The appendices are structured as follows: App. A presents the setting and the existing results. In particular, we start by introducing the setting of the mirror-prox algorithm in §A.1 and detail the relation between solving this problem and finding Nash equilibria in convex n-player games §A.2. We then present the proofs of our theorems in App. B. We analyze the DSEG algorithm (Alg. 1) and study its variance-reduction version. App. D presents further experimental results and details.

A. Existing results

A.1. Mirror-prox

Mirror-prox and mirror descent are the formulation of the extra-gradient method and gradient descent for non-Euclidean (Banach) spaces. Bubeck (2015) (which is a good reference for this subsection) and Juditsky et al. (2011) study extra-gradient/mirror-prox in this setting. We provide an introduction to the topic for completeness.

Setting and notations. We consider a Banach space E and a compact set $\Theta \subset E$. We define an open convex set $\mathcal{D}$ such that Θ is included in its closure, that is $\Theta \subseteq \bar{\mathcal{D}}$ and $\mathcal{D} \cap \Theta \neq \emptyset$. The Banach space E is characterized by a norm $\|\cdot\|$. Its conjugate norm $\|\cdot\|_*$ is defined as $\|\xi\|_* = \max_{z:\|z\|\leqslant 1} \langle \xi, z\rangle$. For simplicity, we assume $E = \mathbb{R}^n$.

We assume the existence of a mirror map for Θ, which is defined as a function $\Phi\colon \mathcal{D} \to \mathbb{R}$ that is differentiable and μ-strongly convex i.e.

$$\forall x, y \in \mathcal{D},\ \langle \nabla\Phi(x) - \nabla\Phi(y), x - y\rangle \geqslant \mu\|x - y\|^2.$$

We can define the *Bregman divergence* in terms of the mirror map.

Definition 2. *Given a mirror map $\Phi\colon \mathcal{D} \to \mathbb{R}$, the Bregman divergence $D : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is defined as*

$$D(x, y) \triangleq \Phi(x) - \Phi(y) - \langle \nabla\Phi(y), x - y\rangle.$$

Note that $D(\cdot, \cdot)$ is always non-negative. For more properties, see e.g. Nemirovsky & Yudin (1983) and references therein. Given that Θ is compact convex space, we define $\Omega = \max_{x\in\mathcal{D}\cap\Theta} \Phi(x) - \Phi(x_1)$. Lastly, for $z \in \mathcal{D}$ and $\xi \in E^*$, we define the prox-mapping as

$$P_z(\xi) \triangleq \operatorname*{argmin}_{u\in\mathcal{D}\cap\Theta}\{\Phi(u) + \langle \xi - \nabla\Phi(z), u\rangle\} = \operatorname*{argmin}_{u\in\mathcal{D}\cap\Theta}\{D(z, u) + \langle \xi, u\rangle\}. \tag{11}$$

The mirror-prox algorithm is the most well-known algorithm to solve convex n-player games in the mirror setting (and variational inequalities, see §A.2). An iteration of mirror-prox consists of:

$$\begin{aligned} &\text{Compute the extrapolated point: } \begin{cases} \nabla\Phi(y_{\tau+1/2}) = \nabla\Phi(\theta_\tau) - \gamma F(\theta_\tau), \\ \theta_{\tau+1/2} = \operatorname{argmin}_{x\in\mathcal{D}\cap\Theta}\ D(x, y_{\tau+1/2}), \end{cases} \\ &\text{Compute a gradient step: } \begin{cases} \nabla\Phi(y_{\tau+1}) = \nabla\Phi(\theta_\tau) - \gamma F(\theta_{\tau+1/2}), \\ \theta_{\tau+1} = \operatorname{argmin}_{x\in\mathcal{D}\cap\Theta}\ D(x, y_{\tau+1}). \end{cases} \end{aligned} \tag{12}$$

Remark that the extra-gradient algorithm defined in equation (3) corresponds to the mirror-prox (12) when choosing $\Phi(x) = \frac{1}{2}\|x\|_2^2$.

Lemma 1. *By using the proximal mapping notation (11), the mirror-prox updates are equivalent to:*

$$\begin{aligned} \textit{Compute the extrapolated point: } \theta_{\tau+1/2} &= P_{\theta_\tau}(\gamma F(\theta_\tau)), \\ \textit{Compute a gradient step: } \theta_{\tau+1} &= P_{\theta_\tau}(\gamma F(\theta_{\tau+1/2})). \end{aligned}$$

Proof. We just show that $\theta_{\tau+1/2} = P_{\theta_\tau}(\gamma F(\theta_\tau))$, as the second part is analogous.

$$\begin{aligned} \theta_{\tau+1/2} &= \operatorname*{argmin}_{x\in\mathcal{D}\cap\Theta}\ D(x, y_{\tau+1/2}) \\ &= \operatorname*{argmin}_{x\in\mathcal{D}\cap\Theta}\ \Phi(x) - \langle \nabla\Phi(y_{\tau+1/2}), x\rangle \\ &= \operatorname*{argmin}_{x\in\mathcal{D}\cap\Theta}\ \Phi(x) - \langle \nabla\Phi(\theta_\tau) - \alpha F(\theta_\tau), x\rangle \\ &= \operatorname*{argmin}_{x\in\mathcal{D}\cap\Theta}\ \langle \alpha F(\theta_\tau), x\rangle + D(x, \theta_\tau). \end{aligned}$$

□

The mirror framework is particularly well-suited for simplex constraints i.e. when the parameter of each player is a probability vector. Such constraints usually arise in matrix games. If Θ_i is the d_i-simplex, we express the negative entropy for player i as

$$\Phi_i(\theta^i) = \sum_{j=1}^{d_i} \theta^i(j) \log \theta^i(j).$$

We can then define $\mathcal{D} \triangleq \operatorname{int} \Theta = \operatorname{int} \Theta_1 \times \cdots \times \operatorname{int} \Theta_n$ and the mirror map as

$$\Phi(\theta) = \sum_{i=1}^{n} \Phi_i(\theta^i).$$

We use this mirror map in the experiments for random monotone quadratic games (§5.1).

A.2. Link between convex games and variational inequalities

As first noted by Rosen (1965), finding a Nash equilibrium in a convex n-player game is related to solving a variational inequality (VI) problem. We consider a space of parameters $\Theta \subseteq \mathbb{R}^d$ that is compact and convex, equipped with the standard scalar product $\langle \cdot, \cdot \rangle$ in $\mathbb{R}^d$.

For convex n-player games (Ass. 1), the simultaneous (sub)gradient F (Eq. 3.1) is a monotone operator.

Definition 3. *An operator $F \colon \Theta \to \mathbb{R}^d$ is monotone if $\forall \theta, \theta' \in \Theta,\ \langle F(\theta) - F(\theta'), \theta - \theta' \rangle \geqslant 0$.*

Assuming continuity of the losses ℓ_i, we then consider the set of solutions to the following vairational inequality problem:

$$\text{Find } \ \theta_* \in \Theta \ \text{ such that } \ \langle F(\theta), \theta - \theta_* \rangle \geqslant 0 \quad \forall \theta \in \Theta. \tag{13}$$

Under Ass. 1, this set coincides with the set of Nash equilibria, and we may solve (13) instead of (1) (Rosen, 1965; Harker & Pang, 1990; Nemirovski et al., 2010). (13) indeed corresponds to the first-order necessary optimality condition applied to the loss of each player.

The quantity used to quantify the inaccuracy of a solution θ to (13) is the dual VI gap defined as $\operatorname{Err}_{\mathrm{VI}}(\theta) = \max_{u \in \Theta} \langle F(u), \theta - u \rangle$. However, the *functional Nash error* (2), also known as the (Nikaidô & Isoda, 1955) function, is the usual performance measure for convex games. We provide the convergence rates in term of functional Nash error but they also apply to the dual VI gap.

B. Proofs and mirror-setting algorithms

We start by proving Corollary 1, that derives from Juditsky et al. (2011) (§B.1). As this result is not instructive, we use the structure of the player sampling noise in (5) to obtain a stronger result in the non-smooth case (§B.3). For this, we directly modify the proof of Theorem 1 from Juditsky et al. (2011), using a few useful lemmas (§B.2). We then turn to the smooth case, for which a variance reduction mechanism proves necessary (§B.4). The proof is original, and builds upon techniques from the variance reduction literature (Defazio et al., 2014).

B.1. Proof of Corollary 1

Player sampling noise modifies the variance of the unbiased gradient estimate. Indeed, in equation (5) $\tilde{F}_i(\theta, \mathcal{P})$ is an unbiased estimate of $\nabla_i \ell_i(\theta)$, and for all $i \in [n]$

$$\mathbb{E}\left[\tilde{F}_i(\theta, \mathcal{P})\right] = \text{Prob}(i \in \mathcal{P}) \frac{n}{b} \mathbb{E}\left[g_i(\theta)\right] = \mathbb{E}\left[g_i(\theta)\right] = \nabla_i \ell_i(\theta).$$

If g_i has variance bounded by σ^2, we can bound the variance of $\tilde{F}_i(\theta, \mathcal{P})$:

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{F}_i(\theta, \mathcal{P}) - \nabla_i \ell_i(\theta)\|^2\right] &= \mathbb{E}\left[\|\tilde{F}_i(\theta, \mathcal{P}) - g_i(\theta) + g_i(\theta) - \nabla_i \ell_i(\theta)\|^2\right] \\
&\leqslant 2\mathbb{E}\left[\|\tilde{F}_i(\theta, \mathcal{P}) - g_i(\theta)\|^2\right] + 2\mathbb{E}\left[\|g_i(\theta) - \nabla_i \ell_i(\theta)\|^2\right] \\
&\leqslant 2\mathbb{E}\left[\|\tilde{F}_i(\theta, \mathcal{P}) - g_i(\theta)\|^2\right] + 2\sigma^2 \\
&= 2\mathbb{E}\left[\frac{b}{n}\left\|\left(\frac{n}{b} - 1\right) g_i(\theta)\right\|^2 + \left(1 - \frac{b}{n}\right)\|g_i(\theta)\|^2\right] + 2\sigma^2 \\
&\leqslant 2\frac{n-b}{b}\mathbb{E}\left[\|g_i(\theta)\|^2\right] + 2\sigma^2 \\
&\leqslant 2\frac{n-b}{b} G^2 + 2\sigma^2.
\end{aligned}$$

Substituting σ^2 by $2\frac{n-b}{b}G^2 + 2\sigma^2$ in equations (7) and (8) yields:

$$\mathbb{E}\left[\text{Err}_N(\hat{\theta}_{t(k)})\right] \leqslant 14n\sqrt{\frac{\Omega}{3k}\left(\frac{4n-3b}{b}G^2 + 2\sigma^2\right)} = \mathcal{O}\left(n\sqrt{\frac{\Omega}{k}\left(\frac{n}{b}G^2 + \sigma^2\right)}\right).$$

$$\mathbb{E}\left[\text{Err}_N(\hat{\theta}_{t(k)})\right] \leqslant \max\left\{\frac{7\Omega L n^{3/2}}{k}, 28n\sqrt{\frac{\Omega((\frac{n}{b}-1)G^2 + \sigma^2)}{3k}}\right\}$$

These bounds are worse than the ones in Theorem 1 when $b \ll n$. This motivates the following derivations, that yields Theorem 2 and 3.

