# Extra-gradient with player sampling for provable fast convergence in $n$-player games

**Samy Jelassi***
Princeton University
sjelassi@princeton.edu

**Carles Domingo Enrich***
CIMS
New York University
cd2754@nyu.edu

**Damien Scieur**
Princeton University
dscieur@princeton.edu

**Arthur Mensch**
École Normale Supérieure, CNRS
CIMS, New York University
arthur.mensch@m4x.org

**Joan Bruna**
CIMS
New York University
bruna@cims.nyu.edu

**Abstract**

Data-driven model training is increasingly relying on finding Nash equilibria with provable techniques, e.g., for GANs and multi-agent RL. In this paper, we analyse a new extra-gradient method, that performs gradient extrapolations and updates on a random subset of players at each iteration. This approach provably exhibits the same rate of convergence as full extra-gradient in non-smooth convex games. We propose an additional variance reduction mechanism for this to hold for smooth convex games. Our approach makes extrapolation amenable to massive multiplayer settings, and brings empirical speed-ups, in particular when using cyclic sampling schemes. We demonstrate the efficiency of player sampling on large-scale non-smooth and non-strictly convex games. We show that the joint use of extrapolation and player sampling allows to train better GANs on CIFAR10.

## 1 Introduction

A growing number of models in machine learning require to optimize over multiple interacting objectives. This is the case of generative adversarial networks (Goodfellow et al., 2014), imaginative agents (Racanière et al., 2017), hierarchical reinforcement learning (Vezhnevets et al., 2017; Wayne and Abbott, 2014) and more generally multi-agent reinforcement learning (Bu et al., 2008). Solving saddle-point problems (see e.g., Rockafellar, 1970), that is key in robust learning (Kim et al., 2006) and image reconstruction (Chambolle and Pock, 2011), also falls in this category. All these examples can be cast as games where players are modules that compete or cooperate to minimize their own objective functions.

Optimizing over several objectives is challenging. To define a principled solution to a multi-objective optimization problem, we may rely on the notion of Nash equilibrium (Nash, 1951). At a Nash equilibrium, no player can improve its objective by unilaterally changing its strategy. In general games, finding a Nash equilibrium is known to be PPAD-complete (Daskalakis et al., 2009). The theoretical section of this paper considers the class of *convex n-player games*, for which Nash equilibria exist (Nemirovski et al., 2010). Finding a Nash equilibrium in this setting is equivalent to solving a variational inequality problem (VI) with a monotone operator (Harker and Pang, 1990; Nemirovski et al., 2010). This VI can be solved using first-order methods, that are prevalent in single-objective optimization for machine learning. Stochastic gradient descent (the simplest first-order method) is indeed known to converge to local minima under mild conditions met by ML problems (Bottou and Bousquet, 2008; Lee et al., 2016). Yet, while gradient descent can be applied simultaneously to different objectives, it may fail in finding a Nash equilibrium in very simple settings (see e.g., Gidel et al., 2019; Letcher et al., 2019). Two alternative modifications of gradient descent are necessary

*Equal contribution

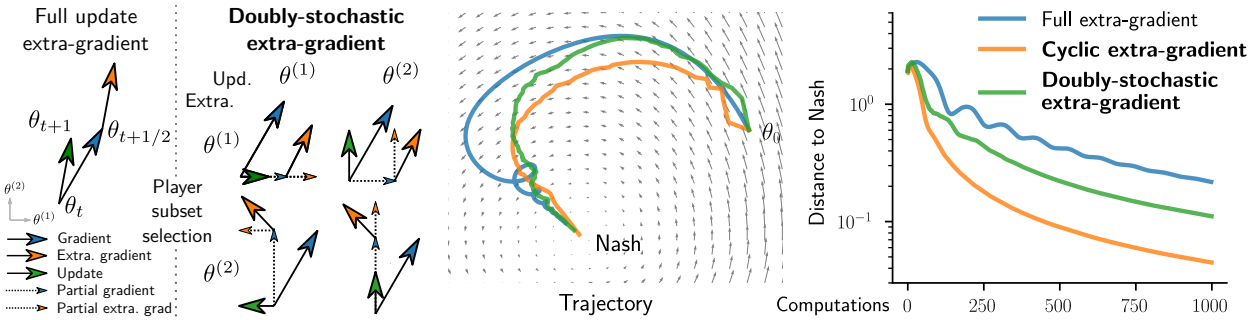



Figure 1: *Left:* We compute masked gradient during the extrapolation and update steps of the extra-gradient algorithm, to perform faster updates. *Right:* Optimization trajectories for doubly stochastic extra-gradient and full-update extra-gradient, on a convex two-player, one-parameter each, convex game. Player sampling improves the expected rate of convergence toward the Nash equilibrium $(0, 0)$. Doubly-stochastic extra-gradient trajectory averaged over 10 runs.

| Algorithm | Non-smooth games | Smooth games |
|---|---|---|
| Juditsky et al. (2011) | $\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_{t(k)})\right] \leqslant \mathcal{O}\left(\sqrt{\frac{\Omega n^2(4G^2+\sigma^2)}{k}}\right)$ | $\mathcal{O}\left(\frac{\Omega L n^{3/2}}{k} + \sqrt{\frac{\Omega n^2\sigma^2}{k}}\right)$ |
| Doubly-stochastic extra-gradient | $\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_{t(k)})\right] \leqslant \mathcal{O}\left(\sqrt{\frac{\Omega n(G^2(3n-b)+\sigma^2 b)}{k}}\right)$ | $\mathcal{O}\left(\frac{\Omega L n^2}{\sqrt{b}k} + \sqrt{\frac{\Omega n b\sigma^2}{k}}\right)$ |

Table 1: New and existing convergence rates with respect to number of computations $k$. Doubly-stochastic extra-gradient *divides the noise contribution with a factor* $\sqrt{n/b}$, where $b$ is the number of sampled players among $n$. $G$ bounds the gradient norm, $L$ is the Lipschitz constant of the gradients of losses, $\sigma^2$ bounds the noise in gradient estimation, $\Omega$ is the diameter of the parameter space.

to solve the VI (hence Nash) problem: *averaging* (Magnanti and Perakis, 1997; Nedić and Ozdaglar, 2009) and *extrapolation* with averaging (introduced as the *extra-gradient* method (Korpelevich, 1976)), which is faster (Nemirovski, 2004). Extrapolation corresponds to an *opponent shaping* step: each player anticipates its opponents' next moves to update its strategy.

In $n$-player games, extra-gradient computes $2n$ single player gradients before performing a parameter update. Whether in massive or simple two-players games, this may be an inefficient update strategy: early gradient information, computed at the beginning of each iteration, could be used to perform eager updates or extrapolations, similar to how alternated training e.g, Goodfellow et al., 2014, for GANs would behave. In this paper, we introduce and analyse new extra-gradient algorithms that extrapolate and update random or carefully selected subsets of players at each iteration (Fig. 1). Contributions are as follow.

- We review the extra-gradient algorithm for convex games and outline its shortcomings (§3.1). We propose a doubly-stochastic extra-gradient (DSEG) algorithm (§3.2) that relies on partially observed gradients of players. It performs faster but noisier updates than original extra-gradient descent. We introduce a variance reduction method to attenuate the added noise for smooth games. We describe an importance and a cyclic sampling scheme that improve convergence speed.
- We propose a sharp analysis of our method's convergence rates (§4), as outlined in Table 1, for the various proposed variants. Those rates outlines the trade-offs of player sampling.
- We demonstrate the performance of player-sampled extra-gradient in controlled settings (quadratic games, §5), showing how our approach overcomes vanilla extra-gradient, especially using cyclic player selection. Most interestingly, compared to vanilla extra-gradient, our approach (with cyclic sampling) is also more efficient for GAN training (CIFAR10, ResNet architecture).

## 2 Related work

**Extra-gradient method.** In this paper, we focus on finding the Nash equilibrium in convex $n$-player games (1), or equivalently the Variational Inequality problem (4) (Harker and Pang, 1990; Nemirovski et al., 2010). This can be done using extrapolated gradient (Korpelevich, 1976), a "cautious" gradient descent approach (described in (3)) that was promoted by Nemirovski (2004) and Nesterov (2007), under the name *mirror-prox*—we review this work in §3.1. Juditsky et al. (2011) propose a stochastic variant of mirror-prox, that assumes access to a noisy gradient oracle. Recently, Bach and Levy (2019) described a smoothness-adaptive variant of this algorithm similar to AdaGrad (Duchi et al., 2011), an approach that can be combined with ours. Yousefian et al. (2018) consider multi-agent games on networks and analyze a stochastic variant of extra-gradient that consists in randomly extrapolating and updating a single player. Compared to them, we analyse more general player sampling strategies. Moreover, our analysis holds for non-smooth losses, and provides better rates for smooth losses, through variance reduction. We also analyse precisely the reasons why player sampling is useful (see §3.2 and comments on rates in §4), which is an original endeavor.

**Finding Nash equilibria in non-convex settings.** A number of algorithms have been proposed in the non-convex setting under restricted assumptions on the game, for example WoLF in two-player two-action games (Bowling and Veloso, 2001), policy prediction in two-player two-action bi-matrix games (Zhang and Lesser, 2010), AWESOME in repeated games (Conitzer and Sandholm, 2007), Optimistic Mirror Descent in two-player bilinear zero-sum games (Daskalakis et al., 2018) and Consensus Optimization in two-player zero-sum games (Mescheder et al., 2017). Mertikopoulos et al. (2019) proved asymptotic convergence results for extra-gradient without averaging in a slightly non-convex setting. Gidel et al. (2019) demonstrated the effectivenes of extra-gradient for non-convex GAN training—in §5, we demonstrate that player sampling improves training speed and effectiveness in the GAN setting.

**Opponent shaping and gradient adjustment.** Extra-gradient can also be understood as an *opponent shaping* method: in the extrapolation step, the player looks one step in the future and anticipates the next moves of his opponents. Several recent works proposed algorithms that make use of the opponents' information to converge to an equilibrium (Foerster et al., 2018; Letcher et al., 2019; Zhang and Lesser, 2010). In particular, the "Learning with opponent-learning awareness" (LOLA) algorithm is known for encouraging cooperation in cooperative games (Foerster et al., 2018). Lastly, some recent works proposed algorithms to modify the dynamics of simultaneous gradient descent by adding an adjustment term in order to converge to the Nash equilibrium (Mazumdar et al., 2019) and avoid oscillations (Balduzzi et al., 2018; Mescheder et al., 2017). One caveat of these works is that they need to estimate the Jacobian of the simultaneous gradient, which may be expensive in large-scale systems or even impossible when dealing with non-smooth losses as we consider in our setting. This is orthogonal to our approach that finds solutions of the original VI problem (4).

## 3 Solving convex games with partial first-order information

We review the framework of Cartesian convex games and the extra-gradient method in §3.1. Building on these, we propose to augment extra-gradient with player sampling and variance reduction in §3.2.

### 3.1 Solving convex $n$-player games with gradients

Each player observes a loss that depends on the independent parameters of all other players.

**Definition 1.** *A Cartesian $n$-player game is given by a set of $n$ players with parameters $\theta = (\theta^1, \dots, \theta^n) \in \Theta \subset \mathbb{R}^d$ where $\Theta$ decomposes into a Cartesian product $\prod_{i=1}^n \Theta_i$. Each player's parameter $\theta_i$ lives in $\Theta_i \subset \mathbb{R}^{d_i}$, where $d = \sum_{i=1}^n d_i$. Each player is given a loss function $\ell_i \colon \Theta \to \mathbb{R}$.*

For example, generative adversarial network (GAN) training can be cast as a Cartesian game between a generator and discriminator that do not share parameters. We make the following assumption over the geometry of losses and constraints, that corresponds to the convexity assumption for one player.

**Assumption 1.** *The parameter spaces $\Theta, \Theta_1, \dots, \Theta_n$ are compact, convex and non-empty. Each player's loss $\ell_i(\theta^i, \theta^{-i})$ is convex in its parameter $\theta^i$ and concave in $\theta^{-i}$, where $\theta^{-i}$ contains all other players' parameters. Moreover, $\sum_{i=1}^n \ell_i(\theta)$ is convex in $\theta$.*

Ass. 1 implies that $\Theta$ has a diameter $\Omega \triangleq \max_{u,z\in\Theta} \|u - z\|_2$. Note that the losses may be non-differentiable. A simple example of Cartesian convex games satisfying Ass. 1, that we will empirically study in §5, are matrix games (e.g., rock-paper-scissors) defined by a positive payoff matrix $A \in \mathbb{R}^{d\times d}$, with parameters $\theta$ corresponding to $n$ mixed strategies $\theta_i$ lying in the probability simplex $\triangle^{d_i}$.

**Nash equilibria.** Joint solutions to minimizing losses $(\ell_i)_i$ are naturally defined as the set of *Nash equilibria* (Nash, 1951) of the game. In this setting, the goal of multi-objective optimization becomes

$$\text{Find} \quad \theta_\star \in \Theta \quad \text{such that} \quad \forall\, i \in [n], \quad \ell_i(\theta^i_\star, \theta^{-i}_\star) = \min_{\theta^i \in \Theta^i} \ell_i(\theta^i, \theta^{-i}_\star). \tag{1}$$

Intuitively, a Nash equilibrium is a point where no player can benefit by changing his strategy while the other players keep theirs unchanged. Ass. 1 implies the existence of a Nash equilibrium (Rosen, 1964). We quantify the inaccuracy of a solution $\theta$ by the *functional Nash error* (Nemirovski, 2004)

$$\mathrm{Err}_N(\theta) \triangleq \textstyle\sum_{i=1}^n \left[\ell_i(\theta) - \min_{z\in\Theta_i} \ell_i(z, \theta^{-i})\right]. \tag{2}$$

This error, computable through convex optimization, quantifies the gain that each player can obtain when deviating alone from the current strategy. In particular, $\mathrm{Err}_N(\theta) = 0$ if and only if $\theta$ is a Nash equilibrium; thus $\mathrm{Err}_N(\theta)$ constitutes a proper indication of convergence for sequence of iterates seeking a Nash equilibrium. It is the value we bound in our convergence analysis (see §4).

