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 modeling increasingly requires to find a Nash equilibrium in multi-player games, e.g. when training GANs. In this paper, we analyse a new extra-gradient method for Nash equilibrium finding, that performs gradient extrapolations and updates on a random subset of players at each iteration. This approach provably exhibits a better rate of convergence than full extra-gradient for non-smooth convex games with noisy gradient oracle. We propose an additional variance reduction mechanism to obtain speed-ups in smooth convex games. Our approach makes extrapolation amenable to massive multiplayer settings, and brings empirical speed-ups, in particular when using a heuristic cyclic sampling scheme. Most importantly, it allows to train faster and better GANs and mixtures of GANs.

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 (Wayne and Abbott, 2014) and 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. These examples can be cast as games where players are parametrized modules that compete or cooperate to minimize their own objective functions.

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. The theoretical section of this paper considers the class of *convex n-player games*, for which Nash equilibria exist (Rosen, 1965). Finding a Nash equilibrium in this setting is equivalent to solving a variational inequality problem (VI) with a monotone operator (Harker and Pang, 1990; Rosen, 1965). 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). 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 to solve the VI (hence Nash) problem: *averaging* (Magnanti and Perakis, 1997; Nedić and Ozdaglar, 2009) or *extrapolation* with averaging. The later was introduced as the *extra-gradient* (EG) method by Korpelevich (1976)); it is faster (Nemirovski, 2004) and can handle noisy gradients (Juditsky et al., 2011). Extrapolation corresponds to an *opponent shaping* step: each player anticipates its opponents' next moves to update its strategy.

[*]Equal contribution
[†]Joint senior author

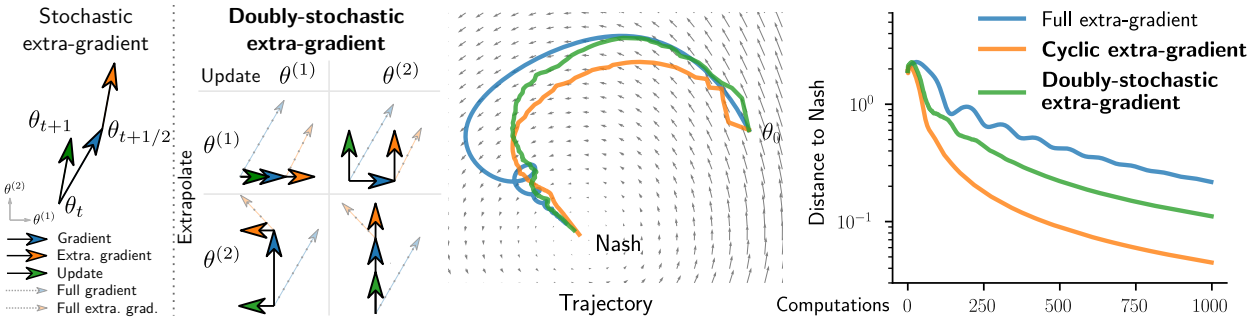


Figure 1: *Left:* We compute masked gradient during the extrapolation and update steps of the extra-gradient algorithm, to perform faster updates. *Right:* Optimization trajectories for doubly stochastic extra-gradient and full-update extra-gradient, on a convex single-parameter two-player convex game.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

Table 1: New and existing (Juditsky et al., 2011) convergence rates for convex games, w.r.t. the number of gradient computations k. Doubly-stochastic extra-gradient (DSEG) *multiplies the noise contribution by a factor* $\alpha \triangleq \sqrt{b/n}$, where b is the number of sampled players among n. G bounds the gradient norm. L: Lip. constant of losses' gradient. σ^2 bounds the gradient estimation noise. Ω: diameter of the param. space.

$\alpha \triangleq \sqrt{b/n}$	Non-smooth	Smooth
DSEG	$\mathcal{O}\left(n\sqrt{\frac{\Omega}{k}(G^2 + \alpha^2\sigma^2)}\right)$	$\mathcal{O}\left(\frac{\Omega L n^{3/2}}{\alpha k} + \alpha n\sigma\sqrt{\frac{\Omega}{k}}\right)$
Full EG	$\mathcal{O}\left(n\sqrt{\frac{\Omega}{k}(G^2 + \sigma^2)}\right)$	$\mathcal{O}\left(\frac{\Omega L n^{3/2}}{k} + n\sigma\sqrt{\frac{\Omega}{k}}\right)$

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 update of each player would behave. Therefore, we introduce and analyse new extra-gradient algorithms that extrapolate and update random or carefully selected subsets of players at each iteration (Figure 1).

- We review the extra-gradient algorithm for differentiable games and outline its shortcomings (§3.1). We propose a doubly-stochastic extra-gradient (DSEG) algorithm (§3.2) that updates the strategies of a subset of players, thus performing *player sampling*. DSEG performs faster but noisier updates than the original full extra-gradient method (full EG, Juditsky et al., 2011), that uses a (once) stochastic gradient oracle. We introduce a variance reduction method to attenuate the noise added by player sampling in smooth games.
- We derive convergence rates for DSEG in the convex setting (§4), as summarized in Table 1. Proofs strongly relies on the specific structure of the noise introduced by player sampling. Our rates exhibit a better dependency on gradient noise compared to stochastic extra-gradient, and are thus interesting in the high-noise regime common in machine learning.
- Empirically, we first validate that DSEG is faster in massive differentiable convex games with noisy gradient oracles. We further show that non-random player selection improves convergence speed, and provide explanations for this phenomenon. In practical non-convex settings, we find that cyclic player sampling improves the speed and performance of GAN training (CIFAR10, ResNet architecture). The positive effects of extrapolation and alternation combine: DSEG should be used to train GANs, and even more to train *mixtures* of GANs.

2 Related work

Extra-gradient method. In this paper, we focus on finding the Nash equilibrium in convex n-player games, or equivalently the Variational Inequality problem (Harker and Pang, 1990; Nemirovski et al., 2010). This can be done using extrapolated gradient (Korpelevich, 1976), a "cautious" gradient descent approach 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. In the convex setting, their results guarantees the convergence of the algorithm we propose, albeit with very slack rates. Our theoretical analysis refines these rates to show the usefulness of player sampling. 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 discussion in §4), an original endeavor.

Extra-gradient in non-convex settings. Extra-gradient has been applied in non-convex settings. Mertikopoulos et al. (2019) proves asymptotic convergence results for extra-gradient without averaging in a slightly non-convex case. Gidel et al. (2019) demonstrate the effectiveness of extra-gradient for GANs. They argue that it allows to escape the potentially chaotic behavior of simultaneous gradient updates (examplified by e.g. Cheung and Piliouras (2019)). Earlier work on GANs propose to replace simultaneous updates with alternated updates, with a comparable improvement (Gulrajani et al., 2017). In §5, we show that alternating player updates while performing opponent extrapolation improves the training speed and quality of GANs.

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 games with gradients

In a game, each player observes a loss that depends on the independent parameters of all other players.

Definition 1. *A standard 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 Θ decomposes into a Cartesian product $\prod_{i=1}^n \Theta^i$. Each player's parameter θ^i lives in $\Theta^i \subset \mathbb{R}^{d_i}$. Each player is given a loss function $\ell_i \colon \Theta \to \mathbb{R}$.*

For example, generative adversarial network (GAN) training is a standard game between a generator and discriminator that do not share parameters. We make the following assumption over the geometry of losses and constraints, that is the counterpart of the convexity assumption in single-objective optimization.

