On Scaled Methods for Saddle Point Problems

Aleksandr Beznosikov
Moscow Institute of Physics and Technology
Ivannikov Institute for System Programming
Innopolis University

Aibek Alanov
HSE University
Artificial Intelligence Research Institute

Martin Takáč
Mohamed bin Zayed University of Artificial Intelligence

Alexander Gasnikov
Innopolis University
Moscow Institute of Physics and Technology
Ivannikov Institute for System Programming

Abstract

Methods with adaptive scaling of different features play a key role in solving saddle point problems, primarily due to Adam's popularity for solving adversarial machine learning problems, including GANS training. This paper carries out a theoretical analysis of the following scaling techniques for solving SPPs: the well-known Adam and RmsProp scaling and the newer AdaHessian and OASIS based on Hutchison approximation. We use the Extra Gradient and its improved version with negative momentum as the basic method. Experimental studies on GANs show good applicability not only for Adam, but also for other less popular methods.

1 Introduction

In this paper, we focus on the saddle point problem (SPP):

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} f(x, y). \tag{1}$$

It has been of concern to the applied mathematics community for a long time. For example, SPPs arise in game theory, optimal control and equilibrium search in economics (Neumann & Morgenstern, 1944; Harker & Pang., 1990; Facchinei & Pang, 2003). The problem (1) is now even more widespread due to the current interest in solving adversarial formulations in machine learning. One of the brightest representatives of such tasks is Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). The main and most popular method of training GANs is Adam (Kingma & Ba, 2015).

Since its appearance, Adam has become a cornerstone learning method not only for GANs but also, for example, for NLPs (Lan et al., 2020) or models with attention (Zhang et al., 2020). There are two features that distinguish Adam from SGD: a heavy-ball momentum as well as a gradient scaling. If momentum is more or less the standard acceleration approach, then scaling is precisely the key of Adam that allows it to be robust. Scaling is often also referred to as adaptivity, but it is different from classical adaptivity, where we simply recalculate the stepsize using information from previous iterations. Scaling modifies the entire gradient by multiplying each component by a different value. Adam was not the first method to use scaling (Duchi et al., 2011; Zeiler, 2012), but the Adam update is the most popular amongst the community.

While Adam was gaining its practical relevance, it was also explored from a theoretical point of view for minimization problems (Reddi et al., 2019; Défossez et al., 2020). Moreover, Adam and its modifications appear in many theoretical papers on SPPs and GANs training. The most famous of these works follow the same pattern (Daskalakis et al., 2017; Gidel et al., 2018; Mertikopoulos et al., 2018; Chvdarova et al., 2019; Liang & Stokes, 2019; Peng et al., 2020). The authors propose studies for methods without scaling, and then without theoretical justifications transfer the new features to Adam and obtain a new more powerful method. For example, (Gidel et al., 2018) adapts the classical and optimal method for SPPs, the Extra Gradient method, and obtains Extra Adam. But while Adam is a good framework into which some theoretical hints and intuitions can be inserted, up to the current time, other methods with scaling are hardly mentioned in the SPPs literature from a theoretical point

of view. Recently, papers on scaling algorithms that provide guarantees of convergence have begun to appear (Liu et al., 2019; Dou & Li, 2021; Barazandeh et al., 2022). Despite the recent progress, there are still many disadvantages to these methods that should be addressed.

1.1 OUR CONTRIBUTION AND RELATED WORKS

Eight methods with theory. In this paper, we present eight methods with scaling for solving (1). In fact, we give two general methods in which different types of scaling can be substituted. We analyze four popular types of scaling in this paper, although the universality of the analysis allows exploring other scaling approaches. Theoretical convergence guarantees for these methods are provided in the strongly convex–strongly concave, convex–concave, and non-convex–non-concave (Minty) cases of the problem (1).

Extra Gradient. Our first method (Algorithm 1) is based on the Extra Gradient method (Korpelevich, 1976; Nemirovski, 2004; Juditsky et al., 2008), an optimal and frequently used method for SPPs. The key feature of this method is the so-called extrapolation/extra step. We add scaling to this method. Unlike most of the works (Daskalakis et al., 2017; Gidel et al., 2018; Mertikopoulos et al., 2018; Chavdarova et al., 2019; Liang & Stokes, 2019; Peng et al., 2020) which connect Extra Gradient and its modifications with Adam-based methods, we use the same scaling in both Extra Gradient steps and obtain to some extent a method that has not been encountered before. Consequently, the theoretical analysis for this kind of method is done for the first time and does not coincide, for example, with the analysis from (Liu et al., 2019; Dou & Li, 2021).

Single call with momentum. Our second algorithm is an improved version of the first one.

Primarily, we solve the global problem of the Extra Gradient method – the double calling of gradients at each iteration. We use a single call approach that allows us to compute the gradient once. Such methods without scaling were explored e.g., in (Popov, 1980; Mokhtari et al., 2020; Hsieh et al., 2019).

Secondly, we add momentum. Methods with momentum have been partially investigated for SPPs. In paper (Gidel et al., 2019), the authors show that for SPPs (including training GANs), one should use negative momentum because classical positive (heavy-ball) momentum can be not only useless but have a negative effect. One can also note that negative momentum was used in works (Alacaoglu & Malitsky, 2021) and (Kovalev et al., 2022) to obtain optimal stochastic variance reduced methods for SPPs.

Scalings. As noted above, we consider different types of scaling approaches: two types of Adam-based scaling (Adam (Kingma & Ba, 2015) itself, as well as RMSProp (Tieleman et al., 2012)) and two types of rather new Hutchinson's approximation (Bekas et al., 2007) scalings (AdaHessian (Yao et al., 2020) and OASIS (Jahani et al., 2021)). None of these methods have been analyzed theory-wise for SPPs before. Moreover, methods based on Hutchinson's approximation are found for the first time in an application to SPPs. Note that AdaHessian (Yao et al., 2020) does not provide convergence analysis at all (not even for minimization problems).

Table 1: Comparison of scaled methods with theoretical analysis for SPPs.

