7-2023

Convergence of Proximal Point and Extragradient-Based Methods Beyond Monotonicity: the Case of Negative Comonotonicity

Eduard Gorbunov
Mohamed Bin Zayed University of Artificial Intelligence

Adrien Taylor
Université PSL

Samuel Horváth
Mohamed Bin Zayed University of Artificial Intelligence

Gauthier Gidel
Montreal Institute for Learning Algorithms

Recommended Citation

E. Gorbunov, A. Taylor, S. Horvath and G. Gidel, "Convergence of Proximal Point and Extragradient-Based Methods Beyond Monotonicity: the Case of Negative Comonotonicity," *Proceedings of Machine Learning Research* , vol. 202, pp. 11614 - 11641, Jul 2023.

Convergence of Proximal Point and Extragradient-Based Methods Beyond Monotonicity: the Case of Negative Comonotonicity

Eduard Gorbunov [1 2] Adrien Taylor [3] Samuel Horváth [1] Gauthier Gidel [4 5]

Abstract

Algorithms for min-max optimization and variational inequalities are often studied under monotonicity assumptions. Motivated by non-monotone machine learning applications, we follow the line of works (Diakonikolas et al., 2021; Lee & Kim, 2021; Pethick et al., 2022; Böhm, 2022) aiming at going beyond monotonicity by considering the weaker *negative comonotonicity* assumption. In this work, we provide tight complexity analyses for the Proximal Point (PP), Extragradient (EG), and Optimistic Gradient (OG) methods in this setup, closing several questions on their working guarantees beyond monotonicity. In particular, we derive the first non-asymptotic convergence rates for PP under negative comonotonicity and star-negative comonotonicity and show their tightness via constructing worst-case examples; we also relax the assumptions for the last-iterate convergence guarantees for EG and OG and prove the tightness of the existing best-iterate guarantees for EG and OG via constructing counter-examples.

1. Introduction

The study of efficient first-order methods for solving variational inequality problems (VIP) have known a surge of interest due to the development of recent machine learning (ML) formulations involving multiple objectives. VIP appears in various ML tasks such as robust learning (Ben-Tal et al., 2009), adversarial training (Madry et al., 2018), Generative Adversarial Networks (Goodfellow et al., 2014), or games with decision-dependent data (Narang et al., 2022). In this work, we focus on unconstrained VIPs[1], which we state formally in the slightly more general form of an *inclusion problem*:

$$\text{find } x^* \in \mathbb{R}^d \text{ such that } 0 \in F(x^*), \tag{IP}$$

where $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is some (possibly set-valued) mapping. In the sequel, we use the slightly abusive shorthand notation $F(x)$ to denote any particular image of x by the mapping F, independently of F being single-valued of not.

Among the main simple first-order methods under consideration for such problems, the extragradient method (EG) (Korpelevich, 1976) and the optimistic gradient method (OG) (Popov, 1980) occupy an important place. These two algorithms have been traditionally analyzed under the assumption that the considered operator is monotone and Lipschitz (Korpelevich, 1976; Popov, 1980) and are often interpreted as an approximation to the proximal point (PP) method (Nemirovski, 2004; Mokhtari et al., 2019). PP can be formally stated as an implicit iterative method generating a sequence $x^1, x^2, \ldots \in \mathbb{R}^d$ when initiated at some $x^0 \in \mathbb{R}^d$:

$$x^{k+1} = x^k - \gamma F(x^{k+1}), \tag{PP}$$

for some well-chosen stepsize $\gamma \in \mathbb{R}$. When F is single-valued, one can instead use explicit methods such as EG:

$$\begin{aligned} \widetilde{x}^k &= x^k - \gamma_1 F(x^k), \\ x^{k+1} &= x^k - \gamma_2 F(\widetilde{x}^k), \end{aligned} \quad \forall k \geq 0, \tag{EG}$$

or OG with the additional initialization $\widetilde{x}^0 = x^0$:

$$\begin{aligned} \widetilde{x}^k &= x^k - \gamma_1 F(\widetilde{x}^{k-1}), \quad \forall k > 0, \\ x^{k+1} &= x^k - \gamma_2 F(\widetilde{x}^k), \quad \forall k \geq 0, \end{aligned} \tag{OG}$$

where $\gamma_1, \gamma_2 \in \mathbb{R}$ are some well-chosen stepsizes. For examples of the usage of extragradient-based methods in practice, we refer to (Daskalakis et al., 2018) who use a variant of OG with Adam (Kingma & Ba, 2014) estimators to train WGAN (Gulrajani et al., 2017) on CIFAR10

[1] Mohamed bin Zayed University of Artificial Intelligence, UAE [2] Moscow Institute of Physics and Technology, Russia (part of this work was done while the author was a researcher at MIPT) [3] INRIA & D.I. École Normale Supérieure, CNRS & PSL Research University, France [4] Université de Montréal and Mila, Canada [5] Canada CIFAR AI Chair. Correspondence to: Eduard Gorbunov <eduard.gorbunov@mbzuai.ac.ae>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] We refer to (Gidel et al., 2019) for the details on how these formulations appear in the real-world problems.

(Krizhevsky et al., 2009), (Brown & Sandholm, 2019) where extragradient-based methods were applied in regret matching, (Farina et al., 2019) for the application to counterfactual regret minimization, and (Anagnostides et al., 2022) where these methods were used for training agents to play poker.

Interestingly, until recently, the convergence rate for the *last iterate* of neither EG nor OG were known even when F is (maximally) monotone and Lipschitz. First results in this direction were obtained by Golowich et al. (2020b;a) under some additional assumptions (namely the Jacobian of F being Lipschitz). Later, Gorbunov et al. (2022b;c); Cai et al. (2022b) closed this question by proving the tight worst-case last iterate convergence rate of these methods under monotonicity and Lipschitzness of F.

As some important motivating applications involve deep neural networks, the operator F under consideration is typically not monotone. However, for general non-monotone problems approximating first-order locally optimal solutions can be intractable (Daskalakis et al., 2021; Diakonikolas et al., 2021). Thus, it is natural to consider assumptions on structured non-monotonicity. Recently Diakonikolas et al. (2021) proposed to analyse EG using a weaker assumption than the traditional monotonicity. In the sequel, this assumption is referred to as *ρ-negative comonotonicity* (with $\rho \geq 0$). That is, for all $x, y \in \mathbb{R}^d$, the operator F satisfies:

$$\langle F(x) - F(y), x - y\rangle \geq -\rho\|F(x) - F(y)\|^2. \quad (1)$$

A number of works have followed the idea of Diakonikolas et al. (2021) and considered (1) as their working assumption, see, e.g., (Yoon & Ryu, 2021; Lee & Kim, 2021; Luo & Tran-Dinh, 2022; Cai et al., 2022a; Gorbunov et al., 2022a). Albeit being a reasonable first step toward the understanding of the behavior of algorithms for (IP) beyond F being monotone, it remains unclear by what means the ρ-negative comonotonicity assumption is general enough to capture complex non-monotone operators. This question is crucial for developing a clean optimization theory that can fully encompass ML applications involving neural networks.

To the best of our knowledge, *ρ(-star)-negative comonotonicity* is the weakest known assumption under which extragradient-type methods can be analyzed for solving (IP). The first part of this work is devoted to providing simple interpretations of this assumption. Then, we close the problem of studying the convergence rate of the PP method in this setting, the base ingredient underlying most algorithms for solving (IP) (which are traditionally interpreted as approximations to PP, see (Nemirovski, 2004)). That is, we provide upper and lower convergence bounds as well as a tight conditions on its stepsize for PP under negative comonotonicity. We eventually consider the last-iterate convergence of EG and OG and provide an almost complete picture in that case, listing the remaining open questions.

Before moving to the next sections, let us mention that many of our results were discovered using the performance estimation approach, first coined by (Drori & Teboulle, 2014) and formalized by (Taylor et al., 2017c;a). The operator version of the framework is due to (Ryu et al., 2020). We used the framework through the packages PESTO (Taylor et al., 2017b) and PEPit (Goujaud et al., 2022), thereby providing a simple way to validate our results numerically.

1.1. Preliminaries

In the context of (IP), we refer to F as being ρ-star-negative comonotone ($\rho \geq 0$) – a relaxation[2] of (1) – if for all $x \in \mathbb{R}^d$ and x^* being a solution to (IP), we have:

$$\langle F(x), x - x^*\rangle \geq -\rho\|F(x)\|^2. \quad (2)$$

Furthermore, similar to monotone operators (see (Bauschke et al., 2011) or (Ryu & Yin, 2020) for details), we assume that the mapping F is *maximal* in the sense that its graph is not strictly contained in the graph of any other ρ-negative comonotone operator (resp., ρ-star-negative comonotone), which ensures the corresponding proximal operator used in the sequel to be well-defined. Some examples of star-negative comonotone operators are given in (Pethick et al., 2022, Appendix C). Moreover, if F is star-monotone or quasi-strongly monotone (Loizou et al., 2021), then F is also star-negative comonotone. The examples of star-monotone/quasi-strongly monotone operators that are not monotone are given in (Loizou et al., 2021, Appendix A.6). Next, there are some studies of the eigenvalues of the Jacobian around the equilibrium of GAN games (Mescheder et al., 2018; Nagarajan & Kolter, 2017; Berard et al., 2019). These studies imply that the corresponding variational inequalities are locally quasi-strongly monotone. Finally, when F is L-Lipschitz it satisfies $\langle F(x), x - x^*\rangle \geq -L\|x - x^*\|^2$. If in addition $\|F(x)\| \geq \eta\|x - x^*\|$ for some $\eta > 0$ (meaning that F changes not "too slowly"), then $\langle F(x), x - x^*\rangle \geq -\frac{L}{\eta^2}\|F(x)\|^2$, i.e., condition (2) holds with $\rho = \frac{L}{\eta^2}$.

For the analysis of the EG and OG, we further assume F to be L-Lipschitz, meaning that for all $x, y \in \mathbb{R}^d$:

$$\|F(x) - F(y)\| \leq L\|x - y\|. \quad (3)$$

Note that in that case, F is a single-valued mapping. In this case, IP transforms into a variational inequality:

$$\text{find } x^* \in \mathbb{R}^d \text{ such that } F(x^*) = 0. \quad \text{(VIP)}$$

1.2. Related Work

Last-iterate convergence rates in the monotone case. Several recent theoretical advances focus on the last-iterate

[2] For the example of star-negative comonotone operator that is not negative comonotone we refer to (Daskalakis et al., 2020, Section 5.1) and (Diakonikolas et al., 2021, Section 2.2).

Table 1: Known and new $\mathcal{O}\left(1/N\right)$ convergence results for PP, EG and OG. Notation: NC = negative comonotonicity, SNC = star-negative comonotonicity, L-Lip. = L-Lipschitzness. Whenever the derived results are completely novel or extend the existing ones, we highlight them in green.

Method	Setup	$\rho \in$	Convergence	Reference	Counter-/Worst-case examples?
PP[(1)]	NC	$[0, +\infty)$	Last-iterate	Theorem 3.1	Theorem 3.2 (worst-case example) & 3.3 (diverge for $\gamma \leq 2\rho$)
	SNC	$[0, +\infty)$	Best-iterate	Theorem 3.1	Theorem 3.2 (worst-case example) & 3.3 (diverge for $\gamma \leq 2\rho$)
EG	NC + L-Lip.	$[0, 1/16L)$	Last-iterate	(Luo & Tran-Dinh, 2022)	✗
	NC + L-Lip.	$[0, 1/8L)$	Last-iterate	Theorem 4.2	Theorem 4.3 (diverge for $\rho \geq 1/2L$ and any $\gamma_1, \gamma_2 > 0$)
	SNC + L-Lip.	$[0, 1/8L)$	Best-iterate	(Diakonikolas et al., 2021)	✗
	SNC + L-Lip.	$[0, 1/2L)$	Best-iterate	(Pethick et al., 2022)	Theorem 3.4 (diverge for $\gamma_1 = 1/L$ and $\rho \geq (1-L\gamma_2)/2L$)
	SNC + L-Lip.	$[0, 1/2L)$	Best-iterate	Theorem 4.2 [(2)]	Theorem 4.3 (diverge for $\rho \geq 1/2L$ and any $\gamma_1, \gamma_2 > 0$)
OG	NC + L-Lip.	$[0, 8/(27\sqrt{6}L))$	Last-iterate	(Luo & Tran-Dinh, 2022)	✗
	NC + L-Lip.	$[0, 5/62L)$	Last-iterate	Theorem 4.4	Theorem 4.5 (diverge for $\rho \geq 1/2L$ and any $\gamma_1, \gamma_2 > 0$)
	SNC + L-Lip.	$[0, 1/2L)$	Best-iterate	(Böhm, 2022)	✗
	SNC + L-Lip.	$[0, 1/2L)$	Best-iterate	Theorem 4.4 [(2)]	Theorem 4.5 (diverge for $\rho \geq 1/2L$ and any $\gamma_1, \gamma_2 > 0$)

[(1)] The best-iterate convergence result can be obtained (Iusem et al., 2003, Lemma 2), and the last-iterate convergence result can also be derived from the non-expansiveness of PP update (Bauschke et al., 2021, Proposition 3.13 (iii)). At the moment of writing our paper, we were not aware of these results.

[(2)] Although these results are not new for the best-iterate convergence of EG and OG, the proof techniques differ from prior works.

convergence of the methods for solving IP/VIP with *monotone* operator F. In particular, He & Yuan (2015) derive the last-iterate $\mathcal{O}(1/N)$ rate[3] for PP and Gu & Yang (2020) show its tightness. Under the additional assumption of Lipschitzness of F and of its Jacobian, Golowich et al. (2020b;a) obtain last-iterate $\mathcal{O}(1/N)$ convergence for EG and OG and prove matching lower bounds for them. Next, Gorbunov et al. (2022b;c); Cai et al. (2022b) prove similar upper bounds for EG/OG without relying on the Lipschitzness (and even existence) of the Jacobian of F. Finally, for this class of problems one can design (accelerated) methods with provable $\mathcal{O}(1/N^2)$ last-iterate convergence rate (Yoon & Ryu, 2021; Bot et al., 2022; Tran-Dinh & Luo, 2021; Tran-Dinh, 2022). Although $\mathcal{O}(1/N^2)$ is much better than $\mathcal{O}(1/N)$, EG/OG are still more popular due to their higher flexibility. Moreover, when applied to non-monotone problems the mentioned accelerated methods may be attracted to “bad” stationary points, see, e.g., (Gorbunov et al., 2022c, Example 1.1).

Best-iterate convergence under ρ-star-negative comonotonicity. The convergence of EG is also studied under ρ-star-negative comonotonicity (and L-Lipschitzness): Diakonikolas et al. (2021) prove best-iterate $\mathcal{O}(1/N)$ convergence of EG with $\gamma_2 < \gamma_1$ for any $\rho < 1/8L$ and Pethick et al. (2022) derive a similar result for any $\rho < 1/2L$. Moreover, Pethick et al. (2022) show that EG is not necessary convergent when $\gamma_1 = 1/L$ and $\rho \geq (1-L\gamma_2)/2L$. Böhm (2022) prove best-iterate $\mathcal{O}(1/N)$ convergence of OG for $\rho < 1/2L$, i.e., for the same range of ρ as in the best-known result for EG.

Last-iterate convergence under ρ-negative comonotonicity. In a very recent work, Luo & Tran-Dinh (2022) prove the first last-iterate $\mathcal{O}(1/N)$ convergence results for EG and OG applied to solve VIP with ρ-negative comonotone L-Lipschitz operator. Both results rely on the usage of $\gamma_1 = \gamma_2$. Next, for EG the result from (Luo & Tran-Dinh, 2022) requires $\rho < 1/16L$ and for OG the corresponding result is proven for $\rho < 4/(27\sqrt{6}L)$. In contrast, for the accelerated (anchored) version of EG Lee & Kim (2021) prove $\mathcal{O}(1/N^2)$ last-iterate convergence rate for any $\rho < 1/2L$, which is a larger range of ρ than in the known results for EG/OG from (Luo & Tran-Dinh, 2022).

1.3. Contributions

◇ **Spectral viewpoint on negative comonotonicity.** Our work provides a spectral interpretation of negative comonotonicity, shedding some light on the relation between this assumption and classical monotonicity, Lipschitzness, and cocoercivity.

◇ **Closer look at the convergence of Proximal Point method.** We derive $\mathcal{O}(1/N)$ last-iterate and best-iterate convergence rates for PP under negative comonotonicity and star-negative comonotonicity assumptions, respectively. These results follow from existing ones (Iusem et al., 2003; Bauschke et al., 2021). However, we go further and show the tightness of the derived results via constructing matching worst-case examples and also propose counter-examples for the case when the stepsize is smaller than 2ρ.