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

Assumption 1 implies that Θ 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 Assumption 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 θ corresponding to n mixed strategies θ_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, we look for equilibria $\theta_\star \in \Theta$ such that

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

A Nash equilibrium is a point where no player can benefit by changing his strategy while the other players keep theirs unchanged. Assumption 1 implies the existence of a Nash equilibrium (Rosen, 1965). We quantify the inaccuracy of a solution θ by the *functional Nash error*, also known as the Nikaidô and Isoda (1955) function:

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

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 θ is a Nash equilibrium; thus $\mathrm{Err}_N(\theta)$ constitutes a propose indication of convergence for sequence of iterates seeking a Nash equilibrium. We bound this value in our convergence analysis (see §4).

First-order methods and extrapolation. In convex games, as the losses ℓ_i are (sub)differentiable, we may solve (1) using first-order methods. We assume access to the *simultaneous gradient* of the game

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

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, and may be noisy. The losses ℓ_i may be non-smooth, in which case the gradients $\nabla_i \ell_i$ can be replaced by any subgradients. Simultaneous gradient descent, that explicitly discretizes the flow of the simultaneous gradient may converge slowly—e.g., in matrix games with skew-symmetric payoff and noiseless gradient oracle, convergence of the average iterate demands *decreasing* step-sizes. The extra-gradient method (Korpelevich, 1976) provides better guarantees (Juditsky et al., 2011; Nemirovski, 2004)—e.g., in the previous example, the step-size can remain constant. We build upon this method.

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

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

where $p_\Theta[\cdot]$ is the Euclidean projection onto the constraint set Θ, 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 Figure 1). The generalization of extra-gradient to general Banach spaces equipped by a Bregman divergence was introduced as the *mirror-prox* algorithm (Nemirovski, 2004). The new convergence results of §4 extend to the *mirror* setting (see §A.1). As recalled in Table 1,

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 $b \in [n]$.
With variance reduction (VR), $R \leftarrow \tilde{F}(\theta_0, [1, n])$ as in (5), i.e. the 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 VR (Algorithm 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 VR
    Gradient step: $\theta_{\tau+1} \leftarrow p_\Theta[\theta_\tau - \gamma_\tau \tilde{F}_{\tau+1}]$.
Return $\hat{\theta}_t = [\sum_{\tau=0}^t \gamma_\tau]^{-1} \sum_{\tau=0}^t \gamma_\tau \theta_\tau$.

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, finding

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

where Assumption 1 ensures that the simultaneous gradient F is a monotone operator (see §A.2 for a review).

3.2 DSEG: Partial extrapolation and update for extra-gradient

The proposed algorithms are theoretically analyzed in the convex setting §4, and empirically validated in convex and non-convex setting in §5.

Caveats of extra-gradient. 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, in our setting, partial gradient observation should be sufficient to perform extrapolation and updates toward the Nash equilibrium

Player sampling. While standard extra-gradient performs at each iteration two passes of player's gradient computation, we therefore compute *doubly-stochastic simultaneous gradient estimates*, where only the gradients of a random subset of players are evaluated. 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 $\tilde{F} \triangleq (\tilde{F}^{(1)}, \dots, \tilde{F}^{(n)})^\top \in \mathbb{R}^d$ where

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

and $g_i(\theta)$ 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 (DSEG) replaces the full gradients in the update (3) by the oracle (5), as detailed in Algorithm 1.

Variance reduction for player noise. To obtain faster rates in convex games with smooth losses, we propose to compute a variance-reduced estimate of the gradient oracle (5). This mitigates the noise due

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 $\bar{F}^{(i)} \leftarrow \tilde{F}^{(i)}(\theta) - (1 - \frac{b}{n})R^{(i)}$
    Update $R^{(i)} \leftarrow \tilde{F}^{(i)}(\theta)$
For $i \notin \mathcal{P}$, set $\bar{F}^{(i)} \leftarrow R^{(i)}$.
Return estimate $\bar{F} = (\bar{F}^{(1)}, ..., \bar{F}^{(n)})$, table R.

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 Algorithm 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, akin to the approach of SAGA (Defazio et al., 2014) to reduce the variance of data noise.

Player sampling strategies. For convergence guarantees to hold, each player must have an equal probability of being sampled (*equiprobable player sampling condition*). 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.

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. This scheme bridges extrapolation and alternated gradient descent: for GANs, it corresponds to extrapolate the generator before updating the discriminator, and vice-versa, cyclically. Although its convergence is not guaranteed, cyclic sampling over players is powerful for convex quadratic games (§5.1) and GANs (§5.2).

4 Convergence for convex games

We derive new rates for DSEG with random player sampling, improving the analysis of Juditsky et al. (2011). Player sampling can be seen as an extra source of noise in the gradient oracle. Hence the results of Juditsky et al. on stochastic extra-gradient guarantees the convergence of DSEG, as we detail in Corollary 1. Unfortunately, the convergence rates in this corollary do not predict any improvement of DSEG over full extra-gradient. Our main theoretical contribution is therefore a refinement of these rates for player-sampling noise. Improvements are obtained both for non-smooth and smooth losses, the latter using the proposed variance reduction approach. Our results predict better performance for DSEG in the high-noise regime. Results are stated here in Euclidean spaces for simplicity; they are proven in the more general mirror setting in Appendix B. In the analysis, we separately consider the two following assumptions on the losses.

Assumption 2a (Non-smoothness)**.** *For each $i \in [n]$, the loss ℓ_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 ℓ_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 $\theta, \theta' \in \Theta$.*

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

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

$$\begin{aligned} \forall\, \theta \in \Theta, \quad & \mathbb{E}[g_i(\theta)] = \nabla_i \ell_i(\theta), \\ & \mathbb{E}[\|g_i(\theta) - \nabla_i \ell_i(\theta)\|_2^2] \leqslant \sigma^2. \end{aligned} \tag{6}$$

To compare DSEG to simple stochastic EG, we must take into account the cost of a single iteration, that we assume proportional to the number b of gradients to estimate at each step. We therefore set $k \triangleq 2\,b\,t$ to be the number of gradients estimates computed up to iteration t, and re-index the sequence of iterate $(\hat{\theta}_t)_{t\in\mathbb{N}}$ as $(\hat{\theta}_k))_{k\in 2b\mathbb{N}}$. We give rates with respect to k in the following propositions.

4.1 Slack rates derived from Juditsky et al.

Let us first recall the rates obtained by Juditsky et al. (2011) with noisy gradients but no player sampling.

Theorem 1 (Adapted from Juditsky et al. (2011))**.** *We consider a convex n-player game where 2a and Assumption 3 hold. We run Algorithm 1 for t iterations without player sampling, thus performing $k = 2\,n\,t$ gradient evaluations. With optimal constant stepsize, the expected Nash error verifies*

$$\mathbb{E}\left[Err_N(\hat{\theta}_k)\right] \leqslant 14n\sqrt{\frac{\Omega}{3k}\,(G^2 + 2\sigma^2)}. \tag{7}$$

Assuming smoothness (2b) and optimal stepsize,

$$\mathbb{E}\left[Err_N(\hat{\theta}_k)\right] \leqslant \max\left\{\frac{7\Omega L n^{3/2}}{k}, 14n\sqrt{\frac{2\Omega\sigma^2}{3k}}\right\}. \tag{8}$$

Player sampling fits within the framework of noisy gradient oracle (6), replacing the gradient estimates $(g_i)_{i\in[n]}$ with the estimates $(\tilde{F}^{(i)})_{i\in[n]}$ from (5), and updating the variance σ^2 accordingly. We thus derive the following corollary.