B.2. Useful lemmas

The following two technical lemmas are proven and used in the proof of Theorem 2 of Juditsky et al. (2011).

Lemma 2. *Let z be a point in $\mathcal{X}$, let χ, η be two points in the dual E^*, let $w = P_z(\chi)$ and $r_+ = P_z(\eta)$. Then,*

$$\|w - r_+\| \leqslant \|\chi - \eta\|_* .$$

Moreover, for all $u \in E$, one has

$$D(u, r_+) - D(u, z) \leqslant \langle \eta, u - w\rangle + \frac{1}{2}\|\chi - \eta\|_*^2 - \frac{1}{2}\|w - z\|^2 .$$

Lemma 3. *Let $\xi_1, \xi_2, \ldots$ be a sequence of elements of E^*. Define the sequence $\{y_\tau\}_{\tau=0}^{\infty}$ in $\mathcal{X}$ as follows:*

$$y_\tau = P_{y_{\tau-1}}(\xi_\tau).$$

Then y_τ is a measurable function of y_0 and $\xi_1, \dots, \xi_\tau$ such that:

$$\forall u \in Z, \qquad \left\langle \sum_{\tau=1}^{t} \xi_t, y_{\tau-1} - u \right\rangle \leqslant D(u, y_0) + \frac{1}{2} \sum_{\tau=1}^{t} \|\xi_\tau\|_*^2.$$

The following lemma stems from convexity assumptions on the losses (Ass. 1) and is proven as an intermediate development of the proof of Theorem 2 of Juditsky et al. (2011).

Lemma 4. *We consider a convex n-player game with players losses ℓ_i where $i \in [n]$. Let a sequence of points $(z_\tau)_{\tau \in [t]} \in \Theta$, the stepsizes $(\gamma_\tau)_{\tau \in [t]} \in (0, \infty)$. We define the average iterate $\hat{z}_\tau = \left[\sum_{\tau=0}^{t} \gamma_\tau\right]^{-1} \sum_{\tau=0}^{t} \gamma_\tau z_\tau$. The functional Nash error evaluated in $\hat{z}_t$ is upper bounded by*

$$\mathrm{Err}_N(\hat{z}_t) \triangleq \sup_{u \in Z} \sum_{i=1}^{n} \ell_i(\hat{z}_t) - \ell_i(u^i, \hat{z}_t^{-i}) \leqslant \sup_{u \in Z} \left(\sum_{\tau=0}^{t} \gamma_\tau \right)^{-1} \sum_{\tau=0}^{t} \langle \gamma_\tau F(z_\tau), z_\tau - u \rangle.$$

The following lemma is a consequence of first-order optimality conditions.

Lemma 5. *Let $(\gamma_t)_{t \in \mathbb{N}}$ be a sequence in $(0, \infty)$ and $A, B > 0$. For any $t \in \mathbb{N}$, we define the function f_t to be*

$$f_t(\alpha) \triangleq \frac{A}{\sum_{\tau=0}^{t} \alpha\gamma_\tau} + \frac{B \sum_{\tau=0}^{t} (\alpha\gamma_\tau)^2}{\sum_{\tau=0}^{t} \alpha\gamma_\tau}.$$

Then, it attains its minimum for $\alpha > 0$ when both terms are equal. Let us call α_* the point at which the minimum is reached. Then,

$$\alpha_* = \sqrt{\frac{A}{B \sum_{\tau=0}^{t} \gamma_\tau^2}}, \quad f_t(\alpha_*) = \frac{2\sqrt{AB \sum_{\tau=0}^{t} \gamma_\tau^2}}{\sum_{\tau=0}^{t} \gamma_\tau}.$$

The next lemma describes the dual norm of the natural Pythagorean norm on a Cartesian product of Banach spaces.

Lemma 6. *Let $(X_1, \|\cdot\|_{X_1}), \dots, (X_n, \|\cdot\|_{X_n})$ be Banach spaces where for each i, $\|\cdot\|_{X_i}$ is the norm associated to X_i. The Cartesian product is $X = X_1 \times X_2 \times \cdots \times X_n$ and has a norm $\|\cdot\|_X$ defined for $y = (y_1, \dots, y_n) \in X$ as*

$$\|y\|_X \triangleq \sqrt{\sum_{i=1}^{n} \|y_i\|_{X_i}^2}.$$

It is known that $(X, \|\cdot\|_X)$ is a Banach space. Moreover, we define the dual spaces $(X_1^*, \|\cdot\|_{X_1^*}, \dots, (X_n^*, \|\cdot\|_{X_n^*})$. The dual space of X is $X^* = X_1^* \times X_2^* \times \dots \times X_n^*$ and has a norm $\|\cdot\|_{X^*}$. Then, for any $a = (a_1, \dots, a_n) \in X^*$, the following inequality holds

$$\|a\|_{X^*}^2 = \sum_{i=1}^{n} \|a_i\|_{X_i^*}^2.$$

Proof. On the one hand,

$$\|a\|_{X^*}^2 = \sup_{y \in X} \frac{|ay|^2}{\|y\|_X^2} = \sup_{y \in X} \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\|y\|_X^2} \leqslant \sup_{y \in X} \frac{\left(\sum_{i=1}^{n} \|a_i\|_{X_i^*} \|y_i\|_{X_i}\right)^2}{\|y\|_X^2},$$

and by Cauchy-Schwarz inequality

$$\|a\|_{X^*}^2 \leqslant \sup_{y \in X} \frac{\left(\sum_{i=1}^{n} \|a_i\|_{X_i^*}^2\right) \left(\sum_{i=1}^{n} \|y_i\|_{X_i}^2\right)}{\|y\|_X^2} = \sum_{i=1}^{n} \|a_i\|_{X_i^*}^2.$$

To prove the other inequality we define $Z_i = \{y_i \in X_i | \|y_i\|_X = \|a_i\|_{X_i^*}\}$.

$$\|a\|_{X^*}^2 \geqslant \sup_{y \in Z_1 \times \cdots \times Z_n} \frac{|ay|^2}{\|y\|_X^2} = \frac{\left(\sum_{i=1}^n \sup_{y_i \in Z_i} a_i y_i\right)^2}{\sum_{i=1}^n \|a_i\|_{X_i^*}^2} = \frac{\left(\sum_{i=1}^n \|a_i\|_{X_i^*}^2\right)^2}{\sum_{i=1}^n \|a_i\|_{X_i^*}^2} = \sum_{i=1}^n \|a_i\|_{X_i^*}^2.$$

□

The following two numerical lemmas will be used in Lemma 11.

Lemma 7. *The following inequality holds for any $j \in \mathbb{N}, p \in \mathbb{R}$ such that $p > 0$:*

$$\frac{(2\lceil (j+1)/2 \rceil - j)(1-p)^{2\lceil (j+1)/2 \rceil - j - 1} p + 2(1-p)^{2\lceil (j+1)/2 \rceil - j}}{p^2} \leqslant \frac{2-p}{p^2}.$$

Proof. For j even, we can write

$$(2\lceil (j+1)/2 \rceil - j)(1-p)^{2\lceil (j+1)/2 \rceil - j - 1} p + 2(1-p)^{2\lceil (j+1)/2 \rceil - j} = 2(1-p)p + 2(1-p)^2 = 2(1-p).$$

For j odd,

$$(2\lceil (j+1)/2 \rceil - j)(1-p)^{2\lceil (j+1)/2 \rceil - j - 1} p + 2(1-p)^{2\lceil (j+1)/2 \rceil - j} = p + 1 - p + 1 - p = 2 - p.$$

Since $p > 0$, $2 - p \geqslant 2(1-p)$. □

Lemma 8. *For all $|\alpha| < 1$,*

$$\sum_{s=q}^{\infty} \alpha^{s-1} s = \frac{q\alpha^{q-1}(1-\alpha) + \alpha^q}{(1-\alpha)^2}.$$

Proof.

$$\sum_{s=q}^{\infty} \alpha^{s-1} s = \left(\sum_{s=q}^{\infty} \alpha^s\right)' = \left(\frac{\alpha^q}{1-\alpha}\right)' = \frac{q\alpha^{q-1}(1-\alpha) + \alpha^q}{(1-\alpha)^2}.$$

□

B.3. Doubly-stochastic mirror-prox—Proof of Theorem 2

B.3.1. ALGORITHM

While Alg. 1 presents the doubly-stochastic algorithm in the Euclidean setting, we consider here its mirror version.

