A VARIATIONAL INEQUALITY PERSPECTIVE ON GENERATIVE ADVERSARIAL NETWORKS

Gauthier Gidel[1,*] **Hugo Berard**[1,3,*] **Gaëtan Vignoud**[1] **Pascal Vincent**[1,2,3]
Simon Lacoste-Julien[1,2]
[1] Mila & DIRO, Université de Montréal. [2] CIFAR fellow.
[3] Facebook Artificial Intelligence Research. Montreal.

ABSTRACT

Generative adversarial networks (GANs) form a generative modeling approach known for producing appealing samples, but they are notably difficult to train. One common way to tackle this issue has been to propose new formulations of the GAN objective. Yet, surprisingly few studies have looked at optimization methods designed for this adversarial training. In this work, we cast GAN optimization problems in the general variational inequality framework. Tapping into the mathematical programming literature, we counter some common misconceptions about the difficulties of saddle point optimization and propose to extend techniques designed for variational inequalities to the training of GANs. We apply *averaging*, *extrapolation* and a novel computationally cheaper variant that we call *extrapolation from the past* to the stochastic gradient method (SGD) and Adam.

1 INTRODUCTION

Generative adversarial networks (GANs) (Goodfellow et al., 2014) form a generative modeling approach known for producing realistic natural images (Karras et al., 2018) as well as high quality super-resolution (Ledig et al., 2017) and style transfer (Zhu et al., 2017). Nevertheless, GANs are also known to be difficult to train, often displaying an unstable behavior (Goodfellow, 2016). Much recent work has tried to tackle these training difficulties, usually by proposing new formulations of the GAN objective (Nowozin et al., 2016; Arjovsky et al., 2017). Each of these formulations can be understood as a two-player game, in the sense of game theory (Von Neumann and Morgenstern, 1944), and can be addressed as a variational inequality problem (VIP) (Harker and Pang, 1990), a framework that encompasses traditional saddle point optimization algorithms (Korpelevich, 1976).

Solving such GAN games is traditionally approached by running variants of stochastic gradient descent (SGD) initially developed for optimizing supervised neural network objectives. Yet it is known that for some games (Goodfellow, 2016, §8.2) SGD exhibits oscillatory behavior and fails to converge. This oscillatory behavior, which does not arise from stochasticity, highlights a fundamental problem: while a direct application of basic gradient descent is an appropriate method for regular minimization problems, it is *not* a sound optimization algorithm for the kind of two-player games of GANs. This constitutes a fundamental issue for GAN training, and calls for the use of more principled methods with more reassuring convergence guarantees.

Contributions. We point out that multi-player games can be cast as *variational inequality problems* (VIPs) and consequently the same applies to any GAN formulation posed as a minimax or non-zero-sum game. We present two techniques from this literature, namely *averaging* and *extrapolation*, widely used to solve VIPs but which have not been explored in the context of GANs before.[1]

We extend standard GAN training methods such as SGD or Adam into variants that incorporate these techniques (Alg. 3, 4 are new). We also explain that the oscillations of basic SGD for GAN training previously noticed (Goodfellow, 2016) can be explained by standard variational inequality optimization results and we illustrate how *averaging* and *extrapolation* can fix this issue.

[*]Equal contribution, correspondence to `firstname.lastname@umontreal.ca`.

[1]Independent works (Mertikopoulos et al., 2018) and (Yazıcı et al., 2018) respectively explored extrapolation and averaging in the context of GANs. More details in related work section §6.

We introduce a new technique, called *extrapolation from the past*, that only requires one gradient computation per iteration compared to *extrapolation* which requires to compute the gradient twice. We *prove its convergence* in the stochastic variational inequality setting, i.e. when applied to SGD.

Finally, we test these techniques in the context of standard GAN training. We observe a 4%-6% improvement on the inception score (Salimans et al., 2016) of WGAN (Arjovsky et al., 2017) and WGAN-GP (Gulrajani et al., 2017) on the CIFAR-10 dataset.

Outline. §2 presents the background on GAN and optimization, and shows how to cast this optimization as a VIP. §3 presents standard techniques to optimize variational inequalities in a batch setting as well as our new one, *extrapolation from the past*. §4 considers these methods in the *stochastic* setting, yielding three corresponding variants of SGD, and provides their respective convergence rates. §5 develops how to combine these techniques with already existing algorithms. §6 discusses the related work and §7 presents experimental results.

2 GAN OPTIMIZATION AS A VARIATIONAL INEQUALITY PROBLEM

2.1 GAN FORMULATIONS

The purpose of generative modeling is to generate samples from a distribution $q_{\boldsymbol{\theta}}$ that matches best the true distribution p of the data. The generative adversarial network training strategy can be understood as a *game* between two players called *generator* and *discriminator*. The former produces a sample that the latter has to classify between real or fake data. The final goal is to build a generator able to produce sufficiently realistic samples to fool the discriminator.

In the original GAN paper (Goodfellow et al., 2014), the GAN objective is formulated as a *zero-sum game* where the cost function of the discriminator $D_{\boldsymbol{\varphi}}$ is given by the negative log-likelihood of the binary classification task between real or fake data generated from $q_{\boldsymbol{\theta}}$ by the generator,

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\varphi}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}) \quad \text{where} \quad \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}) \stackrel{\text{def}}{=} -\mathop{\mathbb{E}}_{\mathbf{x}\sim p}[\log D_{\boldsymbol{\varphi}}(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x}'\sim q_{\boldsymbol{\theta}}}[\log(1 - D_{\boldsymbol{\varphi}}(\mathbf{x}'))] \,. \tag{1}$$

However Goodfellow et al. (2014) recommends to use in practice a second formulation, called *non-saturating GAN*. This formulation is a *non-zero-sum game* where the aim is to jointly minimize:

$$\mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\theta}, \boldsymbol{\varphi}) \stackrel{\text{def}}{=} -\mathop{\mathbb{E}}_{\mathbf{x}'\sim q_{\boldsymbol{\theta}}} \log D_{\boldsymbol{\varphi}}(\mathbf{x}') \quad \text{and} \quad \mathcal{L}^{(\boldsymbol{\varphi})}(\boldsymbol{\theta}, \boldsymbol{\varphi}) \stackrel{\text{def}}{=} -\mathop{\mathbb{E}}_{\mathbf{x}\sim p} \log D_{\boldsymbol{\varphi}}(\mathbf{x}) - \mathop{\mathbb{E}}_{\mathbf{x}'\sim q_{\boldsymbol{\theta}}} \log(1 - D_{\boldsymbol{\varphi}}(\mathbf{x}')) \,. \tag{2}$$

The dynamics of this formulation have the same *stationary points* as the zero-sum one (1) but are claimed to provide "much stronger gradients early in learning" (Goodfellow et al., 2014) .

2.2 EQUILIBRIUM

The minimax formulation (1) is theoretically convenient because a large literature on games studies this problem and provides guarantees on the existence of equilibria. Nevertheless, practical considerations lead the GAN literature to consider a different objective for each player as formulated in (2). In that case, the *two-player game problem* (Von Neumann and Morgenstern, 1944) consists in finding the following *Nash equilibrium*:

$$\boldsymbol{\theta}^* \in \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\theta}, \boldsymbol{\varphi}^*) \quad \text{and} \quad \boldsymbol{\varphi}^* \in \operatorname*{arg\,min}_{\boldsymbol{\varphi}\in\Phi} \mathcal{L}^{(\boldsymbol{\varphi})}(\boldsymbol{\theta}^*, \boldsymbol{\varphi}) \,. \tag{3}$$

Only when $\mathcal{L}^{(\boldsymbol{\theta})} = -\mathcal{L}^{(\boldsymbol{\varphi})}$ is the game called a *zero-sum game* and (3) can be formulated as a minimax problem. One important point to notice is that the two optimization problems in (3) are *coupled* and have to be considered *jointly* from an optimization point of view.

Standard GAN objectives are non-convex (i.e. each cost function is non-convex), and thus such (pure) equilibria may not exist. As far as we know, not much is known about the existence of these equilibria for non-convex losses (see Heusel et al. (2017) and references therein for some results). In our theoretical analysis in §4, our assumptions (monotonicity (24) of the operator and convexity of the constraints set) imply the existence of an equilibrium.

In this paper, we focus on ways to optimize these games, assuming that an equilibrium exists. As is often standard in non-convex optimization, we also focus on finding points satisfying the necessary

stationary conditions. As we mentioned previously, one difficulty that emerges in the optimization of such games is that the two different cost functions of (3) have to be minimized jointly in $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$. Fortunately, the optimization literature has for a long time studied so-called *variational inequality problems*, which generalize the stationary conditions for two-player game problems.

2.3 Variational inequality problem formulation

We first consider the local necessary conditions that characterize the solution of the *smooth* two-player game (3), defining *stationary points*, which will motivate the definition of a variational inequality. In the unconstrained setting, a *stationary point* is a couple $(\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*)$ with zero gradient:

$$\|\nabla_{\boldsymbol{\theta}} \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*)\| = \|\nabla_{\boldsymbol{\varphi}} \mathcal{L}^{(\boldsymbol{\varphi})}(\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*)\| = 0\,. \tag{4}$$

When constraints are present,[2] a *stationary point* $(\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*)$ is such that the directional derivative of each cost function is non-negative in any feasible direction (i.e. there is no feasible descent direction):

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \geq 0 \quad \text{and} \quad \nabla_{\boldsymbol{\varphi}} \mathcal{L}^{(\boldsymbol{\varphi})}(\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*)^\top (\boldsymbol{\varphi} - \boldsymbol{\varphi}^*) \geq 0\,, \quad \forall\, (\boldsymbol{\theta}, \boldsymbol{\varphi}) \in \Theta \times \Phi. \tag{5}$$

Defining $\boldsymbol{\omega} \stackrel{\text{def}}{=} (\boldsymbol{\theta}, \boldsymbol{\varphi})$, $\boldsymbol{\omega}^* \stackrel{\text{def}}{=} (\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*)$, $\Omega \stackrel{\text{def}}{=} \Theta \times \Phi$, Eq. (5) can be compactly formulated as:

$$F(\boldsymbol{\omega}^*)^\top (\boldsymbol{\omega} - \boldsymbol{\omega}^*) \geq 0\,, \quad \forall \boldsymbol{\omega} \in \Omega \quad \text{where} \quad F(\boldsymbol{\omega}) \stackrel{\text{def}}{=} \begin{bmatrix} \nabla_{\boldsymbol{\theta}} \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\theta}, \boldsymbol{\varphi}) & \nabla_{\boldsymbol{\varphi}} \mathcal{L}^{(\boldsymbol{\varphi})}(\boldsymbol{\theta}, \boldsymbol{\varphi}) \end{bmatrix}^\top . \tag{6}$$

These stationary conditions can be generalized to any continuous vector field: let $\Omega \subset \mathbb{R}^d$ and $F : \Omega \to \mathbb{R}^d$ be a continuous mapping. The *variational inequality problem* (Harker and Pang, 1990) (depending on F and Ω) is:

$$\text{find } \boldsymbol{\omega}^* \in \Omega \quad \text{such that} \quad F(\boldsymbol{\omega}^*)^\top (\boldsymbol{\omega} - \boldsymbol{\omega}^*) \geq 0\,, \quad \forall \boldsymbol{\omega} \in \Omega\,. \tag{VIP}$$

We call *optimal set* the set Ω^* of $\boldsymbol{\omega} \in \Omega$ verifying (VIP). The intuition behind it is that any $\boldsymbol{\omega}^* \in \Omega^*$ is a *fixed point* of the *constrained* dynamic of F (constrained to Ω).

We have thus showed that both saddle point optimization and non-zero sum game optimization, which encompasses the large majority of GAN variants proposed in the literature, can be cast as VIPs. In the next section, we turn to suitable optimization techniques for such problems.

3 Optimization of Variational Inequalities (batch setting)

Let us begin by looking at techniques that were developed in the optimization literature to solve (VIP). We present the intuitions behind them as well as their performance on a simple bilinear problem (see Fig. 1). Our goal here is to provide mathematical insights into the techniques of *averaging* (§3.1) and *extrapolation* (§3.2), to inspire their application to extending other optimization algorithm. We then propose a novel variant of the extrapolation technique in §3.3 *extrapolation from the past*. We here treat the batch setting, i.e. considering that the operator $F(\boldsymbol{\omega})$ as defined in Eq. 6 yields an exact full gradient. We will present extensions of these techniques to the stochastic setting later in §4.

The two standard methods studied in the VIP literature are the *gradient method* (Bruck, 1977) and the *extragradient method* (Korpelevich, 1976). The iterates of the basic gradient method are given by $\boldsymbol{\omega}_{t+1} = P_\Omega[\boldsymbol{\omega}_t - \eta F(\boldsymbol{\omega}_t)]$ where $P_\Omega[\cdot]$ is the *projection onto the constraints set* (if constraints are present) associated to (VIP). These iterates are known to converge linearly under an additional assumption on the operator[3] (Chen and Rockafellar, 1997), but oscillate for a bilinear operator as shown in Fig. 1. On the other hand, the *uniform average* of these iterates converge for any bounded monotone operator with a $O(1/\sqrt{t})$ rate (Nedić and Ozdaglar, 2009), motivating the presentation of *averaging* in §3.1. By contrast, the *extragradient method* (extrapolated gradient) does not require any averaging to converge for monotone operators (in the batch setting), and can even converge at the faster $O(1/t)$ rate (Nesterov, 2007). The idea of this method is to compute a lookahead step (see intuition on *extrapolation* in §3.2) in order to compute a more stable direction to follow.

[2] An example of constraint for GANs is to clip the parameters of the discriminator (Arjovsky et al., 2017).

[3] Strong monotonicity, a generalization of strong convexity. See §A.

3.1 AVERAGING

More generally, we consider a *weighted averaging* scheme with weights $\rho_t \geq 0$. This *weighted averaging* scheme have been proposed for the first time for (batch) VIP by Bruck (1977),

$$\bar{\boldsymbol{\omega}}_T \stackrel{\text{def}}{=} \frac{\sum_{t=0}^{T-1} \rho_t \boldsymbol{\omega}_t}{S_T}\,, \quad S_T \stackrel{\text{def}}{=} \sum_{t=0}^{T-1} \rho_t\,. \tag{7}$$

Averaging schemes can be efficiently implemented in an online fashion noticing that,

$$\bar{\boldsymbol{\omega}}_t = (1-\tilde{\rho}_t)\bar{\boldsymbol{\omega}}_{t-1} + \tilde{\rho}_t \boldsymbol{\omega}_t \quad \text{where} \quad 0 \leq \tilde{\rho}_t \leq 1\,. \tag{8}$$

For instance, setting $\tilde{\rho}_t = \frac{1}{t}$ provides *uniform averaging* ($\rho_t = 1$) and $\tilde{\rho}_t = 1-\beta < 1$ provides *geometric averaging* also known as *exponential moving averaging* ($\rho_t = \beta^t$). Averaging is experimentally compared with the other techniques presented in this section in Fig. 1.

In order to illustrate how averaging tackles the oscillatory behavior in game optimization, we consider a toy example where the discriminator and the generator are linear: $D_{\boldsymbol{\varphi}}(\mathbf{x}) = \boldsymbol{\varphi}^T\mathbf{x}$ and $G_{\boldsymbol{\theta}}(\mathbf{z}) = \boldsymbol{\theta}\mathbf{z}$ (implicitly defining $q_{\boldsymbol{\theta}}$). By replacing these expressions in the WGAN objective,[4] we get the following bilinear objective:

$$\min_{\boldsymbol{\theta}\in\Theta} \max_{\boldsymbol{\varphi}\in\Phi, ||\boldsymbol{\varphi}||\leq 1} \boldsymbol{\varphi}^T\mathbb{E}[\mathbf{x}] - \boldsymbol{\varphi}^T\boldsymbol{\theta}\mathbb{E}[\mathbf{z}]\,. \tag{9}$$

A similar task was presented by Nagarajan and Kolter (2017) where they consider a quadratic discriminator instead of a linear one, and show that gradient descent is not necessarily asymptotically stable. The bilinear objective has been extensively used (Goodfellow, 2016; Mescheder et al., 2018; Yadav et al., 2018) to highlight the difficulties of gradient descent for saddle point optimization. Yet, ways to cope with this issue have been proposed decades ago in the context of mathematical programming. Simplifying further by setting the dimension to 1 and centering the equilibrium to the origin, Eq. 9 becomes:

$$\min_{\theta\in\mathbb{R}} \max_{\phi\in\mathbb{R}} \; \theta\cdot\phi \qquad \text{and} \qquad (\theta^*, \phi^*) = (0,0)\,. \tag{10}$$

The operator associated with this minimax game is $F(\theta,\phi) = (\phi, -\theta)$. There are several ways to compute the discrete updates of this dynamics. The two most common ones are the *simultaneous* and the *alternating* gradient update rules,

$$\text{Simultaneous update:} \quad \begin{cases} \theta_{t+1} = \theta_t - \eta\phi_t \\ \phi_{t+1} = \phi_t + \eta\theta_t \end{cases}, \quad \text{Alternating update:} \quad \begin{cases} \theta_{t+1} = \theta_t - \eta\phi_t \\ \phi_{t+1} = \phi_t + \eta\theta_{t+1} \end{cases}. \tag{11}$$

Interestingly, these two choices give rise to have a completely different behavior. The norm of the *simultaneous* updates diverges geometrically whereas the alternating iterates are bounded but do not converge to the equilibrium. As a consequence, their respective uniform average have a different behavior, as highlighted in the following proposition (more details and proof in §B.1):

Proposition 1. *The* simultaneous *iterates diverge geometrically and the* alternating *iterates defined in* (11) *are bounded but do not converge to 0 as*

$$\text{Simultaneous:} \;\; \theta_{t+1}^2 + \phi_{t+1}^2 = (1+\eta^2)(\theta_t^2+\phi_t^2)\,, \quad \text{Alternating:} \;\; \theta_t^2+\phi_t^2 = \Theta(\theta_0^2+\phi_0^2) \tag{12}$$

where $u_t = \Theta(v_t) \Leftrightarrow \exists \alpha, \beta > 0 : \alpha v_t \leq u_t \leq \beta v_t$.

The uniform average $(\bar{\theta}_t, \bar{\phi}_t) \stackrel{\text{def}}{=} \frac{1}{t}\sum_{s=0}^{t-1}(\theta_s, \phi_s)$ *of the* simultaneous *updates (resp. the* alternating updates*) diverges (resp. converges to 0) as,*

$$\text{Simultaneous:} \;\; \bar{\theta}_t^2 + \bar{\phi}_t^2 = \Theta\left(\frac{\theta_0^2+\phi_0^2}{\eta^2 t^2}(1+\eta^2)^t\right), \quad \text{Alternating:} \;\; \bar{\theta}_t^2 + \bar{\phi}_t^2 = \Theta\left(\frac{\theta_0^2+\phi_0^2}{\eta^2 t^2}\right). \tag{13}$$

This sublinear convergence result, proved in §B, underlines the benefits of averaging when the sequence of iterates is bounded (i.e. for *alternating* update rule). When the sequence of iterates is not bounded (i.e. for *simultaneous* updates) averaging fails to ensure convergence. This theorem also shows how *alternating* updates may have better convergence properties than *simultaneous* updates.

[4] Wasserstein GAN (WGAN) proposed by Arjovsky et al. (2017) boils down to the following minimax formulation: $\min_{\boldsymbol{\theta}\in\Theta}\max_{\boldsymbol{\varphi}\in\Phi, ||D_{\boldsymbol{\varphi}}||_L\leq 1} \mathbb{E}_{\mathbf{x}\sim p}[D_{\boldsymbol{\varphi}}(\mathbf{x})] - \mathbb{E}_{\mathbf{x}'\sim q_{\boldsymbol{\theta}}}[D_{\boldsymbol{\varphi}}(\mathbf{x}')]$.

