Extra-Gradient and Optimistic Gradient Descent Converge in Iterates Faster than $\mathcal{O}(1/\sqrt{T})$ in All Monotone Lipschitz Variational Inequalities

Kimon Antonakopoulos
LIONS-EPFL
kimon.antonakopoulos@epfl.ch

Gabriele Farina
MIT
gfarina@mit.edu

Volkan Cevher
LIONS-EPFL
volkan.cevher@epfl.ch

Abstract

In this work, we show that the extra-gradient and optimistic gradient descent-ascent methods, arguably the de facto algorithms for solving variational inequalities (VIs) and saddle-point optimization problems, converge in iterates at a rate strictly faster than the established lower bound of $\Omega(1/\sqrt{T})$ in all monotone smooth variational inequalities. These rates apply to both constant and line-search-free, AdaGrad-type step sizes, where no decrease property is generally guaranteed. We believe these results are especially noteworthy in light of recent lower bounds which established that, for any *fixed* time horizon T, an adversary could construct a monotone Lipschitz VI in which the the iterate at time T is bounded $\Omega(1/\sqrt{T})$ away from the solution. Our results imply that, even on such adversarially constructed examples, the extra-gradient type method's slow convergence is only transitory.

1. Introduction

Variational inequalities (VI) are a general and flexible framework for "optimization beyond minimization". They capture large classes of optimization problems, including saddle point computation, equilibrium finding in multiplayer games, fixed-point computation, and complementarity problems [16, 46]. These classes of problems arise naturally in machine learning, for example in GANs [23], robust reinforcement learning [41], and other adversarial models. The case of *monotone* VIs is especially important, since these can be solved in polynomial time while capturing a number of useful applications, including computation of Nash equilibria in two-player zero-sum games, and minimization of *convex* functions. Positive complexity results for monotone VIs were first given by [27] constructively, who showed that the *extra-gradient* (EG) method—a two-step variant of projected gradient descent—produces iterates whose average converges to a solution. The demand of solving more complex problems steming from online convex optimization, and convex/ concave min-max problem triggered a wide range of literature like [50], [45], [13] and references therein.

On the other hand, the Extra-gradient and their optimistic variants have a long history in the field of optimization. The Extra-gradient, is known to achieve an optimal rate of order $\mathcal{O}(1/T)$ in monotone VIs. This method has been further extended in [35, 38] by introducing Mirror-prox and its primal-dual counterpart Dual-extrapolation. However, all these methods require two oracle calls per iteration (one for the extrapolation and one for the update step) which makes them more expensive than the standard Forward/Backward methods. The first issue to address this issue was Popov's modified Arrow–Hurwicz algorithm [42]. To that end, several extensions have been proposed such as Past Extra-Gradient (PEG) of [10, 19], Reflected Gradient (RG) of [9, 11, 29], Optimistic Gradient (OG) of [12, 33, 34, 40] and Golden Ratio method of [30]. Since then, several generalizations and improvements have been proposed over the EG method, and several authors have set out to strengthen EG analysis beyond the case of average iterates. Two prior results are especially important in the context of our paper.

- Tseng [51] showed that, under certain conditions on the structure of the VI (or being more precise a favorable geometry of the problems domain), the extra-gradient method guarantees asymptotic convergence to a solution of the VI at a speed of convergence *at least* exponentially fast with respect to the number of iterations of the method. This result applies in particular in the case of VIs with affine operators and polyhedral feasible sets, as is the case in normal-form and extensive-form games.

- A recent paper by Golowich, Pattathil, Daskalakis, and Ozdaglar [22] showed that, among the class of L-Lipschitz VIs with solution norm bounded by D, for any number of iterations T an adversary could always construct a VI in the class such that the approximation of the solution obtained by EG (or, in fact, any method from 1,2-SCLI[3] after T iterations is never better than $\Omega(LD/\sqrt{T})$.

Contributions In seeming defiance to the paper by Golowich et al. [22] discussed above, in this paper we show the following result:

In all monotone Lipshchitz VIs, EG and OGDA converge (in iterates) strictly faster than $\mathcal{O}(1/\sqrt{T})$.

We remark that, unlike the result of Tseng [51]'s, our results apply to *all* monotone Lipschitz VIs under any arbitrary convex domains, although such generality comes at the cost of the linear rate shown in that paper. We suspect that such a gap can hardly be overcome, given the generality of our setting. One might wonder why our result is not precluded by the recent lower bound of Golowich et al. [22]. The resolution is in the following observation: our results says that, for any VI, there exists a time instant T (dependent on the specific VI) after which EG exhibits asymptotic convergence faster than $\mathcal{O}(1/\sqrt{T})$. Golowich et al. [22]'s result, on the other hand, says that, for any T, there exists a VI (dependent on the specific T) such that EG exhibits slow convergence up to time T. Our results imply that, even on such adversarially-constructed examples, extra-gradient's (and OGDA) slow convergence can only be transitory. This results completes our understanding of these two classical methods, namely the (EG)/(OGDA) algorithms regarding their last iterate performance. Furthermore, we provide a novel framework which additionally allows us to tackle "non-decreasing" methods with controllable error.

In a nutshell, our contributions can be summarized as follows.

1. We provide a refinement of the existing work on last iterate convergence rates and provide a strictly faster speed of convergence of order $o(1/\sqrt{T})$ for (EG)/(OGDA).
2. We provide a last iterate convergence rate of the same order for a family of one fly adaptive step size family in the sense of [15]. As far as our knowledge goes this is the first result that provides the last iterate rate for these types of adaptive step sizes. More importantly, we establish a faster than the $1/\sqrt{T}$ by employing a common adaptive mechanism for both methods.
3. We provide some initial numerical evidence of our theoretical results on an actual Poker game in the Appendix.

2. Problem Setup and Notation

In this preliminary section, we shall illustrate the generic variational inequality problem formulation along with its interplay with popular "convex-structured" problems as its special cases. Moreover, we introduce the main performance criterion for the case of (VI), in view of the lack of a particular potential as in the case of the simple minimization framework.

2.1. The Variational Inequality Problem

Our main objective is to solve the following variational inequality given by the following abstract formulation:

$$\text{Find } \boldsymbol{x}^* \in \mathcal{X} \text{ such that } \langle \mathcal{A}(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0 \text{ for all } \boldsymbol{x} \in \mathcal{X} \tag{VI}$$

where $\mathcal{X}$ is a convex and closed (but not necessarily bounded) subset of $\mathbb{R}^d$. For the case where $\mathcal{X} = \mathbb{R}^d$, the (VI) problem boils down to the classical problem of finding a zero-*equilibrium point*-of $\mathcal{A}$, *i.e.*:

$$\text{Find } \boldsymbol{x}^* \in \mathbb{R}^d \text{ such that } \mathcal{A}(\boldsymbol{x}^*) = \mathbf{0} \tag{Zer}$$

Moreover, in terms of structural and regularity conditions concerning the defining operator $\mathcal{A} : \mathcal{X} \to \mathbb{R}^d$, we make the following blanket assumptions:

- *Monotonicity*: $\langle \mathcal{A}(\boldsymbol{x}) - \mathcal{A}(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}' \rangle \geq 0$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$.
- *Lipschitz continuity*: $\|\mathcal{A}(\boldsymbol{x}) - \mathcal{A}(\boldsymbol{x}')\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{x}'\|_2$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$.

The (VI) formulation offers a versatile representation that embraces multiple structured problems as its special cases. Let us bring about some of the prominent examples which are covered by this powerful unifying framework.

2.2. Restricted Gap Function and the Oracle Model

To that end, a widely used performance metric to evaluate a candidate solution of (VI) is the so-called *restricted gap function*:

$$\mathrm{Gap}_{\mathcal{C}}(\hat{\boldsymbol{x}}) = \sup_{\boldsymbol{x} \in \mathcal{C}} \langle \mathcal{A}(\boldsymbol{x}), \hat{\boldsymbol{x}} - \boldsymbol{x} \rangle, \tag{Gap}$$

where the "test domain" $\mathcal{C}$ is a non-empty compact subset of $\mathbb{R}^d$. This particular construction of the merit function characterizes the solutions of the (VI) through its zeros.

Proposition 1 *Let $\mathcal{C}$ be a non-empty, convex and compact subset of $\mathcal{X}$. Then, the following hold true:* $\operatorname{Gap}_{\mathcal{C}}(\hat{\boldsymbol{x}}) \geq 0$, *whenever $\hat{\boldsymbol{x}} \in \mathcal{C}$. Conversely, if* $\operatorname{Gap}_{\mathcal{C}}(\hat{\boldsymbol{x}}) = 0$ *and $\mathcal{C}$ contains a neighborhood of $\hat{\boldsymbol{x}}$ in $\mathcal{X}$, then $\hat{\boldsymbol{x}}$ is a solution of* (VI).

The above is a generalization of the merit function by [37] (see also [1, 39] and references therein) for the general convex and compact test domains. We defer the proof of the above to Appendix B.

From an algorithmic point of view, our aim is to solve (VI) using iterative methods that require access to a so-called *first order oracle* (FO) [36]. This means that, at each stage of the algorithmic process, the optimizer can query some black-box mechanism that returns the operator $\mathcal{A}$ at the queried point. In other words, this mechanism has no prior knowledge of the operator and/or any favorable geometrical or regularity properties which $\mathcal{A}$ may or may not possess and only computes the value of $\mathcal{A}$ at a particular point in the problem domain $\mathcal{X}$.

2.3. Overview of the methods and their last iterate performance

Perhaps the most popular numerical method for games and variational inequalities (VIs) is the so-called extra-gradient (EG) algorithm, defined by the following recursive formula:

$$\boldsymbol{x}^{(t+1/2)} := \Pi_{\mathcal{X}}\big(\boldsymbol{x}^{(t)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t)})\big), \qquad \boldsymbol{x}^{(t+1)} := \Pi_{\mathcal{X}}\big(\boldsymbol{x}^{(t)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t+1/2)})\big), \tag{EG}$$

initially introduced by [27] and later developed by [35, 38]. Heuristically, the main idea is that, at each $t = 1, 2, \ldots$, the oracle is called at the algorithm's base state $\boldsymbol{x}^{(t)}$ to generate an intermediate, *leading* state $\boldsymbol{x}^{(t+1/2)}$; subsequently, the base state is updated with oracle information from the leading state $\boldsymbol{x}^{(t+1/2)}$ and the process repeats. In this way, (EG) essentially tries to "anticipate" the change of $\mathcal{A}$ along a projection step, and to exploit this "forward" information. This anticipatory mechanism, along with the gradual variation of the operator $\mathcal{A}$, hence the need for Lipschitz continuity, leads to a faster convergence rate than ordinary forward-backward/gradient descent schemes in the classical analysis of the algorithm [26, 35, 37]. Furthermore, the so-called *optimistic version* of (EG) was introduced by [42] and is defined as follows:

$$\boldsymbol{x}^{(t+1/2)} := \Pi_{\mathcal{X}}\big(\boldsymbol{x}^{(t)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t-1/2)})\big), \quad \boldsymbol{x}^{(t+1)} := \Pi_{\mathcal{X}}\big(\boldsymbol{x}^{(t)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t+1/2)})\big). \tag{OGDA}$$

Note, that for this case the leading state is generated by reusing the previous oracle query and therefore it exhibits the advantage of requiring only one oracle call per iteration in contrast to the (EG). Regarding the convergence rate guarantees, the typical analysis, considers either the *time average* or the *ergodic average* as the respective output of (EG)/(OGDA),*i.e.*,

$$\bar{\boldsymbol{x}}^{(T+1/2)} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}^{(t+1/2)} \text{ and } \tilde{\boldsymbol{x}}^{(T+1/2)} = \left(\sum_{t=1}^{T}\gamma^{(t)}\right)^{-1}\sum_{t=1}^{T}\gamma^{(t)}\boldsymbol{x}^{(t+1/2)} \tag{1}$$

In that case, an order optimal $\mathcal{O}(1/T)$ rate relative to the (Gap) is obtained. On the other hand, the behaviour of the actual iterates $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$, was only studied in the context of asymptotic convergence to a solution (see, [2, 25, 32]); and hence a particular rate relative to the (Gap) was largely unknown. This open problem has been tackled in [8, 21, 22, 24] for the specific case of (EG) run with a constant step size $\gamma \leq 1/L$.

If one considers the so-called *tangent residual*:

$$r(\boldsymbol{x}) = \min_{\zeta \in \mathcal{N}_{\mathcal{X}}(\boldsymbol{x})} \|\mathcal{A}(\boldsymbol{x}) + \zeta\|_2 \tag{2}$$

$$= \left\|\mathcal{A}\boldsymbol{x} + \Pi_{\mathcal{N}_{\mathcal{X}}(\boldsymbol{x})}(-\mathcal{A}(\boldsymbol{x}))\right\|_2 \tag{3}$$

with $\mathcal{N}_{\mathcal{X}}(\boldsymbol{x})$ being the *normal cone* of $\mathcal{X}$ at $\boldsymbol{x}$:

$$\mathcal{N}_{\mathcal{X}}(\boldsymbol{x}) = \{\zeta \in \mathbb{R}^d \mid \langle \zeta, \boldsymbol{x}' - \boldsymbol{x} \rangle \leq 0 \text{ for all } \boldsymbol{x}' \in \mathcal{X}\} \tag{4}$$

Proposition 2 *[8] Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterations generated by (EG)/ (OGDA) run with a constant step size $\gamma = 1/2L$. Moreover, assume that $\mathcal{C}$ is a compact, non-empty subset of $\mathcal{X}$ which contains a neighborhood of a solution of (VI). Then, the following estimation holds:*

$$\mathrm{Gap}_{\mathcal{C}}(\boldsymbol{x}^{(T+1/2)}) \leq \frac{DL}{\sqrt{T}}$$

Regarding the tightness, the respective performance is matched by a respective "uniform" lower bound on all (VI) problems associated with a monotone and smooth operator $\mathcal{A}$ for a whole class of 1-SCLI algorithms, introduced by [3], which includes the template (EG). More precisely, considering for simplicity the unconstrained (Zer) problem, and setting $\mathrm{Lip}(L)$ is the set of all monotone and L Lipschitz constant operators, in [22] one has:

$$\sup_{\mathcal{A} \in \mathrm{Lip}(L)} \left\|\mathcal{A}(\boldsymbol{x}^{(T)})\right\|_2^2 \geq \frac{DL^2}{20T}$$

Similarly, the lower bound holds also for the case of (OGDA). We investigate in more detail the fragility of these lower bounds for the case where a fixed operator $\mathcal{A}$ in appendix D is used.

3. Faster Rates for (EG)/(OGDA)'s Last Iterate Convergence

Having all this at hand, we aim to address the following fundamental issue: Imagine that instead of considering a uniform, in all Lipschitz and monotone (VI), worst-case lower bound, we face a fixed smooth variational inequality problem. Then we seek to answer the following question:

Is it possible to provide provable faster rates which outperform the existing "uniform" lower bounds?

More precisely, given a Lipschitz mnonotone (VI) problem should we expect a better performance for (EG)/(OGDA) templates than the $1/\sqrt{T}$ lower bound in the long-run?

We answer this question affirmatively by providing a strictly faster rate of order $o(1/\sqrt{T})$. In doing so, we begin with the simplest case of (EG)/(OGDA) methods run with a constant step size; more precisely where the optimizer has access to an exact estimation of the Lipschitz constant of $\mathcal{A}$. This intermediate result refines existing upper bounds obtained in the existing works of [8, 22].

Moving forward, we present the full power of our results to more practise-oriented scenarios, where the calculation of the Lipschitz constant of $\mathcal{A}$ may become a significant computational bottleneck, we first propose a family of line-search free, on-the-fly step size policies in the sense of [43, 44]. In turn, we extend the last iterate faster convergence rate for the said step size.

