Stable Nonconvex-Nonconcave Training via Linear Interpolation

Thomas Pethick
EPFL (LIONS)
thomas.pethick@epfl.ch

Wanyun Xie
EPFL (LIONS)
wanyun.xie@epfl.ch

Volkan Cevher
EPFL (LIONS)
volkan.cevher@epfl.ch

Abstract

This paper presents a theoretical analysis of linear interpolation as a principled method for stabilizing (large-scale) neural network training. We argue that instabilities in the optimization process are often caused by the nonmonotonicity of the loss landscape and show how linear interpolation can help by leveraging the theory of nonexpansive operators. We construct a new optimization scheme called relaxed approximate proximal point (RAPP), which is the first 1-SCLI method to achieve last iterate convergence rates for ρ-comonotone problems while only requiring $\rho > -\frac{1}{2L}$. The construction extends to constrained and regularized settings. By replacing the inner optimizer in RAPP we rediscover the family of Lookahead algorithms for which we establish convergence in cohypomonotone problems even when the base optimizer is taken to be gradient descent ascent. The range of cohypomonotone problems in which Lookahead converges is further expanded by exploiting that Lookahead inherits the properties of the base optimizer. We corroborate the results with experiments on generative adversarial networks which demonstrates the benefits of the linear interpolation present in both RAPP and Lookahead.

1 Introduction

Stability is a major concern when training large scale models. In particular, generative adversarial networks (GANs) are known to be notoriously difficult to train. To stabilize training, the Lookahead algorithm of Zhang et al. (2019) was recently proposed for GANs Chavdarova et al. (2020) which linearly interpolates with a slow moving iterate. The mechanism has enjoyed superior empirical performance in both minimization and minimax problems, but it largely remains a heuristic with little theoretical motivation.

One major obstacle for providing a theoretical treatment, is in capturing the (fuzzy) notion of stability. Loosely speaking, a training dynamics is referred to as *unstable* in practice when the iterates either cycle indefinitely or (eventually) diverge—as has been observed for the Adam optimizer (see e.g. Gidel et al. (2018, Fig. 12) and Chavdarova et al. (2020, Fig. 6) respectively). Conversely, a *stable* dynamics has some bias towards stationary points. The notion of stability so far (e.g. in Chavdarova et al. (2020, Thm. 2-3)) is based on the spectral radius and thus inherently *local*.

In this work, we are interested in establishing *global* convergence properties, in which case some structural assumptions are needed. One (nonmonotone) structure that lends itself well to the study of stability is that of cohypomonotonicity studied in Combettes & Pennanen (2004); Diakonikolas et al. (2021), since even the extragradient method has been shown to cycle and diverge in this problem class (see Pethick et al. (2022, Fig. 1) and Pethick et al. (2023, Fig. 2) respectively). We provide a geometric intuition behind these difficulties in Figure 1. Biasing the optimization schemes towards stationary points becomes a central concern and we demonstrate in Figure 2 that Lookahead can indeed converge for such nonmonotone problems.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Table 1: Overview of last iterate results with our contribution highlighted in blue. Prior to this work there existed no rates for 1-SCLI schemes handling ρ-comonotone problems with $\rho \in (-1/2L, \infty)$ and no global convergence guarantees for Lookahead beyond bilinear games.

	Method	Setting	ρ	Handles constraints	ρ-independent rates	Reference
Implicit	PP	Comonotone	$(-1/2L, \infty)$	✓	✗	(Gorbunov et al., 2022b, Thm. 3.1)
	Relaxed PP	Comonotone	$(-1/2L, \infty)$	✓	✓	Theorem 6.2
Extrapolate	EG	Comonotone & Lips.	$(-1/8L, \infty)$	✗	✗	(Gorbunov et al., 2022b, Thm. 4.1)
	EG+	Comonotone & Lips.			Unknown rates	
	RAPP	Comonotone & Lips.	$(-1/2L, \infty)$	✓	✓	Corollary 6.4
Lookahead	LA-GDA	Local	-	✗	-	(Chavdarova et al., 2020, Thm. 2)
		Bilinear	-	✗	-	(Ha & Kim, 2022, Cor. 7)
		Comonotone & Lips.	$(-1/3\sqrt{3}L, \infty)$	✗	-	Theorem 7.1
	LA-EG	Bilinear	-	✗	-	(Ha & Kim, 2022, Cor. 8)
		Monotone & Lips.	-	✓	-	Theorem F.1
	LA-CEG+	Comonotone & Lips.	$(-1/2L, \infty)$	✓	-	Corollary 7.7

A principled approach to cohypomonotone problems is the extragradient+ algorithm (EG+) proposed by Diakonikolas et al. (2021). However, the only known rates are on the best iterate, which can be problematic to pick in practice. It is unclear whether *last* iterate rates for EG+ are possible even in the monotone case (see discussion prior to Thm. 3.3 in Gorbunov et al. (2022a)). For this reason, the community has instead resorted to showing last iterate of extragradient (EG) method of Korpelevich (1977), despite originally being developed for the monotone case. Maybe not surprisingly, EG only enjoys a last iterate guarantee under mild form of cohypomonotonicity and have so far only been studied in the unconstrained case (Luo & Tran-Dinh; Gorbunov et al., 2022b). Recently, last iterate rate were established for the same (tight) range of cohypomonotone problems for which EG+ has best iterate guarantees. However, the analyzed scheme is *implicit* and the complexity blows up with increasing cohypomonotonicity (Gorbunov et al., 2022b). This leaves the questions: *Can an explicit scheme enjoy last iterate rates for the same range of cohypomonotone problems? Can the rate be agnostic to the degree of cohypomonotonicity?* We answer both in the affirmative.

This work focuses on 1-SCLI schemes (Arjevani et al., 2015; Golowich et al., 2020), whose update rule only depends on the previous iterate in a time-invariant fashion. Another approach to establishing last iterate is Halpern-type methods with an explicit scheme developed in Lee & Kim (2021) for cohypomonotone problems and later extended to the constrained case in Cai et al. (2022) (c.f. Appendix A).

As will become clear, a principled mechanism behind convergence in this nonmonotone class is the linear interpolation also used in Lookahead. This iterative interpolation is more broadly referred to as the Krasnosel'skiĭ-Mann (KM) iteration in the theory of nonexpansive operators. We show that the extragradient+ algorithm (EG+) of Diakonikolas et al. (2021), our proposed relaxed approximate proximal point method (RAPP), and Lookahead based algorithms are all instances of the (inexact) KM iteration and provide simple proofs of these schemes in the cohypomonotone case.

More concretely we make the following contributions:

1. We prove global convergence rates for the last iterate of our proposed algorithm RAPP which additionally handles constrained and regularized settings. This makes RAPP the first 1-SCLI scheme to have non-asymptotic guarantees for ρ-comonotone problems while only requiring $\rho > -1/2L$. As a byproduct we obtain a last iterate convergence rate for an implicit scheme that is *independent* of the degree of cohypomonotonicity. The last iterate rates are established by showing monotonic decrease of the operator norm–something which is not possible for EG+. This contrast is maybe surprising, since RAPP can be viewed as an extension of EG+, which simply takes multiple extrapolation steps.
2. By replacing the inner optimization routine in RAPP with gradient descent ascent (GDA) and extragradient (EG) we rediscover the Lookahead algorithms considered in Chavdarova et al. (2020). We obtain guarantees for the Lookahead variants by deriving nonexpansive properties of the base optimizers. By casting Lookahead as a KM iteration we find that the optimal interpolation constant is $\lambda = 0.5$. This choice corresponds to the default value used in practice for both minimization and minimax—thus providing theoretical motivation for the parameter value.

3. For $\tau = 2$ inner iterations we observe that LA-GDA reduces to a linear interpolation between GDA and EG+ which allows us to obtain global convergence in ρ-comonotone problems when $\rho > -1/3\sqrt{3}L$. However, for τ large, we provide a counterexample showing that LA-GDA cannot be guaranteed to converge. This leads us to instead propose LA-CEG+ which corrects the inner optimization to guarantee global convergence for ρ-comonotone problems when $\rho > -1/2L$.
4. We test the methods on a suite of synthetic examples and GAN training where we confirm the stabilizing effect. Interestingly, RAPP seems to provide a similar benefit as Lookahead, which suggest that linear interpolation could play a key role also experimentally.

An overview of the theoretical results is provided in Table 1 and Figure 5§B.

2 Related work

Lookahead The Lookahead algorithm was first introduced for minimization in Zhang et al. (2019). In the context of Federated Averaging in federated learning (McMahan et al., 2017) and the Reptile algorithm in meta-learning (Nichol et al., 2018), the method can be seen as a single worker and single task instance respectively. Analysis for Lookahead was carried out for nonconvex minimization (Wang et al., 2020; Zhou et al., 2021) and a nested variant proposed in (Pushkin & Barba, 2021). Chavdarova et al. (2020) popularized the Lookahead algorithm for minimax training by showing state-of-the-art performance on image generation tasks. Apart from the original local convergence analysis in Chavdarova et al. (2020) and the bilinear case treated in Ha & Kim (2022) we are not aware of any convergence analysis for Lookahead for minimax problems and beyond.

Cohypomonotone Cohypomontone problems were first studied in Iusem et al. (2003); Combettes & Pennanen (2004) for proximal point methods and later expanded on in greater detail in Bauschke et al. (2021). The condition was relaxed to the star-variant referred to as the weak Minty variational inequality (MVI) in Diakonikolas et al. (2021) and the extragradient+ algorithm (EG+) was analyzed. The analysis of EG+ was later tightened and extended to the constrained case in Pethick et al. (2022).

Proximal point The proximal point method (PP) has a long history. For maximally monotone operators (and thus convex-concave minimax problems) convergence of PP follows from Opial (1967). The first convergence analysis of *inexact* PP dates back to Rockafellar (1976); Brézis & Lions (1978). It was later shown that convergence also holds for the *relaxed* inexact PP as defined in (8) (Eckstein & Bertsekas, 1992). In recent times, PP has gained renewed interest due to its success for certain nonmonotone structures. Inexact PP was studied for cohypomontone problems in Iusem et al. (2003). Asymptotic convergence was established of the relaxed inexact PP for a sum of cohypomonotone operators in Combettes & Pennanen (2004), and later considered in Grimmer et al. (2022) without inexactness. Last iterate rates were established for PP in ρ-comonotone problems (with a dependency on ρ) (Gorbunov et al., 2022b). Explicit approximations of PP through a contractive map was used for convex-concave minimax problems in Cevher et al. (2023) and was the original motivation for MirrorProx of Nemirovski (2004). See Appendix A for additional references in the stochastic setting.

3 Setup

We are interested in finding a zero of an operator $S : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ which decomposes into a Lipschitz continuous (but possibly nonmonotone) operator F and a maximally monotone operator A, i.e. find $z \in \mathbb{R}^d$ such that,

$$0 \in Sz := Az + Fz. \tag{1}$$

Most relevant in the context of GAN training is that (1) includes constrained minimax problems.

Example 3.1. *Consider the following minimax problem*

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \phi(x, y). \tag{2}$$

The problem can be recast as the inclusion problem (1) by defining the joint iterates $z = (x, y)$, the stacked

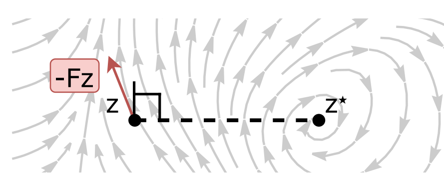

Figure 1: Consider $\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \phi(z)$ with $z = (x, y)$. As opposed to convex-concave minimax problems, the cohypomonotone condition allows the gradients $Fz = (\nabla_x\phi(z), -\nabla_y\phi(z))$ to point away from the solutions (see Appendix B.1 for the relationship between cohypomonotonicity and the weak MVI). This can lead to instability issues for standard algorithms such as the Adam optimizer.

gradients $Fz = (\nabla_x \phi(x,y), -\nabla_y \phi(x,y))$, and $A = (\mathcal{N}_\mathcal{X}, \mathcal{N}_\mathcal{Y})$ where $\mathcal{N}$ denotes the normal cone. As will become clear (cf. Algorithm 1), A will only be accessed through the resolvent $J_{\gamma A} := (\operatorname{id} + \gamma A)^{-1}$ which reduces to the proximal operator. More specifically $J_{\gamma A}(z) = (\operatorname{proj}_\mathcal{X}(x), \operatorname{proj}_\mathcal{Y}(y))$.

We will rely on the following assumptions (see Appendix B for any missing definitions).

Assumption 3.2. *In problem (1),*

(i) The operator $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximally monotone.

(ii) The operator $F : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz, i.e. for some $L \in [0, \infty)$,

$$\|Fz - Fz'\| \le L\|z - z'\| \quad \forall z, z' \in \mathbb{R}^d.$$

(iii) The operator $S := F + A$ is ρ-comonotone for some $\rho \in (-1/2L, \infty)$, i.e.

$$\langle v - v', z - z' \rangle \ge \rho \|v - v'\|^2 \quad \forall (v, z), (v', z') \in \operatorname{grph} S.$$

Remark 3.3. Assumption 3.2(iii) is also known as $|\rho|$-cohypomonotonicity when $\rho < 0$, which allows for increasing nonmonotonicity as $|\rho|$ grows. See Appendix B.1 for the relationship with weak MVI.

When only stochastic feedback $\hat{F}_\sigma(\cdot, \xi)$ is available we make the following classical assumptions.

Assumption 3.4. *For the operator $\hat{F}_\sigma(\cdot, \xi) : \mathbb{R}^d \to \mathbb{R}^d$ the following holds.*

(i) Unbiased: $\mathbb{E}_\xi\big[\hat{F}_\sigma(z, \xi)\big] = Fz \quad \forall z \in \mathbb{R}^d$.