3.2 EXTRAPOLATION

Another technique used in the variational inequality literature to prevent oscillations is *extrapolation*. This concept is anterior to the extragradient method since Korpelevich (1976) mentions that the idea of *extrapolated* "prices" to give "stability" had been already formulated by Polyak (1963, Chap. II). The idea behind this technique is to compute the gradient at an (extrapolated) point different from the current point from which the update is performed, stabilizing the dynamics:

$$\text{Compute extrapolated point:} \quad \boldsymbol{\omega}_{t+1/2} = P_\Omega[\boldsymbol{\omega}_t - \eta F(\boldsymbol{\omega}_t)]\,, \tag{14}$$

$$\text{Perform update step:} \quad \boldsymbol{\omega}_{t+1} = P_\Omega[\boldsymbol{\omega}_t - \eta F(\boldsymbol{\omega}_{t+1/2})]\,. \tag{15}$$

Note that, even in the *unconstrained case*, this method is intrinsically different from Nesterov's momentum[5] (Nesterov, 1983, Eq. 2.2.9) because of this lookahead step for the gradient computation:

$$\text{Nesterov's method:} \quad \boldsymbol{\omega}_{t+1/2} = \boldsymbol{\omega}_t - \eta F(\boldsymbol{\omega}_t)\,, \qquad \boldsymbol{\omega}_{t+1} = \boldsymbol{\omega}_{t+1/2} + \beta(\boldsymbol{\omega}_{t+1/2} - \boldsymbol{\omega}_t)\,. \tag{16}$$

Nesterov's method does not converge when trying to optimize (10). One intuition explaining why *extrapolation* provides better convergence properties than the standard gradient method comes from Euler's method framework (see for instance (Atkinson, 2003) for more details on that topic). Actually, if we consider a first order approximation of $\boldsymbol{\omega}_{t+1/2}$, we have $\boldsymbol{\omega}_{t+1/2} \approx \boldsymbol{\omega}_{t+1} + o(\eta)$ and consequently, the update step (15) is close to an *implicit method* step:

$$\text{Implicit step:} \quad \boldsymbol{\omega}_{t+1} = \boldsymbol{\omega}_t - \eta F(\boldsymbol{\omega}_{t+1})\,. \tag{17}$$

In the literature on Euler's method, *implicit methods* are known to be more stable and to benefit from better convergence properties (Atkinson, 2003) than *explicit methods*. They are not often used in practice though since they require to solve a potentially non-linear system at each iteration.

Taking back the simplified WGAN toy example (10) from §3.1 we get the following update rules,

$$\text{Implicit:} \quad \begin{cases} \theta_{t+1} = \theta_t - \eta\phi_{t+1} \\ \phi_{t+1} = \phi_t + \eta\theta_{t+1} \end{cases}, \qquad \text{Extrapolation:} \begin{cases} \theta_{t+1} = \theta_t - \eta(\phi_t + \eta\theta_t) \\ \phi_{t+1} = \phi_t + \eta(\theta_t - \eta\phi_t) \end{cases}. \tag{18}$$

In the following proposition, we will see that the respective convergence rates of the *implicit method* and *extrapolation* are highly similar. Keeping in mind that the latter has the major advantage of being more practical, this proposition clearly underlines the benefits of *extrapolation* (more details and proof in §B.2),

Proposition 2. *The squared norm of the iterates $N_t \overset{def}{=} \theta_t^2 + \phi_t^2$, where the update rule of θ_t and ϕ_t are defined in (18), decreases geometrically for any $\eta < 1$ as,*

$$\text{Implicit:} \quad N_{t+1} = \big(1 - \eta^2 + \eta^4 + \mathcal{O}(\eta^6)\big)N_t\,, \qquad \text{Extrapolation:} \quad N_{t+1} = (1 - \eta^2 + \eta^4)N_t\,. \tag{19}$$

3.3 EXTRAPOLATION FROM THE PAST

One issue with extrapolation is that the algorithm "wastes" a gradient (14). Indeed we need to compute the gradient at two different positions for every single update of the parameters. We thus propose a new technique that we call *extrapolation from the past* which only requires to compute one gradient for every update. The idea of this technique is to store and re-use the previous extrapolated gradient to compute the new extrapolation point:

$$\text{Extrapolation from the past:} \quad \boldsymbol{\omega}_{t+1/2} = P_\Omega[\boldsymbol{\omega}_t - \eta F(\boldsymbol{\omega}_{t-1/2})] \tag{20}$$

$$\text{Perform update step:} \quad \boldsymbol{\omega}_{t+1} = P_\Omega[\boldsymbol{\omega}_t - \eta F(\boldsymbol{\omega}_{t+1/2})] \;\text{ and store: }\; F(\boldsymbol{\omega}_{t+1/2}) \tag{21}$$

This update scheme can be related to (Chiang et al., 2012, Alg. 1) in the context of online learning and to the *optimistic mirror descent* (Rakhlin and Sridharan, 2013; Daskalakis et al., 2018). In the *unconstrained* case, (20) and (21) reduce to:

$$\text{Optimistic mirror descent (OMD):} \quad \boldsymbol{\omega}_{t+1/2} = \boldsymbol{\omega}_{t-1/2} - 2\eta F(\boldsymbol{\omega}_{t-1/2}) + \eta F(\boldsymbol{\omega}_{t-3/2}) \tag{22}$$

However our technique comes from a different perspective, it was motivated by VIP and inspired from the extragradient method. Furthermore our technique extends to *constrained* optimization as shown in (20) and (21). It is not clear whether or not a single projection added to (22) provides a provably converging algorithm. Using the VIP point of view we are able to prove a linear convergence rate for a *projected version of the extrapolation from the past* (see details and proof of Theorem 1 in §B.3). We also extend these results to the stochastic operator setting in §4.

[5] Sutskever (2013, §7.2) showed the equivalence between "standard momentum" and Nesterov's formulation.

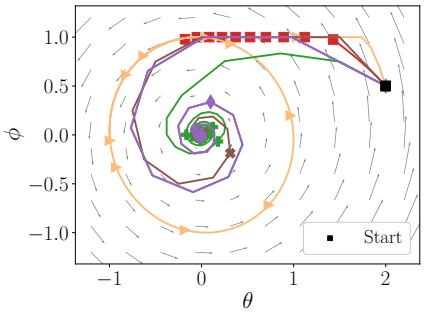


Figure 1: Comparison of the basic gradient method (as well as Adam) with the techniques presented in §3 on the optimization of (9). Only the algorithms advocated in this paper (Averaging, Extrapolation and Extrapolation from the past) converge quickly to the solution. Each marker represents 20 iterations. We compare these algorithms on a non-convex objective in §G.1.

Algorithm 1 AvgSGD

Let $\boldsymbol{\omega}_0 \in \Omega$
for $t = 0 \dots T-1$ **do**
  $\xi_t \sim P$ *(minibatch)*
  $\boldsymbol{d}_t \leftarrow F(\boldsymbol{\omega}_t, \xi_t)$
  $\boldsymbol{\omega}_{t+1} \leftarrow P_\Omega[\boldsymbol{\omega}_t - \eta_t \boldsymbol{d}_t]$
end for
Return $\bar{\boldsymbol{\omega}}_T \leftarrow \frac{\sum_{t=0}^{T-1} \eta_t \boldsymbol{\omega}_t}{\sum_{t=0}^{T-1} \eta_t}$

Algorithm 2 AvgExtraSGD

for $t = 0 \dots T-1$ **do**
  $\xi_t, \xi'_t \sim P$ *(minibatches)*
  $\boldsymbol{d}_t \leftarrow F(\boldsymbol{\omega}_t, \xi_t)$
  $\boldsymbol{\omega}'_t \leftarrow P_\Omega[\boldsymbol{\omega}_t - \eta_t \boldsymbol{d}_t]$
  $\boldsymbol{d}'_t \leftarrow F(\boldsymbol{\omega}'_t, \xi'_t)$
  $\boldsymbol{\omega}_{t+1} \leftarrow P_\Omega[\boldsymbol{\omega}_t - \eta_t \boldsymbol{d}'_t]$
end for
Return $\bar{\boldsymbol{\omega}}_T \leftarrow \frac{\sum_{t=0}^{T-1} \eta_t \boldsymbol{\omega}'_t}{\sum_{t=0}^{T-1} \eta_t}$

Algorithm 3 AvgPastExtraSGD

Let $\boldsymbol{\omega}_0 \in \Omega$
for $t = 0 \dots T-1$ **do**
  $\xi_t \sim P$ *(minibatch)*
  $\boldsymbol{\omega}'_t \leftarrow P_\Omega[\boldsymbol{\omega}_t - \eta_t \boldsymbol{d}_{t-1}]$
  $\boldsymbol{d}_t \leftarrow F(\boldsymbol{\omega}'_t, \xi_t)$
  $\boldsymbol{\omega}_{t+1} \leftarrow P_\Omega[\boldsymbol{\omega}_t - \eta_t \boldsymbol{d}_t]$
end for
Return $\bar{\boldsymbol{\omega}}_T \leftarrow \frac{\sum_{t=0}^{T-1} \eta_t \boldsymbol{\omega}'_t}{\sum_{t=0}^{T-1} \eta_t}$

Figure 2: Three variants of SGD using the techniques introduced in §3.

Theorem 1 (Linear convergence of *extrapolation from the past*)**.** *If F is μ-strongly monotone (see §A for the definition of strong monotonicity) and L-Lipschitz, then the updates (20) and (21) with $\eta = \frac{1}{4L}$ provide linearly converging iterates,*

$$\|\boldsymbol{\omega}_t - \boldsymbol{\omega}^*\|_2^2 \leq \left(1 - \frac{\mu}{4L}\right)^t \|\boldsymbol{\omega}_0 - \boldsymbol{\omega}^*\|_2^2 , \quad \forall t \geq 0 . \tag{23}$$

In comparison to the results from (Daskalakis et al., 2018) that hold only for a bilinear objective, we provide a faster convergence rate (linear vs sublinear) on the last iterate for a general (strongly monotone) operator F and any projection on a convex Ω. One thing to notice is that the operator of a bilinear objective is *not* strongly monotone, but in that case one can use the standard extrapolation method (14) which converges linearly for a (constrained or not) bilinear game (Tseng, 1995, Cor. 3.3).

4 Optimization of VIP with stochastic gradients

In this section, we consider extensions of the techniques presented in section §3 for optimizing (VIP), to the context of a *stochastic* operator. In this case, at each time step we no longer have access to the exact gradient $F(\boldsymbol{\omega})$ but to an unbiased *stochastic* estimate of it $F(\boldsymbol{\omega}, \xi)$ where $\xi \sim P$ and $F(\boldsymbol{\omega}) := \mathbb{E}_{\xi \sim P}[F(\boldsymbol{\omega}, \xi)]$. This is motivated from the GAN formulation where we only have access to a finite sample estimate of the expected gradient, computed on a mini-batch. For GANs, ξ is thus a mini-batch of points coming from the true data distribution p and the generator distribution $q_{\boldsymbol{\theta}}$.

For our analysis, we require at least one of the two following assumptions on the stochastic operator:

Assumption 1. *Bounded variance by σ^2:* $\mathbb{E}_\xi[\|F(\boldsymbol{\omega}) - F(\boldsymbol{\omega}, \xi)\|^2] \leq \sigma^2 , \quad \forall \boldsymbol{\omega} \in \Omega .$

Assumption 2. *Bounded expected squared norm by M^2:* $\mathbb{E}_\xi[\|F(\boldsymbol{\omega}, \xi)\|^2] \leq M^2, \ \forall \boldsymbol{\omega} \in \Omega.$

Assump. 1 is standard in stochastic variational analysis, while Assump. 2 is a stronger assumption sometimes made in stochastic convex optimization. To illustrate how strong Assump. 2 is, note that it does not hold for an unconstrained bilinear objective like in our example 10 in §3. It is thus mainly reasonable for bounded constraint sets. Note that in practice we have $\sigma \ll M$.

We now present and analyse three algorithms, variants of SGD that are appropriate to solve (VIP). The first one Alg. 1 (AvgSGD) is the stochastic extension of the gradient method for solving (VIP); Alg. 2 (AvgExtraSGD) uses *extrapolation* and Alg. 3 (AvgPastExtraSGD) uses *extrapolation from the past*. A fourth variant (Alg.5) is proposed in §D. These three algorithms return an *average* of the iterates. The proofs of the theorems presented in this section are in §F.

To handle constraints such as parameter clipping (Arjovsky et al., 2017), we present a *projected* version of theses algorithms, where $P_\Omega[\boldsymbol{\omega}']$ denotes the projection of $\boldsymbol{\omega}'$ onto Ω (see §A). Note that when $\Omega = \mathbb{R}^d$, the projection is the identity mapping (unconstrained setting). In order to prove the convergence of these three algorithms we will assume that F is monotone:

$$(F(\boldsymbol{\omega}) - F(\boldsymbol{\omega}'))^\top (\boldsymbol{\omega} - \boldsymbol{\omega}') \geq 0 \quad \forall \boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega\,. \tag{24}$$

If F can be written as (6), it implies that the cost functions are convex[6]. Note however that in general GANs parametrized with neural networks lead to non-monotone VIPs.

Assumption 3. *F is* monotone *and Ω is a compact convex set, such that* $\max_{\boldsymbol{\omega},\boldsymbol{\omega}'\in\Omega} \|\boldsymbol{\omega}-\boldsymbol{\omega}'\|^2 \leq R^2$.

In that setting the quantity $g(\boldsymbol{\omega}^*) := \max_{\boldsymbol{\omega}\in\Omega} F(\boldsymbol{\omega})^\top(\boldsymbol{\omega}^* - \boldsymbol{\omega})$ is well defined and is equal to 0 if and only if $\boldsymbol{\omega}^*$ is a solution of (VIP). Moreover, if we are optimizing a *zero-sum game*, we have $\boldsymbol{\omega} = (\boldsymbol{\theta}, \boldsymbol{\varphi})$, $\Omega = \Theta \times \Phi$ and $F(\boldsymbol{\theta}, \boldsymbol{\varphi}) = [\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}) \;\; -\nabla_{\boldsymbol{\varphi}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi})]^\top$. Hence, the quantity $h(\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*) := \max_{\boldsymbol{\varphi}\in\Phi} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\varphi}) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}^*)$ is well defined and equal to 0 if and only if $(\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*)$ is a *Nash equilibrium* of the game. The two functions g and h are called *merit functions* (more details on the concept of *merit functions* in §C). In the following, we call,

$$\mathrm{Err}(\boldsymbol{\omega}) \stackrel{\text{def}}{=} \begin{cases} \max\limits_{\boldsymbol{\theta}',\boldsymbol{\varphi}'\in\Theta\times\Phi} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}') - \mathcal{L}(\boldsymbol{\theta}', \boldsymbol{\varphi}) & \text{if } \; F(\boldsymbol{\theta}, \boldsymbol{\varphi}) = [\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}) \;\; -\nabla_{\boldsymbol{\varphi}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi})]^\top \\ \max\limits_{\boldsymbol{\omega}'\in\Omega} F(\boldsymbol{\omega}')^\top(\boldsymbol{\omega} - \boldsymbol{\omega}') & \text{otherwise.} \end{cases} \tag{25}$$

Averaging. Alg. 1 (AvgSGD) presents the stochastic gradient method with *averaging*, which reduces to the standard (simultaneous) SGD updates for the two-player games used in the GAN literature, but returning an *average* of the iterates.

Theorem 2. *Under Assump. 1, 2 and 3, SGD with averaging (Alg. 1) with a constant step-size gives,*

$$\mathbb{E}[\mathrm{Err}(\bar{\boldsymbol{\omega}}_T)] \leq \frac{R^2}{2\eta T} + \eta\frac{M^2 + \sigma^2}{2} \quad \textit{where} \quad \bar{\boldsymbol{\omega}}_T \stackrel{\textit{def}}{=} \frac{1}{T}\sum_{t=0}^{T-1} \boldsymbol{\omega}_t\,, \quad \forall T \geq 1\,. \tag{26}$$

Thm. 2 uses a similar proof as (Nemirovski et al., 2009). The constant term $\eta(M^2 + \sigma^2)/2$ in (26) is called the *variance term*. This type of bound is standard in stochastic optimization. We also provide in §F a similar rate with an extra log factor when $\eta_t = \frac{\eta}{\sqrt{t}}$. We show that this variance term is smaller than the one of *SGD with prediction method* (Yadav et al., 2018) in §E.

Extrapolations. Alg. 2 (AvgExtraSGD) adds an extrapolation step compared to Alg. 1 in order to reduce the oscillations due to the game between the two players. A theoretical consequence is that it has a smaller variance term than (26). As discussed previously, Assump. 2 made in Thm. 2 for the convergence of Alg. 1 is very strong in the unbounded setting. One advantage of SGD with *extrapolation* is that Thm. 3 does not require this assumption.

Theorem 3. *(Juditsky et al., 2011, Thm. 1) Under Assump. 1 and 3, if $\mathbb{E}_\xi[F]$ is L-Lipschitz, then SGD with* extrapolation and averaging *(Alg. 2) using a constant step-size $\eta \leq \frac{1}{\sqrt{3}L}$ gives,*

$$\mathbb{E}[\mathrm{Err}(\bar{\boldsymbol{\omega}}_T)] \leq \frac{R^2}{\eta T} + \frac{7}{2}\eta\sigma^2 \quad \textit{where} \quad \bar{\boldsymbol{\omega}}_T \stackrel{\textit{def}}{=} \frac{1}{T}\sum_{t=0}^{T-1} \boldsymbol{\omega}'_t\,, \quad \forall T \geq 1\,. \tag{27}$$

Since in practice $\sigma \ll M$, the variance term in (27) is significantly smaller than the one in (26). To summarize, SGD with *extrapolation* provides better convergence guarantees but requires two gradient computations and samples per iteration. This motivates our new method, Alg. 3 (AvgPastExtraSGD) which uses *extrapolation from the past* and achieves *the best of both worlds*.

Theorem 4. *Under Assump. 1 and 3, if $\mathbb{E}_\xi[F]$ is L-Lipschitz then SGD with* extrapolation from the past *using a constant step-size $\eta \leq \frac{1}{2\sqrt{3}L}$, gives that the averaged iterates converge as,*

$$\mathbb{E}[\mathrm{Err}(\bar{\boldsymbol{\omega}}_T)] \leq \frac{R^2}{\eta T} + \frac{13}{2}\eta\sigma^2 \quad \textit{where} \quad \bar{\boldsymbol{\omega}}_T \stackrel{\textit{def}}{=} \frac{1}{T}\sum_{t=0}^{T-1} \boldsymbol{\omega}'_t \quad \forall T \geq 1\,. \tag{28}$$

The bounds is similar to the one provided in Thm. 3 but each iteration of Alg. 3 is computationally half the cost of an iteration of Alg. 2.

[6]The convexity of the cost functions in (3) is a necessary condition (not sufficient) for the operator to be monotone. In the context of a zero-sum game, the convexity of the cost functions is a sufficient condition.

5 COMBINING THE TECHNIQUES WITH ESTABLISHED ALGORITHMS

In the previous sections we presented several techniques that converge on a simple bilinear example. These techniques can be combined in practice with existing algorithms. We propose to combine them to two standard algorithms used for training deep neural networks: the Adam optimizer (Kingma and Ba, 2015) and the SGD optimizer (Robbins and Monro, 1951). Note that in the case of a two-player game (3), the previous results can be generalized to gradient updates with a different step-size for each player by simply rescaling the objectives $\mathcal{L}^{(\boldsymbol{\theta})}$ and $\mathcal{L}^{(\boldsymbol{\varphi})}$ by a different scaling factor. A detailed pseudo-code for Adam with extrapolation step (Extra-Adam) is given in Algorithm 4.

Algorithm 4 Extra-Adam: proposed Adam with extrapolation step.

input: step-size η, decay rates for moment estimates β_1, β_2, access to the stochastic gradients $\nabla \ell_t(\cdot)$ and to the projection $P_\Omega[\cdot]$ onto the constraint set Ω, initial parameter $\boldsymbol{\omega}_0$, averaging scheme $(\rho_t)_{t\geq 1}$
for $t = 0 \dots T-1$ **do**
  Option 1: Standard extrapolation.
    Sample new minibatch and compute stochastic gradient: $g_t \leftarrow \nabla \ell_t(\boldsymbol{\omega}_t)$
  Option 2: Extrapolation from the past
    Load previously saved stochastic gradient: $g_t = \nabla \ell_{t-1/2}(\boldsymbol{\omega}_{t-1/2})$
  Update estimate of first moment for extrapolation: $m_{t-1/2} \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$
  Update estimate of second moment for extrapolation: $v_{t-1/2} \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^2$
  Correct the bias for the moments: $\hat{m}_{t-1/2} \leftarrow m_{t-1/2}/(1-\beta_1^{2t-1})$, $\hat{v}_{t-1/2} \leftarrow v_{t-1/2}/(1-\beta_2^{2t-1})$
  Perform *extrapolation* step from iterate at time t: $\boldsymbol{\omega}_{t-1/2} \leftarrow P_\Omega[\boldsymbol{\omega}_t - \eta \frac{\hat{m}_{t-1/2}}{\sqrt{\hat{v}_{t-1/2}}+\epsilon}]$
  Sample new minibatch and compute stochastic gradient: $g_{t+1/2} \leftarrow \nabla \ell_{t+1/2}(\boldsymbol{\omega}_{t+1/2})$
  Update estimate of first moment: $m_t \leftarrow \beta_1 m_{t-1/2} + (1-\beta_1) g_{t+1/2}$
  Update estimate of second moment: $v_t \leftarrow \beta_2 v_{t-1/2} + (1-\beta_2) g_{t+1/2}^2$
  Compute bias corrected for first and second moment: $\hat{m}_t \leftarrow m_t/(1-\beta_1^{2t})$, $\hat{v}_t \leftarrow v_t/(1-\beta_2^{2t})$
  Perform *update* step from the iterate at time t: $\boldsymbol{\omega}_{t+1} \leftarrow P_\Omega[\boldsymbol{\omega}_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon}]$
end for
Output: $\boldsymbol{\omega}_{T-1/2}$, $\boldsymbol{\omega}_T$ or $\bar{\boldsymbol{\omega}}_T = \sum_{t=0}^{T-1} \rho_{t+1} \boldsymbol{\omega}_{t+1/2} / \sum_{t=0}^{T-1} \rho_{t+1}$ (see (8) for online averaging)

6 RELATED WORK

The extragradient method is the standard algorithm to optimize variational inequalities. This algorithm has been originally introduced by Korpelevich (1976) and extended by Nesterov (2007) and Nemirovski (2004). Stochastic versions of the extragradient have been recently analyzed (Juditsky et al., 2011; Yousefian et al., 2014; Iusem et al., 2017) for stochastic variational inequalities with *bounded constraints*. A linearly convergent variance reduced version of the stochastic gradient method has been proposed by Palaniappan and Bach (2016) for strongly monotone variational inequalities.

