Two steps at a time — taking GAN training in stride with Tseng's method*

Axel Böhm†, Michael Sedlmayer†,‡, Ernö Robert Csetnek†, and Radu Ioan Boţ†

Abstract. Motivated by the training of Generative Adversarial Networks (GANs), we study methods for solving minimax problems with additional nonsmooth regularizers. We do so by employing *monotone operator* theory, in particular the *Forward-Backward-Forward (FBF)* method, which avoids the known issue of limit cycling by correcting each update by a second gradient evaluation and does so requiring less projection steps compared to the Extragradient method in the presence of constraints. Furthermore, we propose a seemingly new scheme which recycles old gradients to mitigate the additional computational cost. In doing so we rediscover a known method, related to *Optimistic Gradient Descent Ascent (OGDA)*. For both schemes we prove novel convergence rates for convex-concave minimax problems via a unifying approach. The derived error bounds are in terms of the gap function for the ergodic iterates. For the deterministic and the stochastic problem we show a convergence rate of $\mathcal{O}(1/k)$ and $\mathcal{O}(1/\sqrt{k})$, respectively. We complement our theoretical results with empirical improvements in the training of Wasserstein GANs on the CIFAR10 dataset.

Key words. min-max, convex-concave, stochastic gradient, GAN

AMS subject classifications. 65K15, 90C15, 90C47

1. Introduction.

Generative Adversarial Networks (GANs) [19] have proven to be a powerful class of generative models, producing for example unseen realistic images. Two neural networks, called generator and discriminator, compete against each other in a game. In the special case of a zero sum game this task can be formulated as a minimax (aka saddle point) problem.

Conventionally, GANs are trained using variants of (stochastic) *Gradient Descent Ascent (GDA)* which are known to exhibit oscillatory behavior [39] and thus fail to converge even for simple bilinear saddle point problems, see [18]. We therefore propose the use of methods with provable convergence guarantees for (stochastic) convex-concave minimax problems, even though GANs are well known to not warrant these properties. Along similar considerations an adaptation of the *Extragradient method (EG)* [29] for the training of GANs was suggested in [15], whereas [11, 12, 31] studied *Optimistic Gradient Descent Ascent (OGDA)* based on *optimistic mirror descent* [48, 49]. We however investigate the *Forward-Backward-Forward (FBF)* method [55] from monotone operator theory, which uses two gradient evaluations per update, similar to EG, in order to circumvent the aforementioned issues but requires less projection/proximal steps per iteration.

Instead of trying to improve GAN performance with new architectures, loss functions, etc., we want to improve their training by contributing to the study of minimax problems. While the landscape of GAN training is far from matching the rigorous setting of monotonicity, the nonconvex-nonconcave setting remains either intractable in its full generality [13] or other simplifying assumption or other simplifying assumptions have to be made, for example the existence of Minty solutions [37, 32], which

‡Research Platform Data Science @ Uni Vienna

*Submitted to the editors March 22, 2022.

Funding: This project has received funding from the doctoral programme *Vienna Graduate School on Computational Optimization (VGSCO)*, FWF (Austrian Science Fund), project W 1260, as well as project P 29809-N32.

†Faculty of Mathematics, University of Vienna, Austria ({axel.boehm, michael.sedlmayer, robert.csetnek, radu.bot}@univie.ac.at)

are again related to monotonicity.

We also want to point out that while extrapolation / optimistic steps are able to combat some of the oscillatory behaviour of minimax problems, another set of problems arises from the stochastic noise of the gradient evaluations which can be dealt with variance reduction techniques [8, 26] or reducing the stepsize (as we do). Other considerations have proposed to use the same sample in the for the two gradient evaluations leading to improved empirical performance [40]. Another approach was explored in [25] to use very different stepsizes for the descent and ascent steps respectively (something that has been observed to be very beneficial in practice as well). Known methods were outfitted with *negative momentum* to improve stability in [16]. An additional technique proposed to deal with the oscillatory behaviour of minimax problems called *crossing-the-curl* was suggested in [14], whose authors used second order information to take a step perpendicular to the direction of rotation.

While our convergence results are stated in terms of the averaged iterates, a technique which can prove beneficial in practice [9, 15], having guarantees on the last iterate would be more in the spirit of nonconvex methods. Such results have been obtained in [11] and [8], but are for bilinear and strongly-convex-strongly-concave problems, respectively, which are known to allow for the derivation of better convergence rates. On the other hand the works [17, 33, 20] have been able to guarantee rates of convergence for the last iterate for convex-concave problems, but all of them only in the unconstrained setting. Furthermore [17] uses additional smoothness assumptions, [33] employs a second-order method and [20] uses the norm of the gradient as the measure of optimality.

Contribution. Establishing the connection between GAN training and *monotone inclusions* motivates to use the FBF method, originally designed to solve this type of problems. This approach allows to naturally extend the constrained setting to a regularized one making use of the proximal operator.

We also propose a variant of FBF reusing previous gradients to reduce the computational cost per iteration, which turns out to be a known method, related to OGDA. By developing a unifying scheme that captures FBF and a generalization of OGDA, we reveal a hitherto unknown connection. Using this approach we prove novel nonasymptotic convergence statements in terms of the minimax gap for both methods in the context of saddle point problems. In the deterministic and stochastic setting we obtain rates of $\mathcal{O}(1/k)$ and $\mathcal{O}(1/\sqrt{k})$, respectively. Concluding, we highlight the relevance of our proposed method as well as the role of regularizers by showing empirical improvements in the training of Wasserstein GANs on the CIFAR10 dataset.

Organization. This paper is structured as follows. In section 2 we highlight the connection of GAN training and monotone inclusions and give an extensive review of methods with convergence guarantees for the latter. The main results as well as a precise definition of the measure of optimality are discussed in section 3. Concluding, section 4 illustrates the empirical performance in the training of GANs as well as solving bilinear problems.

2. GAN training as monotone inclusion.

The GAN objective was originally cast as a two-player zero-sum game between the discriminator D_y and the generator G_x [19] given by

$$\min_x \max_y \mathbb{E}_{\rho\sim q}[\log(D_y(\rho))] + \mathbb{E}_{\zeta\sim p}[\log(1 - D_y(G_x(\zeta)))],$$

exhibiting the aforementioned minimax structure. Due to problems with vanishing gradients in the training of such models, a successful alternative formulation called *Wasserstein GAN (WGAN)* [1] has been proposed. In this case the minimization tries to reduce the Wasserstein distance between the true distribution q and the one learned by the generator. Reformulating this distance via the Kantorovich Rubinstein duality leads to an inner maximization over 1-Lipschitz functions which are approximated via neural networks, yielding the saddle point problem

$$\min_{x} \max_{y:\|D_y\|_{\mathrm{Lip}}\leq 1} \mathbb{E}_{\rho\sim q}[D_y(\rho)] - \mathbb{E}_{\zeta\sim p}[D_y(G_x(\zeta))].$$

2.1. Convex-concave minimax problems. Due to the observations made in the previous paragraph we study the following abstract minimax problem

$$\min_{x\in\mathbb{R}^d} \max_{y\in\mathbb{R}^n} \Psi(x,y) := f(x) + \mathbb{E}_{\xi\sim Q}\left[\Phi(x,y;\xi)\right] - h(y), \tag{2.1}$$

where the convex-concave coupling function $\Phi(x,y) := \mathbb{E}_{\xi\sim Q}\left[\Phi(x,y;\xi)\right]$, which hides the stochasticity for ease of notation, is differentiable with L-Lipschitz continuous gradient. The proper, convex and lower semicontinuous functions $f : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ and $h : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ act as regularizers. A solution of (2.1) is given by a so-called *saddle point* (x^*, y^*) fulfilling for all x and y

$$\Psi(x^*, y) \leq \Psi(x^*, y^*) \leq \Psi(x, y^*). \tag{2.2}$$

In the context of two-player games this corresponds to a pair of strategies, where no player can be better off by changing just their own strategy.

For the purpose of this motivating section, we will restrict ourselves for now to the special case of the *deterministic constrained* version of (2.1), given by

$$\min_{x\in X} \max_{y\in Y} \Phi(x,y), \tag{2.3}$$

where f and h are given by indicator functions of closed convex sets X and Y, respectively. The indicator function δ_C of a set C is defined as $\delta_C(z) = 0$ for $z \in C$ and $\delta_C(z) = +\infty$ otherwise.

2.2. Minimax problems as monotone inclusions. If the coupling function Φ is convex-concave and differentiable then solving (2.1) is equivalent to solving the first order optimality conditions which can be written as a so-called *monotone inclusion* with $w = (x, y) \in \mathbb{R}^m$ and $m = d + n$, given by

$$0 \in F(w) + N_\Omega(w). \tag{2.4}$$

The entities involved are

$$F(x,y) := (\nabla_x\Phi(x,y), -\nabla_y\Phi(x,y)), \tag{2.5}$$

and the *normal cone* N_Ω of the convex set $\Omega := X \times Y$. The normal cone mapping is given by

$$N_\Omega(w) = \{v \in \mathbb{R}^m : \langle v, w' - w\rangle \leq 0 \quad \forall w' \in \Omega\}, \tag{2.6}$$

for $w \in \Omega$ and $N_\Omega(w) = \emptyset$ for $w \notin \Omega$. Here, the operators F and N_Ω satisfy well known properties from convex analysis [4], in particular the first one is monotone (and Lipschitz if $\nabla\Phi$ is so) whereas the latter one is maximal monotone. We call a, possibly *set-valued*, operator A from $\mathbb{R}^m$ to itself monotone if

$$\langle u-u', z-z'\rangle \geq 0 \quad \forall u \in A(z), u' \in A(z'). \tag{2.7}$$

We say A is maximal monotone, if there exists no monotone operator A' such that the graph of A is properly contained in the graph of A'.

Problems of type (2.4) have been studied thoroughly in convex optimization, with the most established solution methods being *Extragradient* [29] and *Forward-Backward-Forward* [55]. Both methods are known to generate sequences of iterates converging to a solution of (2.4). Note that in the unconstrained setting (i.e. if Ω is the entire space) both of these algorithms even produce the same iterates.