◇ **New results for Extragradient-Based Methods.** We derive $\mathcal{O}(1/N)$ last-iterate convergence of EG and OG under milder assumptions on the negative comonotonicity parameter ρ than in the prior work by Luo & Tran-Dinh (2022), see the details in Table 1. We also provide alternative analyses of the best-iterate convergence of EG and OG under star-negative comonotonicity and recover

[3] Here and below we mean the rates of convergence in terms of the squared residual $\|x^N - x^{N-1}\|^2$ in the case of set-valued operators and $\|F(x^N)\|^2$ in the case of single-valued ones.

the best-known results in this case (Pethick et al., 2022; Böhm, 2022). Finally, we show that the range of allowed ρ cannot be improved for EG and OG via constructing counter-examples for these methods.

◇ **Constructive proofs.** We derive the proofs for the last-iterate convergence of PP, EG, and OG as well as worst-case examples for PP using using the performance estimation technique (Drori & Teboulle, 2014; Taylor et al., 2017c;a). In particular, it required us to extend some theoretical and program tools to handle negative comonotone and star-negative comonotone problems; see the details in App. B and Github-repository https://github.com/eduardgorbunov/Proximal_Point_and_Extragradient_based_methods_negative_comonotonicity, containing the codes for generating worst-case examples for PP, numerical verification of the derived results and symbolical verification of certain technical derivations. We believe that these tools are important on its own and can be applied in future works studying the convergence of different methods under negative comonotonicity.

2. A Closer Look at Negative Comonotonicity

Negative comonotonicity (also known as cohypomonotonicity) was originally introduced as a relaxation of monotonicity that is sufficient for the convergence of PP (Pennanen, 2002). This assumption is relatively weak: one can show that F is ρ-negative comonotone in a neighborhood of solution x^* for large enough ρ, if the (possibly set-valued) operator $F^{-1} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ has a Lipschitz localization around $(0, x^*) \in G_{F^{-1}}$, where $G_{F^{-1}}$ denotes the graph of F^{-1} (Pennanen, 2002, Proposition 7). The next lemma characterizes negative comonotone operators; it is technically very close to (Bauschke et al., 2011, Proposition 4.2) (on cocoercive operators).

Lemma 2.1. *$F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximally ρ-negative comonotone ($\rho \geq 0$) if and only if operator $\mathrm{Id} + 2\rho F$ is expansive.*

The proof of this lemma follows directly from the definition of negative comonotonicity. Among others, it implies the following result about the spectral properties of the Jacobian of negative comonotone operator (when it exists).

Theorem 2.2. *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be a continuously differentiable. Then, the following statements are equivalent:*

- *F is ρ-negative comonotone,*
- $\mathrm{Re}(1/\lambda) \geq -\rho$ *for all* $\lambda \in \mathrm{Sp}(\nabla F(x))$, $\forall x \in \mathbb{R}^d$.

We notice that the above theorem holds for any continuously differentiable operator F. In the case of the linear operator F, this result is known (Bauschke et al., 2021, Proposition 5.1). The condition $\mathrm{Re}(1/\lambda) \geq -\rho$ means that λ lies *outside* the disc in $\mathbb{C}$ centered at $-1/2\rho$ and having radius $1/2\rho$, see Figure 1. In particular, for the case of twice differentiable functions ρ-negative comonotonicity forbids the Hessian to have eigenvalues in $(-1/\rho, 0)$, i.e., eigenvalues of the Hessian have to be either negative with sufficiently large absolute value or non-negative. An alternate interpretation of Figure 1 can be formally made in terms of scaled relative graphs, see (Ryu et al., 2022); see also older references using such illustrations (Eckstein, 1989; Eckstein & Bertsekas, 1992), or (Giselsson & Boyd, 2016, arXiv version 1 to 3).

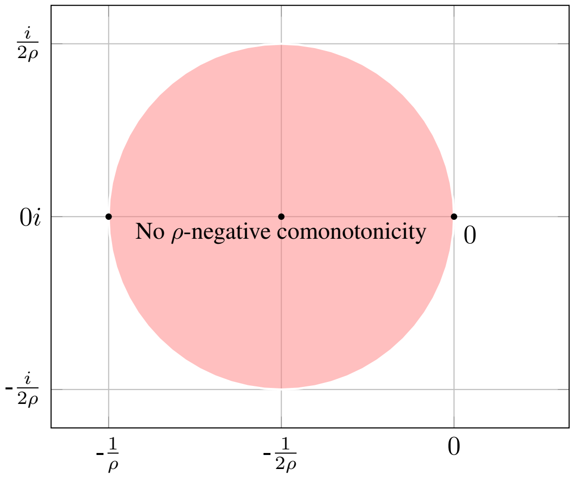


Figure 1: Visualization of Theorem 2.2. Red open disc corresponds to the constraint $\mathrm{Re}(1/\lambda) < -\rho$ that defines the set such that all eigenvalues the Jacobian of ρ-negative comonotone operator should lie outside this set.

Finally, we touch the following informal question: *to what extent negative comonotone operators are non-monotone?* To formalize a bit we consider a way more simpler question: *can negative comonotone operator have isolated zeros/solutions of VIP?* Unfortunately, the answer is no.

Theorem 2.3 (Corollary 3.15 from (Bauschke et al., 2021)[4]). *If $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximally ρ-negative comonotone, then the solution set $X^* = F^{-1}(0)$ is convex.*

Proof. The proof follows from the observations provided by Pennanen (2002). First, notice that F and its Yosida regularization $(F^{-1} + \rho \cdot \mathrm{Id})^{-1}$ have the same set of the solutions: $((F^{-1} + \rho \cdot \mathrm{Id})^{-1})^{-1}(0) = (F^{-1} + \rho \cdot \mathrm{Id})(0) = F^{-1}(0)$. Next, by definition (1) we have that maximal ρ-negative comonotonicity of F implies maximal monotonicity of $F^{-1} + \rho \cdot \mathrm{Id}$ that is equivalent to maximal monotonicity of $(F^{-1} + \rho \cdot \mathrm{Id})^{-1}$. Since the set of zeros of maximal mono-

[4] We were not aware of the results from (Bauschke et al., 2021) during the work on our paper.

tone operator is convex (Bauschke et al., 2011, Proposition 23.39), we have the result. □

Therefore, despite its apparent generality, negative comonotonicity is not satisfied (globally) for the many practical tasks that have isolated optima. Nevertheless, studying the convergence of traditional methods under negative comonotonicity can be seen as a natural step towards understanding their behaviors in more complicated non-monotonic cases.

3. Proximal Point Method

In this section, we consider Proximal Point method (Martinet, 1970; Rockafellar, 1976), which is usually written as $x^{k+1} = (F + \gamma\,\mathrm{Id})^{-1}\, x^k$ (where we assume here that $\gamma > 0$ is large enough so that the iteration is well and uniquely defined) or equivalently:

$$x^{k+1} = x^k - \gamma F(x^{k+1}). \qquad \text{(PP)}$$

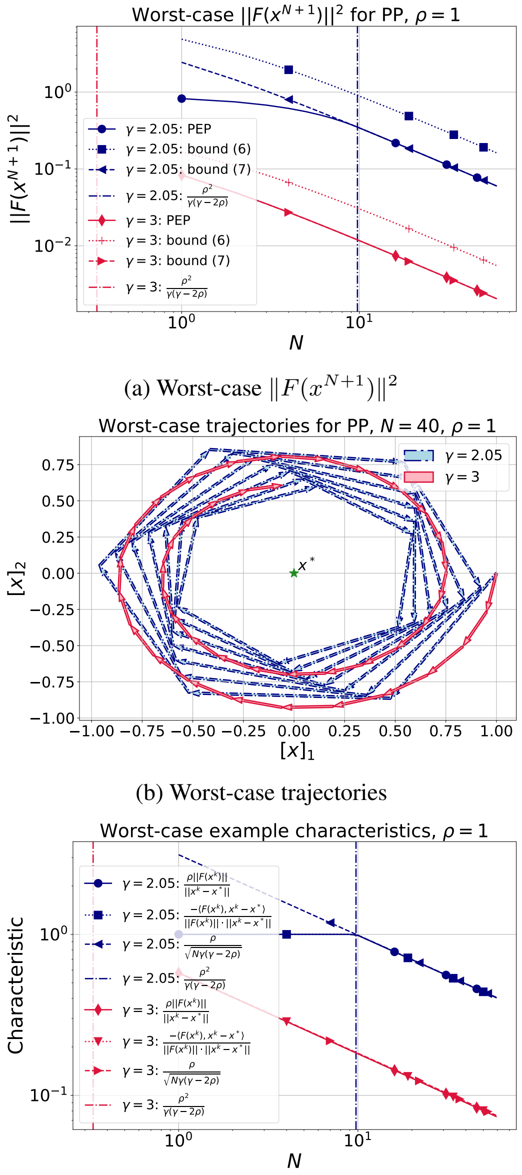
In particular, for given values of $N \in \mathbb{N}$, $R > 0$, $\rho > 0$, and $\gamma > 0$ we focus on the following question: *what guarantees can we prove on $\|x^N - x^{N-1}\|^2$ (in particular: as a function of N), where $\{x^k\}_{k=0}^N$ is generated by* PP *with stepsize γ after $N \geq 1$ iterations of solving IP with $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ being ρ-negative comonotone and $\|x^0 - x^*\|^2 \leq R^2$?* This kind of question can naturally be reformulated as an explicit optimization problem looking for worst-case problem instances, often referred to as *performance estimation problems* (PEPs), as introduced and formalized in (Drori & Teboulle, 2014; Taylor et al., 2017c;a):

$$\begin{aligned} \max_{F, x^0} \quad & \|x^N - x^{N-1}\|^2 && (4)\\ \text{s.t.} \quad & F \text{ satisfies } (1), \\ & \|x^0 - x^*\|^2 \leq R^2,\ 0 \in F(x^*), \\ & x^{k+1} = x^k - \gamma F(x^{k+1}), \quad k = 0, 1, \ldots, N-1. \end{aligned}$$

As we show in Appendix B, (4) can be *formulated* as a semidefinite program (SDP). For constructing and solving this SDP problem corresponding to (4) numerically, one can use the PEPit package (Goujaud et al., 2022) (after adding the class of ρ-negative comonotone operators to it), which thereby allows constructing worst-case guarantees and examples, numerically. Figure 2a shows the numerical results obtained by solving (4) for different values of N. We observe that worst-case value of (4) behaves as $\mathcal{O}(1/N)$ similarly to the monotone case.

Motivated by these numerical results, we derive the following convergence rates for PP.

Theorem 3.1. *Let $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be maximally ρ-star-negative comonotone. Then, for any $\gamma > 2\rho$ the iterates*



(a) Worst-case $\|F(x^{N+1})\|^2$

(b) Worst-case trajectories

(c) Study of the norms and angles

Figure 2: In (a), we report the solution of (4) for different values of γ and N. The plot illustrates that for the considered range of N and values of γ PP converges as $\mathcal{O}(1/N)$ in terms of $\|F(x^N)\|^2$. In (b), we show the worst-case trajectories of PP for $N = 40$. The form of trajectories hints that the worst-case operator is a rotation operator. For each particular choice of N and $\gamma > 2\rho$ we observed numerically that quantities $\rho\|F(x^k)\|^2/\|x^k - x^*\|$ and $-\langle F(x^k), x^k - x^*\rangle/(\|F(x^k)\| \cdot \|x^k - x^*\|)$ remain the same during the run of the method (the standard deviation of arrays $\{\rho\|F(x^k)\|^2/\|x^k - x^*\|\}_{k=1}^N$ and $\{-\langle F(x^k), x^k - x^*\rangle/(\|F(x^k)\| \cdot \|x^k - x^*\|)\}_{k=1}^N$ is of the order $10^{-6} - 10^{-7}$). Finally, in (c), we illustrate that these characteristics coincide with $\rho/\sqrt{N\gamma(\gamma - 2\rho)}$ as long as the total number of steps N is sufficiently large ($N \geq \max\{\rho^2/\gamma(\gamma-2\rho), 1\}$).

produced by PP *are well-defined and satisfy* $\forall N \geq 1$*:*

$$\frac{1}{N}\sum_{k=1}^{N}\|x^k - x^{k-1}\|^2 \leq \frac{\gamma\|x^0 - x^*\|^2}{(\gamma - 2\rho)N}. \tag{5}$$

If $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ *is maximally* ρ*-negative comonotone, then for any* $\gamma > 2\rho$ *and any* $k \geq 1$ *the iterates produced by* PP *satisfy*

$$\|x^{k+1} - x^k\| \leq \|x^k - x^{k-1}\|$$

and for any $N \geq 1$*:*

$$\|x^N - x^{N-1}\|^2 \leq \frac{\gamma\|x^0 - x^*\|^2}{(\gamma - 2\rho)N}. \tag{6}$$

Proof. We start with ρ-star-negative comonotone case. From the update rule of PP we have

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &= \|x^k - x^* - (x^k - x^{k+1})\|^2 \\ &= \|x^k - x^*\|^2 - 2\langle x^k - x^*, x^k - x^{k+1}\rangle \\ &\quad + \|x^k - x^{k+1}\|^2 \\ &= \|x^k - x^*\|^2 - 2\langle x^{k+1} - x^*, x^k - x^{k+1}\rangle \\ &\quad - \|x^k - x^{k+1}\|^2.\end{aligned}$$

Since $x^k - x^{k+1} = \gamma F(x^{k+1})$, where $F(x^{k+1})$ is some value of operator F at point x^{k+1}, we can apply ρ-star-negative comonotonicity and get

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \left(1 - \frac{2\rho}{\gamma}\right)\|x^k - x^{k+1}\|^2.$$

Telescoping the above inequality for $k = 0, \ldots, N-1$ and changing the index in the summation, we obtain (5). Next, to get the last-iterate convergence we use ρ-negative comonotonicity (1) inequality written for x^k and x^{k+1}:

$$\begin{aligned}\frac{1}{\gamma}\langle x^{k-1} - x^k - (x^k - x^{k+1}), x^k - x^{k+1}\rangle \\ \geq -\frac{\rho}{\gamma^2}\|x^{k-1} - x^k - (x^k - x^{k+1})\|^2,\end{aligned}$$

where we use that $(x^{k-1} - x^k)/\gamma$ and $(x^k - x^{k+1})/\gamma$ belongs to the values of F at points x^k and x^{k+1} respectively. Multiplying both sides by γ^2 and rearranging the terms, we get

$$\begin{aligned}\gamma\|x^k - x^{k+1}\|^2 &\leq \gamma\langle x^{k-1} - x^k, x^k - x^{k+1}\rangle \\ &\quad + \rho\|x^{k-1} + x^{k+1} - 2x^k\|^2.\end{aligned}$$

Finally, using $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$, which holds for any $a, b \in \mathbb{R}^d$, and rearranging the terms, we derive

$$\begin{aligned}\frac{\gamma}{2}\|x^k - x^{k+1}\|^2 &\leq \frac{\gamma}{2}\|x^{k-1} - x^k\|^2 \\ &\quad - \left(\frac{\gamma}{2} - \rho\right)\|x^{k-1} + x^{k+1} - 2x^k\|^2.\end{aligned}$$

Taking into account $\gamma > 2\rho$, we obtain $\|x^{k+1} - x^k\| \leq \|x^k - x^{k-1}\|$. Together with (5) it implies (6). □

First, the result from (5) implies only best-iterate $\mathcal{O}(1/N)$ rate – this result follows from (Iusem et al., 2003, Lemma 2)[5]. Such kind of guarantees are weaker the last-iterate ones but they do hold under the more general star-negative comonotonicity assumption. We notice that the result from (6) can be also obtained from non-expansiveness of PP update (Bauschke et al., 2021, Proposition 3.13 (iii))[6]. Note that the guarantee (6) matches the best-known guarantee for the monotone case (up to the factor $\gamma/(\gamma-2\rho)$) from (He & Yuan, 2015; Gu & Yang, 2020), and it is therefore natural to ask whether it is possible to improve factor $\gamma/(\gamma-2\rho)$ in the convergence guarantee of PP for the ρ-negative comonotone case.

