategy NE. Figure 4 presents the Mean GQRE Gap, i.e., the gap from the GQRE for Algorithm 1 (labelled FW Smoothed LP), and we see that our algorithm achieves a very low gap in about 1000 iterations, an order of magnitude better than the next best.

Additional results are reported for strongly-monotone games in Figures 5 and 6. Again, the performance of Algorithm 1 stands out. Similarly, in Figure 7 we present the peformance of Algorithm 1 relative to the baselines for games of different ranks. Here, Algorithm 1 show comparable performance but much better stability.

C.2 ADDITIONAL NUMERICAL RESULTS WITH UNIFORM-MIX FRANK–WOLFE

We evaluate the Uniform-Mix Frank–Wolfe (FW) update against Algorithm 1 and other baselines on Jordan's three-player, two-action game (Jordan, 1993) in Figure 9.

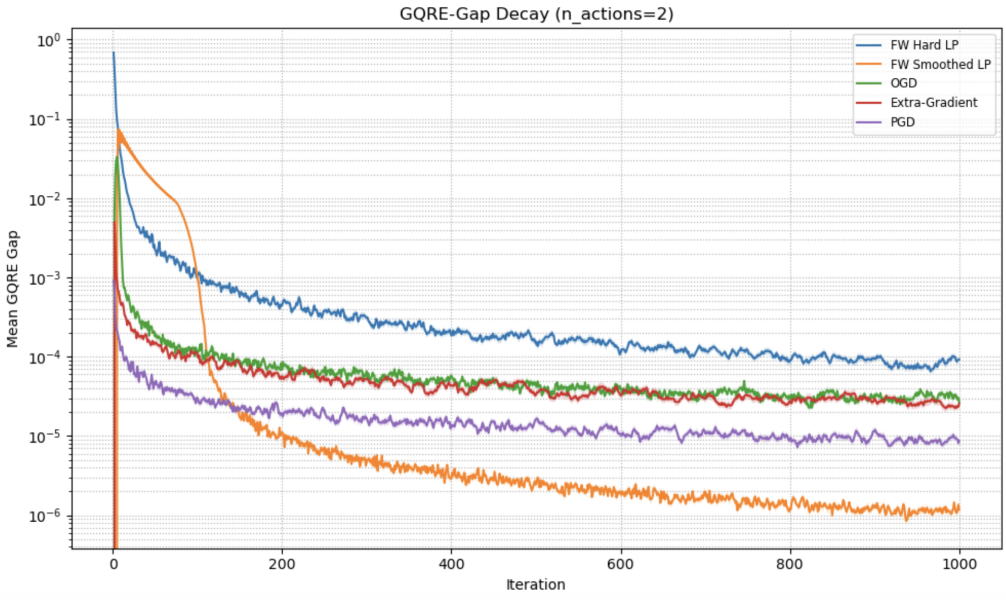


Figure 4: GQRE gap in the matching pennies game.

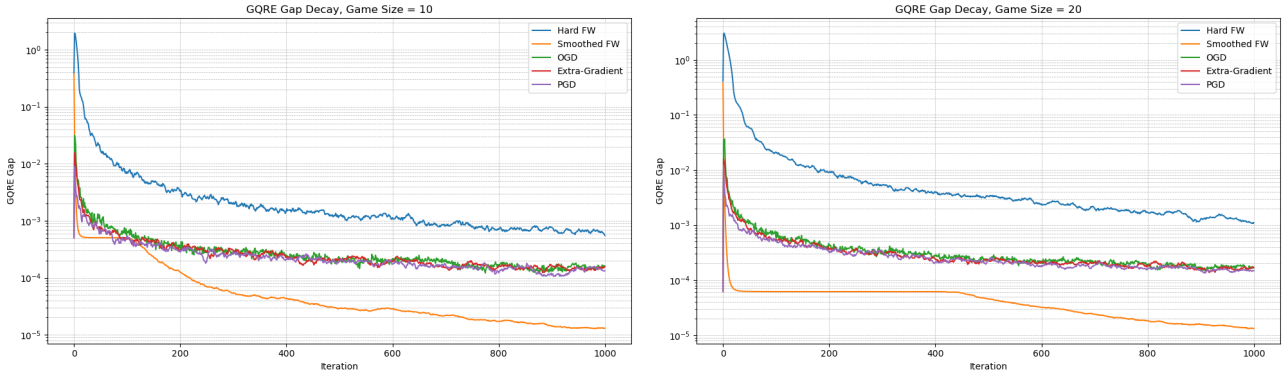


Figure 5: GQRE gap decay for strongly monotone games with sizes $n = 10, 20$.

C.3 JORDAN'S 3-PLAYER GAME: STABILITY VS. λ AND SAMPLING M

Again, we study Jordan's three-player, two-action game from (Jordan, 1993), in which the unique symmetric QRE is $(0.5, 0.5, 0.5)$. For the smoothed Frank–Wolfe (FW) update, Assumption 2.2 (ensuring uniqueness and local stability) is satisfied for small λ (here $\lambda < 1$). For larger λ, uniqueness may fail and equilibria can become unstable, consistent with the instability phenomena relative to uncoupled dynamics (Hart and Mas-Colell, 2003).