(ii) Bounded variance: $\mathbb{E}_\xi\big[\|\hat{F}_\sigma(z, \xi) - Fz\|^2\big] \le \sigma^2 \quad \forall z, z' \in \mathbb{R}^d$.

4 Inexact Krasnosel'skiĭ-Mann iterations

The main work horse we will rely on is the inexact Krasnosel'skiĭ-Mann (IKM) iteration from monotone operators (also known as the *averaged* iteration), which acts on an operator $T : \mathbb{R}^d \to \mathbb{R}^d$ with inexact feedback,

$$z^{k+1} = (1 - \lambda) z^k + \lambda (T z^k + e^k), \tag{IKM}$$

where $\lambda \in (0, 1)$ and e^k is a random variable with dependency on all variables up until (and including) k. The operator $\widetilde{T}_k : z \mapsto Tz + e^k$ can crucially be an iterative optimization scheme in itself. This is important, since we can obtain RAPP, LA-GDA and LA-CEG+ by plugging in different optimization routines. In fact, RAPP is derived by taking $\widetilde{T}_k$ to be a (contractive) fixed point iteration in itself, which approximates the resolvent.

We note that also the extragradient+ (EG+) method of Diakonikolas et al. (2021), which converges for cohypomonotone and Lipschitz problems, can be seen as a Krasnosel'skiĭ-Mann iteration on an extragradient step

$$\begin{aligned} \operatorname{EG}(z) &= z - \gamma F(z - \gamma F z) \\ z^{k+1} &= (1 - \lambda) z^k + \lambda \operatorname{EG}(z^k) \end{aligned} \tag{EG+}$$

where $\lambda \in (0, 1)$. We provide a proof of EG+ in Theorem G.1 which extends to the constrained case using the construction from Pethick et al. (2022) but through a simpler argument under fixed stepsize.

Essentially, the IKM iteration leads to a conservative update that stabilizes the update using the previous iterate. This is the key mechanism behind showing convergence in the nonmonotone setting known as cohypomonotonicity. Very generally, it is possible to provide convergence guarantees for IKM when the following holds (Theorem C.1 is deferred to the appendix due to space limitations).

Definition 4.1. *An operator $T : \mathbb{R}^n \to \mathbb{R}^d$ is said to be quasi-nonexpansive if*

$$\|Tz - z'\| \le \|z - z'\| \quad \forall z \in \mathbb{R}^d, \forall z' \in \operatorname{fix} T. \tag{3}$$

Remark 4.2. This notion is crucial to us since the resolvent $J_B := (\operatorname{id} + B)^{-1}$ is (quasi)-nonexpansive if B is $1/2$-cohypomonotone (Bauschke et al., 2021, Prop. 3.9(iii)).

5 Approximating the resolvent

As apparent from Remark 4.2, the IKM iteration would provide convergence to a zero of the cohypomonotone operator S from Assumption 3.2 by using its resolvent $T = J_{\gamma S}$. However, the update is implicit, so we will instead approximate $J_{\gamma S}$. Given $z \in \mathbb{R}^d$ we seek $z' \in \mathbb{R}^d$ such that

$$z' = J_{\gamma S}(z) = (\mathrm{id} + \gamma S)^{-1} z = (\mathrm{id} + \gamma A)^{-1}(z - \gamma F z')$$

This can be approximated with a fixed point iteration of

$$Q_z : w \mapsto (\mathrm{id} + \gamma A)^{-1}(z - \gamma F w) \tag{4}$$

which is a contraction for small enough γ since F is Lipschitz continuous. It follows from Banach's fixed-point theorem Banach (1922) that the sequence converges linearly. We formalize this in the following theorem, which additionally applies when only stochastic feedback is available.

$$w^{t+1} = (\mathrm{id} + \gamma A)^{-1}(z - \gamma \hat{F}_\sigma(w^t, \xi_t)) \quad \xi_t \sim \mathcal{P} \tag{5}$$

Lemma 5.1. *Suppose Assumptions 3.2(i), 3.2(ii) and 3.4. Given $z \in \mathbb{R}^d$, the iterates generated by (5) with $\gamma \in (0, 1/L)$ converges to a neighborhood linearly, i.e.,*

$$\mathbb{E}\big[\|w^\tau - J_{\gamma S}(z)\|^2\big] \leq (\gamma L)^{2\tau} \|w^0 - w^\star\|^2 + \tfrac{\gamma^2}{(1-\gamma L)^2} \sigma^2. \tag{6}$$

The resulting update in (5) is identical to GDA but crucially always steps from z. We use this as a subroutine in RAPP to get convergence under a cohypomonotone operator while only suffering a logarithmic factor in the rate.

Interpretation In the special case of the constrained minimax problem in (2), the application of the resolvent $J_{\gamma S}(z)$ is equivalent to solving the following optimization problem

$$\min_{x' \in \mathcal{X}} \max_{y' \in \mathcal{Y}} \Big\{ \phi_\mu(x', y') := \phi(x', y') + \frac{1}{2\mu} \|x' - x\|^2 - \frac{1}{2\mu} \|y' - y\|^2 \Big\}. \tag{7}$$

for appropriately chosen $\mu \in (0, \infty)$. (5) can thus be interpreted as solving a particular regularized subproblem. Later we will drop this regularization to arrive at the Lookahead algorithm.

6 Last iterate under cohypomonotonicity

As stated in Section 5, we can obtain convergence using the approximate resolvent through Theorem C.1. The convergence is provided in terms of the average, so additional work is needed for a last iterate result. IKM iteration on the approximate resolvent (i.e. $\widetilde{T}_k(z) = J_{\gamma S}(z) + e^k$) becomes,

$$\bar{z}^k = z^k - v^k \quad \text{with} \quad v^k \in \gamma S(\bar{z}^k) \tag{8a}$$

$$z^{k+1} = (1 - \lambda) z^k + \lambda(\bar{z}^k + e^k) \tag{8b}$$

with $\lambda \in (0, 1)$ and $\gamma > 0$ and error $e^k \in \mathbb{R}^d$. Without error, (8) reduces to relaxed proximal point

$$z^{k+1} = (1 - \lambda) z^k + \lambda J_{\gamma S}(z^k) \tag{Relaxed PP}$$

For a last iterate result it remains to argue that the residual $\|J_{\gamma S}(z^k) - z^k\|$ is monotonically decreasing (up to an error we can control). Showing monotonic decrease is fairly straightforward if $\lambda = 1$ (see Lemma E.1 and the associated proof). However, we face additional complication due to the averaging, which is apparent both from the proof and the slightly more complicated error term in the following lemma.

Lemma 6.1. *If S is ρ-comonotone with $\rho > -\frac{\gamma}{2}$ then (8) satisfies for all $z^\star \in \operatorname{zer} S$,*

$$\|J_{\gamma S}(z^k) - z^k\|^2 \leq \|J_{\gamma S}(z^{k-1}) - z^{k-1}\|^2 + \delta_k(z^\star)$$

where $\delta_k(z) := 4\|e^k\|(\|z^{k+1} - z\| + \|z^k - z\|)$.

The above lemma allows us to obtain last iterate convergence for IKM on the inexact resolvent by combing the lemma with Theorem C.1.

Algorithm 1 Relaxed approximate proximal point method (RAPP)

Require: $z^0 \in \mathbb{R}^n$ $\lambda \in (0,1)$, $\gamma \in (\lfloor -2\rho \rfloor_+, {}^1/L)$
Repeat for $k = 0, 1, \ldots$ until convergence
1: $w_k^0 = z^k$
2: **for all** $t = 0, 1, \ldots, \tau - 1$ **do**
3: $\quad \xi_{k,t} \sim \mathcal{P}$
4: $\quad w_k^{t+1} = (\mathrm{id} + \gamma A)^{-1}(z^k - \gamma \hat{F}_{\sigma_k}(w_k^t, \xi_{k,t}))$
5: $z^{k+1} = (1-\lambda) z^k + \lambda w_k^\tau$
Return z^{k+1}

Theorem 6.2 (Last iterate of inexact resolvent)**.** *Suppose Assumptions 3.2 and 3.4 with σ_k. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by (8) with $\lambda \in (0,1)$ and $\rho > -\frac{\gamma}{2}$. Then, for all $z^\star \in \operatorname{zer} S$,*

$$\mathbb{E}[\|J_{\gamma S}(z^K) - z^K\|^2] \le \frac{\|z^0 - z^\star\|^2 + \sum_{k=0}^{K-1} \varepsilon_k(z^\star)}{\lambda(1-\lambda)K} + \frac{1}{K}\sum_{k=0}^{K-1}\sum_{j=k}^{K-1} \delta_j(z^\star),$$

where $\varepsilon_k(z) := 2\lambda \mathbb{E}[\|e^k\|\|z^k - z\|] + \lambda^2 \mathbb{E}[\|e^k\|^2]$ and $\delta_k(z) := 4\mathbb{E}[\|e^k\|(\|z^{k+1} - z\| + \|z^k - z\|)]$.

Remark 6.3. Notice that the rate in Theorem 6.2 has *no* dependency on ρ. Specifically, it gets rid of the factor $\gamma/(\gamma + 2\rho)$ which Gorbunov et al. (2022b, Thm. 3.2) shows is unimprovable for PP. Theorem 6.2 requires that the iterates stays bounded. In Corollary 6.4 we will assume bounded diameter for simplicity, but it is relatively straightforward to show that the iterates can be guaranteed to be bounded by controlling the inexactness (see Lemma E.2).

All that remains to get convergence of the explicit scheme in RAPP, is to expand and simplify the errors $\varepsilon_k(z)$ and $\delta_k(z)$ using the approximation of the resolvent analyzed in Lemma 5.1.

Corollary 6.4 (Explicit inexact resolvent)**.** *Suppose Assumption 3.2 holds. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by RAPP with deterministic feedback and $\rho > -\frac{\gamma}{2}$. Then, for all $z^\star \in \operatorname{zer} S$ with $D := \sup_{j\in\mathbb{N}} \|z^j - z^\star\| < \infty$,*

(i) with $\tau = \frac{\log K}{\log(1/\gamma L)}$: $\frac{1}{K}\sum_{i=0}^{K-1} \|J_{\gamma S}(z^k) - z^k\|^2 = \mathcal{O}\Big(\frac{\|z^0 - z^\star\|^2}{\lambda(1-\lambda)K} + \frac{D^2}{(1-\lambda)K}\Big)$.

(ii) with $\tau = \frac{\log K^2}{\log(1/\gamma L)}$: $\|J_{\gamma S}(z^K) - z^K\|^2 = \mathcal{O}\Big(\frac{\|z^0 - z^\star\|^2}{\lambda(1-\lambda)K} + \frac{D^2}{K} + \frac{D^2}{(1-\lambda)K^2},\Big)$.

Remark 6.5. Corollary 6.4(ii) implies an oracle complexity of $\mathcal{O}\big(\log(\varepsilon^{-2})\varepsilon^{-1}\big)$ for ensuring that the last iterate satisfies $\|J_{\gamma S}(z^K) - z^K\|^2 \le \varepsilon$. A stochastic extension is provided in Corollary E.3 by taking the batch size increasing. Notice that RAPP, for $\tau = 2$ inner steps, reduces to EG+ in the unconstrained case where $A \equiv 0$.

7 Analysis of Lookahead

The update in RAPP leads to a fairly conservative update in the inner loop, since it corresponds to optimizing a highly regularized subproblem as noted in Section 5. Could we instead replace the optimization procedure with gradient descent ascent (GDA)? If we replace the inner optimization routine we recover what is known as the Lookahead (LA) algorithm

$$\begin{aligned} w_k^0 &= z^k \\ w_k^{t+1} &= w_k^t - \gamma F w_k^t \quad \forall t = 0, ..., \tau - 1 \\ z^{k+1} &= (1-\lambda) z^k + \lambda w_k^\tau \end{aligned} \tag{LA-GDA}$$

We empirically demonstrate that this scheme can converge for nonmonotone problems for certain choices of parameters (see Figure 2). However, what global guarantees can we provide theoretically?

It turns out that for LA-GDA with two inner steps ($\tau = 2$) we have an affirmative answer. After some algebraic manipulation it is not difficult to see that the update can be simplified as follows

$$z^{k+1} = \tfrac{1}{2}(z^k - 2\lambda\gamma F z^k) + \tfrac{1}{2}(z^k - 2\lambda\gamma F(z^k - \gamma F z^k)). \tag{9}$$

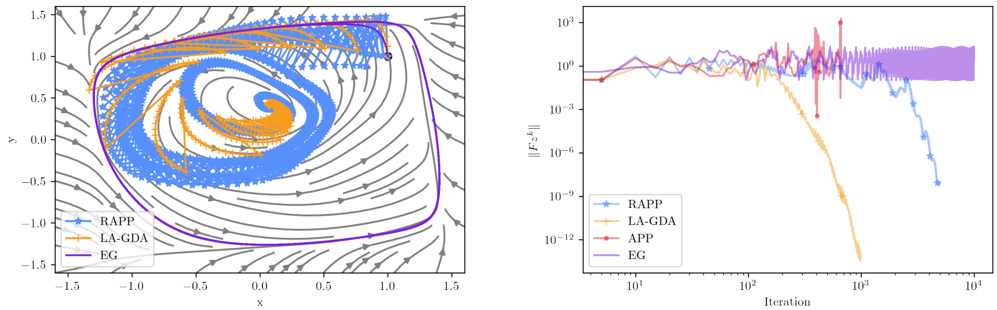

Figure 2: LA-GDA and RAPP can converge for Hsieh et al. (2021, Ex. 5.2). Interestingly, we can set the stepsize γ larger than $1/L$ while RAPP remains stable. Approximate proximal point (APP) with the same stepsize diverges (the iterates of APP are deferred to Figure 6). In this example, it is apparent from the rates, that there is a benefit in replacing the conservative inner update in RAPP with GDA in LA-GDA as explored in Section 7.