To answer this question, one can use performance estimation again. In particular, using the trace heuristic for trying to find low-dimensional worst-case examples to (4), we obtain 2-dimensional worst-case examples for different values of γ and N, see Figure 2b and Figure 2c. These figures illustrate that the worst-case examples found numerically correspond to the scaled rotation operators (similar to Gu & Yang (2020) but with different angles). Moreover, the rotation angle and scaling parameter have non-trivial dependencies on number of iterations. These observations lead to the following result, which shows that the multiplicative cannot be removed asymptotically as N grows.

Theorem 3.2. *For any* $\rho > 0, \gamma > 2\rho$*, and* $N \geq \max\{\rho^2/\gamma(\gamma-2\rho), 1\}$ *there exists* ρ*-negatively comonotone single-valued operator* $F : \mathbb{R}^d \to \mathbb{R}^d$ *such that after* N *iterations* PP *with stepsize* γ *produces* x^{N+1} *satisfying*

$$\|F(x^{N+1})\|^2 \geq \frac{\|x^0 - x^*\|^2}{\gamma(\gamma - 2\rho)N\left(1 + \frac{1}{N}\right)^{N+1}}. \tag{7}$$

Indeed, one can pick the two-dimensional $F : \mathbb{R}^2 \to \mathbb{R}$*:* $F(x) = \alpha A x$ *with*

$$A = \begin{pmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{pmatrix}, \quad \alpha = \frac{|\cos\theta|}{\rho}$$

for $\theta \in (\pi/2, \pi)$ *such that* $\cos\theta = -\frac{\rho}{\sqrt{N\gamma(\gamma-2\rho)}}$.

Proof. Consider the linear operator $F(x) = \alpha A x$ described above. First, we verify its ρ-negative comonotonicity: for any $x, y \in \mathbb{R}^d$

$$\begin{aligned}\langle F(x) - F(y), x - y\rangle &= \alpha\langle A(x - y), x - y\rangle \\ &= \alpha\|A(x - y)\| \cdot \|x - y\| \cdot \cos\theta \\ &= -\frac{\cos^2\theta}{\rho}\|A(x - y)\|^2 \\ &= -\rho\|F(x) - F(y)\|^2,\end{aligned}$$

[5] We were not aware of the results from (Iusem et al., 2003) during the work on our paper.

[6] We were not aware of the results from (Bauschke et al., 2021) during the work on our paper.

where we use $\|A(x-y)\| = \|x-y\|$, since A is the rotation matrix. Next, one can check that $(I+\gamma\alpha A)^{-1}$ equals

$$\frac{1}{1+\gamma\alpha^2(\gamma-2\rho)}\begin{pmatrix}1+\gamma\alpha\cos\theta & \gamma\alpha\sin\theta\\ -\gamma\alpha\sin\theta & 1+\gamma\alpha\cos\theta\end{pmatrix}.$$

Since $x^{k+1} = (I+\gamma\alpha A)^{-1}x^k$, one can verify via direct computations that

$$\|x^{k+1}\|^2 = \frac{1}{1+\gamma\alpha^2(\gamma-2\rho)}\|x^k\|^2.$$

Unrolling this identity for $k = N, N-1, \ldots, 0$ and using $x^* = 0$, $\|F(x^{k+1})\| = \alpha\|Ax^{k+1}\| = \alpha\|x^{k+1}\|$, we get

$$\|F(x^{k+1})\|^2 = \alpha^2\left(\frac{1}{1+\gamma\alpha^2(\gamma-2\rho)}\right)^{N+1}\|x^0-x^*\|^2.$$

Maximizing the right-hand side in α we get that the optimal value is $\alpha = {}^1\!/\!\sqrt{N\gamma(\gamma-2\rho)}$. Since $\alpha\rho = |\cos\theta|$, we should assume that $N \geq {}^{\rho^2}\!/_{\gamma(\gamma-2\rho)}$. Plugging $\alpha = {}^1\!/\!\sqrt{N\gamma(\gamma-2\rho)}$ in the above formula for $\|F(x^{k+1})\|^2$ we get the result. □

Since $\exp(1) \leq (1+{}^1\!/_N)^{N+1} \leq 4$, the above result implies the tightness (up to a multiplicative constant) of Theorem 3.1. One should note again that both Theorem 3.1 and 3.2 rely on the assumption that $\gamma > 2\rho$ for the proximal operation to be well-defined. That is, these results are valid only for *large enough stepsizes*. This is a relatively rare phenomenon in optimization and variational inequalities. As the next theorem states, PP *is not guaranteed to converge if the stepsize is too small*, even if the proximal operation is well-defined.

Theorem 3.3. *For any $\rho > 0$ there exists ρ-negatively comonotone single-valued operator $F : \mathbb{R}^d \to \mathbb{R}^d$ such that* PP *does not converge to the solution of* VIP *for any $0 < \gamma \leq 2\rho$. In particular, one can take $F(x) = -{}^x\!/_\rho$.*

Proof. First, $F(x) = -{}^x\!/_\rho$ is ρ-negative comonotone: for any $x, y \in \mathbb{R}^d$ we have $\langle F(x)-F(y), x-y\rangle = -({}^1\!/_\rho)\|x-y\|^2 = -\rho\|F(x)-F(y)\|^2$. Next, the iterates of PP satisfy $x^{k+1} = x^k + \gamma{}^{x^{k+1}}\!/_\rho$. If $\gamma = \rho$, the next iterate is undefined. If $\gamma = 2\rho$, then $x^{k+1} = x^k$. Finally, when $\gamma \in (0,\rho)\cup(\rho,2\rho)$ we have $x^{k+1} = \frac{x^k}{(1-{}^\gamma\!/_\rho)}$ implying $\|x^{k+1}\| > \|x^k\|$, i.e., PP diverges. □

As a summary, Theorem 3.1 and Theorem 3.3 provide a complete picture of the convergence of PP under negative comonotonicity, including the upper bounds, and worst-case examples and counter-examples justifying the need of using large enough stepsizes for PP applied to ρ-negative comonotone IP/VIP.

4. Extragradient-Based Methods

Extragradient. The update rule of Extragradient method (Korpelevich, 1976) is defined as follows:

$$\begin{aligned}\widetilde{x}^k &= x^k - \gamma_1 F(x^k),\\ x^{k+1} &= x^k - \gamma_2 F(\widetilde{x}^k),\end{aligned}\quad \forall k \geq 0. \tag{EG}$$

In its pure form, EG has the same extrapolation (γ_1) and update (γ_2) stepsizes, i.e., $\gamma_1 = \gamma_2$. However, the existing analysis of EG under ρ-(star-)negative comonotonicity relies on the usage of $\gamma_2 < \gamma_1$ (Diakonikolas et al., 2021; Pethick et al., 2022). The following lemma sheds some light on this phenomenon.

Lemma 4.1. *Let F be L-Lipschitz and ρ-star-negative comonotone. Then, for any $k \geq 0$ the iterates produced by* EG *after $k \geq 0$ iterations satisfy*

$$\begin{aligned}\|x^{k+1}-x^*\|^2 \quad\leq\quad & \|x^k-x^*\|^2\\ & -\gamma_2(\gamma_1-2\rho-\gamma_2)\|F(\widetilde{x}^k)\|^2 \quad (8)\\ & -\gamma_1\gamma_2(1-L^2\gamma_1^2)\|F(x^k)\|^2. \quad (9)\end{aligned}$$

Proof sketch. The proof follows a quite standard pattern: we start with expanding the square $\|x^{k+1}-x^*\|^2$ and then rearrange the terms to get $\|x^k-x^*\|^2 - 2\gamma_2\langle\widetilde{x}^k - x^*, F(\widetilde{x}^k)\rangle - 2\gamma_1\gamma_2\langle F(x^k), F(\widetilde{x}^k)\rangle + \gamma_2^2\|F(\widetilde{x}^k)\|^2$ in the right-hand side. After that, it remains to estimate inner products. From ρ-star-negative comonotonicity we have $-2\gamma_2\langle\widetilde{x}^k - x^*, F(\widetilde{x}^k)\rangle \leq 2\rho\gamma_2\|F(\widetilde{x}^k)\|^2$. For the second inner product $-2\gamma_1\gamma_2\langle F(x^k), F(\widetilde{x}^k)\rangle$ we use $2\langle a,b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$, which holds for any $a, b \in \mathbb{R}^d$, and then apply L-Lipschitzness to upper bound the term $\gamma_1\gamma_2\|F(x^k) - F(\widetilde{x}^k)\|^2$. Finally, we rearrange the terms, see the full proof in Appendix C.1. □

From the above result one can easily notice that the choice of $\gamma_2 \leq \gamma_1 - 2\rho$ and $\gamma_1 < {}^1\!/_L$ implies best-iterate convergence in terms of the squared norm of the operator. However, in this proof, γ_2 should be positive, i.e., this proof is valid only for $\gamma_1 > 2\rho$. In other words, one can derive best-iterate $\mathcal{O}({}^1\!/_N)$ rate for EG whenever $\rho < {}^1\!/_{2L}$, which is also known from Pethick et al. (2022) (though Pethick et al. (2022) do not provide analogs of Lemma 4.1).

Next, to get the last-iterate convergence of EG we assume ρ-negative comonotonicity, since even for PP – a simpler algorithm – we need to do this. Moreover, even in the monotone case the existing proofs of the last-iterate convergence of EG rely on the usage of same stepsizes $\gamma_1 = \gamma_2 = \gamma$ (Gorbunov et al., 2022b; Cai et al., 2022b). This partially can be explained by the following fact: $\|F(x^{k+1})\|$ can be larger than $\|F(x^k)\|$ if $\gamma_1 \neq \gamma_2$ (Gorbunov et al., 2022b). Therefore, we also assume that $\gamma_1 = \gamma_2 = \gamma$ to derive last-iterate convergence rate.

However, as Lemma 4.1 indicates, the choice $\gamma_1 = \gamma_2 = \gamma$ may complicate the proof because the term from (8) becomes non-negative. Moreover, it is natural to expect that the proof will work for smaller range of ρ. Nevertheless, using computer-assisted approach, we derive that for any $\rho \leq {}^1/_{8L}$ and $4\rho \leq \gamma \leq {}^1/_{2L}$ EG the iterates of EG satisfy $\|F(x^{k+1})\| \leq \|F(x^k)\|$ which is the main building block of the obtained proof.

We summarized the derived upper-bounds for EG in the following result.

Theorem 4.2. *Let F be L-Lipschitz and ρ-star-negative comonotone with $\rho < {}^1/_{2L}$. Then, for any $2\rho < \gamma_1 < {}^1/_L$ and $0 < \gamma_2 \leq \gamma_1 - 2\rho$ the iterates produced by EG after $N \geq 0$ iteration satisfy*

$$\frac{1}{N+1}\sum_{k=0}^{N}\|F(x^k)\|^2 \leq \frac{\|x^0 - x^*\|^2}{\gamma_1\gamma_2(1 - L^2\gamma_1^2)(N+1)}. \quad (10)$$

If, in addition, F is ρ-negative comonotone with $\rho \leq {}^1/_{8L}$ and $\gamma_1 = \gamma_2 = \gamma$ such that $4\rho \leq \gamma \leq {}^1/_{2L}$, then for any $k \geq 0$ the iterates produced by EG satisfy $\|F(x^{k+1})\| \leq \|F(x^k)\|$ and for any $N \geq 1$

$$\|F(x^N)\|^2 \leq \frac{28\|x^0 - x^*\|^2}{N\gamma^2 + 320\gamma\rho}. \quad (11)$$

The results similar to (10) are known in the literature: Diakonikolas et al. (2021) derives best-iterate $\mathcal{O}({}^1/_N)$ convergence for $\rho < {}^1/_{8L}$ and Pethick et al. (2022) generalizes it to the case of any $\rho < {}^1/_{2L}$. In this sense, (10) recovers the one from Pethick et al. (2022), though the proof is different.

Next, the last-iterate convergence result from (11) holds for any $\rho \leq {}^1/_{8L}$, which is much smaller than the range $\rho < {}^1/_{2L}$ allowed for the best-iterate result. Nevertheless, the previous best-known last-iterate rate requires ρ to be smaller than ${}^1/_{16L}$ (Luo & Tran-Dinh, 2022), which is 2 times smaller than what is allowed for (11).

This discussion naturally leads us to the following question: *for given $L > 0$ what is the maximal possible ρ for which there exists a choice of stepsizes in EG such that it converges for any ρ-negative comonotone L-Lipschitz operator F?* This question is partially addressed by Pethick et al. (2022), who prove that if $\gamma_1 = {}^1/_L$, then for $\rho \geq {}^{(1-L\gamma_2)}/_{2L}$ EG does not necessary converge. Guided by the results obtained for PP, we make a further step and derive the following statement.

Theorem 4.3. *For any $L > 0$, $\rho \geq {}^1/_{2L}$, and any choice of stepsizes $\gamma_1, \gamma_2 > 0$ there exists ρ-negative comonotone L-Lipschitz operator F such that EG does not necessary converges on solving VIP with this operator F. In particular, for $\gamma_1 > {}^1/_L$ it is sufficient to take $F(x) = Lx$, and for $0 < \gamma_1 \leq {}^1/_L$ one can take $F(x) = LAx$, where $x \in \mathbb{R}^2$,*

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad \theta = \frac{2\pi}{3}.$$

This result corroborates Theorem 3.3 and known relationship between EG and PP. That is, from the one side, it is known that EG can be seen as an approximation of PP (Mishchenko et al., 2020, Theorem 1). Since for PP converges only for the stepsizes larger than 2ρ, it is natural to expect that EG also needs to have at least one stepsize larger than 2ρ (otherwise, it can be seen as an approximation of PP with stepsize not larger than 2ρ that is known to be non-convergent). From the other side, unlike PP, EG does not converge for arbitrary large stepsizes, which is a standard phenomenon for explicit methods in optimization. In particular, one has to take $\gamma_1 \leq {}^1/_L$ (otherwise there exists a "very good" – L-cocoercive – operator such that EG diverges). These two observations explain the intuition behind Theorem 4.3.

Optimistic gradient. Optimistic gradient (Popov, 1980) is a single-call version of EG having the following form:

$$\begin{aligned} \widetilde{x}^k &= x^k - \gamma_1 F(\widetilde{x}^{k-1}), \quad \forall k > 0, \\ x^{k+1} &= x^k - \gamma_2 F(\widetilde{x}^k), \quad \forall k \geq 0, \end{aligned} \quad \text{(OG)}$$

where $\widetilde{x}^0 = x^0$. Guided by the results and intuition developed for EG, here we also deviate from the original form of OG, which has $\gamma_1 = \gamma_2$, and allow γ_1 and γ_2 being different. The main goal of the rest of this section is in the obtaining the results on the convergence of OG similar to what are derived for EG earlier in this section.

Before we move on, we would like to highlight the challenges in the analysis of OG. Although EG and OG can both be seen as approximations of PP (Mokhtari et al., 2020), they have some noticeable theoretical differences going beyond algorithmic ones. For example, even for monotone L-Lipschitz operator F the iterates produced by OG do not satisfy $\|F(x^{k+1})\| \leq \|F(x^k)\|$ or $\|F(\widetilde{x}^k)\| \leq \|F(x^k)\|$ in general (Gorbunov et al., 2022c), while for EG $\|F(x^{k+1})\| \leq \|F(x^k)\|$ holds (Gorbunov et al., 2022b). This fact makes the analysis of OG more complicated than in the case of EG. Moreover, the known convergence results in the monotone case for OG require smaller stepsizes than for EG (Gorbunov et al., 2022c; Cai et al., 2022b). In view of the obtained results for PP and EG, this fact highlights non-triviality of obtaining convergence results for OG under ρ-negative comonotonicity for the same range of allowed values ρ as for EG.

Nevertheless, we obtain the best-iterate $\mathcal{O}({}^1/_N)$ convergence of OG for any $\rho < {}^1/_{2L}$, i.e., for the same range of ρ as for EG. We also derive last-iterate $\mathcal{O}({}^1/_N)$ convergence of OG but for $\rho \leq {}^{5}/_{62L}$, which is a smaller range than we have for EG. The results are summarized below.