Figure 10 shows simulation results for $\lambda \in \{0.5, 5, 50\}$. The top panel plots the Lyapunov gap (log scale), while the bottom panel tracks Player 1's probability of playing action 0. For $\lambda = 0.5$, the gap decays rapidly and the strategy converges toward the symmetric GQRE. At $\lambda = 5$, convergence occurs more slowly and is sensitive to sampling noise. At $\lambda = 50$, the dynamics fail to converge and drift toward the boundary of the simplex, illustrating instability.

Convergence vs. sampling M. We conducted 100 independent experiments for each (λ, M) and measured the proportion with final gap $< 10^{-3}$, and report the results in Table 2. When $\lambda = 0.5$, convergence is robust for all M. For $\lambda = 5$, larger M substantially improves convergence, while for $\lambda = 50$ convergence never occurs, even at high M.

These results confirm the stability picture: when Assumption 2.2 holds (small λ), convergence is guaranteed; in the intermediate regime (moderate λ), variance reduction via larger M can restore convergence; for large λ, equilibria are unstable and convergence is not possible, in line with the impossibility results for uncoupled dynamics Hart and Mas-Colell

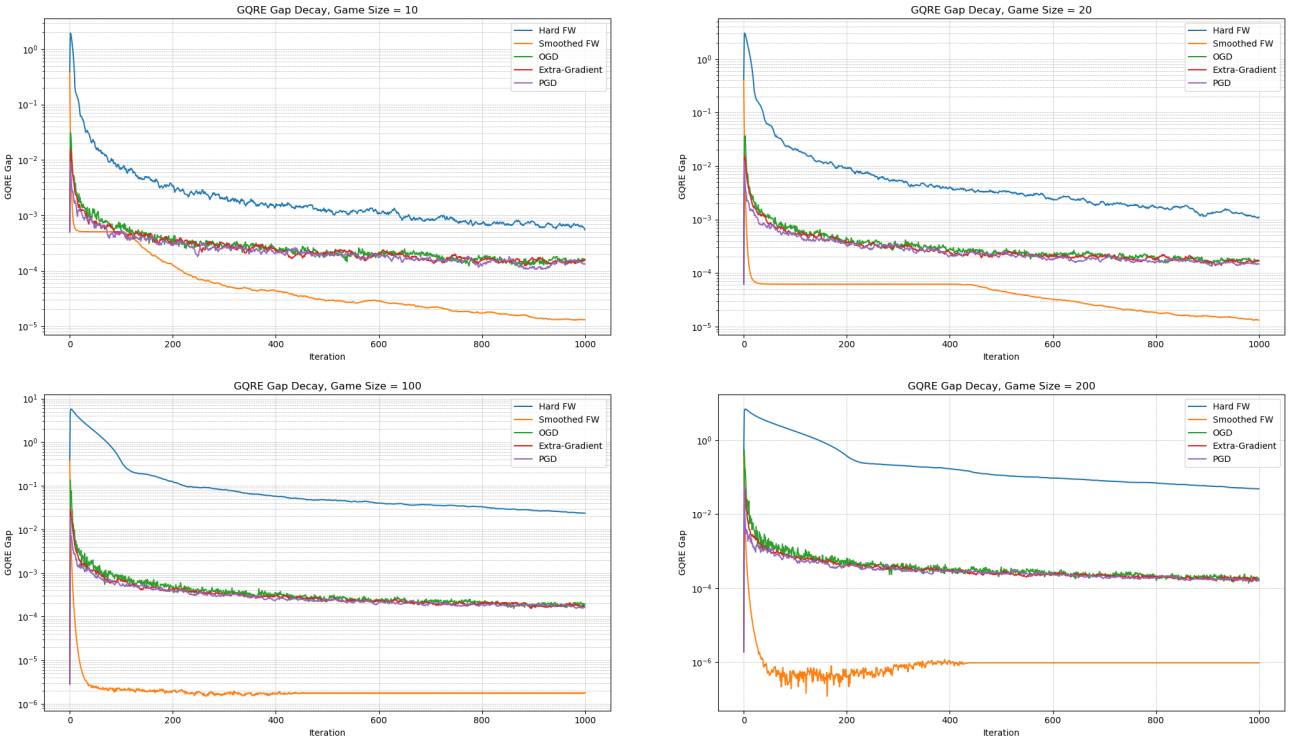


Figure 6: GQRE gap decay on strongly monotone games of increasing size. Smoothed methods consistently outperform in both convergence rate and stability across larger games.

M	$\lambda = 0.5$	$\lambda = 5$	$\lambda = 50$
500	100.0%	37.0%	0.0%
1000	100.0%	58.0%	0.0%
5000	100.0%	98.0%	0.0%

Table 2: Convergence rates (final gap $< 10^{-3}$) across λ and M in Jordan's game.

(2003).