3.1. Faster Convergence Rate with Constant Step Sizes

In this section, we primarily consider the case where the (EG)/(OGDA) are run with a constant step size upper bounded by $1/L$. In particular, we assume that the respective step size satisfies:

$$\gamma^{(t)} \equiv \gamma \leq 1/\sqrt{32}L \tag{Constant}$$

To that end, the following theorem holds:

Theorem 3 *Assume $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates generated by (EG)/(OGDA) run with a constant step size policy satisfying (Constant). Moreover, assume that $\mathcal{C}$ a non-empty, convex and compact subset of $\mathcal{X}$ which contains a neighborhood of a solution of (VI). Then, the following hold:*

$$\lim_{T\to+\infty} \sqrt{T}\,\mathrm{Gap}_{\mathcal{C}}(\boldsymbol{x}^{(T)}) = \lim_{T\to+\infty} \sqrt{T}\,\mathrm{Gap}_{\mathcal{C}}(\boldsymbol{x}^{(T+1/2)}) = 0$$

Furthermore, we have:

$$\liminf_{T\to+\infty} \sqrt{T\log T}\,\mathrm{Gap}_{\mathcal{C}}(\boldsymbol{x}^{T}) = \liminf_{T\to+\infty} \sqrt{T\log T}\,\mathrm{Gap}_{\mathcal{C}}(\boldsymbol{x}^{T+1/2}) = 0. \tag{5}$$

The above theorem provides a more fine-grained and strictly faster asymptotic convergence rate compared to the one established in [8, 22].

3.2. Last Iterate with Adaptive Step Sizes

We turn now our attention towards establishing our main result; namely the last iterate convergence rate guarantees for (EG) run with an adaptive step size inspired by line-search free AdaGrad step size [2, 15, 31].

In particular computing the exact Lipschitz constant of $\mathcal{A}$ in order to properly fine-tune the respective step size, can become a very tedious task. Moreover, even the slightest miscalculation of L may lead to catastrophic oscillations of the methods and cycling phenomena. To that end, we shall study a family of adaptive step sizes of the following form:

$$\gamma^{(t)} = \left(\gamma^{(0)} + \sum_{j=1}^{t-1} j \left\| \boldsymbol{x}^{(j+1)} - \boldsymbol{x}^{(j+1/2)} \right\|_2^2\right)^{-1/2} \tag{Adapt}$$

where $\gamma^{(0)} > 0$. Having established the main ingredients, we are in a position to present the full potency of our analysis. More precisely, we show that (EG) run with (Adapt) exhibit a last iterate of order $o(1/\sqrt{T})$, if one considers (Gap) as a performance metric. Formally, this is captured by the following theorem.

Theorem 4 *Assume $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates generated by (EG) or (OGDA) run with the adaptive step size policy (Adapt). Moreover, assume that $\mathcal{C}$ a convex and compact subset of $\mathcal{X}$ which contains a neighborhood of a solution of (VI). Then, the following hold:*

$$\lim_{T\to+\infty} \sqrt{T}\,\mathrm{Gap}_{\mathcal{C}}(\boldsymbol{x}^{(T)}) = \lim_{T\to+\infty} \sqrt{T}\,\mathrm{Gap}_{\mathcal{C}}(\boldsymbol{x}^{(T+1/2)}) = 0$$

Furthermore, we have:

$$\liminf_{T\to+\infty} \sqrt{T\log T}\,\mathrm{Gap}_{\mathcal{C}}(\boldsymbol{x}^{(T)}) = \liminf_{T\to+\infty} \sqrt{T\log T}\,\mathrm{Gap}_{\mathcal{C}}(\boldsymbol{x}^{(T+1/2)}) = 0 \tag{6}$$

Our analysis builds around the following inequalities for (EG)/(OGDA) run with a generic, non-negative, non-increasing step size policy $\gamma^{(t)}$. This general approach provides in addition a unifying framework for examining (EG)/(OGDA) run for both constant and adaptive step size policies. To do this, we start by showing a novel template for (EG) relative to the tangent residual (3).

The second step is to prove a quasi-Fejér type inequality [6] for the "weighted" tangent residual sequence $(t\gamma^{(t)}r^{(t)})_{t\in\mathbb{N}}$ for the case of (EG) run with a non-negative, non-increasing step size $\gamma^{(t)}$.

Therefore the final step would be to show that the employed step size is bounded away from zero. While for the case of a constant step size is self-evident, we additionally show that this desired property is also satisfied by (Adapt). We defer the particular technical details to the Appendix.

4. Conclusions

In this paper, we have provided faster convergence rate guarantees for the popular (EG) run with both constant and adaptive step size policies. A fruitful direction is to further investigate the fragility of the respective lower bound in the long run. This line of research may provide further theoretical evidence which may bridge the worst-case theoretical guarantees with actual practical performance. We defer this open questions to future work.

5. Acknowledgments

K.A. was supported by the Swiss National Science Foundation (SNSF) under grant number 200021_205011. V.C. was supported by Hasler Foundation Program: Hasler Responsible AI (project number 21043), Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-24-1-0048.

References

[1] K. Antonakopoulos, E.V. Belmega, and P. Mertikopoulos. An adaptive mirror-prox algorithm for variational inequalities with singular operators. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Processing Information Systems*, 2019.

[2] Kimon Antonakopoulos, E. Veronica Belmega, and Panayotis Mertikopoulos. Adaptive extra-gradient methods for min-max optimization and games. In *ICLR '21: Proceedings of the 2021 International Conference on Learning Representations*, 2021.

[3] Yossi Arjevani, Shai Shalev-Shwartz, and Ohad Shamir. On lower and upper bounds in smooth and strongly convex optimization. *The Journal of Machine Learning Research*, 17(1): 4303–4353, 2016.

[4] Hedy Attouch and Juan Peypouquet. The rate of convergence of nesterov's accelerated forward-backward method is actually faster than $1/k^2$. *SIAM Journal on Optimization*, 26(3):1824–1834, January 2016. ISSN 1095-7189. doi: 10.1137/15m1046095. URL http://dx.doi.org/10.1137/15M1046095.

[5] Francis Bach and Kfir Y Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. *arXiv preprint arXiv:1902.01637*, 2019.

[6] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, NY, USA, 2 edition, 2017.

[7] Radu Ioan Bot, Dang-Khoa Nguyen, and Chunxiang Zong. Fast forward-backward splitting for monotone inclusions with a convergence rate of the tangent residual of $o(1/k)$, 2023.

[8] Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Tight last-iterate convergence of the extragradient and the optimistic gradient descent-ascent algorithm for constrained monotone variational inequalities, 2022.

[9] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, May 2011.

[10] Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *COLT '12: Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

[11] Shisheng Cui and Uday V. Shanbhag. On the analysis of reflected gradient and splitting methods for monotone stochastic variational inequality problems. In *CDC '16: Proceedings of the 57th IEEE Annual Conference on Decision and Control*, 2016.

[12] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*, 2018.

[13] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism, 2018.

[14] Jelena Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities, 2020.

[15] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[16] Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research. Springer, 2003.

[17] Gabriele Farina, Ioannis Anagnostides, Haipeng Luo, Chung-Wei Lee, Christian Kroer, and Tuomas Sandholm. Near-optimal no-regret learning dynamics for general convex games. In *Neural Information Processing Systems (NeurIPS)*, 2022.

[18] A.V. Gasnikov, P.E. Dvurechensky, F.S. Stonyakin, and A.A. Titov. An adaptive mirror-prox algorithm for variational inequalities. *Computational Mathematics and Mathematical Physics*, 59:836–841, 2019.

[19] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*, 2019.

[20] Andrew Gilpin. *Algorithms for abstracting and solving imperfect information games*. PhD thesis, Carnegie Mellon University, 2009.

[21] Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games, 2020.

[22] Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems, 2020.

[23] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS '14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014.

[24] Eduard Gorbunov, Adrien Taylor, and Gauthier Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities, 2022.

[25] Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos. Adaptive learning in continuous games: Optimal regret bounds and convergence to nash equilibrium. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2388–2422. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/hsieh21a.html.

[26] Anatoli Juditsky, Arkadi Semen Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[27] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[28] H. W. Kuhn. *A SIMPLIFIED TWO-PERSON POKER*, pages 97–104. Princeton University Press, Princeton, 1951.

[29] Yura Malitsky. Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520, 2015.

[30] Yura Malitsky. Golden ratio algorithms for variational inequalities, 2019.

[31] H Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. *COLT 2010*, page 244, 2010.

[32] Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*, 2019.

[33] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: proximal point approach. https://arxiv.org/abs/1901.08511v2, 2019.

[34] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. Convergence rate of $\mathcal{O}(1/k)$ for optimistic gradient and extra-gradient methods in smooth convex-concave saddle point problems. https://arxiv.org/pdf/1906.01115.pdf, 2019.

[35] Arkadi Nemirovski. Prox-method with rate of convergence o(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[36] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Number 87 in Applied Optimization. Kluwer Academic Publishers, 2004.

[37] Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

[38] Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Math. Program.*, 109(2–3):319–344, mar 2007. ISSN 0025-5610.

[39] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

[40] Wei Peng, Yu-Hong Dai, Hui Zhang, and Lizhi Cheng. Training GANs with centripetal acceleration. https://arxiv.org/abs/1902.08949, 2019.

[41] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *ICML '17: Proceedings of the 34th International Conference on Machine Learning*, 2017.

[42] Leonid Denisovich Popov. 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.

[43] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *COLT '13: Proceedings of the 26th Annual Conference on Learning Theory*, 2013.

[44] Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *NIPS '13: Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2013.

[45] Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013.

[46] Gesualdo Scutari, Francisco Facchinei, Daniel Pérez Palomar, and Jong-Shi Pang. Convex optimization, game theory, and variational inequality theory in multiuser communication systems. 27(3):35–49, May 2010.

[47] Michael Sedlmayer, Dang-Khoa Nguyen, and Radu Ioan Bot. A fast optimistic method for monotone variational inequalities, 2023.

[48] Fedor Stonyakin, Alexander Gasnikov, Alexander Tyurin, Dmitry Pasechnyuk, Artem Agafonov, Pavel Dvurechensky, Darina Dvinskikh, Alexey Kroshnin, and Victorya Piskunova. Inexact model: A framework for optimization and variational inequalities. https://arxiv.org/abs/1902.00990v6.

[49] Fedor Stonyakin, Alexander Gasnikov, Pavel Dvurechensky, Mohammad Alkousa, and Alexander Titov. Generalized mirror prox for monotone variational inequalities: Universality and inexact oracle. https://arxiv.org/abs/1806.05140, 2018.

[50] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/7fea637fd6d02b8f0adf6f7dc36aed93-Paper.pdf.

[51] Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, June 1995.

[52] Bernhard von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2):220–246, 1996.

[53] TaeHo Yoon and Ernest K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm, 2021.

Contents

Appendix A. Further related work on adaptive methods for (VI)

The literature on the topic is too vast to be summarized here. We mostly focus on the papers that are strongly related to one or more aspects of our results.

Theoretical guarantees which eventually outperform the respective lower bounds were first investigated for the case of simple smooth convex minimization in [4]. In that regard, a faster rate $o(1/T^2)$ is established for the case of convex smooth minimization, which seemingly outperforms the lower bound $1/T^2$, if the iterations of the algorithm exceed the dimension of the respective problem. This effect takes place because the construction of the respective "bad instances" in [36] heavily depend on the underlying dimensionality of the problem. Therefore, whenever the method iterations go beyond this threshold may exhibit faster rates On the other hand, this is not self-evident for our case. For example,(EG) remains unclear whether it can achieve faster asymptotic rates if we consider the time average as the method's output since even in the most favorable case one must always suffer the $\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\|_2^2/T$; this prohibits an asymptotically faster rate. That being said, for the case of the last iterate, our result is obtained by attacking the uniformity upon which the respective lower bound of (EG)/(OGDA) is built in [3, 22]. More precisely, instead of considering the lower bound in the monotone and Lipschitz continuous set (VI), we treat the performance of the (EG)/(OGDA) methods in each individual (VI).

On the other hand, other works focus on moving away from the slow last iterate performance of (EG)/(OGDA) by introducing novel algorithmic schemes, mostly based on the so-called Halpern iteration concept. More precisely, [53] suggests an anchored version of (EG) that is capable of performing as $\mathcal{O}(1/T)$ for unconstrained problems. Furthermore, in [14] a rate of order $\mathcal{O}(\log T/T)$ is established for the Halpern iterates. Finally, a recent line of work focuses on establishing faster $o(1/T)$ rates in [7, 47] for non-adaptive, beyond (EG) type methods. That being said, all these results go beyond the scope of this paper since they do not aim to establish new insights regarding any refinements for the actual last iterate of the traditional (EG) and (OGDA) algorithms. Furthermore, they do not study the convergence behaviour of AdaGrad type step size policies.

Several works have been performed to improve the guarantees of the original (EG)/(MP) template. All of them consider a time-average iterate as their output. We review some of these works below. Because many of these works appear in the literature on VI [16], we also use this language in the sequel. In unconstrained problems with an operator that is locally Lipschitz continuous (but not necessarily globally so), the (GRAAL) [30] achieves convergence without requiring prior knowledge of the problem's Lipschitz parameter.

However, (GRAAL) provides no rate guarantees for non-smooth problems – and hence, a fortiori, no interpolation guarantees either. By contrast, such guarantees are provided in problems with a bounded domain by the GMP algorithm of [49] under the umbrella of Hölder continuity.

Still, nothing is known about the convergence of (GRAAL) / (GMP) in problems with singularities (i.e. when the vector field defining the problem blows up at a boundary point of the problem domain). Singularities of this type were treated in a recent series of papers [1, 18, 48] using a "Bregman continuity" or "Lipschitz-like condition ". These methods are order-optimal in the smooth case, without requiring any knowledge of the problem's smoothness modulus. On the other hand, like (GRAAL) (but unlike (GMP), they do not provide any rate interpolation guarantees between smooth and non-smooth problems. Another method that simultaneously achieves an $\mathcal{O}(1/\sqrt{T})$ rate in non-smooth problems and an $\mathcal{O}(1/T)$ rate in smooth ones is the recent algorithm of [5].

The (BL) algorithm employs an Adagrad-like adaptive step size policy which allows the method to interpolate between the two regimes –and this, even with noisy gradient feedback. On the negative side, the (BL) algorithm requires a bounded domain with a diameter (Bregman) known in advance; as a result, its theoretical guarantees do not apply to unbounded problems.

Appendix B. Preliminaries

We consider a generic monotone variational inequality (VI) problem

$$\text{find} \quad \boldsymbol{x}^* \in \mathcal{X} \quad \text{such that} \quad \langle \mathcal{A}(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0 \quad \forall \boldsymbol{x} \in \mathcal{X}, \tag{VI}$$

where the operator $\mathcal{A}$ is Lipschitz continuous, that is, $\|\mathcal{A}\boldsymbol{x} - \mathcal{A}\boldsymbol{x}'\| \leq L\|\boldsymbol{x} - \boldsymbol{x}'\|$ for an appropriate Lipschitz constant $L > 0$, and $\mathcal{X} \subseteq \mathbb{R}^d$ is a closed and convex set; but *not* necessarily compact. Moreover, we define the so-called normal cone:

$$\mathcal{N}_{\mathcal{X}}(\boldsymbol{x}) = \{v \in \mathbb{R}^d : \langle v, \boldsymbol{x}' - \boldsymbol{x} \rangle \leq 0 \ \text{ for all } \ \boldsymbol{x}' \in \mathcal{X}\}. \tag{7}$$

We investigate the last-iterate convergence properties of the *Extragradient algorithm*, given by

$$\boldsymbol{x}^{(t+1/2)} := \Pi_{\mathcal{X}}\big(\boldsymbol{x}^{(t)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t)})\big), \boldsymbol{x}^{(t+1)} := \Pi_{\mathcal{X}}\big(\boldsymbol{x}^{(t)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t+1/2)})\big), \tag{EG}$$

and its optimistic counterpart:

$$\begin{aligned} \boldsymbol{x}^{(t+1/2)} &:= \Pi_{\mathcal{X}}\big(\boldsymbol{x}^{(t)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t-1/2)})\big), \\ \boldsymbol{x}^{(t+1)} &:= \Pi_{\mathcal{X}}\big(\boldsymbol{x}^{(t)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t+1/2)})\big), \end{aligned} \tag{OGDA}$$

where the learning rates $\{\gamma^{(t)}\}_{t\in\mathbb{N}}$ for a non-negative, non-increasing sequence.