Theorem 4.4. *Let F be L-Lipschitz and ρ-star-negative comonotone with $\rho < {}^{1}/_{2L}$. Then, for any $2\rho < \gamma_1 < {}^{1}/_{L}$ and $0 < \gamma_2 \leq \min\{{}^{1}/_{L} - \gamma_1, \gamma_1 - 2\rho\}$ the iterates produced by OG after $N \geq 0$ iteration satisfy*

$$\frac{1}{N+1}\sum_{k=0}^{N}\|F(x^k)\|^2 \leq \frac{\|x^0 - x^*\|^2}{\gamma_1\gamma_2(1 - L^2(\gamma_1+\gamma_2)^2)(N+1)}. \tag{12}$$

If, in addition, F is ρ-negative comonotone with $\rho \leq {}^{5}/_{62L}$ and $\gamma_1 = \gamma_2 = \gamma$ such that $4\rho \leq \gamma \leq {}^{10}/_{31L}$, then for any $N \geq 1$ the iterates produced by OG satisfy

$$\|F(x^N)\|^2 \leq \frac{717\|x^0 - x^*\|^2}{N\gamma(\gamma - 3\rho) + 800\gamma^2}. \tag{13}$$

The derived best-iterate rate (12) for OG is not new: Böhm (2022) proves a similar result for the same range of ρ, though the proof that we provide differs from the proof by Böhm (2022). Similarly to the case of EG, it is valid for any $\rho < {}^{1}/_{2L}$. Next, the last-iterate $\mathcal{O}({}^{1}/_{N})$ rate is recently obtained for OG by Luo & Tran-Dinh (2022). It holds for any $\rho < {}^{8}/_{(27\sqrt{6}L)}$, while the rate that we obtain is valid for any $\rho \leq {}^{5}/_{62L}$, which is ≈ 1.33 times larger range.

Finally, as for EG, we derive the following result about the largest possible range for ρ in the case of OG.

Theorem 4.5. *For any $L > 0$, $\rho \geq {}^{1}/_{2L}$, and any choice of stepsizes $\gamma_1, \gamma_2 > 0$ there exists ρ-negative comonotone L-Lipschitz operator F such that OG does not necessary converges on solving VIP with this operator F. In particular, for $\gamma_1 > {}^{1}/_{L}$ it is sufficient to take $F(x) = Lx$, and for $0 < \gamma_1 \leq {}^{1}/_{L}$ one can take $F(x) = LAx$, where $x \in \mathbb{R}^2$,*

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad \theta = \frac{2\pi}{3}.$$

Note that the counter-examples are exactly the same as for EG. Moreover, since OG can be seen as an approximation of PP, this result is expected and the same has the same intuition behind as Theorem 4.3.

5. Discussion

In this work, we studied worst-case convergence of methods for solving IP/VIP with (star-)negative-comonotone operators, which we believe is an important first step for going beyond the very popular monotonocity assumption, that is often not satisfied in modern applications.

Namely, we study the proximal point (PP), the extragradient (EG), and the optimistic gradient (OG) methods. Although the basic understanding of the convergence of PP and best-iterate convergence of EG and OG is relatively complete, several open-questions about last-iterate convergence of EG and OG remain. In particular, it is unclear what is the largest possible range for ρ for which one can guarantee last-iterate $\mathcal{O}({}^{1}/_{N})$ convergence of EG/OG under ρ-negative comonotonicity and L-Lipschitzness.

Moreover, another important direction for future research is identifying weaker assumptions allowing to prove non-asymptotic convergence rates for PP/EG/OG and at the same time allowing to have isolated optima or non-convex solution sets, as discussed in Section 2. Finally, it would be very important to extend the results to the stochastic case; see (Pethick et al., 2023) for the recent advances in this direction.

Acknowledgements

We thank Axel Böhm for pointing us to the reference (Böhm, 2022), which we were not aware of during the work on the paper. We also thank an anonymous reviewer for pointing us to the references (Bauschke et al., 2021; Iusem et al., 2003), which we were also not aware of during the work on the paper. The research of E. Gorbunov was partially supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Moscow Institute of Physics and Technology dated November 1, 2021 No. 70-2021-00138.

References

Anagnostides, I., Panageas, I., Farina, G., and Sandholm, T. On last-iterate convergence beyond zero-sum games. In *International Conference on Machine Learning*, pp. 536–581. PMLR, 2022. (Cited on page 2)

Bauschke, H. H., Combettes, P. L., et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011. (Cited on pages 2, 4, 5, and 15)

Bauschke, H. H., Moursi, W. M., and Wang, X. Generalized monotone operators and their averaged resolvents. *Mathematical Programming*, 189:55–74, 2021. (Cited on pages 3, 4, 6, and 9)

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. *Robust optimization*. Princeton university press, 2009. (Cited on page 1)

Berard, H., Gidel, G., Almahairi, A., Vincent, P., and Lacoste-Julien, S. A closer look at the optimization landscapes of generative adversarial networks. *arXiv preprint arXiv:1906.04848*, 2019. (Cited on page 2)

Böhm, A. Solving nonconvex-nonconcave min-max problems exhibiting weak minty solutions. *arXiv preprint arXiv:2201.12247*, 2022. (Cited on pages 1, 3, 4, and 9)

Bot, R. I., Csetnek, E. R., and Nguyen, D.-K. Fast OGDA in continuous and discrete time. *arXiv preprint arXiv:2203.10947*, 2022. (Cited on page 3)

Brown, N. and Sandholm, T. Solving imperfect-information games via discounted regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1829–1836, 2019. (Cited on page 2)

Cai, Y., Oikonomou, A., and Zheng, W. Accelerated algorithms for monotone inclusions and constrained nonconvex-nonconcave min-max optimization. *arXiv preprint arXiv:2206.05248*, 2022a. (Cited on page 2)

Cai, Y., Oikonomou, A., and Zheng, W. Tight last-iterate convergence of the extragradient and the optimistic gradient descent-ascent algorithm for constrained monotone variational inequalities. *arXiv preprint arXiv:2204.09228*, 2022b. (Cited on pages 2, 3, 7, 8, and 25)

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training gans with optimism. In *International Conference on Learning Representations*, 2018. (Cited on page 1)

Daskalakis, C., Foster, D. J., and Golowich, N. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33:5527–5540, 2020. (Cited on page 2)

Daskalakis, C., Skoulakis, S., and Zampetakis, M. The complexity of constrained min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1466–1478, 2021. (Cited on page 2)

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*. PMLR, 2021. (Cited on pages 1, 2, 3, 7, and 8)

Drori, Y. and Teboulle, M. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1):451–482, 2014. (Cited on pages 2, 4, 5, and 15)

Eckstein, J. *Splitting methods for monotone operators with applications to parallel optimization*. PhD thesis, Massachusetts Institute of Technology, 1989. (Cited on page 4)

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. (Cited on page 4)

Farina, G., Kroer, C., Brown, N., and Sandholm, T. Stable-predictive optimistic counterfactual regret minimization. In *International conference on machine learning*, pp. 1853–1862. PMLR, 2019. (Cited on page 2)

Gidel, G., Berard, H., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial nets. In *ICLR*, 2019. (Cited on page 1)

Giselsson, P. and Boyd, S. Linear convergence and metric selection for douglas-rachford splitting and admm. *IEEE Transactions on Automatic Control*, 62(2):532–544, 2016. (Cited on page 4)

Golowich, N., Pattathil, S., and Daskalakis, C. Tight last-iterate convergence rates for no-regret learning in multiplayer games. *arXiv preprint arXiv:2010.13724*, 2020a. (Cited on pages 2 and 3)

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, 2020b. (Cited on pages 2 and 3)

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. (Cited on page 1)

Gorbunov, E., Danilova, M., Dobre, D., Dvurechensky, P., Gasnikov, A., and Gidel, G. Clipped stochastic methods for variational inequalities with heavy-tailed noise. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022a. (Cited on page 2)

Gorbunov, E., Loizou, N., and Gidel, G. Extragradient method: $O(1/k)$ last-iterate convergence for monotone variational inequalities and connections with cocoercivity. pp. 366–402, 2022b. (Cited on pages 2, 3, 7, 8, and 16)

Gorbunov, E., Taylor, A., and Gidel, G. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. *arXiv preprint arXiv:2205.08446*, 2022c. (Cited on pages 2, 3, 8, and 25)

Goujaud, B., Moucer, C., Glineur, F., Hendrickx, J., Taylor, A., and Dieuleveut, A. Pepit: computer-assisted worst-case analyses of first-order optimization methods in python. *arXiv preprint arXiv:2201.04040*, 2022. (Cited on pages 2, 5, and 16)

Gu, G. and Yang, J. Tight sublinear convergence rate of the proximal point algorithm for maximal monotone inclusion problems. *SIAM Journal on Optimization*, 30(3): 1905–1921, 2020. (Cited on pages 3 and 6)

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. (Cited on page 1)

He, B. and Yuan, X. On the convergence rate of douglas–rachford operator splitting method. *Mathematical Programming*, 153(2):715–722, 2015. (Cited on pages 3 and 6)

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. (Cited on pages 3, 6, and 9)

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. (Cited on page 1)

Korpelevich, G. M. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976. (Cited on pages 1 and 7)

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009. (Cited on page 2)

Lee, S. and Kim, D. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 34, 2021. (Cited on pages 1, 2, and 3)

Loizou, N., Berard, H., Gidel, G., Mitliagkas, I., and Lacoste-Julien, S. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *arXiv preprint arXiv:2107.00052*, 2021. (Cited on page 2)

Luo, Y. and Tran-Dinh, Q. Last-iterate convergence rates and randomized block-coordinate variant of extragradient-type methods for co-monotone equations. *preprint*, 2022. (Cited on pages 2, 3, 8, 9, and 25)

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *ICLR 2018*, 2018. (Cited on page 1)

Martinet, B. Regularisation d'inequations variationelles par approximations successives. *Revue Francaise d'Informatique et de Recherche Operationelle*, 4:154–159, 1970. (Cited on page 5)

Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for gans do actually converge? In *International conference on machine learning*, pp. 3481–3490. PMLR, 2018. (Cited on page 2)

Mishchenko, K., Kovalev, D., Shulgin, E., Richtárik, P., and Malitsky, Y. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pp. 4573–4582. PMLR, 2020. (Cited on page 8)

Mokhtari, A., Ozdaglar, A., and Pattathil, S. Proximal point approximations achieving a convergence rate of $O(1/k)$ for smooth convex-concave saddle point problems: Optimistic gradient and extra-gradient methods. *arXiv preprint arXiv:1906.01115*, 2019. (Cited on page 1)

Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pp. 1497–1507. PMLR, 2020. (Cited on page 8)

Nagarajan, V. and Kolter, J. Z. Gradient descent gan optimization is locally stable. *Advances in neural information processing systems*, 30, 2017. (Cited on page 2)

Narang, A., Faulkner, E., Drusvyatskiy, D., Fazel, M., and Ratliff, L. J. Multiplayer performative prediction: Learning in decision-dependent games. *arXiv preprint arXiv:2201.03398*, 2022. (Cited on page 1)

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. (Cited on pages 1 and 2)

Pennanen, T. Local convergence of the proximal point algorithm and multiplier methods without monotonicity. *Mathematics of Operations Research*, 27(1):170–191, 2002. (Cited on page 4)

Pethick, T., Latafat, P., Patrinos, P., Fercoq, O., and Cevherå, V. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In *International Conference on Learning Representations*, 2022. (Cited on pages 1, 2, 3, 4, 7, 8, and 17)

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. (Cited on page 9)

Popov, L. D. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980. (Cited on pages 1 and 8)

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

Ryu, E. and Yin, W. Large-scale convex optimization via monotone operators, 2020. (Cited on page 2)

Ryu, E. K., Taylor, A. B., Bergeling, C., and Giselsson, P. Operator splitting performance estimation: Tight contraction factors and optimal parameter selection. *SIAM Journal on Optimization*, 30(3):2251–2271, 2020. (Cited on pages 2, 15, and 16)

Ryu, E. K., Hannah, R., and Yin, W. Scaled relative graphs: Nonexpansive operators via 2d euclidean geometry. *Mathematical Programming*, 194(1):569–619, 2022. (Cited on page 4)

Taylor, A. B., Hendrickx, J. M., and Glineur, F. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 27 (3):1283–1313, Jan 2017a. ISSN 1095-7189. (Cited on pages 2, 4, 5, and 15)

Taylor, A. B., Hendrickx, J. M., and Glineur, F. Performance estimation toolbox (pesto): automated worst-case analysis of first-order optimization methods. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pp. 1278–1283. IEEE, 2017b. (Cited on pages 2 and 16)

Taylor, A. B., Hendrickx, J. M., and Glineur, F. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1):307–345, 2017c. (Cited on pages 2, 4, 5, and 15)

Tran-Dinh, Q. The connection between nesterov's accelerated methods and halpern fixed-point iterations. *arXiv preprint arXiv:2203.04869*, 2022. (Cited on page 3)

Tran-Dinh, Q. and Luo, Y. Halpern-type accelerated and splitting algorithms for monotone inclusions. *arXiv preprint arXiv:2110.08150*, 2021. (Cited on page 3)

Yoon, T. and Ryu, E. K. Accelerated algorithms for smooth convex-concave minimax problems with $\mathrm{O}(1/k^2)$ rate on squared gradient norm. In *International Conference on Machine Learning*, pp. 12098–12109. PMLR, 2021. (Cited on pages 2 and 3)

Contents

A. Missing Proofs and Details From Section 2

Lemma A.1 (Lemma 2.1)**.** *$F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is ρ-negative comonotone ($\rho \geq 0$) if and only if operator $\mathrm{Id} + 2\rho F$ is expansive.*

Proof. Expansiveness of operator $\mathrm{Id} + 2\rho F$ means that for any $x, y \in \mathbb{R}^d$

$$\|x + 2\rho F(x) - y - 2\rho F(y)\|^2 \geq \|x - y\|^2,$$

where $F(x)$ and $F(y)$ represent the arbitrary elements from the values of F measured at x and y, respectively. Expanding the square in the left-hand side of the above inequality, we get

$$\|x - y\|^2 + 4\rho\langle F(x) - F(y), x - y\rangle + 4\rho^2\|F(x) - F(y)\|^2 \geq \|x - y\|^2.$$

The above inequality is equivalent to ρ-negative comonotonicity (1). □

Theorem A.2 (Theorem 2.2)**.** *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be a continuously differentiable. Then, the following statements are equivalent:*

- *F is ρ-negative comonotone,*
- $\mathrm{Re}(1/\lambda) \geq -\rho$ *for all* $\lambda \in \mathrm{Sp}(\nabla F(x))$, $\forall x \in \mathbb{R}^d$.

Proof. For a complex number λ condition $\mathrm{Re}(1/\lambda) \geq -\rho$ is equivalent to $|\lambda + 1/2\rho| \geq 1/2\rho$. Indeed, for $\lambda = \lambda_1 + i\lambda_2$, $\lambda_1, \lambda_2 \in \mathbb{C}$ we have

$$\mathrm{Re}\left(\frac{1}{\lambda}\right) = \frac{\lambda_1}{\lambda_1^2 + \lambda_2^2} \geq -\rho \quad \Longleftrightarrow \quad \lambda_1^2 + \lambda_2^2 + \frac{\lambda_1}{\rho} \geq 0 \quad \Longleftrightarrow \quad \left|\lambda + \frac{1}{2\rho}\right| \geq \frac{1}{2\rho}, \tag{14}$$

i.e., $\mathrm{Re}(1/\lambda) \geq -\rho$ means that λ lies in the disc in $\mathbb{C}$ centered at $(-1/2\rho)$ and radius $1/2\rho$. From the other side, ρ-negative comonotonicity of F is equivalent to expansiveness of $\mathrm{Id} + 2\rho F$ (Lemma 2.1), which is equivalent to $|\lambda| \geq 1$ for any $\lambda \in \mathrm{Sp}(I + 2\rho\nabla F(x))$ and for any $x \in \mathbb{R}^d$. Since

$$\mathrm{Sp}(I + 2\rho\nabla F(x)) = \{1 + 2\rho\lambda \mid \lambda \in \mathrm{Sp}(\nabla F(x))\},$$

we get that $|1 + 2\rho\lambda| \geq 1$ for any $\lambda \in \mathrm{Sp}(\nabla F(x))$ and any $x \in \mathbb{R}^d$. Taking into account (14), we obtain the desired result. □

B. Missing Details on PEP From Section 3

On PEP formulation (4). To find the tight convergence rate of PP and build worst-case examples of ρ-negative comonotone operators for PP, we consider problem (4), which we restate below for convenience:

$$\begin{aligned} \max_{F,d,x^0} \quad & \|x^N - x^{N-1}\|^2 \\ \text{s.t.} \quad & F:\mathbb{R}^d \rightrightarrows \mathbb{R}^d \text{ is } \rho\text{-negative comonotone}, \\ & \|x^0 - x^*\|^2 \le R^2,\ 0 \in F(x^*), \\ & x^{k+1} = x^k - \gamma F(x^{k+1}), \quad k = 0,1,\dots,N-1. \end{aligned} \tag{15}$$

The above problem requires maximization over *infinitely-dimensional* space of ρ-negative comonotone operators. To solve such a problem numerically, one can properly reformulate it to a *finite-dimensional* one. Whereas PEPs were introduced by Drori & Teboulle (2014), thie reformulation technique was provided in Taylor et al. (2017c;a) in the context of optimization problems, and was extended to problems involving (monotone) operators in Ryu et al. (2020). In particular, instead of (15), one can consider an equivalent finite-dimensional problem

$$\max_{\substack{d \\ x^*,x^0,x^1,\dots,x^N \in \mathbb{R}^d \\ g^*,g^0,g^1,\dots,g^N \in \mathbb{R}^d}} \quad \|x^N - x^{N-1}\|^2 \tag{16}$$

$$\text{s.t.} \quad F:\mathbb{R}^d \rightrightarrows \mathbb{R}^d \text{ is } \rho\text{-negative comonotone}, \tag{17}$$

$$g^k \in F(x^k), \quad k = *,0,1,\dots,N, \quad g^* = 0, \tag{18}$$

$$\|x^0 - x^*\|^2 \le R^2,$$

$$x^{k+1} = x^k - \gamma g^{k+1}, \quad k = 0,1,\dots,N-1.$$

Although the above problem is finite-dimensional, it has non-trivial constraints (17)-(18), which can be handled via the following result.