C.4 HETEROGENEOUS REGULARIZERS: MIXED f_i FUNCTIONS ACROSS PLAYERS

To understand the potential benefit of assigning different f_i functions to different agents, we ran simulations with heterogeneous regularizers. Specifically, for the rank-k game of size 5, we considered all pairwise combinations of four divergences: Shannon entropy (KL), Rényi ($\alpha = 0.5$), $0.5\|\cdot\|_2^2$ relative to uniform, and the total variation (TV) distance to uniform. In each run, we fixed $\lambda = 1$ and used our smoothed Frank–Wolfe algorithm.

Distances from uniform. We also computed the KL, L2, and TV distances of each resulting distribution from the uniform distribution. Results (Table 3) show that entropy-based equilibria remain closest to uniform, Rényi pushes farthest away, and TV regularization produces distributions almost indistinguishable from uniform.

Convergence surrogate. To evaluate convergence, we use the gap value at $T = 1000$ iterations as a surrogate measure. Table 4 shows these values. Mixed regularizers yield notable improvements: Entropy–TV and L2–Entropy significantly reduce the final gap compared to their homogeneous counterparts. In contrast, homogeneous cases of Rényi and TV divergences do not benefit from mixing.

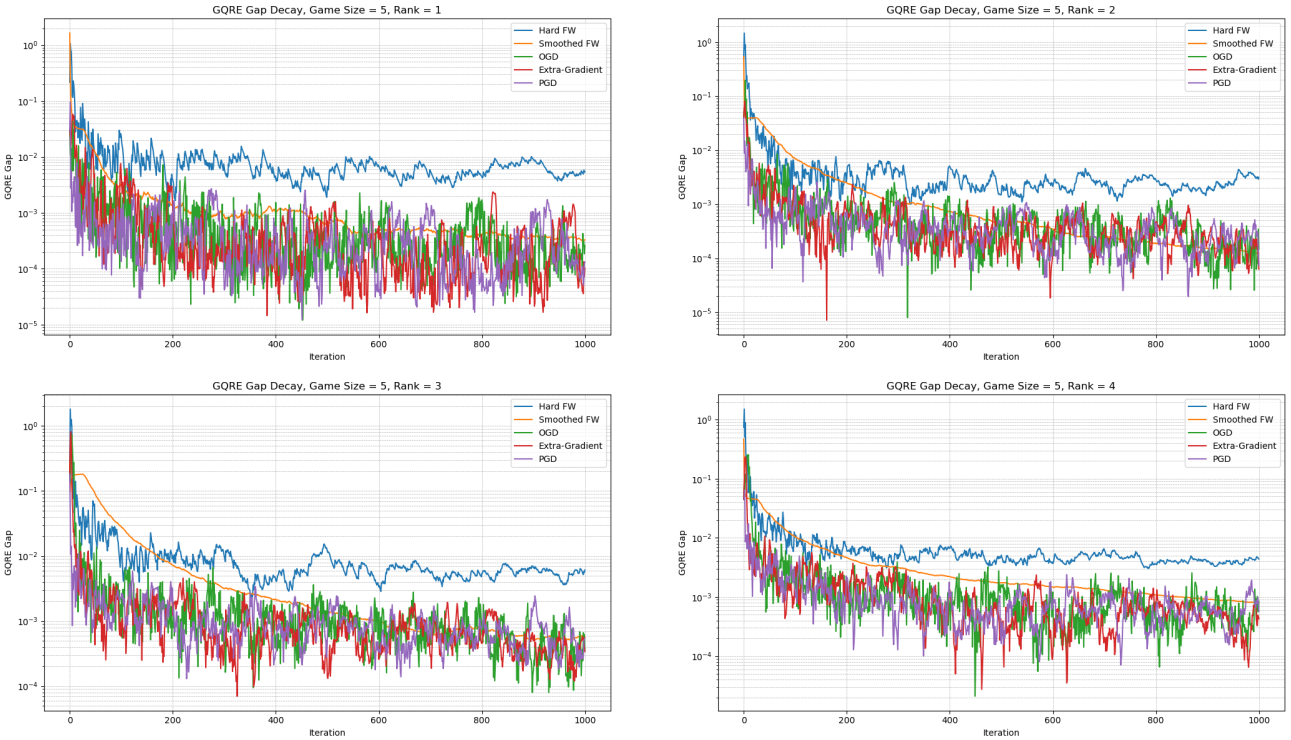


Figure 7: GQRE gap decay for games with size 5 and different ranks.

f_1–f_2	KL divergence	L2 distance	TV distance
Entropy–Entropy	0.0319	0.1180	0.1022
Entropy–TV	0.0119	0.0712	0.0620
L2–Entropy	0.0751	0.1855	0.1631
L2–TV	0.0333	0.1213	0.1067
Rényi–Rényi	0.1992	0.3156	0.2822
TV–TV	0.0000	0.0008	0.0008

Table 3: Distances of Player 1 distributions from uniform under different regularizer pairs. Entropy and TV combinations yield the closest-to-uniform distributions, while Rényi divergence pushes farthest apart.

	Entropy	L2	Rényi	TV
Entropy	0.000081	0.054651	0.603757	0.049424
L2	0.052204	0.078403	0.622428	0.098707
Rényi	0.313565	0.276359	0.967499	0.813753
TV	0.039242	0.136138	0.855256	0.039518

Table 4: Surrogate convergence: gap at $T = 1000$ for each pair (f_1, f_2). Mixed combinations (e.g., Entropy–TV, L2–Entropy, L2–TV) improve convergence rates relative to homogeneous choices.