This is the average of GDA and EG+ (when $\lambda \in (0, 1/2)$). This observation allows us to show convergence under cohypomonotonicity. This positive result for nonmonotone problems partially explains the stabilizing effect of LA-GDA.

Theorem 7.1. *Suppose Assumption 3.2 holds. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by LA-GDA with $\tau = 2$, $\gamma \leq 1/L$ and $\lambda \in (0, 1/2)$. Furthermore, suppose that*

$$2\rho > -(1-2\lambda)\gamma \quad \textit{and} \quad 2\rho \geq 2\lambda\gamma - (1-\gamma^2L^2)\gamma. \tag{10}$$

Then, for all $z^\star \in \operatorname{zer} F$,

$$\frac{1}{K}\sum_{k=0}^{K-1}\|F\bar{z}^k\|^2 \leq \frac{\|z^0 - z^\star\|^2}{\lambda\gamma\big((1-2\lambda)\gamma + 2\rho\big)K}. \tag{11}$$

Remark 7.2. For $\lambda \to 0$ and $\gamma = c/L$ where $c \in (0, \infty)$, sufficient condition reduces to $\rho \geq -\gamma(1-\gamma^2L^2)/2 = -c(1-c^2)/2L$, of which the minimum is attained with $c = 1/\sqrt{3}$, leading to the requirement $\rho \geq -1/3\sqrt{3}L$. A similar statement is possible for z^k. Thus, (LA-GDA) improves on the range of ρ compared with EG (see Table 1).

For larger τ, LA-GDA does not necessarily converge (see Figure 3 for a counterexample). We next ask what we would require of the base optimizer to guarantee convergence for any τ. To this end, we replace the inner iteration with some abstract algorithm $\operatorname{Alg} : \mathbb{R}^d \to \mathbb{R}^d$, i.e.

$$\begin{aligned} w_k^0 &= z^k \\ w_k^{t+1} &= \operatorname{Alg}(w_k^t) \quad \forall t = 0, ..., \tau-1 \\ z^{k+1} &= (1-\lambda)z^k + \lambda w_k^\tau \end{aligned} \tag{LA}$$

Convergence follows from quasi-nonexpansiveness.

Theorem 7.3. *Suppose $\operatorname{Alg} : \mathbb{R}^d \to \mathbb{R}^d$ is quasi-nonexpansive. Then $(z^k)_{k\in\mathbb{N}}$ generated by (LA) converges to some $z^\star \in \operatorname{fix}\operatorname{Alg}$.*

Remark 7.4. Even though the base optimizer Alg might not converge (since nonexpansiveness is not sufficient), Theorem 7.3 shows that the outer loop converges. Interestingly, this aligns with the benefit observed in practice of using the outer iteration of Lookahead (see Figure 4).

Cocoercive From Theorem 7.3 we almost immediately get converge of LA-GDA for coercive problems since $V = \operatorname{id} - \gamma F$ is nonexpansive iff γF is $1/2$-cocoercive.

Corollary 7.5. *Suppose F is $1/L$-cocoercive. Then $(z^k)_{k\in\mathbb{N}}$ generated by LA-GDA with $\gamma \leq 2/L$ converges to some $z^\star \in \operatorname{zer} F$.*

Remark 7.6. Corollary 7.5 can trivially be extended to the constrained case by observing that also $V = (\operatorname{id} + \gamma A)^{-1}(\operatorname{id} - \gamma F)$ is nonexpansive when A is maximally monotone. As a special case this captures constrained convex and gradient Lipschitz minimization problems.

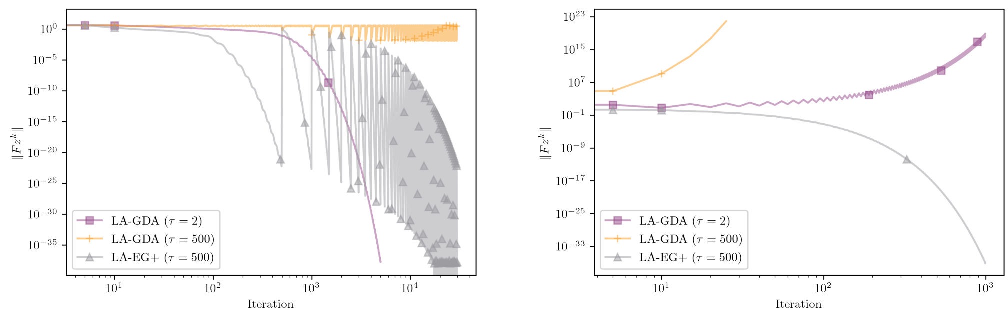

Figure 3: We test the Lookahead variants on Pethick et al. (2022, Ex. 3(iii)) where $\rho \in (-1/8L, -1/10L)$ (left) and Pethick et al. (2022, Ex. 5) with $\rho = -1/3$ (right). For the left example LA-GDA (provably) converges for $\tau = 2$, but may be nonconvergent for larger τ as illustrate. Both variants of LA-GDA diverges in the more difficult example on the right, while LA-CEG+ in contrast provably converges. It seems that LA-CEG+ trades off a constant slowdown in the rate for convergence in a larger class.

Monotone When only monotonicity and Lipschitz holds we may instead consider the following extragradient based version of Lookahead (first empirically investigated in Chavdarova et al. (2020))

$$\begin{aligned} w_k^0 &= z^k \\ w_k^{t+1} &= \mathrm{EG}(w_t^k) \quad \forall t = 0, ..., \tau - 1 \\ z^{k+1} &= (1-\lambda)z^k + \lambda w_k^\tau \end{aligned} \tag{LA-EG}$$

where $\mathrm{EG}(z) = z - \gamma F(z - \gamma F z)$. We show in Theorem F.1 that the EG-operator of the inner loop is quasi-nonexpansive, which implies convergence of LA-EG through Theorem 7.3. Theorem F.1 extends even to cases where $A \not\equiv 0$ by utilizing the forward-backward-forward construction of Tseng (1991). This providing the first global convergence guarantee for Lookahead beyond bilinear games.

Cohypomonotone For cohypomonotone problems large τ may prevent LA-GDA from converging (see Figure 3 for a counterexample). Therefore we propose replacing the inner optimization loop in LA-GDA with the method proposed in (Pethick et al., 2022, Alg. 1). Let $H = \mathrm{id} - \gamma F$. We can write one step of the inner update with $\alpha \in (0, 1)$ as

$$\mathrm{CEG}^+(w) = w + 2\alpha(H\bar{w} - Hw) \quad \text{with} \quad \bar{w} = (\mathrm{id} + \gamma A)^{-1} Hw. \tag{12}$$

The usefulness of the operator $\mathrm{CEG}^+ : \mathbb{R}^d \to \mathbb{R}^d$ comes from the fact that it is quasi-nonexpansive under Assumption 3.2 (see Theorem G.1). Thus, Theorem 7.3 applies even when F is only cohypomonotone if we make the following modification to LA-GDA

$$\begin{aligned} w_k^0 &= z^k \\ w_k^{t+1} &= \mathrm{CEG}^+(w_k^t) \quad \forall t = 0, ..., \tau - 1 \\ z^{k+1} &= (1-\lambda)z^k + \lambda w_k^\tau \end{aligned} \tag{LA-CEG+}$$

In the unconstrained case ($A \equiv 0$) this reduces to using the EG+ algorithm of Diakonikolas et al. (2021) for the inner loop. We have the following convergence guarantee.

Corollary 7.7. *Suppose Assumption 3.2 holds. Then $(z^k)_{k\in\mathbb{N}}$ generated by LA-CEG+ with $\lambda \in (0, 1)$, $\gamma \in (\lfloor -2\rho \rfloor_+, 1/L)$ and $\alpha \in (0, 1 + \frac{2\rho}{\gamma})$ converges to some $z^\star \in \operatorname{zer} S$.*

8 Experiments

This section demonstrates that linear interpolation can lead to an improvement over common baselines.

Synthetic examples Figures 2 and 3 demonstrate RAPP, LA-GDA and LA-CEG+ on a host of nonmonotone problems (Hsieh et al. (2021, Ex. 5.2), Pethick et al. (2022, Ex. 3(iii)), Pethick et al. (2022, Ex. 5)). See Appendix H.2 for definitions and further details.

Table 2: Adam-based. The combination of Lookahead and extragradient-like methods performs the best.

	FID	ISC
Adam	21.04±2.20	7.61±0.15
ExtraAdam	18.23±1.13	7.79±0.08
ExtraAdam+	22.94±1.93	7.65±0.13
LA-Adam	17.63±0.65	7.86±0.07
LA-ExtraAdam	**15.88±0.67**	7.97±0.12
LA-ExtraAdam+	17.86±1.03	**8.08±0.15**

Table 3: GDA-based. Both RAPP and Lookahead increases the scores substantially.

	FID	ISC
GDA	19.36±0.08	7.84±0.07
EG	18.94±0.60	7.84±0.02
EG+	19.35±4.28	7.74±0.44
LA-GDA	**16.87±0.18**	**8.01±0.08**
LA-EG	16.91±0.66	7.97±0.12
LA-EG+	17.20±0.44	7.94±0.11
RAPP	17.76±0.82	7.98±0.08

Image generation We replicate the experimental setup of Chavdarova et al. (2020); Miyato et al. (2018), which uses hinge version of the non-saturated loss and a ResNet with spectral normalization for the discriminator (see Appendix H.2 for details). To evaluate the performance we rely on the commonly used Inception score (ISC) (Salimans et al., 2016) and Fréchet inception distance (FID) (Heusel et al., 2017) and report the best iterate. We demonstrate the methods on the CIFAR10 dataset (Krizhevsky et al., 2009). The aim is *not* to beat the state-of-the-art, but rather to complement the already exhaustive numerical evidence provided in Chavdarova et al. (2020).

For a fair *computational* comparison we count the number of *gradient computations* instead of iterations k as in Chavdarova et al. (2020). Maybe surprisingly, we find that the extrapolation methods such as EG and RAPP still outperform the baseline, despite having fewer effective iterations. RAPP improves over EG, which suggest that it can be worthwhile to spend more computation on refining the updates at the cost of making fewer updates.

The first experiment we conduct matches the setting of Chavdarova et al. (2020) by relying on the Adam optimizer and using and update ratio of $5:1$ between the discriminator and generator. We find in Table 2 that LA-ExtraAdam+ has the highest ISC (8.08) while LA-ExtraAdam has the lowest FID (15.88). In contrast, we confirm that Adam is unstable while Lookahead prevents divergence as apparent from Figure 4, which is in agreement with Chavdarova et al. (2020). In addition, the *outer* loop of Lookahead achieves better empirical performance, which corroborate the theoretical result (cf. Remark 7.4). Notice that ExtraAdam+ has slow convergence (without Lookahead), which is possibly due to the $1/2$-smaller stepsize.

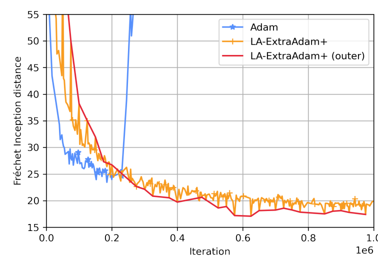

Figure 4: Adam eventually diverges on CIFAR10 while Lookahead is stable with the outer iterate enjoying superior performance.

We additionally simplify the setup by using GDA-based optimizers with an update ratio of $1:1$, which avoids the complexity of diagonal adaptation, gradient history and multiple steps of the discriminator as in the Adam-based experiments. The results are found in Table 3. The learning rates are tuned for GDA and we use those parameters fixed across all other methods. Despite being tuned on GDA, we find that extragradient methods, Lookahead-based methods and RAPP all *still* outperform GDA in terms of FID. The biggest improvement comes from the linear interpolation based methods Lookahead and RAPP (see Figure 8 for further discussion on EG+). Interesting, the Lookahead-based methods are roughly comparable with their Adam variants (Table 2) while GDA even performs better than Adam.

9 Conclusion & limitations

We have precisely characterized the stabilizing effect of linear interpolation by analyzing it under cohypomonotonicity. We proved last iterate convergence rates for our proposed method RAPP. The algorithm is double-looped, which introduces a log factor in the rate as mentioned in Remark E.4. It thus remains open whether last iterate is possible using only $\tau = 2$ inner iterations (for which RAPP reduces to EG+ in the unconstrained case). By replacing the inner solver we subsequently rediscovered and analyzed Lookahead using nonexpansive operators. In that regard, we have only dealt with compositions of operators. It would be interesting to further extend the idea to understanding and developing both Federated Averaging and the meta-learning algorithm Reptile (of which Lookahead can be seen as a single client and single task instance respectively), which we leave for future work.

10 Acknowledgements

Thanks to Leello Dadi and Stratis Skoulakis for helpful discussion. This work was supported by Google. This work was supported by Hasler Foundation Program: Hasler Responsible AI (project number 21043). This work was supported by the Swiss National Science Foundation (SNSF) under grant number 200021_205011. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 725594 - time-data).

References

Arjevani, Y., Shalev-Shwartz, S., and Shamir, O. On lower and upper bounds for smooth and strongly convex optimization problems. *arXiv preprint arXiv:1503.06833*, 2015.

Banach, S. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fund. math*, 3(1):133–181, 1922.

Bauschke, H. H. and Combettes, P. L. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics. Springer, 2017. ISBN 978-3-319-48310-8.

Bauschke, H. H., Moursi, W. M., and Wang, X. Generalized monotone operators and their averaged resolvents. *Mathematical Programming*, 189(1):55–74, 2021.