Corollary 1. *We consider a convex n-player game where 2a and Assumption 3 hold. We run Algorithm 1 for t iterations with equiprobable player sampling, thus performing $k = 2\,b\,t$ gradient evaluations. With optimal constant stepsize, the expected Nash error verifies*

$$\mathbb{E}\left[\mathrm{Err}_N(\hat{\theta}_k)\right] \leqslant \mathcal{O}\left(n\sqrt{\frac{\Omega}{k}\left(\frac{n}{b}G^2 + \sigma^2\right)}\right).$$

Assuming smoothness (2b) and optimal stepsize,

$$\mathbb{E}\left[Err_N(\hat{\theta}_k)\right] \leqslant \mathcal{O}\left(\frac{\Omega L n^{3/2}}{k} + n\sqrt{\frac{\Omega}{k}(\frac{n}{b}L^2\Omega^2 + \sigma^2)}\right).$$

The proof is in §B.1. The notation $\mathcal{O}(\cdot)$ hides numerical constants. Whether in the smooth or non-smooth case, the upper-bounds from Corollary 1 does not predict any improvement due to player sampling, as the factor before the gradient size G or $L\Omega$ is increased, and the factor before the noise variance σ remains constant.

4.2 Tighter rates using noise structure

Fortunately, a more cautious analysis allows to improve these bounds, by taking into account the noise structure induced by sampling in (5). We provide a new result in the non-smooth case, proven in §B.3.

Theorem 2. *We consider a convex n-player game where 2a and Assumption 3 hold. We run Algorithm 1 for t iterations with equiprobable player sampling, thus performing $k = 2\,b\,t$ gradient evaluations. With optimal constant stepsize, the expected Nash error verifies*

$$\mathbb{E}\left[Err_N(\hat{\theta}_k)\right] \leqslant \mathcal{O}\left(n\sqrt{\frac{\Omega}{k}\left(G^2 + \frac{b}{n}\sigma^2\right)}\right). \tag{9}$$

Compared to Corollary 1, we obtain a factor $\sqrt{\frac{b}{n}}$ in front of the noise term $\frac{\sigma}{\sqrt{k}}$, without changing the constant before the gradient size G. We can thus expect faster convergence with noisy gradients. (9) is tightest when sampling a single player, i.e. when $b = 1$.

A similar improvement can be obtained with smooth losses thanks to the variance reduction technique proposed in Algorithm 2. This is made clear in the following result, proven in §B.4.

Theorem 3. *We consider a convex n-player game where 2a and Assumption 3 hold. We run Algorithm 1 for t iterations with equiprobable player sampling, thus performing $k = 2\,b\,t$ gradient evaluations. Algorithm 2 yields gradient estimates. With optimal constant stepsize, the expected Nash error verifies*

$$\mathbb{E}\left[Err_N(\hat{\theta}_k)\right] \leqslant \mathcal{O}\Big(\sqrt{\frac{n}{b}}\frac{\Omega L n^{3/2}}{k} + \sqrt{\frac{b}{n}}n\sqrt{\frac{\Omega\sigma^2}{k}}\Big). \tag{10}$$

The upper-bound (10) should be compared with the bound of full extra-gradient (8)—that it recovers for $b = n$. With player sampling, the constant before the gradient size $L\Omega$ is bigger of a factor $\sqrt{\frac{n}{b}}$. On the other hand, the constant before the noise term σ is smaller of a factor $\sqrt{\frac{n}{b}}$. Player sampling is therefore beneficial when the noise term dominates, which is the case whenever the number of iterations is such that $k \geqslant \frac{\Omega L^2 n}{\sigma^2}\left(\frac{n}{b}\right)^2$. For $k \to \infty$, the bound (10) is once again tightest by sampling a random single player.

To sum up, doubly-stochastic extra-gradient convergence is controlled with a better rate than stochastic extra-gradient (EG) with non-smooth losses; with smooth losses, DSEG exhibits the same rate structure in $\frac{1}{k} + \frac{1}{\sqrt{k}}$ as stochastic EG, with a better dependency on the noise but worse dependency on the gradient smoothness. In the high noise regime, or equivalently when demanding high precision results, DSEG brings the same improvement of a factor $\sqrt{\frac{b}{n}}$ before the constant $\frac{\sigma}{\sqrt{k}}$, for both smooth and non-smooth problems.

Step-sizes. The stepsizes of the previous propositions are assumed to be constant and are optimized knowing the geometry of the problem. They are explicit in Appendix B. As in full extra-gradient, convergence can be guaranteed without such knowledge using decreasing step-sizes. In experiments, we perform a grid-search over stepsizes to obtain the best results given a computational budget k.

5 Convex and non-convex applications

We show the performance of doubly-stochastic extra-gradient in the setting of quadratic games, comparing different sampling schemes. We assess the speed and final performance of DSEG in the practical context of GAN training. A *PyTorch*/*Numpy* package is attached.

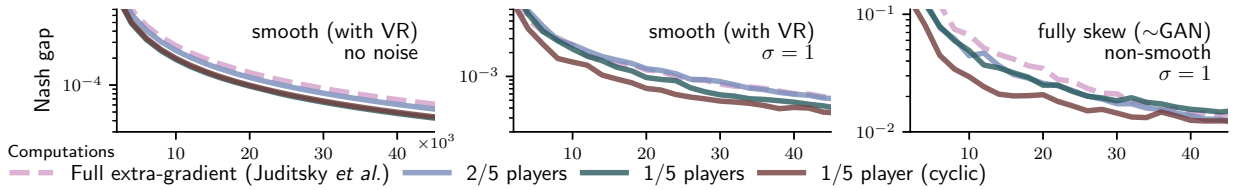
5.1 Random convex quadratic games

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 ℓ_i of each player is defined as its expected payoff given the n mixed strategies $(\theta^1, \dots, \theta^n)$, i.e. $\forall\, i \in [n]$, $\forall\, \theta \in \Theta = \triangle^{d_1} \times \cdots \times \triangle^{d_n}$,

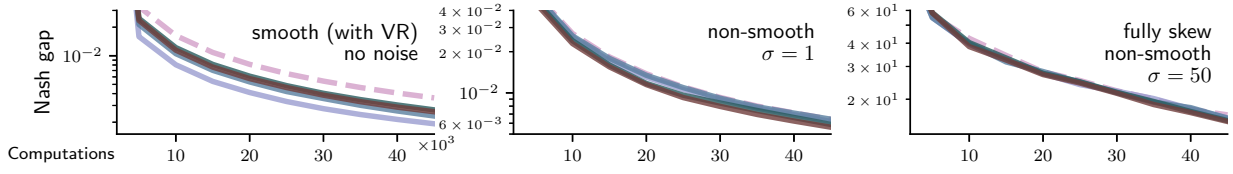
$$\ell_i(\theta^i, \theta_{-i}) = {\theta^i}^\top A_i \theta + \lambda \|\theta^i - \frac{1}{d}\|_1,$$