Theorem B.1. *Let $\{(x^k,g^k)\}_{k=0}^N \subseteq \mathbb{R}^d \times \mathbb{R}^d$ be some finite set of pairs of points in $\mathbb{R}^d$. There exists a maximal ρ-negative comonotone operator $F:\mathbb{R}^d \rightrightarrows \mathbb{R}^d$ such that $g^k \in F(x^k)$, $k = 0,\dots,N$ if and only if*

$$\langle g^i - g^j, x^i - x^j\rangle \ge -\rho\|g^i - g^j\|^2 \quad \forall i,j = 0,\dots,N. \tag{19}$$

Proof. Following Ryu et al. (2020), we say that the set $\{(x^k,g^k)\}_{k=0}^N \subseteq \mathbb{R}^d \times \mathbb{R}^d$ is $\mathcal{M}$-interpolable if there exists a maximal monotone (0-negative comonotone) operator $F:\mathbb{R}^d \rightrightarrows \mathbb{R}^d$ such that $g^k \in F(x^k)$, $k = 0,\dots,N$. One can introduce a similar notion for ρ-negative comonotone case, i.e., we say that the set $\{(x^k,g^k)\}_{k=0}^N \subseteq \mathbb{R}^d \times \mathbb{R}^d$ is $\mathcal{NM}_\rho$-interpolable if there exists a maximal ρ-negative comonotone operator $F:\mathbb{R}^d \rightrightarrows \mathbb{R}^d$ such that $g^k \in F(x^k)$, $k = 0,\dots,N$. Next, for convenience we denote the classes of maximal monotone and maximal ρ-negative comonotone operators as $\mathcal{M}$ and $\mathcal{NM}_\rho$ respectively. Then, in view of the maximal monotone extension theorem (Bauschke et al., 2011, Theorem 20.21), the set $\{(x^k,g^k)\}_{k=0}^N \subseteq \mathbb{R}^d \times \mathbb{R}^d$ is $\mathcal{M}$-interpolable if and only if $\langle g^i - g^j, x^i - x^j\rangle \ge 0$ for any $i,j = 0,\dots,N$.

For obtaining the desired result, we simply reduce the problem of finding a maximal ρ-negative comonotone interpolating operator for the set $\{(x^k,g^k)\}_{k=0}^N$ as that of finding a maximal monotone operator interpolating $\{(x^k + \rho g^k, g^k)\}_{k=0}^N$, which is a consequence of the following equivalence: an operator $F:\mathbb{R}^d \rightrightarrows \mathbb{R}$ is maximal ρ-negatively monotone if and only if $F^{-1} + \rho\mathrm{Id}$ is maximal monotone. More precisely, the reasoning is as follows:

$$\begin{aligned} \langle g^i - g^j, x^i - x^j\rangle \quad & \ge \quad -\rho\|g^i - g^j\|^2 \quad \forall i,j = 0,\dots,N \\ & \Longleftrightarrow \quad \langle g^i - g^j, x^i + \rho g^i - (x^j + \rho g^j)\rangle \ge 0 \quad \forall i,j = 0,\dots,N \\ & \Longleftrightarrow \quad \exists\, T \in \mathcal{M}:\ x^i + \rho g^i \in T(g^i) \quad \forall i = 0,\dots,N \\ & \overset{Q=T-\rho\mathrm{Id}}{\Longleftrightarrow} \quad \exists\, Q:\ Q + \rho\mathrm{Id} \in \mathcal{M} \text{ and } x^i \in Q(g^i) \quad \forall i = 0,\dots,N \\ & \overset{F=Q^{-1}}{\Longleftrightarrow} \quad \exists\, F \in \mathcal{NM}_\rho \text{ and } g^i \in F(x^i) \quad \forall i = 0,\dots,N, \end{aligned}$$

where the last equivalence follows from the following fact: F is ρ-negative comonotone if and only if $F^{-1} + \rho\mathrm{Id}$ is monotone, thereby concluding the proof. □

In view of the above theorem, one can replace (17)-(18) constraints by $(N+1)(N+2)$ inequalities of the type (19) and get the following finite-dimensional problem, which is equivalent to (16):

$$\max_{\substack{d\\ x^*,x^0,x^1,\ldots,x^N\in\mathbb{R}^d\\ g^*,g^0,g^1,\ldots,g^N\in\mathbb{R}^d}} \quad \|x^N - x^{N-1}\|^2 \tag{20}$$

$$\text{s.t.} \quad \langle g^i - g^j, x^i - x^j\rangle \geq -\rho\|g^i - g^j\|^2, \quad i,j = *,0,1,\ldots,N, \quad g^* = 0, \tag{21}$$

$$\|x^0 - x^*\|^2 \leq R^2, \tag{22}$$

$$x^{k+1} = x^k - \gamma g^{k+1}, \quad k = 0,1,\ldots,N-1.$$

We notice that $x^1,\ldots,x^N$ are linear combinations of $x^0, g^0, g^1,\ldots,g^N$ and we also have constraint $g^* = 0$. Therefore, one can reduce the number of maximization vector-variables to $N+3$: $x^*, x^0, g^0, g^1, \ldots, g^N$. Moreover, the above problem is linear w.r.t. the inner products of all possible pairs of vectors $x^*, x^0, g^0, g^1, \ldots, g^N$. This means that (20) is linear w.r.t. the elements of matrix $G = V^\top V$, where $V = (x^*, x^0, g^0, g^1, \ldots, g^N)$, and one can reformulate the problem (20) as the following semidefinite programming (SDP)

$$\max_{G\in\mathbb{S}_+^{N+3}} \quad \mathrm{Tr}(M_0 G) \tag{23}$$

$$\text{s.t.} \quad \mathrm{Tr}(M_i G) \leq 0, \quad i = 1,2\ldots,(N+2)(N+3),$$

$$\mathrm{Tr}(M_{-1}G) \leq R^2.$$

Here $\mathbb{S}_+^{N+3}$ denotes the set of symmetric positive semidefinite matrices of size $(N+3)\times(N+3)$ and matrices M_0, $\{M_i\}_{i=1}^{(N+2)(N+3)}$, and M_{-1} encode the objective (20) and constraints (21)-(22), respectively. We do not provide the exact formulas for these matrices and refer to the examples of how they can be constructed provided in (Ryu et al., 2020; Gorbunov et al., 2022b). We note that in toolboxes like PESTO (Taylor et al., 2017b) and PEPit (Goujaud et al., 2022), the process of constructing matrices M_0, $\{M_i\}_{i=1}^{(N+2)(N+3)}$, M_{-1} is fully automated.

On low-dimensional worst-case examples. It is worth mentioning that for any $G\in\mathbb{S}_+^{N+3}$ one can reconstruct vectors $x^*, x^0, g^0, g^1, \ldots, g^N \in \mathbb{R}^{N+3}$ such that G is their Gram matrix, i.e., find $V = (x^*, x^0, g^0, g^1, \ldots, g^N) \in \mathbb{R}^{(N+3)\times(N+3)}$ such that $G = V^\top V$. More precisely, if $\mathrm{rank}(G) = r \leq N+3$, then one can find $x^*, x^0, g^0, g^1, \ldots, g^N \in \mathbb{R}^r$ such that G is the Gram matrix of this set of vectors.

Therefore, to obtain low-dimensional worst-case trajectories like ones illustrated in Figure 2b, we need to find low-rank solution of (23). To do so, we apply *trace heuristic* (Taylor et al., 2017b), where we first find numerically an approximate optimal value v_* of problem (23) and then solve the following problem:

$$\min_{G\in\mathbb{S}_+^{N+3}} \quad \mathrm{Tr}(G) \tag{24}$$

$$\text{s.t.} \quad \mathrm{Tr}(M_i G) \leq 0, \quad i = 1,2\ldots,(N+2)(N+3),$$

$$\mathrm{Tr}(M_{-1}G) \leq R^2,$$

$$\mathrm{Tr}(M_0 G) = v_*. \tag{25}$$

Constraint (25) enforces that by solving the above problem we find numerically an approximate solution for (23) of a comparable quality and minimization of $\mathrm{Tr}(G)$ can be seen as an “approximate minimization” of $\mathrm{rank}(G)$.

C. Missing Proofs and Details From Section 4

This appendix provides the complete proofs of the results of EG and OG.

C.1. Extragradient method

C.1.1. Guarantees for the averaged squared norm of the operator

Theorem 4.2 consists of the two results: one requires only star negative comonotonicity and gives the rate in terms of the averaged squared norms of the operator along the trajectory and the other one requires negative comonotonicity but gives last-iterate convergence guarantee. We start with the first result which is a simplification of Theorem 3.1 from Pethick et al. (2022). Our proof is a bit more explicit in terms of why we need γ_1 to be large, because it relies on the following lemma.

Lemma C.1. *Let F be L-Lipschitz and ρ-star-negative comonotone. Then, for any $k \geq 0$ the iterates produced by EG after $k \geq 0$ iterations satisfy*

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \gamma_2\left(\gamma_1 - 2\rho - \gamma_2\right)\|F(\widetilde{x}^k)\|^2 - \gamma_1\gamma_2(1 - L^2\gamma_1^2)\|F(x^k)\|^2. \tag{26}$$

Proof. By the definition of x^{k+1} and $\widetilde{x}^k$ we have

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - 2\gamma_2\langle x^k - x^*, F(\widetilde{x}^k)\rangle + \gamma_2^2\|F(\widetilde{x}^k)\|^2\\ &= \|x^k - x^*\|^2 - 2\gamma_2\langle \widetilde{x}^k - x^*, F(\widetilde{x}^k)\rangle - 2\gamma_1\gamma_2\langle F(x^k), F(\widetilde{x}^k)\rangle + \gamma_2^2\|F(\widetilde{x}^k)\|^2.\end{aligned}$$

Next, we estimate the second term in the right-hand side using star-negative comonotonicity:

$$\|x^{k+1} - x^*\|^2 \overset{(2)}{\leq} \|x^k - x^*\|^2 + \gamma_2\left(2\rho + \gamma_2\right)\|F(\widetilde{x}^k)\|^2 - 2\gamma_1\gamma_2\langle F(x^k), F(\widetilde{x}^k)\rangle.$$

Finally, we handle the last term in the right-hand side of the above inequality using $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$, which holds for any $a, b \in \mathbb{R}^d$, and then applying L-Lipschitzness of F:

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &\leq \|x^k - x^*\|^2 - \gamma_2\left(\gamma_1 - 2\rho - \gamma_2\right)\|F(\widetilde{x}^k)\|^2 - \gamma_1\gamma_2\|F(x^k)\|^2\\ &\quad +\gamma_1\gamma_2\|F(x^k) - F(\widetilde{x}^k)\|^2\\ &\overset{(3)}{\leq} \|x^k - x^*\|^2 - \gamma_2\left(\gamma_1 - 2\rho - \gamma_2\right)\|F(\widetilde{x}^k)\|^2 - \gamma_1\gamma_2\|F(x^k)\|^2\\ &\quad +\gamma_1\gamma_2 L^2\|x^k - \widetilde{x}^k\|^2.\end{aligned}$$

Taking into account $x^k - \widetilde{x}^k = \gamma_1 F(x^k)$ and rearranging the terms, we get the result. □

This lemma implies the first part of Theorem 4.2 and even a bit more.

Theorem C.2 (First part of Theorem 4.2)**.** *Let F be L-Lipschitz and ρ-star-negative comonotone with $\rho < 1/2L$. If $2\rho < \gamma_1 < 1/L$ and $0 < \gamma_2 \leq \gamma_1 - 2\rho$, then the iterates produced by EG after $N \geq 0$ iteration satisfy*

$$\frac{1}{N+1}\sum_{k=0}^{N}\|F(x^k)\|^2 \leq \frac{\|x^0 - x^*\|^2}{\gamma_1\gamma_2(1 - L^2\gamma_1^2)(N+1)}. \tag{27}$$

If $2\rho < \gamma_1 \leq 1/L$ and $0 < \gamma_2 < \gamma_1 - 2\rho$, then the iterates produced by EG after $N \geq 0$ iteration satisfy

$$\frac{1}{N+1}\sum_{k=0}^{N}\|F(\widetilde{x}^k)\|^2 \leq \frac{\|x^0 - x^*\|^2}{\gamma_2(\gamma_1 - 2\rho - \gamma_2)(N+1)}. \tag{28}$$

Proof. First, we consider the case when $2\rho < \gamma_1 < 1/L$ and $0 < \gamma_2 \leq \gamma_1 - 2\rho$. In this case, $\gamma_2(\gamma_1 - 2\rho - \gamma_2) \geq 0$ and $\gamma_1\gamma_2(1 - L^2\gamma_1^2) > 0$. Therefore, Lemma 4.1 implies

$$\gamma_1\gamma_2(1 - L^2\gamma_1^2)\|F(x^k)\|^2 \leq \|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2.$$

Summing up the above inequality for $k = 0, \ldots, N$, dividing the result by $\gamma_1\gamma_2(1 - L^2\gamma_1^2)(N+1)$, and using $-\|x^{N+1} - x^*\|^2 \leq 0$, we get (27).

Next, we consider the case when $2\rho < \gamma_1 \leq {}^1\!/\!{}_L$ and $0 < \gamma_2 < \gamma_1 - 2\rho$. In this case, $\gamma_2(\gamma_1 - 2\rho - \gamma_2) > 0$ and $\gamma_1\gamma_2(1 - L^2\gamma_1^2) \geq 0$. Therefore, Lemma 4.1 implies

$$\gamma_2(\gamma_1 - 2\rho - \gamma_2)\|F(\widetilde{x}^k)\|^2 \leq \|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2.$$

Summing up the above inequality for $k = 0, \ldots, N$, dividing the result by $\gamma_2(\gamma_1 - 2\rho - \gamma_2)(N+1)$, and using $-\|x^{N+1} - x^*\|^2 \leq 0$, we get (28). □

C.1.2. Last-iterate guarantees

We start with the following lemma:

Lemma C.3. *Let F be L-Lipschitz and ρ-negative comonotone. Then for any $k \geq 0$ the iterates produced by EG with $\gamma_1 = \gamma_2 = \gamma > 0$ satisfy*

$$\begin{aligned} \|F(x^{k+1})\|^2 \quad \leq \quad & \|F(x^k)\|^2 - \left(\frac{1}{2} - 2L^2\gamma^2\right)\|F(\widetilde{x}^k) - F(x^k)\|^2 \\ & - \left(\frac{1}{2} - \frac{\rho}{\gamma}\right)\|F(\widetilde{x}^k) - F(x^{k+1})\|^2 - \left(\frac{1}{2} - \frac{2\rho}{\gamma}\right)\|F(x^k) - F(x^{k+1})\|^2. \end{aligned} \tag{29}$$

If additionally $\gamma \leq {}^1\!/\!{}_{2L}$ and $\gamma \geq 4\rho$, then we have $\|F(x^{k+1})\| \leq \|F(x^k)\|$.