Algorithm 3 Doubly-stochastic mirror-prox

Input: initial point $\theta_0 \in \mathbb{R}^d$, stepsizes $(\gamma_\tau)_{\tau \in [t]}$, mini-batch size over the players $b \in [n]$.
for $\tau = 0, \ldots, t$ **do**
  Sample the random matrices $M_\tau, M_{\tau+1/2} \in \mathbb{R}^{d \times d}$.
  Compute $\tilde{F}_{\tau+1/2} = \frac{n}{b} \cdot M_\tau \hat{F}(\theta_\tau)$.
  Extrapolation step: $\theta_{\tau+1/2} = P_{\theta_\tau}(\gamma_\tau \tilde{F}_{\tau+1/2})$.
  Compute $\tilde{F}_{\tau+1} = \frac{n}{b} \cdot M_{\tau+1/2} \hat{F}(\theta_{\tau+1/2})$.
  Gradient step: $\theta_{\tau+1} = P_{\theta_\tau}(\gamma_\tau \tilde{F}_{\tau+1})$.
Return $\hat{\theta}_t = \left[\sum_{\tau=0}^t \gamma_\tau\right]^{-1} \sum_{\tau=0}^t \gamma_\tau \theta_\tau$.

Notation. We introduce the noisy simultaneous gradient $\hat{F}(\theta)$ defined as

$$\hat{F}(\theta) = (\hat{F}^{(1)}(\theta), \ldots, \hat{F}^{(n)}(\theta))^\top \triangleq (g_1, \ldots, g_n)^\top \in \mathbb{R}^d,$$

where g_i is a noisy unbiased estimate of $\nabla_i l_i(\theta)$ with variance bounded by σ^2. We are abusing the notation because $\hat{F}(\theta)$ is a random variable indexed by Θ and not a function, but we do so for the sake of clarity.

For our convenience, we also define the ratio $p = b/n$.

Differences with Alg. 1 The notation in Alg. 3 differs in a few aspects. First, we model the sampling over the players by using the random block-diagonal matrices M_τ and $M_{\tau+1/2}$ in $\mathbb{R}^{d\times d}$. More precisely, at each iteration, we select according to a uniform distribution b diagonal blocks and assign them to the identity matrix. Remark that we add a factor n/b in front of the random matrices to ensure the unbiasedness of the gradient estimates $\tilde{F}_\tau$ and $\tilde{F}_{\tau+1/2}$. Note that the matrices M_τ and $M_{\tau+1/2}$ are just used for the convenience of the analysis. In practice, sampling over players is not performed in this way.

Moreover, while the update in Alg. 1 involve Euclidean projections, we use the proximal mapping (11) in Alg. 3. The new notation will be used throughout the appendix.

We first proceed to the analysis of Alg. 3 in the case of non-smooth losses.

B.3.2. Convergence rate under Assumption 2a (non-smoothness)—proof of Theorem 2

The following Theorem 4 generalizes Theorem 2 to the mirror setting.

Theorem 4. *We consider a convex n-player game where Ass. 2a holds. Assume that Alg. 3 is run with constant stepsizes $\gamma_\tau = \gamma$. Let $t(k) = k/(2b)$ be the number of iterations corresponding to k gradient computations. Setting*

$$\gamma = \sqrt{\frac{2\Omega}{n\left(\frac{(3n-b)G^2}{b} + \sigma^2\right)(t(k)+1)}},$$

the rate of convergence in expectation at iteration $t(k)$ is

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_{t(k)})\right] = 4\sqrt{\frac{\Omega n\left(3G^2 n + b(\sigma^2 - G^2)\right)}{k+2b}}. \tag{14}$$

Proof. The strategy of the proof is similar to the proof of Theorem 2 and part of Theorem 1 from Juditsky et al. (2011). It consists in bounding $\sum_{\tau=0}^t \langle \gamma_\tau F(\theta_{\tau+1/2}), \theta_{\tau+1/2} - u\rangle$, which by Lemma 4 is itself a bound of the functional Nash error.

By using Lemma 2 with $z = \theta_\tau$, $\chi = \gamma_\tau \tilde{F}_{\tau+1/2}$, $\eta = \gamma_\tau \tilde{F}_{\tau+1}$ (so that $w = \theta_{\tau+1/2}$ and $r_+ = \theta_{\tau+1}$), we have for any $u \in \Theta$

$$\begin{aligned}\langle \gamma_\tau \tilde{F}_{\tau+1}, \theta_{\tau+1/2} - u\rangle + D(u, \theta_{\tau+1}) - D(u,\theta_\tau) &\leqslant \frac{\gamma_\tau^2}{2}\|\tilde{F}_{\tau+1} - \tilde{F}_{\tau+1/2}\|_*^2 - \frac{1}{2}\|\theta_{\tau+1/2} - \theta_\tau\|_*^2 \\ &\leqslant \frac{\gamma_\tau^2}{2}\|\tilde{F}_{\tau+1} - \tilde{F}_{\tau+1/2}\|_*^2. \end{aligned} \tag{15}$$

When summing up from $\tau = 0$ to $\tau = t$ in equation (15), we get

$$\sum_{\tau=0}^t \langle \gamma_\tau \tilde{F}_{\tau+1}, \theta_{\tau+1/2} - u\rangle \leqslant D(u,\theta_0) - D(u,\theta_{t+1}) + \sum_{\tau=0}^t \frac{\gamma_\tau^2}{2}\|\tilde{F}_{\tau+1} - \tilde{F}_{\tau+1/2}\|_*^2. \tag{16}$$

By decomposing the right-hand side (16), we obtain

$$\begin{aligned}\sum_{\tau=0}^t \langle \gamma_\tau F(\theta_{\tau+1/2}), \theta_{\tau+1/2} - u\rangle &\leqslant D(u,\theta_0) - D(u,\theta_{t+1}) + \sum_{\tau=0}^t \frac{\gamma_\tau^2}{2}\|\tilde{F}_{\tau+1} - \tilde{F}_{\tau+1/2}\|_*^2 \\ &\quad + \sum_{\tau=0}^t \left\langle \gamma_\tau (F(\theta_{\tau+1/2}) - \tilde{F}_{\tau+1}), \theta_{\tau+1/2} - u \right\rangle \\ &\leqslant \Omega + \sum_{\tau=0}^t \frac{\gamma_\tau^2}{2}\|\tilde{F}_{\tau+1} - \tilde{F}_{\tau+1/2}\|_*^2 \\ &\quad + \sum_{\tau=0}^t \gamma_\tau \left\langle F(\theta_{\tau+1/2}) - \tilde{F}_{\tau+1}, \theta_{\tau+1/2} - y_\tau \right\rangle \\ &\quad + \sum_{\tau=0}^t \gamma_\tau \left\langle F(\theta_{\tau+1/2}) - \tilde{F}_{\tau+1}, y_\tau - u \right\rangle, \end{aligned} \tag{17}$$

where we used $D(u, \theta_0) \leqslant \Omega$ and defined $y_{\tau+1} = P_{y_\tau}(\gamma_\tau \Delta_\tau)$ with $y_0 = \theta_0$ and $\Delta_\tau = F(\theta_{\tau+1/2}) - \tilde{F}_{\tau+1}$. So far, we followed the same steps as Juditsky et al. (2011). We aim at bounding the left-hand side of equation (17) in expectation. To this end, we will now bound the expectation of each of the right-hand side terms. These steps represent the main difference with the analysis by Juditsky et al. (2011).

We first define the filtrations $\mathcal{F}_\tau = \sigma(\theta_{\tau'} : \tau' \leqslant \tau + 1/2)$ and $\mathcal{F}_\tau = \sigma(\theta_{\tau'} : \tau' \leqslant \tau)$. We now bound the third term on the right-hand side of (17) in expectation.

$$
\begin{aligned}
\mathbb{E}\left[\|\tilde{F}_{\tau+1} - \tilde{F}_{\tau+1/2}\|_*^2\right] &\leqslant 2\left(\mathbb{E}\left[\|\tilde{F}_{\tau+1}\|_*^2\right] + \mathbb{E}\left[\|\tilde{F}_{\tau+1/2}\|_*^2\right]\right) \\
&= \frac{2}{p^2}\left(\mathbb{E}\left[\mathbb{E}\left[\|M_{\tau+1/2}\hat{F}(\theta_{\tau+1/2})\|_*^2|\mathcal{F}_\tau\right]\right] + \mathbb{E}\left[\mathbb{E}\left[\|M_\tau \hat{F}(\theta_\tau)\|_*^2|\mathcal{F}'_\tau\right]\right]\right) \\
&= \frac{2}{p^2}\sum_{i=1}^n \left(\mathbb{E}\left[\mathbb{E}\left[\|M^{(i)}_{\tau+1/2}\hat{F}^{(i)}(\theta_{\tau+1/2})\|_*^2|\mathcal{F}_\tau\right]\right]\right. \\
&\qquad \left. + \mathbb{E}\left[\mathbb{E}\left[\|M^{(i)}_{\tau}\hat{F}^{(i)}(\theta_{\tau})\|_*^2|\mathcal{F}'_\tau\right]\right]\right) \\
&\leqslant \frac{2}{p}\sum_{i=1}^n \mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{\tau+1/2})\|_*^2\right] + \mathbb{E}\left[\|\hat{F}^{(i)}(\theta_\tau)\|_*^2\right] \\
&\leqslant \frac{4nG^2}{p},
\end{aligned}
\tag{18}
$$

where we used $\|a+b\|_*^2 \leqslant 2\|a\|_*^2 + 2\|b\|_*^2$ in the first inequality and applied Lemma 6 in the second equality. Now, we compute the expectation of the fourth term of equation (17).