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

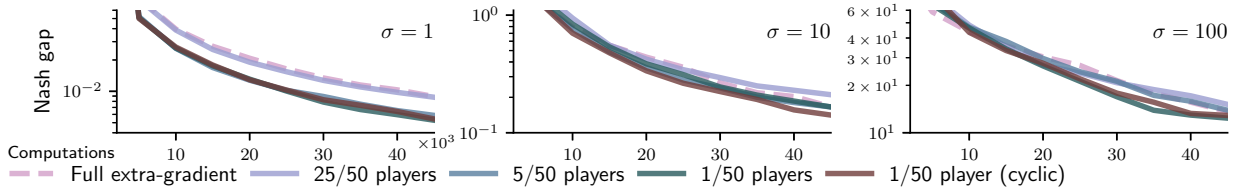
Experiments. We sample A as the weighted sum of a random symmetric positive definite matrix and a skew matrix. We compare the convergence speeds of extra-gradient algorithms, with or without player sampling. We vary three parameters: the variance σ of the noise in the gradient oracle (we add a Gaussian noise on each gradient coordinate), the non-smoothness λ of the loss, and the skewness of the matrix. We consider



(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). Higher sampling ratio is beneficial in high noise regime (c), Curves averaged over 5 games and 5 runs.

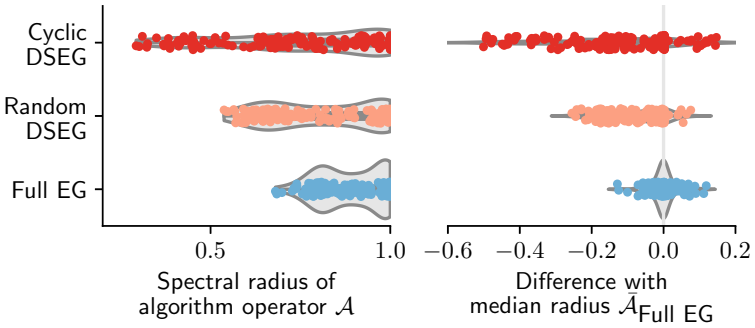


Figure 3: *Left:* Spectral radii of operators for random 2-player matrix games. *Right:* each radius is compared to the median radius obtained for full extra-gradient, within each category of skewness and conditioning of random payoff matrices. Cyclic sampling lowers spectral radii and improve convergence rates.

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 Appendix D). We use variance reduction when $\lambda = 0$ (smooth case). We also consider cyclic sampling in our benchmarks, as described in §3.2.

Results. Figure 2 compares the convergence speed of player-sampled extra-gradient for the various settings and sampling schemes. As predicted by Theorem 2 and 3, the regime of convergence in $1/\sqrt{k}$ in the presence of noise is unchanged with player sampling. DSEG always brings a benefit in the convergence constants (Figure 2a-b), in particular for smooth noisy problems (Figure 2a center, Figure 2b left). Most interestingly, cyclic player selection improves upon random sampling for small number of players (Figure 2a).

Figure 2c highlights the trade-offs in Theorem 3: as the noise increase, the size of player batches should

Figure 4: Training curves and samples using doubly-stochastic extragradient on CIFAR10 with WGAN-GP losses, for the 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.

be reduced. Not that for skew-games with many players (Figure 2b col. 3), our approach only becomes beneficial in the high-noise regime. As predicted in §4, full EG should be favored with noiseless oracles (see Appendix D).

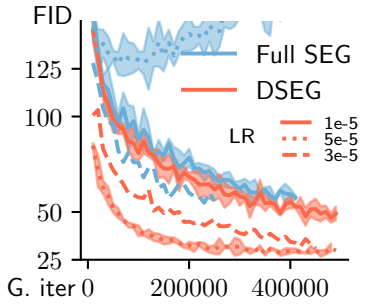
Spectral study of sampling schemes. The benefit of cyclic sampling can be explained for simple quadratic games. We consider a two-player quadratic game where $\ell_i(\theta) = {\theta^i}^\top A\theta$ for $i = 1, 2$, $\theta = (\theta^1, \theta^2)$ is an unconstrained vector of $\mathbb{R}^{2\times d}$, and gradients are noiseless. In this setting, full EG and DSEG expected iterates follows a linear recursion $\mathbb{E}[\theta_{k+4}] = \mathcal{A}(\mathbb{E}[\theta_k])$, where k is the number of gradient evaluation and $\mathcal{A}$ is a linear "algorithm operator", computable in closed form. A lower spectral radius for $\mathcal{A}$ yields a better convergence rate for $(\mathbb{E}[\theta_k])_k$, in light of Gelfand (1941) formula—we compare spectral radii across methods.

We sample random payoff matrices A of varying skewness and condition number, and compare the spectral radius $\mathcal{A}$ associated to full EG, and DSEG with cyclic and random player selection. As summarized in Figure 3, player sampling reduces the spectral radius of $\mathcal{A}$ on average; most interestingly, the reduction is more important using cyclic sampling. Spectral radii are not always in the same order across methods, hinting that sampling can be harmful in the worst cases. Yet cyclic sampling will perform best on average in this (simple) setting. We report details and further figures in Appendix C.

5.2 Generative adversarial networks (GANs)

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. Our theoretical analysis indeed shows a $1/\sqrt{2}$ speed-up for noisy monotonous 2-player games—the following suggests that speed-up also arises in a non-convex setting. We compare the full stochastic extra-gradient (SEG) 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 Appendix D). We use the Adam (Kingma and Ba, 2015) refinement of extra-gradient (Gidel et al., 2019) for both the baseline and proposed methods. The notion of functional Nash error does not exist in the non-convex setting. We estimate the convergence speed toward an equilibrium by measuring a quality criterion for the generator. We therefore evaluate the Inception Score (Salimans et al., 2016) and Fréchet Inception Distance (FID, Heusel et al. (2017) along training, and report their final values.

Results. We report training curves versus wall-clock time in Figure 4. Cyclic sampling allows faster and better training, especially with respect to FID, which is more correlated to human appreciation (Heusel et al., 2017). Figure 5 (right) compares our result to full extra-gradient with uniform averaging. It shows substantial improvements in FID, with results less sensitive to randomness. SEG itself slightly outperforms optimistic mirror descent (Gidel et al., 2019; Mertikopoulos et al., 2019).



Method	FID (50k)
SEG	19.69 ± 1.53
DSEG	$\mathbf{17.10 \pm 1.07}$
	IS
SEG	$8.26 \pm .16$
DSEG	$\mathbf{8.38 \pm .06}$

Figure 5: *Left*: Player sampling allows faster training of mixtures of GANs. *Right*: Player sampling trains better ResNet WGAN-GP. FID and IS computed on 50k samples, averaged over 5 runs.

Interpretation. Without extrapolation, alternated training is known to perform better than simultaneous updates in WGAN-GP (Gulrajani et al., 2017). Full extrapolation has been shown to perform similarly to alternated updates (Gidel et al., 2019). Our approach combine extrapolation with an alternated schedule. It thus performs better than extrapolating with simultaneous updates. It remains true across every learning rate we tested. Echoing our findings of §5.1, deterministic sampling is crucial for performance, as random player selection performs poorly (score 6.2 IS).