To that end, we have the following general proposition.

Proposition 5 *Assume that* $\boldsymbol{x}^{+} = \Pi_{\mathcal{X}}(\boldsymbol{x})$. *Then, the following holds:*

$$\boldsymbol{x} - \boldsymbol{x}^{+} = \zeta \in \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}) \tag{8}$$

In order to streamline our presentation, for this fairly standard result we refer the reader to [6]. On the other we shall use as a performance criterion, the so-called *restricted gap function*:

$$\operatorname{Gap}_{\mathcal{C}}(\hat{\boldsymbol{x}}) = \sup_{\boldsymbol{x}\in\mathcal{C}}\langle\mathcal{A}(\boldsymbol{x}), \hat{\boldsymbol{x}} - \boldsymbol{x}\rangle, \tag{Gap}$$

Proposition 6 *Let* $\mathcal{C} \subseteq \mathbb{R}^d$ *be a compact, convex and nonempty subset of* $\mathcal{X}$. *Then, the following hold true:* $\operatorname{Gap}_{\mathcal{C}}(\hat{\boldsymbol{x}}) \geq 0$, *as long as* $\hat{\boldsymbol{x}} \in \mathcal{X}$. *If* $\operatorname{Gap}_{\mathcal{C}}(\hat{\boldsymbol{x}}) = 0$ *and* $\mathcal{C}$ *contains a neighbourhood of* $\hat{\boldsymbol{x}}$, *then* $\hat{\boldsymbol{x}}$ *is a solution of* (VI)

Proof Let $\boldsymbol{x}^{*} \in \mathcal{X}$ be a solution of (VI) so $\langle\mathcal{A}(\boldsymbol{x}^{*}), \boldsymbol{x} - \boldsymbol{x}^{*}\rangle \geq 0$ for all $\boldsymbol{x} \in \mathcal{X}$. Then, by monotonicity, we get:

$$\begin{aligned} \langle\mathcal{A}(\boldsymbol{x}), \boldsymbol{x}^{*} - \boldsymbol{x}\rangle &\leq \langle\mathcal{A}(\boldsymbol{x}) - \mathcal{A}(\boldsymbol{x}^{*}), \boldsymbol{x}^{*} - \boldsymbol{x}\rangle + \langle\mathcal{A}(\boldsymbol{x}^{*}), \boldsymbol{x}^{*} - \boldsymbol{x}\rangle \\ &= -\langle\mathcal{A}(\boldsymbol{x}^{*}) - \mathcal{A}(\boldsymbol{x}), \boldsymbol{x}^{*} - \boldsymbol{x}\rangle - \langle\mathcal{A}(\boldsymbol{x}^{*}), \boldsymbol{x} - \boldsymbol{x}^{*}\rangle \leq 0, \end{aligned} \tag{9}$$

so $\operatorname{Gap}_{\mathcal{C}}(\boldsymbol{x}^{*}) \leq 0$. On the other hand, if $\boldsymbol{x}^{*} \in \mathcal{C}$, we also get $\operatorname{Gap}_{\mathcal{C}}(\boldsymbol{x}^{*}) \geq \langle\mathcal{A}(\boldsymbol{x}^{*}), \boldsymbol{x}^{*} - \boldsymbol{x}^{*}\rangle = 0$, so we conclude that:

$$\operatorname{Gap}_{\mathcal{C}}(\boldsymbol{x}^{*}) = 0 \tag{10}$$

For the converse statement, assume that $\operatorname{Gap}_{\mathcal{C}}(\boldsymbol{x}^{+}) = 0$ for some $\boldsymbol{x}^{+} \in \mathcal{C}$ and suppose that $\mathcal{C}$ contains a neighborhood of $\boldsymbol{x}^{+}$ in $\mathcal{X}$. First, we claim that the following inequality holds:

$$\left\langle\mathcal{A}(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^{+}\right\rangle \geq 0 \quad \text{for all } \boldsymbol{x} \in \mathcal{C}. \tag{11}$$

Indeed, assume to the contrary that there exists some $\boldsymbol{x}^{(1)} \in \mathcal{C}$ such that

$$\left\langle\mathcal{A}(\boldsymbol{x}^{(1)}), \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(+)}\right\rangle < 0. \tag{12}$$

or, equivalently $\left\langle\mathcal{A}(\boldsymbol{x}^{(1)}), \boldsymbol{x}^{+} - \boldsymbol{x}^{(1)}\right\rangle > 0$. This would then give the following.

$$0 = \operatorname{Gap}_{\mathcal{C}}(\boldsymbol{x}^{+}) \geq \left\langle\mathcal{A}(\boldsymbol{x}^{(1)}), \boldsymbol{x}^{+} - \boldsymbol{x}^{(1)}\right\rangle > 0, \tag{13}$$

which is a contradiction. Now, we further claim that $\boldsymbol{x}^+$ is a solution of (VI),i.e.:

$$\langle \mathcal{A}(\boldsymbol{x}^+), \boldsymbol{x} - \boldsymbol{x}^+ \rangle \geq 0 \text{ for all } \boldsymbol{x} \in \mathcal{X}. \tag{14}$$

If we suppose that there exists some $z_1 \in \mathcal{X}$ such that $\langle \mathcal{A}\boldsymbol{x}^+, z_1 - \boldsymbol{x}^+ \rangle < 0$, then, by the continuity of $\mathcal{A}$, there exists a neighborhood $\mathcal{D}$ of $\boldsymbol{x}^+$ in $\mathcal{X}$ such that

$$\langle \mathcal{A}(\boldsymbol{x}), z_1 - \boldsymbol{x} \rangle < 0 \quad \text{for all } \boldsymbol{x} \in \mathcal{D}. \tag{15}$$

Hence, assuming without loss of generality that $\mathcal{D} \subset \mathcal{D}' \subset \mathcal{C}$ (the latter assumption due to the assumption that $\mathcal{C}$ contains a neighborhood of $\boldsymbol{x}^+$), and taking $\lambda > 0$ sufficiently small so that $\boldsymbol{x} = \boldsymbol{x}^+ + \lambda(z_1 - \boldsymbol{x}^+) \in \mathcal{D}$, we get $\langle \mathcal{A}\boldsymbol{x}, \boldsymbol{x} - \boldsymbol{x}^+ \rangle = \lambda \langle \mathcal{A}(\boldsymbol{x}), z_1 - \boldsymbol{x}^+ \rangle < 0$, in contradiction to (11). We conclude that $\boldsymbol{x}^+$ is a solution of (VI), as claimed. ■

Appendix C. Experiments

We validate our results numerically by running the extra-gradient method to compute a max-min strategy (Nash equilibrium) in two bilinear games.

EG and OGDA on Kuhn Poker. We investigated EG in Kuhn poker [28], a standard reference in the literature on extensive form games of imperfect information. The variational inequality corresponding to the vector field formulation of the equilibrium has dimension 26, and is constrained on a polyhedral set—the *sequence-form* polytope [52]—-with 14 linear constraints. The projections on the sequence-form polytope were computed using the algorithm laid out by Farina et al. [17] based on ideas from Gilpin [20].

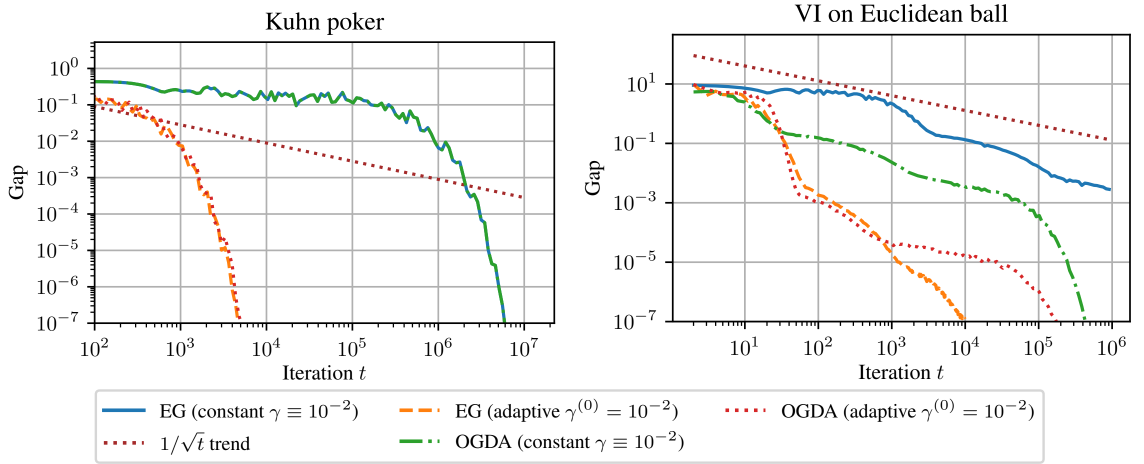

Figure 1: Gap of the iterates produced by the extra-gradient (EG) and optimistic gradient descent-ascent (OGDA) algorithms set up with constant and adaptive stepsizes, in two benchmark games.

fig. 1 (Left) shows the performance of EG and OGDA in the case in which the stepsize is set to the *constant* value $\gamma \equiv 10^{-2}$ and *adaptive* stepsizes (see section 3.2) starting from the same initial value $\gamma^{(0)} = 10^{-2}$. Both axes are on a logarithmic scale, with the x-axis displaying the number of iterations t of the algorithm, and the y-axis showing the value of $\mathrm{Gap}_{\mathcal{X}}(\boldsymbol{x}^{(t)})$ produced by EG and OGDA. A dotted line shows the slope of a function of the form $f(t) = 1/\sqrt{t}$; since the plot is on a logarithmic scale, the plot of a generic function $c/\sqrt{t}$ would be parallel to the dotted line. Hence, it becomes clear from the plot that the rate of decrease of the gap is not compatible with a rate of $1/\sqrt{t}$, and becomes significantly asymptotically faster as the number of iterations increases. This validates our results in theorem 3 and theorem 4. We remark that the adaptive choice of stepsize consistently produces significantly better-approximated equilibria (up to 7 orders of magnitude) for the same computational budget, validating the importance of studying adaptive schedules in the context of last-iterate convergence.

EG and OGDA on a Bilinear Game with Euclidean Ball Strategy Sets. fig. 1 (Right) follows a similar setup, but with strategy sets for the players that are set to be Euclidean balls of radius one. The payoff matrix is chosen with independently uniform entries in the range $[0, 1]$. The choice of strategy sets was guided by the desire to investigate EG beyond polyhedral domains. We observe that all conclusions drawn in appendix C hold verbatim. Once again, we observe that the performance of the adaptive stepsize scheduling is significantly better than that of constant stepsize.

Appendix D. On the Lower Bounds

To illustrate some of the fragility of the known lower bounds for the extra-gradient, we use the very adversarial example that was introduced by Golowich et al. [22] to show slow convergence.

In particular, Golowich et al. [22] consider the following monotone Lipschitz VI parameterized by the Lipschitzness constant $L = \nu > 0$:

$$\mathcal{A}_\nu : \boldsymbol{x} \mapsto \mathbf{A}_\nu \boldsymbol{x} + \boldsymbol{b}_\nu, \qquad \text{where} \quad \mathbf{A}_\nu := \begin{pmatrix} 0 & \nu \\ -\nu & 0 \end{pmatrix}, \quad \boldsymbol{b}_\nu := \begin{pmatrix} \nu \\ \nu \end{pmatrix}.$$

For any ν, the iterates produced by extra gradient with fixed learning rate $\gamma^{(t)} := \eta$ are therefore

$$\begin{aligned} \boldsymbol{x}^{(t+1/2)} &= \boldsymbol{x}^{(t)} - \eta(\mathbf{A}_\nu \boldsymbol{x}^{(t)} + \boldsymbol{b}_\nu), \\ \boldsymbol{x}^{(t+1)} &= \boldsymbol{x}^{(t)} - \eta(\mathbf{A}_\nu \boldsymbol{x}^{(t+1/2)} + \boldsymbol{b}_\nu), \end{aligned}$$

which implies, by substituting $\boldsymbol{x}^{(t+1/2)}$ into the expression for $\boldsymbol{x}^{(t+1)}$,

$$\boldsymbol{x}^{(t+1)} = \left(\mathbf{I} - \eta \mathbf{A}_\nu + \eta^2 \mathbf{A}_\nu^2\right) \boldsymbol{x}^{(t)} - \eta\big(\mathbf{I} - \eta \mathbf{A}_\nu\big)\boldsymbol{b}_\nu.$$

Assuming the initial point $\boldsymbol{x}^{(0)} = \mathbf{0}$, a closed-form solution to the above recurrence equation is given by

$$\boldsymbol{x}^{(t)} = \Big(\big(\mathbf{I} - \eta \mathbf{A}_\nu + \eta^2 \mathbf{A}_\nu^2\big)^t - \mathbf{I}\Big) \mathbf{A}_\nu^{-1} \boldsymbol{b}_\nu.$$

Plugging the above expression into the Hamiltonian leads to

$$\begin{aligned} \mathrm{Ham}(\boldsymbol{x}^{(t)}) &:= \|\mathbf{A}_\nu \boldsymbol{x}^{(t)} + \boldsymbol{b}_\nu\|_2^2 \\ &= \left\| \big(\mathbf{I} - \eta \mathbf{A}_\nu + \eta^2 \mathbf{A}_\nu^2\big)^t \boldsymbol{b}_\nu \right\|_2^2. \end{aligned}$$

The matrix $\mathbf{M} \coloneqq \mathbf{I} - \eta\mathbf{A}_\nu + \eta^2\mathbf{A}_\nu^2$ admits the unitary diagonalization $\mathbf{M} = \mathbf{C}\,\mathrm{diag}(\lambda_1, \lambda_2)\,\mathbf{C}^H$, where

$$\mathbf{C} \coloneqq \frac{1}{\sqrt{2}}\begin{pmatrix} -i & i \\ 1 & 1 \end{pmatrix}, \qquad \begin{aligned} \lambda_1 &\coloneqq \big(1 - (\eta\nu)^2\big) - (\eta\nu)i, \\ \lambda_2 &\coloneqq \big(1 - (\eta\nu)^2\big) + (\eta\nu)i. \end{aligned}$$

Hence,

$$\begin{aligned} \mathrm{Ham}(\boldsymbol{x}^{(t)}) &= \left\| \begin{pmatrix} \lambda_1^t & 0 \\ 0 & \lambda_2^t \end{pmatrix} \mathbf{C}^H \boldsymbol{b}_\nu \right\|_2^2 \\ &= \frac{\nu^2}{2} \left\| \begin{pmatrix} \lambda_1^t & 0 \\ 0 & \lambda_2^t \end{pmatrix} \begin{pmatrix} 1+i \\ 1-i \end{pmatrix} \right\|_2^2 \\ &= \nu^2\big(|\lambda_1|^t + |\lambda_2|^t\big) \\ &= 2\nu^2(1 - (\eta\nu)^2 + (\eta\nu)^4)^t. \end{aligned} \tag{16}$$

Suppose now that an upper bound $L > 0$ on the maximum possible Lipschitzness constant of $\mathcal{A}_\nu$ has been set. In other words, the adversary is constrained to only using game instances where $\nu \leq L$. Then, there are two cases.