Summary. These results show that heterogeneous choice of regularizers can be beneficial. In particular, Entropy–TV and L2–based mixtures consistently improve convergence compared to homogeneous choices. This provides a concrete example where using different f_i across players yields different GQRE and also enhances the robustness of the algorithm.

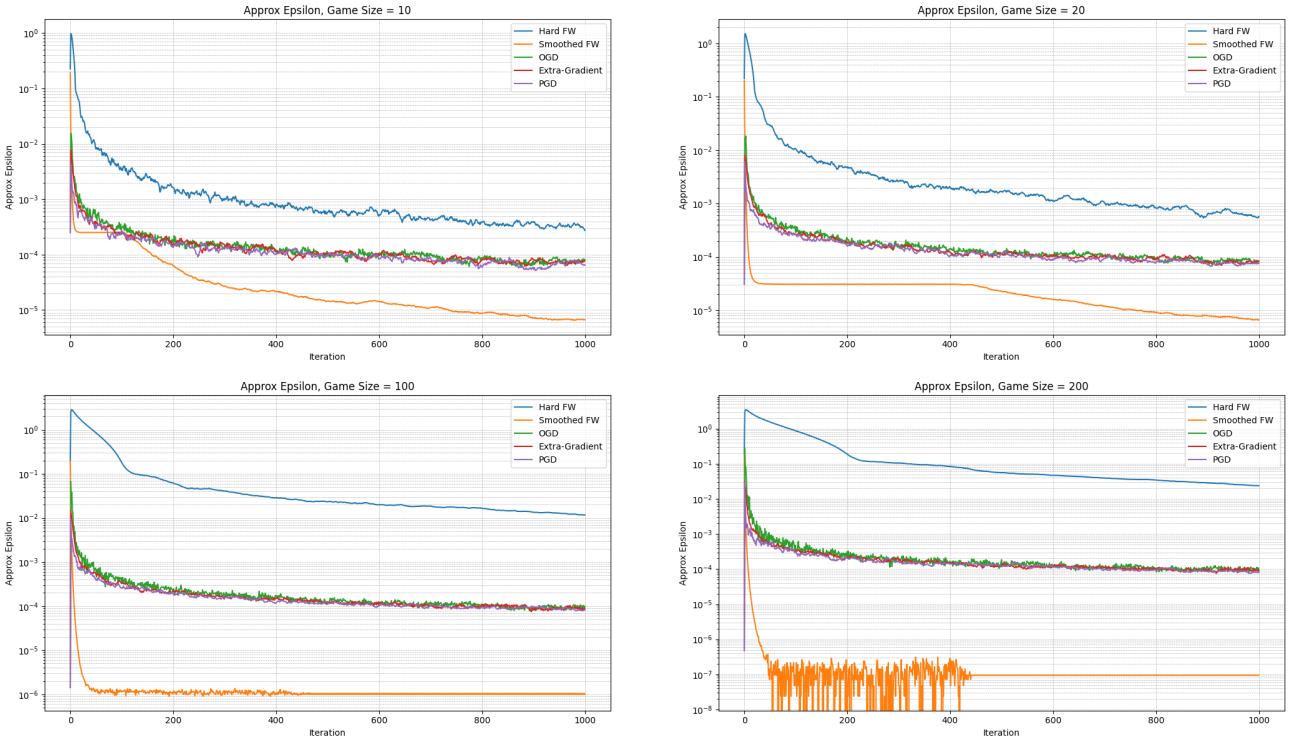
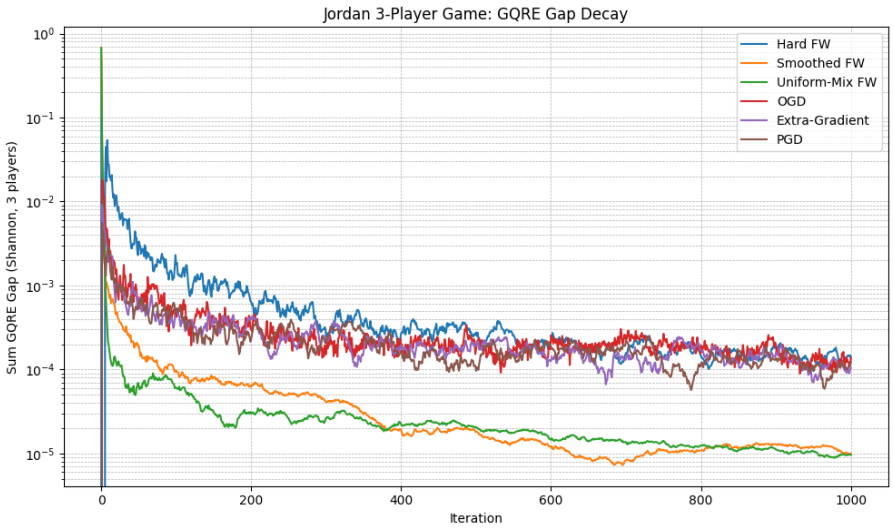


Figure 8: Approx ϵ-GQRE vs iteration for strongly monotone games.