Proof. From L-Lipschitzness and ρ-negative comonotonicity of F we have

$$\begin{aligned} \|F(\widetilde{x}^k) - F(x^{k+1})\|^2 &\leq L^2\|\widetilde{x}^k - x^{k+1}\|^2, \\ \langle F(\widetilde{x}^k) - F(x^{k+1}), \widetilde{x}^k - x^{k+1}\rangle &\geq -\rho\|F(\widetilde{x}^k) - F(x^{k+1})\|^2, \\ \langle F(x^k) - F(x^{k+1}), x^k - x^{k+1}\rangle &\geq -\rho\|F(x^k) - F(x^{k+1})\|^2. \end{aligned}$$

Taking into account $\widetilde{x}^k - x^{k+1} = \gamma(F(\widetilde{x}^k) - F(x^k))$ and $x^k - x^{k+1} = \gamma F(\widetilde{x}^k)$, we get

$$\begin{aligned} \|F(\widetilde{x}^k) - F(x^{k+1})\|^2 &\leq L^2\gamma^2\|F(\widetilde{x}^k) - F(x^k)\|^2, \\ \gamma\langle F(\widetilde{x}^k) - F(x^{k+1}), F(\widetilde{x}^k) - F(x^k)\rangle &\geq -\rho\|F(\widetilde{x}^k) - F(x^{k+1})\|^2, \\ \gamma\langle F(x^k) - F(x^{k+1}), F(\widetilde{x}^k)\rangle &\geq -\rho\|F(x^k) - F(x^{k+1})\|^2. \end{aligned}$$

Next, we sum up the above inequalities with weights 2, ${}^1\!/\!{}_\gamma$, and ${}^2\!/\!{}_\gamma$ respectively:

$$\begin{aligned} 2\|F(\widetilde{x}^k) - F(x^{k+1})\|^2 - \frac{\rho}{\gamma}\|F(\widetilde{x}^k) - F(x^{k+1})\|^2 - \frac{2\rho}{\gamma}\|F(x^k) - F(x^{k+1})\|^2 \\ \leq 2L^2\gamma^2\|F(\widetilde{x}^k) - F(x^k)\|^2 + \langle F(\widetilde{x}^k) - F(x^{k+1}), F(\widetilde{x}^k) - F(x^k)\rangle \\ + 2\langle F(x^k), F(\widetilde{x}^k)\rangle - 2\langle F(x^{k+1}), F(\widetilde{x}^k)\rangle. \end{aligned}$$

To get rid of the inner products, we use $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$, which holds for any $a, b \in \mathbb{R}^d$. Using this, we continue our derivation as follows:

$$\begin{aligned} 2\|F(\widetilde{x}^k) - F(x^{k+1})\|^2 - \frac{\rho}{\gamma}\|F(\widetilde{x}^k) - F(x^{k+1})\|^2 - \frac{2\rho}{\gamma}\|F(x^k) - F(x^{k+1})\|^2 \\ \leq 2L^2\gamma^2\|F(\widetilde{x}^k) - F(x^k)\|^2 + \frac{1}{2}\|F(\widetilde{x}^k) - F(x^{k+1})\|^2 + \frac{1}{2}\|F(\widetilde{x}^k) - F(x^k)\|^2 \\ - \frac{1}{2}\|F(x^k) - F(x^{k+1})\|^2 + \|F(x^k)\|^2 + \|F(\widetilde{x}^k)\|^2 - \|F(\widetilde{x}^k) - F(x^k)\|^2 \\ - \|F(x^{k+1})\|^2 - \|F(\widetilde{x}^k)\|^2 + \|F(\widetilde{x}^k) - F(x^{k+1})\|^2. \end{aligned}$$

Rearranging the terms we get

$$\begin{aligned} \|F(x^{k+1})\|^2 \quad \leq \quad & \|F(x^k)\|^2 - \left(\frac{1}{2} - 2L^2\gamma^2\right)\|F(\widetilde{x}^k) - F(x^k)\|^2 \\ & - \left(\frac{1}{2} - \frac{\rho}{\gamma}\right)\|F(\widetilde{x}^k) - F(x^{k+1})\|^2 - \left(\frac{1}{2} - \frac{2\rho}{\gamma}\right)\|F(x^k) - F(x^{k+1})\|^2, \end{aligned}$$

which concludes the proof. □

Using this lemma we construct the potential-based proof of the last-iterate convergence of EG.

Theorem C.4 (Second part of Theorem 4.2)**.** *Let F be L-Lipschitz and ρ-negative comonotone. Then, for any $k \geq 0$ the iterates produced by* EG *with $\gamma_1 = \gamma_2 = \gamma$ such that $4\rho \leq \gamma \leq 1/2L$ satisfy*

$$\Phi_{k+1} \leq \Phi_k, \quad \text{where} \quad \Phi_k = \|x^k - x^*\|^2 + \left(k\gamma^2\left(1 - \frac{5\rho}{2\gamma} - L^2\gamma^2\right) + 40\gamma\rho\right)\|F(x^k)\|^2. \tag{30}$$

That is, under the introduced assumptions on γ and ρ for any $N \geq 1$ the iterates produced by EG *satisfy*

$$\|F(x^N)\|^2 \leq \frac{(1 + 40\gamma\rho L^2)\|x^0 - x^*\|^2}{N\gamma^2\left(1 - \frac{5\rho}{2\gamma} - L^2\gamma^2\right) + 40\gamma\rho}. \tag{31}$$

Proof. From (26) with $\gamma_1 = \gamma_2 = \gamma$ we have

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 + 2\gamma\rho\|F(\widetilde{x}^k)\|^2 - \gamma^2\left(1 - L^2\gamma^2\right)\|F(x^k)\|^2.$$

Next, taking into account that $4\rho \leq \gamma \leq 1/2L$, we also have from Lemma C.3 the following inequality:

$$\|F(x^{k+1})\|^2 \leq \|F(x^k)\|^2 - \left(\frac{1}{2} - \frac{\rho}{\gamma}\right)\|F(\widetilde{x}^k) - F(x^{k+1})\|^2. \tag{32}$$

Using these two inequalities, we derive the following upper bound on Φ_{k+1}:

$$\begin{aligned}
\Phi_{k+1} &= \|x^{k+1} - x^*\|^2 + \left((k+1)\gamma^2\left(1 - \frac{5\rho}{2\gamma} - L^2\gamma^2\right) + 40\gamma\rho\right)\|F(x^{k+1})\|^2 \\
&\leq \|x^k - x^*\|^2 + 2\gamma\rho\|F(\widetilde{x}^k)\|^2 - \gamma^2\left(1 - L^2\gamma^2\right)\|F(x^k)\|^2 \\
&\quad + \left((k+1)\gamma^2\left(1 - \frac{5\rho}{2\gamma} - L^2\gamma^2\right) + 40\gamma\rho\right)\left(\|F(x^k)\|^2 - \left(\frac{1}{2} - \frac{\rho}{\gamma}\right)\|F(\widetilde{x}^k) - F(x^{k+1})\|^2\right) \\
&= \Phi_k + 2\gamma\rho\|F(\widetilde{x}^k)\|^2 - \frac{5}{2}\gamma\rho\|F(x^k)\|^2 \\
&\quad - \left((k+1)\gamma^2\left(1 - \frac{5\rho}{2\gamma} - L^2\gamma^2\right) + 40\gamma\rho\right)\left(\frac{1}{2} - \frac{\rho}{\gamma}\right)\|F(\widetilde{x}^k) - F(x^{k+1})\|^2 \\
&\leq \Phi_k + 2\gamma\rho\|F(\widetilde{x}^k)\|^2 - \frac{5}{2}\gamma\rho\|F(x^k)\|^2 - 20\rho(\gamma - 2\rho)\|F(\widetilde{x}^k) - F(x^{k+1})\|^2.
\end{aligned}$$

Finally, we apply $\|a + b\|^2 \leq (1 + \beta)\|a\|^2 + (1 + \beta^{-1})\|b\|^2$, which holds $\forall a, b \in \mathbb{R}^d$, $\beta > 0$, with $\beta = 1/4$ to upper bound the second term in the right-hand side of the above inequality and continue our derivation as follows:

$$\begin{aligned}
\Phi_{k+1} &\leq \Phi_k + 2\gamma\rho\|F(x^{k+1}) + F(\widetilde{x}^k) - F(x^{k+1})\|^2 - \frac{5}{2}\gamma\rho\|F(x^k)\|^2 \\
&\quad -20\rho(\gamma - 2\rho)\|F(\widetilde{x}^k) - F(x^{k+1})\|^2 \\
&\leq \Phi_k + 2\gamma\rho\left(1 + \frac{1}{4}\right)\|F(x^{k+1})\|^2 + 2\gamma\rho\left(1 + 4\right)\|F(\widetilde{x}^k) - F(x^{k+1})\|^2 - \frac{5}{2}\gamma\rho\|F(x^k)\|^2 \\
&\quad -20\rho(\gamma - 2\rho)\|F(\widetilde{x}^k) - F(x^{k+1})\|^2 \\
&= \Phi_k + \frac{5}{2}\gamma\rho\|F(x^{k+1})\|^2 - \frac{5}{2}\gamma\rho\|F(x^k)\|^2 - 10\rho\left(\gamma - 4\rho\right)\|F(\widetilde{x}^k) - F(x^{k+1})\|^2.
\end{aligned}$$

Taking into account $\|F(x^{k+1})\|^2 \overset{(32)}{\leq} \|F(x^k)\|^2$ and $\gamma \geq 4\rho$, we get (30). Next, we unroll (30) and derive (31):

$$\begin{aligned}\|F(x^N)\|^2 &\leq \frac{1}{N\gamma^2\left(1-\frac{5\rho}{2\gamma}-L^2\gamma^2\right)+40\gamma\rho}\Phi_N\\ &\leq \frac{1}{N\gamma^2\left(1-\frac{5\rho}{2\gamma}-L^2\gamma^2\right)+40\gamma\rho}\Phi_{N-1} \leq \ldots \leq \frac{1}{N\gamma^2\left(1-\frac{5\rho}{2\gamma}-L^2\gamma^2\right)+40\gamma\rho}\Phi_0\\ &= \frac{\|x^0-x^*\|^2+40\gamma\rho\|F(x^0)\|^2}{N\gamma^2\left(1-\frac{5\rho}{2\gamma}-L^2\gamma^2\right)+40\gamma\rho}\\ &\overset{(3)}{\leq} \frac{(1+40\gamma\rho L^2)\|x^0-x^*\|^2}{N\gamma^2\left(1-\frac{5\rho}{2\gamma}-L^2\gamma^2\right)+40\gamma\rho},\end{aligned}$$

which concludes the proof of (31). Moreover, (11) follows from (31) since $4\rho \leq \gamma \leq 1/2L$ implies $1-(5\rho/2\gamma)-L^2\gamma^2 \geq 1/8$ and $1+40\gamma\rho L^2 \leq 7/2$. □

C.1.3. COUNTER-EXAMPLES

Theorem C.5 (Theorem 4.3)**.** *For any $L>0$, $\rho \geq 1/2L$, and any choice of stepsizes $\gamma_1,\gamma_2>0$ there exists ρ-negative comonotone L-Lipschitz operator F such that EG does not necessary converges on solving VIP with this operator F. In particular, for $\gamma_1 > 1/L$ it is sufficient to take $F(x)=Lx$, and for $0<\gamma_1\leq 1/L$ one can take $F(x)=LAx$, where $x\in\mathbb{R}^2$,*

$$A=\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix},\quad \theta=\frac{2\pi}{3}.$$

Proof. Assume that $L>0$ and $\rho\geq 1/2L$. We start with the case when $\gamma_1>1/L$. Consider operator $F(x)=Lx$. This operator is L-Lipschitz. Moreover, F is monotone and, as the result, it is ρ-negative comonotone for any $\rho\geq 0$. The iterates produced by EG with $x^0\neq 0$ satisfy

$$\widetilde{x}^k=(1-L\gamma_1)x^k,\quad x^{k+1}=x^k-L\gamma_2\widetilde{x}^k=(1-L\gamma_2+L^2\gamma_2\gamma_1)x^k$$

implying that

$$\|x^{k+1}-x^*\|=\|x^{k+1}\|=|1-L\gamma_2+L^2\gamma_2\gamma_1|\cdot\|x^k\|>\|x^k\|=\|x^k-x^*\|,$$

since $1-L\gamma_2+L^2\gamma_2\gamma_1>1-L\gamma_2+L\gamma_2=1$. That is, if $x^0\neq 0$, then EG diverges in this case.

Next, assume that $\gamma_1<1/L$ and consider $F(x)=LAx$, where $x\in\mathbb{R}^2$,

$$A=\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix},\quad \theta=\frac{2\pi}{3}.$$

Operator F is L-Lipschitz and $(1/2L)$-negative comonotone: for any $x,y\in\mathbb{R}^d$

$$\begin{aligned}\|F(x)-F(y)\| &= L\|A(x-y)\|=L\|x-y\|,\\ \langle F(x)-F(y),x-y\rangle &= \|F(x)-F(y)\|\cdot\|x-y\|\cdot\cos\theta\\ &= \|F(x)-F(y)\|\cdot\|A(x-y)\|\cdot\cos\frac{2\pi}{3}\\ &= -\frac{1}{2L}\|F(x)-F(y)\|^2\end{aligned}$$

where we use the fact that A is a rotation matrix. That is, $F(x)$ satisfies the conditions of the theorem. Taking into account that

$$A=\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}=\begin{pmatrix}-\frac{1}{2} & -\frac{\sqrt{3}}{2}\\ \frac{\sqrt{3}}{2} & \frac{1}{2}\end{pmatrix},\quad A^2=\begin{pmatrix}\cos(2\theta) & -\sin(2\theta)\\ \sin(2\theta) & \cos(2\theta)\end{pmatrix}=\begin{pmatrix}-\frac{1}{2} & \frac{\sqrt{3}}{2}\\ -\frac{\sqrt{3}}{2} & \frac{1}{2}\end{pmatrix},$$

we rewrite the update rule of EG as follows:

$$\begin{aligned} x^{k+1} &= x^k - \gamma_2 F\left(x^k - \gamma_2 F(x^k)\right) \\ &= x^k - \gamma_2 LA\left(x^k - \gamma_2 LAx^k\right) \\ &= \left(I - \gamma_2 LA + \gamma_1\gamma_2 L^2 A^2\right) x^k. \end{aligned}$$

To prove the divergence of EG, it remains to show that $\exists \lambda \in \mathrm{Sp}(I - \gamma_2 LA + \gamma_1\gamma_2 L^2 A^2)$ such that $|\lambda| > 1$. Indeed, we have

$$\begin{aligned} I - \gamma_2 LA + \gamma_1\gamma_2 L^2 A^2 &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \gamma_2 L \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix} + \gamma_1\gamma_2 L^2 \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix} \\ &= \begin{pmatrix} 1 + \frac{\gamma_2 L}{2}(1 - \gamma_1 L) & \frac{\sqrt{3}\gamma_2 L}{2}(1 + \gamma_1 L) \\ -\frac{\sqrt{3}\gamma_2 L}{2}(1 + \gamma_1 L) & 1 + \frac{\gamma_2 L}{2}(1 - \gamma_1 L). \end{pmatrix} \end{aligned}$$

The above matrix has eigenvalues $\lambda_{1,2} = 1 + \frac{\gamma_2 L}{2}(1 - \gamma_1 L) \pm i \cdot \frac{\sqrt{3}\gamma_2 L}{2}(1 + \gamma_1 L)$. Since $\gamma_2 > 0$ and $\gamma_1 \leq {}^1\!/_L$ we have that $|\lambda_{1,2}|^2 = \left(1 + \frac{\gamma_2 L}{2}(1 - \gamma_1 L)\right)^2 + \frac{3}{4}\gamma_2^2 L^2 (1 + \gamma_1 L)^2 > 1$. This concludes the proof. □

C.2. Optimistic gradient

C.2.1. Guarantees for the averaged squared norm of the operator

In this section, we proceed in an analogous way to the previous section on the Extragradient method. For notation convenience, we assume that $F(\widetilde{x}^{-1}) = 0$. Then, the update rule for OG can be written as

$$\begin{aligned} \widetilde{x}^k &= x^k - \gamma_1 F(\widetilde{x}^{k-1}), \\ x^{k+1} &= x^k - \gamma_2 F(\widetilde{x}^k), \end{aligned} \quad \forall k \geq 0. \tag{OG}$$

Lemma C.6. *Let F be ρ-star-negative comonotone. Then, for any $k \geq 0$ the iterates produced by OG after $k \geq 0$ iterations satisfy*

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &\leq \|x^k - x^*\|^2 - \gamma_2\left(\gamma_1 - 2\rho - \gamma_2\right)\|F(\widetilde{x}^k)\|^2 - \gamma_1\gamma_2\|F(\widetilde{x}^{k-1})\|^2 \\ &\quad + \gamma_1\gamma_2\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2. \end{aligned} \tag{33}$$

Proof. By the definition of x^{k+1} and $\widetilde{x}^k$ we have

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - 2\gamma_2\langle x^k - x^*, F(\widetilde{x}^k)\rangle + \gamma_2^2\|F(\widetilde{x}^k)\|^2 \\ &= \|x^k - x^*\|^2 - 2\gamma_2\langle \widetilde{x}^k - x^*, F(\widetilde{x}^k)\rangle - 2\gamma_1\gamma_2\langle F(\widetilde{x}^{k-1}), F(\widetilde{x}^k)\rangle + \gamma_2^2\|F(\widetilde{x}^k)\|^2. \end{aligned}$$

Next, we estimate the second term in the right-hand side using star-negative comonotonicity:

$$\|x^{k+1} - x^*\|^2 \overset{(2)}{\leq} \|x^k - x^*\|^2 + \gamma_2\left(2\rho + \gamma_2\right)\|F(\widetilde{x}^k)\|^2 - 2\gamma_1\gamma_2\langle F(\widetilde{x}^{k-1}), F(\widetilde{x}^k)\rangle.$$

Finally, we handle the last term in the right-hand side of the above inequality using $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$, which holds for any $a, b \in \mathbb{R}^d$:

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &\leq \|x^k - x^*\|^2 - \gamma_2\left(\gamma_1 - 2\rho - \gamma_2\right)\|F(\widetilde{x}^k)\|^2 - \gamma_1\gamma_2\|F(\widetilde{x}^{k-1})\|^2 \\ &\quad + \gamma_1\gamma_2\|F(\widetilde{x}^{k-1}) - F(\widetilde{x}^k)\|^2. \end{aligned}$$

□

This lemma is a building block of the first part of Theorem 4.4.