Bertsekas, D. P. Incremental proximal methods for large scale convex optimization. *Mathematical programming*, 129(2):163–195, 2011.

Bianchi, P. On the convergence of a stochastic proximal point algorithm. In *CAMSAP*, 2015.

Bravo, M. and Cominetti, R. Stochastic fixed-point iterations for nonexpansive maps: Convergence and error bounds. *arXiv preprint arXiv:2208.04193*, 2022.

Brézis, H. and Lions, P. L. Produits infinis de résolvantes. *Israel Journal of Mathematics*, 29(4): 329–345, 1978.

Cai, Y., Oikonomou, A., and Zheng, W. Accelerated algorithms for monotone inclusions and constrained nonconvex-nonconcave min-max optimization. *arXiv preprint arXiv:2206.05248*, 2022.

Cevher, V., Piliouras, G., Sim, R., and Skoulakis, S. Min-max optimization made simple: Approximating the proximal point method via contraction maps, 2023. URL https://arxiv.org/abs/2301.03931.

Chavdarova, T., Pagliardini, M., Stich, S. U., Fleuret, F., and Jaggi, M. Taming GANs with Lookahead-Minmax. *arXiv preprint arXiv:2006.14567*, 2020.

Combettes, P. L. Quasi-Fejérian analysis of some optimization algorithms. In *Studies in Computational Mathematics*, volume 8, pp. 115–152. Elsevier, 2001.

Combettes, P. L. and Pennanen, T. Proximal methods for cohypomonotone operators. *SIAM journal on control and optimization*, 43(2):731–742, 2004.

Diakonikolas, J. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *Conference on Learning Theory*, pp. 1428–1451. PMLR, 2020.

Diakonikolas, J., Daskalakis, C., and Jordan, M. I. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2746–2754. PMLR, 2021.

Eckstein, J. and Bertsekas, D. P. On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, 1992.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

Golowich, N., Pattathil, S., Daskalakis, C., and Ozdaglar, A. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *Conference on Learning Theory*, pp. 1758–1784. PMLR, 2020.

Gorbunov, E., Loizou, N., and Gidel, G. Extragradient method: O (1/k) last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *International Conference on Artificial Intelligence and Statistics*, pp. 366–402. PMLR, 2022a.

Gorbunov, E., Taylor, A., Horváth, S., and Gidel, G. Convergence of proximal point and extragradient-based methods beyond monotonicity: the case of negative comonotonicity. *arXiv preprint arXiv:2210.13831*, 2022b.

Grimmer, B., Lu, H., Worah, P., and Mirrokni, V. The landscape of the proximal point method for nonconvex–nonconcave minimax optimization. *Mathematical Programming*, pp. 1–35, 2022.

Ha, J. and Kim, G. On convergence of Lookahead in smooth games. In *International Conference on Artificial Intelligence and Statistics*, pp. 4659–4684. PMLR, 2022.

Halpern, B. Fixed points of nonexpanding maps. 1967.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Hsieh, Y.-P., Mertikopoulos, P., and Cevher, V. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In *International Conference on Machine Learning*, pp. 4337–4348. PMLR, 2021.

Iusem, A. N., Pennanen, T., and Svaiter, B. F. Inexact variants of the proximal point algorithm without monotonicity. *SIAM Journal on Optimization*, 13(4):1080–1097, 2003.

Korpelevich, G. Extragradient method for finding saddle points and other problems. *Matekon*, 13(4): 35–49, 1977.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lee, S. and Kim, D. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 34:22588–22600, 2021.

Lieder, F. On the convergence rate of the halpern-iteration. *Optimization letters*, 15(2):405–418, 2021.

Luo, Y. and Tran-Dinh, Q. Extragradient-type methods for co-monotone root-finding problems.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Nemirovski, A. 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(1):229–251, 2004.

Nichol, A., Achiam, J., and Schulman, J. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Obukhov, A., Seitzer, M., Wu, P.-W., Zhydenko, S., Kyl, J., and Lin, E. Y.-J. High-fidelity performance metrics for generative models in PyTorch, 2020. URL https://github.com/toshas/torch-fidelity. Version: 0.3.0, DOI: 10.5281/zenodo.4957738.

Opial, Z. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bulletin of the American Mathematical Society*, 73(4):591–597, 1967.

Patrascu, A. and Irofti, P. Stochastic proximal splitting algorithm for composite minimization. *Optimization Letters*, 15(6):2255–2273, 2021.

Patrascu, A. and Necoara, I. Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *The Journal of Machine Learning Research*, 18(1):7204–7245, 2017.

Pethick, T., Latafat, P., Patrinos, P., Fercoq, O., and Cevher, V. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In *International Conference on Learning Representations*, 2022.

Pethick, T., Fercoq, O., Latafat, P., Patrinos, P., and Cevher, V. Solving stochastic weak Minty variational inequalities without increasing batch size. In *The Eleventh International Conference on Learning Representations*, 2023.

Pushkin, D. and Barba, L. Multilayer Lookahead: a nested version of Lookahead. *arXiv preprint arXiv:2110.14254*, 2021.

Rockafellar, R. T. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training GANs. *Advances in neural information processing systems*, 29, 2016.

Toulis, P., Horel, T., and Airoldi, E. M. The proximal Robbins-Monro method. *arXiv preprint arXiv:1510.00967*, 2015.

Toulis, P., Tran, D., and Airoldi, E. Towards stability and optimality in stochastic gradient descent. In *Artificial Intelligence and Statistics*, pp. 1290–1298. PMLR, 2016.

Tseng, P. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM Journal on Control and Optimization*, 29(1):119–138, 1991.

Wang, J., Tantia, V., Ballas, N., and Rabbat, M. Lookahead converges to stationary points of smooth non-convex functions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8604–8608. IEEE, 2020.

Yoon, T. and Ryu, E. K. Accelerated algorithms for smooth convex-concave minimax problems with o (1/kˆ 2) rate on squared gradient norm. In *International Conference on Machine Learning*, pp. 12098–12109. PMLR, 2021.

Zhang, M., Lucas, J., Ba, J., and Hinton, G. E. Lookahead optimizer: k steps forward, 1 step back. *Advances in neural information processing systems*, 32, 2019.

Zhou, P., Yan, H., Yuan, X., Feng, J., and Yan, S. Towards understanding why lookahead generalizes better than SGD and beyond. *Advances in Neural Information Processing Systems*, 34:27290–27304, 2021.

Appendix

Table of Contents

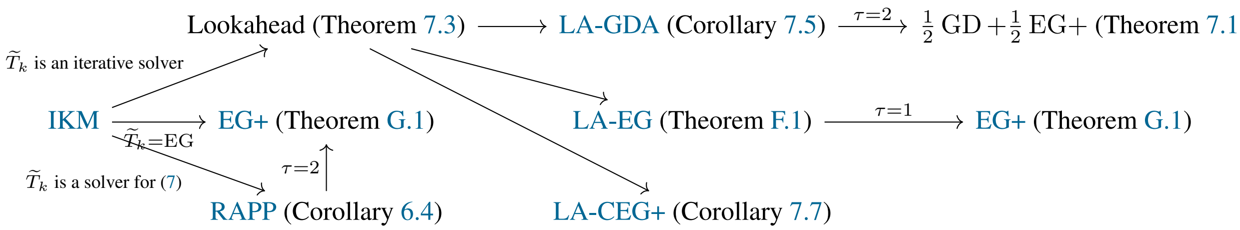

Figure 5: Overview of results and relationship between methods.

A Additional related work

Stochastic feedback There are several ways in which a stochastic variant of PP can be devised. *Incremental proximal methods* were pioneered for convex minimization in (Bertsekas, 2011), which uses an implicit update conditioned on the current randomness. Related approaches include Patrascu & Necoara (2017); Bianchi (2015); Patrascu & Irofti (2021); Toulis et al. (2016). Alternatively, (Toulis et al., 2015) assumes noisy access to the *full batch* implicit update in what they call the *proximal Robbins-Monro precedure*, which is similar to the approach taken in Bravo & Cominetti (2022) concerning Krasnoselskii-Mann iterations. Toulis et al. (2015) explicitly approximate the implicit update in the *proximal stochastic fixed-point* algorithm which is closely related to the approximation in Section 5. In the cohypomonotone case it is common to rely on increasing batchsizes (see e.g. (Diakonikolas et al., 2021, Thm. 4.5) and (Lee & Kim, 2021, Thm. 6.1)) similarly to Corollary E.3. Very recently, (Pethick et al., 2023) showed that convergence in stochastic weak MVI (and thus cohypomonotone problems) is possible for an extragradient-type scheme if the Lipschitz conditions are further tightened to a mean-squared smoothness assumption on the stochastic oracles.

Halpern-type Halpern iteration introduced in Halpern (1967), in contrast with IKM, linearly interpolates with the initial point using a time-varying stepsize, i.e. $z^{k+1} = (1-\lambda_k)z^0 - \lambda_k T z^k$. A $\mathcal{O}(1/k^2)$ convergence rate for the squared fixed point residual was shown in Lieder (2021) for nonexpansive operators. By directly approximating the Halpern iteration, an explicit scheme for monotone problems was later proposed in Diakonikolas (2020), but it suffered a logarithmic factor in the rate. The logarithmic factor was later removed by means of an extragradient variant (Yoon & Ryu, 2021). The scheme was extended to unconstrained cohypomonotone problems in Lee & Kim (2021) and subsequently the constrained case in Cai et al. (2022) while only requiring a single projection.

For a detailed discussion on how Halpern-type methods are not 1-SCLI algorithms see Yoon & Ryu (2021, Appendix E.2), which specifically addresses the anchored extragradient method. The extragradient method, on the other hand, can be written as an 1-SCLI algorithm (c.f. Golowich et al. (2020, Def. 5) and the subsequent discussion). This argument extends to the multistep extragradient construction used in RAPP.

B Preliminaries

The distance from $z \in \mathbb{R}^d$ to a set $\mathcal{Z} \subseteq \mathbb{R}^d$ is defined as $\operatorname{dist}(z, \mathcal{Z}) := \min_{z' \in \mathcal{Z}} \|z - z'\|$. The normal cone is defined as $\mathcal{N}_{\mathcal{Z}}(z) := \{ v \mid \langle v, z' - z \rangle \leq 0 \quad \forall z' \in \mathcal{Z} \}$ and the projection as $\mathbf{\Pi}_{\mathcal{Z}}(z) := \min_{w \in \mathcal{Z}} \|z - w\|^2$. We will denote the natural filtration up to iteration k as $\mathcal{F}_k$ and use $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_k]$.

We restate here some common definitions from monotone and nonexpansive operator for convenience (for further details see Bauschke & Combettes (2017)). An operator $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$ maps each point $z \in \mathbb{R}^d$ to a subset $Az \subseteq \mathbb{R}^n$, where the notation $A(z)$ and Az will be used interchangably. We denote the domain of A by $\operatorname{dom} A := \{z \in \mathbb{R}^d \mid Az \neq \emptyset\}$, its graph by $\operatorname{grph} A := \{(z, v) \in \mathbb{R}^d \times \mathbb{R}^n \mid v \in Az\}$. The inverse of A is defined through its graph, $\operatorname{grph} A^{-1} := \{(v, z) \mid (z, v) \in \operatorname{grph} A\}$ and the set of its zeros by $\operatorname{zer} A := \{z \in \mathbb{R}^d \mid 0 \in Az\}$. The set of fixed points is defined as $\operatorname{fix} T := \{z \in \mathbb{R}^d \mid z \in Tz\}$ for the operator $T : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$.

Definition B.1. *A single-valued operator $T: \mathbb{R}^d \to \mathbb{R}^d$ is said to be*

(i) nonexpansive if $\|Tz - Tz'\| \leq \|z - z'\| \quad \forall z, z' \in \mathbb{R}^d$.

(ii) quasi-nonexpansive if $\|Tz - z^\star\| \leq \|z - z^\star\| \quad \forall z \in \mathbb{R}^d$ and $\forall z^\star \in \operatorname{fix} T$.

(iii) firmly nonexpansive if $\|Tz - Tz'\|^2 \leq \|z - z'\|^2 - \|(z - z') - (Tz - Tz')\|^2 \quad \forall z, z' \in \mathbb{R}^d$.

The resolvent operator $J_A := (\operatorname{id} + A)^{-1}$ is firmly nonexpansive (with $\operatorname{dom} J_A = \mathbb{R}^d$) iff A is maximally monotone.

Definition B.2 ((co)monotonicity Bauschke et al. (2021))**.** *An operator $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is called monotone if,*

$$\langle v - v', z - z' \rangle \geq 0 \quad \forall (z, v), (z', v') \in \operatorname{grph} A,$$

and the operator A is called ρ-comonotone (also referred to as $|\rho|$-cohypomonotonicity when $\rho < 0$) if

$$\langle v - v', z - z' \rangle \geq \rho \|v - v'\|^2 \quad \forall (z, v), (z', v') \in \operatorname{grph} A.$$

The operator A is maximally *(co)monotone if no other (co)monotone operator B exists for which* $\operatorname{grph} A \subset \operatorname{grph} B$.

Definition B.3 (Lipschitz continuity and cocoercivity)**.** *Let $\mathcal{D} \subseteq \mathbb{R}^d$ be a nonempty set. A single-valued operator $A : \mathcal{D} \to \mathbb{R}^n$ is said to be L-Lipschitz continuous if for any $z, z' \in \mathcal{D}$*

$$\|Az - Az'\| \leq L \|z - z'\|,$$

and β-cocoercive if

$$\langle z - z', Az - Az' \rangle \geq \beta \|Az - Az'\|^2.$$

The forward step $H = \operatorname{id} - \gamma F$ is $1/2$-cocoercive when F is Lipschitz continuity and γ is sufficiently small.