(a) Jordan's 3-player game.

Figure 9: Sum GQRE gap decay in Jordan's 3-player cyclic game. Uniform-Mix FW closely tracks Smoothed FW, converging to near-zero gaps in a non-monotone multi-agent setting.

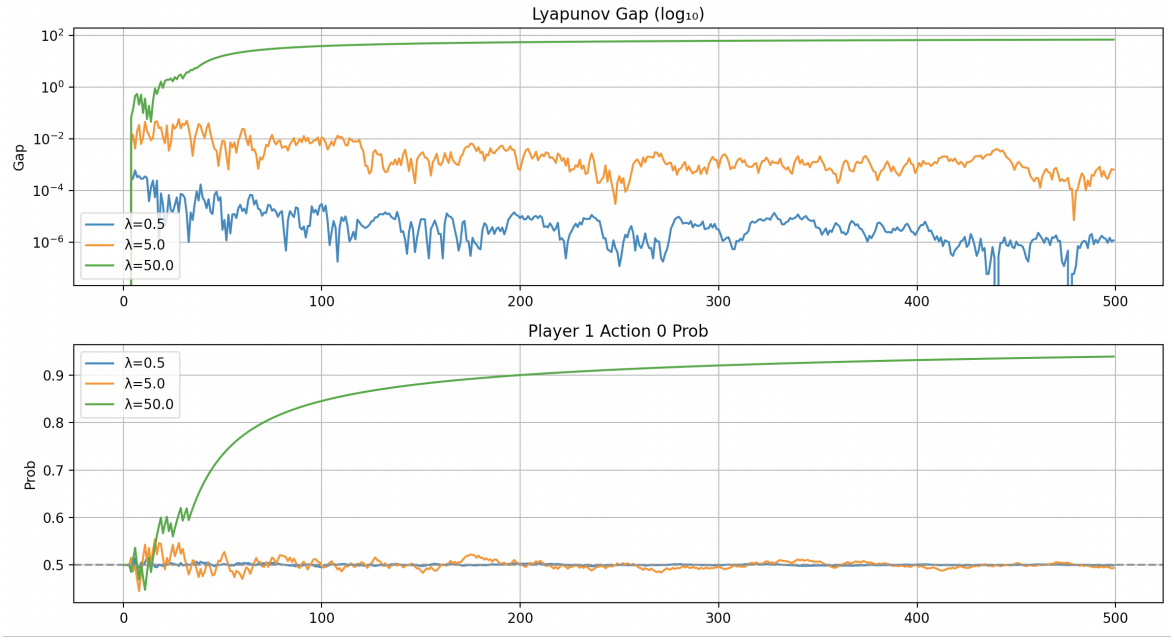


Figure 10: Jordan's three-player, two-action game with $\lambda \in \{0.5, 5, 50\}$. Top: Lyapunov gap ($\log_{10}$). Bottom: Player 1's action-0 probability. Small λ yields robust convergence, moderate λ produces mixed outcomes, and large λ leads to instability.

D EXAMPLES OF PENALTY FUNCTIONS AND RESPONSES

D.1 FAIR QUANTAL RESPONSE

We aim to solve the optimization problem:

$$\max_{\mathbf{p}} \quad \sum_{a\in A} p(a)\lambda u(a) - \frac{1}{2}\sum_{a\in A}\left|p(a) - \frac{1}{|A|}\right| \text{ s.t. } \sum_{a\in A} p(a) = 1, \quad p(a) \geq 0 \quad \forall a \in A,$$

which we rewrite through scaling as

$$\begin{aligned}\max_{\mathbf{p}} \quad & \sum_{a\in A} p(a)2\lambda u(a) - \sum_{a\in A}\left|p(a) - \frac{1}{|A|}\right| \\ \text{s.t.} \quad & \sum_{a\in A} p(a) = 1, \quad p(a) \geq 0 \quad \forall a \in A.\end{aligned}$$

Let $n = |A|$, so the uniform distribution is given by $p(a)^{\text{uniform}} = \frac{1}{n}$. Introducing the Lagrange multiplier μ for the sum probability constraint, the Lagrangian is given by:

$$\mathcal{L} = \sum_{a\in A} p(a)2\lambda u(a) - \sum_{a\in A}\left|p(a) - \frac{1}{n}\right| - \mu\left(\sum_{a\in A} p(a) - 1\right)$$