$$
\begin{aligned}
&\mathbb{E}\left[\gamma_\tau \sum_{\tau=0}^t \left\langle F(\theta_{\tau+1/2}) - \tilde{F}_{\tau+1}, \theta_{\tau+1/2} - y_\tau \right\rangle\right] \\
&= \mathbb{E}\left[\sum_{\tau=0}^t \mathbb{E}\left[\left\langle \gamma_\tau\left(I - \frac{M_{\tau+1/2}}{p}\right)\hat{F}(\theta_{\tau+1/2}), \theta_{\tau+1/2} - y_\tau\right\rangle \middle| \mathcal{F}_\tau\right]\right] \\
&= \mathbb{E}\left[\sum_{\tau=0}^t \left\langle \gamma_\tau \mathbb{E}\left[\left(I - \frac{M_{\tau+1/2}}{p}\right)\middle|\mathcal{F}_\tau\right]\mathbb{E}\left[\hat{F}(\theta_{\tau+1/2})\middle|\mathcal{F}_\tau\right], \theta_{\tau+1/2} - y_\tau\right\rangle\right] \\
&= 0,
\end{aligned}
\tag{19}
$$

where we used the independence property of the random variables in the second equality and $\mathbb{E}[\frac{k}{n} \cdot M_{\tau+1/2}] = I_d$ in the third equality. Regarding the fifth term of (17), by using the sequences $\{y_\tau\}$ and $\{\xi_\tau = \gamma_\tau \Delta_\tau\}$ in Lemma 3 (as done in Juditsky et al. (2011)), we obtain:

$$
\sum_{\tau=0}^t \langle \gamma_\tau \Delta_\tau, y_\tau - u\rangle \leqslant D(u, \theta_0) + \sum_{\tau=0}^t \frac{\gamma_\tau^2}{2}\|\Delta_\tau\|_*^2 \leqslant \Omega + \sum_{\tau=0}^t \frac{\gamma_\tau^2}{2}\|F(\theta_{\tau+1/2}) - \tilde{F}_{\tau+1}\|_*^2. \tag{20}
$$

We now bound the expectation of $\|F(\theta_{\tau+1/2}) - \tilde{F}_{\tau+1}\|_*^2$ using the filtration $\mathcal{F}_\tau$. By using Lemma 6 in the first equality,

$\|a+b\|_*^2 \leqslant 2\|a\|_*^2 + 2\|b\|_*^2$ in the second inequality and the bound on the variance (Ass. 3) in the third inequality, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\|F(\theta_{\tau+1/2}) - \tilde{F}_{\tau+1}\|_*^2\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[\|F^{(i)}(\theta_{\tau+1/2}) - \tilde{F}_{\tau+1}^{(i)}\|_*^2\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[\left\|F^{(i)}(\theta_{\tau+1/2}) - \frac{M_{\tau+1}^{(i)}}{p}\hat{F}^{(i)}(\theta_{\tau+1/2})\right\|_*^2\right] \\
&\leqslant \sum_{i=1}^n 2\mathbb{E}\left[\left\|\left(I - \frac{M_{\tau+1}^{(i)}}{p}\right)\hat{F}^{(i)}(\theta_{\tau+1/2})\right\|_*^2\right] + \sum_{i=1}^n 2\mathbb{E}\left[\left\|F^{(i)}(\theta_{\tau+1/2}) - \hat{F}^{(i)}(\theta_{\tau+1/2})\right\|_*^2\right] \\
&\leqslant \sum_{i=1}^n 2\mathbb{E}\left[p\left\|\frac{p-1}{p}\hat{F}^{(i)}(\theta_{\tau+1/2})\right\|_*^2 + (1-p)\|\hat{F}^{(i)}(\theta_{\tau+1/2})\|_*^2\right] + 2n\sigma^2 \\
&= \sum_{i=1}^n 2\left(1 - p + \frac{(1-p)^2}{p}\right)\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{\tau+1/2})\|_*^2\right] + 2n\sigma^2 \\
&= \sum_{i=1}^n 2\left(\frac{1}{p} - 1\right)\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{\tau+1/2})\|_*^2\right] + 2n\sigma^2 \\
&\leqslant \frac{2nG^2(1-p)}{p} + 2n\sigma^2. \qquad (21)
\end{aligned}
$$

Therefore, by taking the expectation in equation (17) and plugging (18), (19), (20) and (21), we finally get:

$$
\mathbb{E}\left[\sup_{u\in Z}\sum_{\tau=0}^t \langle \gamma_\tau F(\theta_{\tau+1/2}), \theta_{\tau+1/2} - u\rangle\right] \leqslant 2\Omega + \sum_{\tau=0}^t \gamma_\tau^2 n\left(\frac{(3-p)G^2}{p} + \sigma^2\right) \qquad (22)
$$

Applying Lemma 4 to equation (22) yields an upper bound on the functional Nash error shown in equation (23).

$$
\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_t)\right] \leqslant \left(\sum_{\tau=0}^t \gamma_\tau\right)^{-1}\left(2\Omega + \sum_{\tau=0}^t \gamma_\tau^2 n\left(\frac{(3n-b)G^2}{b} + \sigma^2\right)\right). \qquad (23)
$$

Now, let us set γ_t constant and optimize the bound (23). Namely, we apply Lemma 5 setting $\gamma_\tau = 1$ for all $\tau \in [t]$, $A = 2\Omega$ and

$$
B = n\left(\frac{(3n-b)G^2}{b} + \sigma^2\right).
$$

The optimal value for γ_τ is

$$
\gamma_\tau = \gamma = \sqrt{\frac{2\Omega}{n\left(\frac{(3n-b)G^2}{b} + \sigma^2\right)(t+1)}}.
$$

and the optimal value of the bound is

$$
\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_t)\right] \leqslant \sqrt{\frac{8\Omega n\left(\frac{(3n-b)G^2}{b} + \sigma^2\right)}{t+1}}. \qquad (24)
$$

The number of iterations t can be expressed in terms of the number of gradient computations k as $t(k) = k/(2b)$. Plugging this expression into (24), we get

$$
\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_{t(k)})\right] = \sqrt{\frac{8\Omega n\left(\frac{3G^2 n}{b} + \sigma^2 - G^2\right)}{\frac{k}{2b} + 1}},
$$

which yields equation (14) after simplification. □

Remark 1. *For constant stepsizes, equation (24) implies that with an appropriate choice of t and γ we can achieve a value of the Nash error arbitrarily close to zero at time t. However, from Equation 23 we see that constant stepsizes do not ensure convergence; the bound has a strictly positive limit. Stepsizes decreasing as $1/\sqrt{\tau}$ do ensure convergence, although we do not make a detailed analysis of this case.*

Remark 2. *Without using any variance reduction technique, the smooth losses assumption Ass. 2b does not yield a significant improvement over the bound from Theorem 4. We do not include the analysis of this case.*

B.4. Doubly-stochastic mirror-prox with variance reduction—Proof of Theorem 3

B.4.1. Algorithm

With the same notations as above, we present a version of Alg. 1 with variance reduction in the mirror framework.

Algorithm 4 Mirror prox with variance reduced player randomness

Input: initial point $\theta_0 \in \mathbb{R}^d$, stepsizes $(\gamma_\tau)_{\tau\in[t]}$, mini-batch size over the players $b \in [n]$.
Set $R_0 = \hat{F}(\theta_0) \in \mathbb{R}^d$
for $\tau = 0, \dots, t$ **do**
  Sample the random matrices $M_\tau, M_{\tau+1/2} \in \mathbb{R}^{d\times d}$.
  Compute $\tilde{F}_{\tau+1/2} = R_\tau + \frac{n}{b} M_\tau(\hat{F}(\theta_\tau) - R_\tau)$
  Set $R_{\tau+1/2} = R_\tau + M_\tau(\hat{F}(\theta_\tau) - R_\tau)$
  Extrapolation step: $\theta_{\tau+1/2} = P_{\theta_\tau}(\gamma_\tau \tilde{F}_{\tau+1/2})$.
  Compute $\tilde{F}_{\tau+1} = R_{\tau+1/2} + \frac{n}{b} M_{\tau+1/2}(\hat{F}(\theta_{\tau+1/2}) - R_{\tau+1/2})$
  Set $R_{\tau+1} = R_{\tau+1/2} + M_{\tau+1/2}(\hat{F}(\theta_{\tau+1/2}) - R_{\tau+1/2})$
  Extra-gradient step: $\theta_{\tau+1} = P_{\theta_\tau}(\gamma_\tau \tilde{F}_{\tau+1})$.
Return $\hat{\theta}_t = \left[\sum_{\tau=0}^t \gamma_\tau\right]^{-1} \sum_{\tau=0}^t \gamma_\tau \theta_\tau$.

$\hat{F}(\theta)$ is defined as in Alg. 3. The random matrices $M_\tau, M_{\tau+1/2}$ are also sampled the same way.

In Alg. 4, we leverage information from a table $(R_\tau)_{\tau\in[t]}$ to produce doubly-stochastic simultaneous gradient estimates with lower variance than in Alg. 3. The table R_τ is updated when possible.

The following Theorem 5 generalizes Theorem 3 in the mirror setting.

Theorem 5. *Assume that for all i between 1 and n, the gradients $\nabla_i \ell_i$ are L-Lipschitz (Ass. 2b) and bounded by G_i (Ass. 2a). Assume Alg. 4 is run for $t(k)$ iterations with constant stepsizes $\gamma_\tau = \gamma$, with $\gamma \leqslant \sqrt{\frac{2}{3nL^2}}$ defined as*

$$\gamma \triangleq \sqrt{\frac{2\Omega}{7\sigma^2 n(t(k)+1)}}$$

where $p \triangleq b/n$, k is the number of gradient computations and $t(k) = k/(2b)$ is the corresponding number of iterations. Then, the convergence rate in expectation at iteration $t(k)$ is

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_{t(k)})\right] \leqslant 4\sqrt{\frac{7\Omega\sigma^2 bn}{k+2b}} + h(\Omega, \sigma^2, G, L, n, b)\frac{1}{(k+2b)^{3/2}},$$

where

$$h(\Omega, \sigma^2, G, L, n, b) \triangleq 42L^2n^2(\sigma^2 + G^2)\left(1 - \frac{b}{n}\right)\left(2\left(\frac{n}{b}\right)^{3/2} - \left(\frac{n}{b}\right)^{1/2}\right)\left(\frac{4\Omega}{7\sigma^2}\right)^{3/2}.$$

Outline of the proof of Theorem 5. We prove intermediate results (Lemma 9, Lemma 10, Lemma 11) and use the same framework as in the smooth case, based on the work of Juditsky et al. (2011).