Theorem C.7 (First part of Theorem 4.4)**.** *Let F be L-Lipschitz and ρ-star-negative comonotone with $\rho < {}^1/_{2L}$. If $2\rho < \gamma_1 < {}^1/_L$ and $0 < \gamma_2 < \min\{{}^1/_L - \gamma_1, \gamma_1 - 2\rho\}$, then the iterates produced by OG after $N \geq 0$ iteration satisfy*

$$\frac{1}{N+1}\sum_{k=0}^{N}\|F(\widetilde{x}^k)\|^2 \leq \frac{\|x^0 - x^*\|^2}{\gamma_1\gamma_2(1 - L^2(\gamma_1+\gamma_2)^2)(N+1)}. \tag{34}$$

Proof. We first upper bound the last term that appeared in Lemma C.6 using L-Lipschitzness:

$$\begin{aligned}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2 &\leq L^2\|\widetilde{x}^k - \widetilde{x}^{k-1}\|^2\\ &= L^2\|(\widetilde{x}^k - x^k) + (x^k - x^{k-1}) + (x^{k-1} - \widetilde{x}^{k-1})\|^2\\ &= L^2\|(\gamma_1+\gamma_2)F(\widetilde{x}^{k-1}) - \gamma_1 F(\widetilde{x}^{k-2})\|^2\\ &= L^2(\gamma_1+\gamma_2)^2\|F(\widetilde{x}^{k-1})\|^2 + L^2\gamma_1^2\|F(\widetilde{x}^{k-2})\|^2\\ &\quad - 2L^2(\gamma_1+\gamma_2)\gamma_1\langle F(\widetilde{x}^{k-1}), F(\widetilde{x}^{k-2})\rangle.\end{aligned}$$

Decomposing the last term using $2\langle a,b\rangle = \|a\|^2 + \|a\|^2 - \|a-b\|^2$ yields

$$\begin{aligned}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2 &\leq L^2(\gamma_1+\gamma_2)\gamma_2\|F(\widetilde{x}^{k-1})\|^2 - L^2\gamma_1\gamma_2\|F(\widetilde{x}^{k-2})\|^2\\ &\quad + L^2\gamma_1(\gamma_1+\gamma_2)\|F(\widetilde{x}^{k-1}) - F(\widetilde{x}^{k-2})\|^2.\end{aligned} \tag{35}$$

The above recurrence holds for $k \geq 2$. For $k = 1$, we have $\widetilde{x}^0 = x^0$ and $\widetilde{x}^1 = x^1 - \gamma_1 F(\widetilde{x}^0) = x^0 - (\gamma_1+\gamma_2)F(x^0)$. Therefore,

$$\|F(\widetilde{x}^1) - F(\widetilde{x}^0)\|^2 \leq L^2(\gamma_1+\gamma_2)^2\|F(x^0)\|^2. \tag{36}$$

Combining (36) and (35), we obtain

$$\begin{aligned}\sum_{k=1}^{N}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2 &\leq \sum_{k=2}^{N}\left(L^2(\gamma_1+\gamma_2)\gamma_2\|F(\widetilde{x}^{k-1})\|^2 - L^2\gamma_1\gamma_2\|F(\widetilde{x}^{k-2})\|^2\right)\\ &\quad + \sum_{k=2}^{N}L^2\gamma_1(\gamma_1+\gamma_2)\|F(\widetilde{x}^{k-1}) - F(\widetilde{x}^{k-2})\|^2 + L^2(\gamma_1+\gamma_2)^2\|F(x^0)\|^2.\end{aligned}$$

We can simplify the terms on the right-hand side. Therefore,

$$\begin{aligned}\sum_{k=1}^{N}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2 &\leq L^2(\gamma_1+\gamma_2)\gamma_2\|F(\widetilde{x}^{N-1})\|^2 + \sum_{k=2}^{N}L^2\gamma_2^2\|F(\widetilde{x}^{k-1})\|^2\\ &\quad + \sum_{k=2}^{N}L^2\gamma_1(\gamma_1+\gamma_2)\|F(\widetilde{x}^{k-1}) - F(\widetilde{x}^{k-2})\|^2\\ &\quad + L^2(\gamma_1^2 + \gamma_1\gamma_2 + \gamma_2^2)\|F(x^0)\|^2\\ &\leq \sum_{k=2}^{N}L^2(\gamma_1+\gamma_2)\gamma_2\|F(\widetilde{x}^{k-1})\|^2\\ &\quad + \sum_{k=1}^{N}L^2\gamma_1(\gamma_1+\gamma_2)\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2\\ &\quad + L^2(\gamma_1^2 + \gamma_1\gamma_2 + \gamma_2^2)\|F(x^0)\|^2.\end{aligned}$$

Using the above equation, we can bound $\sum_{k=1}^{N}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2$, i.e.,

$$\begin{aligned}(1 - L^2\gamma_1(\gamma_1+\gamma_2))\sum_{k=1}^{N}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2 &\leq \sum_{k=1}^{N}L^2(\gamma_1+\gamma_2)\gamma_2\|F(\widetilde{x}^{k-1})\|^2\\ &\quad + L^2\gamma_1^2\|F(x^0)\|^2.\end{aligned} \tag{37}$$

Let us now apply (33) recursively ($F(\widetilde{x}^{-1}) = 0$ for simpler notation), which leads to

$$\begin{aligned}\|x^N - x^*\|^2 &\leq \|x^0 - x^*\|^2 - \sum_{k=0}^{N-1}\big(\gamma_2\left(\gamma_1 - 2\rho - \gamma_2\right)\|F(\widetilde{x}^k)\|^2 + \gamma_1\gamma_2\|F(\widetilde{x}^{k-1})\|^2\big)\\ &\quad + \sum_{k=0}^{N-1}\gamma_1\gamma_2\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2)\\ &\leq \|x^0 - x^*\|^2 - \sum_{k=0}^{N-2}\gamma_2\left(2\gamma_1 - 2\rho - \gamma_2\right)\|F(\widetilde{x}^k)\|^2 + \sum_{k=1}^{N-1}\gamma_1\gamma_2\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2\\ &\quad + \gamma_1\gamma_2\|F(x^0)\|^2.\end{aligned}$$

Plugging (37) to the above with $1 - L^2\gamma_1(\gamma_1 + \gamma_2) > 0$, which follows from $\gamma_1 < \nicefrac{1}{L}$ and $\gamma_2 < \nicefrac{1}{L} - \gamma_1$, leads to

$$\begin{aligned}\|x^N - x^*\|^2 &\leq \|x^0 - x^*\|^2 - \sum_{k=0}^{N-2}\gamma_2\left(2\gamma_1 - 2\rho - \gamma_2\right)\|F(\widetilde{x}^k)\|^2 + \gamma_1\gamma_2\|F(x^0)\|^2\\ &\quad + \frac{\gamma_1\gamma_2^2(\gamma_1+\gamma_2)L^2}{1 - L^2\gamma_1(\gamma_1+\gamma_2)}\sum_{k=0}^{N-2}\|F(\widetilde{x}^k)\|^2 + \frac{\gamma_1^3\gamma_2L^2}{1 - L^2\gamma_1(\gamma_1+\gamma_2)}\|F(x^0)\|^2.\end{aligned}$$

Since $\gamma_1 - 2\rho - \gamma_2 > 0$, we have

$$\begin{aligned}\|x^N - x^*\|^2 &\leq \|x^0 - x^*\|^2 - \gamma_1\gamma_2\sum_{k=0}^{N-2}\|F(\widetilde{x}^k)\|^2 + \gamma_1\gamma_2\|F(x^0)\|^2\\ &\quad + \frac{\gamma_1\gamma_2^2(\gamma_1+\gamma_2)L^2}{1 - L^2\gamma_1(\gamma_1+\gamma_2)}\sum_{k=0}^{N-2}\|F(\widetilde{x}^k)\|^2 + \frac{\gamma_1^3\gamma_2L^2}{1 - L^2\gamma_1(\gamma_1+\gamma_2)}\|F(x^0)\|^2.\end{aligned}$$

Rearranging terms and applying $\|x^N - x^*\|^2 \geq 0$ plus $\|F(x^0)\|^2 \leq L^2\|x^0 - x^*\|^2$, we obtain

$$\begin{aligned}\gamma_1\gamma_2(1 - L^2(\gamma_1+\gamma_2)^2)\sum_{k=0}^{N-2}\|F(\widetilde{x}^k)\|^2 &\leq (1 - L^2\gamma_1(\gamma_1+\gamma_2))\|x^0 - x^*\|^2\\ &\quad + \gamma_1\gamma_2L^2(1 - L^2\gamma_1\gamma_2)\|x^0 - x^*\|^2\\ &\leq (1 - L^2\gamma_1^2)\|x^0 - x^*\|^2 \leq \|x^0 - x^*\|^2.\end{aligned}$$

The above inequality implies (34), since one can replace N with $N+2$. □

C.2.2. Last-iterate guarantees

Our proof is based on the following lemma.

Lemma C.8. *Let F be L-Lipschitz and ρ-negative comonotone such that $\rho \leq \nicefrac{5}{62L}$. Then for any $k \geq 1$ the iterates produced by* EG *with $\gamma_1 = \gamma_2 = \gamma > 0$ such that $4\rho \leq \gamma \leq \nicefrac{10}{31L}$ satisfy*

$$\begin{aligned}\|F(x^{k+1})\|^2 + \|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 &\leq \|F(x^k)\|^2 + \|F(x^k) - F(\widetilde{x}^{k-1})\|^2\\ &\quad - \frac{1}{100}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2. \end{aligned} \tag{38}$$

Proof. From ρ-negative comonotonicity and L-Lipschitzness of F we have

$$\begin{aligned}-\rho\|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 &\leq \langle F(x^{k+1}) - F(\widetilde{x}^k), x^{k+1} - \widetilde{x}^k\rangle,\\ -\rho\|F(x^k) - F(x^{k+1})\|^2 &\leq \langle F(x^k) - F(x^{k+1}), x^k - x^{k+1}\rangle,\\ \|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 &\leq L^2\|x^{k+1} - \widetilde{x}^k\|^2.\end{aligned}$$

Using $x^{k+1} - \widetilde{x}^k = \gamma(F(\widetilde{x}^{k-1}) - F(\widetilde{x}^k))$ and $x^{k+1} - x^k = -\gamma F(\widetilde{x}^k)$, we rewrite the above inequalities as

$$-\frac{\rho}{\gamma}\|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 \leq \langle F(x^{k+1}) - F(\widetilde{x}^k), F(\widetilde{x}^{k-1}) - F(\widetilde{x}^k)\rangle, \tag{39}$$

$$-\frac{\rho}{\gamma}\|F(x^k) - F(x^{k+1})\|^2 \leq \langle F(x^k) - F(x^{k+1}), F(\widetilde{x}^k)\rangle, \tag{40}$$

$$\|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 \leq L^2\gamma^2\|F(\widetilde{x}^{k-1}) - F(\widetilde{x}^k)\|^2. \tag{41}$$

Next, we apply $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$, which holds for any $a, b \in \mathbb{R}^d$, and from the first two inequalities get

$$\begin{aligned}
-\frac{2\rho}{3\gamma}\|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 &\overset{(39)}{\leq} \frac{1}{3}\|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 + \frac{1}{3}\|F(\widetilde{x}^{k-1}) - F(\widetilde{x}^k)\|^2 \\
&\quad -\frac{1}{3}\|F(x^{k+1}) - F(\widetilde{x}^{k-1})\|^2, \qquad (42)
\end{aligned}$$

$$\begin{aligned}
-\frac{2\rho}{\gamma}\|F(x^k) - F(x^{k+1})\|^2 &\overset{(40)}{\leq} 2\langle F(x^k), F(\widetilde{x}^k)\rangle - 2\langle F(x^{k+1}), F(\widetilde{x}^k)\rangle \\
&= \|F(x^k)\|^2 + \|F(\widetilde{x}^k)\|^2 - \|F(x^k) - F(\widetilde{x}^k)\|^2 \\
&\quad -\|F(x^{k+1})\|^2 - \|F(\widetilde{x}^k)\|^2 + \|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 \\
&= \|F(x^k)\|^2 + \|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 - \|F(x^{k+1})\|^2 \\
&\quad -\|F(x^k) - F(\widetilde{x}^k)\|^2. \qquad (43)
\end{aligned}$$

Summing up (41), (42), and (43) with weights 3, 1, and 1 respectively, we derive

$$\begin{aligned}
\left(3 - \frac{2\rho}{3\gamma}\right)\|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 - \frac{2\rho}{\gamma}\|F(x^k) - F(x^{k+1})\|^2 &\leq 3L^2\gamma^2\|F(\widetilde{x}^{k-1}) - F(\widetilde{x}^k)\|^2 \\
&\quad +\frac{1}{3}\|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 \\
&\quad +\frac{1}{3}\|F(\widetilde{x}^{k-1}) - F(\widetilde{x}^k)\|^2 \\
&\quad -\frac{1}{3}\|F(x^{k+1}) - F(\widetilde{x}^{k-1})\|^2 \\
&\quad +\|F(x^k)\|^2 + \|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 \\
&\quad -\|F(x^{k+1})\|^2 - \|F(x^k) - F(\widetilde{x}^k)\|^2 \\
&= \left(\frac{1}{3} + 3L^2\gamma^2\right)\|F(\widetilde{x}^{k-1}) - F(\widetilde{x}^k)\|^2 \\
&\quad +\frac{4}{3}\|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 \\
&\quad -\frac{1}{3}\|F(x^{k+1}) - F(\widetilde{x}^{k-1})\|^2 \\
&\quad +\|F(x^k)\|^2 - \|F(x^{k+1})\|^2 \\
&\quad -\|F(x^k) - F(\widetilde{x}^k)\|^2.
\end{aligned}$$

To simplify further derivations, we introduce new notation: $\Psi_k = \|F(x^k)\|^2 + \|F(x^k) - F(\widetilde{x}^{k-1})\|^2, \forall k \geq 1$. Rearranging the terms in the above inequality and using the new notation, we arrive at

$$\begin{aligned}
\Psi_{k+1} - \Psi_k &\leq T_k, \quad \text{where} \\
T_k &\overset{\text{def}}{=} \frac{2\rho}{\gamma}\|F(x^k) - F(x^{k+1})\|^2 + \left(\frac{1}{3} + 3L^2\gamma^2\right)\|F(\widetilde{x}^{k-1}) - F(\widetilde{x}^k)\|^2 \\
&\quad -\frac{2}{3}\left(1 - \frac{\rho}{\gamma}\right)\|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 - \|F(x^k) - F(\widetilde{x}^{k-1})\|^2 \\
&\quad -\frac{1}{3}\|F(x^{k+1}) - F(\widetilde{x}^{k-1})\|^2 - \|F(x^k) - F(\widetilde{x}^k)\|^2.
\end{aligned}$$