Several methods to stabilize GANs consist in transforming a zero-sum formulation into a more general game that can no longer be cast as a saddle point problem. This is the case of the *non-saturating* formulation of GANs (Goodfellow et al., 2014; Fedus et al., 2018), the DCGANs (Radford et al., 2016), the *gradient penalty*[7] for WGANs (Gulrajani et al., 2017). Yadav et al. (2018) propose an optimization method for GANs based on AltSGD using a momentum based step on the generator. Daskalakis et al. (2018) proposed a method inspired from game theory. Li et al. (2017) suggest to dualize the GAN objective to reformulate it as a maximization problem and Mescheder et al. (2017) propose to add the norm of the gradient in the objective and provide an interesting perspective on GANs, interpreting the training as the search of a two-player game equilibrium. A study of the continuous version of two player games has been conducted by Ratliff et al. (2016). Interesting non-convex results were proved, for a new notion of regret minimization, by Hazan et al. (2017) and in the context of GANs by Grnarova et al. (2018).

The technique of *unrolling* steps proposed by Metz et al. (2017) can be confused with extrapolation but is actually fundamentally different: the perspective is try to construct the "true generator objective function" unrolling for K steps the updates of the generator and then update the discriminator.

[7]The gradient penalty is only added to the discriminator cost function. Since this gradient penalty depends also on the generator, WGAN-GP cannot be cast as a SP problem and is actually a non-zero sum game.

Nevertheless the fact that this "true generator function" may not be found with a satisfying accuracy may lead to a different behavior than the one expected.

Regarding the averaging technique, some recent work appear to have already successfully used *geometric averaging* (7) for GANs in practice, but only briefly mention it (Karras et al., 2018; Mescheder et al., 2018). By contrast the present work formally motivates and justifies the use of averaging for GANs by relating them to the VIP perspective, and sheds light on its underlying intuitions in §3.1. Another independent work (Yazıcı et al., 2018) made a similar attempt but in the context of regret minimization in games. Mertikopoulos et al. (2018) also independently explored extrapolation providing asymptotic convergence results (i.e. without any rate of convergence) in the context of *coherent saddle point*. The coherence assumption is slightly weaker than monotonicity.

7 Experiments

Our goal in this experimental section is not to provide new state-of-the art results with architectural improvements or a new GAN formulation but to show that using the *techniques* (with theoretical guarantees in the monotone case) that we introduced earlier allow us to optimize standard GANs in a better way. These techniques, which are orthogonal to the design of new formulations of GAN optimization objectives, and to architectural choices, can potentially be used for the training of any type of GAN. We will compare the following optimization algorithms: baselines are SGD and Adam using either simultaneous updates on the generator and on the discriminator (denoted **SimAdam** and **SimSGD**) or k updates on the discriminator alternating with 1 update on the generator (denoted **AltSGD**$\{k\}$ and **AltAdam**$\{k\}$)[8]. Variants that use *extrapolation* are denoted **ExtraSGD** (Alg. 2) and **ExtraAdam** (Alg. 4). Variants using *extrapolation from the past* are **PastExtraSGD** (Alg. 3) and **PastExtraAdam** (Alg. 4). We also present results using as output the *averaged* iterates, adding **Avg** as a prefix of the algorithm name when we use (uniform) *averaging*.

7.1 Bilinear saddle point (stochastic)

We evaluate the performance of the various stochastic algorithms first on a simple ($n = 10^3, d = 10^3$) finite sum bilinear objective (a monotone operator) constrained to $[-1, 1]^d$:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\theta}^\top \boldsymbol{M}^{(i)}\boldsymbol{\varphi} + \boldsymbol{\theta}^\top \boldsymbol{a}^{(i)} + \boldsymbol{\varphi}^\top \boldsymbol{b}^{(i)}\right) \qquad (29)$$

$$\text{solved by } (\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*) \text{ s.t. } \begin{cases} \bar{\boldsymbol{M}}\boldsymbol{\varphi}^* = -\bar{\boldsymbol{a}} \\ \bar{\boldsymbol{M}}^T\boldsymbol{\theta}^* = -\bar{\boldsymbol{b}} \end{cases},$$

where $\bar{\boldsymbol{a}} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{a}^{(i)}$, $\bar{\boldsymbol{b}} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{b}^{(i)}$ and $\bar{\boldsymbol{M}} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{M}^{(i)}$. Marices and vectors $\boldsymbol{M}^{(i)}_{kj}$, $\boldsymbol{a}^{(i)}_k$, $\boldsymbol{b}^{(i)}_k$; $1 \leq i \leq n$, $1 \leq j, k \leq d$ were randomly generated, but ensuring that $(\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*)$ would belong to $[-1, 1]^d$. Results are shown in Fig. 3. We can see that AvgSGD and AvgPastExtraSGD perform the best on this task.

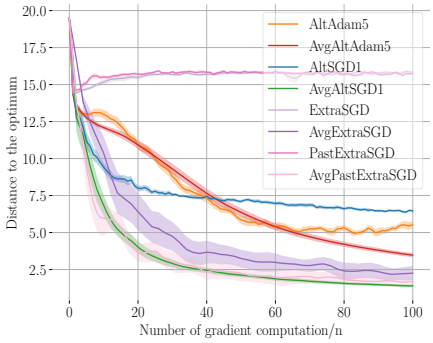


Figure 3: Performance of the considered stochastic optimization algorithms on the bilinear problem (29). Each method uses its respective optimal step-size found by grid-search.

7.2 WGAN and WGAN-GP on CIFAR10

We now evaluate the proposed techniques in the context of GAN training, which is a challenging stochastic optimization problem where the objectives of both players are non-convex. We propose to evaluate the Adam variants of the different optimization algorithms (see Alg. 4 for Adam with *extrapolation*) by training two different architectures on the CIFAR10 dataset (Krizhevsky and Hinton, 2009). First we consider a constrained zero-sum game by training the DCGAN architecture (Radford et al., 2016) with the WGAN objective and weight clipping as proposed in Arjovsky et al. (2017). Then we compared the different methods on a state of the art architecture, by training a ResNet with

[8]In the original WGAN (Arjovsky et al., 2017) paper the authors use $k = 5$.

Model	WGAN (DCGAN)			WGAN-GP (ResNet)	
Method	no averaging	uniform avg	EMA	no averaging	uniform avg
SimAdam	*6.05 ± .12*	5.85 ± .16	6.08 ± .10	*7.54 ± .21*	7.74 ± .27
AltAdam5	*5.45 ± .08*	5.72 ± .06	5.49 ± .05	*7.20 ± .06*	7.67 ± .15
ExtraAdam	**6.38 ± .09**	**6.38 ± .20**	**6.37 ± .08**	7.79 ± .09	**8.26 ± .12**
PastExtraAdam	5.98 ± .15	6.07 ± .19	6.01 ± .11	7.71 ± .12	7.84 ± .18
OptimAdam	*5.74 ± .10*	5.80 ± .08	5.78 ± .05	*7.80 ± .07*	**7.99 ± .12**

Table 1: Best inception scores (averaged over 5 runs) achieved on CIFAR10 for every considered Adam variant. OptimAdam is the related *Optimistic Adam* (Daskalakis et al., 2018) algorithm. EMA denotes *exponential moving average* (with $\beta = 0.999$, see Eq. 8). We see that the techniques of extrapolation and averaging consistently enable improvements over the baselines (in italic).

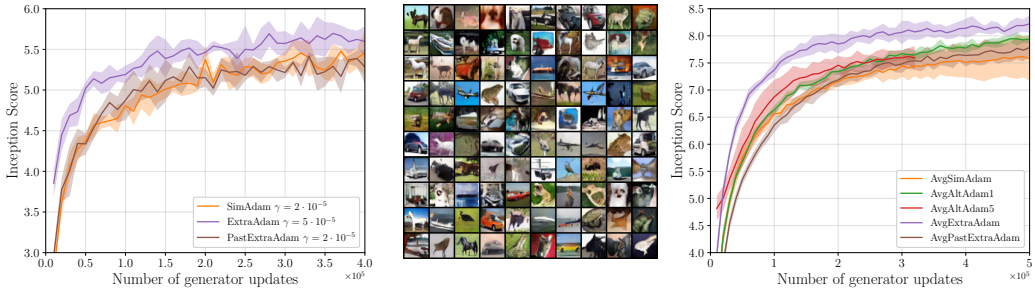


Figure 4: **Left:** Mean and standard deviation of the inception score computed over 5 runs for each method on WGAN trained on CIFAR10. To keep the graph readable we show only SimAdam but AltAdam performs similarly. **Middle:** Samples from a ResNet generator trained with the WGAN-GP objective using AvgExtraAdam. **Right:** WGAN-GP trained on CIFAR10: mean and standard deviation of the inception score computed over 5 runs for each method using the best performing learning rate plotted over wall-clock time; all experiments were run on a NVIDIA Quadro GP100 GPU. We see that ExtraAdam converges faster than the Adam baselines.

the WGAN-GP objective similar to Gulrajani et al. (2017). Models are evaluated using the inception score (Salimans et al., 2016) computed on 10,000 samples.

For each algorithm we did an extensive search over the hyperparameters of Adam, we fixed $\beta_1 = 0.5$ and $\beta_2 = 0.9$ for all methods as they seemed to perform well. We want to emphasize that as proposed by Heusel et al. (2017) it is quite important to set different learning rates for the generator and the discriminator. Experiments were run with 5 random seeds for 500,000 updates of the generator.

Table 1 reports the best inception score achieved on this problem by each considered method. We see that the techniques of *extrapolation* and *averaging* consistently enable improvements over the baselines (see §G.4 for more experiments on *averaging*). Fig. 4 shows training curves for each method (for their best performing learning rate), as well as samples from a ResNet generator trained with ExtraAdam on a WGAN-GP objective. For both tasks, using an *extrapolation step* and averaging with Adam (ExtraAdam) outperformed all other methods. ExtraAdam with averaging is able to reach state of the art results on CIFAR10 similar to Miyato et al. (2018). We also observed that methods based on *extrapolation* are less sensitive to the choice of learning rate and can be used with higher learning rates with less degradation; see App. §G.3 for more details.

8 Conclusion

We newly addressed GAN objectives in the framework of variational inequality. We tapped into the optimization literature to provide more principled sound techniques to optimize such games. We leveraged these techniques to develop practical optimization algorithms suitable for a wide range of GAN training objectives (including non-zero sum games and projections onto constraints). We experimentally verified that this could yield better trained models, achieving to our knowledge the best inception score when optimizing a WGAN objective on the reference unmodified DCGAN architecture (Radford et al., 2016). The presented techniques address a fundamental problem in GAN training in a principled way, and are orthogonal to the design of new GAN architectures and objectives. They are thus likely to be widely applicable, and benefit future development of GANs.

ACKNOWLEDGMENTS.

This research was partially supported by the Canada Excellence Research Chair in "Data Science for Realtime Decision-making" and by the NSERC Discovery Grant RGPIN-2017-06936. Gauthier Gidel would like to acknowledge Benoît Joly and Florestan Martin-Baillon for bringing a fresh point of view on the proof of Proposition 1.

REFERENCES

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

K. E. Atkinson. *An introduction to numerical analysis*. John Wiley & Sons, 2003.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

R. E. Bruck. On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in hilbert space. *Journal of Mathematical Analysis and Applications*, 1977.

G. H. Chen and R. T. Rockafellar. Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 1997.

C.-K. Chiang, T. Yang, C.-J. Lee, M. Mahdavi, C.-J. Lu, R. Jin, and S. Zhu. Online optimization with gradual variations. In *COLT*, 2012.

G. P. Crespi, A. Guerraggio, and M. Rocca. Minty variational inequality and optimization: Scalar and vector case. In *Generalized Convexity, Generalized Monotonicity and Applications*, 2005.

C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism. In *ICLR*, 2018.

W. Fedus, M. Rosca, B. Lakshminarayanan, A. M. Dai, S. Mohamed, and I. Goodfellow. Many paths to equilibrium: GANs do not need to decrease a divergence at every step. In *ICLR*, 2018.

G. Gidel, T. Jebara, and S. Lacoste-Julien. Frank-wolfe algorithms for saddle point problems. In *AISTATS*, 2017.

I. Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv:1701.00160*, 2016.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

P. Grnarova, K. Y. Levy, A. Lucchi, T. Hofmann, and A. Krause. An online learning approach to generative adversarial networks. In *ICLR*, 2018.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein GANs. In *NIPS*, 2017.

P. T. Harker and J.-S. Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming*, 1990.

E. Hazan, K. Singh, and C. Zhang. Efficient regret minimization in non-convex games. In *ICML*, 2017.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NIPS*, 2017.

A. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 2017.

A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 2011.

T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

G. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12, 1976.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, Canada, 2009.

T. Larsson and M. Patriksson. A class of gap functions for variational inequalities. *Math. Program.*, 1994.

C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

Y. Li, A. Schwing, K.-C. Wang, and R. Zemel. Dualing GANs. In *NIPS*, 2017.

P. Mertikopoulos, H. Zenati, B. Lecouat, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. Mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv*, 2018.

L. Mescheder, S. Nowozin, and A. Geiger. The numerics of GANs. In *NIPS*, 2017.

L. Mescheder, A. Geiger, and S. Nowozin. Which training methods for GANs do actually converge? In *ICML*, 2018.

L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein. Unrolled generative adversarial networks. In *ICLR*, 2017.

T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.

V. Nagarajan and J. Z. Kolter. Gradient descent GAN optimization is locally stable. In *NIPS*, 2017.

A. Nedić and A. Ozdaglar. Subgradient methods for saddle-point problems. *J Optim Theory Appl*, 2009.

A. Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optim.*, 2004.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 2009.

Y. Nesterov. *Introductory Lectures On Convex Optimization.* Springer, 1983.

Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Math. Program.*, 2007.

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *NIPS*, 2016.

B. Palaniappan and F. Bach. Stochastic variance reduction methods for saddle-point problems. In *NIPS*, 2016.

B. T. Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 1963.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *COLT*, 2013.

L. J. Ratliff, S. A. Burden, and S. S. Sastry. On the characterization of local nash equilibria in continuous games. 2016.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 1951.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *NIPS*, 2016.

I. Sutskever. *Training recurrent neural networks*. PhD thesis, 2013.

P. Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 1995.

J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior.* Princeton University Press, 1944.

A. Yadav, S. Shah, Z. Xu, D. Jacobs, and T. Goldstein. Stabilizing adversarial nets with prediction methods. In *ICLR*, 2018.

Y. Yazıcı, C.-S. Foo, S. Winkler, K.-H. Yap, G. Piliouras, and V. Chandrasekhar. The unusual effectiveness of averaging in gan training. *arXiv preprint arXiv:1806.04498*, 2018.

F. Yousefian, A. Nedić, and U. V. Shanbhag. Optimal robust smoothing extragradient algorithms for stochastic variational inequality problems. In *CDC*. IEEE, 2014.

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

A DEFINITIONS

In this section we recall usual definitions and lemmas from convex analysis. We start with the definitions and lemmas regarding the projection mapping.

A.1 PROJECTION MAPPING

Definition 1. *The projection P_Ω onto Ω is defined as,*

$$P_\Omega(\boldsymbol{\omega}') \in \arg\min_{\boldsymbol{\omega}' \in \Omega} \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2^2 \,. \tag{30}$$

When Ω is a convex set this projection is unique. This is a consequence of the following lemma that we will use in the following sections: the *non-expansiveness* of the projection onto a convex set.

Lemma 1. *Let Ω a convex set, the projection mapping $P_\Omega : \mathbb{R}^d \to \Omega$ is nonexpansive, i.e.,*

$$\|P_\Omega(\boldsymbol{\omega}) - P_\Omega(\boldsymbol{\omega}')\|_2 \leq \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2 \,, \quad \forall \boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega \,. \tag{31}$$

This is standard convex analysis result which can be found for instance in (Boyd and Vandenberghe, 2004). The following Lemma is also standard in convex analysis and its proof uses similar arguments as the proof of Lemma 1.

Lemma 2. *Let $\boldsymbol{\omega} \in \Omega$ and $\boldsymbol{\omega}^+ \stackrel{def}{=} P_\Omega(\boldsymbol{\omega} + \boldsymbol{u})$ then for all $\boldsymbol{\omega}' \in \Omega$ we have,*

$$\|\boldsymbol{\omega}^+ - \boldsymbol{\omega}'\|_2^2 \leq \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2^2 + 2\boldsymbol{u}^\top(\boldsymbol{\omega}^+ - \boldsymbol{\omega}') - \|\boldsymbol{\omega}^+ - \boldsymbol{\omega}\|_2^2 \,. \tag{32}$$

Proof of Lemma 2. We start by simply developing,

$$\begin{aligned}\|\boldsymbol{\omega}^+ - \boldsymbol{\omega}'\|_2^2 = \|(\boldsymbol{\omega}^+ - \boldsymbol{\omega}) + (\boldsymbol{\omega} - \boldsymbol{\omega}')\|_2^2 &= \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2^2 + 2(\boldsymbol{\omega}^+ - \boldsymbol{\omega})^\top(\boldsymbol{\omega} - \boldsymbol{\omega}') + \|\boldsymbol{\omega}^+ - \boldsymbol{\omega}\|_2^2 \\ &= \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2^2 + 2(\boldsymbol{\omega}^+ - \boldsymbol{\omega})^\top(\boldsymbol{\omega}^+ - \boldsymbol{\omega}') - \|\boldsymbol{\omega}^+ - \boldsymbol{\omega}\|_2^2 \,.\end{aligned}$$

Then since $\boldsymbol{\omega}^+$ is the projection onto the convex set Ω of $\boldsymbol{\omega} + \boldsymbol{u}$ we have that $(\boldsymbol{\omega}^+ - (\boldsymbol{\omega} + \boldsymbol{u}))^\top(\boldsymbol{\omega}^+ - \boldsymbol{\omega}') \leq 0 \,,\ \forall\, \boldsymbol{\omega}' \in \Omega$, leading to the result of the Lemma. □

A.2 SMOOTHNESS AND MONOTONICITY OF THE OPERATOR

Another important property used is the Lipschitzness of an operator.