Definition 4. *For a given j and i (which we omit), let us define K_j as the random variable indicating the highest $q \in \mathbb{N}$ strictly lower than j such that $M^{(i)}_{q/2}$ is the identity (and $K_j = 0$ if there exists no such q).*

In other words, K_j is the last step q before j at which the sequence $(R^{(i)}_{q/2})_{q\in\mathbb{N}}$ was updated with a new value $\hat{F}^{(i)}(\theta_{q/2})$. That is, $R_{j/2,i} = \hat{F}^{(i)}(\theta_{K_j/2})$.

Lemma 9. *For a given j, $j - K_j$ is a random variable that has a geometric distribution with parameter p and support between 1 and j, i.e., for all q such that $j - 1 \geqslant q \geqslant 1$,*

$$P(K_j = q) = p(1-p)^{j-1-q},$$

and $P(K_j = 0) = 1 - \sum_{q=1}^{j-1} P(K_j = q) = (1-p)^{j-1}$.

Proof. $M^{(i)}_{q/2}$ is Bernoulli distributed with parameter p among zero and the identity, for all q. □

Lemma 10. *The following equalities hold:*

$$\mathbb{E}\left[\|F^{(i)}(\theta_\tau) - \tilde{F}^{(i)}_{\tau+1/2}\|^2_\star\right] = \frac{2(1-p)}{p}\mathbb{E}\left[\|R^{(i)}_\tau - \hat{F}^{(i)}(\theta_\tau)\|^2_\star\right] + 2\sigma^2,$$
$$\mathbb{E}\left[\|\tilde{F}^{(i)}_{\tau+1} - F^{(i)}(\theta_{\tau+1/2})\|^2_\star\right] = \frac{2(1-p)}{p}\mathbb{E}\left[\|R^{(i)}_{\tau+1/2} - \hat{F}^{(i)}(\theta_{\tau+1/2})\|^2_\star\right] + 2\sigma^2.$$

Proof. Using the conditional expectation with respect to the filtration up to w_τ,

$$\begin{aligned}
&\mathbb{E}\left[\|\tilde{F}^{(i)}_{\tau+1} - F^{(i)}(\theta_{\tau+1/2})\|^2_\star\right] \\
&= 2\mathbb{E}\left[\left\|R^{(i)}_{\tau+1/2} + \frac{M^{(i)}_{\tau+1/2}}{p}(\hat{F}^{(i)}(\theta_{\tau+1/2}) - R^{(i)}_{\tau+1/2}) - \hat{F}^{(i)}(\theta_{\tau+1/2})\right\|^2_\star\right] \\
&+ 2\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{\tau+1/2}) - F^{(i)}(\theta_{\tau+1/2})\|^2_\star\right] \\
&= 2\mathbb{E}\left[\left\|\left(I - \frac{M^{(i)}_{\tau+1/2}}{p}\right)(R^{(i)}_{\tau+1/2} - \hat{F}^{(i)}(\theta_{\tau+1/2}))\right\|^2_\star\right] + 2\sigma^2 \\
&= 2\mathbb{E}\left[p\left\|\frac{p-1}{p}(R^{(i)}_{\tau+1/2} - \hat{F}^{(i)}(\theta_{\tau+1/2}))\right\|^2_\star + (1-p)\|R^{(i)}_{\tau+1/2} - \hat{F}^{(i)}(\theta_{\tau+1/2})\|^2_\star\right] + 2\sigma^2 \\
&= 2\left(1 - p + \frac{(1-p)^2}{p}\right)\mathbb{E}\left[\|R^{(i)}_{\tau+1/2} - \hat{F}^{(i)}(\theta_{\tau+1/2})\|^2_\star\right] + 2\sigma^2 \\
&= \frac{2(1-p)}{p}\mathbb{E}\left[\|R^{(i)}_{\tau+1/2} - \hat{F}^{(i)}(\theta_{\tau+1/2})\|^2_\star\right] + 2\sigma^2.
\end{aligned}$$

The second equality is derived analogously. □

Let us define the change of variables $j = 2\tau$. Parametrized by j, the sequences that we are dealing with are $(M^{(i)}_{j/2})_{j\in\mathbb{N}}$, $(R^{(i)}_{j/2})_{j\in\mathbb{N}}$ and $(\theta_{j/2})_{j\in\mathbb{N}}$. In this scope i is a fixed integer between 1 and n.

Lemma 11. *Assume that $(\gamma_\tau)_{\tau\in\mathbb{N}}$ is non-increasing and upper bounded by γ. Then, the following holds:*

$$\sum_{\tau=0}^{t}\gamma_\tau^2\mathbb{E}\left[\|R^{(i)}_\tau - \hat{F}^{(i)}(\theta_\tau)\|^2_\star\right] \leqslant 6L^2\gamma^4 n(\sigma^2 + G^2)\left(\frac{2}{p^2} - \frac{1}{p}\right)(t+1)$$
$$\sum_{\tau=0}^{t}\gamma_\tau^2\mathbb{E}\left[\|R^{(i)}_{\tau+1/2} - \hat{F}^{(i)}(\theta_{\tau+1/2})\|^2_\star\right] \leqslant 6L^2\gamma^4 n(\sigma^2 + G^2)\left(\frac{2}{p^2} - \frac{1}{p}\right)(t+1)$$

Proof. We can write

$$\begin{aligned}
\mathbb{E}\left[\|R_\tau^{(i)} - \hat{F}^{(i)}(\theta_\tau)\|_\star^2\right] &= \mathbb{E}\left[\|R_{2\tau/2}^{(i)} - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\|R_{2\tau/2}^{(i)} - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2 \middle| K_{2\tau}\right]\right] \\
&= \sum_{q=0}^{2\tau-1} P(K_{2\tau}=q)\mathbb{E}\left[\|R_{2\tau/2}^{(i)} - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2 \middle| K_{2\tau}=q\right] \\
&= \sum_{q=1}^{2\tau-1} p(1-p)^{2\tau-1-q}\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{q/2}) - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2 \middle| K_{2\tau}=q\right] \\
&\quad + (1-p)^{2\tau-1}\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_0) - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2 \middle| K_{2\tau}=0\right].
\end{aligned}$$

For $0 \leqslant q \leqslant 2\tau - 1$,

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{q/2}) - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2 \middle| K_{2\tau}=q\right] \\
&= \mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{q/2}) - F^{(i)}(\theta_{q/2}) + F^{(i)}(\theta_{q/2}) - F^{(i)}(\theta_{2\tau/2}) + F^{(i)}(\theta_{2\tau/2}) - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2 \middle| K_{2\tau}=q\right] \\
&= 3\mathbb{E}\left[\|F^{(i)}(\theta_{q/2}) - F^{(i)}(\theta_{2\tau/2})\|_\star^2 \middle| K_{2\tau}=q\right] + 6\sigma^2 \\
&\leqslant 3L^2\mathbb{E}\left[\|\theta_{q/2} - \theta_{2\tau/2}\|_\star^2 \middle| K_{2\tau}=q\right] + 6\sigma^2 \\
&\leqslant 3L^2\mathbb{E}\left[\left\|\sum_{j=q}^{2\tau-1}\gamma_{j/2}\hat{F}(\theta_{q/2})\right\|_\star^2\right] + 6\sigma^2 = 3L^2\left(\sum_{j=q}^{2\tau-1}\gamma_{j/2}\right)^2\mathbb{E}\left[\left\|\hat{F}(\theta_{q/2})\right\|_\star^2\right] + 6\sigma^2 \\
&= 3L^2\left(\sum_{j=q}^{2\tau-1}\gamma_{j/2}\right)^2 2n(\sigma^2+G^2) + 6\sigma^2
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sum_{\tau=0}^{t}\gamma_\tau^2\mathbb{E}\left[\|R_\tau^{(i)} - \hat{F}^{(i)}(\theta_\tau)\|_\star^2\right] \\
&= \sum_{\tau=0}^{t}\gamma_\tau^2\sum_{q=1}^{2\tau-1}p(1-p)^{2\tau-1-q}\left(3L^2\left(\sum_{j=q}^{2\tau-1}\gamma_{j/2}\right)^2 2n(\sigma^2+G^2) + 6\sigma^2\right) \\
&\quad + \gamma_\tau^2(1-p)^{2\tau-1}\left(3L^2\left(\sum_{j=0}^{2\tau-1}\gamma_{j/2}\right)^2 2n(\sigma^2+G^2) + 6\sigma^2\right) \\
&\leqslant \sum_{\tau=0}^{t}\gamma^2\sum_{q=1}^{2\tau-1}p(1-p)^{2\tau-1-q}\left(3L^2\gamma^2(2\tau-q)^2 2n(\sigma^2+G^2) + 6\sigma^2\right) \\
&\quad + \gamma^2(1-p)^{2\tau-1}\left(3L^2\gamma^2(2\tau)^2 2n(\sigma^2+G^2) + 6\sigma^2\right)
\end{aligned}$$

We simplify the right hand side. On the one hand,

$$6\sigma^2\gamma^2\sum_{\tau=0}^{t}\sum_{q=1}^{2\tau-1}p(1-p)^{2\tau-1-q} + (1-p)^{2\tau-1} = 6\sigma^2\gamma^2(t+1)$$

On the other hand, we compute

$$\begin{aligned}
&\sum_{q=1}^{2\tau-1} p(1-p)^{2\tau-1-q}\left(2\tau-q\right)^2+(1-p)^{2\tau-1}\left(2\tau\right)^2 \leqslant \sum_{q=1}^{\infty} p(1-p)^{q-1}q^2\\
&= p\left((1-p)\sum_{q=1}^{\infty}(1-p)^{q-2}q(q-1)+\sum_{q=1}^{\infty}(1-p)^{q-1}q\right)\\
&= p\left((1-p)\frac{d^2}{dp^2}\left(\sum_{q=1}^{\infty}(1-p)^q\right)-\frac{d}{dp}\left(\sum_{q=1}^{\infty}(1-p)^q\right)\right)\\
&= p\left((1-p)\frac{d^2}{dp^2}\left(\frac{1-p}{p}\right)-\frac{d}{dp}\left(\frac{1-p}{p}\right)\right)\\
&= p\left(\frac{2(1-p)}{p^3}+\frac{1}{p^2}\right)=\frac{2(1-p)}{p^2}+\frac{1}{p}=\frac{2}{p^2}-\frac{1}{p}
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sum_{\tau=0}^{t}\gamma_\tau^2\mathbb{E}\left[\|R_\tau^{(i)}-\hat{F}^{(i)}(\theta_\tau)\|_\star^2\right]\\
&= 6L^2\gamma^4 n(\sigma^2+G^2)\left(\sum_{\tau=0}^{t}\sum_{q=1}^{2\tau-1} p(1-p)^{2\tau-1-q}\left(2\tau-q\right)^2+(1-p)^{2\tau-1}\left(2\tau\right)^2\right)\\
&\leqslant 6L^2\gamma^4 n(\sigma^2+G^2)\left(\frac{2}{p^2}-\frac{1}{p}\right)(t+1)
\end{aligned}$$