Since we have separability, we can equivalently solve the following piece-wise linear continuous function with a kink at $\frac{1}{|A|}$ for each action $a \in A$:

$$\max_{p\in[0,1]} p\left(2\lambda u(a) - \mu\right) - |p - 1/n|$$

In the segment $[0, 1/n]$, the function is

$$-1/n + p\left(2\lambda u(a) - \mu + 1\right),$$

and in the segment $[1/n, 1]$, the function is

$$1/n + p\left(2\lambda u(a) - \mu - 1\right).$$

Hence, we have the maximizer for each μ given by

$$p^*(a;\mu) \in \begin{cases} \{0\} & \text{if } \mu > 2\lambda u(a) + 1 \\ [0, 1/n] & \text{if } \mu = 2\lambda u(a) + 1 \\ \{1/n\} & \text{if } \mu \in (2\lambda u(a) - 1, 2\lambda u(a) + 1) \text{ and } 2\lambda u(a) > 1 \\ [1/n, 1] & \text{if } \mu = 2\lambda u(a) - 1 \text{ and } 2\lambda u(a) \geq 1 \\ \{1\} & \text{if } \mu < 2\lambda u(a) - 1 \text{ and } 2\lambda u(a) > 1 \\ \{1/n\} & \text{if } \mu \in [0, 2\lambda u(a) + 1) \text{ and } 2\lambda u(a) < 1 \end{cases}$$

Thereafter, we choose μ to be a value such that $p^*(\cdot;\mu)$ vector lies in the probability simplex. For this we can start with $1 + \frac{2\max_a u(a)}{\lambda}$ and then decrease μ until we get a probability vector. At points where we have multiple solutions, a careful selection may be needed.

D.2 WASSERSTEIN DISTANCE PENALTY

Let $\mathbf{p}, \mathbf{p}_0 \in \Delta_n$ be discrete probability distributions over a uniform grid of ordered support points $x_1 < x_2 < \cdots < x_n \in \mathbb{R}$, denoted by the vector $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$.

We consider the optimization problem:

$$\max_{\mathbf{p} \in \Delta_n} \left\{ \lambda \mathbf{v}^\top \mathbf{p} - W_2^2(\mathbf{p}, \mathbf{p}_0) \right\}, \tag{22}$$

where:

- $\mathbf{v} \in \mathbb{R}^n$ is a utility vector,
- $\lambda > 0$ is a regularization parameter,
- $W_2^2(\mathbf{p}, \mathbf{p}_0)$ is the squared Wasserstein-2 distance between $\mathbf{p}$ and $\mathbf{p}_0$.

This formulation can be generalized when using a Wasserstein p-distance as:

- $p \in \Delta_n$ be a probability distribution over n actions (decision variable),
- $q \in \Delta_n$ be a reference distribution,
- $u \in \mathbb{R}^n$ be the utility vector,
- $d(i, j)$ denote the ground distance between actions i and j,
- $\lambda > 0$ be the regularization parameter.

D.3 SQUARED DISTANCE PENALTY

The optimization problem here is given by:

$$\max_{\boldsymbol{p} \in \Delta_n} \left\{ \lambda \boldsymbol{v}^\top \boldsymbol{p} - \left(\boldsymbol{x}^\top (\boldsymbol{p} - \boldsymbol{p}_0) \right)^2 \right\}, \tag{23}$$

where:

- $\boldsymbol{p} \in \Delta_n$ is a probability vector (i.e., $\boldsymbol{p} \geq 0, \sum_i p_i = 1$),
- $\boldsymbol{v} \in \mathbb{R}^n$ is a utility vector,
- $\boldsymbol{x} \in \mathbb{R}^n$ is a vector of support points (e.g., $\boldsymbol{x} = (1, 2, \ldots, n)^\top$),
- $\boldsymbol{p}_0 \in \Delta_n$ is a reference distribution,
- $\lambda > 0$ is a regularization parameter.

Let us define:

$$a := \boldsymbol{x}, \qquad b := \boldsymbol{x}^\top \boldsymbol{p}_0,$$

then the objective becomes:

$$\min_{\boldsymbol{p}\in\Delta_n} \left\{-\lambda \boldsymbol{v}^\top \boldsymbol{p} + (a^\top \boldsymbol{p} - b)^2\right\}. \tag{24}$$

Expanding the square term:

$$(a^\top \boldsymbol{p} - b)^2 = \boldsymbol{p}^\top a a^\top \boldsymbol{p} - 2ba^\top \boldsymbol{p} + b^2.$$

Therefore, the optimization problem becomes:

$$\min_{\boldsymbol{p}\in\Delta_n} \left\{\boldsymbol{p}^\top Q \boldsymbol{p} + \boldsymbol{c}^\top \boldsymbol{p}\right\} + \text{const}, \tag{25}$$

where:

$$Q = aa^\top, \qquad \boldsymbol{c} = -\lambda \boldsymbol{v} - 2ba.$$

This is a quadratic program (QP) over the simplex:

$$\text{minimize} \quad \boldsymbol{p}^\top Q \boldsymbol{p} + \boldsymbol{c}^\top \boldsymbol{p}, \tag{26}$$

$$\text{subject to} \quad \boldsymbol{p} \geq 0, \quad \mathbf{1}^\top \boldsymbol{p} = 1. \tag{27}$$

Application to approximation of Wasserstein-2 Distance in 1D: In 1D, with ordered supports and under some assumptions to be detailed below, the Wasserstein-2 distance can be approximated by the squared difference in expected values:

$$W_2^2(\mathbf{p}, \mathbf{p}_0) \approx (\mathbb{E}_{\mathbf{p}}[\mathbf{x}] - \mathbb{E}_{\mathbf{p}_0}[\mathbf{x}])^2 = \left(\mathbf{x}^\top (\mathbf{p} - \mathbf{p}_0)\right)^2. \tag{28}$$