Lemma B.4 (Pethick et al. (2022, Lm. A.3(i)))**.** *Suppose Assumption 3.2(ii) holds and $\gamma \leq 1/L$. Then, the mapping $H = \operatorname{id} - \gamma F$ is $1/2$-cocoercive for all $u \in \mathbb{R}^d$. Specifically,*

$$\langle Hz' - Hz, z' - z \rangle \geq \tfrac{1}{2} \|Hz' - Hz\|^2 + \tfrac{1}{2}(1 - \gamma^2 L^2) \|z' - z\|^2 \quad \forall z, z' \in \mathbb{R}^d. \tag{13}$$

Proof. By expanding,

$$Hz - Hz' = (z - z') - \gamma(Fz - Fz'). \tag{14}$$

Using (14) we get,

$$\begin{aligned} \langle Hz' - Hz, z' - z \rangle &= \langle Hz' - Hz, Hz' - Hz - \gamma(Fz - Fz') \rangle \\ \text{(14)} &= \tfrac{1}{2} \|Hz' - Hz\|^2 + \tfrac{1}{2} \|z' - z\|^2 - \tfrac{\gamma^2}{2} \|Fz - Fz'\|^2 \\ \text{(Assumption 3.2(ii))} &\geq \tfrac{1}{2} \|Hz' - Hz\|^2 + \tfrac{1}{2}(1 - \gamma^2 L^2) \|z' - z\|^2 \end{aligned} \tag{15}$$

This completes the proof. □

B.1 Relationship between weak Minty variational inequilities and cohypomonotonicity

Let $F: \mathbb{R}^d \to \mathbb{R}^d$ be a single-valued operator. In the unconstrained case, the weak Minty variational inequality (MVI) with parameter $\rho \in (-1/2L, \infty)$ is defined as

$$\langle Fz, z - z^\star \rangle \geq \rho \|Fz\|^2 \quad \forall z \in \mathbb{R}^d, \forall z^\star \in \operatorname{zer} F. \tag{16}$$

For $\rho < 0$, this condition allows the operator $-Fz$ to point away from the solution set as illustrated in Figure 1.

Notice that since $z^\star \in \operatorname{zer} F$ we could equivalently write

$$\langle Fz - Fz^\star, z - z^\star \rangle \geq \rho \|Fz - Fz^\star\|^2. \quad \forall z \in \mathbb{R}^d \tag{17}$$

In contrast, ρ-comonotonicity of F states that the above condition should hold for all pairs of point in the domain, i.e.

$$\langle Fz - Fz', z - z' \rangle \geq \rho \|Fz - Fz'\|^2 \quad \forall z, z' \in \mathbb{R}^d.$$

For $\rho < 0$, ρ-comonotonicity is also referred to as $|\rho|$-cohypomonotonicity. We say that the weak MVI is a *star-variant* of comonotonicity. This is analogue to the relationship between convexity and star-convexity.

For simplicity we state all results in terms of comonotonicity. However, note that *almost all results in this paper trivially extends to the more relaxed notion of weak MVI*. The only exception is the last iterate rates in Theorem 6.2 which relies on cohypomonotonicity to prove monotonic decrease through Lemma 6.1.

C Proofs for Section 4 (Inexact Krasnosel'skiĭ-Mann iterations)

The IKM iteration is well studied (see Combettes (2001)). The following result refurbishes sub-results of Combettes (2001, Prop. 4.2) to establish a rate of convergence under potentially stochastic feedback.

Theorem C.1 (Convergence of IKM)**.** *Suppose $T : \mathbb{R}^d \to \mathbb{R}^d$ is quasi-nonexpansive. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by IKM with $\lambda \in (0, 1)$. Then, for all $z^\star \in \operatorname{fix} T$*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|Tz^k - z^k\|^2 \leq \frac{\|z^0 - z^\star\|^2 + \sum_{k=0}^{K-1} \varepsilon_k(z^\star)}{\lambda(1-\lambda)K}. \tag{18}$$

where $\varepsilon_k(z) = 2\lambda\mathbb{E}[\|e^k\|\|z^k - z\|] + \lambda^2\mathbb{E}[\|e^k\|^2]$. Furthermore, $z^k \to z^\star$ a.s. as long as $\sum_{k=0}^{\infty} \varepsilon_k(z^\star) < \infty$.

Remark C.2. Notice that the optimal choice of λ in the upper bound is $\lambda = 0.5$, which is the default used for the Lookahead algorithm in both for minimax problems (Chavdarova et al., 2020) and minimization (Zhang et al., 2019) (see Section 7 for a treatment of Lookahead).

Proof. We will denote the natural filtration up to iteration k as $\mathcal{F}_k$ and use $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_k]$. Consider one exact step

$$s^k = (1-\lambda)z^k + \lambda T z^k \tag{19}$$

Then

$$\begin{aligned}\|s^k - z^\star\|^2 &= (1-\lambda)\|z^k - z^\star\|^2 + \lambda\|Tz^k - z^\star\|^2 - \lambda(1-\lambda)\|Tz^k - z^k\|^2 \\ \text{(quasi-nonexpansive)} &\leq (1-\lambda)\|z^k - z^\star\|^2 + \lambda\|z^k - z^\star\|^2 - \lambda(1-\lambda)\|Tz^k - z^k\|^2 \\ &= \|z^k - z^\star\|^2 - \lambda(1-\lambda)\|Tz^k - z^k\|^2 \end{aligned} \tag{20}$$

So

$$\|s^k - z^\star\| \leq \|z^k - z^\star\| \tag{21}$$

By using triangle inequality and the update rule IKM we have,

$$\begin{aligned}\mathbb{E}_k[\|z^{k+1} - z^\star\|^2] &\leq \mathbb{E}_k[(\|s^k - z^\star\| + \lambda\|e^k\|)^2] \\ &= \|s^k - z^\star\|^2 + 2\lambda\mathbb{E}_k[\|e^k\|]\|s^k - z^\star\| + \lambda^2\mathbb{E}_k[\|e^k\|^2] \\ (21) &\leq \|s^k - z^\star\|^2 + 2\lambda\mathbb{E}_k[\|e^k\|]\|z^k - z^\star\| + \lambda^2\mathbb{E}_k[\|e^k\|^2] \\ (20) &\leq \|z^k - z^\star\|^2 - \lambda(1-\lambda)\|Tz^k - z^k\|^2 + 2\lambda\mathbb{E}_k[\|e^k\|]\|z^k - z^\star\| + \lambda^2\mathbb{E}_k[\|e^k\|^2]. \end{aligned} \tag{22}$$

Using law of total expectation and telescoping obtains the claimed rate. The claimed asymptotic result follows from the Robbins-Siegmund supermartingale theorem (Bertsekas, 2011, Prop. 2). This completes the proof. ☐

D Proofs for Section 5 (Approximating the resolvent)

Lemma 5.1. *Suppose Assumptions 3.2(i), 3.2(ii) and 3.4. Given $z \in \mathbb{R}^d$, the iterates generated by (5) with $\gamma \in (0, 1/L)$ converges to a neighborhood linearly, i.e.,*

$$\mathbb{E}\big[\|w^\tau - J_{\gamma S}(z)\|^2\big] \leq (\gamma L)^{2\tau}\|w^0 - w^\star\|^2 + \tfrac{\gamma^2}{(1-\gamma L)^2}\sigma^2. \tag{6}$$

Proof. Let $\zeta^t = Fw^t - \hat{F}_\sigma(w^t, \xi_t)$. Then the stochastic update in (5) can be written as

$$w^{t+1} = (\mathrm{id} + \gamma A)^{-1}(z - \gamma F w^t + \gamma \zeta^t) \tag{23}$$

Let $w^\star \in \operatorname{fix} Q_z$ such that

$$\|w^{t+1} - w^\star\|^2 = \|w^{t+1} - Q_z(w^\star)\|^2. \tag{24}$$

Due to (firmly) nonexpansiveness of $(\mathrm{id} + \gamma A)^{-1}$ when A is maximally monotone we can go on as

$$\begin{aligned}
\|w^{t+1} - Q_z(w^\star)\|^2 &= \|(\mathrm{id} + \gamma A)^{-1}(z - \gamma F w^t + \gamma \zeta^t) - (\mathrm{id} + \gamma A)^{-1}(z - \gamma F w^\star)\|^2 \\
&\leq \|(z - \gamma F w^t + \gamma \zeta^t) - (z - \gamma F w^\star)\|^2 \\
&= \gamma^2 \|F w^t - F w^\star\|^2 + \gamma^2 \|\zeta^t\|^2 + 2\gamma^2 \langle \zeta^t, F w^\star - F w^t \rangle \\
&\leq \gamma^2 L^2 \|w^t - w^\star\|^2 + \gamma^2 \|\zeta^t\|^2 + 2\gamma^2 \langle \zeta^t, F w^\star - F w^t \rangle
\end{aligned} \tag{25}$$

where the last inequality follows from Lipschitz continuity of F.

Taking expectation and using unbiasedness and bounded variance from Assumption 3.4 we get

$$\mathbb{E}\big[\|w^{t+1} - w^\star\|^2 \mid \mathcal{F}_t\big] \leq \gamma^2 L^2 \|w^t - w^\star\|^2 + \gamma^2 \sigma^2 \tag{26}$$

By law of total expectation

$$\begin{aligned}
\mathbb{E}\big[\|w^\tau - w^\star\|\big] &\leq \gamma^2 L^2 \mathbb{E}\big[\|w^{\tau-1} - w^\star\|^2\big] + \gamma^2 \sigma^2 \\
&\leq \gamma^4 L^4 \mathbb{E}\big[\|w^{\tau-2} - w^\star\|^2\big] + \gamma^2 (1 + \gamma^2 L^2) \sigma^2 \\
&\leq \cdots \leq (\gamma L)^{2\tau} \mathbb{E}\big[\|w^0 - w^\star\|^2\big] + \gamma^2 \sigma^2 \sum_{t=0}^{\tau-1} (\gamma L)^{2t} \\
&\leq (\gamma L)^{2\tau} \|w^0 - w^\star\|^2 + \tfrac{\gamma^2}{(1-\gamma L)^2} \sigma^2
\end{aligned}$$

where the last inequality follows from $\sum_{t=0}^\infty a^t = \frac{1}{1-a}$ when $a < 1$.

By construction $\operatorname{fix} Q_z = \{J_{\gamma S}(z)\}$ which completes the proof. ∎

E Proofs for Section 6 (Last iterate under cohypomonotonicity)

Lemma 6.1. *If S is ρ-comonotone with $\rho > -\frac{\gamma}{2}$ then (8) satisfies for all $z^\star \in \operatorname{zer} S$,*

$$\|J_{\gamma S}(z^k) - z^k\|^2 \leq \|J_{\gamma S}(z^{k-1}) - z^{k-1}\|^2 + \delta_k(z^\star)$$

where $\delta_k(z) := 4\|e^k\|(\|z^{k+1} - z\| + \|z^k - z\|)$.

Proof. Rearranging the update (8b) and subsequently using (8a),

$$z^k - z^{k+1} = \lambda(z^k - \bar{z}^k - e^k) = \lambda(v^k - e^k). \tag{27}$$

Since γS is $\frac{1}{2}$-cohypomonotone

$$\begin{aligned}
-\tfrac{1}{2}\|v^k - v^{k+1}\|^2 &\leq \langle v^k - v^{k+1}, \bar{z}^k - \bar{z}^{k+1} \rangle \\
\text{(8a)} &= \langle v^k - v^{k+1}, z^k - v^k - (z^{k+1} - v^{k+1}) \rangle \\
&= \langle v^k - v^{k+1}, z^k - z^{k+1} \rangle - \|v^k - v^{k+1}\|^2 \\
\text{(27)} &= \lambda \langle v^k - v^{k+1}, v^k - e^k \rangle - \|v^k - v^{k+1}\|^2 \\
&= \lambda \|v^k\|^2 - \lambda \langle v^{k+1}, v^k \rangle - \|v^k - v^{k+1}\|^2 + \lambda \langle v^{k+1} - v^k, e^k \rangle
\end{aligned} \tag{28}$$

Rearranging

$$\begin{aligned}
0 &\leq \lambda \|v^k\|^2 - \lambda \langle v^{k+1}, v^k \rangle - \tfrac{1}{2}\|v^k - v^{k+1}\|^2 + \lambda \langle v^{k+1} - v^k, e^k \rangle \\
&\leq \lambda \|v^k\|^2 - \lambda \langle v^{k+1}, v^k \rangle - \tfrac{\lambda}{2}\|v^k - v^{k+1}\|^2 + \lambda \langle v^{k+1} - v^k, e^k \rangle \\
&= \lambda \|v^k\|^2 - \tfrac{\lambda}{2}\|v^k\|^2 - \tfrac{\lambda}{2}\|v^{k+1}\|^2 + \lambda \langle v^{k+1} - v^k, e^k \rangle
\end{aligned} \tag{29}$$

where the second inequality follows from observing that $\frac{1}{2} > \frac{\lambda}{2}$ since $\lambda \in (0, 1)$. It remain to bound the error term. Since γS is $1/2$-cohypomonotone the resolvent $J_{\gamma S}$ is nonexpansive. Thus,

$$\|\bar{z}^k - z^\star\| \leq \|z^k - z^\star\|. \tag{30}$$

Using Cauchy-Schwarz and the triangle inequality,

$$\begin{aligned}
\langle v^{k+1} - v^k, e^k \rangle &\leq \|e^k\| \|v^{k+1} - v^k\| \\
&\leq \|e^k\| (\|v^{k+1}\| + \|v^k\|) \\
&= \|e^k\| (\|\bar{z}^{k+1} - z^{k+1}\| + \|\bar{z}^k - z^k\|) \\
&\leq \|e^k\| (\|z^{k+1} - z^\star\| + \|z^k - z^\star\| + \|\bar{z}^{k+1} - z^\star\| + \|\bar{z}^k - z^\star\||) \\
\text{(30)} &\leq 2\|e^k\| (\|z^{k+1} - z^\star\| + \|z^k - z^\star\|)
\end{aligned} \tag{31}$$

Combining (29) and (31),

$$\tfrac{1}{2}\|v^{k+1}\|^2 \leq \tfrac{1}{2}\|v^k\|^2 + 2\|e^k\| (\|z^{k+1} - z^\star\| + \|z^k - z^\star\|). \tag{32}$$

Substituting in the resolvent using (8a) completes the proof. □

The proof of Lemma 6.1 simplifies for $\lambda = 1$. Consider one application of the inexact resolvent with error $e \in \mathbb{R}^d$,

$$z' = J_{\gamma S}(z) + e, \tag{33}$$

where $\lambda \in (0, 1)$ and $\gamma > 0$.