2.3. Solving monotone inclusions. The connection between monotone inclusions and saddle point problems is of course not new. The application of Extragradient (EG) to minimax problems has been studied in the seminal paper [43] under the name of *Mirror Prox* and a convergence rate of $\mathcal{O}(1/k)$ in terms of the function values has been proven. Even a stochastic version of the Mirror Prox algorithm has been studied in [27] with a convergence rate of $\mathcal{O}(1/\sqrt{k})$. Applied to problem (2.4), with P_Ω being the projection onto Ω, it iterates

$$\text{EG:} \left\lfloor \begin{array}{l} w_k = P_\Omega[z_k - \alpha_k F(z_k)] \\ z_{k+1} = P_\Omega[z_k - \alpha_k F(w_k)]. \end{array} \right. \tag{2.8}$$

The Forward-Backward-Forward (FBF) method, introduced and convergence of the iterates established in [55], has not been studied rigorously for minimax problems in terms of function values yet, despite promising applications in [7] and *its advantage of it only requiring one projection*, whereas EG needs two. It is given by

$$\text{FBF:} \left\lfloor \begin{array}{l} w_k = P_\Omega[z_k - \alpha_k F(z_k)] \\ z_{k+1} = w_k + \alpha_k(F(z_k) - F(w_k)). \end{array} \right. \tag{2.9}$$

Both, EG and FBF, have the "disadvantage" of needing two gradient evaluations per iteration. A possible remedy — suggested in [15] for EG under the name of *extrapolation from the past* — is to reuse previous gradients. In a similar fashion we consider

$$\text{FBFp:} \left\lfloor \begin{array}{l} w_k = P_\Omega[z_k - \alpha_k F(w_{k-1})] \\ z_{k+1} = w_k + \alpha_k(F(w_{k-1}) - F(w_k)), \end{array} \right. \tag{2.10}$$

where we replaced $F(z_k)$ by $F(w_{k-1})$ twice in (2.9). As a matter of fact, the above method can be written exclusively in terms of the first variable w_k by incrementing the index k in the first update and then substituting in the second line. This results in

$$w_{k+1} = P_\Omega\Big[w_k - \alpha_{k+1}F(w_k) + \alpha_k(F(w_{k-1}) - F(w_k))\Big]. \tag{2.11}$$

This way we rediscover a known method which was studied in [36] and convergence of the iterates established for general monotone inclusions under the name of *forward-reflected-backward*. It reduces to *optimistic mirror descent* [48, 49] in the unconstrained case with constant step size $\alpha_k = \alpha$, giving

$$w_{k+1} = w_k - \alpha(2F(w_k) - F(w_{k-1})) \tag{2.12}$$

which has been proposed for the training of GANs under the name of *Optimistic Gradient Descent Ascent (OGDA)*, see [11, 12, 31]. Only recently a method very related to (2.11) was proposed in [10] and is characterized by applying the correction term after the projection

$$w_{k+1} = P_\Omega\Big[w_k - \alpha F(w_k)\Big] - \alpha(F(w_k) - F(w_{k-1})). \tag{2.13}$$

Evidently, in the unconstrained case and for constant stepsize the methods (2.10), (2.11), (2.12) and (2.13) are all equivalent.

All of the above methods and extensions rely solely on the monotone operator formulation of the saddle point problem where the two components x and y play a symmetric role. Taking the special minimax structure into consideration, [22] showed convergence of a method that uses an optimistic step (2.12) in one component and a regular gradient step in the other, thus requiring less storing of past gradients in comparison to (2.11).

On the downside, however, by reducing the number of required gradient evaluations per iteration, the largest possible step size is reduced from $1/L$ (see [29] or section 3) to $1/2L$ (see [15, 36, 35] or section 3). To summarize, the number of required gradient evaluations is halved, but so is the step size, resulting in no clear net gain.

2.4. Regularizers. The role of regularizers is well studied in many fields such as statistics [54], signal processing [45] or inverse problems [52]. They serve different purposes such as inducing sparsity in the solution or conditioning of the problem. In the context of deep learning this has been explored from different perspectives, e.g. in incremental convex neural networks where neurons with zero weights are removed from the network and new ones are inserted according to different policies, see [2, 5, 51, 46]. Other examples include the box-constraints for WGANs with weight clipping (see [1]) or spectral normalization (see [41]) which has so far rather been considered as part of the architecture, but could at the same time be seen as a regularization step or as a projection onto the set of matrices with spectral norm less than 1 (again not rigorously).

In the framework of monotone operator theory the optimality condition of the regularized minimax problem (2.1) can be written as

$$0 \in F(w) + \partial r(w), \tag{2.14}$$

where r is given by $(x, y) \mapsto f(x) + h(y)$. The possibly set-valued operator ∂r denotes the subdifferential of r and is given by

$$\partial r(w) := \{v \in \mathbb{R}^m : \langle v, w' - w\rangle + r(w) \le r(w') \quad \forall w' \in \mathbb{R}^m\}. \tag{2.15}$$

The monotone inclusion (2.14) generalizes (2.4) in a natural way, since $N_\Omega = \partial\delta_\Omega$. Similarly, the projection constitutes a special case of the so-called *proximal mapping* which for the function r and $\lambda > 0$ is given by

$$\operatorname{prox}_{\lambda r}(w) := \operatorname*{arg\,min}_{w' \in \mathbb{R}^m} \Big\{ r(w') + \frac{1}{2\lambda}\|w' - w\|^2 \Big\}. \tag{2.16}$$

In particular, the proximal mapping of the indicator δ_Ω yields the projection onto the set Ω, i.e. $\operatorname{prox}_{\lambda\delta_\Omega} = P_\Omega$.

3. Main results.

Motivated by the considerations above we study the inclusion problem

$$0 \in F(w) + \partial r(w), \tag{3.1}$$

where $F : \mathbb{R}^m \to \mathbb{R}^m$ is a monotone and Lipschitz operator and $r : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is a proper convex lower semicontinuous function.

3.1. Measure of optimality.

There are two common quantities measuring the quality of a point with respect to the monotone inclusion (2.14). The most natural one is the distance to the solution set for which typically only asymptotic convergence can be proved. If F arises from a saddle point problem (2.1) meaning that F has the form (2.5), we want to use a more problem specific measure, the *minimax gap*, which for a point $w = (u, v) \in \mathbb{R}^d \times \mathbb{R}^n$ is given by

$$\sup_{y\in\mathbb{R}^n} \Psi(u, y) - \inf_{x\in\mathbb{R}^d} \Psi(x, v)\Big(= \sup_{x\in\mathbb{R}^d, y\in\mathbb{R}^n} \Psi(u, y) - \Psi(x, v)\Big). \tag{3.2}$$

This minimax gap can be interpreted from a game theoretic standpoint as the sum of the maximal payoffs achievable by the two players by playing their respective best responses, given the current strategy of the opponent. In the more general monotone inclusion setting where no function values are available, an appropriate generalization of (3.2) is given for any $w \in \mathbb{R}^m$ by

$$\sup_{z\in\mathbb{R}^m} \langle F(z), w - z\rangle + r(w) - r(z). \tag{3.3}$$

If r is the indicator δ_Ω of the compact and convex set Ω it is clear that the supremum is only taken over $z \in \Omega$ and will thus be finite.

The restricted gap. Since the problem (3.1) is in general unconstrained and the supremum can be infinite we consider instead, as done for example in [44], the *restricted* gap where the above supremum is taken over an auxiliary compact set $B \subset \mathbb{R}^m$ instead of the entire space. Note that the restricted gap is in general only a reasonable measure of optimality for elements of B. It is nonnegative on B and zero for points of B which solve (3.1). Additionally we want to be able to conclude that if a point w^* has zero gap it solves (3.1). This is for example the case if w^* is in the interior of B, which can always be ensured if B is chosen large enough.

In order to capture both at the same time we define the following unifying gap

$$G_B(w) := \begin{cases} \sup_{(x,y)\in B} \Psi(u, y) - \Psi(x, v) & \text{if } F \text{ and } r \text{ come from (2.1)} \\ \sup_{z\in B} \langle F(z), w - z\rangle + r(w) - r(z) & \text{otherwise.} \end{cases} \tag{3.4}$$

3.2. Methods.

We now present a novel unifying scheme for solving problem (3.1), which generalizes FBF (2.9) and in addition recovers the method motivated in (2.10) as FBFp. Let us point out again that the latter algorithm was already introduced in [36] and corresponds to OGDA [48, 11, 12] if F stems from the minimax setting (2.5).

Algorithm 3.1 (generalized FBF). *For a starting point $z_0 \in \mathbb{R}^m$ and step sizes $\alpha_k > 0$ we consider for all $k \geq 0$*

$$\left\lfloor \begin{array}{l} w_k = \operatorname{prox}_{\alpha_k r}(z_k - \alpha_k F(\diamondsuit_k)) \\ z_{k+1} = w_k + \alpha_k(F(\diamondsuit_k) - F(w_k)). \end{array} \right. \tag{3.5}$$

For $\diamondsuit_k = z_k$ this reduces to the well known FBF method, whereas $\diamondsuit_k = w_{k-1}$, with the additional initial condition $w_{-1} = z_0$, recycles previous gradients (FBFp).

Consider the scenario where F is given as an expectation $\mathbb{E}_\xi[F(\cdot\,;\xi)]$, e.g. coming from (2.1), and only a stochastic estimator $F(\cdot\,;\xi)$ is accessible instead of F itself. In this case we adapt Algorithm 3.1 in the following way.