5.3 Mixtures of GANs

Finally, we consider a simple multi-player GAN setting, akin to Ghosh et al. (2018), where n different generators $(g_{\theta_i})_i$ seeks to fool m different discriminators $(f_{\varphi_j})_j$. We minimize $\sum_j \ell(g_{\theta_i}, f_{\varphi_j})$ for all i, and maximize $\sum_i \ell(g_{\theta_i}, f_{\varphi_j})$ for all j. Fake data is then sampled from mixture $\sum_{i=1}^n \delta_{i=J} g_{\theta_i}(\varepsilon)$, where J is sampled uniformly in $[n]$ and $\varepsilon \sim \mathcal{N}(0, I)$. We compare two methods: (i) SEG extrapolates and updates all $(g_{\theta_i})_i$, $(f_{\varphi_j})_j$ at the same time; (ii) DSEG extrapolates and updates successive pairs $(g_{\theta_j}, f_{\varphi_j})$ alternating the 4-step updates from §5.2.

Results. We compare the training curves of both SEG and DSEG in Figure 5, for a range of learning rates. DSEG outperform SEG for all learning rates; more importantly, higher learning rates can be used for DSEG, allowing for faster training. DSEG is thus appealing for mixtures of GANs, that are useful to mitigate mode collapse in generative modeling. We report generated images in Appendix D.

6 Conclusion

We propose and analyse a doubly-stochastic extra-gradient approach for finding Nash equilibria. According to our convergence results, updating and extrapolating random sets of players in extra-gradient brings speed-up in noisy and non-smooth convex problems. Numerically, doubly-stochastic extra-gradient indeed brings speed-ups in convex settings, especially with noisy gradients. It brings speed-ups and *improve solutions* when training non-convex GANs and mixtures of GANs, thus combining the benefits of alternation and extrapolation in adversarial training. Numerical experiments show the importance of *sampling schemes*. We take a first step towards understanding the good behavior of *cyclic* player sampling through spectral analysis. We foresee interesting developments using player sampling in reinforcement learning: the policy gradients obtained using multi-agent actor critic methods (Lowe et al., 2017) are noisy estimates, a setting in which it is beneficial.

References

Bach, Francis and Kfir Levy (2019). "A universal algorithm for variational inequalities adaptive to smoothness and noise". In: *Proceedings of the Conference on Learning Theory*.

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.

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: *Advances in Neural Information Processing Systems*.

Cheung, Yun Kuen and Georgios Piliouras (2019). "Vortices Instead of Equilibria in MinMax Optimization: Chaos and Butterfly Effects of Online Learning in Zero-Sum Games". In: *Proceedings of the Conference on Learning Theory*.

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.

Ghosh, Arnab et al. (2018). "Multi-Agent Diverse Generative Adversarial Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

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.

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

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.

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

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.

Nikaidô, Hukukane and Kazuo Isoda (1955). "Note on Non-Cooperative Convex Games". In: *Pacific Journal of Mathematics* 5.Suppl. 1, pp. 807–815.

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. B. (1965). "Existence and Uniqueness of Equilibrium Points for Concave N-Person Games". In: *Econometrica* 33.3, pp. 520–534.

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

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

A Existing results

A.1 Mirror-prox

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

□

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

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

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

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

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

A.2 Link between convex games and variational inequalities

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

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

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

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

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

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

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

B Proofs and mirror-setting algorithms

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

B.1 Proof of Corollary 1

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

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

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

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

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

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

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

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

B.2 Useful lemmas

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

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

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

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

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

Lemma 3. *Let $\xi_1, \xi_2, \dots$ 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_τ is a measurable function of y_0 and $\xi_1, \dots, \xi_\tau$ such that:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Proof. On the one hand,

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

and by Cauchy-Schwarz inequality

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

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

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

□

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

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

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

Proof. For j even, we can write

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

For j odd,

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

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

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

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

Proof.

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

□

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

B.3.1 Algorithm

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 (Assumption 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 shown in equation (23).

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

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

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

The optimal value for γ_τ is

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

and the optimal value of the bound is

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

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

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

which yields equation (14) after simplification. □

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

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

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

B.4.1 Algorithm

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

Algorithm 4 Mirror prox with variance reduced player randomness

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

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

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

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

Theorem 5. *Assume that for all i between 1 and n, the gradients $\nabla_i \ell_i$ are L-Lipschitz (2b). Assume Algorithm 4 is run with constant stepsizes $\gamma_\tau = \gamma$, with γ defined as*

$$\gamma \triangleq \min\left\{\frac{p^{3/2}}{\sqrt{(1-p)(2-p)}}\frac{1}{12L\sqrt{n}}, \frac{1}{L}\sqrt{\frac{5}{27n+12}}, \frac{1}{2}\sqrt{\frac{\Omega}{13n\sigma^2 t(k)}}\right\},$$

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

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

Outline of the proof of Theorem 5.

- Lemma 12 provides a bound for $\mathbb{E}\left[\sum_{\tau=0}^{t}\gamma_\tau^2\|\tilde{F}_{\tau+1} - F(\theta_{\tau+1/2})\|_\star^2 + \gamma_\tau^2\|F(\theta_\tau) - \tilde{F}_{\tau+1/2}\|_\star^2\right]$ and it is the keystone of the proof. It specifically uses the structure of player sampling and the introduced variance reduction mechanism.
- Lemma 10 and 11 are intermediate steps in the proof of Lemma 12. Lemma 9 and Lemma 8 are used in the proof of Lemma 11.
- We prove Theorem 5 by refining base inequalities established by Juditsky et al. (2011), using the results from Lemma 12.

Definition 4. *For a given j and i (which we omit), let us define K_j as the random variable indicating the highest $q \in \mathbb{N}$ strictly lower than j such that $M_{q/2}^{(i)}$ 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 $(R_{q/2}^{(i)})_{q\in\mathbb{N}}$ was updated with a new value $\hat{F}^{(i)}(\theta_{q/2})$. That is, $R_{j/2,i} = \hat{F}^{(i)}(\theta_{K_j/2})$.

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

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

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

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

Lemma 10. *The following equalities hold:*

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

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

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

The second equality is derived analogously. □

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

Lemma 11. *Let us define $h : \mathbb{R} \to \mathbb{R}$ as*

$$h(p) \triangleq \frac{2-p}{p^2}. \tag{25}$$

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

$$\begin{aligned}
\sum_{\tau=0}^{t}\gamma_\tau^2\mathbb{E}\left[\|R^{(i)}_\tau - \hat{F}^{(i)}(\theta_\tau)\|^2_\star\right] &\leqslant \sum_{j=0}^{2t-1} h(p)\gamma^2_{\lfloor j/2\rfloor}\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{j/2}) - \hat{F}^{(i)}(\theta_{(j+1)/2})\|^2_\star\right], \\
\sum_{\tau=0}^{t}\gamma_\tau^2\mathbb{E}\left[\|R^{(i)}_{\tau+1/2} - \hat{F}^{(i)}(\theta_{\tau+1/2})\|^2_\star\right] &\leqslant \sum_{j=0}^{2t-1} h(p)\gamma^2_{\lfloor j/2\rfloor}\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{j/2}) - \hat{F}^{(i)}(\theta_{(j+1)/2})\|^2_\star\right].
\end{aligned} \tag{26}$$