Definition 2. *A mapping $F : \mathbb{R}^p \to \mathbb{R}^d$ is said to be L-Lipschitz if,*

$$\|F(\boldsymbol{\omega}) - F(\boldsymbol{\omega}')\|_2 \leq L\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2 \,, \quad \forall \boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega \,. \tag{33}$$

Definition 3. *A differentiable function $f : \Omega \to \mathbb{R}$ is said to be μ-*strongly convex *if*

$$f(\boldsymbol{\omega}) \geq f(\boldsymbol{\omega}') + \nabla f(\boldsymbol{\omega}')^\top(\boldsymbol{\omega} - \boldsymbol{\omega}') + \frac{\mu}{2}\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2^2 \quad \forall \boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega \,. \tag{34}$$

Definition 4. *A function $(\boldsymbol{\theta}, \boldsymbol{\varphi}) \mapsto \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi})$ is said convex-concave if $\mathcal{L}(\cdot, \boldsymbol{\varphi})$ is convex for all $\boldsymbol{\varphi} \in \Phi$ and $\mathcal{L}(\boldsymbol{\theta}, \cdot)$ is concave for all $\boldsymbol{\theta} \in \Theta$. An $\mathcal{L}$ is said to be μ-strongly convex concave if $(\boldsymbol{\theta}, \boldsymbol{\varphi}) \mapsto \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}) - \frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2 + \frac{\mu}{2}\|\boldsymbol{\varphi}\|_2^2$ is convex concave.*

Definition 5. *For $\mu > 0$, an operator $F : \Omega \to \mathbb{R}^d$ is said to be μ-strongly monotone if*

$$(F(\boldsymbol{\omega}) - F(\boldsymbol{\omega}'))^\top(\boldsymbol{\omega} - \boldsymbol{\omega}') \geq \mu\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2^2 \,. \tag{35}$$

B GRADIENT METHODS ON UNCONSTRAINED BILINEAR GAMES

In this section we will prove the results provided in §3, namely Proposition 1, Proposition 2 and Theorem 1. For Proposition 1 and 2 let us recall the context. We wanted to derive properties of some gradient methods on the following simple illustrative example

$$\min_{\theta \in \mathbb{R}} \max_{\phi \in \mathbb{R}} \ \theta \cdot \phi \tag{36}$$

B.1 PROOF OF PROPOSITION 1

Let us first recall the proposition:

Proposition' 1. *The* simultaneous *iterates diverge geometrically and the* alternating *iterates defined in* (11) *are bounded but do not converge to 0 as*

$$\text{Simultaneous: } \theta_{t+1}^2 + \phi_{t+1}^2 = (1+\eta^2)(\theta_t^2 + \phi_t^2)\,, \quad \text{Alternating: } \theta_t^2 + \phi_t^2 = \Theta(\theta_0^2 + \phi_0^2) \tag{37}$$

where $u_t = \Theta(v_t) \Leftrightarrow \exists \alpha, \beta > 0 : \alpha v_t \leq u_t \leq \beta v_t$.

The uniform average $(\bar{\theta}_t, \bar{\phi}_t) \overset{def}{=} \frac{1}{t}\sum_{s=0}^{t-1}(\theta_s, \phi_s)$ *of the* simultaneous *updates (resp. the* alternating updates*) diverges (resp. converges to 0) as,*

$$\text{Simultaneous: } \bar{\theta}_t^2 + \bar{\phi}_t^2 = \Theta\left(\frac{\theta_0^2 + \phi_0^2}{\eta^2 t^2}(1+\eta^2)^t\right), \quad \text{Alternating: } \bar{\theta}_t^2 + \bar{\phi}_t^2 = \Theta\left(\frac{\theta_0^2 + \phi_0^2}{\eta^2 t^2}\right). \tag{38}$$

Proof. Let us start with the *simultaneous* update rule:

$$\begin{cases} \theta_{t+1} = \theta_t - \eta\phi_t \\ \phi_{t+1} = \phi_t + \eta\theta_t\,. \end{cases} \tag{39}$$

Then we have,

$$\theta_{t+1}^2 + \phi_{t+1}^2 = (\theta_t - \eta\phi_t)^2 + (\phi_t + \eta\theta_t)^2 \tag{40}$$

$$= (1+\eta^2)(\theta_t^2 + \phi_t^2)\,. \tag{41}$$

The update rule (39) also gives us,

$$\begin{cases} \eta\phi_t = \theta_t - \theta_{t+1} \\ \eta\theta_{t+1} = \phi_{t+1} - \phi_t\,. \end{cases} \tag{42}$$

Summing these equation for $0 \leq t \leq T-1$ we get,

$$\bar{\phi}_T^2 + \bar{\theta}_T^2 = (\theta_0 - \theta_T)^2 + (\phi_0 - \phi_T)^2 \tag{43}$$

$$= ((1+\eta^2)^T + 1)(\theta_0^2 + \phi_0^2) - 2\theta_0\theta_T - 2\phi_0\phi_T \tag{44}$$

$$= \Theta\left((1+\eta^2)^T((\theta_0^2 + \phi_0^2)\right) \tag{45}$$

Let continue start with the *alternating* update rule

$$\begin{cases} \theta_{t+1} = \theta_t - \eta\phi_t \\ \phi_{t+1} = \phi_t + \eta\theta_{t+1} = \phi_t + \eta(\theta_t - \eta\phi_t) \end{cases} \tag{46}$$

Then we have

$$\begin{bmatrix} \theta_{t+1} \\ \phi_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & -\eta \\ \eta & 1-\eta^2 \end{bmatrix} \begin{bmatrix} \theta_t \\ \phi_t \end{bmatrix} \tag{47}$$

by simple linear algebra, for $\eta < 2$, the matrix $M \overset{\text{def}}{=} \begin{bmatrix} 1 & -\eta \\ \eta & 1-\eta^2 \end{bmatrix}$ has two complex conjugate eigenvalues which are

$$\lambda_\pm = 1 - \eta\frac{\eta \pm i\sqrt{4-\eta^2}}{2} \tag{48}$$

and their squared magnitude is equal to $\det(M) = 1 - \eta^2 + \eta^2 = 1$. We can diagonalize M meaning that there exist P an invertible matrix such that $M = P^{-1}\operatorname{diag}(\lambda_+, \lambda_-)P$. Then, we have

$$\begin{bmatrix} \theta_t \\ \phi_t \end{bmatrix} = M^t \begin{bmatrix} \theta_0 \\ \phi_0 \end{bmatrix} = P^{-1}\operatorname{diag}(\lambda_+^t, \lambda_-^t)P \begin{bmatrix} \theta_0 \\ \phi_0 \end{bmatrix} \tag{49}$$

and consequently,

$$\theta_t^2 + \phi_t^2 = \left\| \begin{bmatrix} \theta_t \\ \phi_t \end{bmatrix} \right\|_{\mathbb{C}}^2 = \left\| P^{-1}\operatorname{diag}(\lambda_+^t, \lambda_-^t)P \begin{bmatrix} \theta_0 \\ \phi_0 \end{bmatrix} \right\|_{\mathbb{C}}^2 \leq \|P^{-1}\|\|P\|(\theta_0^2 + \phi_0^2) \tag{50}$$

where $\|\cdot\|_{\mathbb{C}}$ is the norm in $\mathbb{C}^2$ and $\|P\| := \max_{u\in\mathbb{C}^2} \frac{\|Pu\|_{\mathbb{C}}}{\|u\|_{\mathbb{C}}}$ is the induced matrix norm. The same way we have

$$\theta_0^2 + \phi_0^2 = \left\| M^{-t} \begin{bmatrix} \theta_t \\ \phi_t \end{bmatrix} \right\|_{\mathbb{C}}^2 = \left\| P^{-1} \operatorname{diag}(\lambda_+^{-t}, \lambda_-^{-t}) P \begin{bmatrix} \theta_t \\ \phi_t \end{bmatrix} \right\|_{\mathbb{C}}^2 \leq \|P^{-1}\|\|P\|(\theta_t^2 + \phi_t^2) \tag{51}$$

Hence, if $\theta_0^2 + \phi_0^2 > 0$ the sequence (θ_t, ϕ_t) is bounded but do not converge to 0. Moreover the update rule gives us,

$$\begin{cases} \eta\phi_t = \theta_t - \theta_{t+1} \\ \eta\theta_t = \phi_t - \phi_{t-1} \end{cases} \Rightarrow \begin{cases} \dfrac{\eta}{T}\displaystyle\sum_{t=0}^{T-1} \phi_t = \dfrac{\theta_0 - \theta_T}{T} \\ \dfrac{\eta}{T}\displaystyle\sum_{t=0}^{T-1} \theta_t = \dfrac{\phi_{T-1} - \phi_0 + \eta\theta_0}{T} \end{cases} \Rightarrow \begin{cases} \bar{\phi}_T = \dfrac{\theta_0 - \theta_T}{\eta T} \\ \bar{\theta}_T = \dfrac{\phi_{T-1} - \phi_0 + \eta\theta_0}{\eta T} \end{cases} \tag{52}$$

Consequently, since $\theta_t^2 + \phi_t^2 = \mathcal{O}(\theta_0^2 + \phi_0^2)$,

$$\sqrt{\bar{\theta}_t^2 + \bar{\phi}_t^2} = \mathcal{O}\left(\frac{\sqrt{\theta_0^2 + \phi_0^2}}{\eta t}\right) \tag{53}$$

□

B.2 Implicit and extrapolation method

In this section we will prove a slightly more precise proposition than Proposition 2,

Proposition' 2. *The squared norm of the iterates $N_t \stackrel{def}{=} \theta_t^2 + \phi_t^2$, where the update rule of θ_t and ϕ_t defined in (18), decrease geometrically for any $0 < \eta < 1$ as,*

$$\text{Implicit: } N_{t+1} = \frac{N_t}{1+\eta^2}\,, \quad \text{Extrapolation: } N_{t+1} = (1 - \eta^2 + \eta^4)N_t\,, \quad \forall t \geq 0 \tag{54}$$

Proof. Let us recall the update rule for the implicit method

$$\begin{cases} \theta_{t+1} = \theta_t - \eta\phi_{t+1} \\ \phi_{t+1} = \phi_t + \eta\theta_{t+1} \end{cases} \Rightarrow \begin{cases} (1+\eta^2)\theta_{t+1} = \theta_t - \eta\phi_t \\ (1+\eta^2)\phi_{t+1} = \phi_t + \eta\theta_t \end{cases} \tag{55}$$

Then,

$$(1+\eta^2)^2(\theta_{t+1}^2 + \phi_{t+1}^2) = (\theta_t - \eta\phi_t)^2 + (\phi_t + \eta\theta_t)^2 \tag{56}$$

$$= \theta_t^2 + \phi_t^2 + +\eta^2(\theta_t^2 + \phi_t^2) \tag{57}$$

Implying that

$$\theta_{t+1}^2 + \phi_{t+1}^2 = \frac{\theta_t^2 + \phi_t^2}{1+\eta^2} \tag{58}$$

For the extrapolation method we have the update rule

$$\begin{cases} \theta_{t+1} = \theta_t - \eta(\phi_t + \eta\theta_t) \\ \phi_{t+1} = \phi_t + \eta(\theta_t - \eta\phi_t) \end{cases} \tag{59}$$

Implying that,

$$\theta_{t+1}^2 + \phi_{t+1}^2 = (\theta_t - \eta(\phi_t + \eta\theta_t))^2 + (\phi_t + \eta(\theta_t - \eta\phi_t))^2 \tag{60}$$

$$= \theta_t^2 + \phi_t^2 - 2\eta^2(\theta^2 + \phi^2) + \eta^2((\theta_t - \eta\phi_t)^2 + (\phi_t + \eta\theta_t)^2) \tag{61}$$

$$= (1 - \eta^2 + \eta^4)(\theta_t^2 + \phi_t^2) \tag{62}$$

□

B.3 EXTRAPOLATION FROM THE PAST

Let us recall what we call *projected extrapolation form the past*, where we used the notation $\boldsymbol{\omega}'_t = \boldsymbol{\omega}_{t+1/2}$ for compactness,

$$\text{Extrapolation from the past:} \quad \boldsymbol{\omega}'_t = P_\Omega[\boldsymbol{\omega}_t - \eta F(\boldsymbol{\omega}'_{t-1})] \tag{63}$$

$$\text{Perform update step:} \quad \boldsymbol{\omega}_{t+1} = P_\Omega[\boldsymbol{\omega}_t - \eta F(\boldsymbol{\omega}'_t)] \;\text{ and store: }\; F(\boldsymbol{\omega}'_t) \tag{64}$$

where $P_\Omega[\cdot]$ is the projection onto the constraint set Ω. An operator $F : \Omega \to \mathbb{R}^d$ is said to be μ-strongly monotone if

$$(F(\boldsymbol{\omega}) - F(\boldsymbol{\omega}'))^\top (\boldsymbol{\omega} - \boldsymbol{\omega}') \geq \mu \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2^2 \,. \tag{65}$$

If F is strongly monotone we can prove the following theorem:

Theorem' 1. *If F is μ-strongly monotone (see §A for the definition of strong monotonicity) and L-Lipschitz, then the updates (20) and (21) with $\eta = \frac{1}{4L}$ provide linearly converging iterates,*

$$\|\boldsymbol{\omega}_t - \boldsymbol{\omega}^*\|_2^2 \leq \left(1 - \frac{\mu}{4L}\right)^t \|\boldsymbol{\omega}_0 - \boldsymbol{\omega}^*\|_2^2 \,, \quad \forall t \geq 0 \,. \tag{66}$$

Proof. In order to prove tis theorem we will prove a slightly more general result,

$$\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}^*\|_2^2 + \|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 \leq \left(1 - \frac{\mu}{4L}\right) (\|\boldsymbol{\omega}_t - \boldsymbol{\omega}^*\|_2^2 + \|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_{t-2}\|_2^2) \,. \tag{67}$$

implying that

$$\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}^*\|_2^2 \leq \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}^*\|_2^2 + \|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 \leq \left(1 - \frac{\mu}{4L}\right)^t \|\boldsymbol{\omega}_0 - \boldsymbol{\omega}^*\|_2^2 \,. \tag{68}$$

with the convention that $\boldsymbol{\omega}'_0 = \boldsymbol{\omega}'_{-1} = \boldsymbol{\omega}'_{-2}$.

Let us first proof three technical lemmas,

Lemma 3. *If F is μ-strongly monotone we have,*

$$\mu \left(\|\boldsymbol{\omega}_t - \boldsymbol{\omega}^*\|_2^2 - 2\|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2 \right) \leq 2F(\boldsymbol{\omega}'_t)^\top (\boldsymbol{\omega}'_t - \boldsymbol{\omega}^*) \,, \quad \forall \boldsymbol{\omega}^* \in \Omega^* \,. \tag{69}$$

Proof. By strong monotonicity and optimality of $\boldsymbol{\omega}^*$,

$$2\mu\|\boldsymbol{\omega}'_t - \boldsymbol{\omega}^*\|_2^2 \leq 2F(\boldsymbol{\omega}^*)^\top (\boldsymbol{\omega}'_t - \boldsymbol{\omega}^*) + 2\mu\|\boldsymbol{\omega}'_t - \boldsymbol{\omega}^*\|_2^2 \leq 2F(\boldsymbol{\omega}'_t)^\top (\boldsymbol{\omega}'_t - \boldsymbol{\omega}^*) \tag{70}$$

and then we use the inequality $2\|\boldsymbol{\omega}'_t - \boldsymbol{\omega}^*\|_2^2 \geq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}^*\|_2^2 - 2\|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2$ to get the result claimed. □

Lemma 4. *If F is L-Lipschitz, we have for any $\boldsymbol{\omega} \in \Omega$,*

$$2\eta_t F(\boldsymbol{\omega}'_t)^\top (\boldsymbol{\omega}'_t - \boldsymbol{\omega}) \leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2 + \eta_t^2 L^2 \|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 \,. \tag{71}$$

Proof. Applying Lemma 2 for $(\boldsymbol{\omega}, \boldsymbol{u}, \boldsymbol{\omega}^+, \boldsymbol{\omega}') = (\boldsymbol{\omega}_t, -\eta_t F(\boldsymbol{\omega}'_t), \boldsymbol{\omega}_{t+1}, \boldsymbol{\omega})$ and $(\boldsymbol{\omega}, \boldsymbol{u}, \boldsymbol{\omega}^+, \boldsymbol{\omega}') = (\boldsymbol{\omega}_t, -\eta_t F(\boldsymbol{\omega}'_{t-1}), \boldsymbol{\omega}'_t, \boldsymbol{\omega}_{t+1})$, we get,

$$\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 \leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_t)^\top (\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}) - \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}_t\|_2^2 \tag{72}$$

and

$$\|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_{t+1}\|_2^2 \leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}_{t+1}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_{t-1})^\top (\boldsymbol{\omega}'_t - \boldsymbol{\omega}_{t+1}) - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2 \,. \tag{73}$$

Summing (72) and (73) we get,

$$\begin{aligned}
\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 &\leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_t)^\top (\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}) && (74)\\
&\quad - 2\eta_t F(\boldsymbol{\omega}'_{t-1})^\top (\boldsymbol{\omega}'_t - \boldsymbol{\omega}_{t+1}) - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2 - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_{t+1}\|_2^2 && (75)\\
&= \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_t)^\top (\boldsymbol{\omega}'_t - \boldsymbol{\omega}) - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2 - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_{t+1}\|_2^2 \\
&\quad - 2\eta_t (F(\boldsymbol{\omega}'_{t-1}) - F(\boldsymbol{\omega}'_t))^\top (\boldsymbol{\omega}'_t - \boldsymbol{\omega}_{t+1}) \,. && (76)
\end{aligned}$$

Then, we can use the Young's inequality $2a^\top b \leq \|a\|_2^2 + \|b\|_2^2$ to get,

$$\begin{aligned}\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 &\leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_t)^\top(\boldsymbol{\omega}'_t - \boldsymbol{\omega}) + \eta_t^2\|F(\boldsymbol{\omega}'_{t-1}) - F(\boldsymbol{\omega}'_t)\|_2^2 \\ &\quad + \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_{t+1}\|_2^2 - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2 - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_{t+1}\|_2^2 \qquad (77)\\ &= \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_t)^\top(\boldsymbol{\omega}'_t - \boldsymbol{\omega}) + \eta_t^2\|F(\boldsymbol{\omega}'_{t-1}) - F(\boldsymbol{\omega}'_t)\|_2^2 - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2 \\ &\leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_t)^\top(\boldsymbol{\omega}'_t - \boldsymbol{\omega}) + \eta_t^2 L^2\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2 . \qquad (78)\end{aligned}$$

□

Lemma 5. *For all $t \geq 0$, if we set $\boldsymbol{\omega}'_{-2} = \boldsymbol{\omega}'_{-1} = \boldsymbol{\omega}'_0$ we have*

$$\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 \leq 4\|\boldsymbol{\omega}_t - \boldsymbol{\omega}'_t\|_2^2 + 4\eta_{t-1}^2 L^2\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_{t-2}\|_2^2 - \|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 . \qquad (79)$$

Proof. We start with $\|a + b\|_2^2 \leq 2\|a\|^2 + 2\|b\|^2$.

$$\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 \leq 2\|\boldsymbol{\omega}_t - \boldsymbol{\omega}'_t\|_2^2 + 2\|\boldsymbol{\omega}_t - \boldsymbol{\omega}'_{t-1}\|_2^2 . \qquad (80)$$

Moreover, since the projection is contractive we have that

$$\begin{aligned}\|\boldsymbol{\omega}_t - \boldsymbol{\omega}'_{t-1}\|_2^2 &\leq \|\boldsymbol{\omega}_{t-1} - \eta_{t-1}F(\boldsymbol{\omega}'_{t-1}) - \boldsymbol{\omega}_{t-1} - \eta_{t-1}F(\boldsymbol{\omega}'_{t-2})\|_2^2 \qquad (81)\\ &= \eta_{t-1}^2\|F(\boldsymbol{\omega}'_{t-1}) - F(\boldsymbol{\omega}'_{t-2})\|_2^2 \qquad (82)\\ &\leq \eta_{t-1}^2 L^2\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_{t-2}\|_2^2 . \qquad (83)\end{aligned}$$

Combining (80) and (83) we get,

$$\begin{aligned}\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 &= 2\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 - \|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 \qquad (84)\\ &\leq 4\|\boldsymbol{\omega}_t - \boldsymbol{\omega}'_t\|_2^2 + 4\|\boldsymbol{\omega}_t - \boldsymbol{\omega}'_{t-1}\|_2^2 - \|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 \qquad (85)\\ &\leq 4\|\boldsymbol{\omega}_t - \boldsymbol{\omega}'_t\|_2^2 + 4\eta_{t-1}^2 L^2\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_{t-2}\|_2^2 - \|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 . \qquad (86)\end{aligned}$$

□

Proof of Theorem 1. Let $\boldsymbol{\omega}^* \in \Omega^*$ be an optimal point of (VIP). Combining Lemma 3 and Lemma 4 we get,

$$\eta_t\mu\left(\|\boldsymbol{\omega}_t - \boldsymbol{\omega}^*\|_2^2 - 2\|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2\right) \leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}^*\|_2^2 - \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}^*\|_2^2 + \eta_t^2 L^2\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2$$

leading to,

$$\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}^*\|_2^2 \leq (1 - \eta_t\mu)\|\boldsymbol{\omega}_t - \boldsymbol{\omega}^*\|_2^2 + \eta_t^2 L^2\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 - (1 - 2\eta_t\mu)\|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2 \qquad (87)$$