The approximation can be shown as follows: Let $\mathbf{p}$ and $\mathbf{p}_0$ be probability measures on $\mathbb{R}$ with finite second moments, i.e., $p, p_0 \in \mathcal{P}_2(\mathbb{R})$. Denote their cumulative distribution functions by F_p, F_{p_0}, and their quantile functions by F_p^{-1}, $F_{p_0}^{-1}$. Define:

$$X := F_p^{-1}(U), \quad Y := F_{p_0}^{-1}(U), \quad U \sim \text{Unif}[0, 1].$$

Then the squared 2-Wasserstein distance is given by:

$$W_2^2(p, p_0) = \int_0^1 \left(F_p^{-1}(u) - F_{p_0}^{-1}(u)\right)^2 du = \mathbb{E}[(X - Y)^2].$$

We apply the identity:

$$\mathbb{E}[(X - Y)^2] = (\mathbb{E}[X - Y])^2 + \text{Var}(X - Y),$$

which yields:

$$W_2^2(p, p_0) = (\mu_1 - \nu_1)^2 + \text{Var}(X - Y),$$

where $\mu_1 := \mathbb{E}[X]$, $\nu_1 := \mathbb{E}[Y]$. The approximation $W_2^2(p, p_0) \approx (\mu_1 - \nu_1)^2$ is valid when $\text{Var}(X - Y) \ll (\mu_1 - \nu_1)^2$, i.e., the variance term is negligible compared to the squared mean difference. This holds when:

- Translation of means:
  If $\nu = \mu(\cdot - \delta)$, then $F_\nu^{-1}(u) = F_\mu^{-1}(u) + \delta$, so

  $$X - Y = -\delta \quad \text{a.s.}, \quad \text{Var}(X - Y) = 0.$$

  Thus:

  $$W_2^2(p, p_0) = \delta^2 = (\mu_1 - \nu_1)^2.$$

- Small Perturbations:
  If $F_\mu^{-1}(u) - F_\nu^{-1}(u) = \delta(u)$ with $\delta(u)$ small and smooth, then

  $$X - Y = \delta(U), \quad \text{Var}(X - Y) = \text{Var}(\delta(U)) \text{ is small.}$$

  Hence,

  $$W_2^2(p, p_0) = (\mu_1 - \nu_1)^2 + o(1).$$

D.4 RENYI DIVERGENCE PENALTY

We aim to maximize an expected utility function while keeping the probability distribution P close to a uniform reference distribution Q using a Rényi divergence penalty. Let:

- X be a discrete random variable with support $\mathcal{X} = \{x_1, x_2, ..., x_n\}$.
- $P = (p_1, p_2, ..., p_n)$ be the probability distribution over X.
- $Q = (q_1, q_2, ..., q_n)$ be the reference uniform distribution, i.e., $q_i = \frac{1}{n}$ for all i.
- $U(x)$ be the utility function associated with choosing x.
- For $\alpha \in (0, 1)$, $D_\alpha(P\|Q)$ be the Rényi divergence penalty.

The Rényi divergence of order α between P and uniform Q is:

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \sum_{i=1}^{n} p_i^\alpha q_i^{1-\alpha}.$$

Since $q_i = \frac{1}{n}$, we substitute:

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \left(\frac{1}{n^{1-\alpha}} \sum_{i=1}^{n} p_i^\alpha \right).$$

The optimization problem is given by:

$$\max_{\boldsymbol{p}} \sum_{i=1}^{n} p_i \lambda U(x_i) - D_\alpha(\boldsymbol{p}\|Q)$$

subject to:

$$\boldsymbol{p} = (p_1, \ldots, p_n) \text{ s.t. } \sum_{i=1}^{n} p_i = 1, \quad p_i \geq 0, \quad \forall i.$$

where $\lambda > 0$ controls the strength of the penalty. The Lagrangian for $\alpha \in (0, 1)$ ($\alpha = 1$ case is QR) is

$$\mathcal{L}(P, \mu) = \sum_{i=1}^{n} p_i \lambda U(x_i) + \frac{1}{1 - \alpha} \log \left(\frac{1}{n^{1-\alpha}} \sum_{i=1}^{n} p_i^\alpha \right) + \mu \left(1 - \sum_{i=1}^{n} p_i \right).$$

Then, taking a derivative with respect to p_i we get

$$\lambda U(x_i) + \frac{\alpha}{1 - \alpha} \frac{p_i^{\alpha - 1}}{\sum_{j=1}^{n} p_j^\alpha} - \mu \leq 0,$$

with equality if $p_i > 0$. From this it follows that

$$p_i^{\alpha - 1} \leq \frac{1 - \alpha}{\alpha} (\sum_{j=1}^{n} p_j^\alpha)(\mu - \lambda U(x_i)).$$

From the above it follows that the Lagrange multiplier $\mu > \lambda \max_i U(x_i)$ and none of the optimal probability values are 0, so that

$$\begin{aligned} p_i^{\alpha - 1} &= \frac{1 - \alpha}{\alpha} (\sum_{j=1}^{n} p_j^\alpha)(\mu - \lambda U(x_i)) \\ \Rightarrow\ p_i &\propto (\mu - \lambda U(x_i))^{\frac{-1}{1-\alpha}}. \end{aligned}$$