**First-order methods and extrapolation.** We consider (sub)differentiable losses $\ell_i$ forming a convex game. The Nash equilibrium can be found using first-order methods, that access the gradients of $\ell_i$. We define the *simultaneous gradient* of the game to be

$$F \triangleq (\nabla_1 \ell_1, \dots, \nabla_n \ell_n)^\top \in \mathbb{R}^d,$$

where we write $\nabla_i \ell_i \triangleq \nabla_{\theta_i} \ell_i$. It corresponds to the concatenation of the gradients of each player's loss with respect to its own parameters. The losses $\ell_i$ may be non-smooth, in which case the gradients $\nabla_i l_i$ should be replaced by subgradients. Simultaneous gradient descent simply approximates the flow of the simultaneous gradient. It fails to converge in very simple settings, in particular in any matrix games for which the payoff is skew-symmetric. An alternative approach with better guarantees is the extra-gradient (Korpelevich, 1976) method, which forms the basis for the algorithms presented in this paper. It has been extensively analyzed under several settings (see §2). In particular, Nemirovski (2004) provides convergence results when gradients are exact, and Juditsky et al. (2011) when gradients are accessed through a noisy oracle.

Extra-gradient consists in two steps: first, we take a gradient step to go to an extrapolated point. We then use the gradient at the extrapolated point to perform a gradient step from the original point:

$$\begin{aligned} \text{At iteration } \tau, \quad \text{(extrapolation)} \qquad & \theta_{\tau+1/2} = p_\Theta[\theta_\tau - \gamma_\tau F(\theta_\tau)], \\ \text{(update)} \qquad & \theta_{\tau+1} = p_\Theta[\theta_\tau - \gamma_\tau F(\theta_{\tau+1/2})], \end{aligned} \tag{3}$$

where $p_\Theta[\cdot]$ is the Euclidean projection onto the constraint set $\Theta$, i.e. $p_\Theta[z] = \mathrm{argmin}_{\theta\in\Theta} \|\theta - z\|_2^2$. This "cautious" approach allows to escape cycling orbits of the simultaneous gradient flow, that may arise around equilibrium points with skew-symmetric Hessians (see Fig. 1). The generalization of extra-gradient to general Banach spaces equipped by a Bregman divergence was introduced as the *mirror-prox* algorithm (Nemirovski,

**Algorithm 1** Doubly-stochastic extra-gradient.

**Input**: initial point $\theta_0 \in \mathbb{R}^d$, stepsizes $(\gamma_\tau)_{\tau \in [t]}$, mini-batch size over the players $b \in [n]$.
If variance reduction, initialize $R \leftarrow \tilde{F}(\theta_0, [n])$ as in equation (5) with full simultaneous gradient.
**for** $\tau = 0, \dots, t$ **do**
    Sample mini-batches of players $\mathcal{P}$, $\mathcal{P}'$.
    Compute $\tilde{F}_{\tau+\frac{1}{2}} = \tilde{F}(\theta_\tau, \mathcal{P})$ using (5) or using variance reduction (Alg. 2).
    Extrapolation step: $\theta_{\tau+\frac{1}{2}} \leftarrow p_\Theta[\theta_\tau - \gamma_\tau \tilde{F}_{\tau+\frac{1}{2}}]$.
    Compute $\tilde{F}_{\tau+1} = \tilde{F}(\theta_{\tau+\frac{1}{2}}, \mathcal{P}')$ using (5) or with variance reduction.
    Gradient step: $\theta_{\tau+1} \leftarrow p_\Theta[\theta_\tau - \gamma_\tau \tilde{F}_{\tau+1}]$.
**Return** $\hat{\theta}_t = [\sum_{\tau=0}^t \gamma_\tau]^{-1} \sum_{\tau=0}^t \gamma_\tau \theta_\tau$.

2004). All new results from §4 extend to this *mirror* setting (see §A.1). As recalled in Table 1, Juditsky et al. (2011) provide rates of convergence for the average iterate $\hat{\theta}_t = \frac{1}{t}\sum_{\tau=1}^t \theta_\tau$. Those rates are introduced for the equivalent variational inequality (VI) problem:

$$\text{Find} \quad \theta_\star \in \Theta \quad \text{such that} \quad F(\theta_\star)^\top (\theta - \theta_\star) \geqslant 0 \quad \forall \theta \in \Theta, \tag{4}$$

where Ass. 1 ensures that the simultaneous gradient $F$ is a monotone operator (see §A.2 for the link between Nash equilibria and solutions of the VI).

**Computational caveats.** In systems with large number of players, an extra-gradient step may be computationally expensive due to the high number of backward passes necessary for gradient computations. Namely, at each iteration, we are required to compute $2n$ gradients before performing a first update. This is likely to be inefficient, as we could use the first computed gradients to perform a first extrapolation or update. This remains true for games down to two players. In a different setting, stochastic gradient descent (Robbins and Monro, 1951) updates model parameters before observing the whole data, assuming that partial observation is sufficient for progress in the optimization loop. Similarly, partial gradient observation should be sufficient to perform extrapolation and updates toward the Nash equilibrium. We therefore propose to compute a few *random* player gradients at each iteration.

## 3.2 Partial extrapolation and update for extra-gradient

We present our main algorithm contribution in this section. While standard extra-gradient requires two full passes over players, we propose to compute *doubly-stochastic simultaneous gradient estimates*. This corresponds to evaluating a simultaneous gradient that is affected by *two* sources of noise. We sample a mini-batch $\mathcal{P}$ of players of size $b \leqslant n$, and compute the gradients for this mini-batch only. Furthermore, we assume that the gradients are noisy estimates, e.g., with noise coming from data sampling. We then compute a doubly-stochastic simultaneous gradient estimate $\tilde{F}$ as

$$\tilde{F} \triangleq (\tilde{F}^{(1)}, \dots, \tilde{F}^{(n)})^\top \in \mathbb{R}^d \text{ where } \tilde{F}^{(i)}(\theta, \mathcal{P}) \triangleq \begin{cases} \frac{n}{b} \cdot g_i(\theta) & \text{if } i \in \mathcal{P} \\ 0_{d_i} & \text{otherwise} \end{cases}, \tag{5}$$

where $g_i(\theta)$ is a noisy unbiased estimate of $\nabla_i \ell_i(\theta)$. The factor $n/b$ in (5) ensures that the doubly-stochastic simultaneous gradient estimate is an unbiased estimator of the simultaneous gradient. Doubly-stochastic extra-gradient replaces the full update (3) by oracle (5), as detailed in Alg. 1.

**Motivation.** Sampling over players introduces a further source of noise in the average iterate sequence $(\bar{\theta}_t)_t$. The convergence of this sequence is already slowed down by noisy gradients or by the non-smoothness of the losses, that both introduce a term in $1/\sqrt{t}$ in the convergence bounds. It is therefore appealing to introduce a further source of noise, hoping that the computational speed-ups provided at each iteration mitigates the approximation errors introduced by player subsampling.

**Algorithm 2** Variance reduced estimate of the simultaneous gradient with doubly-stochastic sampling

**Input**: point $\theta \in \mathbb{R}^d$, mini-batch $\mathcal{P}$, table of previous gradient estimates $R \in \mathbb{R}^d$.
Compute $\tilde{F}(\theta, \mathcal{P})$ as specified in equation (5).
**for** $i \in \mathcal{P}$ **do**
    Compute $\bar{F}^{(i)} \leftarrow \tilde{F}^{(i)}(\theta) - (1 - \frac{b}{n})R^{(i)}$ and update $R^{(i)} \leftarrow \tilde{F}^{(i)}(\theta)$
**for** $i \notin \mathcal{P}$ **do**
    Set $\bar{F}^{(i)} \leftarrow R^{(i)}$.
**Return** the estimate $\bar{F} = (\bar{F}^{(1)}, ..., \bar{F}^{(n)})$, updated table $R$.

**Variance reduction for player noise.** To obtain faster rates in convex games with smooth losses, we propose to compute variance reduced estimate of the simultaneous gradient. This mitigates the noise due to player sampling. Variance reduction is a technique known to accelerate convergence under smoothness assumptions in similar settings. While Chavdarova et al. (2019), Iusem et al. (2017), and Palaniappan and Bach (2016) apply variance reduction on the noise coming from the gradient estimates, we apply it to the noise coming from the sampling over the players. We implement this idea in Alg. 2. We keep an estimate of $\nabla_i \ell_i$ for each player in a table $R$, which we use to compute *unbiased* gradient estimates with lower variance, similar to SAGA (Defazio et al., 2014).

**Sampling strategies.** In the basic version of the algorithm, the sampling over players can be performed using any distribution with uniform marginals, i.e such that all players have equal probability of being sampled. Sampling uniformly over $b$-subsets of $[n]$ is a reasonable way to fulfill this condition as all players have probability $p = b/n$ of being chosen. One faster alternative is to perform importance sampling. Namely, we sample each player $i$ with a probability $p_i$ proportional to the uniform bound of $\|\nabla_i \ell_i\|$. This technique achieve faster convergence (see §B.3) when the gradient bounds for the different losses differ.

As a strategy to accelerate convergence, we propose to cycle over the $n(n-1)$ pairs of different players (with $b = 1$). At each iteration, we extrapolate the first player of the pair and update the second one. We shuffle the order of pairs once the block has been entirely seen. By excluding $(i, i)$ pairs, we avoid players extrapolating themselves, which is never useful to reduce $\ell_i$. This scheme bridges extrapolation and alternated gradient descent: for GANs, it corresponds to extrapolate the generator before updating the discriminator, and vice-versa, cyclically. Sampling over players proves powerful for quadratic games (§5.1) and GANs (§5.2). In App. C, we provide a first explanation for this fact, based on studying the spectral radius of recursion operators (echoing recent work on understanding cyclic coordinate descent (X. Li et al., 2018)).

# 4 Sharp analysis of convergence rates

We state our main convergence results in this section. As announced, we derive rates for the algorithms mentioned in §3.2 following the analysis by Juditsky et al. (2011). We compare them with the rates achieved by stochastic extra-gradient introduced by Juditsky et al. (2011), which also assumes noisy gradients but no player subsampling. While in the main paper the theorems are provided in the Euclidean setting, the proofs in the appendices are written in the mirror setting. In the analysis, we separately consider the two following assumptions on the losses.

**Assumption 2a** (Non-smoothness)**.** *For each $i \in [n]$, the loss $\ell_i$ has a bounded subgradient, namely $\max_{h \in \partial_i \ell_i(\theta)} \|h\|_2 \leqslant G_i$ for all $\theta \in \Theta$. In this case, we also define the quantity $G = \sqrt{\sum_{i=1}^n G_i^2 / n}$.*

**Assumption 2b** (Smoothness)**.** *For each $i \in [n]$, the loss $\ell_i$ is once-differentiable and $L$-smooth, i.e. $\|\nabla_i \ell_i(\theta) - \nabla_i \ell_i(\theta')\|_2 \leqslant L \|\theta - \theta'\|_2$, for $\theta, \theta' \in \Theta$.*

Classically, similar to Juditsky et al. (2011) and Robbins and Monro (1951), we assume unbiasedness and boundedness of the variance.

**Assumption 3.** *For each player $i$, the noisy gradient $g_i$ is unbiased and has bounded variance:*

$$\forall\, \theta \in \Theta, \quad \mathbb{E}[g_i(\theta)] = \nabla_i \ell_i(\theta), \quad \mathbb{E}[\|g_i(\theta) - \nabla_i \ell_i(\theta)\|_2^2] \leqslant \sigma^2.$$

In stochastic gradient-based methods, comparing rates in terms of number of iterations is not appropriate since the complexity per iteration increases with the size $b$ of player mini-batches. Instead, we define $k(t)$ as the number of gradients estimates $g_i$ computed up to iteration $t$. At each iteration in Alg. 1, the doubly-stochastic simultaneous gradient estimate is computed twice and requires $k$ gradient estimates. Therefore, $k(t) = 2bt$ which implies the number of iterations in terms of gradient computations is $t(k) = k/2b$. We give the rates in terms of $k$ in the statement of the theorems. We first state the convergence result for doubly-stochastic extra-gradient under Ass. 2a.

**Theorem 1.** *We consider a convex $n$-player game where Ass. 2a holds. Assume that Alg. 1 is run without variance reduction and constant stepsize $\gamma$. The expected $Err_N(\hat{\theta}_{t(k)})$ is upper bounded as*

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_{t(k)})\right] \leqslant \mathcal{O}\left(\sqrt{\frac{\Omega n}{k}\left(nG^2 + b\sigma^2\right)}\right), \quad \textit{setting} \quad \gamma \in \mathcal{O}\left(\sqrt{\frac{b\Omega}{n\left(nG^2 + b\sigma^2\right)t(k)}}\right).$$

The following results holds when the losses are once-differentiable and smooth (Ass. 2b).

**Theorem 2.** *We consider a convex $n$-player game where Ass. 2b holds. Assume that we run Alg. 1 with variance reduction and constant stepsize $\gamma$. The expected $Err_N(\hat{\theta}_{t(k)})$ is upper bounded as*

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_{t(k)})\right] \leqslant \mathcal{O}\left(\frac{\Omega L n^2}{\sqrt{b}k} + \sqrt{\frac{\Omega n b \sigma^2}{k}}\right), \ \textit{setting } \gamma \in \min\left\{\mathcal{O}\Big(\frac{b^{3/2}}{Ln^2}\Big), \mathcal{O}\Big(\sqrt{\frac{\Omega}{n\sigma^2 t(k)}}\Big)\right\}.$$

Those rates should be compared to the rate of Juditsky et al., 2011, Corollary 1, that we recall in §A.3 and Table 1. Corollary 3 and 7 in §B.2.2 and §B.4 contain the statements of Theorem 1 and 2 in more detail.

(i) Under Ass. 2a, Alg. 1 performs with a rate similar to stochastic extra-gradient. In both cases the rate is $1/\sqrt{k}$, and the subgradient bound $G$ and noise bound $\sigma^2$ appear on the numerator. Doubly-stochastic extra-gradient is more robust to noisy gradient estimates, because the dependency of its rate on $\sigma^2$ is weaker than for full extragradient.

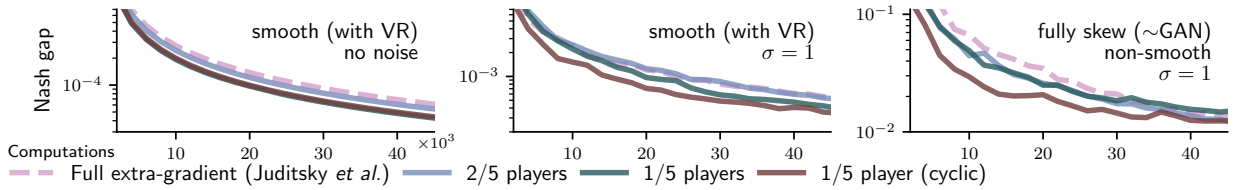
(ii) Under Ass. 2b, the deterministic $\mathcal{O}(1/k)$ term of the rate is $\sqrt{n/b}$ times larger compared to stochastic extra-gradient while the noisy $\mathcal{O}(1/\sqrt{k})$ term is $\sqrt{b/n}$ times smaller. For long runs (large $k$), the noise term dominates the deterministic one, which advocates for the use of small batch sizes: when $b = 1$, the rate is asymptotically $1/\sqrt{n}$ times smaller. Setting $\sigma$ to zero in the noise term, doubly-stochastic extra-gradient with variance reduction recovers the rate from (Nemirovski, 2004).