Method	Reference	Scaling	Single call?	Momentum	Assumptions
OAdaGrad	(Liu et al., 2019)	AdaGrad	yes	no	bounded gradients
Extra Gradient AMSGrad	(Dou & Li, 2021)	Adam-based	no	positive	bounded gradients
AMMO	(Barazandeh et al., 2022)	Adam-based	no	positive	no theory [1]
Extra Adam/RMSProp	Ours	Adam-based	no	no	bounded gradients
Extra AdaHessian/OASIS	Ours	Hutchinson	no	no	
SCM Extra Adam/RMSProp	Ours	Adam-based	yes	negative	bounded gradients
SCM Extra AdaHessian/OASIS	Ours	Hutchinson	yes	negative	

[1] There is a theory in the paper, but it raises serious concerns.

Main competitors. One can highlight three main papers that also deal with the theoretical analysis of methods with scaling for SPPs. We give a comparison in Table 1.

The work (Liu et al., 2019) presents an analysis of the Optimistic version of AdaGrad. The Optimistic approach is a single call, making the method more robust and faster in computation. Meanwhile, the authors do not add momentum to this method. The authors also need gradient boundedness as an assumption in the theoretical analysis; we do not need to make such an additional assumption for Hutchinson scaling methods.

The authors of (Dou & Li, 2021) propose an analysis for the Adam-Based (but not Adam) method. Their method is based on the Extra Gradient method with the AMSGrad update (Reddi et al., 2019). Unlike some of our methods and the method from paper (Liu et al., 2019), this method is not single-call, which in practice can slow down the running time by almost a factor of two. In this paper, the boundedness of the gradients is assumed.

Finally, both papers (Liu et al., 2019) and (Dou & Li, 2021) give convergence only in the non-convex case. We give an analysis for the strongly-convex and convex cases as well. It seems that this is the way to see a more complete picture of how the methods work. This is due to the fact that any guarantees in the non-convex case (not only for SPPs but also for minimization problems) are just convergence to some (most likely the nearest) stationary point, but in practice, for neural network problems, we converge not just to some stationary point, but probably to a good neighborhood of global solution. The practical behavior of the methods lies somewhere between the general non-convex and convex cases.

The paper (Barazandeh et al., 2022) presents several Adam-based methods, but its theoretical foundation is disturbing. We found no proofs in the paper (the authors refer to the arxiv version, which does not exist at the current moment). Moreover, the theorems postulate convergence but not the rate of convergence.

Practical performance. In practice, we explore four methods of scaling (RMSProp, Adam, AdaHessian, OASIS) for training GANs. We test how these methods are affected by the extra step technique that arises from the theory. We further investigate the effect of negative momentum on final learning results.

2 PROBLEM SETUP AND ASSUMPTION

We write $\langle x, y \rangle = \sum_{i=1}^{d} x_i y_i$ to denote the standard inner product of vectors $x, y \in \mathbb{R}^d$, where x_i corresponds to the i-th component of x in the standard basis in $\mathbb{R}^d$. This induces the standard ℓ_2-norm in $\mathbb{R}^d$ in the following way: $\|x\| = \sqrt{\langle x, x \rangle}$. For a given positive definite matrix $D \in \mathcal{S}_{++}^d$, the weighted Euclidean norm is defined to be $\|x\|_D^2 = \langle x, Dx \rangle$, where $x \in \mathbb{R}^d$. We also introduce $\lambda_{\max}(D)$ as the maximum eigenvalue of D, $\lambda_{\min}(D)$ as the minimum eigenvalue of D. Using this notation, for a given matrix $A \in \mathbb{R}^{d \times d}$, we define the spectral norm as $\|A\| = \sqrt{\lambda_{\max}(A^T A)}$, and the infinity norm as $\|A\|_\infty = \max_{1 \le i \le d} \sum_{j=1}^{d} |a_{ij}|$, where $\{a_{ij}\}_{i=1, j=1}^{n, n}$ are elements of A. The symbol $\odot$ denotes the component-wise product between two vectors, and $\mathtt{diag}(x)$ denotes the $d \times d$ diagonal matrix whose diagonal entries are the components of the vector $x \in \mathbb{R}^d$.

We consider the problem (1) and assume that for the function $f(x, y)$ we have only access to stochastic oracle $f(x, y, \xi)$. For example, ξ could be the number of the batch for which the loss is computed. Additionally, only stochastic gradients are available $\nabla_x f(x, y, \xi)$, $\nabla_y f(x, y, \xi)$. Next, we list the main assumptions on (1) and stochastic gradients.

Assumption 1 (Convexity-concavity). We consider three cases of convexity-concavity of (1).

(SC) Strong convexity– strong concavity. The function f is μ-strongly convex–strongly concave, i.e. for all $x_1, x_2 \in \mathbb{R}^{d_x}$ and $y_1, y_2 \in \mathbb{R}^{d_y}$ we have

$$\langle \nabla_x f(x_1, y_1) - \nabla_x f(x_2, y_2), x_1 - x_2 \rangle - \langle \nabla_y f(x_1, y_1) - \nabla_y f(x_2, y_2), y_1 - y_2 \rangle \ge \mu \left(\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2 \right). \tag{2}$$

(C) Convexity–concavity. The function f is convex–concave, i.e. for all $x_1, x_2 \in \mathbb{R}^{d_x}$ and $y_1, y_2 \in \mathbb{R}^{d_y}$ we have

$$\langle \nabla_x f(x_1, y_1) - \nabla_x f(x_2, y_2), x_1 - x_2 \rangle - \langle \nabla_y f(x_1, y_1) - \nabla_y f(x_2, y_2), y_1 - y_2 \rangle \ge 0. \tag{3}$$

(NC) Non-convexity–non-concavity (Minty). The function f satisfies the Minty condition, if and only if there exists $x^* \in \mathbb{R}^{d_x}$ and $y^* \in \mathbb{R}^{d_y}$ such that for all $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}^{d_y}$ we have

$$\langle \nabla_x f(x, y), x - x^* \rangle - \langle \nabla_y f(x, y), y - y^* \rangle \ge 0. \tag{4}$$

The last assumption is not a general non-convexity–non-concavity, but it is firmly associated and regarded in the literature as a good relaxation of non-convexity–non-concavity. (4) can be found

under the names of the Minty or variational stability condition (Minty, 1962; Liu et al., 2019; Hsieh et al., 2020; Dou & Li, 2021; Diakonikolas et al., 2021).