Now using Lemma 5 we get,

$$\begin{aligned}\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}^*\|_2^2 &\leq (1 - \eta_t\mu)\|\boldsymbol{\omega}_t - \boldsymbol{\omega}^*\|_2^2 + \eta_t^2 L^2(4\eta_{t-1}^2 L^2\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_{t-2}\|_2^2 - \|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2) \\ &\quad - (1 - 2\eta_t\mu - 4\eta_t^2 L^2)\|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2 \qquad (88)\end{aligned}$$

Now with $\eta_t = \frac{1}{4L} \leq \frac{1}{4\mu}$ we get,

$$\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}^*\|_2^2 \leq \left(1 - \frac{\mu}{4L}\right)\|\boldsymbol{\omega}_t - \boldsymbol{\omega}^*\|_2^2 + \frac{1}{16}\left(\frac{1}{4}\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_{t-2}\|_2^2 - \|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2\right)$$

Hence, using the fact that $\frac{\mu}{4L} \leq \frac{1}{4}$ we get,

$$\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}^*\|_2^2 + \tfrac{1}{16}\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 \leq \left(1 - \frac{\mu}{4L}\right)\left(\|\boldsymbol{\omega}_t - \boldsymbol{\omega}^*\|_2^2 + \tfrac{1}{16}\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_{t-2}\|_2^2\right) . \qquad (89)$$

□

C MORE ON MERIT FUNCTIONS

In this section, we will present how to handle *unbounded* constaints set Ω with a more refined *merit function* than (25) used in the main paper. Let F the continuous operator and Ω the constraints set associated with the VIP

$$\text{find } \boldsymbol{\omega}^* \in \Omega \quad \text{such that} \quad F(\boldsymbol{\omega}^*)^\top(\boldsymbol{\omega} - \boldsymbol{\omega}^*) \geq 0\,, \quad \forall \boldsymbol{\omega} \in \Omega\,. \tag{VIP}$$

When the operator F is monotone, we have that $F(\boldsymbol{\omega}^*)^\top(\boldsymbol{\omega} - \boldsymbol{\omega}^*) \leq F(\boldsymbol{\omega})^\top(\boldsymbol{\omega} - \boldsymbol{\omega}^*)\,,\ \forall \boldsymbol{\omega}, \boldsymbol{\omega}^*$. Hence, in this case (VIP) implies a stronger formulation sometimes called *Minty variational inequality* (Crespi et al., 2005):

$$\text{find } \boldsymbol{\omega}^* \in \Omega \quad \text{such that} \quad F(\boldsymbol{\omega})^\top(\boldsymbol{\omega} - \boldsymbol{\omega}^*) \geq 0\,, \quad \forall \boldsymbol{\omega} \in \Omega\,. \tag{MVI}$$

This formulation is stronger in the sense that if (MVI) holds for some $\boldsymbol{\omega}^* \in \Omega$, then (VIP) holds too. A *merit function* useful for our analysis can be derived from this formulation. Roughly, a merit function is a convergence measure. More formally, a function $g : \Omega \to \mathbb{R}$ is called a *merit function* if g is non-negative such that $g(\boldsymbol{\omega}) = 0 \Leftrightarrow \boldsymbol{\omega} \in \Omega^*$ (Larsson and Patriksson, 1994). A way to derive a merit function from (MVI) would be to use $g(\boldsymbol{\omega}^*) = \sup_{\boldsymbol{\omega} \in \Omega} F(\boldsymbol{\omega})^\top(\boldsymbol{\omega}^* - \boldsymbol{\omega})$ which is zero if and only if (MVI) holds for $\boldsymbol{\omega}^*$. To deal with unbounded constraint sets (leading to a potentially infinite valued function outside of the optimal set), we use the *restricted merit function* (Nesterov, 2007):

$$\mathrm{Err}_R(\boldsymbol{\omega}_t) \stackrel{\text{def}}{=} \max_{\boldsymbol{\omega} \in \Omega, \|\boldsymbol{\omega} - \boldsymbol{\omega}_0\| \leq R} F(\boldsymbol{\omega})^\top(\boldsymbol{\omega}_t - \boldsymbol{\omega})\,. \tag{90}$$

This function acts as merit function for (VIP) on the interior of the open ball of radius R around $\boldsymbol{\omega}_0$, as shown in Lemma 1 of Nesterov (2007). That is, let $\Omega_R \stackrel{\text{def}}{=} \Omega \cap \{\boldsymbol{\omega} : \|\boldsymbol{\omega} - \boldsymbol{\omega}_0\| < R\}$. Then for any point $\hat{\boldsymbol{\omega}} \in \Omega_R$, we have:

$$\mathrm{Err}_R(\hat{\boldsymbol{\omega}}) = 0 \Leftrightarrow \hat{\boldsymbol{\omega}} \in \Omega^* \cap \Omega_R. \tag{91}$$

The reference point $\boldsymbol{\omega}_0$ is arbitrary, but in practice it is usually the initialization point of the algorithm. R has to be big enough to ensure that Ω_R contains a solution. Err_R measures how much (MVI) is violated on the restriction Ω_R. Such merit function is standard in the variational inequality literature. A similar one is used in (Nemirovski, 2004; Juditsky et al., 2011). When F is derived from the gradients (5) of a zero-sum game, we can define a more interpretable merit function. One has to be careful though when extending properties from the minimization setting to the saddle point setting (e.g. the merit function used by Yadav et al. (2018) is vacuous for a bilinear game as explained in App C.2).

In the appendix we adopt a set of assumptions a little more general than the one in the main paper

Assumption 4.
- *F is* monotone *and Ω is convex and closed.*
- *R is set big enough such that $R > \|\boldsymbol{\omega}_0 - \boldsymbol{\omega}^*\|$ and F is a* monotone *operator.*

Contrary to Assumption 3, in Assumption 4 the constraints set in no longer assumed to be bounded. Assumption 4 is more general than Assumption 3 since that if Ω is bounded then with R set to the diameter of Ω, Assumption 4 is true.

C.1 MORE GENERAL MERIT FUNCTIONS

In the appendix we will note $\mathrm{Err}_R^{(\mathrm{VI})}$ the *restricted merit function* defined in (90). Let us recall its definition,

$$\mathrm{Err}_R^{(\mathrm{VI})}(\boldsymbol{\omega}_t) \stackrel{\text{def}}{=} \max_{\boldsymbol{\omega} \in \Omega, \|\boldsymbol{\omega} - \boldsymbol{\omega}_0\| \leq R} F(\boldsymbol{\omega})^\top(\boldsymbol{\omega}_t - \boldsymbol{\omega})\,. \tag{92}$$

When the objective is a saddle point problem i.e.,

$$F(\boldsymbol{\theta}, \boldsymbol{\varphi}) = [\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}) \ \ -\nabla_{\boldsymbol{\varphi}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi})]^\top \tag{93}$$

and $\mathcal{L}$ is *convex-concave* (see Definition 4 in §A), we can use another merit function than (92) on Ω_R that is more interpretable and more directly related to the cost function of the minimax formulation:

$$\mathrm{Err}_R^{(\mathrm{SP})}(\boldsymbol{\theta}_t, \boldsymbol{\varphi}_t) \stackrel{\text{def}}{=} \max_{\substack{\boldsymbol{\varphi} \in \Phi, \boldsymbol{\theta} \in \Theta \\ \|(\boldsymbol{\theta}, \boldsymbol{\varphi}) - (\boldsymbol{\theta}_0, \boldsymbol{\varphi}_0)\| \leq R}} \mathcal{L}(\boldsymbol{\theta}_t, \boldsymbol{\varphi}) - \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}_t)\,. \tag{94}$$

In particular, if the equilibrium $(\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*) \in \Omega^* \cap \Omega_R$ and we have that $\mathcal{L}(\cdot, \boldsymbol{\varphi}^*)$ and $-\mathcal{L}(\boldsymbol{\theta}^*, \cdot)$ are μ-strongly convex (see §A), then the merit function for saddle points upper bounds the distance for $(\boldsymbol{\theta}, \boldsymbol{\varphi}) \in \Omega_R$ to the equilibrium as:

$$\mathrm{Err}_R^{(\mathrm{SP})}(\boldsymbol{\theta}, \boldsymbol{\varphi}) \geq \frac{\mu}{2}(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 + \|\boldsymbol{\varphi} - \boldsymbol{\varphi}^*\|_2^2)\,. \tag{95}$$

In the appendix, we provide our convergence results with the merit functions (92) and (94), depending on the setup:

$$\mathrm{Err}_R(\boldsymbol{\omega}) \stackrel{\text{def}}{=} \begin{cases} \mathrm{Err}_R^{(\mathrm{SP})}(\boldsymbol{\omega}) & \text{if } F \text{ is a SP operator (5)} \\ \mathrm{Err}_R^{(\mathrm{VI})}(\boldsymbol{\omega}) & \text{otherwise.} \end{cases} \tag{96}$$

C.2 On the importance of the merit function

In this section, we illustrate the fact that one has to be careful when extending results and properties from the minimization setting to the minimax setting (and consequently to the variational inequality setting). Another candidate as a merit function for saddle point optimization would be to naturally extend the suboptimality $f(\boldsymbol{\omega}) - f(\boldsymbol{\omega}^*)$ used in standard minimization (i.e. find $\boldsymbol{\omega}^*$ the minimizer of f) to the gap $P(\boldsymbol{\theta}, \boldsymbol{\varphi}) = \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}^*) - \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\varphi})$. In a previous analysis of a modification of the stochastic gradient descent (SGD) method for GANs, Yadav et al. (2018) gave their convergence rate on P that they called the "primal-dual" gap. Unfortunately, if we do not assume that the function $\mathcal{L}$ is strongly convex-concave (a stronger assumption defined in §A and which fails for bilinear objective e.g.), P may not be a *merit function*. It can be 0 for a non optimal point, see for instance the discussion on the differences between (94) and P in (Gidel et al., 2017, Section 3). In particular, for the simple 2D bilinear example $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}) = \boldsymbol{\theta} \cdot \boldsymbol{\varphi}$, we have that $\boldsymbol{\theta}^* = \boldsymbol{\varphi}^* = 0$ and thus $P(\boldsymbol{\theta}, \boldsymbol{\varphi}) = 0 \quad \forall \boldsymbol{\theta}, \boldsymbol{\varphi}$.

C.3 Variational inequalities for non-convex cost functions

When the cost functions defined in (3) are non-convex, the operator F is no longer monotone. Nevertheless, (VIP) and (MVI) can still be defined, though a solution to (MVI) is less likely to exist. We note that (VIP) is a local condition for F (as only evaluating F at the points $\boldsymbol{\omega}^*$). On the other hand, an appealing property of (MVI) is that it is a global condition. In the context of minimization of a function f for example (where $F = \nabla f$), if $\boldsymbol{\omega}^*$ solves (MVI) then $\boldsymbol{\omega}^*$ is a *global* minimum of f (and not just a stationary point for the solution of (MVI); see Proposition 2.2 from Crespi et al. (2005)).

A less restrictive way to consider variational inequalities in the non-monotone setting is to use a local version of (MVI). If the cost functions are locally convex around the optimal couple $(\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*)$ and if our iterates eventually fall and stay into that neighborhood, then we can consider our restricted merit function $\mathrm{Err}_R(\cdot)$ with a well suited constant R and apply our convergence results for monotone operators.

D Another way of implementing extrapolation to SGD

We now introduce another way to combine extrapolation and SGD. This extension is very similar to Alg. 5, the only difference is that it re-uses the minibatch sample of the extrapolation step for the update of the current point. The intuition is that it correlates the estimator of the gradient of the extrapolation step and the one of the update step leading to a better correction of the oscillations which are also due to the stochasticity. One emerging issue (for the analysis) of this method is that since $\boldsymbol{\omega}'_t$ depend on ξ_t, the quantity $F(\boldsymbol{\omega}'_t, \xi_t)$ is a biased estimator of $F(\boldsymbol{\omega}'_t)$.

Algorithm 5 Re-used minibatches for stochastic extrapolation (ReExtraSGD)

1: Let $\boldsymbol{\omega}_0 \in \Omega$
2: **for** $t = 0 \dots T-1$ **do**
3: Sample $\xi_t \sim P$
4: $\boldsymbol{\omega}'_t \stackrel{\text{def}}{=} P_\Omega[\boldsymbol{\omega}_t - \eta_t F(\boldsymbol{\omega}_t, \xi_t)]$ ▷ Extrapolation step
5: $\boldsymbol{\omega}_{t+1} \stackrel{\text{def}}{=} P_\Omega[\boldsymbol{\omega}_t - \eta_t F\big(\boldsymbol{\omega}'_t, \xi_t\big)]$ ▷ Update step wit the **same** sample
6: **end for**
7: Return $\bar{\boldsymbol{\omega}}_T = \sum_{t=0}^{T-1} \eta_t \boldsymbol{\omega}'_t / \sum_{t=0}^{T-1} \eta_t$

Theorem 5. *Assume that $\|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_0\| \leq R$, $\forall t \geq 0$ where $(\boldsymbol{\omega}'_t)_{t\geq 0}$ are the iterates of Alg. 5. Under Assumption 1 and 4, for any $T \geq 1$, Alg. 5 with constant step-size $\eta \leq \frac{1}{\sqrt{2}L}$ has the following convergence properties:*

$$\mathbb{E}[\mathrm{Err}_R(\bar{\boldsymbol{\omega}}_T)] \leq \frac{R^2}{\eta T} + \eta \frac{\sigma^2 + 4L^2(4R^2 + \sigma^2)}{2} \quad \textit{where} \quad \bar{\boldsymbol{\omega}}_T \stackrel{\textit{def}}{=} \frac{1}{T} \sum_{t=0}^{T-1} \boldsymbol{\omega}'_t \,.$$

Particularly, $\eta_t = \frac{\eta}{\sqrt{T}}$ gives $\mathbb{E}[\mathrm{Err}_R(\bar{\boldsymbol{\omega}}_T)] \leq \frac{O(1)}{\sqrt{T}}$.

The assumption that the sequence of the iterates provided by the algorithm is bounded is strong, but has been also done for instance in (Yadav et al., 2018). The proof of this result is provided in §F.

E Variance comparison between AvgSGD and SGD with prediction method

To compare the variance term of AvgSGD in (26) with the one of the *SGD with prediction method* (Yadav et al., 2018), we need to have the same convergence certificate. Fortunately, their proof can be adapted to our convergence criterion (using Lemma 6 in §F), revealing an extra $\sigma^2/2$ in the variance term from their paper. The resulting variance can be summarized with our notation as $(M^2(1+L)+\sigma^2)/2$ where the L is the Lipschitz constant of the operator F. Since $M \gg \sigma$, their variance term is then $1+L$ time larger than the one provided by the AvgSGD method.

F Proof of Theorems

This section is dedicated on the proof of the theorems provided in this paper in a slightly more general form working with the merit function defined in (96). First we prove an additional lemma necessary to the proof of our theorems.

Lemma 6. *Let F be a monotone operator and let $(\boldsymbol{\omega}_t), (\boldsymbol{\omega}'_t), (\boldsymbol{z}_t), (\Delta_t), (\xi_t)$ and (ζ_t) be six random sequences such that, for all $t \geq 0$*

$$2\eta_t F(\boldsymbol{\omega}'_t)^\top (\boldsymbol{\omega}'_t - \boldsymbol{\omega}) \leq N_t - N_{t+1} + \eta_t^2 (M_1(\boldsymbol{\omega}_t, \xi_t) + M_2(\boldsymbol{\omega}'_t, \zeta_t)) + 2\eta_t \Delta_t^\top (\boldsymbol{z}_t - \boldsymbol{\omega}) \,,$$

where $N_t = N(\boldsymbol{\omega}_t, \boldsymbol{\omega}'_{t-1}, \boldsymbol{\omega}'_{t-2}) \geq 0$ and we extend $(\boldsymbol{\omega}'_t)$ with $\boldsymbol{\omega}'_{-2} = \boldsymbol{\omega}'_{-1} = \boldsymbol{\omega}'_0$. Let also assume that with $N_0 \leq R$, $\mathbb{E}[\|\Delta_t\|_2^2] \leq \sigma^2$, $\mathbb{E}[\Delta_t | \boldsymbol{z}_t, \Delta_0, \dots, \Delta_{t-1}] = 0$, $\mathbb{E}[M_1(\boldsymbol{\omega}_t, \xi_t)] \leq M_1$ and $\mathbb{E}[M_2(\boldsymbol{\omega}'_t, \zeta_t)] \leq M_2$, then,

$$\mathbb{E}[\mathrm{Err}_R(\bar{\boldsymbol{\omega}}_T)] \leq \frac{R^2}{S_T} + \frac{M_1 + M_2 + \sigma^2}{2S_T} \sum_{t=0}^{T-1} \eta_t^2 \tag{97}$$

where $\bar{\boldsymbol{\omega}}_T \stackrel{\textit{def}}{=} \sum_{t=0}^{T-1} \eta_t \boldsymbol{\omega}'_t / S_T$ and $S_T \stackrel{\textit{def}}{=} \sum_{t=0}^{T-1} \eta_t$.

Proof of Lemma 6. We sum (6) for $0 \leq t \leq T-1$ to get,

$$2\sum_{t=0}^{T-1} \eta_t F(\boldsymbol{\omega}'_t)^\top (\boldsymbol{\omega}'_t - \boldsymbol{\omega}) \leq \sum_{t=0}^{T-1} \left[(N_t - N_{t+1}) + \eta_t^2((M_1(\boldsymbol{\omega}_t, \xi_t) + M_2(\boldsymbol{\omega}'_t, \zeta_t)) + 2\eta_t \Delta_t^\top (\boldsymbol{z}_t - \boldsymbol{\omega}) \right] . \tag{98}$$

We will then upper bound each sum in the right-hand side,

$$\Delta_t^\top (\boldsymbol{z}_t - \boldsymbol{\omega}) = \Delta_t^\top (\boldsymbol{z}_t - \boldsymbol{u}_t) + \Delta_t^\top (\boldsymbol{u}_t - \boldsymbol{\omega})$$

where $\boldsymbol{u}_{t+1} \stackrel{\text{def}}{=} P_\Omega(\boldsymbol{u}_t - \eta_t \Delta_t)$ and $\boldsymbol{u}_0 \stackrel{\text{def}}{=} \boldsymbol{\omega}_0$. Then,

$$\|\boldsymbol{u}_{t+1} - \boldsymbol{\omega}\|_2^2 \leq \|\boldsymbol{u}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t \Delta_t^\top (\boldsymbol{u}_t - \boldsymbol{\omega}) + \eta_t^2 \|\Delta_t\|_2^2$$

leading to

$$2\eta_t \Delta_t^\top (\boldsymbol{z}_t - \boldsymbol{\omega}) \leq 2\eta_t \Delta_t^\top (\boldsymbol{z}_t - \boldsymbol{u}_t) + \|\boldsymbol{u}_t - \boldsymbol{\omega}\|_2^2 - \|\boldsymbol{u}_{t+1} - \boldsymbol{\omega}\|_2^2 + \eta_t^2 \|\Delta_t\|_2^2 \tag{99}$$

Then noticing that $\boldsymbol{z}_0 \stackrel{\text{def}}{=} \boldsymbol{\omega}_0$, back to (98) we get a telescoping sum,

$$2\sum_{t=0}^{T-1} \eta_t F(\boldsymbol{\omega}'_t)^\top (\boldsymbol{\omega}'_t - \boldsymbol{\omega}) \leq 2N_0 + \sum_{t=0}^{T-1} \left[\eta_t^2((M_1(\boldsymbol{\omega}_t, \xi_t) + M_2(\boldsymbol{\omega}'_t, \zeta_t)) + \|\Delta_t\|_2^2) + 2\eta_t \Delta_t^\top (\boldsymbol{z}_t - \boldsymbol{u}_t) \right] . \tag{100}$$