The same argument works for the second inequality. □

Proof of Theorem 5. We rewrite equation (17):

$$\begin{aligned}
&\langle\gamma_\tau\tilde{F}_{\tau+1},\theta_{\tau+1/2}-u\rangle+D(u,\theta_{\tau+1})-D(u,\theta_\tau)\\
&\leqslant\frac{\gamma_\tau^2}{2}\|\tilde{F}_{\tau+1}-\tilde{F}_{\tau+1/2}\|_\star^2-\frac{1}{2}\|\theta_{\tau+1/2}-\theta_\tau\|^2\\
&\leqslant\frac{3\gamma_\tau^2}{2}\|\tilde{F}_{\tau+1}-F(\theta_{\tau+1/2})\|_\star^2+\frac{3\gamma_\tau^2}{2}\|F(\theta_\tau)-\tilde{F}_{\tau+1/2}\|_\star^2+\frac{3\gamma_\tau^2}{2}\|F(\theta_{\tau+1/2})-F(\theta_\tau)\|_\star^2\\
&-\frac{1}{2}\|\theta_{\tau+1/2}-\theta_\tau\|^2.
\end{aligned}$$

We rewrite equation (20). We have $\Delta_\tau = F(\theta_{\tau+1/2})-\tilde{F}_{\tau+1}$ and $y_{\tau+1}=P_{y_\tau}(\gamma_\tau\Delta_\tau)$ with $y_0=\theta_0$.

$$\begin{aligned}
\sum_{\tau=0}^{t}\langle\gamma_\tau\Delta_\tau,y_\tau-u\rangle &\leqslant D(u,\theta_0)+\sum_{\tau=0}^{t}\frac{\gamma_\tau^2}{2}\|\Delta_\tau\|_\star^2\\
&= D(u,\theta_0)+\sum_{\tau=0}^{t}\frac{\gamma_\tau^2}{2}\|F(\theta_{\tau+1/2})-\tilde{F}_{\tau+1}\|_\star^2. \qquad (25)
\end{aligned}$$

Using equation (25) and the analogous equation to (19), we reach the following inequality:

$$\begin{aligned}
&\mathbb{E}\left[\sup_{u\in Z}\sum_{\tau=0}^{t}\langle\gamma_\tau F(\theta_{\tau+1/2}),\theta_{\tau+1/2}-u\rangle\right]\leqslant\mathbb{E}\left[\sup_{u\in Z}2D(u,\theta_0)-D(u,\theta_{t+1})-\sum_{\tau=0}^{t}\frac{1}{2}\|\theta_{\tau+1/2}-\theta_\tau\|_2^2\right]\\
&+\mathbb{E}\left[\sum_{\tau=0}^{t}2\gamma_\tau^2\|\tilde{F}_{\tau+1}-F(\theta_{\tau+1/2})\|_\star^2+\frac{3\gamma_\tau^2}{2}\|F(\theta_\tau)-\tilde{F}_{\tau+1/2}\|_\star^2+\frac{3\gamma_\tau^2}{2}\|F(\theta_{\tau+1/2})-F(\theta_\tau)\|_\star^2\right]
\end{aligned}$$

We use Lemma 9 and Lemma 11 and the Lipschitz property of F to bound the right hand side by:

$$
\begin{aligned}
&2\Omega + \mathbb{E}\left[\sum_{\tau=0}^{t}\left(\frac{3\gamma_\tau^2}{2}nL^2 - 1\right)\|\theta_{\tau+1/2}) - \theta_\tau\|_2^2\right] \\
&+ \mathbb{E}\left[\sum_{\tau=0}^{t} 2\gamma_\tau^2\left(2n\sigma^2 + \sum_{i=1}^{n}\frac{2(1-p)}{p}\mathbb{E}\left[\|R_\tau^{(i)} - \hat{F}^{(i)}(\theta_\tau)\|_\star^2\right]\right) + \frac{3\gamma_\tau^2}{2}\left(2n\sigma^2 + \sum_{i=1}^{n}\frac{2(1-p)}{p}\mathbb{E}\left[\|R_{\tau+1/2}^{(i)} - \hat{F}^{(i)}(\theta_{\tau+1/2})\|_\star^2\right]\right)\right] \\
&\leqslant 2\Omega + \mathbb{E}\left[\sum_{\tau=0}^{t}\left(\frac{3\gamma^2}{2}nL^2 - 1\right)\|\theta_{\tau+1/2}) - \theta_\tau\|_2^2\right] + 7\sigma^2 n\gamma^2(t+1) \\
&+ \frac{42(1-p)}{p}L^2\gamma^4 n^2(\sigma^2+G^2)\left(\frac{2}{p^2} - \frac{1}{p}\right)(t+1)
\end{aligned}
$$

If we take $\gamma > 0$ such that $\frac{3\gamma^2}{2}nL^2 < 1$, we can upper bound this expression by

$$
2\Omega + 7\sigma^2 n\gamma^2(t+1) + \frac{42(1-p)}{p}L^2\gamma^4 n^2(\sigma^2+G^2)\left(\frac{2}{p^2} - \frac{1}{p}\right)(t+1)
$$

Thus,

$$
\mathbb{E}\left[\sup_{u\in Z}\sum_{\tau=1}^{t}\langle\gamma_\tau F(\theta_{\tau+1/2}), \theta_{\tau+1/2} - u\rangle\right] \leqslant 2\Omega + 7\sigma^2 n\gamma^2(t+1) + \frac{42(1-p)}{p}L^2\gamma^4 n^2(\sigma^2+G^2)\left(\frac{2}{p^2} - \frac{1}{p}\right)(t+1).
$$

By Lemma 4, if we pick $\gamma_t = \gamma$ for all $t \geqslant 0$, we obtain

$$
\mathrm{Err}_N(\hat{\theta}_t) \leqslant \frac{2\Omega}{\gamma(t+1)} + 7\sigma^2 n\gamma + \frac{42(1-p)}{p}L^2 n^2(\sigma^2+G^2)\left(\frac{2}{p^2} - \frac{1}{p}\right)\gamma^3 =: A/\gamma + B\gamma + C\gamma^3 \tag{26}
$$

For $A, B, C > 0$, the value of $\gamma > 0$ minimizing $A/\gamma + B\gamma + C\gamma^3$ is

$$
\gamma = \sqrt{\frac{-B + \sqrt{B^2 + 12AC}}{6C}}
$$

Notice that when $t >> 0$, $A << B, C$. If we take the first-order Taylor approximation $\sqrt{1+x} \approx 1 + x/2$ of the square root around 1, we obtain

$$
\gamma = \sqrt{\frac{-B + B\sqrt{1 + \frac{12AC}{B^2}}}{6C}} \approx \sqrt{\frac{-B + B\left(1 + \frac{6AC}{B^2}\right)}{6C}} = \sqrt{\frac{\frac{6AC}{B}}{6C}} = \sqrt{\frac{A}{B}} = \sqrt{\frac{2\Omega}{7\sigma^2 n(t+1)}}
$$

If we plug this expression into (26), we get

$$
2\sqrt{AB} + C\left(\frac{A}{B}\right)^{3/2} = 2\sqrt{\frac{14\Omega\sigma^2 n}{t+1}} + 42L^2n^2(\sigma^2+G^2)(1-p)\left(\frac{2}{p^3} - \frac{1}{p^2}\right)\left(\frac{2\Omega}{7\sigma^2 n(t+1)}\right)^{3/2}
$$

Now, let us plug $p = b/n$ and $t(k) = k/(2b)$:

$$
\begin{aligned}
&2\sqrt{\frac{14\Omega\sigma^2 n}{\frac{k}{2b}+1}} + 42L^2n^2(\sigma^2+G^2)\left(1 - \frac{b}{n}\right)\left(\frac{2n^3}{b^3} - \frac{n^2}{b^2}\right)\left(\frac{2\Omega}{7\sigma^2 n(\frac{k}{2b}+1)}\right)^{3/2} \\
&= 4\sqrt{\frac{7\Omega\sigma^2 bn}{k+2b}} + 42L^2n^2(\sigma^2+G^2)\left(1 - \frac{b}{n}\right)\left(2\left(\frac{n}{b}\right)^{3/2} - \left(\frac{n}{b}\right)^{1/2}\right)\left(\frac{4\Omega}{7\sigma^2(k+2b)}\right)^{3/2} \\
&= 4\sqrt{\frac{7\Omega\sigma^2 bn}{k+2b}} + h(\Omega, \sigma^2, G, L, n, b)\frac{1}{(k+2b)^{3/2}}
\end{aligned}
$$

□

C. Spectral convergence analysis for non-constrained 2-player games

We observed in the experimental section that player sampling tended to be empirically faster than full extra-gradient, and that cyclic sampling had a tendency to be better than random sampling.

To have more insight on this finding, let us study a simplified version of the random two-player quadratic games. Let $A \in \mathbb{R}^{2d\times 2d}$ be formed by stacking the matrices $A_i \in \mathbb{R}^{d\times 2d}$ for each $i \in [d]$. We assume that A is invertible and has a positive semidefinite symmetric part. For $i \in \{1, 2\}$, we define the loss of the i-th player ℓ_i as

$$\ell_i(\theta^i, \theta^{-i}) = {\theta^i}^\top A_i \theta - \frac{1}{2} {\theta^i}^\top A_{ii} \theta^i,$$

where $A_{ii} \in \mathbb{R}^d$ and $\theta_i \in \mathbb{R}^{d_i}$. Contrary to the random quadratic games setting in §5.1, we do not enforce here any parameter constraints nor regularization. Therefore, this places us in the extra-gradient (Euclidean) setting. We restrict our attention to the non-noisy regime.