Assumption 2 (*L*-smoothness). For any ξ the function $f(x, y, \xi)$ is L-smooth, i.e. for all $x_1, x_2 \in \mathbb{R}^{d_x}$ and $y_1, y_2 \in \mathbb{R}^{d_y}$ we have

$$\|\nabla_x f(x_1, y_1, \xi) - \nabla_x f(x_2, y_2, \xi)\|^2 + \|\nabla_y f(x_1, y_1, \xi) - \nabla_y f(x_2, y_2, \xi)\|^2$$
$$\leq L^2 \left(\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2 \right). \tag{5}$$

Assumption 3 (Bounded variance). Stochastic gradients of $\nabla_x f(x, y, \xi)$, $\nabla_y f(x, y, \xi)$ are unbiased and their variance are bounded, i.e. there exists $\sigma^2 \geq 0$ such that for all $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}^{d_y}$ we have

$$\mathbb{E}\left[\nabla_x f(x, y, \xi)\right] = \nabla_x f(x, y), \quad \mathbb{E}\left[\nabla_y f(x, y, \xi)\right] = \nabla_y f(x, y),$$

$$\mathbb{E}\left[\|\nabla_x f(x, y, \xi) - \nabla_x f(x, y)\|^2 + \|\nabla_y f(x, y, \xi) - \nabla_y f(x, y)\|^2\right] \leq \frac{\sigma^2}{b}, \tag{6}$$

where the parameter b stands for the size of the batch in the stochastic gradients.

3 SCALED METHODS

Unlike SGD-type methods that use a stochastic gradient g_t for an update, scaled methods additionally compute a so-called preconditioning matrix $\hat{D}_t$ and produce a scaled gradient $\hat{g}_t = \hat{D}_t^{-1} g_t$. For SPPs, obtaining preconditioning matrices can be written as follows:

$$(D_t^x)^2 = \beta_t (D_{t-1}^x)^2 + (1 - \beta_t)(H_t^x)^2, \quad (D_t^y)^2 = \beta_t (D_{t-1}^y)^2 + (1 - \beta_t)(H_t^y)^2, \tag{7}$$

where $\beta_t \in [0; 1]$ is a preconditioning momentum parameter (typically close to 1), $(H_t^x)^2$, $(H_t^y)^2$, $(D_0^x)^2$ and $(D_0^y)^2$ are some diagonal matrices. The update (7) is satisfied by Adam-based methods with $(H_t^x)^2 = \mathrm{diag}(\nabla_x f(x_t, y_t, \xi_t^D) \odot \nabla_x f(x_t, y_t, \xi_t^D))$ and $(H_t^y)^2 = \mathrm{diag}(\nabla_y f(x_t, y_t, \xi_t^D) \odot \nabla_y f(x_t, y_t, \xi_t^D))$ (here ξ_t^D is an independent random variable specially generated to calculate H_t^x and H_t^y), e.g. the original Adam (Kingma & Ba, 2015) ($\beta_t = \frac{\beta - \beta^{t+1}}{1 - \beta^{t+1}}$) and the earlier method, RMSProp (Tieleman et al., 2012) ($\beta_t \equiv \beta$). The rule (7) also applies to AdaHessian (Yao et al., 2020) – method with Hutchinson scaling, here one should take $\beta_t = \frac{\beta - \beta^{t+1}}{1 - \beta^{t+1}}$, $(H_t^x)^2 = \mathrm{diag}(v_t^x \odot \nabla_{xx} f(x_t, y_t, \xi_t^D) v_t^x)^2$ and $(H_t^y)^2 = \mathrm{diag}(v_t^y \odot \nabla_{yy} f(x_t, y_t, \xi_t^D) v_t^y)^2$, where v_t^x, v_t^y are from Rademacher distribution.[1] Let us stress that despite AdaHessian includes a stochastic Hessian, it does not need to be calculated explicitly as it needs just Hessian-vector product. Hence, the matrices H_t^x and H_t^y can be computed by performing back-propagation twice: firstly, we compute the stochastic gradients $\nabla_x f(x_t, y_t, \xi_t^D)$ and $\nabla_y f(x_t, y_t, \xi_t^D)$, then compute the gradients of the functions $\langle \nabla_x f(x_t, y_t, \xi_t^D), v_t^x \rangle$ and $\langle \nabla_y f(x_t, y_t, \xi_t^D), v_t^y \rangle$.

The following update can be used as an analogue of (7):

$$D_t^x = \beta_t D_{t-1}^x + (1 - \beta_t) H_t^x, \quad D_t^y = \beta_t D_{t-1}^y + (1 - \beta_t) H_t^y. \tag{8}$$

For example, such a rule is augmented by OASIS (Jahani et al., 2021) with $\beta_t \equiv \beta$ and $H_t^x = \mathrm{diag}(v_t^x \odot \nabla_{xx} f(x_t, y_t, \xi_t^D) v_t^x)$ and $H_t^y = \mathrm{diag}(v_t^y \odot \nabla_{yy} f(x_t, y_t, \xi_t^D) v_t^y)$. It turns out that expression (8) can easily be calculated by differentiating twice. Firstly, we find $\nabla_x f(x_t, y_t, \xi_t^D)$, and then we compute the gradient again, but for $(\nabla_x f(x_t, y_t, \xi_t^D))^T v_t^x$. As a result, we have $\nabla_{xx} f(x_t, y_t, \xi_t^D) v_t^x$. Using the component-wise product between v_t^x and $\nabla_{xx} f(x_t, y_t, \xi_t^D) v_t^x$, we get the diagonal of D_t^x.

Moreover, (8) is a good approximation of the Hessians' diagonals. Indeed, it is easy to see that in the expectation the non-diagonal elements of the matrices D_t^x, D_t^y are equal to 0 and the diagonal elements are equal the corresponding Hessians' elements.