- If the step size η is set too large, say $\eta > 1/L$, then by picking $\nu = L$ the adversary can guarantee exponential Hamiltonian growth. The algorithm does not converge at all!
- On the other hand, if $\eta \leq \frac{1}{L}$, then the adversary can pick

  $$\nu \coloneqq \frac{L}{2\sqrt{T}}$$

  and obtain that at iteration T the Hamiltonian is

  $$\begin{aligned} \mathrm{Ham}(\boldsymbol{x}^{(t)}) &= 2\frac{L^2}{4T}\left(1 - \frac{(\nu L)^2}{4T} + \frac{(\nu L)^4}{16T^2}\right)^T \\ &> 2\frac{L^2}{4T}\left(1 - \frac{(\nu L)^2}{4T}\right)^T \\ &\geq 2\frac{L^2}{4T}\left(1 - \frac{1}{4T}\right)^T \qquad \text{(since } \eta L \leq 1\text{)} \\ &\geq \frac{L^2}{2T} \cdot \frac{1}{4}. \end{aligned}$$

It is important to note, however, the fragility of such an approach. eq. (16) shows that, *for a fixed VI instance*, the Hamiltonian will in general decrease *exponentially fast*. In other words, the lower bound of Golowich et al. [22] requires that the adversary knows the number of iterations upfront, and that the extragradient method be used for exactly that number of iterations.

Appendix E. Analysis

In this section, we present the technical components of our analysis. In particular, we provide the proofs which lead to the faster convergence rates of (EG)/(OGDA) run with both (Constant) or (Adapt).

E.1. General regret inequalities

We start by showing a generic regret analysis of the (EG) method. More precisely, we have the following result.

Proposition 7 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates of* (EG) *run with a non-increasing, non-negative step size $\gamma^{(t)}$. Then, for all $\boldsymbol{x} \in \mathcal{X}$ the following inequality holds:*

$$\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x} \right\rangle \leq \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\right\|_2^2$$
$$+\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t+1/2)}) - \boldsymbol{x}^{(t+1)} \right\rangle - \frac{1}{2}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2$$

Proof Starting with the update rule of (EG) we have for all $\boldsymbol{x} \in \mathcal{X}$:

$$\left\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(\boldsymbol{x}^{(t+1/2)})}), \boldsymbol{x}^{(t+1)} - \boldsymbol{x} \right\rangle \geq 0 \tag{17}$$

and hence we get:

$$\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1)} - \boldsymbol{x} \right\rangle \leq \left\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)}, \boldsymbol{x}^{(t+1)} - \boldsymbol{x} \right\rangle \tag{18}$$

Now, by expanding the (RHS) of the above, we have:

$$\left\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)}, \boldsymbol{x}^{(t+1)} - \boldsymbol{x} \right\rangle = \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right\|_2^2 \tag{19}$$

So, we have:

$$\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1)} - \boldsymbol{x} \right\rangle \leq \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right\|_2^2 \tag{20}$$

Therefore, for all $\boldsymbol{x} \in \mathcal{X}$ by writing:

$$\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x} \right\rangle = \gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)} \right\rangle$$
$$+ \gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1)} - \boldsymbol{x} \right\rangle \tag{21}$$

and hence combining with (29), we have:

$$\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x} \right\rangle \leq \gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)} \right\rangle + \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\right\|_2^2$$
$$- \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right\|_2^2 \tag{22}$$

On the other hand by the extrapolation step of (EG) we have for all $\boldsymbol{x} \in \mathcal{X}$:

$$\left\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(\boldsymbol{x}^{(t)})}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x} \right\rangle \geq 0 \tag{23}$$

and hence we have:

$$\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x} \right\rangle \leq \left\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x} \right\rangle$$
$$= \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2$$

and by setting $\boldsymbol{x} = \boldsymbol{x}^{(t+1)}$ we have that:

$$\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)} \right\rangle \leq \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 \tag{24}$$

Therefore, by adding (31) and (33), we have:

$$\begin{aligned}&\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x} \right\rangle \leq \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\right\|_2^2 \\ &+\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t+1/2)}) - \boldsymbol{x}^{(t+1)} \right\rangle - \frac{1}{2}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 - \frac{1}{2}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2^2\end{aligned} \tag{25}$$

and hence the result follows. ■

The analog regret analysis for the (OGDA) algorithms is formalized by the following proposition.

Proposition 8 *Assume that $\boldsymbol{x}^{t+1/2}, \boldsymbol{x}^t$ are the iterations of (OGDA) run with a non-increasing, non-negative step size γ^t. Then, for all $x \in \mathcal{X}$ the following inequality holds:*

$$\begin{aligned}&\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x} \right\rangle \leq \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\right\|_2^2 \\ &+\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(t-1/2)}), \boldsymbol{x}^{(t+1/2)}) - \boldsymbol{x}^{(t+1)} \right\rangle - \frac{1}{2}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2\end{aligned}$$

Proof Starting with the update rule of (OGDA) we have for all $\boldsymbol{x} \in \mathcal{X}$:

$$\left\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(\boldsymbol{x}^{(t+1/2)})}), \boldsymbol{x}^{(t+1)} - \boldsymbol{x} \right\rangle \geq 0 \tag{26}$$

and hence we get:

$$\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1)} - \boldsymbol{x} \right\rangle \leq \left\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)}, \boldsymbol{x}^{(t+1)} - \boldsymbol{x} \right\rangle \tag{27}$$

Now, by expanding the (RHS) of the above, we have:

$$\left\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)}, \boldsymbol{x}^{(t+1)} - \boldsymbol{x} \right\rangle = \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right\|_2^2 \tag{28}$$

So, we have:

$$\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1)} - \boldsymbol{x} \right\rangle \leq \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right\|_2^2 \tag{29}$$

Therefore, for all $\boldsymbol{x} \in \mathcal{X}$ by writing:

$$\begin{aligned}\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x} \right\rangle &= \gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)} \right\rangle \\ &\quad + \gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1)} - \boldsymbol{x} \right\rangle\end{aligned} \tag{30}$$

and hence combining with (29), we have:

$$\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x} \right\rangle \leq \gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)} \right\rangle + \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\right\|_2^2 \\ - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right\|_2^2 \quad (31)$$

On the other hand by the extrapolation step of (OGDA) we have for all $\boldsymbol{x} \in \mathcal{X}$:

$$\left\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(\boldsymbol{x}^{(t-1/2)})}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x} \right\rangle \geq 0 \quad (32)$$

and hence we have:

$$\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t-1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x} \right\rangle \leq \left\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x} \right\rangle \\ = \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2$$

and by setting $\boldsymbol{x} = \boldsymbol{x}^{(t+1)}$ we have that:

$$\gamma^{(t-1/2)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)} \right\rangle \leq \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)}\right\|_2^2 \\ - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 \quad (33)$$

Therefore, by adding (31) and (33), we have:

$$\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x} \right\rangle \leq \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\right\|_2^2 \\ +\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(t-1/2)}, \boldsymbol{x}^{(t+1/2)}) - \boldsymbol{x}^{(t+1)} \right\rangle - \frac{1}{2}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 - \frac{1}{2}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 \quad (34)$$

and hence the result follows. ■

We move forward by providing a novel template inequality for (EG). In particular, we have the following proposition.

Proposition 9 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates of (EG) run with a non-increasing non-negative step size policy $\gamma^{(t)}$. Moreover let,*

$$\boldsymbol{c}^{(t+1)} := \Pi_{\mathcal{N}(\boldsymbol{x}^{(t+1)})}\big(-\mathcal{A}(\boldsymbol{x}^{(t+1)})\big), \qquad r^{(t+1)} := \left\|\mathcal{A}\boldsymbol{x}^{(t+1)} + \boldsymbol{c}^{(t+1)}\right\|_2.$$

Then, for all $\boldsymbol{x} \in \mathcal{X}$, we have

$$\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x} \right\rangle \leq \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\right\|_2^2 + (2\gamma^{(t)}L)^2\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 \\ - \min\left\{\frac{1}{4}, \frac{1}{4\gamma^{(0)}L}\right\}(\gamma^{(t)}r^{(t+1)})^2 - \frac{1}{8}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2.$$

Proof From the standard analysis of (EG), for all $\boldsymbol{x} \in \mathcal{X}$ we have :

$$\begin{aligned}\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x} \right\rangle &\leq \frac{1}{2}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\|_2^2 - \frac{1}{2}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\|_2^2 \\ +\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(t)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)} \right\rangle &- \frac{1}{2}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 - \frac{1}{2}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2^2.\end{aligned} \tag{35}$$

Moreover, we have that:

$$\begin{aligned}\gamma^{(t)}\left\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(t)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)} \right\rangle &\leq \gamma^{(t)}\|\mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(t)})\|_2 \cdot \|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\|_2 \\ &\leq \frac{1}{4L^2}\|\mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(t)})\|_2^2 + 4(\gamma^{(t)}L)^2\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\|_2^2 \qquad (36) \\ &\leq \frac{1}{4}\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t)}\|_2^2 + 4(\gamma^{(t)}L)^2\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\|_2^2, \qquad (37)\end{aligned}$$

where (36) follows from applying Young's inequality and (37) by Lipschitz continuity of $\mathcal{A}$.

Moreover, we have

$$\begin{aligned}&\frac{1}{2}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\|^2 + \frac{1}{2}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|^2 \\ &\quad= \frac{1}{4}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 + \frac{1}{8}\left[2\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\|_2 + 2\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2^2\right] + \frac{1}{4}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 \\ &\quad\geq \frac{1}{4}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 + \frac{1}{8}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)}\|_2^2 + \frac{1}{4}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 \qquad (38) \\ &\quad= \frac{1}{4}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 + \frac{1}{8}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)}\|_2^2 + \frac{1}{8}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 + \frac{1}{8}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 \\ &\quad\geq \frac{1}{4}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 + \frac{1}{8}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)}\|_2^2 + \frac{1}{8}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 \\ &\qquad\qquad + \frac{(\gamma^{(t)})^2}{8(\gamma^{(0)}L)^2}\|\mathcal{A}\boldsymbol{x}^{(t+1)} - \mathcal{A}\boldsymbol{x}^{(t+1/2)}\|_2^2 \qquad (39) \\ &\quad= \frac{1}{4}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 + \frac{1}{8}\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\|_2^2 \\ &\qquad\qquad + \frac{1}{16}\left[2\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)}\|_2^2 + \frac{2(\gamma^{(t)})^2}{(\gamma^{(0)}L)^2}\|\mathcal{A}(\boldsymbol{x}^{(t+1)}) - \mathcal{A}(\boldsymbol{x}^{(t+1/2)})\|_2^2\right] \\ &\quad\geq \frac{1}{4}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 + \frac{1}{8}\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\|_2^2 \\ &\qquad\qquad + \frac{\min\{1, 1/(\gamma^{(0)}L)^2\}}{16}\left\|\gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t+1)}) + (\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t+1/2)}))\right\|_2^2 \qquad (40)\end{aligned}$$

where (38) follows from the triangle inequality, (39) follows from the L-Lipschitzness of $\mathcal{A}$. Moreover, by denoting

$$\zeta^{(t+1)} = \boldsymbol{x}^{(t)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \boldsymbol{x}^{(t+1)} \in \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^{(t+1)}) \text{ and } u^{t+1} = \frac{1}{\gamma^{(t)}}\zeta^{(t+1)} \in \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^{(t+1)})$$

we have:

$$\begin{aligned}\left\|\gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t+1)}) + (\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)} - \gamma^{(t)}\mathcal{A}\boldsymbol{x}^{(t+1/2)})\right\|_2^2 &= (\gamma^{(t)})^2\left\|\mathcal{A}(\boldsymbol{x}^{(t+1)}) + u_{t+1}\right\|_2^2 &(41)\\ &\geq (\gamma^{(t)})^2\left\|\mathcal{A}(\boldsymbol{x}^{(t+1)}) + \boldsymbol{c}^{(t+1)}\right\|_2^2 &(42)\\ &= (\gamma^{(t)}r^{(t+1)})^2 &(43)\end{aligned}$$

with (42) being obtained by the definition of $\boldsymbol{c}^{(t+1)}$. Now, combining (43) with (40)

$$\begin{aligned}\gamma^{(t)}\left\langle\mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}\right\rangle \leq \frac{1}{2}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\|_2^2 - \frac{1}{2}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\|_2 + \frac{1}{4}\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t)}\|_2^2\\ + 4(\gamma^{(t)}L)^2\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\|_2^2\\ - \frac{1}{4}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 - \frac{1}{8}\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\|_2^2 - \frac{\min\{1, 1/(\gamma^{(0)}L)^2\}}{16}(\gamma^{(t)}r^{(t+1)})^2 \quad (44)\end{aligned}$$

which in turn yields:

$$\begin{aligned}\gamma^{(t)}\left\langle\mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}\right\rangle \leq \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\right\|_2^2 + 4(\gamma^{(t)}L)^2\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\|_2^2\\ - \frac{1}{8}\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\|_2^2 - \frac{\min\{1, 1/(\gamma^{(0)}L)^2\}}{16}(\gamma^{(t)}r^{(t+1)})^2 \quad (45)\end{aligned}$$

and hence the result follows. ■

Moving forward, we have the following novel template regarding the (OGDA) method, formalized in the following proposition.