Proof. We can write

$$\begin{aligned}
\mathbb{E}\left[\|R^{(i)}_\tau - \hat{F}^{(i)}(\theta_\tau)\|^2_\star\right] &= \mathbb{E}\left[\|R^{(i)}_{2\tau/2} - \hat{F}^{(i)}(\theta_{2\tau/2})\|^2_\star\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\|R^{(i)}_{2\tau/2} - \hat{F}^{(i)}(\theta_{2\tau/2})\|^2_\star \middle| K_{2\tau}\right]\right] \\
&= \sum_{q=0}^{2\tau-1} P(K_{2\tau} = q)\mathbb{E}\left[\|R^{(i)}_{2\tau/2} - \hat{F}^{(i)}(\theta_{2\tau/2})\|^2_\star \middle| K_{2\tau} = q\right] \\
&= \sum_{q=1}^{2\tau-1} p(1-p)^{2\tau-1-q}\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{q/2}) - \hat{F}^{(i)}(\theta_{2\tau/2})\|^2_\star\right] \\
&+ (1-p)^{2\tau-1}\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_0) - \hat{F}^{(i)}(\theta_{2\tau/2})\|^2_\star\right].
\end{aligned} \tag{27}$$

As seen in equation (27), the point of conditioning with respect to the sigma-field generated by $K_{2\tau}$ (see Definition 4) is that we can write the expression for $R_{2\tau/2,i}$. We have used Lemma 9.

Now, using the rearrangement inequality,

$$\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{q/2}) - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2\right] = \mathbb{E}\left[\left\|\sum_{j=q}^{2\tau-1} \hat{F}^{(i)}(\theta_{j/2}) - \hat{F}^{(i)}(\theta_{(j+1)/2})\right\|_\star^2\right] \tag{28}$$
$$\leqslant \sum_{j=q}^{2\tau-1} (2\tau - q)\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{j/2}) - \hat{F}^{(i)}(\theta_{(j+1)/2})\|_\star^2\right].$$

Using equations (27) and (28) we can now write

$$\begin{aligned}
&\sum_{\tau=0}^{t} \gamma_\tau^2 \mathbb{E}\left[\|R_\tau^{(i)} - \hat{F}^{(i)}(\theta_\tau)\|_\star^2\right] \qquad (29)\\
&= \sum_{\tau=0}^{t} \gamma_\tau^2 \sum_{q=1}^{2\tau-1} p(1-p)^{2\tau-1-q}\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{q/2}) - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2\right]\\
&+ \gamma_\tau^2(1-p)^{2\tau-1}\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_0) - \hat{F}^{(i)}(\theta_{2\tau/2})\|_\star^2\right]\\
&\leqslant \sum_{\tau=0}^{t} \gamma_\tau^2 \sum_{q=1}^{2\tau-1} p(1-p)^{2\tau-1-q} \sum_{j=q}^{2\tau-1} (2\tau - q)\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{j/2}) - \hat{F}^{(i)}(\theta_{(j+1)/2})\|_\star^2\right]\\
&+ \gamma_\tau^2(1-p)^{2\tau-1} \sum_{j=0}^{2\tau-1} 2\tau\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{j/2}) - \hat{F}^{(i)}(\theta_{(j+1)/2})\|_\star^2\right].
\end{aligned}$$

Given j between 0 and $2t-1$ the right hand side of equation (29) 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 (29) 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 (27):

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

Using the same reasoning we get an inequality that is analogous to (26):

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

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

$$
\chi(p, \gamma) = 1 - 36 \frac{1-p}{p} n h(p) L^2 \gamma^2. \tag{30}
$$

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

$$\mathbb{E}\left[\sum_{\tau=0}^{t}\gamma_\tau^2\|\tilde{F}_{\tau+1}-F(\theta_{\tau+1/2})\|_\star^2+\gamma_\tau^2\|F(\theta_\tau)-\tilde{F}_{\tau+1/2}\|_\star^2\right] \tag{31}$$
$$\leqslant 104n\sigma^2\sum_{\tau=0}^{t}\gamma_\tau^2+\frac{1-p}{p\chi(p,\gamma)}(12L^2+36L^4\gamma^2)nh(p)\sum_{\tau=0}^{t}\gamma_\tau^2\mathbb{E}\left[\|\theta_\tau-\theta_{\tau+1/2}\|_\star^2\right].$$

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

$$\begin{aligned}\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{j/2})-\hat{F}^{(i)}(\theta_{(j+1)/2})\|_\star^2\right]&=\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_\tau)-\hat{F}^{(i)}(\theta_{\tau+1/2})\|_\star^2\right]\\&\leqslant 3\mathbb{E}\left[\|F^{(i)}(\theta_\tau)-F^{(i)}(\theta_{\tau+1/2})\|_\star^2\right]\\&\quad+3\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_\tau)-F^{(i)}(\theta_{\tau+1/2})\|_\star^2\right]\\&\quad+3\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{\tau+1/2})-F^{(i)}(\theta_{\tau+1/2})\|_\star^2\right]\\&\leqslant 3L^2\mathbb{E}\left[\|\theta_\tau-\theta_{\tau+1/2}\|_\star^2\right]+6\sigma^2.\end{aligned} \tag{32}$$

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):

$$\begin{aligned}\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{j/2})-\hat{F}^{(i)}(\theta_{(j+1)/2})\|_\star^2\right]&=\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{\tau+1/2})-\hat{F}^{(i)}(\theta_{\tau+1})\|_\star^2\right]\\&\leqslant 3\mathbb{E}\left[\|F^{(i)}(\theta_{\tau+1/2})-F^{(i)}(\theta_{\tau+1})\|_\star^2\right]\\&\quad+3\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{\tau+1/2})-F^{(i)}(\theta_{\tau+1/2})\|_\star^2\right]\\&\quad+3\mathbb{E}\left[\|\hat{F}^{(i)}(\theta_{\tau+1})-F^{(i)}(\theta_{\tau+1})\|_\star^2\right]\\&\leqslant 3L^2\mathbb{E}\left[\|\theta_{\tau+1/2}-\theta_{\tau+1}\|_\star^2\right]+6\sigma^2\\&=3L^2\mathbb{E}\left[\|P_{\theta_\tau}(\gamma_\tau\tilde{F}_{\tau+1/2})-P_{\theta_\tau}(\gamma_\tau\tilde{F}_{\tau+1})\|_\star^2\right]+6\sigma^2\\&\leqslant 3L^2\gamma_\tau^2\mathbb{E}\left[\|\tilde{F}_{\tau+1/2}-\tilde{F}_{\tau+1}\|_\star^2\right]+6\sigma^2\\&\leqslant 9L^2\gamma_\tau^2\Big(\mathbb{E}\left[\|\tilde{F}_{\tau+1/2}-F(\theta_\tau)\|_\star^2\right]\\&\qquad+\mathbb{E}\left[\|F(\theta_{\tau+1/2})-\tilde{F}_{\tau+1}\|_\star^2\right]\\&\qquad+\mathbb{E}\left[\|F(\theta_\tau)-F(\theta_{\tau+1/2})\|_\star^2\right]\Big)+6\sigma^2.\end{aligned} \tag{33}$$