C.1. Recursion operator for the different sampling schemes

We study the "algorithm operator" $\mathcal{A}$ that appears in the recursion $\theta_{k+4} = \mathcal{A}(\theta_k)$ for the different sampling schemes. k is the number of gradient computations. We consider steps of 4 evaluation as this corresponds to a single iteration of full extra-gradient.

Full extrapolation and update. We have $\nabla_i \ell_i(\theta) = A_i\theta$. Since A is invertible, $\theta = 0$ is the only Nash equilibrium. The full extra-gradient updates with constant stepsize are

$$\begin{cases} \theta^{\text{full}}_{k+2} = \theta^{\text{full}}_k - \gamma A \theta^{\text{full}}_k, \\ \theta^{\text{full}}_{k+4} = \theta^{\text{full}}_k - \gamma A \theta^{\text{full}}_{k+2}. \end{cases} \tag{27}$$

By introducing $\mathcal{A}^{(\gamma)}_{\text{full}} := I - \gamma A + \gamma^2 A^2$, (27) is simply $\theta^{\text{full}}_{k+4} = \mathcal{A}^{(\gamma)}_{\text{full}} \theta^{\text{full}}_k$.

Cyclic sampling. Defining the matrices $M_1, M_2 \in \mathbb{R}^{2d\times 2d}$

$$M_1 = \begin{bmatrix} I_d & 0_{d\times d} \\ 0_{d\times d} & 0_{d\times d} \end{bmatrix}, \quad M_2 = \begin{bmatrix} 0_{d\times d} & 0_{d\times d} \\ 0_{d\times d} & I_d \end{bmatrix},$$

the updates becomes

$$\begin{cases} \theta^{\text{cyc}}_{k+1} = \theta^{\text{cyc}}_k - \gamma M_1 A \theta^{\text{cyc}}_k, \\ \theta^{\text{cyc}}_{k+2} = \theta^{\text{cyc}}_k - \gamma M_2 A \theta^{\text{cyc}}_{k+1}, \\ \theta^{\text{cyc}}_{k+3} = \theta^{\text{cyc}}_{k+2} - \gamma M_2 A \theta^{\text{cyc}}_{k+2}, \\ \theta^{\text{cyc}}_{k+4} = \theta^{\text{cyc}}_{k+2} - \gamma M_1 A \theta^{\text{cyc}}_{k+3}. \end{cases} . \tag{28}$$

Remark that (28) contains two iterations of Alg. 1; θ_{k+1} and θ_{k+3} are extrapolations and θ_{k+2} and θ_{k+4} are updates. Defining $\mathcal{A}^{(\gamma)}_{ij} := I - \gamma M_i A + \gamma^2 M_i A M_j A$ and $\mathcal{A}^{(\gamma)}_{\text{cyc}} := \mathcal{A}^{\gamma}_{12}\mathcal{A}^{(\gamma)}_{21}$, we have $\theta^{\text{cyc}}_{k+4} = \mathcal{A}^{(\gamma)}_{\text{cyc}} \theta^{\text{cyc}}_k$.

Random sampling. Extra-gradient with random subsampling ($b = 1$) rewrites as

$$\begin{cases} \theta^{\text{rand}}_{k+1} = \theta^{\text{rand}}_k - \gamma M_{S_{k+1}} A \theta^{\text{rand}}_k, \\ \theta^{\text{rand}}_{k+2} = \theta^{\text{rand}}_k - \gamma M_{S_{k+2}} A \theta^{\text{rand}}_{k+1}, \\ \theta^{\text{rand}}_{k+3} = \theta^{\text{rand}}_{k+2} - \gamma M_{S_{k+3}} A \theta^{\text{rand}}_{k+2}, \\ \theta^{\text{rand}}_{k+4} = \theta^{\text{rand}}_{k+2} - \gamma M_{S_{k+3}} A \theta^{\text{rand}}_{k+3}. \end{cases} .$$

where $S_{k+1}, S_{k+2}, S_{k+3}, S_{k+4}$ take values 1 and 2 with equal probability and pairwise are independent. Note that we also enroll two iterations of sampled extra-gradient, as we consider a budget of 4 gradient evaluations. Let $\mathcal{F}_k = \sigma(S_{k'} : k' \leqslant k)$.

For extra-gradient with random player sampling, we can write

$$\begin{aligned}
\mathbb{E}\left[\theta_{k+4}^{\text{rand}}\right] &= \mathbb{E}\left[\mathcal{A}^{(\gamma)}_{S_{k+1}S_{k+3}}\mathcal{A}^{(\gamma)}_{S_{k+2}S_{k+1}}\theta_k^{\text{rand}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathcal{A}^{(\gamma)}_{S_{k+1}S_{k+3}}\mathcal{A}^{(\gamma)}_{S_{k+2}S_{k+1}}\theta_k^{\text{rand}}\middle|\mathcal{F}_k\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathcal{A}^{(\gamma)}_{S_{k+1}S_{k+3}}\mathcal{A}^{(\gamma)}_{S_{k+2}S_{k+1}}\middle|\mathcal{F}_k\right]\theta_k^{\text{rand}}\right] \\
&= \mathbb{E}\left[\mathcal{A}^{(\gamma)}_{S_{k+4}S_{k+3}}\mathcal{A}^{(\gamma)}_{S_{k+2}S_{k+1}}\right]\mathbb{E}\left[\theta_k^{\text{rand}}\right] \\
&= \frac{1}{16}\sum_{j_1,j_2,j_3,j_4\in\{1,2\}}\mathcal{A}^{(\gamma)}_{j_1j_2}\mathcal{A}^{(\gamma)}_{j_3j_4}\mathbb{E}\left[\theta_k^{\text{rand}}\right] \\
&= \frac{1}{16}\left(4I - 2\gamma A + \gamma^2 A^2\right)^2\mathbb{E}\left[\theta_k^{\text{rand}}\right] \triangleq \mathcal{A}^{(\gamma)}_{\text{rand}}\mathbb{E}\left[\theta_k^{\text{rand}}\right]
\end{aligned}$$

C.2. Convergence behavior through spectral analysis

The following well-known result proved by Gelfand (1941) relates matrix norms with spectral radii.

Theorem 6 (Gelfand's formula)**.** *Let $\|\cdot\|$ be a matrix norm on $\mathbb{R}^n$ and let $\rho(A)$ be the spectral radius of $A \in \mathbb{R}^n$ (the maximum absolute value of the eigenvalues of A). Then,*

$$\lim_{t\to\infty}\|A^t\|^{1/t} = \rho(A).$$

In our case, we thus have the following results, that describes the expected rate of convergence of the last iterate sequence $(\theta_t)_t$ towards 0. It is governed by the spectral radii $\rho(\mathcal{A}^{(\eta)})$ whenever the later is strictly lower than 1.

Corollary 2. *The behavior of θ_t^{full}, θ_t^{cyc} and θ_t^{rand} is related to the corresponding operators by the following expressions:*

$$\begin{aligned}
\lim_{t\to\infty}\left(\sup_{\theta_0^{\text{full}}\in\mathbb{R}^{2d}}\frac{\|\theta_t^{\text{full}}\|_2}{\|\theta_0^{\text{full}}\|_2}\right)^{1/t} &= \rho\left(\mathcal{A}^{(\gamma)}_{\text{full}}\right), \\
\lim_{t\to\infty}\left(\sup_{\theta_0^{\text{cyc}}\in\mathbb{R}^{2d}}\frac{\|\theta_t^{\text{cyc}}\|_2}{\|\theta_0^{\text{cyc}}\|_2}\right)^{1/t} &= \rho\left(\mathcal{A}^{(\gamma)}_{\text{cyc}}\right), \\
\lim_{t\to\infty}\left(\sup_{\theta_0^{\text{rand}}\in\mathbb{R}^{2d}}\frac{\|\mathbb{E}\left[\theta_t^{\text{rand}}\right]\|_2}{\|\theta_0^{\text{rand}}\|_2}\right)^{1/t} &= \rho\left(\mathcal{A}^{(\gamma)}_{\text{rand}}\right).
\end{aligned}$$

Proof. The proof is analogous for the three cases. Using the definition of operator norm,

$$\lim_{t\to\infty}\left(\sup_{\theta_0^{\text{full}}\in\mathbb{R}^{2d}}\frac{\|\theta_t^{\text{full}}\|}{\|\theta_0^{\text{full}}\|}\right)^{1/t} = \lim_{t\to\infty}\left(\sup_{\theta_0^{\text{full}}\in\mathbb{R}^{2d}}\frac{\left\|\left(\mathcal{A}^{(\gamma)}_{\text{full}}\right)^t\theta_0^{\text{full}}\right\|}{\|\theta_0^{\text{full}}\|}\right)^{1/t} = \lim_{t\to\infty}\left\|\left(\mathcal{A}^{(\gamma)}_{\text{full}}\right)^t\right\|^{1/t},$$

which is equal to $\rho\left(\mathcal{A}^{(\gamma)}_{\text{full}}\right)$ by Gelfand's formula. □

C.3. Empirical distributions of the spectral radii

Comparing the cyclic, random and full sampling schemes thus requires to compare the values

$$\mathcal{A}^\star_{\text{full}} \triangleq \min_{\gamma\in\mathbb{R}^+}\rho(\mathcal{A}^{(\gamma)}_{\text{full}}), \quad \mathcal{A}^\star_{\text{cyc}} \triangleq \min_{\gamma\in\mathbb{R}^+}\rho(\mathcal{A}^{(\gamma)}_{\text{cyc}}), \quad \mathcal{A}^\star_{\text{rand}} \triangleq \min_{\gamma\in\mathbb{R}^+}\rho(\mathcal{A}^{(\gamma)}_{\text{rand}}), \tag{29}$$

for all matrix games with positive payoff matrix $A \in \mathbb{R}^{2d\times 2d}$. This is not tractable in closed form. However, we may study the distribution of these values for random games.