To make the matrix positive definite, one can use the following transformation that is common in literature:

$$(\hat{D}_t^x)_{ii} = \max\left\{e, |(D_t^x)_{ii}|\right\}, \quad (\hat{D}_t^y)_{ii} = \max\left\{e, |(D_t^y)_{ii}|\right\}, \tag{9}$$

---

[1] All components of vectors v_t^x, v_t^y are independent and equal to ± 1 with probability $1/2$.

where e is a positive parameter. It was suggested e.g., in the original Adam paper (Kingma & Ba, 2015). Subsequently, the critical necessity of (9) was shown in (Reddi et al., 2019). One can alternatively define $(\hat{D}_t^x)_{ii} = |(D_t^x)_{ii}| + e$ and $(\hat{D}_t^y)_{ii} = |(D_t^y)_{ii}| + e$. It can be seen that all of the above scaling approaches construct diagonal matrices with positive elements. We formulate the properties of such matrices via the following lemma.

Lemma 1. Let us assume that D_0^x, D_0^y and for all t the H_t^x, H_t^y are diagonal matrices with elements not greater than Γ in absolute value. Then for matrices $\hat{D}_t^x, \hat{D}_t^y$ obtained by rules (7) − (9), the following holds

1) $\hat{D}_t^x, \hat{D}_t^y$ are diagonal matrices with non-negative elements and $eI \preccurlyeq \hat{D}_t^x \preccurlyeq \Gamma I$, $eI \preccurlyeq \hat{D}_t^y \preccurlyeq \Gamma I$;

2) $\hat{D}_{t+1}^x \preccurlyeq \left(1 + \frac{(1-\beta_{t+1})\Gamma^2}{2e^2}\right)\hat{D}_t^x$ and $\hat{D}_{t+1}^y \preccurlyeq \left(1 + \frac{(1-\beta_{t+1})\Gamma^2}{2e^2}\right)\hat{D}_t^y$ for (7);

3) $\hat{D}_{t+1}^x \preccurlyeq \left(1 + \frac{2(1-\beta_{t+1})\Gamma}{e}\right)\hat{D}_t^x$ and $\hat{D}_{t+1}^y \preccurlyeq \left(1 + \frac{2(1-\beta_{t+1})\Gamma}{e}\right)\hat{D}_t^y$ for (8).

Let us formally prove that all scalings: Adam, RMSProp, AdaHessian, and OASIS, satisfy the conditions of Lemma 1.

Lemma 2. Let Assumption 2 be satisfied, then the AdaHessian and OASIS preconditioners satisfy the conditions of the Lemma 1 with $\Gamma = \sqrt{d_x + d_y}L$. If $\|\nabla_x f(x,y,\xi)\| \leq M$ and $\|\nabla_y f(x,y,\xi)\| \leq M$ for all x, y, ξ, then the Adam and RMSProp preconditioners satisfy the conditions of the Lemma 1 with $\Gamma = M$.

The rule chains (7) + (9) or (8) + (9) allow to construct preconditioning matrices. Next, we present Algorithm 1. As mentioned in Section 1.1, we use the stochastic Extra Gradient as the base algorithm. Such an algorithm is a modification of SGD for SPPs. Its main difference is the presence of an extrapolation step (lines 6 and 7) before the main update (lines 9 and 10). This complicates the algorithm, because it is necessary to double count the gradient at each iteration, but meanwhile the original SGD is not optimal for SPPs.

Algorithm 1 `Scaled Stochastic Extra Gradient`

1: **Input:** initial point x_0, y_0
2: **for** $t = 0, 1, 2, \ldots T - 1$ **do**
3:      Generate independently ξ_t^D for $\hat{D}_t^x, \hat{D}_t^y$
4:      Compute $\hat{D}_t^x$ and $\hat{D}_t^y$ according to (7)–(9)
5:      Generate independently ξ_t for
       $\nabla_x f(x_t, y_t, \xi_t), \nabla_y f(x_t, y_t, \xi_t)$
6:      $x_{t+1/2} = x_t - \gamma(\hat{D}_t^x)^{-1}\nabla_x f(x_t, y_t, \xi_t)$
7:      $y_{t+1/2} = y_t + \gamma(\hat{D}_t^y)^{-1}\nabla_y f(x_t, y_t, \xi_t)$
8:      Generate independently $\xi_{t+1/2}$ for
       $\nabla_x f(x_{t+1/2}, y_{t+1/2}, \xi_{t+1/2}), \nabla_y f(x_{t+1/2}, y_{t+1/2}, \xi_{t+1/2})$
9:      $x_{t+1} = x_t - \gamma(\hat{D}_t^x)^{-1}\nabla_x f(x_{t+1/2}, y_{t+1/2}, \xi_{t+1/2})$
10:     $y_{t+1} = y_t + \gamma(\hat{D}_t^y)^{-1}\nabla_y f(x_{t+1/2}, y_{t+1/2}, \xi_{t+1/2})$
11: **end for**

Moreover, it can perform worse and diverge even on the simplest problems (see Sections 7.2 and 8.2 of (Goodfellow, 2016)). We add the same preconditioners to both steps of the Extra Gradient. These matrices are computed at the beginning of each iteration (lines 3 and 4).

The following theorem gives a convergence estimate for Algorithm 1.