To prove (38), it remains to show that $T_k \leq -\frac{1}{100}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2$ for all $k \geq 1$. Taking into account $4\rho \leq \gamma \leq {}^{10}\!/_{31L}$, we upper bound T_k as follows:

$$\begin{aligned} T_k &\leq \frac{1}{2}\|F(x^k) - F(x^{k+1})\|^2 + \frac{121}{93}\|F(\widetilde{x}^{k-1}) - F(\widetilde{x}^k)\|^2 - \frac{1}{2}\|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 \\ &\quad -\|F(x^k) - F(\widetilde{x}^{k-1})\|^2 - \frac{1}{3}\|F(x^{k+1}) - F(\widetilde{x}^{k-1})\|^2 - \|F(x^k) - F(\widetilde{x}^k)\|^2 \\ &= \begin{pmatrix} F(x^{k+1}) \\ F(x^k) \\ F(\widetilde{x}^k) \\ F(\widetilde{x}^{k-1}) \end{pmatrix}^\top \left(\begin{pmatrix} -\frac{1}{3} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{3} \\ -\frac{1}{2} & -\frac{3}{2} & 1 & 1 \\ \frac{1}{2} & 1 & -\frac{4927}{5766} & -\frac{1861}{2883} \\ \frac{1}{3} & 1 & -\frac{1861}{2883} & -\frac{661}{961} \end{pmatrix} \otimes I_d \right) \begin{pmatrix} F(x^{k+1}) \\ F(x^k) \\ F(\widetilde{x}^k) \\ F(\widetilde{x}^{k-1}) \end{pmatrix}, \end{aligned} \tag{44}$$

where I_d is d-dimensional identity matrix and $A \otimes B$ denotes the Kronecker product of two matrices A and B. One can show numerically (see our codes) that

$$\begin{pmatrix} -\frac{1}{3} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{3} \\ -\frac{1}{2} & -\frac{3}{2} & 1 & 1 \\ \frac{1}{2} & 1 & -\frac{4927}{5766} & -\frac{1861}{2883} \\ \frac{1}{3} & 1 & -\frac{1861}{2883} & -\frac{661}{961} \end{pmatrix} \preccurlyeq -\frac{1}{100} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}.$$

Therefore, in view of (44), we have

$$\begin{aligned} T_K &\leq -\frac{1}{100} \begin{pmatrix} F(x^{k+1}) \\ F(x^k) \\ F(\widetilde{x}^k) \\ F(\widetilde{x}^{k-1}) \end{pmatrix}^\top \left(\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix} \otimes I_d \right) \begin{pmatrix} F(x^{k+1}) \\ F(x^k) \\ F(\widetilde{x}^k) \\ F(\widetilde{x}^{k-1}) \end{pmatrix} \\ &= -\frac{1}{100}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2, \end{aligned}$$

which concludes the proof. □

We emphasize that similar potential $\|F(x^k)\|^2 + \|F(x^k) - F(\widetilde{x}^{k-1})\|^2$ is used in Cai et al. (2022b) to prove the last-iterate convergence of OG for *monotone* L-Lipschitz F. However, our proof non-trivially differs from the one from Cai et al. (2022b): we use ρ-negative comonotonicity for pairs $(x^{k+1}, \widetilde{x}^k)$, (x^{k+1}, x^k) and L-Lipschitzness for $(x^{k+1}, \widetilde{x}^k)$, while Cai et al. (2022b) use *monotonicity* (x^{k+1}, x^k) and L-Lipschitzness for $(x^{k+1}, \widetilde{x}^k)$. Next, the only known last-iterate $\mathcal{O}({}^1\!/_N)$ convergence rate for OG under ρ-negative comonotonicity and L-Lipschitzness (Luo & Tran-Dinh, 2022) is based on a different potential $\|F(x^{k+1})\|^2 + \frac{2\gamma - 3\rho}{2\gamma}\|F(x^{k+1}) - F(\widetilde{x}^k)\|^2$, which coincides with the one used by Gorbunov et al. (2022c) in the monotone case. Moreover, the proof from Luo & Tran-Dinh (2022) is based on 2 inequalities: ρ-negative comonotonicity for pairs (x^{k+1}, x^k) and L-Lipschitzness for $(x^{k+1}, \widetilde{x}^k)$. In contrast, our proof is based on 3 inequalities, i.e., it uses more information about the problem. This might be the reason, why our proof allows ρ to be larger than in the proof by Luo & Tran-Dinh (2022).

Using Lemma C.8, we can proceed with the potential-based proof of the last-iterate convergence of OG.

Theorem C.9 (Second part of Theorem 4.4)**.** *Let F be L-Lipschitz and ρ-negative comonotone. Then, for any $k \geq 0$ the iterates produced by* OG *with $\gamma_1 = \gamma_2 = \gamma$ such that $4\rho \leq \gamma \leq {}^{10}\!/_{31L}$ satisfy*

$$\Phi_{k+1} \leq \Phi_k, \quad \textit{where} \quad \Phi_k = \|x^k - x^*\|^2 + \left(k\frac{\gamma(\gamma - 3\rho)}{2 + 6L^2\gamma^2} + 400\gamma^2\right)\Psi_k, \tag{45}$$

where $\Psi_k = \|F(x^k)\|^2 + \|F(x^k) - F(\widetilde{x}^{k-1})\|^2$. That is, under the introduced assumptions on γ and ρ for any $N \geq 1$ the iterates produced by OG *satisfy*

$$\|F(x^N)\|^2 \leq \frac{717\|x^0 - x^*\|^2}{N\gamma(\gamma - 3\rho) + 800\gamma^2}. \tag{46}$$

Proof. From (33) with $\gamma_1 = \gamma_2 = \gamma$ we have

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 + 2\rho\gamma\|F(\widetilde{x}^k)\|^2 - \gamma^2\|F(\widetilde{x}^{k-1})\|^2 + \gamma^2\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2.$$

Next, taking into account that $4\rho \leq \gamma \leq {}^{10}\!/_{31L}$, we also have from Lemma C.8 the following inequality:

$$\|F(x^{k+1})\|^2 + \|F(x^{k+1}) - F(\widetilde{x}^k)\|^2 \leq \|F(x^k)\|^2 + \|F(x^k) - F(\widetilde{x}^{k-1})\|^2 - \frac{1}{100}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2.$$

Using these two inequalities, we derive the following upper bound on Φ_{k+1}:

$$\begin{aligned}
\Phi_{k+1} &= \|x^{k+1} - x^*\|^2 + \left((k+1)\frac{\gamma(\gamma - 3\rho)}{2 + 6L^2\gamma^2} + 400\gamma^2\right)\Psi_{k+1} \\
&\leq \|x^k - x^*\|^2 + 2\rho\gamma\|F(\widetilde{x}^k)\|^2 - \gamma^2\|F(\widetilde{x}^{k-1})\|^2 + \gamma^2\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2 \\
&\quad + \left((k+1)\frac{\gamma(\gamma - 3\rho)}{2 + 6L^2\gamma^2} + 400\gamma^2\right)\left(\|F(x^k)\|^2 + \|F(x^k) - F(\widetilde{x}^{k-1})\|^2 - \frac{1}{100}\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2\right) \\
&\leq \Phi_k + 2\rho\gamma\|F(\widetilde{x}^k)\|^2 - \gamma^2\|F(\widetilde{x}^{k-1})\|^2 + \gamma^2\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2 \\
&\quad + \frac{\gamma(\gamma - 3\rho)}{2 + 6L^2\gamma^2}\left(\|F(x^k)\|^2 + \|F(x^k) - F(\widetilde{x}^{k-1})\|^2\right) - 4\gamma^2\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2 \\
&= \Phi_k + 2\rho\gamma\|F(\widetilde{x}^k)\|^2 - \gamma^2\|F(\widetilde{x}^{k-1})\|^2 \\
&\quad + \frac{\gamma(\gamma - 3\rho)}{2 + 6L^2\gamma^2}\left(\|F(x^k)\|^2 + \|F(x^k) - F(\widetilde{x}^{k-1})\|^2\right) - 3\gamma^2\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2.
\end{aligned}$$

In the next step, we use following upper bounds based on L-Lipschitzness, (OG), and $\|a + b\|^2 \leq (1 + \beta)\|a\|^2 + (1 + \beta^{-1})\|b\|^2$, which holds $\forall a, b \in \mathbb{R}^d$, $\beta > 0$:

$$\begin{aligned}
\|F(\widetilde{x}^k)\|^2 &\leq 3\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2 + {}^{3}\!/_{2}\|F(\widetilde{x}^{k-1})\|^2, \\
\|F(x^k) - F(\widetilde{x}^{k-1})\|^2 &\leq 2\|F(x^k) - F(\widetilde{x}^k)\|^2 + 2\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2 \\
&\leq 2L^2\gamma^2\|F(\widetilde{x}^{k-1})\|^2 + 2\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2, \\
\|F(x^k)\|^2 &\leq 2\|F(x^k) - F(\widetilde{x}^{k-1})\|^2 + 2\|F(\widetilde{x}^{k-1})\|^2 \\
&\leq 2(1 + 2L^2\gamma^2)\|F(\widetilde{x}^{k-1})\|^2 + 4\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2.
\end{aligned}$$

Applying the above yields

$$\begin{aligned}
\Phi_{k+1} &\leq \Phi_k + \left(\frac{\gamma(\gamma - 3\rho)}{2 + 6L^2\gamma^2}(2 + 6L^2\gamma^2) + 3\rho\gamma - \gamma^2\right)\|F(\widetilde{x}^{k-1})\|^2 \\
&\quad + \left(6\rho\gamma + \frac{3\gamma(\gamma - 3\rho)}{1 + 3L^2\gamma^2} - 3\gamma^2\right)\|F(\widetilde{x}^k) - F(\widetilde{x}^{k-1})\|^2 \\
&\leq \Phi_k,
\end{aligned}$$

where we use $1 + 3L^2\gamma^2 \geq 1$. This applies that $\Phi_N \leq \Phi_1$, therefore

$$\|F(x^N)\|^2 \leq \frac{\|x^1 - x^*\|^2 + \left(\frac{\gamma(\gamma-3\rho)}{2+6L^2\gamma^2} + 400\gamma^2\right)\left(\|F(x^1)\|^2 + \|F(x^1) - F(x^0)\|^2\right)}{N\frac{\gamma(\gamma-3\rho)}{2+6L^2\gamma^2} + 400\gamma^2}$$

In the next step, we bound everything with respect to $\|x^0 - x^*\|^2$ using ρ-negative comonotonicity, L-Lipschitzness, (OG), and $\|a + b\|^2 \leq (1 + \beta)\|a\|^2 + (1 + \beta^{-1})\|b\|^2$:

$$\begin{aligned}
\|x^1 - x^*\|^2 &\overset{(33)}{\leq} \|x^0 - x^*\|^2 + \gamma(2\rho + \gamma)\|F(x^0)\|^2 \\
&\leq (1 + L^2\gamma(2\rho + \gamma))\|x^0 - x^*\|^2 \leq 2\|x^0 - x^*\|^2, \\
\|F(x^1)\|^2 + \|F(x^1) - F(x^0)\|^2 &\leq 3\|F(x^1)\|^2 + 2\|F(x^0)\|^2 \\
&\leq L^2(3\|x^1 - x^*\|^2 + 2\|x^0 - x^*\|^2) \leq 8L^2\|x^0 - x^*\|^2, \\
2 \leq 2 + 6L^2\gamma^2 &\leq 3.
\end{aligned}$$

Putting all together yields

$$\|F(x^N)\|^2 \leq \frac{(2 + 401L^2 \cdot 8\gamma^2)\|x^0 - x^*\|^2}{N\frac{\gamma(\gamma-3\rho)}{2} + 400\gamma^2} \leq \frac{717\|x^0 - x^*\|^2}{N\gamma(\gamma - 3\rho) + 800\gamma^2},$$

which concludes the proof. □

C.2.3. COUNTER-EXAMPLES

Theorem C.10 (Theorem 4.5)**.** *For any $L > 0$, $\rho \geq 1/2L$, and any choice of stepsizes $\gamma_1, \gamma_2 > 0$ there exists ρ-negative comonotone L-Lipschitz operator F such that OG does not necessary converges on solving VIP with this operator F. In particular, for $\gamma_1 > 1/L$ it is sufficient to take $F(x) = Lx$, where $x \in \mathbb{R}$, and for $0 < \gamma_1 \leq 1/L$ one can take $F(x) = LAx$, where $x \in \mathbb{R}^2$,*

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad \theta = \frac{2\pi}{3}.$$

Proof. Assume that $L > 0$ and $\rho \geq 1/2L$. We start with the case when $\gamma_1 > 1/L$. Consider operator $F(x) = Lx$. This operator is L-Lipschitz. Moreover, F is monotone and, as the result, it is ρ-negative comonotone for any $\rho \geq 0$. The iterates produced by OG with $x^0 \neq 0$ satisfy

$$\widetilde{x}^k = x^k - L\gamma_1\widetilde{x}^{k-1}, \quad x^{k+1} = x^k - L\gamma_2\widetilde{x}^k = (1 - L\gamma_2)x^k + L^2\gamma_2\gamma_1\widetilde{x}^{k-1}$$

implying that

$$\begin{bmatrix} x^{k+1} \\ \widetilde{x}^k \end{bmatrix} = \begin{pmatrix} 1 - \gamma_2 L & \gamma_1\gamma_2 L^2 \\ 1 & -\gamma_1 L \end{pmatrix} \begin{bmatrix} x^k \\ \widetilde{x}^{k-1} \end{bmatrix}$$

The eigenvalues of the above 2×2 matrix can be computed analytically. One of them has the form

$$-\frac{L\gamma_1 + L\gamma_2 + \sqrt{L^2\gamma_1^2 + L^2\gamma_2^2 + 2L^2\gamma_1\gamma_2 + 2L\gamma_1 - 2L\gamma_2 + 1} - 1}{2}$$
$$< -\frac{1 + \sqrt{1 + 2L\gamma_2 + 2 - 2L\gamma_2 + 1} - 1}{2} = -1.$$

The derivation above is verified symbolically in our codes. That means we can select such starting setup x^0, $\widetilde{x}^0$, such that OG diverges.

Next, assume that $\gamma_1 \leq 1/L$ and consider $F(x) = LAx$, where $x \in \mathbb{R}^2$,

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad \theta = \frac{2\pi}{3}.$$

Operator F is L-Lipschitz and $(1/2L)$-negative comonotone: for any $x, y \in \mathbb{R}^d$

$$\begin{aligned} \|F(x) - F(y)\| &= L\|A(x-y)\| = L\|x - y\|, \\ \langle F(x) - F(y), x - y\rangle &= \|F(x) - F(y)\| \cdot \|x - y\| \cdot \cos\theta \\ &= \|F(x) - F(y)\| \cdot \|A(x-y)\| \cdot \cos\frac{2\pi}{3} \\ &= -\frac{1}{2L}\|F(x) - F(y)\|^2 \end{aligned}$$

where we use the fact that A is a rotation matrix. That is, $F(x)$ satisfies the conditions of the theorem. Taking into account that

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}, \quad A^2 = \begin{pmatrix} \cos(2\theta) & -\sin(2\theta) \\ \sin(2\theta) & \cos(2\theta) \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix},$$

we rewrite the update rule of OG similarly to the case $\gamma_1 > 1/L$:

$$\begin{bmatrix} x^{k+1} \\ \widetilde{x}^k \end{bmatrix} = \begin{pmatrix} I - \gamma_2 LA & \gamma_1\gamma_2 L^2 A^2 \\ I & -\gamma_1 LA \end{pmatrix} \begin{bmatrix} x^k \\ \widetilde{x}^{k-1} \end{bmatrix} = B \begin{bmatrix} x^k \\ \widetilde{x}^{k-1} \end{bmatrix}.$$

To prove the divergence of OG, we show that B is expansive, i.e., its spectral norm $\|B\| > 1$, therefore, there exists a starting point for which OG does not converge. The spectral norm of B has the following form

$$\|B\|^2 = c + \sqrt{c^2 - L^2\gamma_1^2},$$

where $c = \frac{L^4\gamma_1^2\gamma_2^2+L^2\gamma_1^2+L^2\gamma_2^2+L\gamma_2}{2} + 1$ (this derivation is verified symbolically in our codes). Therefore, $\|B\|$ is well defined since $c > 1$ and $L^2\gamma_1^2 \leq 1$, and, moreover, $\|B\| > 1$, which concludes the proof. □