Lemma E.1. *If S is ρ-comonotone with $\rho > -\frac{\gamma}{2}$ then (33) satisfies $\|J_{\gamma S}(z') - z'\| \leq \|J_{\gamma S}(z) - z\| + 2\|e\|$.*

Proof. Since γS is $1/2$-cohypomonotone the resolvent $J_{\gamma S}$ is nonexpansive. Thus,

$$\begin{aligned}
\|J_{\gamma S}(z') - z'\| &= \|J_{\gamma S}(z') - J_{\gamma S}(z) - e\| \\
\text{(triangle ineq.)} &\leq \|J_{\gamma S}(z') - J_{\gamma S}(z)\| + \|e\| \\
\text{(nonexpansive)} &\leq \|z' - z\| + \|e\| \\
\text{(triangle ineq.)} &\leq \|J_{\gamma S}(z) - z\| + 2\|e\|
\end{aligned}$$

This completes the proof. □

Furthermore, the iterates of (8) are bounded in the following sense.

Lemma E.2. *Consider the sequence $(z^k)_{k \in \mathbb{N}}$ generated by (8) with $\lambda \in (0, 1)$ and $\rho > -\frac{\gamma}{2}$. Then for any $z^\star \in \operatorname{zer} S$,*

$$\|z^{k+1} - z^\star\| \leq \|z^0 - z^\star\| + \lambda \sum_{j=0}^{k} \|e^j\|. \tag{34}$$

Proof. Since γS is $1/2$-cohypomonotone the resolvent $J_{\gamma S}$ is nonexpansive. Thus,

$$\|\bar{z}^k - z^\star\| \leq \|z^k - z^\star\|. \tag{35}$$

We use the update rule

$$\begin{aligned}
\|z^{k+1} - z^\star\| &= \|(1 - \lambda) z^k + \lambda(\bar{z}^k + e^k) - z^\star\| \\
&\leq \|(1 - \lambda)(z^k - z^\star) + \lambda(\bar{z}^k - z^\star)\| + \lambda\|e^k\| \\
&\leq (1 - \lambda)\|z^k - z^\star\| + \lambda\|\bar{z}^k - z^\star\| + \lambda\|e^k\| \\
\text{(35)} &\leq \|z^k - z^\star\| + \lambda\|e^k\|
\end{aligned} \tag{36}$$

By recursively applying (36) we obtain the claim. □

Theorem 6.2 (Last iterate of inexact resolvent)**.** *Suppose Assumptions 3.2 and 3.4 with σ_k. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by (8) with $\lambda \in (0,1)$ and $\rho > -\frac{\gamma}{2}$. Then, for all $z^\star \in \operatorname{zer} S$,*

$$\mathbb{E}[\|J_{\gamma S}(z^K) - z^K\|^2] \leq \frac{\|z^0 - z^\star\|^2 + \sum_{k=0}^{K-1} \varepsilon_k(z^\star)}{\lambda(1-\lambda)K} \quad + \frac{1}{K}\sum_{k=0}^{K-1}\sum_{j=k}^{K-1} \delta_j(z^\star),$$

where $\varepsilon_k(z) := 2\lambda\mathbb{E}[\|e^k\|\|z^k - z\|] + \lambda^2\mathbb{E}[\|e^k\|^2]$ and $\delta_k(z) := 4\mathbb{E}[\|e^k\|(\|z^{k+1} - z\| + \|z^k - z\|)]$.

Proof. By taking $T = J_{\gamma S}$ in Theorem C.1 we have

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\|J_{\gamma S}(z^k) - z^k\|^2] \leq \frac{\|z^0 - z^\star\|^2 + \sum_{k=0}^{K-1} \varepsilon_k(z^\star)}{\lambda(1-\lambda)K}. \tag{37}$$

From Lemma 6.1 (and law of total expectation) we obtain,

$$K\mathbb{E}[\|J_{\gamma S}(z^K) - z^K\|^2] \leq \sum_{k=0}^{K-1} \mathbb{E}[\|J_{\gamma S}(z^k) - z^k\|^2] + \sum_{k=0}^{K-1}\sum_{j=k}^{K-1} \delta_j(z^\star). \tag{38}$$

Dividing by K and combining with (37) yields the rate. Noticing that $\operatorname{fix} J_{\gamma S} = \operatorname{zer} S$ completes the proof. □

Corollary E.3 (Explicit inexact stochastic resolvent)**.** *Suppose Assumptions 3.2 and 3.4 with σ_k for all $k \in \mathbb{N}$. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by RAPP with $\rho > -\frac{\gamma}{2}$. Then, for all $z^\star \in \operatorname{zer} S$ with $D := \sup_{j\in\mathbb{N}} \|z^j - z^\star\| < \infty$,*

(i) with $\sigma_k^2 = \sigma_0^2/k^2$ and $\tau = \frac{\log K}{\log(1/\gamma L)}$,

$$\frac{1}{K}\sum_{i=0}^{K-1} \mathbb{E}[\|J_{\gamma S}(z^k) - z^k\|^2] \leq \frac{\|z^0 - z^\star\|^2}{\lambda(1-\lambda)K} + \mathcal{O}\Big(\max\big\{\tfrac{D^2}{(1-\lambda)K}, \tfrac{\gamma D\sigma_0}{(1-\gamma L)(1-\lambda)K}\big\} + \tfrac{\lambda D^2}{(1-\lambda)K} + \tfrac{\lambda\gamma^2\sigma_0^2}{(1-\gamma L)^2(1-\lambda)K^2}\Big).$$

(ii) with $\sigma_k^2 = \sigma_0^2/k^3$ and $\tau = \frac{\log K^2}{\log(1/\gamma L)}$,

$$\begin{aligned}\mathbb{E}[\|J_{\gamma S}(z^K) - z^K\|^2] \leq \frac{\|z^0 - z^\star\|^2}{\lambda(1-\lambda)K} &+ \mathcal{O}\left(\max\{\tfrac{D^2}{K}, \tfrac{8\gamma D\sigma_0}{(1-\gamma L)\sqrt{K}}\}\right) \\ &+ \mathcal{O}\Big(\max\big\{\tfrac{D^2}{(1-\lambda)K^2}, \tfrac{2\gamma D\sigma_0}{(1-\gamma L)(1-\lambda)K^{3/2}}\big\} \\ &+ \tfrac{\lambda D^2}{(1-\lambda)K^2} + \tfrac{\lambda\gamma^2\sigma_0^2}{(1-\gamma L)^2(1-\lambda)K^3}\Big)\end{aligned} \tag{39}$$

Remark E.4. The assumption on the noise $\sigma_k^2 = \sigma_0^2/n_k$ can be achieved by taking the batch size as n_k, i.e.

$$\hat{F}_{\sigma_k}(z,\xi) = \frac{1}{n_k}\sum_{i=0}^{n_k} \hat{F}_{\sigma_0}(z,\xi_i). \tag{40}$$

This is clear by simple computation. Observe that the random variable $X_i := \hat{F}_\sigma(z,\xi_i) - Fz$ is i.i.d. with $\operatorname{Var}(X_i) = \sigma^2$. Then, the average, $\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$, has a variance as follows

$$\operatorname{Var}(\overline{X}_n) = \operatorname{Var}(\tfrac{1}{n}(X_1 + \cdots + X_n)) = \frac{1}{n^2}\operatorname{Var}(X_1 + \cdots + X_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

We note that increasing batch size might be unfavorable in some applications, but the alternative of diminishing stepsize only leads to only asymptotic convergence of the last iterate (as in e.g. Pethick et al. (2023)).

Proof. The theorem follows from combing Lemma 5.1 with Theorems 6.2 and C.1. Invoke Theorems 6.2 and C.1 with $e^k = w_k^\tau - J_{\gamma S}(z^k)$ and $\sigma = \sigma_k$ and note that the error e^k can be bounded through Lemma 5.1 as

$$\begin{aligned}\mathbb{E}_k[\|e^k\|^2] = \|w_k^\tau - J_{\gamma S}(z^k)\|^2 &\leq \gamma^{2\tau} L^{2\tau} \|w_k^0 - J_{\gamma S}(z^k)\|^2 + \tfrac{\gamma^2}{(1-\gamma L)^2}\sigma_k^2 \\ &= \gamma^{2\tau} L^{2\tau} \|z^k - J_{\gamma S}(z^k)\|^2 + \tfrac{\gamma^2}{(1-\gamma L)^2}\sigma_k^2. \end{aligned} \tag{41}$$

The former term can in turn be bounded through the triangle inequality

$$\|z^k - J_{\gamma S}(z^k)\| \leq \|z^k - z^\star\| + \|J_{\gamma S}(z^k) - z^\star\| \leq 2\|z^k - z^\star\| \leq 2D. \tag{42}$$

with $D := \sup_{j\in\mathbb{N}} \|z^j - z^\star\|$ and where the second last inequality follows from $z^\star \in \operatorname{fix} J_{\gamma S}$ and nonexpansiveness of $J_{\gamma S}$. Plugging into (41) we have,

$$\mathbb{E}_k[\|e^k\|^2] \leq 4\gamma^{2\tau} L^{2\tau} D^2 + \tfrac{\gamma^2}{(1-\gamma L)^2}\sigma_k^2, \tag{43}$$

and

$$\mathbb{E}_k[\|e^k\|] \leq \sqrt{4\gamma^{2\tau} L^{2\tau} D^2 + \tfrac{\gamma^2}{(1-\gamma L)^2}\sigma_k^2} \leq \max\{2\gamma^\tau L^\tau D, \tfrac{\gamma}{1-\gamma L}\sigma_k\}. \tag{44}$$

Substituting into the expression of $\delta_k(z^\star)$ and $\varepsilon_k(z^\star)$ yields,

$$\begin{aligned}\delta_k(z^\star) &\leq \max\{16\gamma^\tau L^\tau D^2, \tfrac{8\gamma}{1-\gamma L}\sigma_k D\} \\ \varepsilon_k(z^\star) &\leq \max\{4\lambda\gamma^\tau L^\tau D^2, \tfrac{2\lambda\gamma}{1-\gamma L}\sigma_k D\} + 4\lambda^2\gamma^{2\tau} L^{2\tau} D^2 + \tfrac{\lambda^2\gamma^2}{(1-\gamma L)^2}\sigma_k^2.\end{aligned}$$

Consequently, with the choice $\sigma_k^2 = \sigma_0^2/k^2$,

$$\frac{\sum_{k=0}^{K-1}\varepsilon_k(z^\star)}{\lambda(1-\lambda)K} \leq \max\Big\{\frac{4\gamma^\tau L^\tau D^2}{1-\lambda}, \tfrac{2\gamma D\sigma_0}{(1-\gamma L)(1-\lambda)K}\Big\} + \frac{4\lambda\gamma^{2\tau} L^{2\tau} D^2}{1-\lambda} + \tfrac{\lambda\gamma^2\sigma_0^2}{(1-\gamma L)^2(1-\lambda)K^2}. \tag{45}$$

We ideally want the terms involving τ to be of order $\mathcal{O}(1/K)$.

$$1/a^\tau = 1/K \iff a^\tau = K \iff \tau\log a = \log K \iff \tau = \frac{\log K}{\log a} \tag{46}$$

Choosing $a = 1/\gamma L$ it thus suffice to pick $\tau = \frac{\log K}{\log(1/\gamma L)}$ in order to have $\gamma^\tau L^\tau = 1/K$. Plugging into the average iterate result of Theorem C.1 yields the claim in Corollary E.3(i).

Additionally, with the choice $\sigma_k^2 = \sigma_0^2/k^3$,

$$\begin{aligned}\frac{\sum_{k=0}^{K-1}\varepsilon_k(z^\star)}{\lambda(1-\lambda)K} &\leq \max\Big\{\frac{4\gamma^\tau L^\tau D^2}{1-\lambda}, \tfrac{2\gamma D\sigma_0}{(1-\gamma L)(1-\lambda)K^{3/2}}\Big\} + \frac{4\lambda\gamma^{2\tau} L^{2\tau} D^2}{1-\lambda} + \tfrac{\lambda\gamma^2\sigma_0^2}{(1-\gamma L)^2(1-\lambda)K^3} \\ \frac{1}{K}\sum_{k=0}^{K-1}\sum_{j=k}^{K-1}\delta_j(z^\star) &\leq \max\{K16\gamma^\tau L^\tau D^2, \tfrac{8\gamma D\sigma_0}{(1-\gamma L)\sqrt{K}}\}.\end{aligned} \tag{47}$$

In order for the terms involving τ to be of order $\mathcal{O}(1/K)$ we need τ to be slightly larger.

$$K/a^\tau = 1/K \iff a^\tau = K^2 \iff \tau\log a = \log K^2 \iff \tau = \frac{\log K^2}{\log a} \tag{48}$$

Choosing $a = 1/\gamma L$ it thus suffice to pick $\tau = \frac{\log K^2}{\log(1/\gamma L)}$ in order to have $K\gamma^\tau L^\tau = 1/K$. Plugging (47) into the last iterate result of Theorem 6.2 completes the proof. □

Corollary 6.4 (Explicit inexact resolvent)**.** *Suppose Assumption 3.2 holds. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by RAPP with deterministic feedback and $\rho > -\frac{\gamma}{2}$. Then, for all $z^\star \in \operatorname{zer} S$ with $D := \sup_{j\in\mathbb{N}} \|z^j - z^\star\| < \infty$,*

(i) with $\tau = \frac{\log K}{\log(1/\gamma L)}$: $\frac{1}{K}\sum_{i=0}^{K-1}\|J_{\gamma S}(z^k) - z^k\|^2 = \mathcal{O}\Big(\frac{\|z^0 - z^\star\|^2}{\lambda(1-\lambda)K} + \frac{D^2}{(1-\lambda)K}\Big)$.