Theorem 1. Consider the problem (1) under Assumptions 2 and 3 and conditions of Lemma 1. Let $\{x_t, y_t\}$ and $\{x_{t+1/2}, y_{t+1/2}\}$ be the sequences generated by Algorithm 1 after T iterations. Also assume that the sequence $\{\beta_t\}$ is not random and $\gamma \leq \frac{e}{4L}$. Then,

● under Assumption 1(SC) it holds that

$$\mathbb{E}\left[R_{t+1}^2\right] \leq \alpha_t \mathbb{E}\left[R_t^2\right] + \theta_t, \tag{10}$$

with $\alpha_t = \left(1 - \frac{\gamma\mu}{\Gamma} + (1 - \beta_{t+1})C\right)$ and $\theta_t = \frac{6\sigma^2(1+(1-\beta_{t+1})C)\gamma^2}{be}$;

● under Assumption 1(C) with $\|x_t\| \leq \Omega, \|y_t\| \leq \Omega$ for all $t \geq 0$ it holds that

$$\mathbb{E}\left[\text{gap}(x_T^{avg}, y_T^{avg})\right] \leq \frac{3\Gamma\Omega^2}{\gamma T} + \frac{2C\Gamma\Omega^2}{\gamma T}\sum_{t=1}^T (1 - \beta_t) + \frac{4\gamma\sigma^2}{eb}; \tag{11}$$

• under Assumption 1(NC) with $\|x_t\| \leq \Omega$, $\|y_t\| \leq \Omega$ for all $t \geq 0$ it holds that

$$\mathbb{E}\big[\|\nabla_x f(\bar{x}_T, \bar{y}_T)\|^2 + \|\nabla_y f(\bar{x}_T, \bar{y}_T)\|^2\big] \leq \frac{12\Gamma^2\Omega^2}{\gamma^2 T} + \frac{21\Gamma\sigma^2}{eb} + \frac{12C\Gamma^2\Omega^2}{\gamma^2 T}\sum_{t=1}^{T}(1-\beta_t).$$

(12)

Here we use $R_t^2 = \|x_t - x^*\|_{\hat{D}_t^x}^2 + \|y_t - y^*\|_{\hat{D}_t^y}^2$, $\text{gap}(x, y) = \max_{y' \in \mathcal{Y}} f(x, y') - \min_{x' \in \mathcal{X}} f(x', y)$, $x_T^{avg} = \frac{1}{T}\sum_{t=0}^{T-1} x_{t+1/2}$, $y_T^{avg} = \frac{1}{T}\sum_{t=0}^{T-1} y_{t+1/2}$, $(\bar{x}_T, \bar{y}_T)$ is picked uniformly from all (x_t, y_t), $C = \frac{\Gamma^2}{2e^2}$ for (7) or $C = \frac{2\Gamma}{e}$ for (8).

The estimate (11) uses a so-called gap function on some bounded set $\mathcal{X} \times \mathcal{Y}$. Eventhough we solve the problem (1) on the whole set $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, this is permissible, since such a version of the criterion is valid if the solution x^*, y^* lies in $\mathcal{X} \times \mathcal{Y}$; for details see (Nesterov, 2007).

Depending on the choice of β_t, we can get convergence for the four methods that are given before. In particular we can estimate the number of iterations to find a solution with ε precision (in the strongly convex–strongly concave case this means that $\|x_t - x^*\|^2 + \|y_t - y^*\|^2 \sim \varepsilon$, in the convex–concave that $\text{gap}(x, y) \sim \varepsilon$, and in the non-convex–non-concave that $\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2 \sim \varepsilon^2$).

Corollary 1 (RMSProp and OASIS). Under the assumptions of Theorem 1 let additionally take $\beta_t \equiv \beta$. Then, the number of iterations to achieve ε-solution (in expectation)

• in the strongly convex–strongly concave case with $\beta \geq 1 - \frac{e\gamma\mu}{2\Gamma C}$ is $\tilde{\mathcal{O}}\left(\frac{\Gamma L}{e\mu}\log\frac{1}{\varepsilon} + \frac{\Gamma^3\sigma^2}{e^2 b\mu^2\varepsilon}\right)$;

• in the convex–concave case with $\beta \geq 1 - \frac{1}{CT}$ is $\mathcal{O}\left(\frac{\Gamma L\Omega^2}{e\varepsilon} + \frac{\Gamma\sigma^2\Omega^2}{eb\varepsilon^2}\right)$;

• in the non-convex–non-concave case with $\beta \geq 1 - \frac{1}{CT}$ and $b \sim \frac{\Gamma\sigma^2}{e\varepsilon^2}$ is $\mathcal{O}\left(\frac{\Gamma^2 L^2\Omega^2}{e^2\varepsilon^2}\right)$.

Corollary 2 (Adam and AdaHessian). Under the assumptions of Theorem 1, let additionally take $\beta_t = \frac{\beta - \beta^{t+1}}{1 - \beta^{t+1}}$. Then, the complexities to achieve ε-solution (in expectation)

• in the str. convex–str. concave case with $\beta = 1 - \left(\frac{\gamma\mu}{2\Gamma C}\right)^2$ is $\mathcal{O}\left(\frac{C^2\Gamma^3 L^2\Omega^2}{e^3\mu^2\varepsilon} + \frac{C^4\Gamma^5\sigma^4\Omega^2}{e^5 b^2\mu^4\varepsilon^3}\right)$;

• in the convex–concave case with $\beta = 1 - \frac{1}{T}$ is $\mathcal{O}\left(\frac{C^2\Gamma^2 L^2\Omega^4}{e^2\varepsilon^2} + \frac{C^2\Gamma^2\sigma^4\Omega^4}{e^2 b^2\varepsilon^4}\right)$;

• in the non-convex–non-concave case with $\beta = 1 - \frac{1}{T}$ and $b \sim \frac{\Gamma\sigma^2}{e\varepsilon^2}$ is $\mathcal{O}\left(\frac{C^2\Gamma^4 L^4\Omega^4}{e^4\varepsilon^4}\right)$.

Corollaries 1 and 2 together with of Lemma 2 give a complete picture for the 4 methods based on Algorithm 1 with Adam, RMSProp, AdaHessian and OASIS scalings.

Discussion. The following is a brief observation of our results.

• In our theoretical estimates, the convergence of methods with time-variable β_t is worse than for methods with constant β. This is observed in other papers as well (Reddi et al., 2019; Défossez et al., 2020; Dou & Li, 2021). Typically, methods with time-variable β_t perform their best in practical problems. But in our experiments (Appendix 5), the performance is the same; RMSProp outperforms Adam, and OASIS wins over AdaHessian.

• For all methods, we need to get the β reasonably close to 1. The same is observed in Adam's analysis for minimization problems (Défossez et al., 2020). Moreover, such a choice is confirmed in the experiments, both in ours and in the classical settings of Adam-based methods (Kingma & Ba, 2015) (recommend $\beta = 0.999$).