Proposition 10 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates of (OGDA) run with a non-increasing, non-negative step-size policy $\gamma^{(t)}$. Moreover, let*

$$\boldsymbol{c}^{(t+1)} := \Pi_{\mathcal{N}(\boldsymbol{x}^{(t+1)})}\big(-\mathcal{A}(\boldsymbol{x}^{(t+1)})\big), \qquad r^{(t+1)} := \left\|\mathcal{A}\boldsymbol{x}^{(t+1)} + \boldsymbol{c}^{(t+1)}\right\|_2.$$

Then, for all $\boldsymbol{x} \in \mathcal{X}$ the following inequality holds:

$$\begin{aligned}\gamma^{(t)}\left\langle\mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(1/2)} - \boldsymbol{x}\right\rangle \leq \frac{1}{2}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\|_2^2 - \frac{1}{2}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\|_2^2\\ + \frac{1}{16}\Big(\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1/2)}\|_2^2 - \|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2^2\Big)\\ - \frac{\min\{1, 1/(\gamma^{(0)}L)^2\}}{16}(\gamma^{(t)}r^{(t+1)})^2 + 8L^2(\gamma^{(t)}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2)^2 - \frac{1}{16}\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\|_2^2\\ (46)\end{aligned}$$

Proof By invoking theorem 8 we have:

$$\begin{aligned}\gamma^{(t)}\left\langle\mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}\right\rangle \leq \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\right\|_2^2\\ + \gamma^{(t)}\left\langle\mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(t-1/2)}), \boldsymbol{x}^{(t+1/2)}) - \boldsymbol{x}^{(t+1)}\right\rangle - \frac{1}{2}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2\\ (47)\end{aligned}$$

Then, we have that:

$$\gamma^{(t)}\Big\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(t-1/2)}), \boldsymbol{x}^{(t+1/2)}) - \boldsymbol{x}^{(t+1)}\Big\rangle \le \|\mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(t-1/2)})\|_2 \\ \gamma^{(t)}\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\|_2 \quad (48)$$

Moreover, by applying the Fenchel-Young inequality, we get:

$$\|\mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(t-1/2)})\|_2\gamma^{(t)}\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\|_2 \le \frac{1}{32L^2}\|\mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(t-1/2)})\|_2^2 \\ + 8L^2(\gamma^{(t)}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2)^2 \quad (49)$$

where in order to obtain the above we used the standard inequality:

$$\alpha\beta \le \frac{1}{2\rho}\alpha^2 + \frac{\rho}{2}\beta^2 \quad (50)$$

for $\rho = 16L^2$. Furthermore, we have:

$$\frac{1}{32L^2}\|\mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(t-1/2)})\|_2^2 \le \frac{1}{32}\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t-1/2)}\|_2^2 \quad (51)$$

$$\le \frac{1}{16}\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t)}\|_2^2 + \frac{1}{16}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1/2)}\|_2^2 \quad (52)$$

On the other hand, we have: Moreover, we have

$$\frac{1}{2}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\|^2 + \frac{1}{2}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|^2 \\ \ge \frac{1}{4}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 + \frac{1}{8}\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\|_2^2 \\ + \frac{\min\{1, 1/(\gamma^{(0)}L)^2\}}{16}\Big\|\gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t+1)}) + (\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t+1/2)}))\Big\|_2^2 \quad (53)$$

Moreover, by denoting

$$\zeta^{(t+1)} = \boldsymbol{x}^{(t)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \boldsymbol{x}^{(t+1)} \in \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^{(t+1)}) \;\text{ and }\; u^{t+1} = \frac{1}{\gamma^{(t)}}\zeta^{(t+1)} \in \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^{(t+1)})$$

we have:

$$\|\gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t+1)}) + (\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t+1/2)}))\|_2^2 \ge (\gamma^{(t)}r^{(t+1)})^2 \quad (54)$$

by the definition $\boldsymbol{c}^{(t+1)}$. Therefore, summarizing we have:

$$\gamma^{(t)}\Big\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(1/2)} - \boldsymbol{x}\Big\rangle \le \frac{1}{2}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\|_2^2 - \frac{1}{2}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\|_2^2 + \frac{1}{16}\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t)}\|_2^2 \\ + \frac{1}{16}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1/2)}\|_2^2 - \frac{1}{4}\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t)}\|_2^2 - \frac{1}{8}\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\|_2^2 - \frac{\min\{1, 1/(\gamma^{(0)}L)^2\}}{16}(\gamma^{(t)}r^{(t+1)})^2 \\ + 8L^2(\gamma^{(t)}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2)^2 \quad (55)$$

which in turn yields:

$$\begin{aligned}\gamma^{(t)}\Big\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(1/2)} - \boldsymbol{x}\Big\rangle &\leq \frac{1}{2}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}\|_2^2 - \frac{1}{2}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}\|_2^2 \\ &+ \frac{1}{16}\Big(\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1/2)}\|_2^2 - \|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2^2\Big) \\ - \frac{\min\{1, 1/(\gamma^{(0)}L)^2\}}{16}(\gamma^{(t)}r^{(t+1)})^2 &+ 8L^2(\gamma^{(t)}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2)^2 - \frac{1}{16}\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\|_2^2\end{aligned} \tag{56}$$

which concludes the proof. ■

Now, turning our attention towards the tangent residual sequence we begin with the (EG) algorithm.

Proposition 11 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates of (EG) run with a non-increasing, non-negative step size $\gamma^{(t)}$. Then, the following inequality holds:*

$$(\gamma^{(t+1)}r^{t+1})^2 - (\gamma^{(t)}r^t)^2 \leq -\Big\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\Big\|_2^2 + (\gamma^{(t)}L)^2\Big\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\Big\|_2^2 \tag{57}$$

Proof We want to have a lower bound for:

$$-(\gamma^{(t+1)}r^{t+1})^2 + (\gamma^{(t)}r^t)^2 \tag{58}$$

Working in the spirit as in [8] we obtain the following inequalities: First, by applying monotonicity, we have:

$$-2\gamma^{(t)}\Big\langle \mathcal{A}\boldsymbol{x}^{(t+1)} - \mathcal{A}\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\Big\rangle \leq 0 \tag{59}$$

Moreover, due to the Lipschitz continuity of $\mathcal{A}$ we have:

$$(\gamma^{(t)})^2\Big\|\mathcal{A}\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\Big\|_2^2 - L^2(\gamma^{(t)})^2\Big\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\Big\|_2^2 \leq 0 \tag{60}$$

Furthermore, by the projection property we have:

$$-2\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)} - \gamma^{(t)}\mathcal{A}\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t)}\rangle \leq 0 \tag{61}$$

and

$$-2\Big\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)} - \gamma^{(t)}\mathcal{A}\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\Big\rangle \leq 0 \tag{62}$$

and

$$-2\Big\langle \boldsymbol{c}^{(t)}, \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)}\Big\rangle \leq 0 \tag{63}$$

and

$$-2\gamma^{(t)}\Big\langle \boldsymbol{c}^{(t+1)} + \mathcal{A}\boldsymbol{x}^{(t+1)}, \boldsymbol{x}^{(t)} - \gamma^{(t)}\mathcal{A}\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\Big\rangle \leq 0 \tag{64}$$

and

$$-2\gamma^{(t)}\Big\langle \boldsymbol{c}^{(t+1)}, -\gamma^{(t)}\boldsymbol{c}^{(t+1)}\Big\rangle \leq 0 \tag{65}$$

Then, by adding (68),(59), (60),(69),(72),(73),(74),(78) and rearranging we get the result. ■

Moreover, we can show the following general inequality, concerning the weighted sequence of the tangent residual generated by (EG).

Proposition 12 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates of (EG) run with a non-increasing, non-negative step size $\gamma^{(t)}$. Then, for all $t = 1, 2, \dots$ the following inequality holds:*

$$(t+1)(\gamma^{(t+1)} r^{(t+1)})^2 - t(\gamma^{(t)} r^{(t)})^2 \leq (\gamma^{(t+1)} r^{(t+1)})^2 - t\left\| \boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)} \right\|_2^2 \\ + t(L\gamma^{(t)})^2 \left\| \boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)} \right\|_2^2$$

Proof We have:

$$(t+1)(\gamma^{(t+1)} r^{(t+1)})^2 - t(\gamma^{(t)} r^{(t)})^2 = (\gamma^{(t+1)} r^{(t+1)})^2 + t\Big((\gamma^{(t+1)} r^{(t+1)})^2 - (\gamma^{(t)} r^{(t)})^2 \Big) \tag{66}$$

and hence by applying theorem 11 we obtain:

$$(t+1)(\gamma^{(t+1)} r^{(t+1)})^2 - t(\gamma^{(t)} r^{(t)})^2 \leq (\gamma^{(t+1)} r^{(t+1)})^2 - t\left\| \boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)} \right\|_2^2 \\ + t(L\gamma^{(t)})^2 \left\| \boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)} \right\|_2^2 \tag{67}$$

and therefore the result follows. ■

On the other hand, regarding the (OGDA) generated tangent residual is formalized by the following proposition.

Proposition 13 *Assume that $\boldsymbol{x}^{t+1/2}, \boldsymbol{x}^{t}$ are the iterates of (OGDA) run with a non-increasing, non-negative step size policy $\gamma^{(t)}$. Then, the following inequality holds:*

$$\Big(\gamma^{(t+1)} r^{(t+1)}\Big)^2 + \Big(\gamma^{t+1} \| \mathcal{A}(\boldsymbol{x}^{(t+1)}) - \mathcal{A}(\boldsymbol{x}^{(t+1/2)}) \|_2 \Big)^2 - \Big(\gamma^{(t)} r^{(t)}\Big)^2 \\ - \Big(\gamma^{(t)} \| \mathcal{A}(\boldsymbol{x}^{(t)}) - \mathcal{A}(\boldsymbol{x}^{(t-1/2)}) \|_2 \Big)^2 \leq -\| \boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1/2} \|_2^2 + 2\Big(\gamma^{(t)} L \| \boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)} \|_2 \Big)^2$$

Proof We would like to bound from below the following quantity:

$$-\Big(\gamma^{(t+1)} r^{(t+1)}\Big)^2 - \Big(\gamma^{t+1} \| \mathcal{A}(\boldsymbol{x}^{(t+1)}) + \mathcal{A}(\boldsymbol{x}^{(t+1/2)}) \|_2 \Big)^2 + \Big(\gamma^{(t)} r^{(t)}\Big)^2 \\ - \Big(\gamma^{(t)} \| \mathcal{A}(\boldsymbol{x}^{(t)}) - \mathcal{A}(\boldsymbol{x}^{(t-1/2)}) \|_2 \Big)^2 \tag{68}$$

In particular, we have the following inequalities. First, due to the monotonicity of $\mathcal{A}$, we have:

$$-2\gamma^{(t)} \Big\langle \mathcal{A}(\boldsymbol{x}^{(t+1)}) - \mathcal{A}(\boldsymbol{x}^{(t)}), \boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)} \Big\rangle \leq 0 \tag{69}$$

Moreover, since $\boldsymbol{x}^{(t+1)} = \Pi_{\mathcal{X}}(\boldsymbol{x}^{(t)} - \gamma^{(t)} \mathcal{A}(\boldsymbol{x}^{(t+1/2)}))$ and $\boldsymbol{x}^{(t+1/2)} = \Pi_{\mathcal{X}}(\boldsymbol{x}^{(t)} - \gamma^{(t)} \mathcal{A}(\boldsymbol{x}^{(t-1/2)}))$ we have the following:

$$\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)} - \gamma^{(t)} \mathcal{A}(\boldsymbol{x}^{(t+1/2)}) \in \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^{(t+1)}) \tag{70}$$

and

$$\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t-1/2)}) \in \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^{(t+1/2)}) \tag{71}$$

These directly yield the following inequalities:

$$-\Big\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)} \Big\rangle \leq 0 \tag{72}$$

and

$$-\Big\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{t-1/2}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{t} \Big\rangle \leq 0 \tag{73}$$

Moreover, since by definition $\boldsymbol{c}^{(t)} \in \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^{(t)})$ we have the following:

$$-\gamma^{(t)}\Big\langle \boldsymbol{c}^{(t)}, \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1/2)} \Big\rangle \leq 0 \tag{74}$$

and

$$-\gamma^{(t)}\Big\langle \boldsymbol{c}^{(t)}, \boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)} \Big\rangle \leq 0 \tag{75}$$

Finally, because of the fact that:

$$\boldsymbol{x}^{(t)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \boldsymbol{x}^{(t+1)} \in \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^{(t+1)}) \text{ and } \boldsymbol{c}^{(t+1)} = \Pi_{\mathcal{N}_{\mathcal{X}}}(-\mathcal{A}(\boldsymbol{x}^{(t+1)})) \tag{76}$$

we get:

$$-2\gamma^{(t)}\Big\langle \boldsymbol{c}^{(t+1)} + \mathcal{A}(\boldsymbol{x}^{(t+1)}), \boldsymbol{x}^{(t)} - \gamma^{(t)}\mathcal{A}(\boldsymbol{x}^{(t+1/2)}) - \boldsymbol{x}^{(t+1)} \Big\rangle \leq 0 \tag{77}$$

and

$$-2\gamma^{(t)}\Big\langle \boldsymbol{c}^{(t+1)} + \mathcal{A}(\boldsymbol{x}^{(t+1)}), -\boldsymbol{c}^{(t+1)} \Big\rangle = 0 \tag{78}$$

Finally, having established the above inequalities, by adding (68),(69),(72),(73),(74),(78) and after rearranging we get:

$$\begin{aligned}&\Big(\gamma^{(t+1)}r^{(t+1)}\Big)^2 + \Big(\gamma^{t+1}\|\mathcal{A}(\boldsymbol{x}^{(t+1))} - \mathcal{A}(\boldsymbol{x}^{(t+1/2)})\|_2\Big)^2 - \Big(\gamma^{(t)}r^{(t)}\Big)^2 \\ &-\Big(\gamma^{(t)}\|\mathcal{A}(\boldsymbol{x}^{(t)}) - \mathcal{A}(\boldsymbol{x}^{(t-1/2)})\|_2\Big)^2 \leq 2\Big(\gamma^{(t)}\|\mathcal{A}(\boldsymbol{x}^{(t+1)}) - \mathcal{A}(\boldsymbol{x}^{(t+1/2)})\|_2\Big)^2 - \|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2^2\end{aligned} \tag{79}$$

Then, the result follows by Lipschitz continuity of $\mathcal{A}$. ■

Moreover, building on theorem 13 may show the following recursive formula for weighted tangent residual generated by (OGDA).

Proposition 14 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates of* (OGDA) *run with a non-increasing, non-negative step-size policy $\gamma^{(t)}$. Then, for all $t = 1, 2, \dots$ the following inequality holds:*

$$\begin{aligned}&(t+1)\Big(\gamma^{t+1}r^{(t+1)}\Big)^2 + (t+1)\Big(\gamma^{t+1}\|\mathcal{A}(\boldsymbol{x}^{(t+1))} - \mathcal{A}(\boldsymbol{x}^{(t+1/2)})\|_2\Big)^2 - t\Big(\gamma^{t+1}r^{(t+1)}\Big)^2 \\ &-t\Big(\gamma^{(t)}\|\mathcal{A}(\boldsymbol{x}^{(t)}) - \mathcal{A}(\boldsymbol{x}^{(t-1/2)})\|_2\Big)^2 \leq -t\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 + 2t\Big(\gamma^{(t)}L\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2\Big)^2 \\ &\qquad + \Big(\gamma^{t+1}r^{(t+1)}\Big)^2 + \Big(\gamma^{t+1}\|\mathcal{A}(\boldsymbol{x}^{(t+1))} - \mathcal{A}(\boldsymbol{x}^{(t+1/2)})\|_2\Big)^2\end{aligned}$$

Proof We have that:

$$
\begin{aligned}
&(t+1)\Big(\gamma^{(t+1)}r^{(t+1)}\Big)^2 + (t+1)\Big(\gamma^{(t+1)}\|\mathcal{A}(\boldsymbol{x}^{(t+1))} - \mathcal{A}(\boldsymbol{x}^{(t+1/2)})\|_2\Big)^2 - t\Big(\gamma^{t+1}r^{(t+1)}\Big)^2 \\
&-t\Big(\gamma^{(t)}\|\mathcal{A}(\boldsymbol{x}^{(t)}) - \mathcal{A}(\boldsymbol{x}^{(t-1/2)})\|_2\Big)^2 = t\bigg(\Big(\gamma^{(t+1)}r^{(t+1)}\Big)^2 + \Big(\gamma^{(t+1)}\|\mathcal{A}(\boldsymbol{x}^{(t+1))} - \mathcal{A}(\boldsymbol{x}^{(t+1/2)})\|_2\Big)^2 \\
&\qquad - \Big(\gamma^{(t)}r^{(t+1)}\Big)^2 - \Big(\gamma^{(t)}\|\mathcal{A}(\boldsymbol{x}^{(t)}) - \mathcal{A}(\boldsymbol{x}^{(t-1/2)})\|_2\Big)^2\bigg) \\
&\qquad\qquad + \Big(\gamma^{(t+1)}r^{(t+1)}\Big)^2 + \Big(\gamma^{(t+1)}\|\mathcal{A}(\boldsymbol{x}^{(t+1))} - \mathcal{A}(\boldsymbol{x}^{(t+1/2)})\|_2\Big)^2
\end{aligned}
$$

Then, the result follows directly by theorem 13. ■

Finally, we have the following proposition regarding the distance of the iterates and a respective solution of the (VI). In particular,

Proposition 15 *Assume that* $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ *are the iterates generated by* (EG)/(OGDA) *run either with a constant step size* $\gamma \leq 1/\sqrt{32}L$ *or with* (Adapt)*. Then, for every* $\boldsymbol{x}^*$ *being a solution of* (VI) *the sequence* $\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|_2$ *is bounded.*

Proof Starting with the generic inequality in theorem 9 and by rearranging we have:

$$\frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\right\|_2^2 \leq \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|_2 + \left((2\gamma^{(t)}L)^2 - \frac{1}{8}\right)\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 \tag{80}$$

So, for the first case of a constant step size, we readily get that:

$$\frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\right\|_2^2 \leq \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|_2 \tag{81}$$

which in turn yields, after telescoping $t = 1, \dots T$ we get:

$$\frac{1}{2}\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^*\right\|_2^2 \leq \frac{1}{2}\left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\right\|_2^2 \tag{82}$$

and so the first case follows. Now, we turn our attention to the adaptive case. Again,in by telescoping from $t = 1, \dots T$ we have:

$$
\begin{aligned}
\frac{1}{2}\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^*\right\|_2^2 &\leq \frac{1}{2}\left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\right\|_2^2 + \gamma^0\sum_{t=1}^{T}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 && (83)\\
&\leq \frac{1}{2}\left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\right\|_2^2 + \gamma^0\sum_{t=1}^{T}t\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 && (84)
\end{aligned}
$$

Then, the result follows by theorem 26 and hence the second is shown. Concerning the respective distance to a solution for the iterates generated by (OGDA), by working in the same spirit as above we again divide the proof into two parts. For the constant step size the result is directly obtained by the appropriate choise of the step size along with the theorem 10 (for $\boldsymbol{x} = \boldsymbol{x}^*$ being a solution of (VI)).Moreover, for the (Adapt) step size iterates of (OGDA) is again obtained in the same spirit by invoking theorem 10 (for $\boldsymbol{x} = \boldsymbol{x}^*$ being a solution of (VI)) and theorem 26. ■

E.2. Constant step size

As a warm-up, we start by presenting the case where (EG)/(OGDA) are run with a constant step size $\gamma^{(t)} \equiv \gamma$. In doing so, we improve the results of [8, 22], by showing a strictly faster asymptotic rate of order $o(1/\sqrt{T})$. First, we show the summability of the tangent residual $(r^{(t)})^2$ generated by (EG). In particular, we have the following proposition.