To sum up, doubly-stochastic extra-gradient provides better convergence guarantees than stochastic extra-gradient under high levels of noise ($\sigma^2 \gg 0$), while it delivers similar or slightly worse theoretical results in the non-noisy regime. Player randomness can be considered in the framework from Juditsky et al. (2011) by including it in the noisy unbiased estimate $g_i$ (increasing $\sigma^2$ in Ass. 3 accordingly). This coarse approach does not yield the sharp bounds of Theorem 1 and 2 (see §A.4).

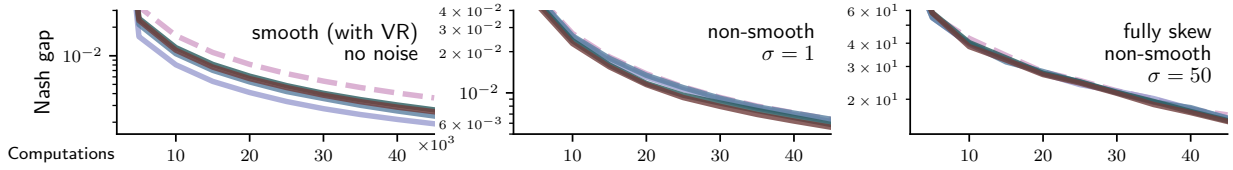
**Importance sampling.** Using importance sampling when choosing player mini-batches yields a better bound by a constant factor (see §B.3). In the non-smooth case, this replaces the constant $G$ with the strictly smaller $\frac{1}{n}\sum_{i=1}^n G_i$, which is useful when the gradient magnitudes are skewed.

# 5 Applications

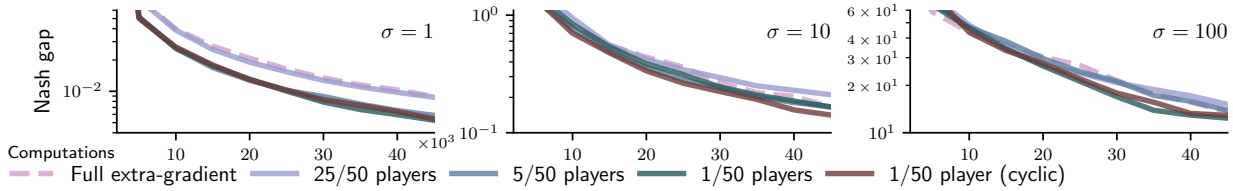
We show the performance of doubly-stochastic extra-gradient in the setting of quadratic games over the simplex, and in the practical context of GAN training. A *PyTorch*/*Numpy* package is attached.




(a) 5-player games with increasingly hard configurations.




(b) 50-player games with increasingly hard configurations.




(c) 50-player smooth game with increasing noise (sampling with variance reduction).

Figure 2: Player sampled extra-gradient outperform vanilla extra-gradient for small noisy/non-noisy smooth/non-smooth games. Cyclic sampling performs better than random sampling, especially for 5 players (a). High noise regime, typical in machine learning, demands stronger subsampling (c). Curves averaged over 5 games and 5 runs for random algorithms.

## 5.1 Random quadratic games

We consider a game where $n$ players can play $d$ actions, with payoffs provided by a matrix $A \in \mathbb{R}^{nd \times nd}$, an horizontal stack of matrices $A_i \in \mathbb{R}^{(d \times nd)}$ (one for each player). The loss function $\ell_i$ of each player is defined as its expected payoff given the $n$ mixed strategies $(\theta^1, \dots, \theta^n)$, i.e.

$$\forall\, i \in [n], \quad \forall\, \theta \in \Theta = \triangle^{d_1} \times \cdots \times \triangle^{d_n}, \quad \ell_i(\theta^i, \theta_{-i}) = {\theta^i}^\top A_i \theta + \lambda \|\theta^i - \frac{1}{d}\|_1,$$

where $\lambda$ is a regularization parameter that introduces non-smoothness and pushes strategies to snap to the simplex center. The positivity of $A$ is equivalent to Ass. 1, i.e. $\theta^\top A \theta \geqslant 0$ for all $\theta \in \Theta$.

**Experiments.** We sample $A$ as the weighted sum of a random symmetric positive definite matrix and a skew matrix. We compare the convergence speeds of extra-gradient algorithms, with or without player subsampling. We vary three parameters: the variance $\sigma$ of the noise in the gradient oracle (we add a Gaussian noise on each gradient coordinate, similar to Langevin dynamics (Neal et al., 2011)), the non-smoothness $\lambda$ of the loss, and the skewness of the matrix. We consider small games and large games ($n \in \{5, 50\}$). We use the (simplex-adapted) mirror variant of doubly-stochastic extra-gradient, and a constant stepsize, selected among a grid (see App. D). We use variance reduction when $\lambda = 0$ (smooth case). We compare random fixed-size sampling with cyclic sampling (§3.2).

**Results.** Fig. 2 compares the convergence speed of player-sampled extra-gradient for the various settings and sampling schemes. As predicted by Theorem 1 and 2, rates of convergence are comparable with and without subsampling. Randomly subsampling players always brings a benefit in the convergence constant (Fig. 2a-b), especially in the smooth noisy regime, using variance reduction (Fig. 2a column 2). Most


Figure 3: Training curves and samples using doubly-stochastic extrapolation on WGAN-GP, with dataset CIFAR10, for best learning rates. Doubly-stochastic extrapolation allows faster and better training, most notably in term of Fréchet Inception Distance (10k). Curves averaged over 5 runs.

| Method | IS | FID (50k) |
|---|---|---|
| EG (Gidel et al., 2019) | $8.26 \pm .16$ | $19.69 \pm 1.53$ |
| **DSEG** | $\mathbf{8.38 \pm .06}$ | $\mathbf{17.10 \pm 1.07}$ |

Table 2: Effect of subsampling on training ResNet WGAN-GP. FID and IS computed on 50k samples, averaged over 5 runs (+ 10 folds for IS).

interestingly, cyclic player selection brings a significant improvement over random sampling for small number of players, allowing larger gain in the rate constants (Fig. 2a).

Fig. 2c highlights the trade-offs in Theorem 2: as the noise increase, the size of player batches should be reduced. Not that for skew-games with many players (Fig. 2b col. 3), which are the hardest games to solve as averaging is *needed* (Mertikopoulos et al., 2019), our approach only becomes beneficial in the high-noise regime (more relevant in ML). Full extra-gradient should be favored in the non-noisy regime (see App. D).

**Spectral effect of sampling.** To better understand the benefit of the cyclic selection scheme, we study the linear "algorithm operator" $\mathcal{A}$ such that $\theta_t \triangleq \mathcal{A}(\theta_{t-1})$ in non constrained two-player bilinear games. The convergence speed of $(\theta_t)_t$ is governed by the spectral radius of $\mathcal{A}$, in light of Gelfand's formula (Gelfand, 1941). In App. C Fig. 4, we consider random matrix games. For these, the algorithm operator of extra-gradient with cyclic player selection has on average a lower spectral radius than with random selection and a fortiori full selection. This leads to faster convergence of cyclic schemes.

## 5.2 Generative adversarial networks (WGAN-GP + ResNet)

We evaluate the performance of the player sampling approach to train a generative model on CIFAR10 (Krizhevsky and Hinton, 2009). We use the WGAN-GP loss (Gulrajani et al., 2017), that defines a non-convex two-player game. We compare the full extra-gradient approach advocated by Gidel et al. (2019) to the cyclic sampling scheme proposed in §3.2 (i.e. *extra. D, upd. G, extra. G, upd. D*). We use the ResNet (He et al., 2016) architecture from Gidel et al. (2019), and select the best performing stepsizes among a grid (see App. D). We use the Adam (Kingma and Ba, 2015) refinement of extra-gradient (Gidel et al., 2019) for both the baseline and proposed methods. We evaluate the Inception Score (Salimans et al., 2016) and Fréchet Inception Distance (FID) (Heusel et al., 2017) along training.

**Results.** We report training curves versus wall-clock time in Fig. 3. Cyclic sampling allows faster and better training, especially with respect to FID, which is more correlated to human appreciation (Heusel et al., 2017). Table 2 compares our result to full extra-gradient with uniform averaging. It shows substantial improvements in FID, with results less sensitive to randomness. Note that scores could be slightly improved by leaving more time for training.

**Interpretation.** Without extrapolation, alternated training is known to perform better than simultaneous updates in WGAN-GP (Gulrajani et al., 2017). Our approach allows to add extrapolation but keep an alternated schedule. It thus performs better than extrapolating with simultaneous updates. It remains true

across every learning rate we tested. Deterministic sampling is crucial for performance, as random player selection performs poorly (best score 6.2 IS). This echoes the good results of cyclic sampling in §5.1.

# 6 Discussion and conclusion

We propose and analyse a doubly-stochastic extra-gradient approach for finding Nash equilibria. According to our convergence analysis, updating/extrapolating subsets of players only is useful in high noise or non-smooth settings, and equivalent otherwise. Numerically, doubly-stochastic extra-gradient leads to speed-ups and improvements in convex and non-convex settings, especially using noisy gradients (as with GANs). Our approach hence combines the advantages of alternated and extrapolation methods over simultaneous gradient descent—we recommend it for training GANs.

Beyond demonstrating the usefulness of sampling, numerical experiments show the importance of *sampling schemes*. We take a first step towards understanding the good performance of *cyclic* player extrapolation and update. A better theoretical analysis of this phenomenon is left for future work.

We foresee interesting developments using player sampling and extrapolation in reinforcement learning: the policy gradients obtained using multi-agent actor critic methods (Lowe et al., 2017; Mnih et al., 2016) are highly noisy estimates, a setting in which sampling over players proves beneficial.

# 7 Acknowledgements

This work was partially supported by NSF grant RI-IIS 1816753, NSF CAREER CIF 1845360, the Alfred P. Sloan Fellowship and Samsung Electronics. The job of C. Domingo Enrich was partially supported by CFIS (UPC). The work of A. Mensch was supported by the European Research Council (ERC project NORIA).

A. Mensch thanks Guillaume Garrigos and Timothée Lacroix for helpful comments.

# References

Bach, Francis and Kfir Levy (2019). "A universal algorithm for variational inequalities adaptive to smoothness and noise". In: *arXiv:1902.01637*.

Balduzzi, David et al. (2018). "The Mechanics of $n$-Player Differentiable Games". In: *Proceedings of the International Conference on Machine Learning*.

Bottou, Léon and Olivier Bousquet (2008). "The tradeoffs of large scale learning". In: *Advances in Neural Information Processing Systems*, pp. 161–168.

Bowling, Michael and Manuela Veloso (2001). "Rational and convergent learning in stochastic games". In: *Proceedings of the International Joint Conference on Artificial intelligence*, pp. 1021–1026.

Bu, Lucian, Robert Babu, Bart De Schutter, et al. (2008). "A comprehensive survey of multi-agent reinforcement learning". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38.2, pp. 156–172.

Bubeck, Sébastien (2015). "Convex Optimization: Algorithms and Complexity". In: *Foundations and Trends in Machine Learning* 8.3-4, pp. 231–357.

Chambolle, Antonin and Thomas Pock (2011). "A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging". en. In: *Journal of Mathematical Imaging and Vision* 40.1, pp. 120–145.

Chavdarova, Tatjana et al. (2019). "Reducing Noise in GAN Training with Variance Reduced Extragradient". In: *arXiv:1904.08598*.

Conitzer, Vincent and Tuomas Sandholm (2007). "AWESOME: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents". In: *Machine Learning* 67.1-2, pp. 23–43.

Daskalakis, Constantinos, Paul W Goldberg, and Christos H Papadimitriou (2009). "The complexity of computing a Nash equilibrium". In: *SIAM Journal on Computing* 39.1, pp. 195–259.

Daskalakis, Constantinos et al. (2018). "Training GANs with Optimism". In: *International Conference on Learning Representations*.

Defazio, Aaron, Francis Bach, and Simon Lacoste-Julien (2014). "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives". In: *Advances in Neural Information Processing Systems*, pp. 1646–1654.

Duchi, John, Elad Hazan, and Yoram Singer (2011). "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". In: *Journal of Machine Learning Research* 12, pp. 2121–2159.

Foerster, Jakob et al. (2018). "Learning with Opponent-Learning Awareness". In: *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*.

Gelfand, Izrail (1941). "Normierte ringe". In: *Matematicheskii Sbornik* 9.1, pp. 3–24.

Gidel, Gauthier et al. (2019). "A variational inequality perspective on generative adversarial networks". In: *International Conference on Learning Representations*.

Goodfellow, Ian et al. (2014). "Generative adversarial nets". In: *Advances in Neural Information Processing Systems*, pp. 2672–2680.

Gulrajani, Ishaan et al. (2017). "Improved training of Wasserstein GANs". In: *Advances in Neural Information Processing Systems*, pp. 5767–5777.

Harker, Patrick T and Jong-Shi Pang (1990). "Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications". In: *Mathematical Programming* 48.1-3, pp. 161–220.

He, Kaiming et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

Heusel, Martin et al. (2017). "GANs trained by a two time-scale update rule converge to a local Nash equilibrium". In: *Advances in Neural Information Processing Systems*, pp. 6626–6637.

Iusem, AN et al. (2017). "Extragradient method with variance reduction for stochastic variational inequalities". In: *SIAM Journal on Optimization* 27.2, pp. 686–724.

Juditsky, Anatoli, Arkadi Nemirovski, and Claire Tauvel (2011). "Solving variational inequalities with stochastic mirror-prox algorithm". In: *Stochastic Systems* 1.1, pp. 17–58.

Kim, Seung-Jean, Alessandro Magnani, and Stephen Boyd (2006). "Robust Fisher Discriminant Analysis". In: *Advances in Neural Information Processing Systems*.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations*.

Korpelevich, GM (1976). "The extragradient method for finding saddle points and other problems". In: *Matecon* 12, pp. 747–756.