Algorithm 3.2 (generalized stochastic FBF). *For a starting point $z_0 \in \mathbb{R}^m$ and step sizes $\alpha_k > 0$ we consider for all $k \geq 0$*

$$\left\lfloor \begin{array}{l} \xi_k \sim Q \quad (\textit{optionally } \eta_k \sim Q) \\ w_k = \operatorname{prox}_{\alpha_k r}\left(z_k - \alpha_k F(\diamondsuit_k; \triangle_k)\right) \\ z_{k+1} = w_k + \alpha_k(F(\diamondsuit_k; \triangle_k) - F(w_k; \xi_k)). \end{array} \right.$$

For $\diamondsuit_k = z_k$ and $\triangle_k = \eta_k$ this results in a stochastic version of FBF, whereas $\diamondsuit_k = w_{k-1}$ and $\triangle_k = \xi_{k-1}$ recycles previous gradients (stochastic FBFp) with the additional initial condition $w_{-1} = z_0$ and $\xi_{-1} = \eta_0$.

Even though both methods encompassed by the unifying scheme Algorithm 3.1 have been studied in the deterministic setting before, the stated convergence results are new. Note that while the rate for FBF is completely new our result for FBFp provides only a generalization of the known rate for OGDA, see [42]. Similarly, the stochastic version of FBF has been considered before in [6] and rates have been obtained, but only in terms of the fixed point residual and not the function values. However, we want to point out that the stochastic version of FBFp has not been considered prior to this work.

3.3. Convergence. Let in the following $B \subset \mathbb{R}^m$ be the compact set of the restricted (unifying) gap function (3.4) with $D := \sup_{w,z\in B} \|z - w\|$ denoting its diameter. For convenience in the estimation we assume that the starting point z_0 of the discussed methods is in B. Recall that $L \geq 0$ denotes the Lipschitz constant of the operator F.

Theorem 3.3 (deterministic). *Let $(w_k)_{k\geq 0}$ be the sequence generated by Algorithm 3.1. If*

(i) FBF, i.e. $\diamondsuit_k = z_k$, with step size $\alpha_k = \alpha \leq 1/L$, or

(ii) FBFp, i.e. $\diamondsuit_k = w_{k-1}$, with step size $\alpha_k = \alpha \leq 1/2L$

is chosen, then for all $K \geq 1$ the averaged iterates $\bar{w}_K := \frac{1}{K}\sum_{k=0}^{K-1} w_k$ fulfill

$$G_B(\bar{w}_K) \leq \frac{D^2}{2\alpha K}, \tag{3.6}$$

where G_B is the restricted gap defined in (3.4).

The proof of the theorem relies on standard techniques making use of the monotinicity of the operator F where we are able to show that the distance to the solution set decreases in every iteration up to an additive error of

$$-\|z_k - w_k\|^2 + \alpha_k^2 L^2 \|\diamondsuit_k - w_k\|. \tag{3.7}$$

Depending on which iterate we use for $\diamondsuit_k$ we choose the stepsize α_k^2 such that the two terms cancel and obtain the desired result.

In order to derive similar convergence statements for the stochastic algorithm we need to assume (standard) properties of the gradient estimator $F(\cdot\,;\xi)$.

Assumption 3.4. *Unbiasedness:* $\mathbb{E}_\xi[F(w;\xi)] = F(w)\,\forall w \in \mathbb{R}^m$.

Assumption 3.5. *Bounded variance:* $\mathbb{E}_\xi[\|F(w;\xi) - F(w)\|^2] \le \sigma^2 \, \forall w \in \mathbb{R}^m$.

We want to point out that the latter assumption is weaker than the commonly used bounded (sub)gradient hypothesis ($\mathbb{E}_\xi[\|F(w;\xi)\|^2] \le \sigma^2$), which is known to conflict with properties like strong monotonicity, see [34].

In particular we actually only need the above assumption to hold for all iterates w_k. Such an hypothesis is in practice difficult to check, but could be exploited in special cases where additional properties of the variance and boundedness of the iterates are known a priori.

Assumption 3.6. *The samples ξ_k are independent of the iterates w_k, for all $k \ge 0$.*

Equipped with these assumptions we are now able to prove the statement.

Theorem 3.7 (stochastic). *Let Assumptions 3.4 to 3.6 hold and let $(w_k)_{k\ge0}$ be the sequence generated by Algorithm 3.2. If*

(i) stochastic FBF, i.e. $\diamondsuit_k = z_k$ and $\triangle_k = \eta_k$, with step size $\alpha_k \le \alpha \le 1/\sqrt{2}L$, or

(ii) stochastic FBFp, i.e. $\diamondsuit_k = w_{k-1}$ and $\triangle_k = \xi_{k-1}$, with step size $\alpha_k \le \alpha \le 1/3L$

is chosen, then for all $K \ge 1$ the averaged iterates $\bar{w}_K := \frac{\sum_{k=0}^{K-1} \alpha_k w_k}{\sum_{k=0}^{K-1} \alpha_k}$ fulfill

$$\mathbb{E}[G_B(\bar{w}_K)] \le \frac{D^2 + 24\sigma^2 \sum_{k=0}^{K-1} \alpha_k^2}{\sum_{k=0}^{K-1} \alpha_k}, \tag{3.8}$$

where G_B is the restricted gap defined in (3.4).

The above theorem exhibits a classical step size dependence [50], yielding convergence for sequences $(\alpha_k)_{k\ge0}$ that are square summable $\sum_{k=0}^{\infty} \alpha_k^2 < +\infty$ but not summable $\sum_{k=0}^{\infty} \alpha_k = +\infty$. Additionally, if in the setting of Theorem 3.7 the step size is chosen to be $\alpha_k = \alpha/\sqrt{k+1}$, a convergence rate can be obtained and is given by

$$\mathbb{E}[G_B(\bar{w}_K)] = \mathcal{O}\Big(\frac{1}{\sqrt{K}}\Big). \tag{3.9}$$

If the step size does not go to zero, the gap can usually not be expected to vanish either. However, we can still show decrease in the gap up to a residual stemming from the variance. In particular, for a constant step size $\alpha_k = \alpha$ we have

$$\mathbb{E}[G_B(\bar{w}_K)] \le \frac{D^2}{\alpha K} + 24\sigma^2\alpha. \tag{3.10}$$

Additionally, if the number of iterations K is fixed beforehand, a conclusion similar to (3.9) can be obtained by choosing $\alpha = 1/\sqrt{K}$ in (3.10).

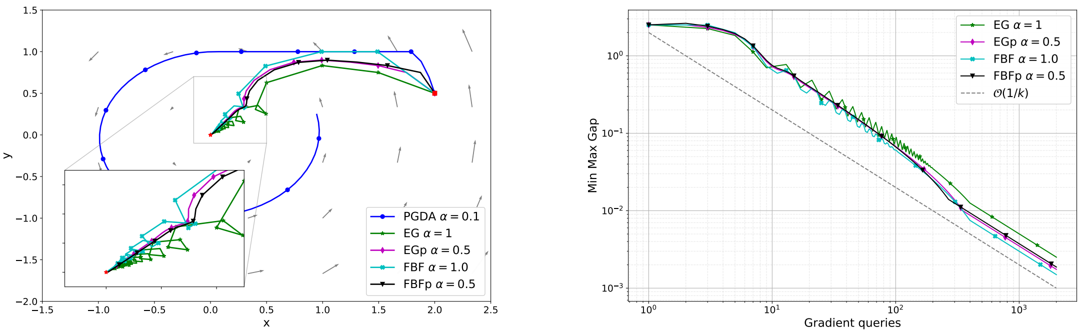
4. Experiments.

The aim of this section is to show how the use of methods with convergence guarantees, albeit only in the monotone setting, can yield better training performance for different architectures and objectives. In particular, we demonstrate that FBF can perform at least as good as EG although requiring less evaluations of the regularizers.

4.1. 2D toy example. Following [18, 38] and others we consider the example $\min_x \max_y xy$, illustrating the cycling behavior of (even bilinear) minimax problems. We augment this approach by adding a nonsmooth L1-regularizer for one player, with $\kappa > 0$, and constraints for the other, resulting in

$$\min_{x\in\mathbb{R}} \max_{y\in[-1,1]} \kappa|x| + xy. \tag{4.1}$$

The aforementioned issue of GDA (and its proximal extension PGDA) cycling around the solution is highlighted in Figure 1. The other methods, for which we display the averaged iterates, however do converge to a solution and show a decrease in the restricted gap according to Theorem 3.3. Interestingly, the two methods using extrapolation from the past seem to exhibit a more monotone decrease than FBF or EG.



(a) Trajectories converging to solution.

(b) Restricted gap function.

Figure 1: A comparison of the methods presented in subsection 2.3 applied to problem (4.1) with $\kappa = 0.01$. *PGDA* denotes (alternating) gradient descent ascent with proximal steps. As mentioned in the introduction it fails to converge (without averaging of the iterates). *EGp* denotes the method presented in [15] as extrapolation from the past. For the restricted gap we use $B_1 = B_2 = [-1, 1]$.

4.2. WGAN trained on CIFAR10. We now apply the above proposed techniques from monotone inclusions to the training of Wasserstein GANs employing DCGAN [47] and ResNet [23] architectures. All models are trained on the CIFAR10 dataset [30] which consists of 60,000 images in 10 different classes (with 50,000 training images and 10,000 test images) using an NVIDIA RTX 2080Ti GPU.

For the DCGAN experiments we work with the original WGAN formulation including weight clipping, since it includes regularizers innately (the indicator of a box for the weights of the discriminator). In addition we propose a modification of the WGAN formulation which replaces the box constraint on the discriminator's weights with an L1-regularization, under the name of *WGAN-L1*. This results in a *soft thresholding* operation instead of the "harsh" clipping.

For the experiments on ResNet we use the WGAN-GP formulation [21] which penalizes the norm of the gradient of the discriminator to enforce the Lipschitz constraint, together with spectral normalization of the weight matrices [41] which can be seen as a projection as argued in subsection 2.4.

Table 1: The best Inception Score (IS)[1] and Fréchet Inception Distance (FID). The column denoted by *WGAN*, *WGAN-L1* and *WGAN-GP* refers to the standard formulation with weight clipping, our regularized implementation using the 1-norm and the formulation with gradient penalty and spectral normalization, respectively.