Now, we use Lemma 6 to break up the dual norms in the right-hand side of (31).

$$\begin{aligned}&\mathbb{E}\left[\sum_{\tau=0}^{t}\gamma_\tau^2\|\tilde{F}_{\tau+1}-F(\theta_{\tau+1/2})\|_\star^2+\gamma_\tau^2\|F(\theta_\tau)-\tilde{F}_{\tau+1/2}\|_\star^2\right]\\&=\mathbb{E}\left[\sum_{\tau=0}^{t}\sum_{i=1}^{n}\gamma_\tau^2\|\tilde{F}_{\tau+1}^{(i)}-F^{(i)}(\theta_{\tau+1/2})\|_\star^2+\gamma_\tau^2\|F^{(i)}(\theta_\tau)-\tilde{F}_{\tau+1/2}^{(i)}\|_\star^2\right],\end{aligned} \tag{34}$$

Hence, from equation (34) and Lemma 10 and 11:

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

We split the last term in summands corresponding to even and odd j, we change variables from j to τ and we apply equations (32) and (33):

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

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

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

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

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

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

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

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

Using equation (35) and the analogous equation to (19), we reach the following inequality:

$$\begin{aligned}&\mathbb{E}\left[\sup_{u \in Z} \sum_{\tau=0}^{t} \langle \gamma_\tau F(\theta_{\tau+1/2}), \theta_{\tau+1/2} - u \rangle\right] \leqslant \mathbb{E}\left[\sup_{u \in Z} 2D(u, \theta_0) - D(u, \theta_{t+1}) - \sum_{\tau=0}^{t} \frac{1}{2} \|\theta_{\tau+1/2} - \theta_\tau\|_2^2\right] \\ &+ \mathbb{E}\left[\sum_{\tau=0}^{t} 2\gamma_\tau^2 \|\tilde{F}_{\tau+1} - F(\theta_{\tau+1/2})\|_\star^2 + \frac{3\gamma_\tau^2}{2} \|F(\theta_\tau) - \tilde{F}_{\tau+1/2}\|_\star^2 + \frac{3\gamma_\tau^2}{2} \|F(\theta_{\tau+1/2}) - F(\theta_\tau)\|_\star^2\right] \end{aligned} \tag{36}$$

Taking the definition of $\chi(p, \gamma)$ in (30), using the definition of $h(p)$ in (25) and rearranging, we obtain

$$\gamma \leqslant \frac{p^{3/2}}{\sqrt{(1-p)(2-p)}} \frac{1}{12L\sqrt{n}} \iff \chi(p, \gamma) \geqslant 3/4 > 0. \tag{37}$$

Hence, the assumptions of Lemma 12 are fulfilled. Starting from the result in (36) and using Lemma 12,

$$\begin{aligned}&\mathbb{E}\left[\sup_{u \in Z} \sum_{\tau=0}^{t} \langle \gamma_\tau F(\theta_{\tau+1/2}), \theta_{\tau+1/2} - u \rangle\right] \\ &\leqslant \mathbb{E}\left[\sup_{u \in Z} 2D(u, \theta_0) - D(u, \theta_t)\right] + 32n\sigma^2 \sum_{\tau=0}^{t} \gamma_\tau^2 \\ &+ 2\frac{1-p}{p\chi(p, \gamma)}(12L^2 + 36L^4\gamma^2)nh(p) \sum_{\tau=0}^{t} \gamma_\tau^2 \mathbb{E}\left[\|\theta_\tau - \theta_{\tau+1/2}\|_\star^2\right] \\ &+ \frac{3nL^2}{2} \sum_{\tau=0}^{t} \gamma_\tau^2 \mathbb{E}\left[\|\theta_\tau - \theta_{\tau+1/2}\|_\star^2\right] - \frac{1}{2} \sum_{\tau=0}^{t} \mathbb{E}\left[\|\theta_\tau - \theta_{\tau+1/2}\|_\star^2\right] \\ &\leqslant 2\Omega + 104n\sigma^2 \sum_{\tau=0}^{t} \gamma_\tau^2 \\ &+ \left((24L^2 + 72L^4\gamma^2)nh(p)\gamma^2 \frac{1-p}{p\chi(p, \gamma)} + \frac{3n\gamma^2L^2}{2} - \frac{1}{2}\right) \sum_{\tau=0}^{t} \mathbb{E}\left[\|\theta_\tau - \theta_{\tau+1/2}\|_\star^2\right]. \end{aligned} \tag{38}$$

Recalling the definition of $h(p)$ in Equation (25), the conditions $\chi(p, \gamma) \geqslant 3/4$ and

$$\gamma \leqslant \frac{1}{L}\sqrt{\frac{5}{27n + 12}}, \tag{39}$$

imply

$$(24L^2 + 72L^4\gamma^2)nh(p)\gamma^2 \frac{1-p}{p\chi(p, \gamma)} + \frac{3n\gamma^2L^2}{2} - \frac{1}{2} \leqslant 0. \tag{40}$$

We show this development:

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

Using Equation (40) on (38) yields

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

By Lemma 4, we conclude

$$
\mathrm{Err}_N(\hat{\theta}_t) \leqslant \left(\sum_{\tau=0}^{t}\gamma_\tau\right)^{-1}\left(2\Omega + 104n\sigma^2\sum_{\tau=0}^{t}\gamma_\tau^2\right) \tag{41}
$$

Now we apply Lemma 5 to equation (41) assuming constant stepsizes. That is, we set $\gamma_\tau = 1$, $A = 2\Omega$ and $B = 104n\sigma^2$. Using the notation from Lemma 5, we get that

$$
\alpha^* = \frac{1}{2}\sqrt{\frac{\Omega}{13n\sigma^2 t}}
$$

and the value of the bound at α^* is

$$
8\sqrt{\frac{13\Omega n\sigma^2}{t}}.
$$

However, γ is also subject to the constraints in equations (37) and (39). Namely,

$$
\gamma \triangleq \min\left\{\frac{p^{3/2}}{\sqrt{(1-p)(2-p)}}\frac{1}{12L\sqrt{n}}, \frac{1}{L}\sqrt{\frac{5}{27n+12}}, \frac{1}{2}\sqrt{\frac{\Omega}{13n\sigma^2 t}}\right\}, \tag{42}
$$

If the minimum in equation (42) is not achieved at α^* (the third term), it is easy to see that the first term of the bound in equation (41) is larger than the second one, which means that $4\Omega/(\gamma t)$ is a looser bound. We conclude

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

Substituting γ for its expression and plugging $t(k) = k/2b$ on equation B.4.1 we get

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

The result follows using $1 - b/n < 1$ and $2 - b/n < 2$. □

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

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

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

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

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

C.1 Recursion operator for the different sampling schemes

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

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

$$\begin{cases} \theta^{\text{full}}_{k+2} = \theta^{\text{full}}_k - \gamma A \theta^{\text{full}}_k, \\ \theta^{\text{full}}_{k+4} = \theta^{\text{full}}_k - \gamma A \theta^{\text{full}}_{k+2}. \end{cases} \tag{43}$$

By introducing $\mathcal{A}^{(\gamma)}_{\text{full}} := I - \gamma A + \gamma^2 A^2$, (43) is simply $\theta^{\text{full}}_{k+4} = \mathcal{A}^{(\gamma)}_{\text{full}} \theta^{\text{full}}_k$.