Krizhevsky, Alex and Geoffrey Hinton (2009). *Learning multiple layers of features from tiny images*. Tech. rep. Citeseer.

Lee, Jason D et al. (2016). "Gradient descent only converges to minimizers". In: *Annual Conference on Learning Theory*, pp. 1246–1257.

Letcher, Alistair et al. (2019). "Stable Opponent Shaping in Differentiable Games". In: *International Conference on Learning Representations*.

Li, Xingguo et al. (2018). "On Faster Convergence of Cyclic Block Coordinate Descent-Type Methods for Strongly Convex Minimization". In: *Journal of Machine Learning Research* 18.184, pp. 1–24.

Lowe, Ryan et al. (2017). "Multi-agent actor-critic for mixed cooperative-competitive environments". In: *Advances in Neural Information Processing Systems*, pp. 6379–6390.

Magnanti, Thomas L and Georgia Perakis (1997). "Averaging schemes for variational inequalities and systems of equations". In: *Mathematics of Operations Research* 22.3, pp. 568–587.

Mazumdar, Eric V, Michael I Jordan, and S Shankar Sastry (2019). "On finding local Nash equilibria (and only local Nash equilibria) in zero-sum games". In: *arXiv:1901.00838*.

Mertikopoulos, Panayotis et al. (2019). "Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile". In: *International Conference on Learning Representations*.

Mescheder, Lars, Sebastian Nowozin, and Andreas Geiger (2017). "The numerics of GANs". In: *Advances in Neural Information Processing Systems*, pp. 1825–1835.

Mnih, Volodymyr et al. (2016). "Asynchronous methods for deep reinforcement learning". In: *Proceedings of the International Conference on Machine Learning*, pp. 1928–1937.

Nash, John (1951). "Non-cooperative games". In: *Annals of Mathematics*, pp. 286–295.

Neal, Radford M et al. (2011). "MCMC using Hamiltonian dynamics". In: *Handbook of Markov Chain Monte Carlo* 2.11, p. 2.

Nedić, Angelia and Asuman Ozdaglar (2009). "Subgradient methods for saddle-point problems". In: *Journal of Optimization Theory and Applications* 142.1, pp. 205–228.

Nemirovski, Arkadi (2004). "Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems". In: *SIAM Journal on Optimization* 15.1, pp. 229–251.

Nemirovski, Arkadi, Shmuel Onn, and Uriel G Rothblum (2010). "Accuracy certificates for computational problems with convex structure". In: *Mathematics of Operations Research* 35.1, pp. 52–78.

Nemirovsky, Arkadii Semenovich and David Borisovich Yudin (1983). *Problem complexity and method efficiency in optimization.* Wiley.

Nesterov, Yurii (2007). "Dual extrapolation and its applications to solving variational inequalities and related problems". In: *Mathematical Programming* 109.2-3, pp. 319–344.

Palaniappan, Balamurugan and Francis Bach (2016). "Stochastic variance reduction methods for saddle-point problems". In: *Advances in Neural Information Processing Systems*, pp. 1416–1424.

Racanière, Sébastien et al. (2017). "Imagination-augmented agents for deep reinforcement learning". In: *Advances in Neural Information Processing Systems*, pp. 5690–5701.

Robbins, Herbert and Sutton Monro (1951). "A Stochastic Approximation Method". In: *The Annals of Mathematical Statistics* 22.3, pp. 400–407.

Rockafellar, R. T. (1970). "Monotone operators associated with saddle-functions and minimax problems". In: *Proceedings of Symposia in Pure Mathematics*. Vol. 18.1, pp. 241–250.

Rosen, J Ben (1964). "Existence and uniqueness of equilibrium points for concave $n$-person games". In: *Econometrica* 3.33.

Salimans, Tim et al. (2016). "Improved Techniques for Training GANs". In: *Advances in Neural Information Processing Systems*, pp. 2234–2242.

Vezhnevets, Alexander Sasha et al. (2017). "Feudal networks for hierarchical reinforcement learning". In: *Proceedings of the International Conference on Machine Learning*, pp. 3540–3549.

Wayne, Greg and LF Abbott (2014). "Hierarchical control using networks trained with higher-level forward models". In: *Neural Computation* 26.10, pp. 2163–2193.

Yousefian, Farzad, Angelia Nedić, and Uday V Shanbhag (2018). "On stochastic mirror-prox algorithms for stochastic Cartesian variational inequalities: Randomized block coordinate and optimal averaging schemes". In: *Set-Valued and Variational Analysis* 26.4, pp. 789–819.

Zhang, Chongjie and Victor Lesser (2010). "Multi-Agent Learning with Policy Prediction". In: *Proceedings of the AAAI Conference on Artificial Intelligence*.

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. After detailing the relation between solving this problem and finding Nash equilibria in convex $n$-player games §A.2, we recall the rates for stochastic mirror-prox obtained by (Juditsky et al., 2011) in §A.3. We then present the proofs of our theorems in App. B. We analyze the doubly-stochastic algorithm (Alg. 1) and separately study two variants of the latter, adding importance sampling (§B.3) and variance-reduction (§B.4). App. C investigates the difference betwen random and cyclic player sampling. 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 $\Theta$ 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 $\Theta$, which is defined as a function $\Phi \colon \mathcal{D} \to \mathbb{R}$ that is differentiable and $\mu$-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 and Yudin (1983) and references therein. Given that $\Theta$ 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

$$P_z(\xi) \triangleq \operatorname*{argmin}_{u\in\mathcal{D}\cap\Theta}\{\Phi(u) + \langle \xi - \nabla\Phi(z), u\rangle\} = \operatorname*{argmin}_{u\in\mathcal{D}\cap\Theta}\{D(z,u) + \langle \xi, u\rangle\}. \tag{6}$$

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:

$$\begin{aligned} \text{Compute the extrapolated point: } & \begin{cases} \nabla\Phi(y_{\tau+1/2}) = \nabla\Phi(\theta_\tau) - \gamma F(\theta_\tau), \\ \theta_{\tau+1/2} = \operatorname{argmin}_{x\in\mathcal{D}\cap\Theta}\ D(x, y_{\tau+1/2}), \end{cases} \\ \text{Compute a gradient step: } & \begin{cases} \nabla\Phi(y_{\tau+1}) = \nabla\Phi(\theta_\tau) - \gamma F(\theta_{\tau+1/2}), \\ \theta_{\tau+1} = \operatorname{argmin}_{x\in\mathcal{D}\cap\Theta}\ D(x, y_{\tau+1}). \end{cases} \end{aligned} \tag{7}$$

Remark that the extra-gradient algorithm defined in equation (3) corresponds to the mirror-prox (7) when choosing $\Phi(x) = \frac{1}{2}\|x\|_2^2$.

**Lemma 1.** *By using the proximal mapping notation (6), 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 $\Theta_i$ 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 used this mirror map in the experiments for random quadratic games (§5.1).

## A.2 Link between convex games and variational inequalities

Finding a Nash equilibrium in a convex $n$-player game is related to solving a variational inequality (VI). Consider a space of parameters $\Theta \subseteq \mathbb{R}^d$ that is compact and convex, and consider a scalar product $\langle\cdot,\cdot\rangle$ in $\mathbb{R}^d$. The strong form of the VI associated to the operator $F : \Theta \to \mathbb{R}^d$ is

$$\text{find } \ \theta_* \in \Theta \ \text{ such that } \ \langle F(\theta_*), \theta - \theta_*\rangle \geqslant 0 \quad \forall \theta \in \Theta. \tag{8}$$

The weak form of the VI is

$$\text{find} \quad \theta_* \in \Theta \quad \text{such that} \quad \langle F(\theta), \theta - \theta_* \rangle \geqslant 0 \quad \forall \theta \in \Theta. \tag{9}$$

We define the concept of monotone operator.

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

If $F$ is monotone, a solution of the strong form of the VI is a solution of the weak form. The reciprocal implication is true when $F$ is continuous.

For convex $n$-player games (Ass. 1), the simultaneous (sub)gradient $F$ (Eq. 3.1) is a monotone operator. Moreover, if we assume continuity of the losses $l_i$, the set of weak solutions to the VI (9) coincides with the set of Nash equilibria. Solving the VI is therefore sufficient to find Nash equilibria (Harker and Pang, 1990; Nemirovski et al., 2010). The intuition behind this result is that equation (8) corresponds to the first-order necessary optimality condition applied to the losses of players.

The quantity that is used in the literature to quantify the inaccuracy of a solution $\theta$ is the dual VI gap defined as $\mathrm{Err}_{\mathrm{VI}}(\theta) = \max_{u \in \Theta} \langle F(u), \theta - u \rangle$. However, the *functional Nash error* (2) is the usual performance measure for convex games. In this article we give the convergence rates for the functional Nash error but they also apply to the dual VI gap. That is because the bound in Lemma 4 applies to the dual VI gap as well.

## A.3 Convergence rates for the stochastic mirror-prox

In this section, we recall the stochastic mirror-prox algorithm and its analysis by (Juditsky et al., 2011). Stochastic mirror-prox corresponds to Alg. 1 without subsampling over players i.e. setting the mini-batch size $b = n$. We start by giving the rates in terms of the number of iterations under Ass. 2a and Ass. 2b.

**Theorem 3** (From Juditsky et al. (2011))**.** *We consider a convex $n$-player game where Ass. 2a and Ass. 3 hold. Assume that Alg. 1 is run for $t$ iterations without subsampling ($b = n$) and with the optimal constant stepsize $\gamma$, the expected $Err_N(\hat{\theta}_t)$ is upper bounded as*

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_t)\right] \leqslant 7\sqrt{\frac{2\Omega n}{3t}(G^2 + 2\sigma^2)}.$$

*Assuming Ass. 2b (instead of Ass. 2a) and setting the optimal constant stepsize $\gamma$,*

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_t)\right] \leqslant \max\left\{\frac{7}{2}\frac{\Omega L}{t}, 14\sqrt{\frac{\Omega n \sigma^2}{3t}}\right\}.$$

To obtain a fair comparison with our results, we state these results in terms of the number of full gradients computations $k$.

**Corollary 1** (From Juditsky et al. (2011))**.** *We consider a convex $n$-player game where Ass. 2a and Ass. 3 hold. Assume that Alg. 1 is run for $t$ iterations without subsampling ($b = n$) and with the optimal constant stepsize $\gamma$, the expected $Err_N(\hat{\theta}_{k(t)})$ is upper bounded as*

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_{t(k)})\right] \leqslant 14n\sqrt{\frac{\Omega}{3k}(G^2 + 2\sigma^2)}. \tag{10}$$

*Assuming Ass. 2b (instead of Ass. 2a) and setting the optimal constant stepsize $\gamma$,*

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_{t(k)})\right] \leqslant \max\left\{\frac{7\Omega L n^{3/2}}{k}, 14n\sqrt{\frac{2\Omega\sigma^2}{3k}}\right\}. \tag{11}$$

## A.4 Player randomness as noise

The easiest way to treat player randomness on the theoretical level is to incorporate it in the unbiased gradient estimate. Indeed, in equation (5) $\tilde{F}_i(\theta, \mathcal{P})$ is an unbiased estimate of $\nabla_i l_i(\theta)$.

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

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

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

Substituting $\sigma^2$ by $2\frac{n-b}{b}G^2 + 2\sigma^2$ on equations (10) and (11) yields:

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

$$\begin{aligned}
\mathbb{E}\left[\text{Err}_N(\hat{\theta}_{t(k)})\right] &\leqslant \max\left\{\frac{7\Omega L n^{3/2}}{k}, 28n\sqrt{\frac{\Omega((n-b)G^2 + b\sigma^2)}{3kb}}\right\} \\
&= \mathcal{O}\left(\frac{\Omega L n^{3/2}}{k} + n\sqrt{\frac{\Omega(nG^2 + b\sigma^2)}{kb}}\right).
\end{aligned}$$

These bounds are clearly worse than the ones in Corollary 1 when $b \ll n$, which motivates the theoretical work in App. B that yields Theorem 1 and 2.

# B Proofs and mirror-setting algorithms

## B.1 Useful lemmas

In this section, we present lemmas that will be frequently used in the analysis of the algorithms in §B.2, §B.3 and §B.4. We first present the following two technical lemmas that are used and proven by Juditsky et al. (2011).

**Lemma 2.** *Let $z$ be a point in $\mathcal{X}$, let $\chi, \eta$ 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 $\xi_1, \xi_2, \dots$ 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 $y_\tau$ 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 provides an upper bound on the Nash functional error $\mathrm{Err}_N$. The following lemma provides an upper bound on the that the Nash functional error $\mathrm{Err}_N$ have the same upper bounds.

**Lemma 4.** *We consider a convex $n$-player game with players losses $\ell_i$ 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*

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

*Proof.* By using the convexity of $\ell_i$ in its parameter and its concavity in the others parameters and applying Jensen's inequality, we obtain:

$$\begin{aligned}\sum_{i=1}^{n} \ell_i(\hat{z}_t) - \ell_i(u^i, \hat{z}_t^{-i}) &= \sum_{i=1}^{n} \ell_i\left(\frac{\sum_{\tau=0}^{t} \gamma_\tau z_\tau}{\sum_{\tau=0}^{t} \gamma_\tau}\right) - \ell_i\left(u^i, \frac{\sum_{\tau=0}^{t} \gamma_\tau z_\tau^{-i}}{\sum_{\tau=0}^{t} \gamma_\tau}\right) \\ &\leqslant \left(\sum_{\tau=0}^{t} \gamma_\tau\right)^{-1} \sum_{\tau=0}^{t} \gamma_\tau \sum_{i=1}^{n} \ell_i(z_\tau) - \ell_i(u^i, z_\tau^{-i}). \end{aligned} \tag{12}$$

As a consequence of the convexity of $\ell_i$ with respect to its parameter, we have $\ell_i(z_\tau) - \ell_i(u^i, z_\tau^{-i}) \leqslant \langle h_i(z_\tau), z_\tau^i - u^i \rangle$ where $h_i(z_\tau) \in \partial_i \ell_i(z_\tau)$. Remark that $F = (h_1, ..., h_n)$. By plugging this inequality in (12), we obtain

$$\begin{aligned}\sum_{i=1}^{n} \ell_i(\hat{z}_t) - \ell_i(u^i, \hat{z}_t^{-i}) &\leqslant \left(\sum_{\tau=0}^{t} \gamma_\tau\right)^{-1} \sum_{\tau=0}^{t} \sum_{i=1}^{n} \langle \gamma_\tau h_i(z_\tau), z_\tau^i - u^i \rangle \\ &= \left(\sum_{\tau=0}^{t} \gamma_\tau\right)^{-1} \sum_{\tau=0}^{t} \langle \gamma_\tau F(z_\tau), z_\tau - u \rangle. \end{aligned}$$

□

**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 $\alpha_*$ the point at which the minimum is reached. The value of $f_t$ evaluated at $\alpha_*$ is*

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

*Proof.* It is sufficient to derive the first-order optimality condition of $f$:

$$-\frac{1}{\alpha_*^2}\frac{A}{\sum_{\tau=0}^t \gamma_\tau} + \frac{B\sum_{\tau=0}^t \gamma_\tau^2}{\sum_{\tau=0}^t \gamma_\tau} = 0,$$

and the result follows. □

**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.* We first prove that the LHS is smaller than the RHS. By definition of the dual norm, we have

$$\|a\|_{X^*}^2 = \sup_{y\in X} \frac{|ay|^2}{\|y\|_X^2} = \sup_{y\in X} \frac{\left(\sum_{i=1}^n a_i y_i\right)^2}{\|y\|_X^2} \leqslant \sup_{y\in X} \frac{\left(\sum_{i=1}^n \|a_i\|_{X_i^*}\|y_i\|_{X_i}\right)^2}{\|y\|_X^2}, \tag{13}$$

where we used Cauchy-Schwarz inequality. By applying again this inequality in (13), we obtain

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

which proves the result. To prove the other inequality we define $Z_i = \left\{y_i \in X_i \middle| \|y_i\|_X = \|a_i\|_{X_i^*}\right\}$.