• It is interesting to see how our estimates depend on the dimension $d = d_x + d_y$. This is relevant especially for large size problems. The dependence on d appears in the constants Γ and C (Lemmas 1,2). For OASIS and AdaHessian, $\Gamma \sim \sqrt{d}$, but for Adam and RMSProp there is no dependence on d. This differs from the results of earlier papers (Défossez et al., 2020; Liu et al., 2019; Dou & Li, 2021), where the convergence estimate for the Adam-based methods depends on d more often linearly.

• As noted earlier, our analysis is universal for any scaled method for which Lemma 1 holds.

• Among the problems of analysis, let us highlight the use of $\|x_t\| \leq \Omega$, $\|y_t\| \leq \Omega$. Only in the strongly convex – strongly concave case is it avoided. At the moment this assumption can be found in the literature on saddle point problems, moreover, all the papers (Liu et al., 2019; Dou & Li, 2021) on methods with scaling also use it.

4 MAKE SCALED METHODS MORE ROBUST

Algorithm 2   Scaled Stochastic Single Call Extra Gradient with Momentum

---

1: **Input:** initial point $x_0 = w_0^x, y_0 = w_0^y$, learning rate γ, probability $p \in [0; 1]$
2: **for** $t = 0, 1, 2, \dots T - 1$ **do**
3:      Compute the preconditioners $\hat{D}_{t-1/2}^x$ and $\hat{D}_{t-1/2}^y$ according to (7)–(9)
4:      $x_{t+1/2} = x_t - \gamma(\hat{D}_{t-1/2}^x)^{-1}\nabla_x f(x_{t-1/2}, y_{t-1/2}, \xi_{t-1/2})$
5:      $y_{t+1/2} = y_t + \gamma(\hat{D}_{t-1/2}^y)^{-1}\nabla_y f(x_{t-1/2}, y_{t-1/2}, \xi_{t-1/2})$
6:      Generate $\xi_{t+1/2}$ for $\nabla_x f(x_{t+1/2}, y_{t+1/2}, \xi_{t+1/2})$, $\nabla_y f(x_{t+1/2}, y_{t+1/2}, \xi_{t+1/2})$ and $\hat{D}_{t+1/2}^x, \hat{D}_{t+1/2}^y$
7:      $x_{t+1} = x_t + \eta(\hat{D}_{t-1/2}^x)^{-1}(w_t^x - x_t) - \gamma(\hat{D}_{t-1/2}^x)^{-1}\nabla_x f(x_{t+1/2}, y_{t+1/2}, \xi_{t+1/2})$
8:      $y_{t+1} = y_t + \eta(\hat{D}_{t-1/2}^y)^{-1}(w_t^y - y_t) + \gamma(\hat{D}_{t-1/2}^y)^{-1}\nabla_y f(x_{t+1/2}, y_{t+1/2}, \xi_{t+1/2})$
9:      $w_{t+1}^x, w_{t+1}^y = \begin{cases} x_t, y_t, & \text{with probability } p \\ w_t^x, w_t^y, & \text{with probability } 1 - p \end{cases}$
10: **end for**

Algorithm 1 as well as Extra Gradient has an important disadvantage, two gradient calculations per iteration. The next method (Algorithm 2) uses the so-called Optimistic approach (Popov, 1980), which solves the problem of Extra Gradient mentioned above. The difference from Algorithm 1 is that in the extrapolation step (lines 4 and 5) we use the gradients from the previous main update (lines 7 and 8) instead of the new gradients at point x_t, y_t. Also Algorithm 2 adds momentum as a useful tool (lines 7 and 8). We use negative momentum with a random update (line 9). The same momentum is investigated in (Alacaoglu & Malitsky, 2021; Kovalev et al., 2022). The usefulness of negative momentum for GANs training is discussed by (Gidel et al., 2019). Both modifications of Algorithm 2 are of a practical nature. From the theoretical point of view, they do not give any improvement in terms of convergence estimates, including the fact that saddle problems cannot be accelerated in terms of ε as minimization problems (Ouyang & Xu, 2019).

Similar to Algorithm 1, we give a convergence theorem and its corollaries for Algorithm 2.

Theorem 2. Consider the problem (1) under Assumptions 2 and 3 and conditions of Lemma 1. Let $\{x_t, y_t\}$ and $\{x_{t+1/2}, y_{t+1/2}\}$ be the sequences generated by Algorithm 2 after T iterations. Then, with $\gamma \leq \frac{e}{10L}, \eta \leq ep$ and $p \leq \frac{1}{4}$

• under Assumption 1(SC) it holds that

$$\mathbb{E}\left[\Psi_{t+1}\right] \leq \alpha_t \mathbb{E}\left[\Psi_t\right] + \frac{12\sigma^2\left(1 + (1 - \beta_{t+1})C\right)\gamma^2}{eb}, \tag{13}$$

with $\alpha_t = \max\left\{\left(1 - \frac{\gamma\mu}{2\Gamma} + (1 - \beta_{t+1})C\right); \left(1 - \frac{\gamma\mu p}{2\eta + \gamma\mu}\right)\right\}$ and $\theta_t = \frac{12\sigma^2(1+(1-\beta_{t+1})C)\gamma^2}{eb}$;

• under Assumption 1(C) with $\|x_t\| \leq \Omega$ and $\|y_t\| \leq \Omega$ for all $t \geq 0$ it holds that

$$\mathbb{E}\left[\text{gap}(x_T^{avg}, y_T^{avg})\right] \leq \frac{50\Gamma\Omega^2}{\gamma T} + \frac{18C\Gamma\Omega^2}{\gamma T}\sum_{t=1}^{T}(1 - \beta_t) + \frac{55\gamma\sigma^2}{eb}; \tag{14}$$

- under Assumption 1(NC) with $\|x_t\| \leq \Omega$ and $\|y_t\| \leq \Omega$ for all $t \geq 0$ it holds that

$$\mathbb{E}[\|\nabla_x f(\bar{x}_{T-1/2}, \bar{y}_{T-1/2})\|^2 + \|\nabla_y f(\bar{x}_{T-1/2}, \bar{y}_{T-1/2})\|^2]$$

$$\leq \frac{96\Gamma^2\Omega^2}{\gamma^2 T} + \frac{48\Gamma L^2\Omega^2}{eT} + \frac{32\Gamma^2 C\Omega^2}{\gamma^2 T}\sum_{t=1}^{T}(1-\beta_t) + \frac{105\Gamma\sigma^2}{eb}. \quad (15)$$

Here we use Ψ_{t+1} defined in (28), $(\bar{x}_{T-1/2}, \bar{y}_{T-1/2})$ is picked uniformly from all $(x_{t-1/2}, y_{t-1/2})$. Other notations are similar to Theorem 1.