	WGAN		WGAN-L1		WGAN-GP	
Method	IS	FID	IS	FID	IS	FID
AltAdam1	4.12±.06	56.44±.62	4.43±.03	50.86±2.17	6.01±.31	28.11±3.65
Extra Adam	4.07±.05	56.67±.61	4.67±.11	47.24±1.21	**6.58**±.08	21.40±.58
FBF Adam	**4.54**±.04	**45.85**±.35	**4.68**±.16	**46.60**±.76	6.57±.10	**21.22**±1.29
Opt. Adam	4.35±.06	50.41±.46	4.63±.13	47.98±1.49	6.42±.10	23.01±.95

Given the ubiquity and dominance of Adam [28] as an optimizer for many deep learning related training tasks, instead of using vanilla SGD we opt for Adam updates. This results in a method we call *FBF Adam*. Analogous approaches have been applied in [15] and [11] resulting in *Extra Adam* and *Optimistic Adam*, respectively. We compare the aforementioned methods with the status-quo in GAN training, namely alternating one Adam step for each network: *AltAdam1*.

Our hyperparameter search was limited to the step sizes when using the WGAN-L1 and WGAN-GP formulation, while all other parameters were kept the same as in [15, 7]. Note that we still report different values for the IS because we used the updated implementation [3]. It seems noteworthy that in the case of soft thresholding bigger step sizes performed better with the only exception of AltAdam1.

The two evaluation metrics used are the Inception Score (IS, higher is better) [53] and the Fréchet inception distance (FID, lower is better) [24], both computed on 50,000 samples.

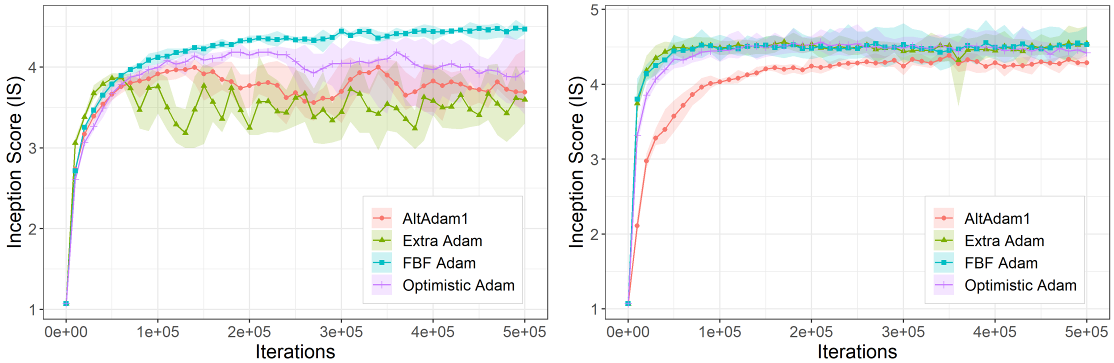
In Table 1 the best IS[1] and FID for each method are reported. FBF Adam performs at least as good as all considered competitors with respect to both evaluation metrics. One can also see that WGAN-L1 using the proximal operator improves the performance of all considered methods. Figure 2 shows the training progress regarding IS for each method and both problem formulations. The graphs suggest that making use of WGAN-L1 objective has a stabilizing effect during training, leading to a smoother and more consistent learning curve — a property that only FBF Adam seems to exhibit for weight clipping. Figure 3 as well as Table 1 show that for the WGAN-GP formulation FBF Adam maintains the improved performance of EG compared to GDA, while only requiring half the amount of spectral normalizations, resulting in time savings of up to 10% as reported in [41]. An even greater improvement by using FBF can be seen in Table 2 where we additionally consider the averaged iterates using an exponential moving average (EMA). As observed by others [15], this can have a very beneficial effect.

5. Conclusion. By highlighting the connection between GAN objectives and monotone inclusions, we are able to tackle their training via the Forward-Backward-Forward method which is known

[1] In the case of the IS we use the updated and corrected implementation from [3], leading to lower reported values.

Table 2: Fréchet Inception Distance (FID) of regular iterates and averaged iterates using an exponential moving average (EMA), from training a ResNet with WGAN-GP formulation.

	WGAN-GP		
Method	non avg.	EMA	best
AltAdam1	28.11±3.65	—	—
Extra Adam	21.40±.58	18.29±0.26	17.86
FBF Adam	**21.22**±1.29	**17.58**±0.73	**16.65**
Opt. Adam	23.01±.95	21.00±1.33	19.28



(a) IS for WGAN (weight clipping).

(b) IS for WGAN-L1 (proximal operator).

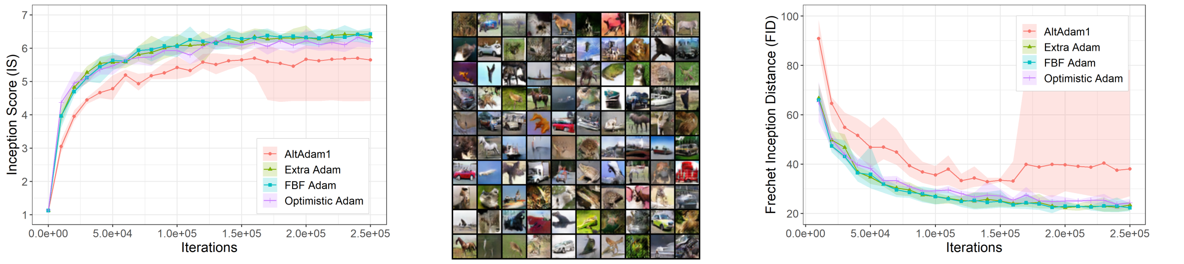
Figure 2: **DCGAN performance.** (a): Average and best/worst IS[1] on the WGAN objective with weight clipping. (b): Average and best/worst IS on the WGAN-L1 objective using the proximal operator (soft thresholding); The WGAN-L1 objective improves the IS in comparison to weight clipping and stabilizes the behavior of all considered methods during the training procedure.

to converge to a solution for convex-concave minimax problems. We extend this theoretical understanding by proving novel convergence rates in terms of the function values and highlighting the connection to other known methods like OGDA. We complement these rigorous considerations by promising practical results, indicating that application of FBF can lead to improved performance and saved computation time (compared to EG).

Appendix A. Definitions.

In subsection 2.4 we require the regularizers to be proper, convex and lower semicontinuous which are common properties in convex analysis. We call a function $r : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ *proper* if it is not constant $+\infty$, which means that it takes a finite value for at least a single point. In addition, we say that r is *lower semicontinuous* if for all $z_0 \in \mathbb{R}^m$

$$\liminf_{z \to z_0} r(z) \geq r(z_0). \tag{A.1}$$



(a) Inception Score. (b) Generator sample. (c) FID.

Figure 3: **ResNet performance**. Average and best/worst results regarding IS[1] and FID, see (a) and (b) respectively, using ResNet architecture on the WGAN-GP objective including spectral normalization. (c): Samples from the generator trained with FBF Adam.

It is easy to see that if $C \subset \mathbb{R}^m$ is nonempty, closed and convex, then the indicator δ_C of this set, given by

$$\delta_C(z) = \begin{cases} 0 & \text{if } z \in C \\ +\infty & \text{otherwise} \end{cases} \tag{A.2}$$

fulfills the assumptions of being proper, convex and lower semicontinuous.

Appendix B. About the gap function.

Typically in monotone inclusions, the distance to the set of solutions is used as a measure of quality of a given point due to the lack of more specific structure in general. Asymptotic convergence of the iterates has been established for FBF and FBFp in [4, Proposition 27.13] and [36], respectively. Furthermore, no convergence rates can be expected without stronger monotonicity assumptions. We want to take into account the special structure of the monotone inclusion coming from the minimax problem (2.1). For this reason we use the following (restricted) *minimax gap*, common for saddle point problems, which for a point (u, v) is given by

$$G_B(u,v) = \sup_{(x,y)\in B} \Psi(u,y) - \Psi(x,v). \tag{B.1}$$

For the general case, i.e. F being an arbitrary monotone and Lipschitz operator this is connected to the other measure of optimality we use in (3.4), for $w \in \mathbb{R}^m$ given by

$$G_B(w) = \sup_{z\in B} \langle F(z), w - z\rangle + r(w) - r(z), \tag{B.2}$$

where we interpret the possible occurrence of $\infty - \infty$ as $+\infty$. It stems from the field of Variational Inequalities where such a function is also known as *merit function* [44]. The relevance of the above two quantities will be made clear by the following statements.

Theorem B.1. *Let $\Phi : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be proper, convex and lower semicontinuous and $B \subset \mathbb{R}^d \times \mathbb{R}^n$. A point (x^*, y^*) in the interior of B solves the saddle point problem (2.1) if and only if its minimax gap (B.1) is zero, $G_B(x^*, y^*) = 0$. For all other elements of B the gap is nonnegative.*

Proof. A saddle point (x^*, y^*) clearly fulfills that $\sup_{(x,y)\in\mathbb{R}^d\times\mathbb{R}^n} \Psi(x^*, y) - \Psi(x, y^*) = 0$. On the other hand let $G_B(x^*, y^*) = 0$. For an arbitrary point (x, y) we can choose $\alpha \in (0, 1)$ large enough such that $(u, v) := \alpha(x^*, y^*) + (1 - \alpha)(x, y)$ is in the interior of B. Therefore,

$$\Psi(x^*, v) - \Psi(u, y^*) = \Psi(x^*, \alpha y^* + (1 - \alpha)y) - \Psi(\alpha x^* + (1 - \alpha)x, y^*) \leq 0. \tag{B.3}$$

Using the convex-concave structure of Ψ we deduce that

$$\alpha\Psi(x^*, y^*) + (1 - \alpha)\Psi(x^*, y) - \alpha\Psi(x^*, y^*) - (1 - \alpha)\Psi(x, y^*) \leq 0, \tag{B.4}$$

which implies that $\Psi(x^*, y) \leq \Psi(x, y^*)$. Since (x, y) was chosen arbitrary (x^*, y^*) is a saddle point. ∎

Similarly, an analogous statement can be shown for (B.2). The proof, however is split up into multiple lemmas to highlight the connection to Variational Inequalities.