(ii) with $\tau = \frac{\log K^2}{\log(1/\gamma L)}$: $\|J_{\gamma S}(z^K) - z^K\|^2 = \mathcal{O}\Big(\frac{\|z^0 - z^\star\|^2}{\lambda(1-\lambda)K} + \frac{D^2}{K} + \frac{D^2}{(1-\lambda)K^2},\Big)$.

Proof. The claim follows directly from Corollary E.3 as a special case with $\sigma_0 = 0$. □

F Proofs for Section 7 (Analysis of Lookahead)

Theorem 7.1. *Suppose Assumption 3.2 holds. Consider the sequence* $(z^k)_{k\in\mathbb{N}}$ *generated by LA-GDA with* $\tau = 2$*,* $\gamma \leq 1/L$ *and* $\lambda \in (0, 1/2)$*. Furthermore, suppose that*

$$2\rho > -(1-2\lambda)\gamma \quad \text{and} \quad 2\rho \geq 2\lambda\gamma - (1-\gamma^2L^2)\gamma. \tag{10}$$

Then, for all $z^\star \in \operatorname{zer} F$*,*

$$\frac{1}{K}\sum_{k=0}^{K-1}\|F\bar{z}^k\|^2 \leq \frac{\|z^0 - z^\star\|^2}{\lambda\gamma\big((1-2\lambda)\gamma + 2\rho\big)K}. \tag{11}$$

Proof. For $\tau = 2$ we can write (LA-GDA) as

$$\begin{aligned} z^{k+1/3} &= z^k - \gamma F z^k \\ z^{k+2/3} &= z^{k+1/3} - \gamma F z^{k+1/3} \\ z^{k+1} &= (1-\lambda)z^k + \lambda z^{k+2/3} \end{aligned} \tag{49}$$

The proof relies on the simplified form of the update rule (9), which can be obtain as follows

$$\begin{aligned} z^{k+1} &= (1-\lambda)z^k + \lambda z^{k+2/3} \\ &= (1-\lambda)z^k + \lambda(z^{k+1/3} - \gamma F z^{k+1/3}) \\ &= (1-\lambda)z^k + \lambda(z^k - \gamma F z^k - \gamma F z^{k+1/3}) \\ &= z^k - \lambda\gamma F z^k - \lambda\gamma F(z^k - \gamma F z^k) \\ &= \tfrac{1}{2}(z^k - 2\lambda\gamma F z^k) + \tfrac{1}{2}(z^k - 2\lambda\gamma F(z^k - \gamma F z^k)). \end{aligned} \tag{50}$$

Define the following operators with $\beta = 2\lambda$

$$\mathrm{EG}^+(z) = z - \beta\gamma F(z - \gamma F z) \tag{51a}$$

$$\mathrm{GDA}(z) = z - \beta\gamma F z \tag{51b}$$

Then, using (50), LA-GDA with $\tau = 2$ can be written as

$$z^{k+1} = \tfrac{1}{2}\,\mathrm{GDA}(z^k) + \tfrac{1}{2}\,\mathrm{EG}^+(z^k) \tag{52}$$

One step of the update can be bounded as

$$\|z^{k+1} - z^\star\|^2 = \|\tfrac{1}{2}\,\mathrm{GDA}(z^k) + \tfrac{1}{2}\,\mathrm{EG}^+(z^k) - z^\star\|^2 \leq \tfrac{1}{2}\|\,\mathrm{GDA}(z^k) - z^\star\|^2 + \tfrac{1}{2}\|\,\mathrm{EG}^+(z^k) - z^\star\|^2, \tag{53}$$

where we have used Young's inequality. The first term can be expanded

$$\|\,\mathrm{GDA}(z^k) - z^\star\|^2 = \|z^k - z^\star\|^2 + \beta^2\gamma^2\|Fz^k\|^2 - 2\beta\gamma\langle Fz^k, z^k - z^\star\rangle \tag{54}$$

For the second term of (53) we will need to bound the following inner product

$$\begin{aligned} \langle \gamma F\bar{z}^k, z^k - \bar{z}^k\rangle &= \tfrac{\gamma^2}{2}\|F\bar{z}^k\|^2 - \tfrac{1}{2}\|\gamma F\bar{z}^k - (z^k - \bar{z}^k)\|^2 + \tfrac{1}{2}\|\bar{z}^k - z^k\|^2 \\ \text{(51a)} &= \tfrac{\gamma^2}{2}\|F\bar{z}^k\|^2 - \tfrac{\gamma^2}{2}\|F\bar{z}^k - Fz^k\|^2 + \tfrac{1}{2}\|\bar{z}^k - z^k\|^2 \\ \text{(Assumption 3.2(ii))} &\geq \tfrac{\gamma^2}{2}\|F\bar{z}^k\|^2 + \tfrac{1}{2}(1-\gamma^2L^2)\|\bar{z}^k - z^k\|^2. \end{aligned} \tag{55}$$

Consequently,

$$\begin{aligned} \gamma\langle F\bar{z}^k, z^k - z^\star\rangle &= \gamma\langle F\bar{z}^k, \bar{z}^k - z^\star\rangle + \gamma\langle F\bar{z}^k, z^k - \bar{z}^k\rangle \\ \text{(55)} &\leq \gamma\langle F\bar{z}^k, \bar{z}^k - z^\star\rangle - \tfrac{\gamma^2}{2}\|F\bar{z}^k\|^2 - \tfrac{1}{2}(1-\gamma^2L^2)\|\bar{z}^k - z^k\|^2. \end{aligned} \tag{56}$$

Finally,

$$\begin{aligned} \|\,\mathrm{EG}^+(z^k) - z^\star\|^2 &= \|z^k - z^\star\|^2 + \beta^2\gamma^2\|F\bar{z}^k\|^2 - 2\beta\gamma\langle F\bar{z}^k, z^k - z^\star\rangle \\ \text{(56)} &\leq \|z^k - z^\star\|^2 - \beta(1-\beta)\gamma^2\|F\bar{z}^k\|^2 - \beta(1-\gamma^2L^2)\|\bar{z}^k - z^k\|^2 - 2\beta\gamma\langle F\bar{z}^k, \bar{z}^k - z^\star\rangle \\ \text{(51a)} &= \|z^k - z^\star\|^2 - \beta(1-\beta)\gamma^2\|F\bar{z}^k\|^2 - \beta(1-\gamma^2L^2)\gamma^2\|Fz^k\|^2 - 2\beta\gamma\langle F\bar{z}^k, \bar{z}^k - z^\star\rangle \end{aligned} \tag{57}$$

Using (54) and (57) in (53), we have

$$
\begin{aligned}
2\|z^{k+1}-z^\star\|^2 &\leq 2\|z^k-z^\star\|^2+\beta^2\gamma^2\|Fz^k\|^2-2\beta\gamma\langle Fz^k, z^k-z^\star\rangle \\
&\quad -\beta(1-\beta)\gamma^2\|F\bar{z}^k\|^2-\beta(1-\gamma^2L^2)\gamma^2\|Fz^k\|^2-2\beta\gamma\langle F\bar{z}^k,\bar{z}^k-z^\star\rangle \\
\text{(Assumption 3.2(iii))} &\leq 2\|z^k-z^\star\|^2-\beta\gamma\big((1-\gamma^2L^2)\gamma+2\rho-\beta\gamma\big)\|Fz^k\|^2 \\
&\quad -\beta\gamma\big((1-\beta)\gamma+2\rho\big)\|F\bar{z}^k\|^2 \qquad (58)
\end{aligned}
$$

To get a recursion it thus suffice to require

$$
(1-\beta)\gamma+2\rho>0 \quad \text{and} \quad (1-\gamma^2L^2)\gamma+2\rho-\beta\gamma\geq 0. \qquad (59)
$$

Rearranging and telescoping (58) achieves the claimed rate. Rearranging (59) completes the proof. □

Theorem 7.3. *Suppose* $\operatorname{Alg}:\mathbb{R}^d\to\mathbb{R}^d$ *is quasi-nonexpansive. Then* $(z^k)_{k\in\mathbb{N}}$ *generated by* (LA) *converges to some* $z^\star\in\operatorname{fix}\operatorname{Alg}$.

Proof. By the composition rule (Bauschke & Combettes, 2017, Prop. 4.49(ii)) Alg^t is also nonexpansive. Since $(z^k)_{k\in\mathbb{N}}$ can be seen as a Krasnosel'skiĭ-Mann iteration of a quasi-nonexpansive operator the iterates converges to $z^\star\in\operatorname{fix}\operatorname{Alg}^t$ by Theorem C.1 with $\varepsilon_k=0$, i.e. $\|z^k-z^\star\|\overset{k\to\infty}{\longrightarrow}0$. By Bauschke & Combettes (2017, Prop. 4.49(i)) it also follows that $\operatorname{fix}\operatorname{Alg}^t=\operatorname{fix}\operatorname{Alg}$, which completes the proof. □

Corollary 7.5. *Suppose* F *is* $1/L$*-cocoercive. Then* $(z^k)_{k\in\mathbb{N}}$ *generated by* LA-GDA *with* $\gamma\leq 2/L$ *converges to some* $z^\star\in\operatorname{zer}F$.

Proof. If F is $1/L$-cocoercive then γF is $1/2$-cocoercive given $\gamma\leq 2/L$, which in turn implies that $V=\operatorname{id}-\gamma F$ is nonexpansive. The claim follows from Theorem 7.3 and by observing that $\operatorname{fix}V=\operatorname{zer}F$. □

Consider the forward-backward-forward (FBF) method of Tseng (1991). We can write one step as follows

$$
\bar{z}=(\operatorname{id}+\gamma A)^{-1}Hz \qquad (60a)
$$
$$
\operatorname{FBF}(z)=z-(Hz-H\bar{z}) \qquad (60b)
$$

where $H=\operatorname{id}-\gamma F$. The extragradient method is obtained as a special case when $A\equiv 0$.

Theorem F.1. *If* $A\colon\mathbb{R}^d\rightrightarrows\mathbb{R}^d$ *is maximally monotone and* $F\colon\mathbb{R}^d\to\mathbb{R}^d$ *is monotone and* L*-Lipschitz continuous then the operator* (60) *with* $\gamma\leq 1/L$ *is quasi-nonexpansive. Furthermore,* $\operatorname{fix}\operatorname{FBF}=\operatorname{zer}S$ *with* $S:=A+F$.

Proof. By $1/2$-cocoercivity from Lemma B.4 we obtain

$$
\begin{aligned}
\langle H\bar{z}-Hz, z-z^\star\rangle &= \langle H\bar{z}-Hz,\bar{z}-z^\star\rangle+\langle H\bar{z}-Hz, z-\bar{z}\rangle \\
\text{(Lemma B.4)} &\leq \langle H\bar{z}-Hz,\bar{z}-z^\star\rangle-\tfrac{1}{2}\|H\bar{z}-Hz\|^2-\tfrac{1}{2}(1-\gamma^2L^2)\|\bar{z}-z\|^2 \\
\text{(monotone)} &\leq -\tfrac{1}{2}\|H\bar{z}-Hz\|^2-\tfrac{1}{2}(1-\gamma^2L^2)\|\bar{z}-z\|^2 \qquad (61)
\end{aligned}
$$

The operator in (60b) satisfies

$$
\begin{aligned}
\|\operatorname{FBF}(z)-z^\star\|^2 &= \|z-z^\star\|^2+\|H\bar{z}-Hz\|^2+2\langle H\bar{z}-Hz, z-z^\star\rangle \\
\text{(61)} &\leq \|z-z^\star\|^2-(1-\gamma^2L^2)\|\bar{z}-z\|^2
\end{aligned}
$$

where the last term is negative due to $\gamma\leq 1/L$. Recognizing the definition of quasi-nonexpansive completes the proof. □

Corollary 7.7. *Suppose Assumption 3.2 holds. Then* $(z^k)_{k\in\mathbb{N}}$ *generated by* LA-CEG+ *with* $\lambda\in(0,1)$, $\gamma\in(\lfloor -2\rho\rfloor_+, 1/L)$ *and* $\alpha\in(0,1+\frac{2\rho}{\gamma})$ *converges to some* $z^\star\in\operatorname{zer}S$.

Proof. Quasi-nonexpansiveness of the operator $\mathrm{CEG}^+ : \mathbb{R}^d \to \mathbb{R}^d$ follows from Theorem G.1(i) provided $\alpha \in (0, 1 + \frac{2\rho}{\gamma})$ so Theorem 7.3 applies.