If F is the operator of a convex-concave saddle point (5), we get, with $\boldsymbol{\omega}'_t = (\boldsymbol{\theta}_t, \boldsymbol{\varphi}_t)$

$$\begin{aligned} F(\boldsymbol{\omega}'_t)^\top (\boldsymbol{\omega}'_t - \boldsymbol{\omega}) &\geq \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_t, \boldsymbol{\varphi}_t)^\top (\boldsymbol{\theta}_t - \boldsymbol{\theta}) - \nabla_{\boldsymbol{\varphi}} \mathcal{L}(\boldsymbol{\theta}_t, \boldsymbol{\varphi}_t)^\top (\boldsymbol{\varphi}_t - \boldsymbol{\varphi}) \\ &\geq \mathcal{L}(\boldsymbol{\theta}_t, \boldsymbol{\varphi}) - \mathcal{L}(\boldsymbol{\theta}_t, \boldsymbol{\varphi}_t) + \mathcal{L}(\boldsymbol{\theta}_t, \boldsymbol{\varphi}_t) - \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}_t) \\ &\qquad \text{(by convexity and concavity)} \\ &= \mathcal{L}(\boldsymbol{\theta}_t, \boldsymbol{\varphi}) - \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}_t) \end{aligned}$$

then by convexity of $\mathcal{L}(\cdot, \boldsymbol{\varphi})$ and concavity of $\mathcal{L}(\boldsymbol{\theta}, \cdot)$, we have that,

$$2S_T \sum_{t=0}^{T-1} \frac{\eta_t}{S_T} F(\boldsymbol{\omega}'_t)^\top (\boldsymbol{\omega}'_t - \boldsymbol{\omega}) \geq 2S_T \sum_{t=0}^{T-1} \frac{\eta_t}{S_T} (\mathcal{L}(\boldsymbol{\theta}_t, \boldsymbol{\varphi}) - \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}_t)) \geq \bar{2} S_T (\mathcal{L}(\bar{\boldsymbol{\theta}}_t, \boldsymbol{\varphi}) - \mathcal{L}(\bar{\boldsymbol{\theta}}, \boldsymbol{\varphi}_t)) \tag{101}$$

Otherwise if the operator F is just monotone since $F(\boldsymbol{\omega}'_t)^\top (\boldsymbol{\omega}'_t - \boldsymbol{\omega}) \geq F(\boldsymbol{\omega}')^\top (\boldsymbol{\omega}'_t - \boldsymbol{\omega})$ we have that

$$2S_T \sum_{t=0}^{T-1} \eta_t F(\boldsymbol{\omega}'_t)^\top (\boldsymbol{\omega}'_t - \boldsymbol{\omega}) \geq 2S_T \sum_{t=0}^{T-1} \eta_t F(\boldsymbol{\omega}')^\top (\boldsymbol{\omega}'_t - \boldsymbol{\omega}) = 2S_T F(\boldsymbol{\omega}')^\top (\bar{\boldsymbol{\omega}}_t - \boldsymbol{\omega}) \tag{102}$$

In both cases, we can now maximize the left hand side respect to $\boldsymbol{\omega}$ (since the RHS does not depend on $\boldsymbol{\omega}$) to get,

$$2S_T \operatorname{Err}_R(\bar{\boldsymbol{\omega}}_t) \leq 2R^2 + \sum_{t=0}^{T-1} \left[\eta_t^2((M_1(\boldsymbol{\omega}_t, \xi_t) + M_2(\boldsymbol{\omega}'_t, \zeta_t)) + \|\Delta_t\|_2^2) + 2\eta_t \Delta_t^\top (\boldsymbol{z}_t - \boldsymbol{u}_t) \right] . \tag{103}$$

Then taking the expectation, since $\mathbb{E}[\Delta_t | \boldsymbol{z}_t, \boldsymbol{u}_t] = \mathbb{E}[\Delta_t | \boldsymbol{z}_t, \Delta_0, \ldots, \Delta_{t-1}] = 0$, $\mathbb{E}_{\zeta_t}[\|\Delta_t\|_2^2] \leq \sigma^2$, $\mathbb{E}_{\xi_t}[M_1(\boldsymbol{\omega}_t, \xi_t)] \leq M_1$ and $\mathbb{E}_{\zeta_t}[M_2(\boldsymbol{\omega}'_t, \zeta_t)] \leq M_2$, we get that,

$$\mathbb{E}[\operatorname{Err}_R(\bar{\boldsymbol{\omega}}_T)] \leq \frac{R^2}{S_T} + \frac{M_1 + M_2 + \sigma^2}{2S_T} \sum_{t=0}^{T-1} \eta_t^2 \tag{104}$$

□

F.1 PROOF OF THM.2

First let us state Theorem 2 in its general form,

Theorem' 2. *Under Assumption 1, 2 and 4, Alg. 1 with constant step-size η has the following convergence rate for all $T \geq 1$,*

$$\mathbb{E}[\mathrm{Err}_R(\bar{\boldsymbol{\omega}}_T)] \leq \frac{R^2}{2\eta T} + \eta\frac{M^2+\sigma^2}{2} \quad \textit{where} \quad \bar{\boldsymbol{\omega}}_T \overset{\textit{def}}{=} \frac{1}{T}\sum_{t=0}^{T-1}\boldsymbol{\omega}_t\,. \tag{105}$$

Particularly, $\eta = \frac{R}{\sqrt{T(M^2+\sigma^2)}}$ gives $\mathbb{E}[\mathrm{Err}_R(\bar{\boldsymbol{\omega}}_T)] \leq \frac{R\sqrt{M^2+\sigma^2}}{\sqrt{T}}$.

Proof of Theorem 2. Let any $\boldsymbol{\omega} \in \Omega$ such that $\|\boldsymbol{\omega}_0 - \boldsymbol{\omega}\|_2 \leq R$,

$$\begin{aligned}
\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 &= \|P_\Omega(\boldsymbol{\omega}_t - \eta_t F(\boldsymbol{\omega}_t, \xi_t)) - \boldsymbol{\omega}\|_2^2 \\
&\leq \|\boldsymbol{\omega}_t - \eta_t F(\boldsymbol{\omega}_t, \xi_t)) - \boldsymbol{\omega}\|_2^2 \\
&\quad \text{(projections are non-contractive, Lemma 1)} \\
&= \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}_t, \xi_t)^\top(\boldsymbol{\omega}_t - \boldsymbol{\omega}) + \|\eta_t F(\boldsymbol{\omega}_t, \xi_t)\|_2^2
\end{aligned}$$

Then we can make appear the quantity $F(\boldsymbol{\omega}_t)^\top(\boldsymbol{\omega}_t - \boldsymbol{\omega})$ on the left-hand side,

$$2\eta_t F(\boldsymbol{\omega}_t)^\top(\boldsymbol{\omega}_t - \boldsymbol{\omega}) \leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 + \eta_t^2\|F(\boldsymbol{\omega}_t, \xi_t)\|_2^2 + 2\eta_t(F(\boldsymbol{\omega}_t) - F(\boldsymbol{\omega}_t, \xi_t))^\top(\boldsymbol{\omega}_t - \boldsymbol{\omega}) \tag{106}$$

we can sum (106) for $0 \leq t \leq T-1$ to get,

$$2\sum_{t=0}^{T-1}\eta_t F(\boldsymbol{\omega}_t)^\top(\boldsymbol{\omega}_t - \boldsymbol{\omega}) \leq \sum_{t=0}^{T-1}\Big[(\|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|^2 - \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|^2) + \eta_t^2\|F(\boldsymbol{\omega}_t, \xi_t)\|_2^2 + 2\eta_t\Delta_t^\top(\boldsymbol{\omega}_t - \boldsymbol{\omega})\Big] \tag{107}$$

where we noted $\Delta_t \overset{\text{def}}{=} F(\boldsymbol{\omega}_t) - F(\boldsymbol{\omega}_t, \xi_t)$.
By monotonicity, $F(\boldsymbol{\omega}_t)^\top(\boldsymbol{\omega}_t - \boldsymbol{\omega}) \geq F(\boldsymbol{\omega})^\top(\boldsymbol{\omega}_t - \boldsymbol{\omega})$ we get,

$$2S_T F(\boldsymbol{\omega})^\top(\bar{\boldsymbol{\omega}}_T - \boldsymbol{\omega}) \leq \sum_{t=0}^{T-1}\Big[(\|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|^2 - \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|^2) + \eta_t^2\|F(\boldsymbol{\omega}_t, \xi_t)\|_2^2 + 2\eta_t\Delta_t^\top(\boldsymbol{\omega}_t - \boldsymbol{\omega})\Big]\,, \tag{108}$$

where $S_T \overset{\text{def}}{=} \sum_{t=0}^{T-1}\eta_t$ and $\bar{\boldsymbol{\omega}}_T \overset{\text{def}}{=} \frac{1}{S_T}\sum_{t=0}^{T-1}\eta_t\boldsymbol{\omega}_t$.
We will then upper bound each sum in the right hand side,

$$\Delta_t^\top(\boldsymbol{\omega}_t - \boldsymbol{\omega}) = \Delta_t^\top(\boldsymbol{\omega}_t - \boldsymbol{u}_t) + \Delta_t^\top(\boldsymbol{u}_t - \boldsymbol{\omega})$$

where $\boldsymbol{u}_{t+1} \overset{\text{def}}{=} P_\Omega(\boldsymbol{u}_t - \eta_t\Delta_t)$ and $\boldsymbol{u}_0 = \boldsymbol{\omega}_0$. Then,

$$\|\boldsymbol{u}_{t+1} - \boldsymbol{\omega}\|_2^2 \leq \|\boldsymbol{u}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t\Delta_t^\top(\boldsymbol{u}_t - \boldsymbol{\omega}) + \eta_t^2\|\Delta_t\|_2^2$$

leading to

$$2\eta_t\Delta_t^\top(\boldsymbol{\omega}_t - \boldsymbol{\omega}) \leq 2\eta_t\Delta_t^\top(\boldsymbol{\omega}_t - \boldsymbol{u}_t) + \|\boldsymbol{u}_t - \boldsymbol{\omega}\|_2^2 - \|\boldsymbol{u}_{t+1} - \boldsymbol{\omega}\|_2^2 + \eta_t^2\|\Delta_t\|_2^2 \tag{109}$$

Then noticing that $\boldsymbol{u}_0 \overset{\text{def}}{=} \boldsymbol{\omega}_0$, back to (108) we get a telescoping sum,

$$\begin{aligned}
2S_T F(\boldsymbol{\omega})^\top(\bar{\boldsymbol{\omega}}_T - \boldsymbol{\omega}) &\leq 2\|\boldsymbol{\omega}_0 - \boldsymbol{\omega}\|^2 + \sum_{t=0}^{T-1}\eta_t^2(\|F(\boldsymbol{\omega}_t, \xi_t)\|_2^2 + \|\Delta_t\|_2^2) + 2\sum_{t=0}^{T-1}\eta_t\Delta_t^\top(\boldsymbol{\omega}_t - \boldsymbol{u}_t) \\
&\leq 2R + \sum_{t=0}^{T-1}\eta_t^2(\|F(\boldsymbol{\omega}_t, \xi_t)\|_2^2 + \|\Delta_t\|_2^2) + 2\sum_{t=0}^{T-1}\eta_t\Delta_t^\top(\boldsymbol{\omega}_t - \boldsymbol{u}_t)
\end{aligned}$$

Then the right hand side does not depends on $\boldsymbol{\omega}$, we can maximize over $\boldsymbol{\omega}$ to get,

$$2S_T \operatorname{Err}_R(\bar{\boldsymbol{\omega}}_T) \leq 2R + \sum_{t=0}^{T-1} \eta_t^2(\|F(\boldsymbol{\omega}_t, \xi_t)\|_2^2 + \|\Delta_t\|_2^2) + 2\sum_{t=0}^{T-1} \eta_t \Delta_t^\top(\boldsymbol{\omega}_t - \boldsymbol{u}_t) \tag{110}$$

Noticing that $\mathbb{E}[\Delta_t|\boldsymbol{\omega}_t, \boldsymbol{u}_t] = 0$ (the estimates of F are unbiased), by Assumption 2 $\mathbb{E}[(\|F(\boldsymbol{\omega}_t, \xi_t)\|_2^2] \leq M^2$ and by Assumption 1 $\mathbb{E}[\|\Delta_t\|_2^2] \leq \sigma^2$ we get,

$$\mathbb{E}[\operatorname{Err}_R(\bar{\boldsymbol{\omega}}_T)] \leq \frac{R}{S_T} + \frac{M^2 + \sigma^2}{2S_T} \sum_{t=0}^{T-1} \eta_t^2 \tag{111}$$

particularly for $\eta_t = \eta$ and $\eta_t = \frac{\eta}{\sqrt{t+1}}$ we respectively get,

$$\mathbb{E}[\operatorname{Err}_R(\bar{\boldsymbol{\omega}}_T)] \leq \frac{2R}{\eta T} + \frac{\eta}{2}(M^2 + \sigma^2) \tag{112}$$

and

$$\mathbb{E}[\operatorname{Err}_R(\bar{\boldsymbol{\omega}}_T)] \leq \frac{4R}{\eta\sqrt{T+1} - 1} + 2\eta \ln(T+1) \frac{M^2 + \sigma^2}{\sqrt{T+1} - 1} \tag{113}$$

□

F.2 Proof of Thm.3

Theorem' 3. *Under Assumption 1 and 4, if $\mathbb{E}_\xi[F]$ is L-Lipschitz, then Alg. 2 with a constant step-size $\eta \leq \frac{1}{\sqrt{3}L}$ has the following convergence rate for any $T \geq 1$,*

$$\mathbb{E}[\operatorname{Err}_R(\bar{\boldsymbol{\omega}}_T)] \leq \frac{R^2}{\eta T} + \frac{7}{2}\eta\sigma^2 \quad \textit{where} \quad \bar{\boldsymbol{\omega}}_T \stackrel{\textit{def}}{=} \frac{1}{T}\sum_{t=0}^{T-1} \boldsymbol{\omega}_t'. \tag{114}$$

Particularly, $\eta = \frac{\sqrt{2}R}{\sigma\sqrt{7T}}$ gives $\mathbb{E}[\operatorname{Err}_R(\bar{\boldsymbol{\omega}}_T)] \leq \frac{\sqrt{14}R\sigma}{\sqrt{T}}$.

Proof of Thm.3. Let any $\boldsymbol{\omega} \in \Omega$ such that $\|\boldsymbol{\omega}_0 - \boldsymbol{\omega}\|_2 \leq R$. Then, the update rules become $\boldsymbol{\omega}_{t+1} = P_\Omega(\boldsymbol{\omega}_t - \eta_t F(\boldsymbol{\omega}_t', \zeta_t))$ and $\boldsymbol{\omega}_t' = P_\Omega(\boldsymbol{\omega}_t - \eta F(\boldsymbol{\omega}_t, \xi_t))$. We start by applying Lemma 2 for $(\boldsymbol{\omega}, \boldsymbol{u}, \boldsymbol{\omega}', \boldsymbol{\omega}^+) = (\boldsymbol{\omega}_t, -\eta F(\boldsymbol{\omega}_t', \zeta_t), \boldsymbol{\omega}, \boldsymbol{\omega}_{t+1})$ and $(\boldsymbol{\omega}, \boldsymbol{u}, \boldsymbol{\omega}', \boldsymbol{\omega}^+) = (\boldsymbol{\omega}_t, -\eta_t F(\boldsymbol{\omega}_t, \xi_t), \boldsymbol{\omega}_{t+1}, \boldsymbol{\omega}_t')$,

$$\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 \leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}_t', \zeta_t)^\top(\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}) - \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}_t\|_2^2$$
$$\|\boldsymbol{\omega}_t' - \boldsymbol{\omega}_{t+1}\|_2^2 \leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}_{t+1}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}_t, \xi_t)^\top(\boldsymbol{\omega}_t' - \boldsymbol{\omega}_{t+1}) - \|\boldsymbol{\omega}_t' - \boldsymbol{\omega}_t\|_2^2$$

Then, summing them we get

$$\begin{aligned}\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 \leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 &- 2\eta_t F(\boldsymbol{\omega}_t', \zeta_t)^\top(\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}) \\ &- 2\eta_t F(\boldsymbol{\omega}_t, \xi_t)^\top(\boldsymbol{\omega}_t' - \boldsymbol{\omega}_{t+1}) - \|\boldsymbol{\omega}_t - \boldsymbol{\omega}_t'\|_2^2 - \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}_t'\|_2^2\end{aligned} \tag{115}$$

leading to

$$\begin{aligned}\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 \leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 &- 2\eta_t F(\boldsymbol{\omega}_t', \zeta_t)^\top(\boldsymbol{\omega}_t' - \boldsymbol{\omega}) \\ &+ 2\eta_t(F(\boldsymbol{\omega}_t', \zeta_t) - F(\boldsymbol{\omega}_t, \xi_t))^\top(\boldsymbol{\omega}_t' - \boldsymbol{\omega}_{t+1}) - \|\boldsymbol{\omega}_t - \boldsymbol{\omega}_t'\|_2^2 - \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}_t'\|_2^2\end{aligned}$$

Then with $2\boldsymbol{a}^\top\boldsymbol{b} \leq \|\boldsymbol{a}\|_2^2 + \|\boldsymbol{b}\|_2^2$ we get

$$\begin{aligned}\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 \leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 &- 2\eta_t F(\boldsymbol{\omega}_t', \zeta_t)^\top(\boldsymbol{\omega}_t' - \boldsymbol{\omega}) \\ &- \|\boldsymbol{\omega}_t - \boldsymbol{\omega}_t'\|_2^2 + \eta_t^2\|F(\boldsymbol{\omega}_t', \zeta_t) - F(\boldsymbol{\omega}_t, \xi_t)\|_2^2\end{aligned}$$

Using the inequality $\|\boldsymbol{a} + \boldsymbol{b} + \boldsymbol{c}\|_2^2 \leq 3(\|\boldsymbol{a}\|_2^2 + \|\boldsymbol{b}\|_2^2 + \|\boldsymbol{c}\|_2^2)$ we get,

$$\begin{aligned}\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 \leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 &- 2\eta_t F(\boldsymbol{\omega}_t', \zeta_t)^\top(\boldsymbol{\omega}_t' - \boldsymbol{\omega}) - \|\boldsymbol{\omega}_t - \boldsymbol{\omega}_t'\|_2^2 \\ &+ 3\eta_t^2(\|F(\boldsymbol{\omega}_t) - F(\boldsymbol{\omega}_t, \xi_t)\|_2^2 + \|F(\boldsymbol{\omega}_t') - F(\boldsymbol{\omega}_t', \zeta_t)\|_2^2 + \|F(\boldsymbol{\omega}_t') - F(\boldsymbol{\omega}_t)\|_2^2)\end{aligned}$$

Then we can use the L-Lipschitzness of F to get,

$$\|\boldsymbol{\omega}_{t+1}-\boldsymbol{\omega}\|_2^2 \le \|\boldsymbol{\omega}_t-\boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_t,\zeta_t)^\top(\boldsymbol{\omega}'_t-\boldsymbol{\omega}) - \|\boldsymbol{\omega}_t-\boldsymbol{\omega}'_t\|_2^2 \\ + 3\eta_t^2(\|F(\boldsymbol{\omega}_t)-F(\boldsymbol{\omega}_t,\xi_t)\|_2^2 + \|F(\boldsymbol{\omega}'_t)-F(\boldsymbol{\omega}'_t,\zeta_t)\|_2^2 + L^2\|\boldsymbol{\omega}_t-\boldsymbol{\omega}'_t\|_2^2)$$

As we restricted the step-size to $\eta_t \le \frac{1}{\sqrt{3}L}$ we get,

$$2\eta_t F(\boldsymbol{\omega}'_t)^\top(\boldsymbol{\omega}'_t-\boldsymbol{\omega}) \le \|\boldsymbol{\omega}_t-\boldsymbol{\omega}\|_2^2 - \|\boldsymbol{\omega}_{t+1}-\boldsymbol{\omega}\|_2^2 + 2\eta_t(F(\boldsymbol{\omega}'_t)-F(\boldsymbol{\omega}'_t,\zeta_t))^\top(\boldsymbol{\omega}'_t-\boldsymbol{\omega}) \\ + 3\eta_t^2\|F(\boldsymbol{\omega}_t)-F(\boldsymbol{\omega}_t,\xi_t)\|_2^2 + 3\eta_t^2\|F(\boldsymbol{\omega}'_t)-F(\boldsymbol{\omega}'_t,\zeta_t)\|_2^2$$

We get a particular case of (6) so we can use Lemma 6 where $N_t = \|\boldsymbol{\omega}_t-\boldsymbol{\omega}\|_2^2$, $M_1(\boldsymbol{\omega}_t,\xi_t) = 3\|F(\boldsymbol{\omega}_t)-F(\boldsymbol{\omega}_t,\xi_t)\|_2^2$, $M_2(\boldsymbol{\omega}'_t,\zeta_t) = 3\|F(\boldsymbol{\omega}'_t)-F(\boldsymbol{\omega}'_t,\zeta_t)\|_2^2$, $\Delta_t = F(\boldsymbol{\omega}'_t)-F(\boldsymbol{\omega}'_t,\zeta_t)$ and $\boldsymbol{z}_t = \boldsymbol{\omega}'_t$. By Assumption 1, $M_1 = M_2 = 3\sigma^2$ and by the fact that $\mathbb{E}[F(\boldsymbol{\omega}'_t) - F(\boldsymbol{\omega}'_t,\zeta_t)\,|\boldsymbol{\omega}'_t,\Delta_0,\dots,\Delta_{t-1}] = \mathbb{E}[\mathbb{E}[F(\boldsymbol{\omega}'_t) - F(\boldsymbol{\omega}'_t,\zeta_t)\,|\boldsymbol{\omega}'_t]|\Delta_0,\dots,\Delta_{t-1}] = 0$ the hypothesis of Lemma 6 hold and we get,

$$\mathbb{E}[\mathrm{Err}_R(\bar{\boldsymbol{\omega}}_T)] \le \frac{R^2}{S_T} + \frac{7\sigma^2}{2S_T}\sum_{t=0}^{T-1}\eta_t^2 \tag{116}$$

□

F.3 Proof of Thm. 4

Theorem' 4. *Under Assumption 1, if $\mathbb{E}_\xi[F]$ is L-Lipschitz, then AvgPastExtraSGD (Alg. 3) with a constant step-size $\eta \le \frac{1}{2\sqrt{3}L}$ has the following convergence rate for any $T \ge 1$,*

$$\mathbb{E}[\mathrm{Err}_R(\bar{\boldsymbol{\omega}}_T)] \le \frac{R^2}{\eta T} + \frac{13}{2}\eta\sigma^2 \quad \textit{where} \quad \bar{\boldsymbol{\omega}}_T \overset{\textit{def}}{=} \frac{1}{T}\sum_{t=0}^{T-1}\boldsymbol{\omega}'_t\,. \tag{117}$$

Particularly, $\eta = \frac{\sqrt{2}R}{\sigma\sqrt{7T}}$ gives $\mathbb{E}[\mathrm{Err}_R(\bar{\boldsymbol{\omega}}_T)] \le \frac{\sqrt{14}R\sigma}{\sqrt{T}}$.