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

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

the updates becomes

$$\begin{cases} \theta^{\text{cyc}}_{k+1} = \theta^{\text{cyc}}_k - \gamma M_1 A \theta^{\text{cyc}}_k, \\ \theta^{\text{cyc}}_{k+2} = \theta^{\text{cyc}}_k - \gamma M_2 A \theta^{\text{cyc}}_{k+1}, \\ \theta^{\text{cyc}}_{k+3} = \theta^{\text{cyc}}_{k+2} - \gamma M_2 A \theta^{\text{cyc}}_{k+2}, \\ \theta^{\text{cyc}}_{k+4} = \theta^{\text{cyc}}_{k+2} - \gamma M_1 A \theta^{\text{cyc}}_{k+3}. \end{cases} . \tag{44}$$

Remark that (44) contains two iterations of Algorithm 1; θ_{k+1} and θ_{k+3} are extrapolations and θ_{k+2} and θ_{k+4} are updates. Defining $\mathcal{A}^{(\gamma)}_{ij} := I - \gamma M_i A + \gamma^2 M_i A M_j A$ and $\mathcal{A}^{(\gamma)}_{\text{cyc}} := \mathcal{A}^{\gamma}_{12} \mathcal{A}^{(\gamma)}_{21}$, we have $\theta^{\text{cyc}}_{k+4} = \mathcal{A}^{(\gamma)}_{\text{cyc}} \theta^{\text{cyc}}_k$.

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

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

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

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

C.2 Convergence behavior through spectral analysis

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

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

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

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

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

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

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

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

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

C.3 Empirical distributions of the spectral radii

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

$$\mathcal{A}^\star_{\text{full}} \triangleq \min_{\gamma\in\mathbb{R}^+} \rho(\mathcal{A}^{(\gamma)}_{\text{full}}), \quad \mathcal{A}^\star_{\text{cyc}} \triangleq \min_{\gamma\in\mathbb{R}^+} \rho(\mathcal{A}^{(\gamma)}_{\text{cyc}}), \quad \mathcal{A}^\star_{\text{rand}} \triangleq \min_{\gamma\in\mathbb{R}^+} \rho(\mathcal{A}^{(\gamma)}_{\text{rand}}), \tag{45}$$

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.

Experiment. We sample matrices A in $\mathbb{R}^{2d\times 2d}$ (with $d = 3$) as the weighted sum of a random positive definite matrix A_{sym} and of a random skew matrix A_{skew}. We refer to Appendix D for a detailed description of the matrix sampling method. We vary the weight $\alpha \in [0, 1]$ of the skew matrix and the lowest eigenvalue μ of the matrix A_{sym}. We sample 300 different games and compute $\mathcal{A}^{(\eta)}$ on a grid of step sizes η, for the three different methods. We thus estimate the best algorithmic spectral radii defined in (45).

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

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

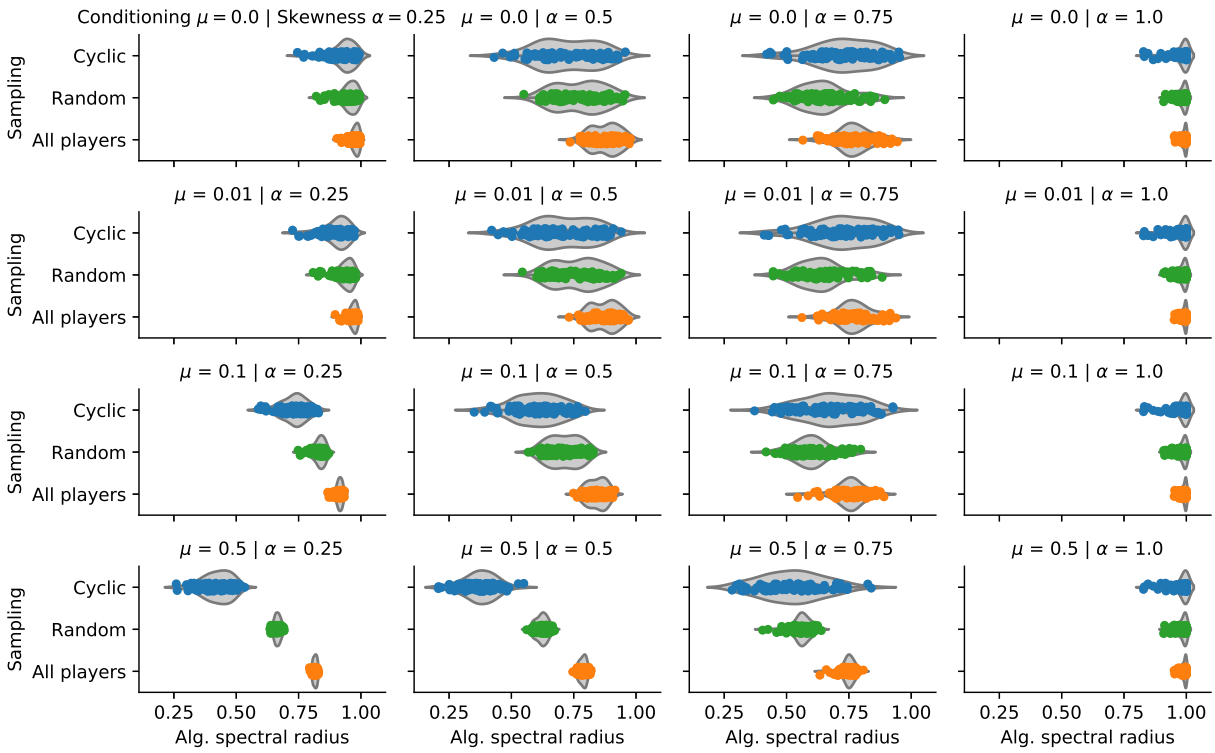


Figure 6: Spectral radii distribution of the algorithmic operator associated to doubly-stochastic and full extra-gradient, in the non-constrained bi-linear two-player game setting, for various conditioning and skewness. Random and cyclic sampling yields lower radius (hence faster rates) for most problem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

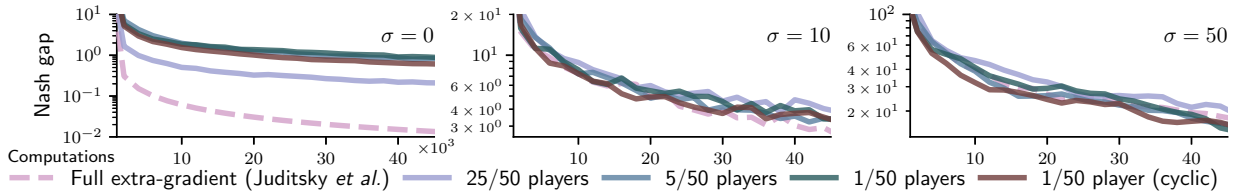


Figure 7: 50-player completely skew smooth game with increasing noise (sampling with variance reduction). In the non-noisy setting, player sampling reduces convergence speed. On the other hand, it provides a speed-up in the high noise regime.

D Experimental results and details

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

D.1 Quadratic games

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

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

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

Table 2: Parameters used in Figure 2 for increasing problem complexity.

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