Corollary 3 (RMSProp and OASIS). Under the assumptions of Theorem 2 let additionally take $\beta_t \equiv \beta$. Then, the number of iterations to achieve ε-solution (in expectation)

- in the strongly convex–strongly concave case with $\beta \geq 1 - \frac{\gamma\mu}{4\Gamma C}$ is $\tilde{\mathcal{O}}\left(\left[p + \frac{\Gamma L}{e\mu}\right]\log\frac{1}{\varepsilon} + \frac{\Gamma^2\sigma^2}{eb\mu^2\varepsilon}\right)$;

- in the convex–concave case with $\beta \geq 1 - \frac{1}{CT}$ is $\mathcal{O}\left(\frac{\Gamma L\Omega^2}{e\varepsilon} + \frac{\Gamma^2\sigma^2\Omega^2}{eb\varepsilon^2}\right)$;

- in the non-convex–non-concave case with $\beta \geq 1 - \frac{1}{CT}$ and $b \sim \frac{\Gamma\sigma^2}{e\varepsilon^2}$ is $\mathcal{O}\left(\frac{\Gamma^2 L^2\Omega^2}{e^2\varepsilon^2}\right)$.

Corollary 4 (Adam and AdaHessian). Under the assumptions of Theorem 2, let additionally take $\beta_t = \frac{\beta-\beta^{t+1}}{1-\beta^{t+1}}$. Then, the complexities to achieve ε-solution (in expectation)

- in the str. convex–str. concave case with $\beta = 1 - \left(\frac{\gamma\mu}{4\Gamma C}\right)^2$ is $\mathcal{O}(\frac{C^2\Gamma^3 L^2\Omega^2}{e^3\mu^2\varepsilon} + \frac{C^2\Gamma^2 L^2\Omega^2}{e^2\mu^2\varepsilon} + \frac{C^2\Gamma^2 L\Omega^2}{e^2 p\mu\varepsilon} + \frac{C^4\Gamma^5\sigma^4\Omega^2}{e^5\mu^4 b^2\varepsilon^3} + \frac{C^4\Gamma^2\sigma^4\Omega^2}{e^2\mu^4 b^2\varepsilon^3} + (\frac{C^5\Gamma^4\sigma^2\Omega^4}{e^4\mu^2 p^3 b\varepsilon^3})^{1/2})$;

- in the convex–concave case with $\beta = 1 - \frac{1}{T}$ is $\mathcal{O}\left(\frac{C^2\Gamma^2 L^2\Omega^4}{e^2\varepsilon^2} + \frac{C^2\Gamma^2\sigma^4\Omega^4}{e^2 b^2\varepsilon^4}\right)$;

- in the non-convex–non-concave case with $\beta = 1 - \frac{1}{T}$ and $b \sim \frac{\Gamma\sigma^2}{e\varepsilon^2}$ is $\mathcal{O}\left(\frac{C^2\Gamma^4 L^4\Omega^4}{e^4\varepsilon^4}\right)$.

The estimates obtained are very similar to Section 3. The same observations can be made about these results as those given in Section 3 after Corollary 2.

5 EXPERIMENTS

This section provides extensive experiments and compares the considered optimization methods for training GANs.

5.1 SETUP

Datasets. We use the standard dataset for training GANs, CIFAR-10 (Krizhevsky et al., 2009) that has 50k of 32x32 resolution images in the training set and 10k images in the test.

Metrics. We use Inception score (IS) (Salimans et al., 2016) and Frećhet Inception distance (FID) (Heusel et al., 2017) as the most common metrics for assessing image synthesis models.

GAN architectures. We implemented ResNet architectures for the generator and the discriminator from (Miyato et al., 2018). See details in Appendix A.

Optimization methods. We employ RMSProp, Adam, AdaHessian, and OASIS, its extragradient versions Extra RMSProp, Extra Adam, Extra AdaHessian, and Extra OASIS. Additionally, we implement OASIS with momentum and Extra OASIS with momentum. We compute the uniform average and exponential moving average (EMA) of weights as well-known techniques to yield faster convergence. The hyperparameter settings of the methods are given in Appendices A.

5.2 RESULTS

The goal of the first experiment is to investigate the four methods for which we give a theory. These include the popular and one of the best-known methods, Adam (the benchmark for the other competitors), the known, but less popular RMSProp, and the very new and never used for GANs training methods with scaling (OASIS and AdaHessian).

For each of the optimizers, we have implemented two versions: classical and with extra step (both Algorithm 1 and Algorithm 2 have extra steps). We compare how the qualities of the four different

scaling approaches relate to each other and also verify the popular idea in the community that using extra step is a necessary trick to increase the learning quality (Mertikopoulos et al., 2018; Gidel et al., 2018; Chavdarova et al., 2019; Liu et al., 2019; Dou & Li, 2021). See results in Table 2 and Figure 1.

We see that three optimizers out of four performed well. The not very popular scaling techniques OASIS and RMSProp perform better than state-of-the-art Adam. The use of the extra step technique can provide improvement but is not an important element for the optimizer.