First let us recall the update rule

$$\begin{cases}\boldsymbol{\omega}_{t+1} = P_\Omega[\boldsymbol{\omega}_t - \eta_t F(\boldsymbol{\omega}'_t,\xi_t)] \\ \boldsymbol{\omega}'_{t+1} = P_\Omega[\boldsymbol{\omega}_{t+1} - \eta_{t+1} F(\boldsymbol{\omega}'_t,\xi_t)]\,.\end{cases} \tag{118}$$

Lemma 7. *We have for any $\boldsymbol{\omega} \in \Omega$,*

$$2\eta F(\boldsymbol{\omega}'_t,\xi_t)^\top(\boldsymbol{\omega}'_t-\boldsymbol{\omega}) \le \|\boldsymbol{\omega}_t-\boldsymbol{\omega}\|_2^2 - \|\boldsymbol{\omega}_{t+1}-\boldsymbol{\omega}\|_2^2 - \|\boldsymbol{\omega}'_t-\boldsymbol{\omega}_t\|_2^2 + 3\eta_t^2L^2\|\boldsymbol{\omega}'_{t-1}-\boldsymbol{\omega}'_t\|_2^2 \\ + 3\eta_t^2\left[\|F(\boldsymbol{\omega}'_{t-1},\xi_{t-1}) - F(\boldsymbol{\omega}'_{t-1})\|_2^2 + \|F(\boldsymbol{\omega}'_t) - F(\boldsymbol{\omega}'_t,\xi_t)\|_2^2\right]. \tag{119}$$

Proof. Applying Lemma 2 for $(\boldsymbol{\omega},\boldsymbol{u},\boldsymbol{\omega}^+,\boldsymbol{\omega}') = (\boldsymbol{\omega}_t, -\eta_t F(\boldsymbol{\omega}'_t,\xi_t), \boldsymbol{\omega}_{t+1}, \boldsymbol{\omega})$ and $(\boldsymbol{\omega},\boldsymbol{u},\boldsymbol{\omega}^+,\boldsymbol{\omega}') = (\boldsymbol{\omega}_t, -\eta_t F(\boldsymbol{\omega}'_{t-1},\xi_{t-1}), \boldsymbol{\omega}'_t, \boldsymbol{\omega}_{t+1})$, we get,

$$\|\boldsymbol{\omega}_{t+1}-\boldsymbol{\omega}\|_2^2 \le \|\boldsymbol{\omega}_t-\boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_t,\xi_t)^\top(\boldsymbol{\omega}_{t+1}-\boldsymbol{\omega}) - \|\boldsymbol{\omega}_{t+1}-\boldsymbol{\omega}_t\|_2^2 \tag{120}$$

and

$$\|\boldsymbol{\omega}'_t-\boldsymbol{\omega}_{t+1}\|_2^2 \le \|\boldsymbol{\omega}_t-\boldsymbol{\omega}_{t+1}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_{t-1},\xi_{t-1})^\top(\boldsymbol{\omega}'_t-\boldsymbol{\omega}_{t+1}) - \|\boldsymbol{\omega}'_t-\boldsymbol{\omega}_t\|_2^2\,. \tag{121}$$

Summing (120) and (121) we get,

$$\|\boldsymbol{\omega}_{t+1}-\boldsymbol{\omega}\|_2^2 \le \|\boldsymbol{\omega}_t-\boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_t,\xi_t)^\top(\boldsymbol{\omega}_{t+1}-\boldsymbol{\omega}) \tag{122}$$

$$- 2\eta_t F(\boldsymbol{\omega}'_{t-1},\xi_{t-1})^\top(\boldsymbol{\omega}'_t-\boldsymbol{\omega}_{t+1}) - \|\boldsymbol{\omega}'_t-\boldsymbol{\omega}_t\|_2^2 - \|\boldsymbol{\omega}'_t-\boldsymbol{\omega}_{t+1}\|_2^2 \tag{123}$$

$$= \|\boldsymbol{\omega}_t-\boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_t,\xi_t)^\top(\boldsymbol{\omega}'_t-\boldsymbol{\omega}) - \|\boldsymbol{\omega}'_t-\boldsymbol{\omega}_t\|_2^2 - \|\boldsymbol{\omega}'_t-\boldsymbol{\omega}_{t+1}\|_2^2 \tag{124}$$

$$- 2\eta_t(F(\boldsymbol{\omega}'_{t-1},\xi_{t-1}) - F(\boldsymbol{\omega}'_t,\xi_t))^\top(\boldsymbol{\omega}'_t-\boldsymbol{\omega}_{t+1})\,. \tag{125}$$

Then, we can use the inequality of arithmetic and geometric means $2a^\top b \leq \|a\|_2^2 + \|b\|_2^2$ to get,

$$\begin{aligned}\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 &\leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_t, \xi_t)^\top(\boldsymbol{\omega}'_t - \boldsymbol{\omega}) + \eta_t^2 \|F(\boldsymbol{\omega}'_{t-1}, \xi_{t-1}) - F(\boldsymbol{\omega}'_t, \xi_t)\|_2^2 \\ &\quad + \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_{t+1}\|_2^2 - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2 - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_{t+1}\|_2^2 \qquad (126)\\ &= \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_t, \xi_t)^\top(\boldsymbol{\omega}'_t - \boldsymbol{\omega}) \qquad (127)\\ &\quad + \eta_t^2 \|F(\boldsymbol{\omega}'_{t-1}, \xi_{t-1}) - F(\boldsymbol{\omega}'_t, \xi_t)\|_2^2 - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2 \,. \qquad (128)\end{aligned}$$

Using the inequality $\|\boldsymbol{a} + \boldsymbol{b} + \boldsymbol{c}\|_2^2 \leq 3(\|\boldsymbol{a}\|_2^2 + \|\boldsymbol{b}\|_2^2 + \|\boldsymbol{c}\|_2^2)$ we get,

$$\begin{aligned}\|F(\boldsymbol{\omega}'_{t-1}, \xi_{t-1}) - F(\boldsymbol{\omega}'_t, \xi_t)\|_2^2 &\leq 3(\|F(\boldsymbol{\omega}'_{t-1}, \xi_{t-1}) - F(\boldsymbol{\omega}'_{t-1})\|_2^2 + \|F(\boldsymbol{\omega}'_{t-1}) - F(\boldsymbol{\omega}'_t)\|_2^2 \\ &\quad + \|F(\boldsymbol{\omega}'_t) - F(\boldsymbol{\omega}'_t, \xi_t)\|_2^2) \qquad (129)\\ &\leq 3(\|F(\boldsymbol{\omega}'_{t-1}, \xi_{t-1}) - F(\boldsymbol{\omega}'_{t-1})\|_2^2 + L^2\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 \\ &\quad + \|F(\boldsymbol{\omega}'_t) - F(\boldsymbol{\omega}'_t, \xi_t)\|_2^2) \,, \qquad (130)\end{aligned}$$

where we used the L-Lipschitzness of F for the last inequality.

Combining (128) with (130) we get,

$$\begin{aligned}\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 &\leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_t, \xi_t)^\top(\boldsymbol{\omega}'_t - \boldsymbol{\omega}) - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2 + 3\eta_t^2 L^2 \|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 \\ &\quad + 3\eta_t^2 \left[\|F(\boldsymbol{\omega}'_{t-1}, \xi_{t-1}) - F(\boldsymbol{\omega}'_{t-1})\|_2^2 + \|F(\boldsymbol{\omega}'_t) - F(\boldsymbol{\omega}'_t, \xi_t)\|_2^2\right] . \qquad (131)\end{aligned}$$

□

Lemma 8. *For all $t \geq 0$, if we set $\boldsymbol{\omega}'_{-2} = \boldsymbol{\omega}'_{-1} = \boldsymbol{\omega}'_0$ we have*

$$\begin{aligned}\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 &\leq 4\|\boldsymbol{\omega}_t - \boldsymbol{\omega}'_t\|_2^2 + 12\eta_{t-1}^2\big(\|F(\boldsymbol{\omega}'_{t-1}, \xi_{t-1}) - F(\boldsymbol{\omega}'_{t-1})\|_2^2 + L^2\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_{t-2}\|_2^2 \\ &\quad + \|F(\boldsymbol{\omega}'_{t-2}) - F(\boldsymbol{\omega}'_{t-2}, \xi_{t-2})\|_2^2)\big) - \|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 \,. \qquad (132)\end{aligned}$$

Proof. We start with $\|a + b\|_2^2 \leq 2\|a\|^2 + 2\|b\|^2$.

$$\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 \leq 2\|\boldsymbol{\omega}_t - \boldsymbol{\omega}'_t\|_2^2 + 2\|\boldsymbol{\omega}_t - \boldsymbol{\omega}'_{t-1}\|_2^2 \,. \qquad (133)$$

Moreover, since the projection is contractive we have that

$$\begin{aligned}\|\boldsymbol{\omega}_t - \boldsymbol{\omega}'_{t-1}\|_2^2 &\leq \|\boldsymbol{\omega}_{t-1} - \eta_{t-1}F(\boldsymbol{\omega}'_{t-1}, \xi_{t-1}) - \boldsymbol{\omega}_{t-1} - \eta_{t-1}F(\boldsymbol{\omega}'_{t-2}, \xi_{t-2})\|_2^2 \qquad (134)\\ &= \eta_{t-1}^2\|F(\boldsymbol{\omega}'_{t-1}, \xi_{t-1}) - F(\boldsymbol{\omega}'_{t-2}, \xi_{t-2})\|_2^2 \qquad (135)\\ &\leq 3\eta_{t-1}^2\big(\|F(\boldsymbol{\omega}'_{t-1}, \xi_{t-1}) - F(\boldsymbol{\omega}'_{t-1})\|_2^2 + L^2\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_{t-2}\|_2^2 \\ &\quad + \|F(\boldsymbol{\omega}'_{t-2}) - F(\boldsymbol{\omega}'_{t-2}, \xi_{t-2})\|_2^2)\big) \,. \qquad (136)\end{aligned}$$

where in the last line we used the same inequality as in (130). Combining (132) and (136) we get,

$$\begin{aligned}\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 &= 2\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 - \|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 \qquad (137)\\ &\leq 4\|\boldsymbol{\omega}_t - \boldsymbol{\omega}'_t\|_2^2 + 4\|\boldsymbol{\omega}_t - \boldsymbol{\omega}'_{t-1}\|_2^2 - \|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 \qquad (138)\\ &\leq 4\|\boldsymbol{\omega}_t - \boldsymbol{\omega}'_t\|_2^2 + 12\eta_{t-1}^2\big(\|F(\boldsymbol{\omega}'_{t-1}, \xi_{t-1}) - F(\boldsymbol{\omega}'_{t-1})\|_2^2 + L^2\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_{t-2}\|_2^2 \\ &\quad + \|F(\boldsymbol{\omega}'_{t-2}) - F(\boldsymbol{\omega}'_{t-2}, \xi_{t-2})\|_2^2)\big) - \|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 \,. \qquad (139)\end{aligned}$$

□

Proof of Theorem 4. Combining Lemma 8 and Lemma 7 we get,

$$\begin{aligned}2\eta_t F(\boldsymbol{\omega}'_t, \xi_t)^\top(\boldsymbol{\omega}'_t - \boldsymbol{\omega}) &\leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 \\ &\quad + 36\eta_t^2\eta_{t-1}^2 L^2\big(\|F(\boldsymbol{\omega}'_{t-1}, \xi_{t-1}) - F(\boldsymbol{\omega}'_{t-1})\|_2^2 + L^2\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_{t-2}\|_2^2 \\ &\quad + \|F(\boldsymbol{\omega}'_{t-2}) - F(\boldsymbol{\omega}'_{t-2}, \xi_{t-2})\|_2^2\big) \\ &\quad - 3\eta_t^2 L^2\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2 + (12\eta_t^2 L^2 - 1)\|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2 \\ &\quad + 3\eta_t^2\left[\|F(\boldsymbol{\omega}'_{t-1}, \xi_{t-1}) - F(\boldsymbol{\omega}'_{t-1})\|_2^2 + \|F(\boldsymbol{\omega}'_t) - F(\boldsymbol{\omega}'_t, \xi_t)\|_2^2\right] . \qquad (140)\end{aligned}$$

Then for $\eta_t \leq \frac{1}{2\sqrt{3}L}$ we have $36\eta_t^2\eta_{t-1}^2L^4 \leq 3\eta_{t-1}^2L^2$,

$$\begin{aligned}
2\eta_t F(\boldsymbol{\omega}'_t)^\top(\boldsymbol{\omega}'_t - \boldsymbol{\omega}) \leq\ & \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 \\
& + 3L^2(\eta_{t-1}^2\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_{t-2}\|_2^2 - \eta_t^2\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_t\|_2^2) \\
& + 2\eta_t(F(\boldsymbol{\omega}'_t) - F(\boldsymbol{\omega}'_t, \xi_t))^\top(\boldsymbol{\omega}'_t - \boldsymbol{\omega}) \\
& + 3\eta_t^2\left[|F(\boldsymbol{\omega}'_{t-2}, \xi_{t-2}) - F(\boldsymbol{\omega}'_{t-2})\|_2^2 + 2\|F(\boldsymbol{\omega}'_{t-1}, \xi_{t-1}) - F(\boldsymbol{\omega}'_{t-1})\|_2^2\right. \\
& \left. + \|F(\boldsymbol{\omega}'_t) - F(\boldsymbol{\omega}'_t, \xi_t)\|_2^2\right]. \qquad (141)
\end{aligned}$$

We can then use Lemma 6 where

$$\begin{aligned}
N_t &= \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 + 3L^3\eta_{t-1}\|\boldsymbol{\omega}'_{t-1} - \boldsymbol{\omega}'_{t-2}\|_2^2, \\
M_1(\boldsymbol{\omega}_t, \xi_t) &= 0 \\
M_2(\boldsymbol{\omega}'_t, \xi_t) &= 3\|F(\boldsymbol{\omega}'_t) - F(\boldsymbol{\omega}'_t, \xi_t)\|_2^2 + 6\|F(\boldsymbol{\omega}'_{t-1}) - F(\boldsymbol{\omega}'_{t-1}, \xi_{t-1})\|_2^2 \\
&\quad + 3\|F(\boldsymbol{\omega}'_{t-2}) - F(\boldsymbol{\omega}'_{t-2}, \xi_{t-2})\|_2^2 \\
\Delta_t &= F(\boldsymbol{\omega}'_t) - F(\boldsymbol{\omega}'_t, \xi_t) \\
\boldsymbol{z}_t &= \boldsymbol{\omega}'_t\,.
\end{aligned}$$

By Assumption 1, $M_2 = 12\sigma^2$ and by the fact that $\mathbb{E}[F(\boldsymbol{\omega}'_t) - F(\boldsymbol{\omega}'_t, \xi_t)\,|\boldsymbol{\omega}'_t, \Delta_0, \ldots, \Delta_{t-1}] = \mathbb{E}[\mathbb{E}[F(\boldsymbol{\omega}'_t) - F(\boldsymbol{\omega}'_t, \xi_t)\,|\boldsymbol{\omega}'_t]|\Delta_0, \ldots, \Delta_{t-1}] = 0$ the hypothesis of Lemma 6 hold and we get,

$$\mathbb{E}[\mathrm{Err}_R(\bar{\boldsymbol{\omega}}_T)] \leq \frac{R^2}{S_T} + \frac{13\sigma^2}{2S_T}\sum_{t=0}^{T-1}\eta_t^2 \qquad (142)$$

□

F.4 Proof of Theorem 5

Theorem 5 has been introduced in §D. This theorem is about Algorithm 5 which consists in another way to implement extrapolation to SGD. Let us first restate this theorem,

Theorem' 5. *Assume that $\|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_0\| \leq R$, $\forall t \geq 0$ where $(\boldsymbol{\omega}'_t)_{t\geq 0}$ are the iterates of Alg. 5. Under Assumption 1 and 4, for any $T \geq 1$, Alg. 5 with constant step-size $\eta \leq \frac{1}{\sqrt{2}L}$ has the following convergence properties:*

$$\mathbb{E}[\mathrm{Err}_R(\bar{\boldsymbol{\omega}}_T)] \leq \frac{R^2}{\eta T} + \eta\frac{\sigma^2 + 4L^2(4R^2 + \sigma^2)}{2} \quad \textit{where} \quad \bar{\boldsymbol{\omega}}_T \overset{\textit{def}}{=} \frac{1}{T}\sum_{t=0}^{T-1}\boldsymbol{\omega}'_t\,.$$

Particularly, $\eta_t = \frac{\eta}{\sqrt{T}}$ gives $\mathbb{E}[\mathrm{Err}_R(\bar{\boldsymbol{\omega}}_T)] \leq \frac{O(1)}{\sqrt{T}}$.