Theorem B.2. *Let $F : \mathbb{R}^m \to \mathbb{R}^m$ be monotone and continuous, $r : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ proper, convex and lower semicontinuous and $B \subset \mathbb{R}^m$. A point w^* in the interior of B solves the monotone inclusion*

$$0 \in F(w) + \partial r(w) \tag{B.5}$$

if and only if its restricted gap (B.2) is zero, $G_B(w^*) = 0$. For all other elements of B the gap is nonnegative.

Let the assumptions of Theorem B.2 hold true for the following lemmas as we break up the proof into separate statements. We do so by making use of the associated *Variational inequality (VI)*

$$\text{find } w \text{ such that} \quad \langle F(w), z - w\rangle + r(z) - r(w) \geq 0 \quad \forall z \in \mathbb{R}^m. \tag{B.6}$$

Lemma B.3. *The monotone inclusion (B.5) is equivalent to the VI (B.6).*

Proof. The equivalence of (B.5) and (B.6) follows immediately from the definition of the subdifferential of r. ∎

The formulation (B.6) is typically referred to as the *strong* form of the VI, whereas

$$\text{find } w \text{ such that} \quad \langle F(z), z - w\rangle + r(z) - r(w) \geq 0 \quad \forall z \in \mathbb{R}^m, \tag{B.7}$$

is known as the *weak* formulation.

Lemma B.4. *Under the given assumptions the notion of weak and strong VI are equivalent.*

Proof. For the monotone operator F it is clear that if w^* is a solution to the strong formulation (B.6), it is also a solution to the weak formulation (B.7). In fact, if F is continuous the reverse implication also holds true. To see this, let w^* be a solution to the weak VI (B.7) and $z = \alpha w^* + (1-\alpha)u$ for an arbitrary $u \in \mathbb{R}^m$ and $\alpha \in (0, 1)$, then

$$\langle F(\alpha w^* + (1 - \alpha)u), (1 - \alpha)(u - w^*)\rangle + r(\alpha w^* + (1 - \alpha)u) - r(w^*) \geq 0. \tag{B.8}$$

This implies by the convexity of r that

$$(1 - \alpha)\langle F(\alpha w^* + (1 - \alpha)u), (u - w^*)\rangle + (1 - \alpha)(r(u) - r(w^*)) \geq 0. \tag{B.9}$$

By dividing by $(1 - \alpha)$ and then taking the limit $\alpha \to 1$ we obtain that w^* is a solution of the strong form (B.6). ∎

With the notion of VIs in mind, the above defined gap (B.2) becomes natural as it measures how much the statement of (B.7) is violated.

Lemma B.5. *G_B is nonnegative on B and zero for solutions of the weak VI.*

Proof. It is clear that $G_B(w) \geq 0$ for $w \in B$ as $z = w$ can be chosen in the supremum. On the other hand if $w^* \in B$ is a solution to the weak VI (B.7) then $G_B(w^*) = 0$. This follows from the fact that for a solution of (B.7) for all $z \in B$

$$\langle F(z), w^* - z\rangle + r(w^*) - r(z) \leq 0. \tag{B.10}$$

Therefore the supremum over the above expression in z is also less than zero, but clearly zero is obtained for $z = w^*$. ∎

For the reverse implication to hold true, we may not use points on the boundary of B.

Lemma B.6. *If a point w^* in the interior of B exhibits zero gap $G_B(w^*) = 0$, then it is a solution to the weak VI (B.7).*

Proof. Since w^* is in the interior of B we can, for an arbitrary $w \in \mathbb{R}^m$, choose $\alpha \in (0, 1)$ large enough such that $z := \alpha w^* + (1-\alpha)w \in B$. Using this z in the supremum of the gap we deduce that

$$\langle F(\alpha w^* + (1-\alpha)w), w^* - \alpha w^* - (1-\alpha)w\rangle + r(w^*) - r(\alpha w^* + (1-\alpha)w) \leq 0. \tag{B.11}$$

This implies that

$$(1-\alpha)\langle F(\alpha w^* + (1-\alpha)w), w - w^*\rangle + (1-\alpha)(r(w) - r(w^*)) \geq 0. \tag{B.12}$$

By dividing by $(1-\alpha)$ and then taking the limit $\alpha \to 1$ we deduce that w^* solves the strong form of the VI (B.6). ∎

Now, we can turn to proving the theorem.

Proof of Theorem B.2. Combine Lemmas B.3 to B.6. ∎

Appendix C. Refined theorems. Recall that restricted (unifying) gap function G_B defined in (3.4) is computed with respect to a set $B \subset \mathbb{R}^m$ where $D := \sup_{w,z\in B} \|z - w\|$ denotes its diameter and it is assumed that $z_0 \in B$. Furthermore, the averaged iterates $\bar{w}_K$ for $K \geq 1$ are given by

$$\bar{w}_K := \frac{\sum_{k=0}^{K-1} \alpha_k w_k}{\sum_{k=0}^{K-1} \alpha_k}. \tag{C.1}$$

C.1. Deterministic statements. The convergence statement of Theorem 3.3 actually holds true not just for a constant step size as presented in section 3, but for variable step sizes as well.

Theorem C.1. *Let $(w_k)_{k\geq 0}$ be the sequence generated by Algorithm 3.1. If*
(i) FBF, i.e. $\diamondsuit_k = z_k$, with step size $0 < \alpha_k \leq \alpha \leq 1/L$, or
(ii) FBFp, i.e. $\diamondsuit_k = w_{k-1}$, with step size $0 < \alpha_k \leq \alpha \leq 1/2L$
is chosen, then for all $K \geq 1$

$$G_B(\bar{w}_K) \leq \frac{D^2}{2\sum_{k=0}^{K-1} \alpha_k}. \tag{C.2}$$

C.2. Stochastic statements. We actually prove a slightly more general version of Theorem 3.7. In particular the step size can be chosen larger than initially claimed, however, at the cost of a worse constant.

Theorem C.2. *Let Assumptions 3.4 to 3.6 hold and let $(w_k)_{k\geq 0}$ be the sequence generated by FBF, i.e. Algorithm 3.2 with $\diamondsuit_k = z_k$ and $\triangle_k = \eta_k$. Let the step size $\alpha_k \leq \alpha < \frac{1}{L}$, then*

$$\mathbb{E}[G_B(\bar{w}_K)] \leq \frac{D^2 + 4(1-\alpha^2 L^2)^{-1}\sigma^2 \sum_{k=0}^{K-1} \alpha_k^2}{2\sum_{k=0}^{K-1} \alpha_k}, \tag{C.3}$$

for all $K \geq 1$.

Theorem 3.7 (i) can be deduced from the above statement by using $\alpha = 1/\sqrt{2}L$ which yields that $(1-\alpha^2 L^2)^{-1} = 2$.

Theorem C.3. *Let Assumptions 3.4 to 3.6 hold and let $(w_k)_{k\geq 0}$ be the sequence generated by FBFp, i.e. Algorithm 3.2 with $\diamondsuit_k = w_{k-1}$ and $\triangle_k = \xi_{k-1}$. Let the step size $\alpha_k \leq \alpha < \frac{1}{2\sqrt{2}L}$, then*

$$\mathbb{E}[G_B(\bar{w}_K)] \leq \frac{D^2 + 6\left(1 + \frac{4\alpha^2L^2}{1-8\alpha^2L^2}\right)\sigma^2 \sum_{k=0}^{K-1} \alpha_k^2}{\sum_{k=0}^{K-1} \alpha_k}, \tag{C.4}$$

for all $K \geq 1$.

Theorem 3.7 (ii) is obtained from the above theorem by using the particular step size bound of $\alpha = 1/3L$, which yields that

$$\frac{4\alpha^2 L^2}{1 - 8\alpha^2 L^2} = 4. \tag{C.5}$$

Although, the step size in the refined statements Theorems C.2 and C.3 can be chosen arbitrarily close to $1/L$ and $1/(2\sqrt{2}L)$ for stochastic FBF and stochastic FBFp, respectively. This does not mean it should be — since the constant in the convergence rate deteriorates when the step size is close to its allowed upper bound.

Appendix D. Proofs.

D.1. Preparations. We introduce the notation connected to the strong formulation of the VI (B.6) associated to the monotone inclusion (3.1), given by

$$g(w,z) := \langle F(w), w - z\rangle + r(w) - r(z), \tag{D.1}$$

for $g : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$. Next we will establish the fact that this function can be used to bound the (restricted) unifying gap function, which we remind, is defined as

$$G_B(w) = \begin{cases} \sup_{(x,y)\in B} \Psi(u,y) - \Psi(x,v) & \text{if } F \text{ is } (2.5) \\ \sup_{z\in B} \langle F(z), w - z\rangle + r(w) - r(z) & \text{otherwise,} \end{cases} \tag{D.2}$$

where in the first case $(u,v) \in \mathbb{R}^d \times \mathbb{R}^n$ is identified with $w \in \mathbb{R}^m$. In particular the dimensions fulfill $d + n = m$, and $r(w)$ is given by $f(u) + h(v)$.

Lemma D.1. *It holds that for all $K \geq 1$*

$$\sup_{z\in B}\left\{\frac{1}{\sum_{k=0}^{K-1}\alpha_k}\sum_{k=0}^{K-1}\alpha_k g(w_k,z)\right\} \geq G_B(\bar{w}_K). \tag{D.3}$$

Proof. First we will prove the case if F is derived from a saddle point problem. Note that from the convex-concave structure of Φ we get that

$$\Phi(u,y) \leq \Phi(u,v) + \langle \nabla_y \Phi(u,v), y-v\rangle \tag{D.4}$$

and

$$\Phi(u,v) + \langle \nabla_x \Phi(u,v), x-u\rangle \leq \Phi(x,v). \tag{D.5}$$

By summing the two up we obtain

$$\Phi(u,y) - \Phi(x,v) \leq \left\langle \begin{matrix} -\nabla_x\Phi(u,v), \\ \nabla_y\Phi(u,v), \end{matrix} \quad \begin{matrix} x-u \\ y-v \end{matrix} \right\rangle. \tag{D.6}$$

We can reformulate the above inequality in terms of g to see that for $z=(x,y)\in\mathbb{R}^d\times\mathbb{R}^n$

$$\langle F(w), w-z\rangle \geq \Phi(u,y) - \Phi(x,v). \tag{D.7}$$

The statement of the first case is obtained by adding $r(w)-r(z)$ on both sides and using the fact that Ψ is convex-concave.