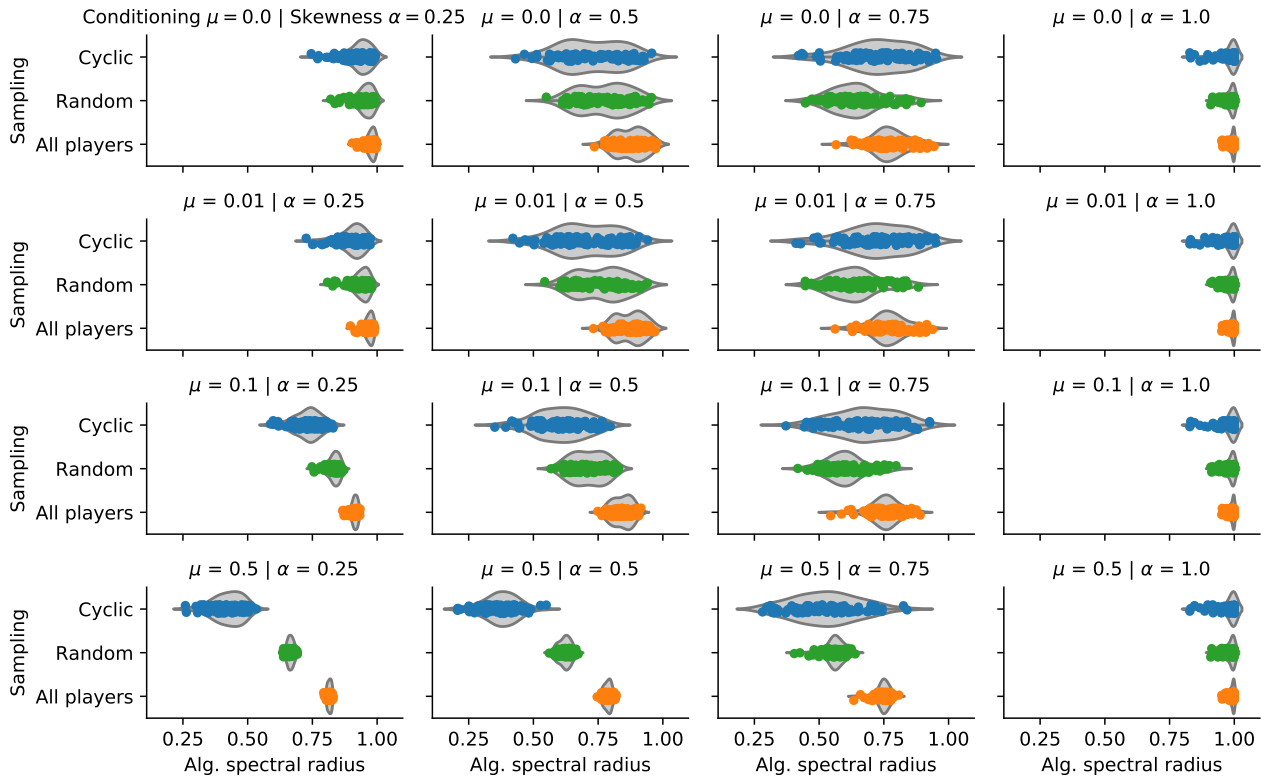

Figure 6. Spectral radii distribution of the algorithmic operator associated to doubly-stochastic and full extra-gradient, in the non-constrained bi-linear two-player game setting, for various conditioning and skewness. Random and cyclic sampling yields lower radius (hence faster rates) for most problem geometry. Cyclic sampling outperforms random sampling in most settings, especially for better conditioned problems.

Experiment. We sample matrices A in $\mathbb{R}^{2d\times 2d}$ (with $d = 3$) as the weighted sum of a random positive definite matrix A_{sym} and of a random skew matrix A_{skew}. We refer to App. D for a detailed description of the matrix sampling method. We vary the weight $\alpha \in [0, 1]$ of the skew matrix and the lowest eigenvalue μ of the matrix A_{sym}. We sample 300 different games and compute $\mathcal{A}^{(\eta)}$ on a grid of step sizes η, for the three different methods. We thus estimate the best algorithmic spectral radii defined in (29).

Results and interpretation. The distributions of algorithm spectral radii are presented in Fig. 6. We observe that the algorithm operator associated with sampling one among two players at each update is systematically more contracting than the standard extra-gradient algorithm operator, providing a further insight for the faster rates observed in §5.1, Fig. 2. Radius tend to be smaller for cyclic sampling than random sampling, in most problem geometry. This is especially true in well conditioned problem (high μ), little-skew problems (skewness $\alpha < .5$) and completely skew problems $\alpha = 1$. The later gives insights to explain the good performance of cyclic player sampling for GANs (§5.2), as those are described by skew games (zero-sum notwithstanding the discriminator penalty in WP-GAN).

On the other hand, we observe that radii are more spread using cyclic sampling for intermediary skew problerm ($\alpha = .75$), hinting that worst-case rates may be better for random sampling.

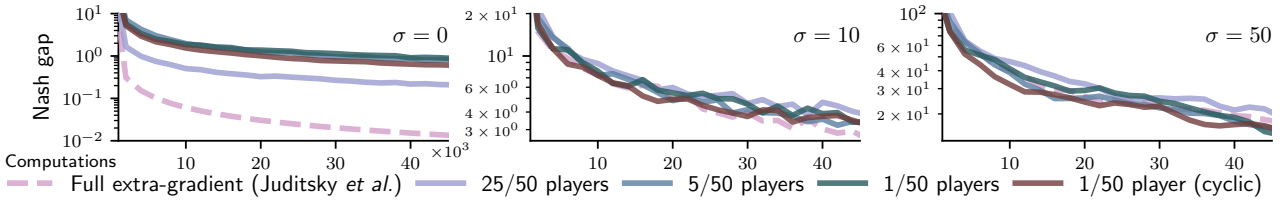

Figure 7. 50-player completely skew smooth game with increasing noise (sampling with variance reduction). In the non-noisy setting, player sampling reduces convergence speed. On the other hand, it provides a speed-up in the high noise regime.

D. Experimental results and details

We provide the necessary details for reproducing the experiments of §5.

D.1. Quadratic games

Generation of random matrices. We sample two random Gaussian matrix G and F in $\mathbb{R}^{nd\times nd}$, where each coefficient $g_{ij}, f_{ij} \sim \mathcal{N}(0,1)$ is sampled independently. We form a symmetric matrix $A_{\text{sym}} = \frac{1}{2}(G + G^T)$, and a skew matrix $A_{\text{skew}} = \frac{1}{2}(F - F^T)$. To make A_{sym} positive definite, we compute its lowest eigenvalue μ_0, and update $A_{\text{sym}} \leftarrow A_{\text{sym}} + (\mu - \mu_0)I_{nd\times nd}$, where μ regulates the conditioning of the problem and is set to 0.01. We then form the final matrix $A = (1-\alpha)A_{\text{sym}} + \alpha A_{\text{skew}}$, where α is a parameter between 0 and 1, that regulates the skewness of the game.

Parameters for quadratic games. Fig. 2 compare rates of convergence for doubly-stochastic extra-gradient and extra-gradient, for increasing problem complexity. Used parameters are reported in Table 2. Note that the conclusion reported in §5.1 regarding the impact of noise and the impact of cyclic sampling holds for all configurations we have tested; we designed increasingly complex experiments for concisely showing the efficiency and limitations of doubly-stochastic extra-gradient.

Grids. For each experiment, we sampled 5 matrices $(A_i)_i$ with skewness parameter α. We performed a grid-search on learning rates, setting $\eta \in \{10^{-5}, \cdots, 1\}$, with 32 logarithmically-spaced values, making sure that the best performing learning rate is always strictly in the tested range.

Limitations in skew non-noisy games. As mentioned in the main section, player sampling can hinder performance in completely skew games ($\alpha = 1$) with non-noisy losses. Those problems are the hardest and slower to solve. They corresponds to *fully adversarial* settings, where sub-game between each pair is zero-sum. We illustrate this finding in Fig. 7, showing how the performance of player sampling improves with noise. We emphasize that the non-noisy setting is not

Table 2. Parameters used in Fig. 2 for increasing problem complexity.

Figure	Players #	Exp.	Skewness α	Noise σ	Reg. λ
Fig. 2a	5	Smooth, no-noise	0.9	0	0
		Smooth, noisy	0.9	1	0.
		Skew, non-smooth, noisy	1.	1	$2 \cdot 10^2$
Fig. 2b	50	Smooth, no-noise	0.9	0	0
		Non-smooth, noisy	0.9	1	$2 \cdot 10^{-2}$
		Skew, non-smooth, noisy	1.	1	$2 \cdot 10^{-2}$
Fig. 2c	50	Smooth, skew, lowest-noise	0.95	1	0.
			0.95	10	0.
		Smooth, skew, highest-noise	0.95	100	0.
Fig. 7	50	Smooth, skew, no-noise	1	0	0.
			1	10	0.
		Smooth, skew, highest-noise	1	50	0

relevant to machine learning or reinforcement learning problems.

D.2. Generative adversarial networks

Models and loss. We use the Residual network architecture for generator and discriminator proposed by Gidel et al. (2019). We use a WGAN-GP loss, with gradient penalty $\lambda = 10$. As advocated by (Gidel et al., 2019), we use a 10 times lower stepsize for the generator. We train the generator and discriminator using the Adam algorithm (Kingma & Ba, 2015), and its straight-forward extension proposed by (Gidel et al., 2019).

Grids. We perform $5 \cdot 10^5$ generator updates. We average each experiments with 5 random seeds, and select the best performing generator learning rate $\eta \in \{2 \cdot 10^{-5}, 5 \cdot 10^{-5}, 8 \cdot 10^{-5}, 1 \cdot 10^{-4}, 2 \cdot 10^{-4}\}$, which turned out to be $5 \cdot 10^{-5}$ for both subsampled and non-subsampled extra-gradient.