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

□

## B.2 Doubly-stochastic mirror-prox

In this section, we present a more general version of Alg. 1. This algorithm (Alg. 3) is then analyzed in the rest of this section and in §B.3.

### B.2.1 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** $\tau = 0, \dots, t$ **do**
  Sample the random matrices $M_\tau, M_{\tau+1/2} \in \mathbb{R}^{d\times d}$.
  Compute $\tilde{F}_{\tau+1/2} = \frac{n}{b} \cdot M_\tau \hat{F}(\theta_\tau)$.
  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

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

where $g_i$ is a noisy unbiased estimate of $\nabla_i l_i(\theta)$ with variance bounded by $\sigma^2$. We are abusing the notation because $\hat{F}(\theta)$ is a random variable indexed by $\Theta$ 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 $M_\tau$ 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 $M_\tau$ 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 (6) 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.2.2 Convergence rate under Assumption 2a (non-smoothness)

**Theorem 4.** *We consider a convex $n$-player game where Ass. 2a holds. Assume that Alg. 3 is run with stepsizes $(\gamma_\tau)_{\tau\in t}$. In terms of the number of iterations $t$, the rate of convergence in expectation is*

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_t)\right] \leqslant \left(\sum_{\tau=0}^t \gamma_\tau\right)^{-1} \left(2\Omega + \sum_{\tau=0}^t \gamma_\tau^2 n \left(\frac{(3n-b)G^2}{b} + \sigma^2\right)\right). \tag{14}$$

*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

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

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 in 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.

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

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).

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

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

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

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

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

Applying Lemma 4 to equation (22) yields an upper bound on the functional Nash error and hence equation (14). □

Setting the constant stepsize that minimizes the rate in (14) implies the following corollary.

**Corollary 2.** *In Theorem 4, the constant stepsize $\gamma_\tau$ minimizing the rate (14) is*

$$\gamma_\tau = \gamma = \sqrt{\frac{2\Omega}{n\left(\frac{(3n-b)G^2}{b} + \sigma^2\right)t}}. \tag{23}$$

*Under this optimal stepsize choice, the rate (14) becomes*

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

*Proof.* For finding the optimal constant stepsize, we apply Lemma 5 to (14). This amounts to substituting $\gamma_\tau = 1$ for all $\tau \in [t]$, $A = 2\Omega$ and

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

Using the notation from Lemma 5, $\alpha_*$ yields the stepsize $\gamma$ in equation (23) and the minimum $f_t(\alpha_*)$ is the value of the bound (24). □

**Remark 1.** *For constant stepsizes, Corollary 2 implies that with an appropriate choice of $t$ and $\gamma$ we can achieve a value of the Nash error arbitrarily close to zero at time $t$. However, from Equation 14 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 but we do not make a detailed analysis of this case.*

**Corollary 3.** *Setting $\gamma_\tau$ as in Corollary 2, the expected convergence rate in terms of the number of gradient computations $k$ is*

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

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

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

which yields the desired result after simplification. □

### B.2.3 Convergence rate under Assumption 2b (smoothness)

In this subsection, we provide a convergence rate for Alg. 3 in the case where the losses are smooth (Ass. 2b). Remark that by continuity of $\nabla_i l_i$ and compactness of $\Omega$, $\forall i \in [n]$, $\exists G_i$ such that $\|\nabla_i \ell_i\|_* \leqslant G_i$. As in §B.2, we define the quantity $G = \sqrt{\sum_{i=1}^n G_i^2/n}$.

This result is not presented in the main paper because it is worse than the rate obtained under similar assumptions in Juditsky et al. (2011) (cf. §A.3). Having weaker guarantees for doubly-stochastic extra-gradient than for stochastic extra-gradient in the smooth case motivates the development of the variance reduction scheme in §B.4. Variance reduction achieves to similar rates to stochastic extra-gradient.

**Theorem 5.** *We consider a convex $n$-player game where Ass. 2b holds. Assume that Alg. 3 is run with stepsizes $(\gamma_\tau)_{\tau\in[t]}$ such that $\gamma_\tau < 1/(L\sqrt{3n})$. In terms of the number of iterations $t$, the rate of convergence in expectation is*

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_t)\right] \leqslant \left(\sum_{\tau=0}^{t}\gamma_\tau\right)^{-1}\left(2\Omega + \sum_{\tau=0}^{t}\frac{13\gamma_\tau^2 n}{2}\left(\frac{G^2(n-b)}{b} + \sigma^2\right).\right) \tag{25}$$

*Proof.* This proof is based on the same steps as in Theorem 4. By starting from equation (26), we obtain

$$\begin{aligned}
&\langle \gamma_\tau \tilde{F}_{\tau+1}, \theta_{\tau+1/2} - u\rangle + D(u, \theta_{\tau+1}) - D(u, \theta_\tau) \\
&\leqslant \frac{\gamma_\tau^2}{2}\|\tilde{F}_{\tau+1} - \tilde{F}_{\tau+1/2}\|_*^2 - \frac{1}{2}\|\theta_{\tau+1/2} - \theta_\tau\|_2^2 \\
&\leqslant \frac{3\gamma_\tau^2}{2}\|\tilde{F}_{\tau+1} - F(\theta_{\tau+1/2})\|_*^2 + \frac{3\gamma_\tau^2}{2}\|F(\theta_\tau) - \tilde{F}_{\tau+1/2}\|_*^2 + \frac{3\gamma_\tau^2}{2}\|F(\theta_{\tau+1/2}) - F(\theta_\tau)\|_*^2 \\
&- \frac{1}{2}\|\theta_{\tau+1/2} - \theta_\tau\|_2^2 \\
&\leqslant \frac{3\gamma_\tau^2}{2}\|\tilde{F}_{\tau+1} - F(\theta_{\tau+1/2})\|_*^2 + \frac{3\gamma_\tau^2}{2}\|F(\theta_\tau) - \tilde{F}_{\tau+1/2}\|_*^2 + \frac{3n\gamma_\tau^2 L^2}{2}\|\theta_{\tau+1/2} - \theta_\tau\|_*^2 \\
&- \frac{1}{2}\|\theta_{\tau+1/2} - \theta_\tau\|_2^2 \\
&\leqslant \frac{3\gamma_\tau^2}{2}\|\tilde{F}_{\tau+1} - F(\theta_{\tau+1/2})\|_*^2 + \frac{3\gamma_\tau^2}{2}\|F(\theta_\tau) - \tilde{F}_{\tau+1/2}\|_*^2, \qquad (26)
\end{aligned}$$

where in the last inequality we have used that $\gamma_t \leqslant 1/(L\sqrt{3n})$. We now apply the same reasoning to upper bound the expectation terms. These calculations lead to

$$\mathbb{E}\left[\|\tilde{F}_{\tau+1} - F(\theta_{\tau+1/2})\|_*^2\right] \leqslant \frac{2nG^2(1-p)}{p} + 2n\sigma^2, \qquad (27)$$

$$\mathbb{E}\left[\|F(\theta_\tau) - \tilde{F}_{\tau+1/2}\|_*^2\right] \leqslant \frac{2nG^2(1-p)}{p} + 2n\sigma^2. \qquad (28)$$

Therefore, by plugging the upper bounds obtained in (27) and (28) into equation (20) we get

$$\mathbb{E}\left[\sup_{u\in Z}\sum_{\tau=0}^{t}\langle \gamma_\tau F(\theta_{\tau+1/2}), \theta_{\tau+1/2} - u\rangle\right] \leqslant 2\Omega + \sum_{\tau=0}^{t}\frac{13\gamma_\tau^2 n}{2}\left(\frac{G^2(1-p)}{p} + \sigma^2\right).$$

By applying Lemma 4, we obtain equation (25). □

**Corollary 4.** *In Theorem 4, the constant stepsize $\gamma_\tau$ minimizing the rate (25) is*

$$\gamma_\tau = \gamma = \min\left\{\frac{1}{L\sqrt{3n}}, 2\sqrt{\frac{\Omega}{13n\left(\frac{G^2(n-b)}{b} + \sigma^2\right)t}}\right\}. \qquad (29)$$

*Under this optimal stepsize choice, the rate (25) becomes*

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_t)\right] \leqslant \max\left\{\frac{4L\Omega\sqrt{3n}}{t}, 2\sqrt{\frac{13\Omega n\left(\frac{G^2(n-b)}{b} + \sigma^2\right)}{t}}\right\}. \qquad (30)$$

*Proof.* We apply Lemma 5 to Theorem 5. We substitute $\gamma_\tau = 1$, $A = 2\Omega$ and

$$B = \frac{13}{2}n\left(\frac{G^2(n-b)}{b} + \sigma^2\right).$$

The second term of the minimum in the RHS of equation (29) is equal to $\alpha_*$ (using the notation from Lemma 5). If $\gamma = \alpha_*$, inequality (30) holds because the second term of the maximum is equal to $f_t(\alpha_*)$.

In the case $\gamma = 1/(L\sqrt{3n})$, we know $\alpha_* \geqslant \gamma$. Since $A/(\alpha_* t) = B\alpha_*$ by Lemma 5, we deduce $A/(\gamma t) \geqslant B\gamma$. Therefore,

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_t)\right] \leqslant \frac{A}{\gamma t} + B\gamma \leqslant \frac{2A}{\gamma t} = \frac{4L\Omega\sqrt{3n}}{t}.$$

□

**Corollary 5.** *Setting $\gamma_\tau$ as in Corollary 4, the expected convergence rate in terms of the number of gradient computations $k$ is*

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_{t(k)})\right] \leqslant \max\left\{\frac{8bL\Omega\sqrt{3n}}{k}, 2\sqrt{\frac{26\Omega\left(G^2n(n-b)+\sigma^2nb\right)}{k}}\right\}$$

*Proof.* It is sufficient to plug $t(k) = k/2b$ into (30) to get the result. □

## B.3 Doubly-stochastic mirror-prox with importance sampling

### B.3.1 Setting

Different sampling strategies over players can be chosen in Alg. 3. In this section, we consider importance sampling and derive convergence rates in this setting. We assume that we select one player among $n$ ($b = 1$). We choose player $i$ with probability

$$p_i = \frac{G_i}{\sum_{k=1}^n G_k}. \tag{31}$$

More formally, this mean that we choose one identity block among $n$ according to the distribution $\mathcal{D}$ in the random matrices $M_\tau$ and $M_{\tau+1/2}$.

### B.3.2 Analysis

**Theorem 6.** *We consider a convex $n$-player game where Ass. 2a holds. Assume that Alg. 3 is run with importance sampling and stepsizes $(\gamma_\tau)_{\tau\in[t]}$. In terms of the number of iterations $t$, the rate of convergence in expectation is*

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_t)\right] \leqslant \left(\sum_{\tau=0}^t \gamma_\tau\right)^{-1}\left(2\Omega + \sum_{\tau=0}^t \gamma_\tau^2\left(3\left(\sum_{i=0}^n G_i\right)^2 - nG^2 + n\sigma^2\right)\right). \tag{32}$$

*Proof.* The proof strategy is similar to the proof of Theorem 4. The only difference lies in the bounds of the expectations terms $\mathbb{E}\left[\|\tilde{F}_{\tau+1} - \tilde{F}_{\tau+1/2}\|_\star^2\right]$ and $\mathbb{E}\left[\|F(\theta_{\tau+1/2}) - \tilde{F}_{\tau+1}\|_\star^2\right]$.