If F is a general monotone operator, then we use its monotonicity to deduce that

$$\langle F(w), w-z\rangle \geq \langle F(z), w-z\rangle. \tag{D.8}$$

The desired result follows from using the linearity of the inner product. ∎

Notation. We denote the error of the stochastic estimator via

$$Z_k := F(\diamondsuit_k;\triangle_k) - F(\diamondsuit_k) \quad \text{and} \quad W_k := F(w_k;\xi_k) - F(w_k). \tag{D.9}$$

Furthermore, we will denote via $\mathbb{E}[\,\cdot\,|\,U]$, the conditional expectation with respect to the random variable U.

We will also need the following lemma.

Lemma D.2. *Let $(p_k)_{k\geq 0} \in \mathbb{R}^d$ be a given sequence and $(v_k)_{k\geq 0}$ recursively defined for all $k\geq 0$ by $v_{k+1} := v_k - p_k$ for some $v_0\in\mathbb{R}^d$, then*

$$\sum_{k=0}^{K-1}\langle p_k, v_k-u\rangle \leq \frac{1}{2}\|v_0-u\|^2 + \frac{1}{2}\sum_{k=0}^{K-1}\|p_k\|^2. \tag{D.10}$$

Proof. From the three point identity it follows immediately that

$$\langle p_k, v_k-u\rangle = \langle v_k - v_{k+1}, v_k-u\rangle = \frac{1}{2}\left(\|v_k-u\|^2 - \|v_{k+1}-u\|^2 + \|v_{k+1}-v_k\|^2\right) \tag{D.11}$$

from which the statement of the lemma follows. ∎

D.2. A unified decrease result. We will start with a unifying proposition which covers the common parts of all convergence proofs.

Proposition D.3. *For a $\gamma > 0$ we have that for all $k \geq 0$ and $z \in \mathbb{R}^m$*

$$\alpha_k g(w_k, z) + \frac{1}{2}\|z_{k+1} - z\|^2 \leq \frac{1}{2}\|z_k - z\|^2 - \frac{1}{2}\|z_k - w_k\|^2 + \frac{1}{2}(1+\gamma)\alpha_k^2 L^2 \|\diamondsuit_k - w_k\|^2 + \alpha_k \langle W_k, z - w_k\rangle + (1+\gamma^{-1})\alpha_k^2(\|W_k\|^2 + \|Z_k\|^2). \tag{D.12}$$

Proof. Let $k \geq 0$ and $z \in \mathbb{R}^m$ be arbitrary. Using the decomposition (D.9) it follows that

$$\langle \alpha_k F(w_k; \xi_k), w_k - z\rangle = \alpha_k \langle W_k, w_k - z\rangle + \alpha_k \langle F(w_k), w_k - z\rangle. \tag{D.13}$$

Since $\operatorname{prox}_{\alpha_k r} = (\mathrm{Id} + \alpha_k \partial r)^{-1}$ we deduce that

$$\langle z - w_k, w_k - z_k + \alpha_k F(\diamondsuit_k; \triangle_k)\rangle \geq \alpha_k (r(w_k) - r(z)). \tag{D.14}$$

Adding (D.13) and (D.14) gives that

$$\langle \alpha_k (F(w_k; \xi_k) - F(\diamondsuit_k; \triangle_k)) + z_k - w_k, w_k - z\rangle \geq \alpha_k \langle W_k, w_k - z\rangle + \alpha_k g(w_k, z), \tag{D.15}$$

which, using the definition of z_{k+1}, is equivalent to

$$\langle z - w_k, z_{k+1} - z_k\rangle \geq \alpha_k \langle W_k, w_k - z\rangle + \alpha_k g(w_k, z). \tag{D.16}$$

We estimate the inner product on the left side of the inequality by inserting and subtracting z_k and using the three point identity twice to deduce

$$\begin{aligned} \langle z - w_k, z_{k+1} - z_k\rangle &= \langle z - z_k + z_k - w_k, z_{k+1} - z_k\rangle \\ &= \frac{1}{2}\left(\|z - z_k\|^2 - \|z_{k+1} - z\|^2 + \|z_{k+1} - w_k\|^2 - \|z_k - w_k\|^2\right). \end{aligned} \tag{D.17}$$

The first two summands are fine as they will telescope, so we are left with estimating $\|z_{k+1} - w_k\|^2$. By the definition of z_{k+1} we have that

$$\begin{aligned} \|z_{k+1} - w_k\|^2 &= \alpha_k^2 \|F(\diamondsuit_k; \triangle_k) - F(w_k; \xi_k)\|^2 \\ &= \alpha_k^2 \|F(\diamondsuit_k) - F(w_k) + Z_k - W_k\|^2 \\ &\leq (1+\gamma)\alpha_k^2 \|F(\diamondsuit_k) - F(w_k)\|^2 + (1+\gamma^{-1})\alpha_k^2 \|Z_k - W_k\|^2 \\ &\leq (1+\gamma)\alpha_k^2 L^2 \|\diamondsuit_k - w_k\|^2 + 2(1+\gamma^{-1})\alpha_k^2 (\|Z_k\|^2 + \|W_k\|^2), \end{aligned} \tag{D.18}$$

where we inserted and subtracted $F(\diamondsuit_k)$ and $F(w_k)$ and applied Young's inequality to deduce the result. Adding (D.18), (D.17) and (D.16) we conclude that ∎

$$\alpha_k g(w_k, z) + \frac{1}{2}\|z_{k+1} - z\|^2 \leq \frac{1}{2}\|z_k - z\|^2 - \frac{1}{2}\|z_k - w_k\|^2 + \frac{1}{2}(1+\gamma)\alpha_k^2 L^2 \|\diamondsuit_k - w_k\|^2 + \alpha_k \langle W_k, z - w_k\rangle + (1+\gamma^{-1})\alpha_k^2(\|W_k\|^2 + \|Z_k\|^2). \tag{D.19}$$

D.3. Forward-Backward-Forward.

Proof for deterministic FBF, Theorem C.1 (i). We start off by plugging $\diamondsuit_k = z_k$ into (D.12). Since $W_k = Z_k = 0$ we can use $\gamma \to 0$ to deduce that for all $k \geq 0$

$$\alpha_k g(w_k, z) + \frac{1}{2}\|z_{k+1} - z\|^2 \leq \frac{1}{2}\|z_k - z\|^2 - \frac{1}{2}(1 - \alpha_k^2 L^2)\|z_k - w_k\|^2. \tag{D.20}$$

From this it is clear that the step size is constrained by $\alpha \leq 1/L$ as stated in the theorem. By summing up from $k = 0$ to $K - 1$ and dividing by $\sum_{k=0}^{K-1} \alpha_k$ we obtain

$$\frac{1}{\sum_{k=0}^{K-1} \alpha_k} \sum_{k=0}^{K-1} \alpha_k g(w_k, z) \leq \frac{\|z_0 - z\|^2}{2\sum_{k=0}^{K-1} \alpha_k}. \tag{D.21}$$

The claimed statement is then derived by taking the supremum in z over B and applying Lemma D.1. ■

Proof for stochastic FBF, Theorem C.2. Plugging $\diamondsuit_k = z_k$ and $\triangle_k = \eta_k$ into (D.12) gives for all $k \geq 0$

$$\begin{aligned} &\alpha_k g(w_k, z) + \frac{1}{2}\|z_{k+1} - z\|^2 \\ &\leq \frac{1}{2}\|z_k - z\|^2 - \frac{1}{2}(1 - (1+\gamma)\alpha_k^2 L^2)\|z_k - w_k\|^2 + \alpha_k \langle W_k, z - v_k\rangle \\ &\quad + \alpha_k \langle W_k, v_k - w_k\rangle + (1 + \gamma^{-1})\alpha_k^2(\|W_k\|^2 + \|Z_k\|^2). \end{aligned} \tag{D.22}$$

By summing this inequality up and applying Lemma D.2 with $v_0 = z_0$, $p_k = -\alpha_k W_k$ and $v_{k+1} := v_k - p_k$ we deduce that

$$\sum_{k=0}^{K-1} \langle -\alpha_k W_k, v_k - z\rangle \leq \frac{1}{2}\|z_0 - z\|^2 + \frac{1}{2}\sum_{k=0}^{K-1} \alpha_k^2 \|W_k\|^2, \tag{D.23}$$

and therefore

$$\sum_{k=0}^{K-1} \alpha_k g(w_k, z) \leq \|z_0 - z\|^2 + \sum_{k=0}^{K} \alpha_k \langle W_k, v_k - w_k\rangle + 2(1 + \gamma^{-1})\alpha_k^2(\|W_k\|^2 + \|Z_k\|^2). \tag{D.24}$$

By choosing γ such that $\alpha = (\sqrt{1+\gamma}L)^{-1}$ we deduce that $1 + \gamma^{-1} = 1/(1 - \alpha^2 L^2)$. Next, we take the supremum over $z \in B$ and the expectation to obtain

$$\mathbb{E}\left[\sup_{z\in B}\left\{\sum_{k=0}^{K-1} \alpha_k g(w_k, z)\right\}\right] \leq D^2 + 4(1 - \alpha^2 L^2)^{-1}\sigma^2 \sum_{k=0}^{K-1} \alpha_k^2, \tag{D.25}$$

where we used that

$$\begin{aligned} \mathbb{E}[\langle W_k, v_k - w_k\rangle] &= \mathbb{E}\Big[\mathbb{E}\big[\langle W_k, v_k - w_k\rangle \,\big|\, w_{[k]}, \xi_{[k-1]}\big]\Big] \\ &= \mathbb{E}\Big[\big\langle \mathbb{E}\big[W_k \,\big|\, w_{[k]}, \xi_{[k-1]}\big], v_k - w_k\big\rangle\Big] = 0, \end{aligned} \tag{D.26}$$

with $\xi_{[k-1]} = (\xi_0, \ldots, \xi_{k-1})$ and $w_{[k]} = (w_0, \ldots, w_k)$. The final statement follows by dividing by $\sum_{k=0}^{K-1} \alpha_k$ and applying Lemma D.1. ■

D.4. Forward-Backward-Forward-past.