Proposition 16 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are iterates of* (EG) *run with constant step size $\gamma \leq 1/\sqrt{32}L$. Then, the following holds:*

$$\sum_{t=1}^{+\infty} (r^{(t+1)})^2 < +\infty \tag{85}$$

Proof By setting $\boldsymbol{x} = \boldsymbol{x}^*$ to be a solution of (VI) in theorem 9, using the fact that:

$$\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0$$

and after rearranging we have:

$$\begin{aligned} \min\left\{\frac{1}{4}, \frac{1}{4\gamma^{(0)}L}\right\} \gamma^2 (r^{(t+1)})^2 &\leq \frac{1}{2}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\right\|_2^2 \\ &+ 4(\gamma^{(t)}L)^2 \|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\|_2^2 - \frac{1}{8}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 \end{aligned} \tag{86}$$

Now, by telescoping (86) $t = 1, \ldots T$ we get:

$$\min\left\{\frac{1}{4}, \frac{1}{4\gamma^{(0)}L}\right\} \gamma^2 \sum_{t=1}^{T} (r^{(t+1)})^2 \leq \frac{1}{2}\left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\right\|_2^2 + \sum_{t=1}^{T}\left(4\gamma^2 L^2 - \frac{1}{8}\right)\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 \tag{87}$$

$$\leq \frac{1}{2}\left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\right\|_2^2 \tag{88}$$

with the last inequality being obtained by the specific choice of γ. Therefore, the result follows by letting $T \to +\infty$.

■

Moreover, we will show that the weighted tangent residual sequence $t\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2$ generated by (EG) is summable. In particular, we have:

Proposition 17 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates of* (EG) *run with a constant step size $\gamma \leq 1/\sqrt{32}L$. Then, the following holds:*

$$\sum_{t=1}^{+\infty} t\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 < +\infty \tag{89}$$

Proof By invoking theorem 12 we have:

$$(t+1)\gamma^2 (r^{(t+1)})^2 \leq t\gamma^2 (r^{(t)})^2 + (\gamma r^{(t+1)})^2 + \left((L\gamma)^2 - 1\right) t\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 \tag{90}$$

and since $\gamma \leq \frac{1}{2L}$, the above yields:

$$(t+1)\gamma^2(r^{(t+1)})^2 \leq t\gamma^2(r^{(t)})^2 + (\gamma r^{(t+1)})^2 - \frac{1}{2}t\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 \tag{91}$$

and hence after rearranging and telescoping $t = 1, \ldots T$ we have:

$$\sum_{t=1}^{T} t\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 \leq \gamma^2(r^{(1)})^2 + \sum_{t=1}^{+\infty}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 < +\infty \tag{92}$$

and hence the result follows by letting $T \to +\infty$. ■

Moreover, we show that the sequence $(t(r^{(t+1)}))^2$ has a limit. In particular, we have the following result.

Proposition 18 *Assume $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates of* (EG) *run with a constant step size $\gamma \leq 1/2L$. Then, the sequence $\left(t(\gamma r^{(t+1)})^2\right)_{t\in\mathbb{N}}$ has a limit.*

Proof By applying theorem 12 we have:

$$(t+1)\gamma^2(r^{(t+1)})^2 \leq t\gamma^2(r^{(t)})^2 + (\gamma r^{(t+1)})^2 + \left((L\gamma)^2 - 1\right)t\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 \tag{93}$$

which in turn yields by the specific choice of γ:

$$(t+1)\gamma^2(r^{(t+1)})^2 \leq t\gamma^2(r^{(t)})^2 + (\gamma r^{(t+1)})^2 \tag{94}$$

The result follows by combing theorem 16 with theorem 31. ■

Following the same methodology for the (OGDA) algorithm we first have the following proposition. Namely

Proposition 19 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates of* (OGDA) *run with a constant step size $\gamma \leq$. Then, the following holds:*

$$\sum_{t=1}^{+\infty}\|\mathcal{A}(\boldsymbol{x}^{(t+1)}) - \mathcal{A}(\boldsymbol{x}^{(t+1/2)})\|_2^2 + (r^{(t+1)})^2 < +\infty \tag{95}$$

Proof By setting $\boldsymbol{x} = \boldsymbol{x}^*$ to be a solution of the (VI) in theorem 10 and using the fact that:

$$\langle \mathcal{A}(\boldsymbol{x}^{(t+1/2)}), \boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^*\rangle \geq 0 \tag{96}$$

Hence, after rearranging and applying the Lipschitz continuity of $\mathcal{A}$ and telescoping $t = 1, \ldots, T$ we get the following.

$$\sum_{t=1}^{T}\|\mathcal{A}(\boldsymbol{x}^{(t+1)}) - \mathcal{A}(\boldsymbol{x}^{(t+1/2)})\|_2^2 + (r^{(t+1)})^2 \leq \frac{1}{2}\left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\right\|_2^2 + \frac{1}{8}\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(1/2)}\|_2^2$$
$$+ \sum_{run=1}^{T}\left(8L^2\gamma^2 - \frac{1}{32}\right)\|\boldsymbol{x}^{(t+1/2)} - \boldsymbol{x}^{(t+1)}\|_2^2 \tag{97}$$

Therefore, by choosing γ we get:

$$\sum_{t=1}^{T}\|\mathcal{A}(\boldsymbol{x}^{(t+1)})-\mathcal{A}(\boldsymbol{x}^{(t+1/2)})\|_2^2+(r^{(t+1)})^2\le\frac{1}{2}\|\boldsymbol{x}^{(1)}-\boldsymbol{x}^*\|_2^2+\frac{1}{8}\|\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(1/2)}\|_2^2 \tag{98}$$

and the result follows. ■

Proposition 20 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates of* (OGDA) *run with a constant step-size γ Then, the following holds:*

$$\sum_{t=1}^{+\infty} t\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\|_2^2<+\infty \tag{99}$$

Proof By invoking theorem 14, after rearranging and telescoping $t=1,\dots,T$ we get that:

$$\frac{1}{2}\sum_{t=1}^{T} t\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\|_2^2\le(\gamma r^{(1)})^2+(\gamma\|\mathcal{A}(\boldsymbol{x}^{1})-\boldsymbol{x}^{(1/2)}\|_2)^2+\frac{1}{2}\sum_{t=1}^{T} t\left(2L^2\gamma^2-1\right)\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\|_2^2 \\ +\sum_{t=1}^{T}\left(\gamma^2(r^{(t+1)})^2+\gamma^2\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\|_2^2\right) \tag{100}$$

Then, the result follows by the appropriate choice of the step size and theorem 19. ■

Proposition 21 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{t}$ are the iterates of* (OGDA) *run with a constant step size γ. Then, the sequence $\left((t+1)(\gamma^2(r^{(t+1)})^2+\gamma^2\|\mathcal{A}(\boldsymbol{x}^{(t+1/2)})-\mathcal{A}(\boldsymbol{x}^{(t+1)})\|_2^2\right)_t$ has a limit.*

Proof Note that that by theorem 14 and theorem 19 and the appropriate choice of γ one directly obtains the fact that the theorem 31 is satisfied and therefore the result follows. ■

We shall now prove the faster convergence rate for $r^{(t+1)}$. More precisely, we have the following proposition.

Proposition 22 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates of* (EG)/(OGDA) *run with a constant step size $\gamma\le 1/\sqrt{32}L$. Then, we have that:*

$$r^{T}=o(1/\sqrt{T}) \tag{101}$$

In other words, we have that:

$$\sqrt{T}r^{T}\to 0 \tag{102}$$

Proof We have by theorem 16:

$$\sum_{t=1}^{+\infty}\frac{1}{t+1}(t+1)(r^{(t+1)})^2<+\infty \tag{103}$$

Moreover, by theorem 18 we have that:

$$\lim_{t\to+\infty}(t+1)(r^{(t+1)})^2 = r_\infty \geq 0 \tag{104}$$

Assume to the contrary that $r_\infty > 0$. Then, there exists some $t_0 \in \mathbb{N}$ such that:

$$(t+1)(r^{(t+1)})^2 > \frac{r_\infty}{2} \quad \text{for all} \quad t > t_0 \tag{105}$$

Now combining the above with (103) we have:

$$+\infty > \sum_{t=1}^{+\infty} \frac{1}{t+1}(t+1)(r^{(t+1)})^2 = \sum_{t=1}^{t_0} \frac{1}{t+1}(t+1)(r^{(t+1)})^2 + \sum_{t=t_0+1}^{+\infty} \frac{1}{t+1}(t+1)(r^{(t+1)})^2 \tag{106}$$

$$\geq \sum_{t=1}^{t_0} \frac{1}{t+1}(t+1)(r^{(t+1)})^2 + \frac{r_\infty}{2}\sum_{t=t_0+1}^{+\infty} \frac{1}{t+1} \tag{107}$$

which is a contradiction. Therefore, $r_\infty = 0$ and hence the result follows. Note that the respective result concerning (OGDA) is obtained by using the same reasoning, ■

Remark 23 *Given the above reasoning, we may in addition extract a subsequence of $\boldsymbol{x}^{(T)}$ such that:*

$$\liminf_{T\to+\infty}\left(T\log T(r^{(T)})^2\right) = 0 \tag{108}$$

This is obtain by the following generic observation. In particular, given any non negative sequence $\beta^{(t)}$ such that $\sum_{t=1}^{+\infty}\frac{\beta^{(t)}}{t} < +\infty$, it cannot be $\liminf_{T\to+\infty}\beta^{(T)}\log T = \varepsilon > 0$, otherwise there $\frac{\beta^{(T)}}{T} \geq \frac{\varepsilon}{T\log T}$ for sufficiently large T, which in turn violates the convergence of the series [4].

Finally, we are in the position to show the final rate of (EG)'s last iterate for the constant step size case. Formally, we have the following result.

Proposition 24 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates generated by (EG)/(OGDA) run with a constant step size $\gamma \leq 1/\sqrt{32}L$. Moreover, assume that $\mathcal{C}$ is a compact neighborhood of a (VI) solution. Then, the following hold:*

$$\operatorname{Gap}_{\mathcal{C}}(\boldsymbol{x}^{(T+1)}) = o(1/\sqrt{T}) \quad \textit{and} \quad \operatorname{Gap}_{\mathcal{C}}(\boldsymbol{x}^{(T+1/2)}) = o(1/\sqrt{T}) \tag{109}$$

Moreover, we have that:

$$\liminf_{T\to+\infty}\sqrt{(T+1)\log(T+1)}\operatorname{Gap}_{\mathcal{C}}(\boldsymbol{x}^{(T+1)}) = \liminf_{T\to+\infty}\sqrt{(T+1)\log(T+1)}\operatorname{Gap}_{\mathcal{C}}(\boldsymbol{x}^{(T+1/2)}) = 0 \tag{110}$$

Proof We first show the last iterate rate for $\boldsymbol{x}^{(T)}$. In particular, for all $\boldsymbol{x} \in \mathcal{C}$ we have:

$$
\begin{aligned}
\sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}), \boldsymbol{x}^{(T+1)} - \boldsymbol{x}\rangle &\leq \sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^{(T+1)} - \boldsymbol{x}\rangle && (111)\\
&\leq \sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}, \boldsymbol{x}^{(T+1)} - \boldsymbol{x}\rangle && (112)\\
&= \sqrt{T+1}\Big(\langle \mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}, \boldsymbol{x}^{(T+1)} - \boldsymbol{x}^*\rangle && (113)\\
&+ \langle \mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}, \boldsymbol{x}^* - \boldsymbol{x}\rangle\Big) && (114)\\
&\leq \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}\right\|_2\left(\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^*\right\|_2 + \|\boldsymbol{x}^* - \boldsymbol{x}\|_2\right) && (115)\\
&\leq \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}\right\|_2\left(\left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(*)}\right\|_2 + \|\boldsymbol{x}^* - \boldsymbol{x}\|_2\right) && (116)
\end{aligned}
$$

with (111) being obtained by monotonicity of $\mathcal{A}$, (112) by the fact that $\boldsymbol{c}^{(T+1)}$ belongs by definition to the normal cone of $\boldsymbol{x}^{(T+1)}$ and finally (116). Hence, by taking suprema on both sides relative to $\boldsymbol{x}$ we get:

$$
\sqrt{T+1}\,\mathrm{Gap}(\boldsymbol{x}^{(T+1)}) \leq \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(t+1)}\right\|_2 D \tag{117}
$$

with $D = \left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\right\|_2 + \sup_{\boldsymbol{x}}\|\boldsymbol{x}^* - \boldsymbol{x}\|_2 < +\infty$. Therefore, the first claim follows directly from theorem 22.