We use the same filtrations $\mathcal{F}_\tau$ and $\mathcal{F}'_\tau$ defined in the proof of Theorem 4. We start by bounding the first expectation term. By using Lemma 6 in the first equality, $\|a+b\|_*^2 \leqslant 2\|a\|_*^2 + 2\|b\|_*^2$ in the second inequality, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{F}_{\tau+1} - \tilde{F}_{\tau+1/2}\|_\star^2\right] &= \sum_{i=1}^n \mathbb{E}\left[\|\tilde{F}^{(i)}_{\tau+1} - \tilde{F}^{(i)}_{\tau+1/2}\|_\star^2\right] \\
&\leqslant \sum_{i=1}^n 2\left(\mathbb{E}\left[\|\tilde{F}^{(i)}_{\tau+1}\|_\star^2\right] + \mathbb{E}\left[\|\tilde{F}^{(i)}_{\tau+1/2}\|_\star^2\right]\right) \\
&= \sum_{i=1}^n \frac{2}{p_i^2}\left(\mathbb{E}\left[\mathbb{E}\left[\|M^{(i)}_{\tau+1/2}\hat{F}^{(i)}(\theta_{\tau+1/2})\|_\star^2|\mathcal{F}_\tau\right]\right]\right. \\
&\qquad \left.+\mathbb{E}\left[\mathbb{E}\left[\|M^{(i)}_{\tau}\hat{F}^{(i)}(\theta_{\tau})\|_\star^2|\mathcal{F}'_\tau\right]\right]\right) \\
&\leqslant \sum_{i=1}^n \frac{2}{p_i}\left(\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{\tau+1/2})\|_\star^2\right] + \mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{\tau})\|_\star^2\right]\right) \\
&\leqslant \sum_{i=1}^n \frac{2\sum_{k=1}^n G_k}{G_i}\cdot 2G_i^2 \\
&= 4\left(\sum_{i=1}^n G_i\right)^2,
\end{aligned} \tag{33}$$

where we used (31) in the last equality. Regarding the second expectation term, by exactly applying the same steps as above, we obtain

$$
\begin{aligned}
\mathbb{E}\left[\|F(\theta_{\tau+1/2}) - \tilde{F}_{\tau+1}\|_\star^2\right] &= \sum_{i=1}^n \mathbb{E}\left[\|F^{(i)}(\theta_{\tau+1/2}) - \tilde{F}^{(i)}_{\tau+1}\|_\star^2\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[\left\|F^{(i)}(\theta_{\tau+1/2}) - \frac{M^{(i)}_{\tau+1}}{p_i}\hat{F}^{(i)}(\theta_{\tau+1/2})\right\|_\star^2\right] \\
&\leqslant \sum_{i=1}^n 2\mathbb{E}\left[\left\|\left(I - \frac{M^{(i)}_{\tau+1}}{p_i}\right)\hat{F}^{(i)}(\theta_{\tau+1/2})\right\|_\star^2\right] \\
&+ \sum_{i=1}^n 2\mathbb{E}\left[\left\|F^{(i)}(\theta_{\tau+1/2}) - \hat{F}^{(i)}(\theta_{\tau+1/2})\right\|_\star^2\right] \\
&\leqslant \sum_{i=1}^n 2\mathbb{E}\left[p_i\left\|\frac{p_i - 1}{p_i}\hat{F}^{(i)}(\theta_{\tau+1/2})\right\|_\star^2 + (1-p_i)\|\hat{F}^{(i)}(\theta_{\tau+1/2})\|_\star^2\right] \\
&+ 2n\sigma^2 \\
&= \sum_{i=1}^n 2\left(1 - p_i + \frac{(1-p_i)^2}{p_i}\right)\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{\tau+1/2})\|_\star^2\right] + 2n\sigma^2 \\
&= \sum_{i=1}^n 2\left(\frac{1}{p_i} - 1\right)G_i^2 + 2n\sigma^2 \\
&\leqslant \sum_{i=1}^n \frac{2\sum_{k=1}^n G_k}{G_i}G_i^2 - 2G_i^2 + 2n\sigma^2 \\
&\leqslant 2\left(\sum_{i=1}^n G_i\right)^2 - 2nG^2 + 2n\sigma^2. \qquad (34)
\end{aligned}
$$

Combining the equations (18) and (19) from the proof of Theorem 4 and the upper bounds obtained in (33) and (34), we finally get the result (32). □

**Remark 2.** *By arithmetic mean-quadratic mean inequality,*

$$\sum_{i=1}^n G_i^2 \geqslant \frac{(\sum_{i=1}^n G_i)^2}{n}.$$

*This implies that the bound with importance sampling (32) is lower or equal than the one with uniform sampling (14).*

**Remark 3.** *Corollary 2 and Corollary 3 hold when we substitute $n^2G^2/b$ by $(\sum_{i=1}^n G_i)^2$.*

## B.4 Doubly-stochastic mirror-prox with variance reduction

### B.4.1 Algorithm

Analogously to previous sections, we present a version of Alg. 1 with variance reduction in the mirror framework and in the new notation.

$\tilde{F}(\theta)$ 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 $R_\tau$ is updated when possible.

This algorithm is relevant because its convergence rate in the smooth case is very similar to the one derived by Juditsky et al. (2011) under the same assumptions (see Theorem 3 and Corollary 1).

**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** $\hat{\theta}_t = \left[\sum_{\tau=0}^t \gamma_\tau\right]^{-1} \sum_{\tau=0}^t \gamma_\tau \theta_\tau$.

### B.4.2 Convergence rate under Assumption 2b (smoothness)

**Preliminaries.** We define $K_j$, a random variable that will be used in Lemma 11.

**Definition 4.** *For a given $j$, 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, $R_{j/2,i} = \hat{F}^{(i)}(\theta_{K_j/2})$.

**Remark 4.** *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 $P(K_j = 0) = 1 - \sum_{q=1}^{j-1} P(K_j = q) = (1-p)^{j-1}$.*

The following inequality 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.*

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

□

**Proof.** We provide an outline of the proof:

- Lemma 9 follows the structure of the proof in §B.2.2 and in (Juditsky et al., 2011). The rest of the proof is particular to variance reduction and not related to the arguments in (Juditsky et al., 2011).
- Lemma 12 provides a bound for $\mathbb{E}\left[\sum_{\tau=0}^{t}\gamma_\tau^2\|\tilde{F}_{\tau+1}-F(\theta_{\tau+1/2})\|_\star^2+\gamma_\tau^2\|F(\theta_\tau)-\tilde{F}_{\tau+1/2}\|_\star^2\right]$, which is one of the terms that appear in the bound of Lemma 9. Lemma 10 and 11 are intermediate steps in the proof of 12.
- Theorem 7 follows from Lemma 9 and 12. It provides the convergence rate for generic stepsize sequences and it is analogous to Theorem 4 from the non-smooth case.
- Corollary 6 provides the convergence rate for constant stepsizes and Corollary 7 gives the convergence rate for constant stepsizes in terms of $\chi$. They are analogous to Corollary 2 and 3.

**Lemma 9.** *The following holds:*

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

*Proof.* 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 $y_0=\theta_0$.

$$\begin{aligned}\sum_{\tau=0}^{t}\langle\gamma_\tau\Delta_\tau,y_\tau-u\rangle&\leqslant D(u,\theta_0)+\sum_{\tau=0}^{t}\frac{\gamma_\tau^2}{2}\|\Delta_\tau\|_\star^2\\&=D(u,\theta_0)+\sum_{\tau=0}^{t}\frac{\gamma_\tau^2}{2}\|F(\theta_{\tau+1/2})-\tilde{F}_{\tau+1}\|_\star^2.\end{aligned}\tag{35}$$

Using equation (35) and the analogous equation to (19), we get the desired inequality. □

The main challenge is to bound $\mathbb{E}\left[\sum_{\tau=0}^{t}\gamma_\tau^2\|\tilde{F}_{\tau+1}-F(\theta_{\tau+1/2})\|_\star^2+\gamma_\tau^2\|F(\theta_\tau)-\tilde{F}_{\tau+1/2}\|_\star^2\right]$, which we can also express as

$$\begin{aligned}&\mathbb{E}\left[\sum_{\tau=0}^{t}\gamma_\tau^2\|\tilde{F}_{\tau+1}-F(\theta_{\tau+1/2})\|_\star^2+\gamma_\tau^2\|F(\theta_\tau)-\tilde{F}_{\tau+1/2}\|_\star^2\right]\\&=\mathbb{E}\left[\sum_{\tau=0}^{t}\sum_{i=1}^{n}\gamma_\tau^2\|\tilde{F}_{\tau+1}^{(i)}-F^{(i)}(\theta_{\tau+1/2})\|_\star^2+\gamma_\tau^2\|F^{(i)}(\theta_\tau)-\tilde{F}_{\tau+1/2}^{(i)}\|_\star^2\right],\end{aligned}\tag{36}$$

where we use Lemma 6.

**Lemma 10.** *The following equalities hold:*

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

*Proof.* Using the conditional expectation with respect to the filtration up to $w_\tau$,

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

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.** *Let us define $h : \mathbb{R} \to \mathbb{R}$ as*

$$h(p) \triangleq \frac{2-p}{p^2}. \tag{37}$$

*Assume that $(\gamma_\tau)_{\tau\in\mathbb{N}}$ is non-increasing. Then, the following holds:*

$$\sum_{\tau=0}^{t}\gamma_\tau^2\mathbb{E}\left[\|R^{(i)}_\tau - \hat{F}^{(i)}(\theta_\tau)\|_\star^2\right] \leqslant \sum_{j=0}^{2t-1} h(p)\gamma^2_{\lfloor j/2\rfloor}\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{j/2}) - \hat{F}^{(i)}(\theta_{(j+1)/2})\|_\star^2\right], \tag{38}$$
$$\sum_{\tau=0}^{t}\gamma_\tau^2\mathbb{E}\left[\|R^{(i)}_{\tau+1/2} - \hat{F}^{(i)}(\theta_{\tau+1/2})\|_\star^2\right] \leqslant \sum_{j=0}^{2t-1} h(p)\gamma^2_{\lfloor j/2\rfloor}\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{j/2}) - \hat{F}^{(i)}(\theta_{(j+1)/2})\|_\star^2\right].$$

*Proof.* We can write

$$\begin{aligned}
\mathbb{E}\left[\|R_\tau^{(i)} - \hat{F}^{(i)}(\theta_\tau)\|_\star^2\right] &= \mathbb{E}\left[\|R_{2\tau/2}^{(i)} - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\|R_{2\tau/2}^{(i)} - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2 \middle| K_{2\tau}\right]\right] \\
&= \sum_{q=0}^{2\tau-1} P(K_{2\tau}=q)\mathbb{E}\left[\|R_{2\tau/2}^{(i)} - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2 \middle| K_{2\tau}=q\right] \\
&= \sum_{q=1}^{2\tau-1} p(1-p)^{2\tau-1-q}\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{q/2}) - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2\right] \\
&+ (1-p)^{2\tau-1}\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_0) - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2\right].
\end{aligned} \tag{39}$$

As seen in equation (39), the point of conditioning with respect to the sigma-field generated by $K_{2\tau}$ is that we can write the expression for $R_{2\tau/2,i}$.

Now, using the rearrangement inequality,

$$\begin{aligned}
\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{q/2}) - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2\right] &= \mathbb{E}\left[\left\|\sum_{j=q}^{2\tau-1} \hat{F}^{(i)}(\theta_{j/2}) - \hat{F}^{(i)}(\theta_{(j+1)/2})\right\|_\star^2\right] \\
&\leqslant \sum_{j=q}^{2\tau-1}(2\tau-q)\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{j/2}) - \hat{F}^{(i)}(\theta_{(j+1)/2})\|_\star^2\right].
\end{aligned} \tag{40}$$

Using equations (39) and (40) we can now write

$$\begin{aligned}
&\sum_{\tau=0}^{t}\gamma_\tau^2\mathbb{E}\left[\|R_\tau^{(i)} - \hat{F}^{(i)}(\theta_\tau)\|_\star^2\right] \\
&= \sum_{\tau=0}^{t}\gamma_\tau^2\sum_{q=1}^{2\tau-1} p(1-p)^{2\tau-1-q}\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{q/2}) - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2\right] \\
&+ \gamma_\tau^2(1-p)^{2\tau-1}\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_0) - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2\right] \\
&\leqslant \sum_{\tau=0}^{t}\gamma_\tau^2\sum_{q=1}^{2\tau-1} p(1-p)^{2\tau-1-q}\sum_{j=q}^{2\tau-1}(2\tau-q)\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{j/2}) - \hat{F}^{(i)}(\theta_{(j+1)/2})\|_\star^2\right] \\
&+ \gamma_\tau^2(1-p)^{2\tau-1}\sum_{j=0}^{2\tau-1} 2\tau\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{j/2}) - \hat{F}^{(i)}(\theta_{(j+1)/2})\|_\star^2\right].
\end{aligned} \tag{41}$$

Given $j$ between 0 and $2t-1$ the right hand side of equation (41) contains the term

$\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{j/2}) - \hat{F}^{(i)}(\theta_{(j+1)/2})\|_2^2\right]$ multiplied by

$$\begin{aligned}
&\sum_{\tau=\lceil (j+1)/2\rceil}^{t} \gamma_\tau^2 \left(\sum_{r=1}^{j}(2\tau - r)p(1-p)^{2\tau-1-r} + 2\tau(1-p)^{2\tau-1}\right) \\
&\leqslant \gamma_{\lfloor j/2\rfloor}^2 \sum_{\tau=\lceil (j+1)/2\rceil}^{t} \sum_{r=1}^{j}(2\tau - r)p(1-p)^{2\tau-1-r} + 2\tau(1-p)^{2\tau-1} \\
&= \gamma_{\lfloor j/2\rfloor}^2 \sum_{\tau=\lceil (j+1)/2\rceil}^{t} p\sum_{r'=0}^{j-1}(1-p)^{2\tau-1-j+r'}(2\tau - j + r') + 2\tau(1-p)^{2\tau-1} \\
&\leqslant \gamma_{\lfloor j/2\rfloor}^2 \sum_{\tau=\lceil (j+1)/2\rceil}^{t} p\sum_{r'=2\tau-j}^{\infty}(1-p)^{r'-1}r' = (*).
\end{aligned}$$

Using Lemma 8 twice:

$$\begin{aligned}
(*) &= \gamma_{\lfloor j/2\rfloor}^2 \sum_{\tau=\lceil (j+1)/2\rceil}^{t} p\frac{(2\tau - j)(1-p)^{2\tau-1-j}p + (1-p)^{2\tau-j}}{p^2} \\
&= \gamma_{\lfloor j/2\rfloor}^2 \sum_{\tau=\lceil (j+1)/2\rceil}^{t} \frac{(2\tau - j)(1-p)^{2\tau-1-j}p + (1-p)^{2\tau-j}}{p} \\
&\leqslant \gamma_{\lfloor j/2\rfloor}^2 \sum_{\tau=2\lceil (j+1)/2\rceil}^{\infty} (\tau - j)(1-p)^{\tau-1-j} + \frac{\gamma_{\lfloor j/2\rfloor}^2}{p}\sum_{\tau=2\lceil (j+1)/2\rceil}^{\infty}(1-p)^{\tau-j} \\
&= \gamma_{\lfloor j/2\rfloor}^2 \sum_{\tau=2\lceil (j+1)/2\rceil-j}^{\infty} \tau(1-p)^{\tau-1} + \frac{\gamma_{\lfloor j/2\rfloor}^2}{p}\sum_{\tau=2\lceil (j+1)/2\rceil-j}^{\infty}(1-p)^{\tau} \\
&= \gamma_{\lfloor j/2\rfloor}^2 \frac{(2\lceil (j+1)/2\rceil - j)(1-p)^{2\lceil (j+1)/2\rceil-j-1}p + 2(1-p)^{2\lceil (j+1)/2\rceil-j}}{p^2}.
\end{aligned}$$

By Lemma 7 we have

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

Hence, from equation (41) we get

$$\sum_{\tau=0}^{t}\gamma_\tau^2\mathbb{E}\left[\|R_\tau^{(i)} - \hat{F}^{(i)}(\theta_\tau)\|_\star^2\right] \leqslant \sum_{j=0}^{2t-1}\gamma_{\lfloor j/2\rfloor}^2 h(p)\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{j/2}) - \hat{F}^{(i)}(\theta_{(j+1)/2})\|_\star^2\right].$$