Proof for deterministic FBFp, Theorem C.1 (ii). We start off by plugging $\diamondsuit_k = z_k$ into (D.12). Since $W_k = Z_k = 0$ we can use $\gamma \to 0$ to conclude that for all $k \geq 0$

$$\alpha_k g(w_k, z) + \frac{1}{2}\|z_{k+1} - z\|^2 \leq \frac{1}{2}\|z_k - z\|^2 - \frac{1}{2}\|z_k - w_k\|^2 + \frac{1}{2}\alpha_k^2 L^2 \|w_{k-1} - w_k\|^2. \tag{D.27}$$

Now we need to bound the term $\|w_{k-1} - w_k\|^2$ by $\|z_k - w_k\|^2$. Since

$$2\|z_k - w_k\|^2 + 2\|z_k - w_{k-1}\|^2 \geq \|w_k - w_{k-1}\|^2 \tag{D.28}$$

we have for all $k \geq 1$

$$\begin{aligned} \|z_k - w_k\|^2 &\geq -\|z_k - w_{k-1}\|^2 + \frac{1}{2}\|w_{k-1} - w_k\|^2 \\ &\geq -\alpha_{k-1}^2 L^2 \|w_{k-1} - w_{k-2}\|^2 + \frac{1}{2}\|w_{k-1} - w_k\|^2 \end{aligned} \tag{D.29}$$

whereas for $k = 0$, since $w_{-1} = z_0$, we have that

$$\|z_0 - w_0\|^2 = \|w_{-1} - w_0\|^2. \tag{D.30}$$

Plugging (D.30) into (D.27) for $k = 0$ we get that

$$\alpha_0 g(w_0, z) + \frac{1}{2}\|z_1 - z\|^2 + \frac{1}{2}(1 - \alpha_0^2 L^2)\|w_0 - w_{-1}\|^2 \leq \frac{1}{2}\|z_0 - z\|^2. \tag{D.31}$$

Plugging (D.29) into (D.27) we get that for all $k \geq 1$

$$\begin{aligned} &\alpha_k g(w_k, z) + \frac{1}{2}\|z_{k+1} - z\|^2 + \frac{1}{2}\left(\frac{1}{2} - \alpha_k^2 L^2\right)\|w_k - w_{k-1}\|^2 \\ \leq &\frac{1}{2}\|z_k - z\|^2 + \frac{1}{2}\alpha_{k-1}^2 L^2 \|w_{k-1} - w_{k-2}\|^2. \end{aligned} \tag{D.32}$$

In order to be able to telescope we need to ensure that for all $k \geq 0$

$$\left(\frac{1}{2} - \alpha_k^2 L^2\right) \geq \alpha_k^2 L^2. \tag{D.33}$$

This is equivalent to the condition $\alpha_k \leq 1/2L$ which was required in the statement of the theorem. Now we sum up (D.32) from $k = 1$ to $K - 1$ which yields

$$\begin{aligned} &\sum_{k=1}^{K-1} \alpha_k g(w_k, z) + \frac{1}{2}\|z_K - z\|^2 + \frac{1}{2}\left(\frac{1}{2} - \alpha_{K-1}^2 L^2\right)\|w_{K-1} - w_{K-2}\|^2 \\ &\leq \frac{1}{2}\|z_1 - z\|^2 + \frac{1}{2}\alpha_0^2 L^2 \|w_0 - w_{-1}\|^2. \end{aligned} \tag{D.34}$$

Adding (D.34) and (D.31) and dividing by $\sum_{k=0}^{K-1} \alpha_k$ to deduce

$$\frac{1}{\sum_{k=0}^{K-1} \alpha_k} \sum_{k=0}^{K-1} \alpha_k g(w_k, z) \leq \frac{\|z_0 - z\|^2}{2\sum_{k=0}^{K-1} \alpha_k}, \tag{D.35}$$

where we used that $1 - \alpha_0^2 L^2 \geq \alpha_0^2 L^2$ to get rid of $\|w_0 - w_{-1}\|^2$. The final statement follows by taking the supremum in z over B and applying Lemma D.1. ∎

Proof for stochastic FBFp, Theorem C.3. By using $\diamondsuit_k = w_{k-1}$ we deduce from (D.12) for all $k \geq 0$ that

(D.36)
$$\alpha_k g(w_k, z) + \frac{1}{2}\|z_{k+1} - z\|^2 \leq \frac{1}{2}\|z_k - z\|^2 - \frac{1}{2}\|z_k - w_k\|^2 + \frac{1}{2}(1+\gamma)\alpha_k^2 L^2 \|w_{k-1} - w_k\|^2 + \alpha_k \langle W_k, z - w_k\rangle + 2(1+\gamma^{-1})\alpha_k^2(\|W_k\|^2 + \|Z_k\|^2).$$

As in (D.23) we can split $\langle \alpha_k W_k, z - w_k\rangle$ into $\langle \alpha_k W_k, z - v_k\rangle + \langle \alpha_k W_k, v_k - w_k\rangle$ and use Lemma D.2 to deduce

$$\sum_{k=0}^{K-1} \alpha_k g(w_k, z) \leq \|z_0 - z\|^2 - \sum_{k=0}^{K-1} \Big(\frac{1}{2}\|z_k - w_k\|^2 + \frac{1}{2}(1+\gamma)\alpha_k^2 L^2 \|w_{k-1} - w_k\|^2 + \langle \alpha_k W_k, v_k - w_k\rangle + 3(1+\gamma^{-1})\alpha_k^2(\|W_k\|^2 + \|Z_k\|^2)\Big). \tag{D.37}$$

Taking now the supremum over $z \in B$ and then the expectation we conclude that the inequality

$$\mathbb{E}\left[\sup_{z\in B}\left\{\sum_{k=0}^{K-1}\alpha_k g(w_k, z)\right\}\right] \leq D^2 - \frac{1}{2}\sum_{k=0}^{K-1}\left(\|z_k - w_k\|^2 - (1+\gamma)\alpha_k^2 L^2\|w_{k-1} - w_k\|^2\right) + 3(1+\gamma^{-1})\sigma^2 \sum_{k=0}^{K-1}\alpha_k^2 \tag{D.38}$$

holds. Let from now on $k \geq 1$ as we will treat the case $k = 0$ separately. Using (D.28) we deduce that

$$\begin{aligned} \|z_k - w_k\|^2 &\geq -\|z_k - w_{k-1}\|^2 + \frac{1}{2}\|w_{k-1} - w_k\|^2 \\ &\geq -\alpha_{k-1}^2\|F(w_{k-1};\xi_{k-1}) - F(w_{k-2};\xi_{k-2})\|^2 + \frac{1}{2}\|w_{k-1} - w_k\|^2. \end{aligned} \tag{D.39}$$

Now we bound the difference of the two estimators by inserting $\pm F(w_{k-1})$, $\pm F(w_{k-2})$ and applying the inequality $\|a+b+c\|^2 \leq 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$ which yields

(D.40)
$$\|F(w_{k-1};\xi_{k-1}) - F(w_{k-2};\xi_{k-2})\|^2 \leq 3\|W_{k-1}\|^2 + 3\|W_{k-2}\|^2 + 3\|F(w_{k-2}) - F(w_{k-1})\|^2.$$

We conclude that

$$\mathbb{E}\big[\|F(w_{k-1};\xi_{k-1}) - F(w_{k-2};\xi_{k-2})\|^2\big] \leq 6\sigma^2 + 3L^2\mathbb{E}\|w_{k-1} - w_{k-2}\|^2. \tag{D.41}$$

Using (D.41) in (D.39) we deduce that

$$\mathbb{E}\|z_k - w_k\|^2 \geq -\alpha_{k-1}^2(6\sigma^2 + 3L^2\mathbb{E}\|w_{k-1} - w_{k-2}\|^2) + \frac{1}{2}\mathbb{E}\|w_{k-1} - w_k\|^2, \tag{D.42}$$

whereas for $k = 0$ we have (D.30). Now we plug (D.42) into (D.38) to conclude that

(D.43)
$$\begin{aligned}
&\mathbb{E}\left[\sup_{z\in B}\left\{\sum_{k=0}^{K-1}\alpha_k g(w_k,z)\right\}\right] \\
&\qquad \le D^2 - \frac{1}{2}\sum_{k=1}^{K-1}\left(-3\alpha_{k-1}^2 L^2 \mathbb{E}\|w_{k-1}-w_{k-2}\|^2 + \left(\frac{1}{2}-(1+\gamma)\alpha_k^2L^2\right)\|w_{k-1}-w_k\|^2\right) \\
&\qquad\quad + \frac{1}{2}((1+\gamma)\alpha_0^2L^2-1)\|w_{-1}-w_0\|^2 + 6(1+\gamma^{-1})\sigma^2\sum_{k=0}^{K-1}\alpha_k^2
\end{aligned}$$