Moving forward, for the second claim we have:

$$
\begin{aligned}
\sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}), \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}\rangle &\leq \sqrt{T+1}\langle \mathcal{A}\boldsymbol{x}^{(T+1/2)}, \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}\rangle && (118)\\
&= \sqrt{T+1}\Big(\langle \mathcal{A}(\boldsymbol{x}^{(T+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}\rangle && (119)\\
&+ \langle \mathcal{A}\boldsymbol{x}^{(T+1)}, \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}\rangle\Big) && (120)
\end{aligned}
$$

We shall each (RHS) term (191) individually. In particular, we have:

$$
\begin{aligned}
\sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}\rangle &= \sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^*\rangle\\
&+ \sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^* - \boldsymbol{x}\rangle && (121)
\end{aligned}
$$

Therefore, by applying Cauchy-Shwartz we get:

$$
\begin{aligned}
\sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}\rangle &\leq \sqrt{T+1}\left\|\mathcal{A}\boldsymbol{x}^{(T+1/2)} - \mathcal{A}\boldsymbol{x}^{(T+1)}\right\|_2 && (122)\\
&\left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^*\right\|_2 + \|\boldsymbol{x}^* - \boldsymbol{x}\|_2\right) && (123)\\
&\leq \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2 && (124)\\
&\left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^*\right\|_2 + \|\boldsymbol{x}^* - \boldsymbol{x}\|_2\right) && (125)
\end{aligned}
$$

with (196) being obtained by L- Lipschitz continuity of $\mathcal{A}$. Furthermore, we have:

$$\begin{aligned}\sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}\rangle &\leq \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2^2 \\ &+ \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2\left(\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^*\right\|_2 + \|\boldsymbol{x}^* - \boldsymbol{x}\|_2\right) \quad (126)\end{aligned}$$

■

On the other hand, we have:

$$\begin{aligned}\sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^* - \boldsymbol{x}\rangle &\leq \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(T+1)})\right\|_2\|\boldsymbol{x}^* - \boldsymbol{x}\|_2 && (127) \\ &\leq \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2\|\boldsymbol{x}^* - \boldsymbol{x}\|_2 && (128)\end{aligned}$$

Therefore, summarizing we have for the first term:

$$\begin{aligned}\sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}\rangle &\leq \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^{(T+1/2)}\right\|_2^2 \\ &+ \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2\left(\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\|_2 + 2\|\boldsymbol{x}^* - \boldsymbol{x}\|_2\right) \quad (129)\end{aligned}$$

For the second term we have,

$$\begin{aligned}\sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}\rangle &\leq \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)})\right\|_2\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}\right\|_2 && (130) \\ &\leq \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}\right\|_2 && (131) \\ &\left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^*\right\|_2 + \|\boldsymbol{x}^* - \boldsymbol{x}\|_2\right) && (132) \\ &\leq \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}\right\|_2 && (133) \\ &\left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2 + \left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^*\right\|_2 + \|\boldsymbol{x}^* - \boldsymbol{x}\|_2\right) && (134) \\ &\leq \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}\right\|_2 && (135) \\ &\left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2 + \left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\right\|_2 + \|\boldsymbol{x}^* - \boldsymbol{x}\|_2\right) && (136) \\ &\leq \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}\right\|_2\left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2 + D\right) && (137)\end{aligned}$$

with $D = \left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\right\|_2 + \sup_{\boldsymbol{x}} \|\boldsymbol{x}^* - \boldsymbol{x}\|_2$. Therefore, summarizing we have:

$$\sqrt{T+1}\left\langle \mathcal{A}(\boldsymbol{x}^{(T+1/2)}), \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x} \right\rangle \leq \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^{(T+1/2)}\right\|_2^2$$
$$+2\sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^{(T+1/2)}\right\|_2 D + \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}\right\|_2 \left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2 + D\right) \tag{138}$$

In turn, this yields:

$$\sqrt{T+1}\left\langle \mathcal{A}(\boldsymbol{x}), \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x} \right\rangle \leq \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^{(T+1/2)}\right\|_2^2 + 2\sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^{(T+1/2)}\right\|_2 D$$
$$+ \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}\right\|_2 \left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2 + D\right) \tag{139}$$

and therefore by taking suprema on both sides we get:

$$0 \leq \sqrt{T+1}\,\mathrm{Gap}(\boldsymbol{x}^{(T+1/2)}) \leq \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^{(T+1/2)}\right\|_2^2 + 2\sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^{(T+1/2)}\right\|_2 D$$
$$+ \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}\right\|_2 \left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2 + D\right) \tag{140}$$

and hence the result follows by theorem 17. The second claim is derived directly from theorem 23

Finally for the (OGDA) algorithm, rate of convergence is obtained via the same line of reasoning.

E.3. Adaptive step size

We now move forward to the more demanding case of the adaptive step size policy. Namely, we will examine the last iterate's behaviour of (EG)/ (OGDA) run with the following generic form:

$$\gamma^{(t)} = \left(\gamma^{(0)} + \sum_{\tau=1}^{t-1} \tau \|\boldsymbol{x}^{(\tau+1)} - \boldsymbol{x}^{(\tau+1/2)}\|_2^2\right)^{-q}, \tag{Adapt}$$

where $q > 0$ and $\gamma^{(0)} > 0$. The main challenge of this framework is that (Adapt) does not guarantee a monotonic decrease of the tangent residual sequence $r^{(t)}$. More precisely, we have the following proposition.

Proposition 25 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates of* (EG)/(OGDA) *run with the step size policy* (Adapt)*. Then, the following holds:*

$$\sum_{t=1}^{+\infty} (\gamma^{(t)} r^{(t+1)})^2 < +\infty \tag{141}$$

Proof First note that $\gamma^{(t)}$ is a nonnegative and nonincreasing sequence. Therefore, its limit exists and moreover we have:

$$\lim_t \gamma^{(t)} = \inf_t \gamma^{(t)} \geq 0 \tag{142}$$

Now, by denoting $\gamma_\infty = \lim_t \gamma^{(t)} = \inf_t \gamma^{(t)}$ we distinguish two cases.

Case 1: $\gamma_\infty > 0$**.** Then, note that by the definition of (Adapt) we have:

$$\sum_{t=1}^{T} t\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 = \left(\frac{1}{\gamma^{(T+1)}}\right)^{1/q} - \gamma^{(0)} \tag{143}$$

Therefore, we have:

$$\begin{aligned}\sum_{t=1}^{+\infty} t\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 &= \lim_{T\to+\infty} \sum_{t=1}^{T} t\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\|_2^2 &(144)\\ &= \lim_{T\to+\infty} \left(\frac{1}{\gamma^{(T+1)}}\right)^{1/q} - \gamma^{(0)} &(145)\\ &= \left(\frac{1}{\gamma_\infty}\right)^{1/q} - \gamma^{(0)} &(146)\\ &< +\infty &(147)\end{aligned}$$

with the last strict inequality being obtained by the fact that $\gamma_\infty > 0$. Therefore, by applying theorem 9 for $\boldsymbol{x} = \boldsymbol{x}^*$, being a solution of (VI), and after rearranging and telescoping for $t = 1, \dots T$ we have:

$$\begin{aligned}\min\left\{\frac{1}{4}, \frac{1}{4\gamma^{(0)}L}\right\} \sum_{t=1}^{T} (\gamma^{(t)} r^{(t+1)})^2 \le \frac{1}{2}\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\|_2^2 + \sum_{t=1}^{T} (2\gamma^{(t)} L)^2 \left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2\\ - \sum_{t=1}^{T} \frac{1}{8}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 \quad (148)\end{aligned}$$

which in turn yields:

$$\begin{aligned}\min\left\{\frac{1}{4}, \frac{1}{4\gamma^{(0)}L}\right\} \sum_{t=1}^{T} (\gamma^{(t)} r^{(t+1)})^2 &\le \frac{1}{2}\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\|_2^2 + \sum_{t=1}^{T} (2\gamma^{(t)} L)^2 \left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 &(149)\\ &\le \frac{1}{2}\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\|_2^2 + 4L^2(\gamma^{(0)})^{2/q} \sum_{t=1}^{T} \left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 &(150)\\ &\le \frac{1}{2}\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\|_2^2 + 4L^2(\gamma^{(0)})^{2/q} \sum_{t=1}^{T} t\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 &(151)\\ &\le \frac{1}{2}\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\|_2^2 + 4L^2(\gamma^{(0)})^{2/q} \sum_{t=1}^{+\infty} t\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1/2)}\right\|_2^2 &(152)\end{aligned}$$

which finally yields:

$$\min\left\{\frac{1}{4},\frac{1}{4\gamma^{(0)}L}\right\}\sum_{t=1}^{T}(\gamma^{(t)}r^{(t+1)})^2 \leq \frac{1}{2}\|\boldsymbol{x}^{(1)}-\boldsymbol{x}^*\|_2^2 + 4L^2(\gamma^{(0)})^{2/q}\left(\left(\frac{1}{\gamma_\infty}\right)^{1/q}-\gamma^{(0)}\right) \tag{153}$$

Hence, the result for the first case follows by taking limits on both sides.

Case 2: $\gamma_\infty = 0$ Now we move to the case where the step-size policy may vanish. This in turn yields that there exists some $t_0 \in \mathbb{N}$ such that:

$$\gamma^{(t)} \leq \frac{1}{\sqrt{32}L} \quad \text{for all} \quad t > t_0 \tag{154}$$

Then, again by invoking theorem 9 in the same spirit as in case 1, we have:

$$\min\left\{\frac{1}{4},\frac{1}{4\gamma^{(0)}L}\right\}\sum_{t=1}^{T}(\gamma^{(t)}r^{(t+1)})^2 \leq \frac{1}{2}\|\boldsymbol{x}^{(1)}-\boldsymbol{x}^*\|_2^2 + \sum_{t=1}^{T}(2\gamma^{(t)}L)^2\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2 \\ - \sum_{t=1}^{T}\frac{1}{8}\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2 \tag{155}$$

which in turn for T large enough:

$$\min\left\{\frac{1}{4},\frac{1}{4\gamma^{(0)}L}\right\}\sum_{t=1}^{T}(\gamma^{(t)}r^{(t+1)})^2 \leq \frac{1}{2}\|\boldsymbol{x}^{(1)}-\boldsymbol{x}^*\|_2^2 + \sum_{t=1}^{T}\left((2\gamma^{(t)}L)^2-\frac{1}{8}\right)\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2 \tag{156}$$

$$= \frac{1}{2}\|\boldsymbol{x}^{(1)}-\boldsymbol{x}^*\|_2^2 + \sum_{t=1}^{t_0}\left((2\gamma^{(t)}L)^2-\frac{1}{8}\right)\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2 \tag{157}$$

$$+ \sum_{t=t_0+1}^{T}\left((2\gamma^{(t)}L)^2-\frac{1}{8}\right)\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2 \tag{158}$$

To that end, by invoking (154) we have that:

$$\min\left\{\frac{1}{4},\frac{1}{4\gamma^{(0)}L}\right\}\sum_{t=1}^{t_0}(\gamma^{(t)}r^{(t+1)})^2 \leq \frac{1}{2}\|\boldsymbol{x}^{(1)}-\boldsymbol{x}^*\|_2^2 + \sum_{t=1}^{T}\left((2\gamma^{(t)}L)^2-\frac{1}{8}\right)\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2 \tag{159}$$

Finally, the result for the second case follows by taking limits on both sides. In order to establish the respective result for (OGDA) the same arguments apply by invoking the template inequality theorem 10 for $\boldsymbol{x} = \boldsymbol{x}^*$. ■

Moving forward, we show the following crucial stepping stones. In particular, we have the following proposition.

Proposition 26 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates of* (EG)/(OGDA) *run with the adaptive step size policy* (Adapt)*. Then, the following hold:*

1. *The adaptive step size policy $\gamma^{(t)}$ is bounded away from zero, i.e.*

$$\lim_{t\to+\infty}\gamma^{(t)}=\inf_t\gamma^{(t)}=\gamma_\infty>0 \tag{160}$$

2. *The sequence $(t\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2)_t$ is summable, i.e.*

$$\sum_{t=1}^{+\infty}t\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2<+\infty \tag{161}$$

Proof For the first claim, assume that $\gamma_\infty=0$. Then, by invoking theorem 12 and rearranging, we have:

$$\frac{1}{2}t\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2\leq t(\gamma^{(t)}r^{(t)})^2-(t+1)(\gamma^{(t+1)}r^{(t+1)})^2+(\gamma^{(t+1)}r^{(t+1)})^2-\frac{1}{2}t\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2 \\ +t(\gamma^{(t)}L)^2\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2 \tag{162}$$

Moreover, since we assumed that $\gamma_\infty=0$, there exists some $t_0\in\mathbb{N}$ such that:

$$\gamma^{(t)}\leq\frac{1}{\sqrt{2}L}\quad\text{for all}\quad t>t_0 \tag{163}$$

So, by telescoping (162) for T large enough, we have:

$$\frac{1}{2}\sum_{t=1}^{T}t\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2\leq(\gamma^{(1)}r^{(1)})^2+\sum_{t=1}^{T}(\gamma^{(t+1)}r^{(t+1)})^2+\sum_{t=1}^{T}\left((\gamma^{(t)}L)^2-\frac{1}{2}\right)t\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2 \tag{164}$$

and therefore by dividing the sum we get:

$$\frac{1}{2}\sum_{t=1}^{T}t\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2\leq(\gamma^{(1)}r^{(1)})^2+\sum_{t=1}^{T}(\gamma^{(t+1)}r^{(t+1)})^2 \\ +\sum_{t=1}^{t_0}\left((\gamma^{(t)}L)^2-\frac{1}{2}\right)t\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2+\sum_{t=t_0+1}^{T}\left((\gamma^{(t)}L)^2-\frac{1}{2}\right)t\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2 \tag{165}$$

which combined with (163) we have:

$$\frac{1}{2}\sum_{t=1}^{T}t\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2\leq(\gamma^{(1)}r^{(1)})^2+\sum_{t=1}^{T}(\gamma^{(t+1)}r^{(t+1)})^2 \tag{166}$$

$$+\sum_{t=1}^{t_0}\left((\gamma^{(t)}L)^2-\frac{1}{2}\right)t\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2 \tag{167}$$

$$<+\infty \tag{168}$$

which the last strict inequality being obtained by theorem 25. On the other hand, by the definition of $\gamma^{(t)}$ we have that:

$$\frac{1}{2}\sum_{t=1}^{T} t\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2=\left(\frac{1}{\gamma^{(T+1)}}\right)^{1/q}-\gamma^{(0)}\to+\infty \tag{169}$$

since we assumed that $\gamma^{(t)}\to 0$ and is non-negative; yielding an contradiction. Therefore, we have $\gamma_\infty>0$ and hence the result follows. Finally, regarding the (OGDA) the respective result is obtained by the same reasoning and invoking theorem 14 and theorem 10. ■

Next step would be to establish the limit existence of the tangent residual. In particular, we have:

Proposition 27 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates of* (EG)/(OGDA) *run with the adaptive step size policy* (Adapt)*. Then, the sequence $t(\gamma^{(t)}r^{(t)})^2$ has a limit.*

Proof By invoking theorem 12 we have:

$$\begin{aligned}(t+1)(\gamma^{(t+1)}r^{(t+1)})^2 &\le t(\gamma^{(t)}r^{(t)})^2+(\gamma^{(t+1)}r^{(t+1)})^2+\left((L\gamma^{(t)})^2-1\right)t\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2 &(170)\\ &\le t(\gamma^{(t)}r^{(t)})^2+(\gamma^{(t+1)}r^{(t+1)})^2+(L\gamma^{(t)})^2t\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2 &(171)\\ &\le t(\gamma^{(t)}r^{(t)})^2+(\gamma^{(t+1)}r^{(t+1)})^2+\gamma^{(0)}L^2t\left\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\right\|_2^2 &(172)\end{aligned}$$

So, since $(\gamma^{(t+1)}r^{(t+1)})^2, \gamma^{(0)}L^2t\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1/2)}\|_2^2$ are summable due to theorem 25 and theorem 26, by applying theorem 31, we directly get the result. ■

Proposition 28 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates of* (EG) *run with the adaptive step size policy* (Adapt)*. Then, we have:*

$$\lim_{T\to+\infty}(Tr^{(T)})^2=0 \tag{173}$$

Proof First, note that by invoking theorem 25 we have:

$$\sum_{t=1}^{+\infty}\frac{1}{t}t(\gamma^{(t)}r^{(t)})^2<+\infty \tag{174}$$

By working in the same spirit as in , we readily get that:

$$\lim_{T\to+\infty}T(\gamma^{(T)}r^{(T)})^2=0 \tag{175}$$

Therefore, we have:

$$\begin{aligned}\lim_{T\to+\infty}T(r^{(T)})^2&=\lim_{T\to+\infty}\frac{T(\gamma^{(T)}r^{(T)})^2}{(\gamma^{(T)})^2} &(176)\\ &=\frac{1}{\gamma_\infty^2}0 &(177)\\ &=0 &(178)\end{aligned}$$

and hence the result follows. Finally, regarding the (OGDA) is directly obtained by the same reasoning by invoking theorem 14. ■

Remark 29 *Following the same spirit as in theorem 23 we readily get:*

$$\liminf_{T\to+\infty}\left(T\log(T+1)(r^{(T)})^2\right)=0 \tag{179}$$

Finally, we are in the position to shown our full result; namely the last iterate of (EG) run with the adaptive step size policy (Adapt). In particular, we have the following theorem.