Analogously to equation (39):

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

Using the same reasoning we get an inequality that is analogous to (38):

$$\sum_{\tau=0}^{t}\gamma_\tau^2\mathbb{E}\left[\|R^{(i)}_{\tau+1/2} - \hat{F}^{(i)}(\theta_{\tau+1/2})\|_\star^2\right] \leqslant \sum_{j=0}^{2t}\gamma^2_{\lfloor j/2\rfloor}h(p)\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{j/2}) - \hat{F}^{(i)}(\theta_{(j+1)/2})\|_\star^2\right]. \qquad \square$$

**Lemma 12.** *Assume that for all $i$ between 1 and $n$, the gradients $\nabla_i \ell_i$ are $L$-Lipschitz. Assume that for all $\tau$ between 0 and $t$, $\gamma_\tau \leqslant \gamma$. Let*

$$\chi(p,\gamma) = 1 - 36\frac{1-p}{p}nh(p)L^2\gamma^2.$$

*If $\gamma$ is small enough that $\chi(p,\gamma)$ is positive, then*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\tau=0}^{t}\gamma_\tau^2\|\tilde{F}_{\tau+1} - F(\theta_{\tau+1/2})\|_\star^2 + \gamma_\tau^2\|F(\theta_\tau) - \tilde{F}_{\tau+1/2}\|_\star^2\right] \\
&\leqslant 104n\sigma^2\sum_{\tau=0}^{t}\gamma_\tau^2 + \frac{1-p}{p\chi(p,\gamma)}(12L^2 + 36L^4\gamma^2)nh(p)\sum_{\tau=0}^{t}\gamma_\tau^2\mathbb{E}\left[\|\theta_\tau - \theta_{\tau+1/2}\|_\star^2\right].
\end{aligned}$$

*Proof.* We first want to bound the terms $\mathbb{E}\left[\|F^{(i)}(\theta_{j/2}) - F^{(i)}(\theta_{(j+1)/2})\|_2^2\right]$. When $j$ is even we can make the change of variables $j/2 = \tau$ (just for simplicity in the notation) and use smoothness. We get

$$\begin{aligned}
\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{j/2}) - \hat{F}^{(i)}(\theta_{(j+1)/2})\|_\star^2\right] &= \mathbb{E}\left[\|\hat{F}^{(i)}(\theta_\tau) - \hat{F}^{(i)}(\theta_{\tau+1/2})\|_\star^2\right] \\
&\leqslant 3\mathbb{E}\left[\|F^{(i)}(\theta_\tau) - F^{(i)}(\theta_{\tau+1/2})\|_\star^2\right] \\
&+ 3\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_\tau) - F^{(i)}(\theta_{\tau+1/2})\|_\star^2\right] \\
&+ 3\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{\tau+1/2}) - F^{(i)}(\theta_{\tau+1/2})\|_\star^2\right] \\
&\leqslant 3L^2\mathbb{E}\left[\|\theta_\tau - \theta_{\tau+1/2}\|_\star^2\right] + 6\sigma^2.
\end{aligned} \tag{42}$$

When $j$ is odd, we can write $j/2 = \tau + 1/2$. We use smoothness and the fact that the prox-mapping is

1-Lipschitz (Lemma 2):

$$\begin{aligned}
\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{j/2}) - \hat{F}^{(i)}(\theta_{(j+1)/2})\|_\star^2\right] &= \mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{\tau+1/2}) - \hat{F}^{(i)}(\theta_{\tau+1})\|_\star^2\right] \\
&\leqslant 3\mathbb{E}\left[\|F^{(i)}(\theta_{\tau+1/2}) - F^{(i)}(\theta_{\tau+1})\|_\star^2\right] \\
&+ 3\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{\tau+1/2}) - F^{(i)}(\theta_{\tau+1/2})\|_\star^2\right] \\
&+ 3\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{\tau+1}) - F^{(i)}(\theta_{\tau+1})\|_\star^2\right] \\
&\leqslant 3L^2\mathbb{E}\left[\|\theta_{\tau+1/2} - \theta_{\tau+1}\|_\star^2\right] + 6\sigma^2 \\
&= 3L^2\mathbb{E}\left[\|P_{\theta_\tau}(\gamma_\tau\tilde{F}_{\tau+1/2}) - P_{\theta_\tau}(\gamma_\tau\tilde{F}_{\tau+1})\|_\star^2\right] + 6\sigma^2 \\
&\leqslant 3L^2\gamma_\tau^2\mathbb{E}\left[\|\tilde{F}_{\tau+1/2} - \tilde{F}_{\tau+1}\|_\star^2\right] + 6\sigma^2 \\
&\leqslant 9L^2\gamma_\tau^2\left(\mathbb{E}\left[\|\tilde{F}_{\tau+1/2} - F(\theta_\tau)\|_\star^2\right]\right. \\
&\qquad +\mathbb{E}\left[\|F(\theta_{\tau+1/2}) - \tilde{F}_{\tau+1}\|_\star^2\right] \\
&\qquad \left.+\mathbb{E}\left[\|F(\theta_\tau) - F(\theta_{\tau+1/2})\|_\star^2\right]\right) + 6\sigma^2.
\end{aligned} \tag{43}$$

Hence, from equation (36) and Lemma 10 and 11:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\tau=0}^{t}\gamma_\tau^2\|\tilde{F}_{\tau+1} - F(\theta_{\tau+1/2})\|_\star^2 + \gamma_\tau^2\|F(\theta_\tau) - \tilde{F}_{\tau+1/2}\|_\star^2\right] \\
&\leqslant 4n\sigma^2\sum_{\tau=0}^{t}\gamma_\tau^2 + \frac{2(1-p)}{p}\mathbb{E}\left[\sum_{\tau=0}^{t}\sum_{i=1}^{n}\gamma_t^2\|R_\tau^{(i)} - \hat{F}^{(i)}(\theta_\tau)\|^2 + \|R_{\tau+1/2}^{(i)} - \hat{F}^{(i)}(\theta_\tau)\|^2\right] \\
&\leqslant 4n\sigma^2\sum_{\tau=0}^{t}\gamma_\tau^2 + \frac{2(1-p)}{p}\sum_{i=1}^{n}\sum_{j=0}^{2t}2\gamma_{\lfloor j/2\rfloor}^2 h(p)\mathbb{E}\left[\|F_i(\theta_{j/2}) - F_i(\theta_{(j+1)/2})\|_\star^2\right] = (**).
\end{aligned}$$

We split the last term in summands corresponding to even and odd $j$, we change variables from $j$ to $\tau$ and

we apply equations (42) and (43):

$$
\begin{aligned}
(**) &= 4n\sigma^2 \sum_{\tau=0}^{t} \gamma_\tau^2 + \frac{2(1-p)}{p} \sum_{i=1}^{n} \sum_{j=0,\ j \text{ even}}^{2t} 2\gamma_{\lfloor j/2 \rfloor}^2 h(p) \mathbb{E}\left[\|F_i(\theta_{j/2}) - F_i(\theta_{(j+1)/2})\|_\star^2\right] \\
&+ \frac{2(1-p)}{p} \sum_{i=1}^{n} \sum_{j=0,\ j \text{ odd}}^{2t} 2\gamma_{\lfloor j/2 \rfloor}^2 h(p) \mathbb{E}\left[\|F_i(\theta_{j/2}) - F_i(\theta_{(j+1)/2})\|_\star^2\right] \\
&= 4n\sigma^2 \sum_{\tau=0}^{t} \gamma_\tau^2 + \frac{2(1-p)}{p} \sum_{i=1}^{n} \sum_{\tau=0}^{t} 2\gamma_\tau^2 h(p) \mathbb{E}\left[\|F_i(\theta_\tau) - F_i(\theta_{\tau+1/2})\|_\star^2\right] \\
&+ \frac{2(1-p)}{p} \sum_{i=1}^{n} \sum_{\tau=0}^{t} 2\gamma_\tau^2 h(p) \mathbb{E}\left[\|F_i(\theta_{\tau+1/2}) - F_i(\theta_{\tau+1})\|_\star^2\right] \\
&\leqslant 52n\sigma^2 \sum_{\tau=0}^{t} \gamma_\tau^2 + \frac{1-p}{p} \sum_{\tau=0}^{t} 12n\gamma_\tau^2 h(p) L^2 \mathbb{E}\left[\|\theta_\tau - \theta_{\tau+1/2}\|_\star^2\right] \\
&+ \frac{1-p}{p} \sum_{\tau=0}^{t} 36nh(p) L^2 \gamma_\tau^4 \left(\mathbb{E}\left[\|\tilde{F}_{\tau+1/2} - F(\theta_\tau)\|_\star^2\right] + \mathbb{E}\left[\|F(\theta_{\tau+1/2}) - \tilde{F}_{\tau+1}\|_\star^2\right]\right) \\
&+ \frac{1-p}{p} \sum_{\tau=0}^{t} 36nh(p) L^4 \gamma_\tau^4 \mathbb{E}\left[\|\theta_\tau - \theta_{\tau+1/2}\|_\star^2\right] = (***).
\end{aligned}
$$

We use that $\gamma_\tau \leqslant \gamma$:

$$
\begin{aligned}
(***) &\leqslant 52n\sigma^2 \sum_{\tau=0}^{t} \gamma_\tau^2 + \frac{1-p}{p}(12L^2 + 36L^4\gamma^2) nh(p) \sum_{\tau=0}^{t} \gamma_\tau^2 \mathbb{E}\left[\|\theta_\tau - \theta_{\tau+1/2}\|_\star^2\right] \\
&+ 36\frac{1-p}{p} nh(p) L^2 \gamma^2 \sum_{\tau=0}^{t} \gamma_\tau^2 \left(\mathbb{E}\left[\|\tilde{F}_{\tau+1/2} - F(\theta_\tau)\|_\star^2\right] + \mathbb{E}\left[\|F(\theta_{\tau+1/2}) - \tilde{F}_{\tau+1}\|_\star^2\right]\right).
\end{aligned}
$$

Rearranging and using $\chi(p,\gamma) > 0$ yields the desired result. □

**Theorem 7.** *Assume that for all $i$ between 1 and $n$, the gradients $\nabla_i \ell_i$ are $L$-Lipschitz. Let $(\hat{\theta}_t)_{t\in\mathbb{N}}$ be defined as in Alg. 4. Choose $(\gamma_t)_{t\in\mathbb{N}}$ such that $\gamma_t \leqslant \gamma$, with $\gamma$ defined as*

$$
\gamma \triangleq \min\left\{ \frac{p^{3/2}}{\sqrt{(1-p)(2-p)}} \frac{1}{12L\sqrt{n}}, \frac{1}{L}\sqrt{\frac{5}{27n+12}} \right\}, \tag{44}
$$

*where $p \triangleq b/n$. Then,*

$$
\mathrm{Err}_N(\hat{\theta}_t) \leqslant \left(\sum_{\tau=0}^{t} \gamma_\tau\right)^{-1} \left(2\Omega + 104n\sigma^2 \sum_{\tau=0}^{t} \gamma_\tau^2\right).
$$

*Proof.* It is easy to see that

$$
\gamma \leqslant \frac{p^{3/2}}{\sqrt{(1-p)(2-p)}} \frac{1}{12L\sqrt{n}} \iff \chi(p,\gamma) \geqslant 3/4 > 0.
$$

Hence, the assumptions of Lemma 12 are fulfilled. Starting from the result in Lemma 9 and using Lemma 12,

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{u\in Z}\sum_{\tau=0}^{t}\langle\gamma_\tau F(\theta_{\tau+1/2}),\theta_{\tau+1/2}-u\rangle\right] \qquad (45)\\
&\leqslant \mathbb{E}\left[\sup_{u\in Z} 2D(u,\theta_0)-D(u,\theta_t)\right]+32n\sigma^2\sum_{\tau=0}^{t}\gamma_\tau^2\\
&+2\frac{1-p}{p\chi(p,\gamma)}(12L^2+36L^4\gamma^2)nh(p)\sum_{\tau=0}^{t}\gamma_\tau^2\mathbb{E}\left[\|\theta_\tau-\theta_{\tau+1/2}\|_\star^2\right]\\
&+\frac{3nL^2}{2}\sum_{\tau=0}^{t}\gamma_\tau^2\mathbb{E}\left[\|\theta_\tau-\theta_{\tau+1/2}\|_\star^2\right]-\frac{1}{2}\sum_{\tau=0}^{t}\mathbb{E}\left[\|\theta_\tau-\theta_{\tau+1/2}\|_\star^2\right]\\
&\leqslant 2\Omega+104n\sigma^2\sum_{\tau=0}^{t}\gamma_\tau^2\\
&+\left((24L^2+72L^4\gamma^2)nh(p)\gamma^2\frac{1-p}{p\chi(p,\gamma)}+\frac{3n\gamma^2L^2}{2}-\frac{1}{2}\right)\sum_{\tau=0}^{t}\mathbb{E}\left[\|\theta_\tau-\theta_{\tau+1/2}\|_\star^2\right].
\end{aligned}
$$

Recalling the definition of $h(p)$ in Equation (37), the conditions $\chi(p,\gamma)\geqslant 3/4$ and

$$\gamma\leqslant\frac{1}{L}\sqrt{\frac{5}{27n+12}},$$

imply

$$(24L^2+72L^4\gamma^2)nh(p)\gamma^2\frac{1-p}{p\chi(p,\gamma)}+\frac{3n\gamma^2L^2}{2}-\frac{1}{2}\leqslant 0. \qquad (46)$$

We show this development because it is not entirely trivial:

$$
\begin{aligned}
&(24L^2+72L^4\gamma^2)n\frac{2-p}{p^2}\gamma^2\frac{1-p}{p\chi(p,\gamma)}+\frac{3n\gamma^2L^2}{2}-\frac{1}{2}\\
&\overset{\chi\geqslant 3/4}{\leqslant}(24L^2+72L^4\gamma^2)n\frac{2-p}{p^2}\gamma^2\frac{4(1-p)}{3p}+\frac{3n\gamma^2L^2}{2}-\frac{1}{2}\\
&=\frac{24+72L^2\gamma^2}{27}(1-\chi(p,\gamma))+\frac{3n\gamma^2L^2}{2}-\frac{1}{2}\\
&\leqslant\frac{2+6L^2\gamma^2}{9}+\frac{3n\gamma^2L^2}{2}-\frac{1}{2}\\
&=\gamma^2\frac{(9n+4)L^2}{6}-\frac{5}{18}.
\end{aligned}
$$