It remains to verify that $\operatorname{fix} \mathrm{CEG}^+ = \operatorname{zer} S$. This follows from

$$\tfrac{1}{\gamma}(Hz - H\bar{z}) \in A(\bar{z}) + F(\bar{z}) = S(\bar{z}), \tag{62}$$

and noticing that the stepsizes are positive, i.e. $\alpha > 0$ and $\gamma > 0$, which completes the proof. □

G Analysis of CEG+

This section provides a simplified convergence proof of the CEG+ scheme proposed in Pethick et al. (2022, Cor. 3.2) without going through adaptivity and a projected interpretation. We additionally provide convergence in terms of the residual $\|z^k - \bar{z}^k\|$. The algorithm can be described with the following recursion

$$\bar{z}^k = (\mathrm{id} + \gamma A)^{-1}(Hz^k) \tag{63a}$$

$$z^{k+1} = z^k - \alpha\left(Hz^k - H\bar{z}^k\right) \tag{63b}$$

where $H = \mathrm{id} - \gamma F$. The EG+ algorithm is obtained as a special case when $A \equiv 0$.

Theorem G.1. *Suppose Assumption 3.2 and $\gamma \in (\lfloor -2\rho \rfloor_+, 1/L]$. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by (63). Then, for all $z^\star \in \operatorname{zer} S$, it follows that*

(i) the iterates $(z^k)_{k\in\mathbb{N}}$ satisfies

$$\|z^{k+1} - z^\star\|^2 \leq \|z^k - z^\star\|^2 - \alpha(1 + \tfrac{2\rho}{\gamma} - \alpha)\|H\bar{z}^k - Hz^k\|^2 - \alpha(1 - \gamma^2 L^2)\|\bar{z}^k - z^k\|^2,$$

and in particular, $\mathrm{CEG}^+ : \mathbb{R}^d \to \mathbb{R}^d$ in (12) is quasi-nonexpansive if $\alpha \in (0, 1 + \frac{2\rho}{\gamma})$.

(ii) for $\alpha \in (0, 1]$ and $\alpha < 1 + \frac{2\rho}{\gamma}$

$$\frac{1}{K}\sum_{k=0}^{K-1} \|z^k - \bar{z}^k\|^2 \leq \frac{\|z^0 - z^\star\|^2}{\alpha(1 - \gamma^2 L^2)K}. \tag{64}$$

(iii) for $\alpha \in (0, 1)$ and $\alpha < 1 + \frac{2\rho}{\gamma}$

$$\frac{1}{K}\sum_{k=0}^{K-1} \operatorname{dist}(0, S\bar{z}^k)^2 \leq \frac{\|z^0 - z^\star\|^2}{\alpha\gamma^2(1 + \frac{2\rho}{\gamma} - \alpha)K}. \tag{65}$$

Proof. By $1/2$-cocoercivity of $H = \mathrm{id} - \gamma F$ from Lemma B.4 we obtain

$$\begin{aligned}\langle H\bar{z}^k - Hz^k, z^k - z^\star\rangle &= \langle H\bar{z}^k - Hz^k, \bar{z}^k - z^\star\rangle + \langle H\bar{z}^k - Hz^k, z^k - \bar{z}^k\rangle \\ &\leq \langle H\bar{z}^k - Hz^k, \bar{z}^k - z^\star\rangle - \tfrac{1}{2}\|H\bar{z}^k - Hz^k\|^2 - \tfrac{1}{2}(1 - \gamma^2 L^2)\|\bar{z}^k - z^k\|^2\end{aligned} \tag{66}$$

The update in (63b) yields

$$\begin{aligned}\|z^{k+1} - z^\star\|^2 &= \|z^k - z^\star\|^2 + \alpha^2\|H\bar{z}^k - Hz^k\|^2 + 2\alpha\langle H\bar{z}^k - Hz^k, z^k - z^\star\rangle \\ \text{(66)} &\leq \|z^k - z^\star\|^2 - 2\alpha\langle Hz^k - H\bar{z}^k, \bar{z}^k - z^\star\rangle \\ &\quad - \alpha(1 - \alpha)\|H\bar{z}^k - Hz^k\|^2 - \alpha(1 - \gamma^2 L^2)\|\bar{z}^k - z^k\|^2.\end{aligned} \tag{67}$$

Noticing that both latter terms are negative. Observe that by (63a) we have

$$\tfrac{1}{\gamma}(Hz^k - H\bar{z}^k) \in A(\bar{z}^k) + F(\bar{z}^k) = S(\bar{z}^k).$$

Therefore, by cohypomonotonicity of $S = A + F$,

$$\tfrac{1}{\gamma}\langle Hz^k - H\bar{z}^k, \bar{z}^k - z^\star\rangle \geq \rho\|Hz^k - H\bar{z}^k\|^2. \tag{68}$$

and consequently (67) leads to Fejér monotonicity,

$$\|z^{k+1} - z^\star\|^2 \leq \|z^k - z^\star\|^2 - \alpha(1 + \tfrac{2\rho}{\gamma} - \alpha)\|H\bar{z}^k - Hz^k\|^2 - \alpha(1 - \gamma^2 L^2)\|\bar{z}^k - z^k\|^2.$$

By telescoping we obtain the two claims. □

H Experiments

H.1 Simulations

We repeat the synthetic examples for convenience below.

Example H.1 (PolarGame (Pethick et al., 2022, Ex. 3(iii)))**.** *Consider*

$$Fz = (\psi(x,y) - y, \psi(y,x) + x)\,,$$

where $\|z\|_\infty \leq {}^{11}/_{10}$ *and* $\psi(x,y) = \frac{1}{16}ax(-1+x^2+y^2)(-9+16x^2+16y^2)$ *with* $a = \frac{1}{3}$.

Example H.2 (Quadratic (Pethick et al., 2022, Ex. 5))**.** *Consider,*

$$\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} \phi(x,y) := axy + \frac{b}{2}x^2 - \frac{b}{2}y^2, \tag{69}$$

where $a \in \mathbb{R}_+$ *and* $b \in \mathbb{R}$.

The problem constants in Example H.2 can easily be computed as $\rho = \frac{b}{a^2+b^2}$ and $L = \sqrt{a^2+b^2}$. We can rewrite Example H.2 in terms of L and ρ by choosing $a = \sqrt{L^2 - L^4\rho^2}$ and $b = L^2\rho$.

We provide below a slight generalization of the Forsaken example (Hsieh et al., 2021, Example 5.2), from which we derive another important case.

Example H.3. *Consider,*

$$\min_{|x|\leq 3/2} \max_{|y|\leq 3/2} \phi(x,y) := x(y-a) + \psi(x) - \psi(y), \tag{70}$$

where $\psi(z) = \frac{1}{4}z^2 - \frac{1}{2}z^4 + \frac{1}{6}z^6$ *and* $a \in \mathbb{R}$. *We have the following important cases:*

(i) for $a = 0.45$ *we recover Forsaken (Hsieh et al., 2021, Example 5.2).*

(ii) for $a = 0.34$ *we ensure that the first-order stationary point is a local Nash equilibrium (LNE), which is apparent from inspection of the Jacobian. We call this new example LNEForsaken.*

In both Example H.2 and Example H.3 the operator F is defined as $Fz = (\nabla_x\phi(x,y), -\nabla_y\phi(x,y))$. For Example H.3 the Lookahead methods use $\tau = 20$, $\lambda = 0.2$ and $\gamma = {}^{1}/_{L} \approx 0.08$ and (R)APP uses $\tau = 10$, $\lambda = 0.2$ and $\gamma = {}^{4}/_{L} \approx 0.32$. In Examples H.1 and H.2 we use $\gamma = {}^{1}/_{L}$, $\lambda = 0.1$ for LA-GDA and EG+ with $\alpha = 0.1$ for the latter. In the constrained examples L refers to the Lipschitz constant constrained to the constraint set.

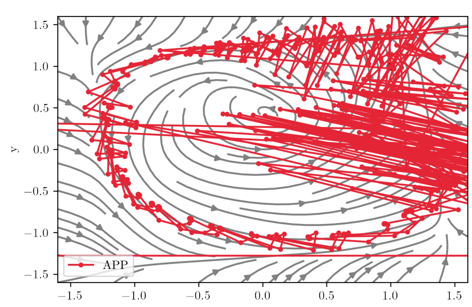

Figure 6: The iterates of APP associated with Figure 2.

H.2 Image generation

Architecture The ResNet uses a 128-dimensional input space for the generator and spectral normalization for the discriminator (see Chavdarova et al. (2020, Table 7)). The models' parameters are initialized using the Xavier initialization as suggested in Miyato et al. (2018).

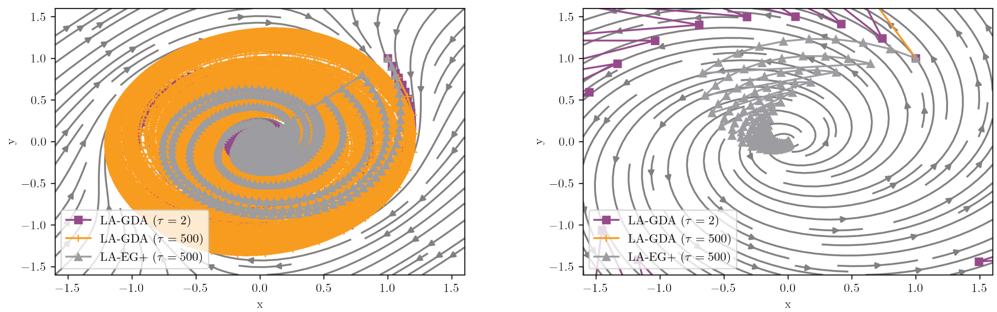

Figure 7: Iterates associated with Figure 3.

Optimizers All methods relies on stochastic gradients computed over a mini-batch. The discriminator and generator is updated in an alternating fashion. We use the same variant of extragradient as Chavdarova et al. (2020) uses in their implementation. The variant only uses the extrapolated point of the *opponent* in the update of the next iterate (x^{k+1}, y^{k+1}) as follows

$$\begin{aligned} \bar{x}^k &= x^k - \gamma_1 \nabla \phi(x^k, y^k) \\ \bar{y}^k &= y^k + \gamma_2 \nabla \phi(x^k, y^k) \\ x^{k+1} &= x^k - \gamma_1 \nabla \phi(x^k, \bar{y}^k) \\ y^{k+1} &= y^k + \gamma_2 \nabla \phi(\bar{x}^k, y^k) \end{aligned} \tag{71}$$

Interestingly, we observed that the classical extragradient method (both a simultaneous and alternating variant) did not perform well under the hinge loss as used in the experiments. We leave investigate of this for future work.

Evaluation We use the Fréchet inception distance (FID) (Heusel et al., 2017) evaluated on $50\,000$ examples and the Inception score (ISC) (Salimans et al., 2016). For consistent and reproducible evaluation we use the `torch-fidelity` Python library (Obukhov et al., 2020) to compute the scores. The mean and standard deviation is computed over 5 and 3 independent execution in Table 2 and Table 3, respectively.

Compute time Producing Table 2 alone takes roughly 6 methods $\times$ 5 runs $\times$ 30 hours $=$ 37.5 days on a NVIDIA A100 GPU.

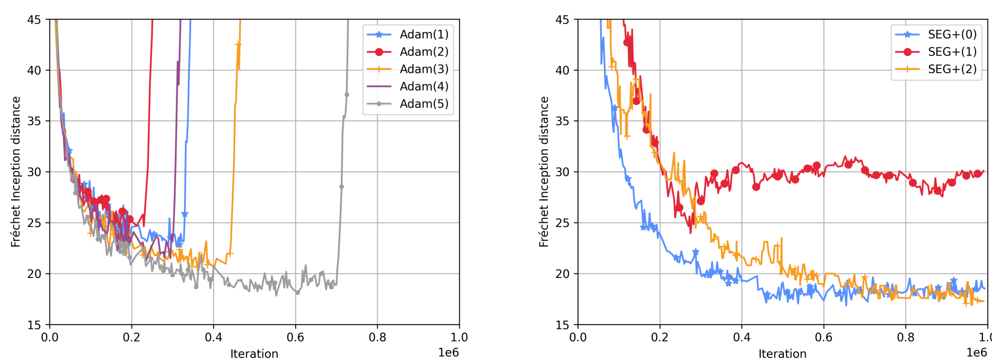

Figure 8: (left) Adam eventually diverges for all 5 runs. See Figure 4 for comparison with Lookahead. (right) In contrast, EG+ increases stability (and thus avoids divergence), but in effect might also be stuck in a local (suboptimal) solution. This explains the high variance and poor performance of EG+. By excluding the locally stuck run, EG+ achieves a FID of 16.88 ± 0.05 and a ISC of 8.0 ± 0.02, which is competitive even with the Lookahead-based methods.

H.2.1 Hyperparameters

Table 4: Training Hyperparameters for Adam-based experiments on CIFAR10

Hyperparameter	Adam	LA-Adam	ExtraAdam+	LA-ExtraAdam+	ExtraAdam	LA-ExtraAdam
lrD	2e-4	2e-4	2e-4	2e-4	2e-4	2e-4
lrG	2e-4	2e-4	2e-4	2e-4	2e-4	2e-4
Batch Size	128	128	128	128	128	128
β_1	0.0	0.0	0.0	0.0	0.0	0.0
D-steps	5	5	5	5	5	5
Lookahead τ		5		5000		5000
Lookahead λ		0.5		0.5		0.5
EG+ α			0.5	0.5		

Table 5: Training Hyperparameters for GDA-based experiments on CIFAR10

Hyperparameter	GDA	LA-GDA	EG+	LA-EG+	EG	LA-EG	RAPP
lrD	0.1	0.1	0.1	0.1	0.1	0.1	0.1
lrG	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Batch Size	128	128	128	128	128	128	128
D-steps	1	1	1	1	1	1	1
Lookahead τ		5000		5000		5000	
Lookahead λ		0.5		0.5		0.5	
EG+ α			0.5	0.5			
RAPP τ							3
RAPP λ							0.9