From this we conclude that in order to be able to telescope we need to enforce

$$\left(\frac{1}{2}-(1+\gamma)\alpha_k^2L^2\right) \ge 3\alpha_k^2L^2 \tag{D.44}$$

which is equivalent to

$$\frac{1}{2(4+\gamma)} \ge \alpha_k^2L^2. \tag{D.45}$$

Since $\alpha_k \le \alpha$, we can ensure this by choosing γ such that

$$\frac{1}{2(4+\gamma)} = \alpha^2L^2. \tag{D.46}$$

With (D.46) in place conclude from (D.43) that the inequality

$$\begin{aligned}
&\mathbb{E}\left[\sup_{z\in B}\left\{\sum_{k=0}^{K-1}\alpha_k g(w_k,z)\right\}\right] \\
&\qquad \le D^2 + \frac{1}{2}((4+\gamma)\alpha_0^2L^2-1)\|w_{-1}-w_0\|^2 + 6(1+\gamma^{-1})\sigma^2\sum_{k=0}^{K-1}\alpha_k^2
\end{aligned} \tag{D.47}$$

Using the fact that $3\alpha_0^2L^2 \le 1-(1+\gamma)\alpha_0^2L^2$ from (D.46) to discard the $\|w_0-w_{-1}\|^2$ term, yields

$$\mathbb{E}\left[\sup_{z\in B}\left\{\sum_{k=0}^{K-1}\alpha_k g(w_k,z)\right\}\right] \le D^2 + 6(1+\gamma^{-1})\sigma^2\sum_{k=0}^{K-1}\alpha_k^2 \tag{D.48}$$

Through (D.46), we can estimate

$$\frac{1}{\gamma} = \frac{2\alpha^2L^2}{1-8\alpha^2L^2}. \tag{D.49}$$

Plugging (D.49) into (D.48), dividing by $\sum_{k=0}^{K-1}\alpha_k$ and applying Lemma D.1, deduces the final statement. ■

REFERENCES

[1] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein GAN*, arXiv:1701.07875, (2017).
[2] F. Bach, *Breaking the curse of dimensionality with convex neural networks*, The Journal of Machine Learning Research, 18 (2017), pp. 629–681.
[3] S. Barratt and R. Sharma, *A note on the inception score*, arXiv:1801.01973, (2018).
[4] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, vol. 408, Springer, 2011.
[5] Y. Bengio, N. L. Roux, P. Vincent, O. Delalleau, and P. Marcotte, *Convex neural networks*, in Advances in Neural Information Processing Systems, 2006, pp. 123–130.
[6] R. I. Boţ, P. Mertikopoulos, M. Staudigl, and P. T. Vuong, *Minibatch forward-backward-forward methods for solving stochastic variational inequalities*, Stochastic Systems, 11 (2021), pp. 112–139.
[7] R. I. Boţ, M. Sedlmayer, and P. T. Vuong, *A relaxed inertial forward-backward-forward algorithm for solving monotone inclusions with application to GANs*, arXiv:2003.07886, (2020).
[8] T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien, *Reducing noise in GAN training with variance reduced extragradient*, in Advances in Neural Information Processing Systems, 2019, pp. 391–401.
[9] T. Chavdarova, M. Pagliardini, S. U. Stich, F. Fleuret, and M. Jaggi, *Taming gans with lookahead-minmax*, arXiv preprint arXiv:2006.14567, (2020).
[10] E. R. Csetnek, Y. Malitsky, and M. K. Tam, *Shadow douglas-rachford splitting for monotone inclusions*, Applied Mathematics & Optimization, 80 (2019), pp. 665–678.
[11] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng, *Training GANs with optimism*, in International Conference on Learning Representations, 2018, https://openreview.net/forum?id=SJJySbbAZ.
[12] C. Daskalakis and I. Panageas, *The limit points of (optimistic) gradient descent in min-max optimization*, in Advances in Neural Information Processing Systems, 2018, pp. 9236–9246.
[13] C. Daskalakis, S. Skoulakis, and M. Zampetakis, *The complexity of constrained min-max optimization*, in Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing, 2021, pp. 1466–1478.
[14] I. Gemp and S. Mahadevan, *Global convergence to the equilibrium of gans using variational inequalities*, arXiv preprint arXiv:1808.01531, (2018).
[15] G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien, *A variational inequality perspective on generative adversarial networks*, in International Conference on Learning Representations, 2019, https://openreview.net/forum?id=r1laEnA5Ym.
[16] G. Gidel, R. A. Hemmat, M. Pezeshki, R. Le Priol, G. Huang, S. Lacoste-Julien, and I. Mitliagkas, *Negative momentum for improved game dynamics*, in The 22nd International Conference on Artificial Intelligence and Statistics, 2019, pp. 1802–1811.
[17] N. Golowich, S. Pattathil, C. Daskalakis, and A. Ozdaglar, *Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems*, in Conference on Learning Theory, PMLR, 2020, pp. 1758–1784.
[18] I. Goodfellow, *Nips 2016 tutorial: Generative adversarial networks*, arXiv:1701.00160, (2016).
[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, in Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[20] E. Gorbunov, N. Loizou, and G. Gidel, *Extragradient method: $o(1/k)$ last-iterate convergence for monotone variational inequalities and connections with cocoercivity*, arXiv preprint arXiv:2110.04261, (2021).
[21] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, *Improved training of Wasserstein GANs*, in Advances in Neural Information Processing Systems, 2017, pp. 5767–5777.
[22] E. Y. Hamedani and N. S. Aybat, *A primal-dual algorithm for general convex-concave saddle point problems*, arXiv:1803.01401, (2018).
[23] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[24] 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 Advances in Neural Information Processing Systems, 2017, pp. 6626–6637.
[25] Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos, *Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling*, arXiv preprint arXiv:2003.10162, (2020).

[26] A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson, *Extragradient method with variance reduction for stochastic variational inequalities*, SIAM Journal on Optimization, 27 (2017), pp. 686–724.

[27] A. Juditsky, A. Nemirovski, and C. Tauvel, *Solving variational inequalities with stochastic mirror-prox algorithm*, Stochastic Systems, 1 (2011), pp. 17–58.

[28] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, arXiv:1412.6980, (2014).

[29] G. Korpelevich, *The extragradient method for finding saddle points and other problems*, Matecon, 12 (1976), pp. 747–756.

[30] A. Krizhevsky, G. Hinton, et al., *Learning multiple layers of features from tiny images*, (2009).

[31] T. Liang and J. Stokes, *Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks*, in The 22nd International Conference on Artificial Intelligence and Statistics, K. Chaudhuri and M. Sugiyama, eds., vol. 89 of Proceedings of Machine Learning Research, PMLR, 2019, pp. 907–915.

[32] Q. Lin, M. Liu, H. Rafique, and T. Yang, *Solving weakly-convex-weakly-concave saddle-point problems as successive strongly monotone variational inequalities*, arXiv:1810.10207, (2018).

[33] N. Loizou, H. Berard, A. Jolicoeur-Martineau, P. Vincent, S. Lacoste-Julien, and I. Mitliagkas, *Stochastic hamiltonian gradient methods for smooth games*, in International Conference on Machine Learning, PMLR, 2020, pp. 6370–6381.

[34] N. Loizou, S. Vaswani, I. H. Laradji, and S. Lacoste-Julien, *Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 1306–1314.

[35] Y. Malitsky, *Projected reflected gradient methods for monotone variational inequalities*, SIAM Journal on Optimization, 25 (2015), pp. 502–520.

[36] Y. Malitsky and M. K. Tam, *A forward-backward splitting method for monotone inclusions without cocoercivity*, SIAM Journal on Optimization, 30 (2020), pp. 1451–1472.

[37] P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras, *Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile*, in International Conference on Learning Representations, 2019, https://openreview.net/forum?id=Bkg8jjC9KQ.

[38] L. Mescheder, A. Geiger, and S. Nowozin, *Which training methods for GANs do actually converge?*, in International Conference on Machine Learning, 2018.

[39] L. Mescheder, S. Nowozin, and A. Geiger, *The numerics of gans*, arXiv preprint arXiv:1705.10461, (2017).

[40] K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtárik, and Y. Malitsky, *Revisiting stochastic extragradient*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 4573–4582.

[41] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, *Spectral normalization for generative adversarial networks*, in International Conference on Learning Representations, 2018, https://openreview.net/forum?id=B1QRgziT-.

[42] A. Mokhtari, A. E. Ozdaglar, and S. Pattathil, *Convergence rate of $O(1/k)$ for optimistic gradient and extra-gradient methods in smooth convex-concave saddle point problems*, SIAM Journal on Optimization, 30 (2020), pp. 3230–3251.

[43] 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 Journal on Optimization, 15 (2004), pp. 229–251.

[44] Y. Nesterov, *Dual extrapolation and its applications to solving variational inequalities and related problems*, Mathematical Programming, 109 (2007), pp. 319–344.

[45] D. P. Palomar and Y. C. Eldar, *Convex optimization in signal processing and communications*, Cambridge university press, 2010.

[46] K. Pieper and A. Petrosyan, *Nonconvex penalization for sparse neural networks*, arXiv:2004.11515, (2020).

[47] A. Radford, L. Metz, and S. Chintala, *Unsupervised representation learning with deep convolutional generative adversarial networks*, arXiv:1511.06434, (2015).

[48] A. Rakhlin and K. Sridharan, *Online learning with predictable sequences*, in Proceedings of the 26th Annual Conference on Learning Theory, 2013, pp. 993–1019.

[49] S. Rakhlin and K. Sridharan, *Optimization, learning, and games with predictable sequences*, in Advances in Neural Information Processing Systems, 2013, pp. 3066–3074.

[50] H. Robbins and S. Monro, *A stochastic approximation method*, The annals of mathematical statistics, (1951), pp. 400–407.

[51] S. Rosset, G. Swirszcz, N. Srebro, and J. Zhu, *ℓ 1 regularization in infinite dimensional feature spaces*, in

International Conference on Computational Learning Theory, Springer, 2007, pp. 544–558.

[52] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: nonlinear phenomena, 60 (1992), pp. 259–268.

[53] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, *Improved techniques for training GANs*, in Advances in Neural Information Processing Systems, 2016, pp. 2234–2242.

[54] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B (Methodological), 58 (1996), pp. 267–288.

[55] P. Tseng, *Applications of a splitting algorithm to decomposition in convex programming and variational inequalities*, SIAM Journal on Control and Optimization, 29 (1991), pp. 119–138.