Using Equation (46) on (45) yields

$$\mathbb{E}\left[\sup_{u\in Z}\sum_{\tau=1}^{t}\langle\gamma_\tau F(\theta_{\tau+1/2}),\theta_{\tau+1/2}-u\rangle\right]\leqslant 2\Omega+104n\sigma^2\sum_{\tau=0}^{t}\gamma_\tau^2.$$

We conclude by Lemma 4. □

**Corollary 6.** *Set $\gamma$ as in Equation (44). For all $\tau$ between 1 and $t$, let*

$$\gamma_\tau=\min\left\{\gamma,\frac{1}{2}\sqrt{\frac{\Omega}{13n\sigma^2t}}\right\},$$

*Then,*

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_t)\right] \leqslant \max\left\{\frac{4\Omega}{\gamma t}, 8\sqrt{\frac{13\Omega n\sigma^2}{t}}\right\}. \tag{47}$$

*Proof.* Same proof as Corollary 4. □

**Corollary 7.** *Setting $\gamma_\tau$ as in Corollary 6, the expected convergence rate in terms of the number of gradient computations $k$ is*

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_{t(k)})\right] \leqslant \max\left\{\frac{96\sqrt{2}\Omega L n^2}{\sqrt{b}k}, 8\Omega bL\sqrt{\frac{27n+12}{5}}\frac{1}{k}, 8\sqrt{\frac{26\Omega nb\sigma^2}{k}}\right\}.$$

*Proof.* Plugging $t(k) = k/2b$ into (47), we get

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_{t(k)})\right] \leqslant \max\left\{\frac{4\Omega}{\frac{\left(\frac{b}{n}\right)^{3/2}}{\sqrt{(1-\frac{b}{n})(2-\frac{b}{n})}}\frac{1}{12L\sqrt{n}}\frac{k}{2b}}, \frac{4\Omega}{\frac{1}{L}\sqrt{\frac{5}{27n+12}}\frac{k}{2b}}, 8\sqrt{\frac{26\Omega nb\sigma^2}{k}}\right\}.$$

If we bound $1 - b/n < 1$ and $2 - b/n < 2$ (for simplicity) we get the desired result. □

# 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 $\ell_i$ 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 transforms $\theta_{t+1} = \mathcal{A}(\theta_t)$ for the different sampling schemes.

**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

$$\begin{cases} \theta^{\text{full}}_{\tau+1/2} = \theta^{\text{full}}_\tau - \gamma A \theta^{\text{full}}_\tau, \\ \theta^{\text{full}}_{\tau+1} = \theta^{\text{full}}_\tau - \gamma A \theta^{\text{full}}_{\tau+1/2}. \end{cases} \tag{48}$$

By introducing $\mathcal{A}^{(\gamma)}_{\text{full}} := I - \gamma A + \gamma^2 A^2$, (48) is simply $\theta^{\text{full}}_{\tau+1} = \mathcal{A}^{(\gamma)}_{\text{full}} \theta^{\text{full}}_\tau$.

**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

$$\begin{cases} \theta^{\text{cyc}}_{\tau+1/4} = \theta^{\text{cyc}}_\tau - \gamma M_1 A \theta^{\text{cyc}}_\tau, \\ \theta^{\text{cyc}}_{\tau+1/2} = \theta^{\text{cyc}}_\tau - \gamma M_2 A \theta^{\text{cyc}}_{\tau+1/4}, \\ \theta^{\text{cyc}}_{\tau+3/4} = \theta^{\text{cyc}}_{\tau+1/2} - \gamma M_2 A \theta^{\text{cyc}}_{\tau+1/2}, \\ \theta^{\text{cyc}}_{\tau+1} = \theta^{\text{cyc}}_{\tau+1/2} - \gamma M_1 A \theta^{\text{cyc}}_{\tau+3/4}. \end{cases} . \tag{49}$$

Remark that (49) contains two iterations of Alg. 1; $\theta_{\tau+1/4}$ and $\theta_{\tau+3/4}$ are extrapolations and $\theta_{\tau+1/2}$ and $\theta_{\tau+1}$ are updates. The reason behind this change is that computing $\theta_{\tau+1}$ from $\theta_\tau$ now requires the same amount of gradient computations in the full and alternated extra-gradient. If we define $\mathcal{A}^{(\gamma)}_{ij} := I - \gamma M_i A + \gamma^2 M_i A M_j A$ and $\mathcal{A}^{(\gamma)}_{\text{cyc}} := \mathcal{A}^{\gamma}_{12} \mathcal{A}^{(\gamma)}_{21}$, we can write $\theta^{\text{cyc}}_{\tau+2} = \mathcal{A}^{(\gamma)}_{\text{cyc}} \theta^{\text{cyc}}_\tau$.

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

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

where $S_{\tau+1/4}, S_{\tau+1/2}, S_{\tau+3/4}, S_{\tau+1}$ take values 1 and 2 with equal probability and pairwise are independent. Note that we also enroll two iterations of sampled extra-gradient, for consistency with cyclic sampling. Let $\mathcal{F}_\tau = \sigma(S_{\tau'} : \tau' \leqslant \tau)$. For extra-gradient with random subsampling, we can write

$$\begin{aligned}
\mathbb{E}\left[\theta_{\tau+1}^{\text{rand}}\right] &= \mathbb{E}\left[\mathcal{A}^{(\gamma)}_{S_{\tau+1}S_{\tau+3/4}}\mathcal{A}^{(\gamma)}_{S_{\tau+1/2}S_{\tau+1/4}}\theta_\tau^{\text{rand}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathcal{A}^{(\gamma)}_{S_{\tau+1}S_{\tau+3/4}}\mathcal{A}^{(\gamma)}_{S_{\tau+1/2}S_{\tau+1/4}}\theta_\tau^{\text{rand}}\middle|\mathcal{F}_\tau\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathcal{A}^{(\gamma)}_{S_{\tau+1}S_{\tau+3/4}}\mathcal{A}^{(\gamma)}_{S_{\tau+1/2}S_{\tau+1/4}}\middle|\mathcal{F}_\tau\right]\theta_\tau^{\text{rand}}\right] \\
&= \mathbb{E}\left[\mathcal{A}^{(\gamma)}_{S_{\tau+1}S_{\tau+3/4}}\mathcal{A}^{(\gamma)}_{S_{\tau+1/2}S_{\tau+1/4}}\right]\mathbb{E}\left[\theta_\tau^{\text{rand}}\right].
\end{aligned}$$

Let us define $\mathcal{A}^{(\gamma)}_{\text{rand}} := \mathbb{E}\left[\mathcal{A}^{(\gamma)}_{S_{\tau+1}S_{\tau+3/4}}\mathcal{A}^{(\gamma)}_{S_{\tau+1/2}S_{\tau+1/4}}\right] = \frac{1}{16}\sum_{j_1,j_2,j_3,j_4\in\{1,2\}}\mathcal{A}^{(\gamma)}_{j_1j_2}\mathcal{A}^{(\gamma)}_{j_3j_4}$. It is easy to see that $\mathcal{A}^{(\gamma)}_{\text{rand}} = \frac{1}{16}\left(4I - 2\gamma A + \gamma^2A^2\right)^2$. We then have $\mathbb{E}\left[\theta^{\text{rand}}_{\tau+2}\right] = \mathcal{A}^{(\gamma)}_{\text{rand}}\mathbb{E}\left[\theta_\tau^{\text{rand}}\right]$.

## C.2 Convergence behavior through spectral analysis

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

**Theorem 8** (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 8.** *The behavior of $\theta_t^{\text{full}}$, $\theta_t^{\text{cyc}}$ and $\theta_t^{\text{rand}}$ 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. □

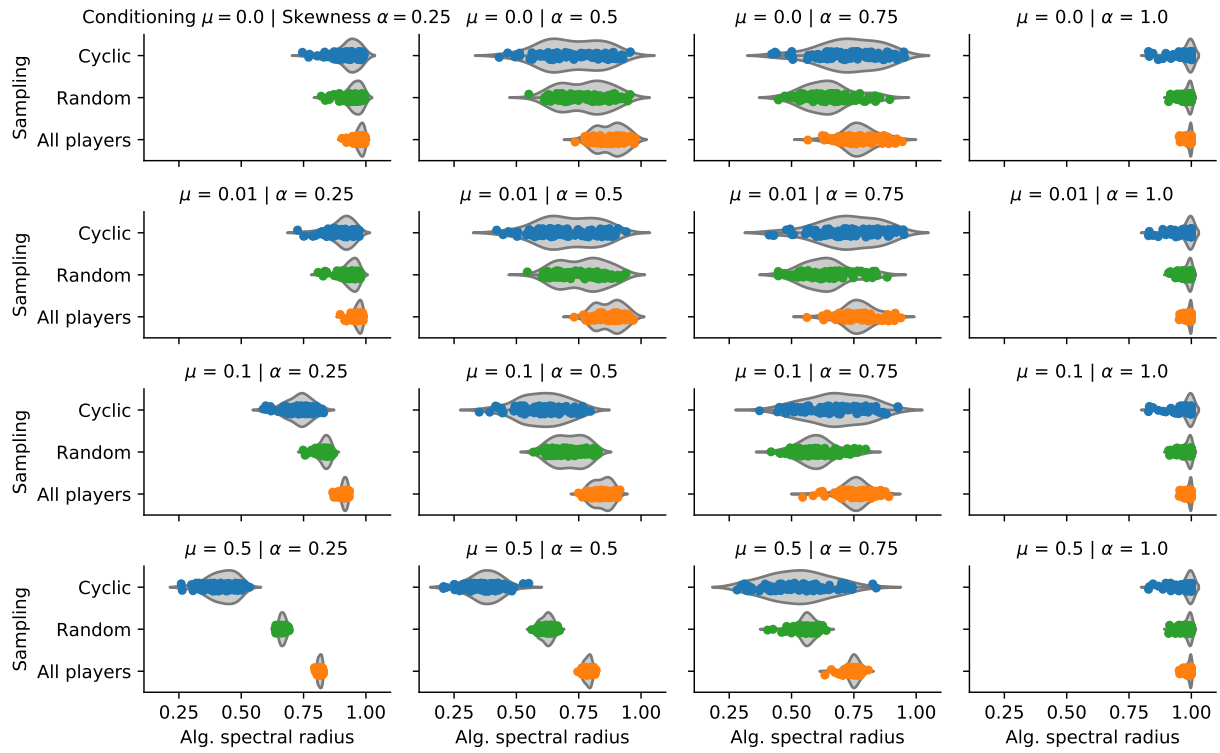



Figure 4: Spectral radii distribution of the algorithmic operator associated to doubly-stochastic and full extra-gradient, in the non-constrained bi-linear two-player game setting, for various conditioning and skewness. Random and cyclic sampling yields lower radius (hence faster rates) for most problem geometry. Cyclic sampling outperforms random sampling in most settings, especially for better conditioned problems.

## C.3 Empirical distributions of the spectral radii

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

$$\mathcal{A}^\star_{\text{full}} \triangleq \min_{\gamma \in \mathbb{R}^+} \rho(\mathcal{A}^{(\gamma)}_{\text{full}}), \quad \mathcal{A}^\star_{\text{cyc}} \triangleq \min_{\gamma \in \mathbb{R}^+} \rho(\mathcal{A}^{(\gamma)}_{\text{cyc}}), \quad \mathcal{A}^\star_{\text{rand}} \triangleq \min_{\gamma \in \mathbb{R}^+} \rho(\mathcal{A}^{(\gamma)}_{\text{rand}}), \tag{50}$$

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, in the setting of §5.1, detailed in App. D.

**Experiment.** Matrices $A$ in $\mathbb{R}^{2d \times 2d}$, with $d = 3$, are sampled as the weighted sum of a random positive definite matrix $A_{\text{sym}}$ and of a random skew matrix $A_{\text{skew}}$. We vary the weight $\alpha \in [0, 1]$ of the skew matrix and the lowest eigenvalue $\mu$ of the matrix $A_{\text{sym}}$. We sample 300 different games and compute $\mathcal{A}^{(\eta)}$ on a grid of step sizes $\eta$, for the three different methods. We thus estimate the best algorithmic spectral radii defined in (50).

**Results and interpretation.** The distributions of algorithm spectral radii are presented in Fig. 4. 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 $\mu$), 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 the GAN game is skew (almost zero-sum notwithstanding the discriminator penalty).

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

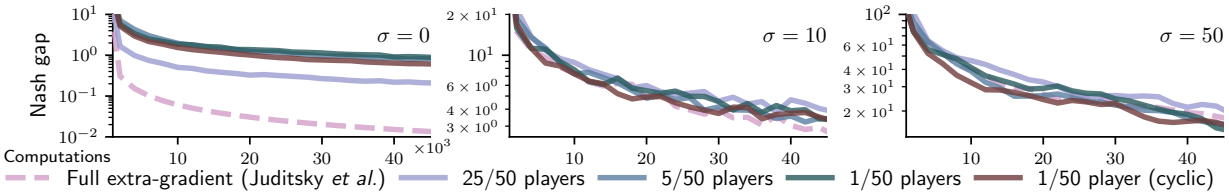



Figure 5: 50-player completely skew smooth game with increasing noise (sampling with variance reduction). In the non-noisy setting, player sampling reduces convergence speed. On the other hand, it provides a speed-up in the high noise regime.

# 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 $A_{\text{sym}}$ positive definite, we compute its lowest eigenvalue $\mu_0$, and update $A_{\text{sym}} \leftarrow A_{\text{sym}} + (\mu - \mu_0)I_{nd\times nd}$, where $\mu$ 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 $\alpha$ 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 3. 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 $\alpha$. We performed a grid-search on learning rates, setting $\eta \in \{10^{-5}, \cdots, 1\}$, with 32 logarithmically-spaced values, making sure

Table 3: Parameters used in Fig. 2 for increasing problem complexity.

| Figure | Players # | Exp. | Skewness $\alpha$ | Noise $\sigma$ | Reg. $\lambda$ |
|---|---|---|---|---|---|
| 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. 5 | 50 | Smooth, skew, no-noise | 1 | 0 | 0. |
| | | | 1 | 10 | 0. |
| | | Smooth, skew, highest-noise | 1 | 50 | 0 |

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. 5, showing how the performance of player sampling improves with noise. We emphasize that the non-noisy setting is not 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 and Ba, 2015), and its straight-forward extension to extrapolation methods 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.