Proof of Thm.5. Let any $\boldsymbol{\omega} \in \Omega$ such that $\|\boldsymbol{\omega}_0 - \boldsymbol{\omega}\|_2 \leq R$. Then, the update rules become $\boldsymbol{\omega}_{t+1} = P_\Omega(\boldsymbol{\omega}_t - \eta_t F(\boldsymbol{\omega}'_t, \xi_t))$ and $\boldsymbol{\omega}'_t = P_\Omega(\boldsymbol{\omega}_t - \eta F(\boldsymbol{\omega}_t, \xi_t))$. We start the same way as the proof of Thm. 3 by applying Lemma 2 for $(\boldsymbol{\omega}, \boldsymbol{u}, \boldsymbol{\omega}', \boldsymbol{\omega}^*) = (\boldsymbol{\omega}_t, -\eta F(\boldsymbol{\omega}'_t, \xi_t), \boldsymbol{\omega}, \boldsymbol{\omega}_{t+1})$ and $(\boldsymbol{\omega}, \boldsymbol{u}, \boldsymbol{\omega}', \boldsymbol{\omega}^+) = (\boldsymbol{\omega}_t, -\eta_t F(\boldsymbol{\omega}_t, \xi_t), \boldsymbol{\omega}_{t+1}, \boldsymbol{\omega}'_t)$,

$$\begin{aligned}
\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 &\leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_t, \xi_t)^\top(\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}) - \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}_t\|_2^2 \\
\|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_{t+1}\|_2^2 &\leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}_{t+1}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}_t, \xi_t)^\top(\boldsymbol{\omega}'_t - \boldsymbol{\omega}_{t+1}) - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2
\end{aligned}$$

Then, summing them we get

$$\begin{aligned}
\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 \leq\ & \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_t, \xi_t)^\top(\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}) \\
& - 2\eta_t F(\boldsymbol{\omega}_t, \xi_t)^\top(\boldsymbol{\omega}'_t - \boldsymbol{\omega}_{t+1}) - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2 - \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}'_t\|_2^2 \qquad (143)
\end{aligned}$$

leading to

$$\begin{aligned}
\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 \leq\ & \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}'_t, \xi_t)^\top(\boldsymbol{\omega}'_t - \boldsymbol{\omega}) \\
& + 2\eta_t(F(\boldsymbol{\omega}'_t, \xi_t) - F(\boldsymbol{\omega}_t, \xi_t))^\top(\boldsymbol{\omega}'_t - \boldsymbol{\omega}_{t+1}) - \|\boldsymbol{\omega}'_t - \boldsymbol{\omega}_t\|_2^2 - \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}'_t\|_2^2
\end{aligned}$$

Then with $2\boldsymbol{a}^\top \boldsymbol{b} \leq \|\boldsymbol{a}\|_2^2 + \|\boldsymbol{b}\|_2^2$ we get

$$\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 \leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}_t', \xi_t)^\top (\boldsymbol{\omega}_t' - \boldsymbol{\omega}) + \eta_t^2 \|F(\boldsymbol{\omega}_t', \xi_t) - F(\boldsymbol{\omega}_t, \xi_t)\|_2^2 - \|\boldsymbol{\omega}_t' - \boldsymbol{\omega}_t\|_2^2$$

Using the Lipschitzness assumption we get

$$\|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 \leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - 2\eta_t F(\boldsymbol{\omega}_t', \xi_t)^\top (\boldsymbol{\omega}_t' - \boldsymbol{\omega}) + (\eta_t^2 L^2 - 1)\|\boldsymbol{\omega}_t - \boldsymbol{\omega}_t'\|_2^2$$

Then we add $2\eta_t F(\boldsymbol{\omega}_t')^\top (\boldsymbol{\omega}_t' - \boldsymbol{\omega})$ in both sides to get,

$$\begin{aligned} 2\eta_t F(\boldsymbol{\omega}_t')^\top (\boldsymbol{\omega}_t' - \boldsymbol{\omega}) \leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 \\ - 2\eta_t (F(\boldsymbol{\omega}_t', \xi_t) - F(\boldsymbol{\omega}_t'))^\top (\boldsymbol{\omega}_t' - \boldsymbol{\omega}) + (\eta_t^2 L^2 - 1)\|\boldsymbol{\omega}_t - \boldsymbol{\omega}_t'\|_2^2 \end{aligned} \quad (144)$$

Here, unfortunately we cannot use Lemma 6 because $F(\boldsymbol{\omega}_t', \xi_t)$ is biased. We will then deal with the quantity $A = (F(\boldsymbol{\omega}_t', \xi_t) - F(\boldsymbol{\omega}'))^\top (\boldsymbol{\omega}_t' - \boldsymbol{\omega})$. We have that,

$$\begin{aligned} A &= (F(\boldsymbol{\omega}_t', \xi_t) - F(\boldsymbol{\omega}_t, \xi_t))^\top (\boldsymbol{\omega} - \boldsymbol{\omega}_t') + (F(\boldsymbol{\omega}_t) - F(\boldsymbol{\omega}_t'))^\top (\boldsymbol{\omega} - \boldsymbol{\omega}_t') \\ &\quad + (F(\boldsymbol{\omega}_t, \xi_t) - F(\boldsymbol{\omega}_t))^\top (\boldsymbol{\omega}_t - \boldsymbol{\omega}_t') + (F(\boldsymbol{\omega}_t, \xi_t) - F(\boldsymbol{\omega}_t))^\top (\boldsymbol{\omega} - \boldsymbol{\omega}_t) \\ &\leq 2L\|\boldsymbol{\omega}_t' - \boldsymbol{\omega}_t\|_2 \|\boldsymbol{\omega}_t' - \boldsymbol{\omega}\|_2 + \|F(\boldsymbol{\omega}_t, \xi_t) - F(\boldsymbol{\omega}_t)\| \|\boldsymbol{\omega}_t' - \boldsymbol{\omega}_t\|_2 \\ &\quad + (F(\boldsymbol{\omega}_t, \xi_t) - F(\boldsymbol{\omega}_t))^\top (\boldsymbol{\omega} - \boldsymbol{\omega}_t) \\ &\qquad \text{(Using Cauchy-Schwarz and the } L\text{-Lip of } F) \end{aligned}$$

Then using $2\|a\|\|b\| \leq \delta \|a\|_2^2 + \frac{1}{\delta}\|b\|_2^2$, for $\delta = 4$,

$$\begin{aligned} -2\eta_t (F(\boldsymbol{\omega}_t', \xi_t) - F(\boldsymbol{\omega}_t'))^\top (\boldsymbol{\omega}_t' - \boldsymbol{\omega}) \leq \frac{1}{2}\|\boldsymbol{\omega}_t' - \boldsymbol{\omega}_t\|^2 + 8\eta_t^2 L^2 \|\boldsymbol{\omega}_t' - \boldsymbol{\omega}\|_2^2 \\ + 4\eta_t^2 \|F(\boldsymbol{\omega}_t, \xi_t) - F(\boldsymbol{\omega}_t)\|_2^2 \\ + \frac{1}{4}\|\boldsymbol{\omega}_t' - \boldsymbol{\omega}_t\|_2^2 + 2\eta_t (F(\boldsymbol{\omega}_t, \xi_t) - F(\boldsymbol{\omega}_t))^\top (\boldsymbol{\omega} - \boldsymbol{\omega}_t) \end{aligned}$$

leading to,

$$\begin{aligned} 2\eta_t F(\boldsymbol{\omega}_t')^\top (\boldsymbol{\omega}_t' - \boldsymbol{\omega}) \leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 + 2\eta_t (F(\boldsymbol{\omega}_t, \xi_t) - F(\boldsymbol{\omega}_t))^\top (\boldsymbol{\omega} - \boldsymbol{\omega}_t) \\ + (\eta_t^2 L^2 - \tfrac{1}{4})\|\boldsymbol{\omega}_t - \boldsymbol{\omega}_t'\|_2^2 \\ + 4\eta_t^2 (2L^2 \|\boldsymbol{\omega}_t' - \boldsymbol{\omega}\|_2^2 + \|F(\boldsymbol{\omega}_t, \xi_t) - F(\boldsymbol{\omega}_t)\|_2^2) \end{aligned}$$

If one assumes finally that $\|\boldsymbol{\omega}_t' - \boldsymbol{\omega}_0\|_2 \leq R$ (assumption of the theorem) and that $\eta_t \leq \frac{1}{2L}$ we get,

$$\begin{aligned} 2\eta_t F(\boldsymbol{\omega}_t')^\top (\boldsymbol{\omega}_t' - \boldsymbol{\omega}) \leq \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2 - \|\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}\|_2^2 + 2\eta_t (F(\boldsymbol{\omega}_t, \xi_t) - F(\boldsymbol{\omega}_t))^\top (\boldsymbol{\omega} - \boldsymbol{\omega}_t) \\ + 4\eta_t^2 (4L^2 R^2 + \|F(\boldsymbol{\omega}_t, \xi_t) - F(\boldsymbol{\omega}_t)\|_2^2) \end{aligned}$$

where we used that $\|\boldsymbol{\omega}_t' - \boldsymbol{\omega}\|_2 \leq \|\boldsymbol{\omega}_t' - \boldsymbol{\omega}_0\|_2 + \|\boldsymbol{\omega}_0 - \boldsymbol{\omega}\|_2 \leq 2R$. Once again this equation is a particular case of Lemma 6 where $N_t = \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_2^2$, $M_1(\boldsymbol{\omega}_t, \xi_t) = 4(4L^2R^2 + \|F(\boldsymbol{\omega}_t, \xi_t) - F(\boldsymbol{\omega}_t)\|_2^2)$, $M_2(\boldsymbol{\omega}_t', \zeta_t) = 0$, $\boldsymbol{z}_t = \boldsymbol{\omega}_t$ and $\Delta_t = F(\boldsymbol{\omega}_t, \xi_t) - F(\boldsymbol{\omega}_t)$. By Assumption 1 $\mathbb{E}[M_1(\boldsymbol{\omega}_t, \xi_t)] \leq 16L^2R^2 + 4\sigma^2$ and $\mathbb{E}[\Delta_t | \boldsymbol{\omega}_t, \Delta_0, \ldots, \Delta_{t-1}] = \mathbb{E}[\mathbb{E}[\Delta_t | \boldsymbol{\omega}_t] | \Delta_0, \ldots, \Delta_{t-1}] = 0$ so we can use Lemma 6 and get,

$$\mathbb{E}[\mathrm{Err}_R(\bar{\boldsymbol{\omega}}_T)] \leq \frac{R^2}{S_T} + \frac{\sigma^2 + 16L^2R^2 + 4\sigma^2}{2S_T} \sum_{t=0}^{T-1} \eta_t^2 \,. \quad (145)$$

□

G ADDITIONAL EXPERIMENTAL RESULTS

G.1 TOY NON-CONVEX GAN (2D AND DETERMINISTIC)

We now consider a task similar to (Mescheder et al., 2018) where the discriminator is linear $D_{\boldsymbol{\varphi}}(\boldsymbol{\omega}) = \boldsymbol{\varphi}^T \boldsymbol{\omega}$, the generator is a Dirac distribution at $\boldsymbol{\theta}$, $q_{\boldsymbol{\theta}} = \delta_{\boldsymbol{\theta}}$ and the distribution we try to match is also a Dirac at $\boldsymbol{\omega}^*$, $p = \delta_{\boldsymbol{\omega}^*}$. The minimax formulation from Goodfellow et al. (2014) gives:

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\varphi}} -\log\left(1 + e^{-\boldsymbol{\varphi}^T \boldsymbol{\omega}^*}\right) - \log\left(1 + e^{\boldsymbol{\varphi}^T \boldsymbol{\theta}}\right) \tag{146}$$

Note that as observed by Nagarajan and Kolter (2017), this objective is concave-concave, making it hard to optimize. We compare the methods on this objective where we take $\boldsymbol{\omega}^* = -2$, thus the position of the equilibrium is shifted towards the position $(\boldsymbol{\theta}, \boldsymbol{\varphi}) = (-2, 0)$. The convergence and the gradient vector field are shown in Figure 5. We observe that depending on the initialization some methods can fail to converge but *extrapolation* (18) seem to perform better than the rest of the methods.

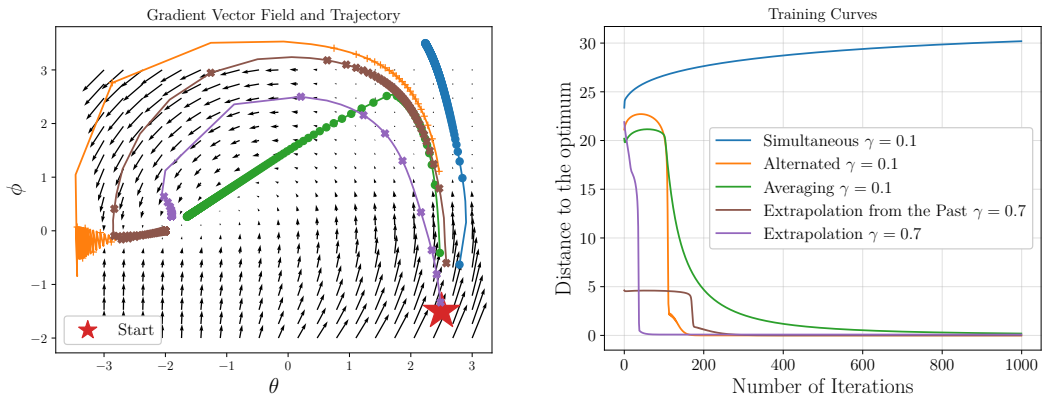


Figure 5: Comparison of five algorithms (described in Section 3) on the non-convex gan objective (146), using the optimal step-size for each method. **Left**: The gradient vector field and the dynamics of the different methods. **Right**:The distance to the optimum as a function of the number of iterations for each methods.

G.2 DCGAN WITH WGAN-GP OBJECTIVE

We also trained the DCGAN architecture of the WGAN experiments but this time with the WGAN-GP objective. The results are shown in Table 2. The best results are achieved with *uniform averaging* of AltAdam5, However its iterations require to update the discriminator 5 times for every generator update. With a small drop in best final score, ExtraAdam can train WGAN-GP significantly faster (see Fig. 6 right) as the discriminator and generator are updated only twice.

Model	WGAN-GP (DCGAN)	
Method	no averaging	uniform avg
SimAdam	$\mathit{6.00 \pm .07}$	$6.01 \pm .08$
AltAdam5	$\mathit{6.25 \pm .05}$	$\mathbf{6.51 \pm .05}$
ExtraAdam	$6.22 \pm .04$	$6.35 \pm .05$
PastExtraAdam	6.27 ± 0.06	6.23 ± 0.13

Table 2: Best inception scores (averaged over 5 runs) achieved on CIFAR10 for every considered Adam variant. OptimAdam is the related *Optimistic Adam* (Daskalakis et al., 2018) algorithm. We see that the techniques of extrapolation and averaging consistently enable improvements over the baselines (in italic).

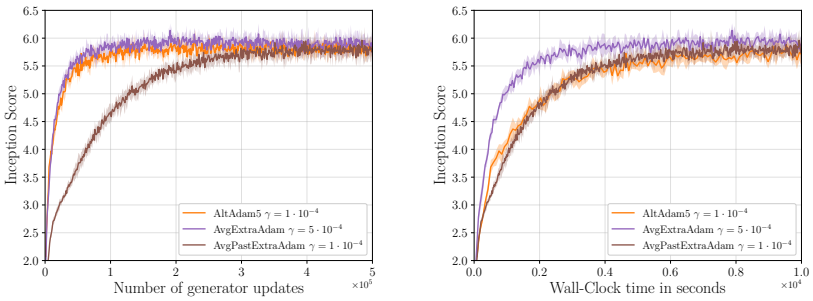


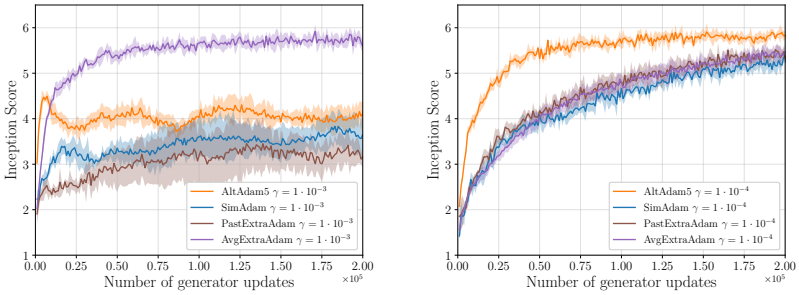
Figure 6: DCGAN architecture with WGAN-GP trained on CIFAR10: mean and standard deviation of the inception score computed over 5 runs for each method using the best performing learning rate plotted over number of generator updates (**Left**) and wall-clock time (**Right**); all experiments were run on a NVIDIA Quadro GP100 GPU. We see that ExtraAdam converges faster than the Adam baselines.

G.3 COMPARISON OF THE METHODS WITH THE SAME LEARNING RATE

In this section we compare how the methods presented in §7 perform with the same step-size. We follow the same protocol as in the experimental section §7. In Figure 7 we plot the inception score provided by each training method as a function of the number of generator updates. Note that these plots advantage **AltAdam5** a bit because each iteration of this algorithm is a bit more costly (since it perform 5 discriminator updates for each generator update). Nevertheless, the goal of this experiment is not to show that **AltAdam5** is faster but to show that that **ExtraAdam** is less sensitive to the choice of learning rate and can be used with higher learning rates with less degradation.

The same way in Figure 8 we compare the sample quality on the WGAN-GP experiment (DCGAN architecture) of **AltAdam5** and **AvgExtraAdam** for different step-sizes. We notice that for **AvgExtraAdam** the sample quality do not significantly change whereas the sample quality of **AltAdam5** seems to be really sensitive to step-size tunning.

We think that robustness to step-size tuning is a key property for an optimization algorithm in order to save as much time as possible to tune other hyperparameters of the learning procedure such as regularization.



(a) learning rate of 10^{-3} (b) learning rate of 10^{-4}

Figure 7: Inception score on CIFAR10 for WGAN-GP (DCGAN) over number of generator updates for different learning rates. We can see that AvgExtraAdam is less sensitive to the choice of learning rate.

(a) AvgExtraAdam with $\eta = 10^{-3}$

(b) AvgExtraAdam with $\eta = 10^{-4}$

(c) AltAdam with $\eta = 10^{-3}$

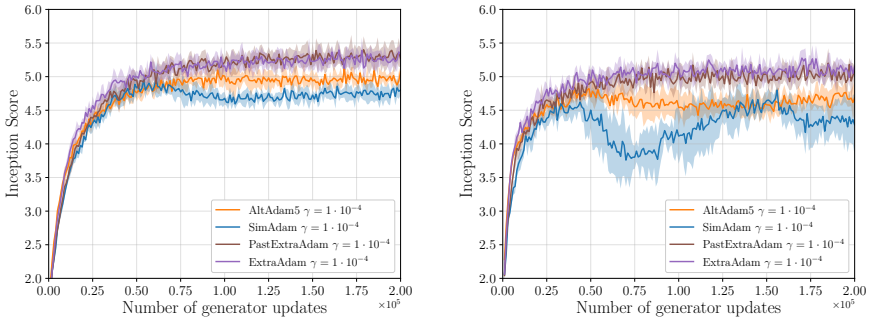
(d) AltAdam with $\eta = 10^{-4}$

Figure 8: Comparison of the samples quality on the WGAN-GP (DCGAN) experiment for different methods and learning rate η.

G.4 COMPARISON OF THE METHODS WITH AND WITHOUT AVERAGING

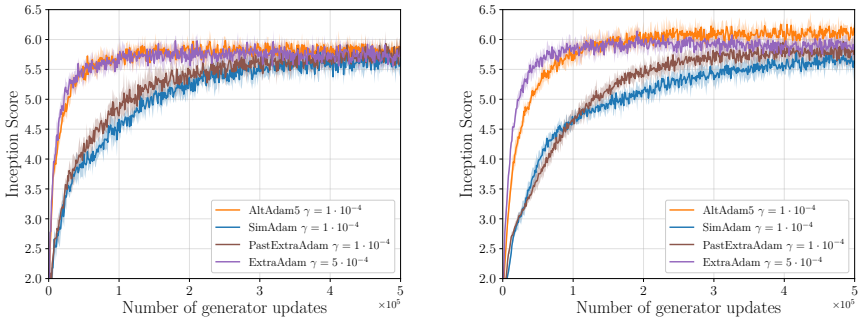
In this section we compare how uniform averaging improve the performance of the methods presented in §7. We follow the same protocol as in the experimental section §7. In Figure 9 and 10, we plot the inception score provided by each training method as a function of the number of generator updates with and without averaging.

We notice that uniform averaging seems to improve the inception score, nevertheless it looks like the sample are a bit more blurry (see Figure 11) this is confirmed by our result (Figure 12) on the Fréchet Inception Distance (FID) which is more sensitive to blurriness.



(a) with averaging (b) without averaging

Figure 9: Inception Score on CIFAR10 for WGAN over number of generator updates with and without averaging. We can see that averaging improve the inception score.



(a) with averaging (b) without averaging

Figure 10: Inception score on CIFAR10 for WGAN-GP (DCGAN) over number of generator updates

(a) PastExtraAdam without averaging

(b) PastExtraAdam with averaging

(c) AltAdam5 without averaging

(d) AltAdam5 with averaging

Figure 11: Comparison of the samples of a WGAN trained with the different methods with and without averaging. Although averaging improves the inception score, the samples seem more blurry

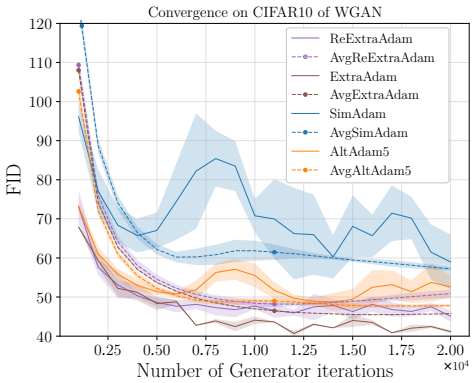


Figure 12: The Fréchet Inception Distance (FID) from Heusel et al. (2017) computed using 50,000 samples, on the WGAN experiments. ReExtraAdam refers to Alg. 5 introduced in §D. We can see that averaging performs worse than when comparing with the Inception Score. We observed that the samples generated by using averaging are a little more blurry and that the FID is more sensitive to blurriness, thus providing an explanation for this observation.