Hence, we have

$$p_i(\mu) = \frac{(\mu - \lambda U(x_i))^{\frac{-1}{1-\alpha}}}{\sum_j (\mu - \lambda U(x_j))^{\frac{-1}{1-\alpha}}}.$$

Since we also have

$$\frac{p_i^\alpha}{\sum_{j=1}^n p_j^\alpha} = \frac{1-\alpha}{\alpha} p_i(\mu - \lambda U(x_i))$$
$$\Rightarrow\ 1 = \frac{1-\alpha}{\alpha} \sum_i p_i(\mu)(\mu - \lambda U(x_i)),$$

The correct value of μ is determined by solving for the unique root of

$$\frac{\sum_j (\mu - \lambda U(x_j))^{\frac{-\alpha}{1-\alpha}}}{\sum_j (\mu - \lambda U(x_j))^{\frac{-1}{1-\alpha}}} = \frac{\alpha}{1-\alpha}$$

with $\mu > \lambda \max_i U(x_i)$.

D.5 HELLINGER DISTANCE PENALTY

We consider a Hellinger distance perturbation with uniform distribution as the reference function. Define:

- $p \in \Delta_n$: decision distribution over n actions,
- $u \in \mathbb{R}^n$: utility vector,
- $q = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right) \in \Delta_n$: uniform reference distribution,
- $\lambda > 0$: regularization parameter.

The squared Hellinger distance is defined as:

$$H^2(p, q) = 1 - \sum_{i=1}^n \sqrt{p_i q_i} = 1 - \frac{1}{\sqrt{n}} \sum_{i=1}^n \sqrt{p_i}$$

The optimization problem is given by

$$\max_{\boldsymbol{p} \in \Delta(A)} \sum_{i=1}^n p_i \lambda u_i + \sqrt{\frac{p_i}{n}}$$

The Lagrangian obtained by relaxing the sum constraint is

$$L(\boldsymbol{p}, \mu) = \sum_{i=1}^n p_i(\lambda u_i - \mu) + \sqrt{\frac{p_i}{n}} + \mu$$

Taking derivatives in p_i, we get

$$\lambda u_i - \mu + \frac{1}{2\sqrt{p_i n}} \leq 0$$
$$\Rightarrow \frac{1}{\sqrt{p_i}} \leq 2\sqrt{n}(\mu - \lambda u_i).$$

Hence, we have the following: at the optimum $p_i > 0$ for all i; and

$$p_i = \frac{1}{4n(\mu - \lambda u_i)^2}$$

where μ is the unique root of

$$\sum_i \frac{1}{(\mu - \lambda u_i)^2} = 4n$$

with $\mu > \lambda \max_i u_i$.

E PROOF OF LEMMA 2.2

Proof. To prove, we can proceed as follows:
Step 1: Fix player $i \in [N]$. From Theorem 2.1, checking whether (π_i^*, π_{-i}^*) is GQRE is equivalent to checking if (π_i^*, π_{-i}^*) is a NE of $\mathcal{G}_f$. This holds if and only if $\pi_i^* \in \arg\max_{\pi \in \Delta(A_i)} u_i^{f_i}(\pi, \pi_{-i}^*)$. The latter is equivalent to the following using the VI characterization:

$$\pi_i^* \text{ satisfies } \langle \nabla_{\pi_i} u_i^{f_i}(\boldsymbol{\pi}^*), \pi_i - \pi_i^* \rangle \leq 0, \ \forall \pi_i \in \Delta(A_i).$$

The LHS is a linear program, and the check needs to be performed only over the set of pure-strategies since $\Delta(A_i)$ is a simplex and the maximum occurs at the extreme points.
Step 2: This follows immediately from the definition of an ϵ-GQRE.

Note that the utilities being multi-linear in the strategies, determining the value of ϵ for a given non-NE strategy profile $\boldsymbol{\pi}$ in a finite normal form game only needs evaluation of pure actions. □

F PROOF OF PROPOSITION 2.1 AND THEOREM 2.1

The argument for the proof of Proposition 2.1 is the following.

Proof. For every given $\boldsymbol{\pi}_{-i}$, the optimization function in (4) is a convex optimization problem when f_i is a convex function. Then, by general results in convex optimization, an optimizer exists and the GQR is well-defined. If, in addition, f_i is strictly convex, then the objective of the optimization problem is strictly convex which establishes the uniqueness of the optimizer. Finally, using Berge's Theorem of the Maximum, as the objective is jointly continuous in $\boldsymbol{\pi}$ and a unique optimizer exists when f_i is strictly convex, the optimizer is a continuous function of the parameter $\boldsymbol{\pi}_{-i}$. □

The argument for the proof of Theorem 2.1 is the following.

Proof. We note that for a strictly convex perturbation f_i, the $\mathcal{G}_f$ is a concave game and therefore, the existence of Nash Equilibrium follows from Rosen (1965). Further, the diagonal domninance criteria used by Rosen (1965) follows from Assumption 2.2. □