Theorem 30 *Assume that $\boldsymbol{x}^{(t+1/2)}, \boldsymbol{x}^{(t)}$ are the iterates of (EG)/(OGDA) run with the adaptive step size policy (Adapt). Moreover, assume that $\mathcal{C}$ is a non-empty, convex and compact subset of $\mathcal{X}$ which contains a neighborhood of a solution of (VI). Then, the following hold:*

$$\operatorname{Gap}_{\mathcal{C}}(\boldsymbol{x}^{(T+1)})=o(1/\sqrt{T}) \ \ \textit{and} \ \ \operatorname{Gap}_{\mathcal{C}}(\boldsymbol{x}^{(T+1/2)})=o(1/\sqrt{T}) \tag{180}$$

Moreover, we have:

$$\liminf_{T\to+\infty}\sqrt{(T+1)\log(T+1)}\operatorname{Gap}_{\mathcal{C}}(\boldsymbol{x}^{(T+1)})=\liminf_{T\to+\infty}\sqrt{(T+1)\log(T+1)}\operatorname{Gap}_{\mathcal{C}}(\boldsymbol{x}^{(T+1/2)})=0 \tag{181}$$

Proof Working in the same spirit with theorem 24 we have:

$$\begin{aligned}
\sqrt{T+1}\langle\mathcal{A}(\boldsymbol{x}),\boldsymbol{x}^{(T+1)}-\boldsymbol{x}\rangle &\le \sqrt{T+1}\langle\mathcal{A}\boldsymbol{x}^{(T+1)},\boldsymbol{x}^{(T+1)}-\boldsymbol{x}\rangle && (182)\\
&\le \sqrt{T+1}\langle\mathcal{A}(\boldsymbol{x}^{(T+1))}+\boldsymbol{c}^{(T+1)},\boldsymbol{x}^{(T+1)}-\boldsymbol{x}\rangle && (183)\\
&= \sqrt{T+1}\Big(\langle\mathcal{A}(\boldsymbol{x}^{(T+1)})+\boldsymbol{c}^{(T+1)},\boldsymbol{x}^{(T+1)}-\boldsymbol{x}^*\rangle && (184)\\
&+\langle\mathcal{A}(\boldsymbol{x}^{(T+1)})+\boldsymbol{c}^{(T+1)},\boldsymbol{x}^*-\boldsymbol{x}\rangle\Big) && (185)\\
&\le \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)})+\boldsymbol{c}^{(T+1)}\right\|_2\left(\left\|\boldsymbol{x}^{(T+1)}-\boldsymbol{x}^*\right\|_2+\|\boldsymbol{x}^*-\boldsymbol{x}\|_2\right) && (186)\\
&\le \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)})+\boldsymbol{c}^{(T+1)}\right\|_2\left(\left\|\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(*)}\right\|_2+\|\boldsymbol{x}^*-\boldsymbol{x}\|_2\right) && (187)
\end{aligned}$$

which finally yields:

$$\sqrt{T+1}\operatorname{Gap}(\boldsymbol{x}^{(T+1)})\le\sqrt{T+1}r^{(T+1)}D \tag{188}$$

and hence the first claim follows by invoking . For the second claim, again by working in the same spirit as in theorem 24 we have:

$$\begin{aligned}
\sqrt{T+1}\langle\mathcal{A}(\boldsymbol{x}),\boldsymbol{x}^{(T+1/2)}-\boldsymbol{x}\rangle &\le \sqrt{T+1}\langle\mathcal{A}(\boldsymbol{x}^{(T+1/2)}),\boldsymbol{x}^{(T+1/2)}-\boldsymbol{x}\rangle && (189)\\
&= \sqrt{T+1}\Big(\langle\mathcal{A}(\boldsymbol{x}^{(T+1/2)})-\mathcal{A}(\boldsymbol{x}^{(T+1)}),\boldsymbol{x}^{(T+1/2)}-\boldsymbol{x}\rangle && (190)\\
&+\langle\mathcal{A}(\boldsymbol{x}^{(T+1)}),\boldsymbol{x}^{(T+1/2)}-\boldsymbol{x}\rangle\Big) && (191)
\end{aligned}$$

We shall bound each (RHS) term (191) individually. In particular, we have:

$$\begin{aligned}\sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1/2)})-\mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^{(T+1/2)}-\boldsymbol{x}\rangle &= \sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1/2)})-\mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^{(T+1/2)}-\boldsymbol{x}^*\rangle \\ &\quad + \sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^* - \boldsymbol{x}\rangle \end{aligned} \tag{192}$$

Therefore, by applying Cauchy-Shwartz we get:

$$\sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}\rangle \leq \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(T+1)})\right\|_2 \tag{193}$$

$$\left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^*\right\|_2 + \|\boldsymbol{x}^* - \boldsymbol{x}\|_2\right) \tag{194}$$

$$\leq \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2 \tag{195}$$

$$\left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^*\right\|_2 + \|\boldsymbol{x}^* - \boldsymbol{x}\|_2\right) \tag{196}$$

with (196) being obtained by L- Lipschitz continuity of $\mathcal{A}$. Furthermore, we have:

$$\begin{aligned}\sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}\rangle &\leq \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2^2 \\ &+ \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2\left(\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^*\right\|_2 + \|\boldsymbol{x}^* - \boldsymbol{x}\|_2\right)\end{aligned} \tag{197}$$

On the other hand, we have:

$$\sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^* - \boldsymbol{x}\rangle \leq \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(T+1)})\right\|_2\|\boldsymbol{x}^* - \boldsymbol{x}\|_2 \tag{198}$$

$$\leq \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2\|\boldsymbol{x}^* - \boldsymbol{x}\|_2 \tag{199}$$

Therefore, summarizing we have for the first term:

$$\begin{aligned}\sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1/2)}) - \mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}\rangle &\leq \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^{(T+1/2)}\right\|_2^2 \\ &+ \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2\left(\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\|_2 + 2\|\boldsymbol{x}^* - \boldsymbol{x}\|_2\right)\end{aligned} \tag{200}$$

For the second term we have,

$$\begin{aligned}
\sqrt{T+1}\langle \mathcal{A}(\boldsymbol{x}^{(T+1)}), \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}\rangle &\leq \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)})\right\|_2 \left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}\right\|_2 && (201)\\
&\leq \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}\right\|_2 \left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^*\right\|_2 + \|\boldsymbol{x}^* - \boldsymbol{x}\|_2\right) && (202)\\
&\leq \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}\right\|_2 \left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2\right. && (203)\\
&\quad \left. + \left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^*\right\|_2 + \|\boldsymbol{x}^* - \boldsymbol{x}\|_2\right) && (204)\\
&\leq \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}\right\|_2 \left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2\right. && (205)\\
&\quad \left. + \left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\right\|_2 + \|\boldsymbol{x}^* - \boldsymbol{x}\|_2\right) && (206)\\
&\leq \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}\right\|_2 \left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2 + D\right) && (207)
\end{aligned}$$

with $D = \left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\right\|_2 + \sup_{\boldsymbol{x}}\|\boldsymbol{x}^* - \boldsymbol{x}\|_2$. Therefore, summarizing we have:

$$\begin{aligned}
&\sqrt{T+1}\left\langle \mathcal{A}(\boldsymbol{x}^{(T+1/2)}), \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}\right\rangle \leq \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^{(T+1/2)}\right\|_2^2\\
&+2\sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^{(T+1/2)}\right\|_2 D + \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}\right\|_2 \left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2 + D\right) \qquad (208)
\end{aligned}$$

In turn, this yields:

$$\begin{aligned}
&\sqrt{T+1}\left\langle \mathcal{A}(\boldsymbol{x}), \boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}\right\rangle \leq \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^{(T+1/2)}\right\|_2^2\\
&+2\sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^{(T+1/2)}\right\|_2 D + \sqrt{T+1}\left\|\mathcal{A}(\boldsymbol{x}^{(T+1)}) + \boldsymbol{c}^{(T+1)}\right\|_2 \left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2 + D\right) \qquad (209)
\end{aligned}$$

and therefore by taking suprema on both sides we get:

$$\begin{aligned}
&0 \leq \sqrt{T+1}\,\mathrm{Gap}(\boldsymbol{x}^{(T+1/2)}) \leq \sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^{(T+1/2)}\right\|_2^2\\
&+2\sqrt{T+1}L\left\|\boldsymbol{x}^{(T+1)} - \boldsymbol{x}^{(T+1/2)}\right\|_2 D + \sqrt{T+1}\left\|\mathcal{A}\boldsymbol{x}^{(T+1)} + \boldsymbol{c}^{(T+1)}\right\|_2 \left(\left\|\boldsymbol{x}^{(T+1/2)} - \boldsymbol{x}^{(T+1)}\right\|_2 + D\right) \qquad (210)
\end{aligned}$$

and hence the result follows by theorem 28. The second claim is directly obtained by theorem 29. ■

Appendix F. Auxiliary lemma

In this section, we briefly present the notion of *quasi-Féjer monotone* sequences. More precisely, sequences which satisfy the following recursion:

$$\alpha^{(t+1)} \leq \alpha^{(t)} - \beta^{(t)} + \delta^{(t)} \ \text{ for all } \ t = 1, 2, \ldots \tag{211}$$

where $\alpha_t, \beta_t, \delta_t$ are non-negative sequences. In particular, we have the following proposition.

Proposition 31 *Let $\chi \in (0, 1]$, $(\alpha^{(t)})_{t\in\mathcal{N}}$, $(\beta^{(t)})_{t\in\mathcal{N}}$ non-negative sequences and $(\delta^{(t)})_{t\in\mathcal{N}}$ such that $t = 1, 2, \ldots$:*

$$\alpha^{(t+1)} \leq \chi\alpha^{(t)} - \beta^{(t)} + \delta^{(t)} \tag{212}$$

Then, $\alpha^{(t)}$ converges.

Proof First, one shows that $\alpha^{(t\in\mathcal{N})}$ is a bounded sequence. Indeed, one can derive directly that:

$$\alpha^{(t+1)} \leq \chi^{t+1}\alpha^{(0)} + \sum_{k=0}^{t} \chi^{t-k}\delta^{(k)} \tag{213}$$

Hence, $(\alpha^{(t)})_{t\in\mathcal{N}}$ lies in $[0, \alpha^{(0)} + \delta]$, with $\delta = \sum_{t=0}^{+\infty} \delta^{(t)}$. Now, one is able to extract a convergent subsequence $(\alpha^{(k_t)})_{t\in\mathcal{N}}$, let say $\lim_{t\to+\infty} \alpha^{(k_t)} = \alpha \in [0, \alpha_0 + \delta]$ and fix $\varepsilon > 0$. Then, one can find some t_0 such that $\alpha^{(k_{t_0})} - \alpha < \frac{\varepsilon}{2}$ and $\sum_{m>t_{k_{t_0}}} \delta^{(m)} < \frac{\varepsilon}{2}$. That said, we have:

$$0 \leq \alpha^{(t)} \leq \alpha^{(k_{t_0})} + \sum_{m>t_{k_{t_0}}} \delta^{(m)} < \frac{\varepsilon}{2} + \alpha + \frac{\varepsilon}{2} = \alpha + \varepsilon \tag{214}$$

Hence, $\limsup_t \alpha^{(t)} \leq \liminf_t \alpha^{(t)} + \varepsilon$. Since, ε is chosen arbitrarily we may let $\varepsilon \to 0$ and hence the result follows. The case of (OGDA) is derived by invoking the same reasoning. ■

Appendix G. Symbolic verification of theorem 11 and theorem 13

We used SymPy, a symbolic algebra system to verify theorem 11.

```
from sympy import *

z0, z1, z2, Fz0, Fz1, Fz2, c0, c2, L, g
= symbols(’z0 z1 z2 Fz0 Fz1 Fz2 c0 c2 L g’)

expression_1 = g**2*(Fz0+ c0)**2 - g**2*(Fz2+ c2)**2
expression_2 = (-1)*g**2*(L**2*(z1 - z2)**2 - (Fz1 - Fz2)**2)
expression_3 = (-2)*g*(Fz2 - Fz0)*(z2 - z0)
expression_4 = (-2)*(z0 - g*Fz0 - z1)*(z1 - z2)
expression_5 = (-2)*(z0 - g*Fz1 - z2)*(z2 - z0)
expression_6 = (-2)*g*c0*(z0 - z1)
expression_7 = (-2)*(g*c2 + g*Fz2)*(z0 - g*Fz1 - z2)
```

```
expression_8 = (-2)*(g*c2 + g*Fz2)*(-g*c2)
expression_9 = (g*Fz0 + g*c0 - z0 + z1)**2
expression_10 = (g*Fz1 + g*c2 - z0 + z2)**2


LHS = (expression_1 + expression_2 + expression_3 + expression_4 +
       expression_5 + expression_6 + expression_7 + expression_8)
RHS = expression_9 + expression_10

print(simplify(LHS-RHS))
```

Moreover, we additionally used SymPy for verifying theorem 13

```
from sympy import *
z1, z0, w0, w1, Fz0, Fz1, Fw0, Fw1, c0, c1
=symbols(’z1 z0 w0 w1 Fz0 Fz1 Fw0 Fw1 c0 c1’)

# Expression (47)
expression_1 = ((Fz0 + c0)**2 + (Fz0 - Fw0)**2 ) -
((Fz1 + c1)**2 + (Fz1 - Fw1)**2)

# LHS of Inequality (48)
expression_2 = (-2)*(Fz1 - Fz0)*(z1 - z0)

# LHS of Inequality (49)
expression_3 = (-2)*(0.25*(z1 - w1)**2 - (Fz1 - Fw1)**2)

# LHS of Inequality (50)
expression_4 = (-1)*(z0 - Fw0 - w1)*(w1 - z1)

# LHS of Inequality (51)
expression_5 = (-2)*(z0 - Fw1 - z1)*(z1 - z0)

# LHS of Inequality (52)
expression_6 = (-1)*c0*(z0 - w1)

# LHS of Inequality (53)
expression_7 = (-1)*c0*(z0 - z1)

# LHS of Inequality (54)
expression_8 = (-2)*(c1 + Fz1)*(z0 - Fw1 - z1)

# LHS of Inequality (55)
expression_9 = (-2)*(c1 + Fz1)*(-c1)

# LHS of Inequality (56)
```

```
expression_10 = ((w1 - z1)/2 + Fw0 - Fz0)**2

# LHS of Inequality (57)
expression_11 = (Fz0 + c0 - z0 + (w1 + z1)/2)**2

# LHS of Inequality (58)
expression_12 = (z0 - Fw1 - z1 - c1)**2

# LHS of the identity
LHS = ( expression_1 + expression_2 + expression_4 + expression_5
      + expression_6 + expression_7 + expression_8 + expression_9)

# RHS of the identity
RHS = expression_10 + expression_11 + expression_12

P = LHS - RHS
Q=simplify(P)
print(Q)
```