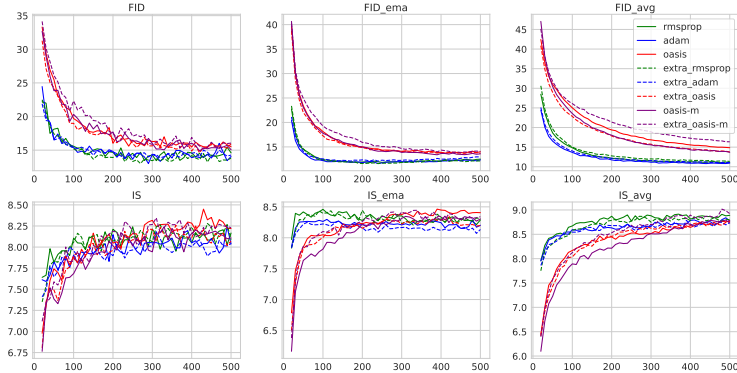In the second group of experiments, we decided to investigate the OASIS scaling technique in more detail, as

Figure 1: Averaged dynamics of FID and IS change by epoch number for different optimizers from Tables 2 and 3.

it is new to GANs in comparison to Adam and RMSProp, and also showed good results in the first group of experiments. In particular, we explore how OASIS is affected by negative momentum and the combination of extra step and negative momentum (Algorithm 2). See results in Table 3 and Figure 1.

We can see that negative momentum can indeed increase the quality of learning, sometimes its combination with extra step is also useful.

Table 2: Comparisons of the considered optimization methods and their extragradient versions using FID and IS, scores obtained at the end of training. EMA and uniform avg denote as exponential moving average and uniform averaging, respectively. Results are averaged over 5 runs. We run each experiment for 500k iterations. For each column the best score is in bold along with any score within its standard deviation reach.

CIFAR-10	Fréchet Inception distance ↓			Inception score ↑		
	no avg	EMA	uniform avg	no avg	EMA	uniform avg
RMSProp	14.54 ± 0.55	12.42 ± 0.08	11.14 ± 0.10	8.15 ± 0.07	8.27 ± 0.03	$\mathbf{8.87 \pm 0.03}$
Adam	14.57 ± 0.96	12.27 ± 0.07	$\mathbf{10.87 \pm 0.05}$	8.05 ± 0.12	8.20 ± 0.05	8.77 ± 0.06
AdaHessian	39.54 ± 4.43	37.71 ± 0.20	53.96 ± 0.19	6.92 ± 0.06	6.85 ± 0.06	6.52 ± 0.05
OASIS	15.76 ± 0.28	13.93 ± 0.13	14.96 ± 0.08	$\mathbf{8.26 \pm 0.05}$	$\mathbf{8.41 \pm 0.02}$	8.74 ± 0.03
Extra RMSProp	$\mathbf{13.68 \pm 0.37}$	$\mathbf{12.08 \pm 0.07}$	11.42 ± 0.05	8.14 ± 0.09	8.30 ± 0.02	8.80 ± 0.02
Extra Adam	14.41 ± 0.36	12.87 ± 0.10	11.09 ± 0.03	8.02 ± 0.05	8.13 ± 0.04	8.72 ± 0.04
Extra AdaHessian	40.11 ± 0.50	40.51 ± 0.43	55.46 ± 0.38	6.72 ± 0.06	6.66 ± 0.04	5.98 ± 0.03
Extra OASIS	15.51 ± 0.25	13.79 ± 0.09	13.95 ± 0.09	8.11 ± 0.06	8.20 ± 0.01	8.73 ± 0.03

Table 3: Comparisons of the OASIS and OASIS with momentum (OASIS-M) optimization methods and their extragradient versions using FID and IS, scores obtained at the end of training. EMA and uniform avg denote as exponential moving average and uniform averaging, respectively. Results are averaged over 5 runs. We run each experiment for 500k iterations. For each column the best score is in bold along with any score within its standard deviation reach.

CIFAR-10	Fréchet Inception distance ↓			Inception score ↑		
	no avg	EMA	uniform avg	no avg	EMA	uniform avg
OASIS	15.76 ± 0.28	13.93 ± 0.13	14.96 ± 0.08	$\mathbf{8.26 \pm 0.05}$	$\mathbf{8.41 \pm 0.02}$	8.74 ± 0.03
OASIS-M	$\mathbf{15.38 \pm 0.23}$	$\mathbf{13.54 \pm 0.10}$	13.91 ± 0.17	8.14 ± 0.04	8.31 ± 0.03	8.77 ± 0.02
Extra OASIS	$\mathbf{15.51 \pm 0.25}$	13.79 ± 0.09	$\mathbf{13.95 \pm 0.09}$	8.11 ± 0.06	8.20 ± 0.01	8.73 ± 0.03
Extra OASIS-M	16.10 ± 0.70	14.32 ± 0.22	17.70 ± 0.27	$\mathbf{8.26 \pm 0.06}$	8.35 ± 0.04	$\mathbf{8.86 \pm 0.03}$

REFERENCES

Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. *arXiv preprint arXiv:2102.08352*, 2021.

Babak Barazandeh, Kristal Curtis, Chandrima Sarkar, Ram Sriharsha, and George Michailidis. On the convergence of adam-type algorithms for solving structured single node and decentralized min-max saddle point games. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4343–4347, 2022. doi: 10.1109/ICASSP43922.2022. 9746485.

C. Bekas, E. Kokiopoulou, and Y. Saad. An estimator for the diagonal of a matrix. *Applied Numerical Mathematics*, 57(11):1214–1229, 2007. Numerical Algorithms, Parallelism and Applications (2).

Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in gan training with variance reduced extragradient. *arXiv preprint arXiv:1904.08598*, 2019.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

Jelena Diakonikolas, Constantinos Daskalakis, and Michael Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2746–2754. PMLR, 2021.

Zehao Dou and Yuanzhi Li. On the one-sided convergence of adam-type algorithms in non-convex non-concave min-max optimization. *arXiv preprint arXiv:2109.14213*, 2021.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159, 2011.

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

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1802–1811. PMLR, 2019.

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *arXiv preprint arXiv:1908.08465*, 2019.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. *Advances in Neural Information Processing Systems*, 33:16223–16234, 2020.

Majid Jahani, Sergey Rusakov, Zheng Shi, Peter Richtárik, Michael Mahoney, and Martin Takáč. Doubly adaptive scaled algorithm for machine learning using second-order information. *arXiv preprint arXiv:2109.05198v1*, 2021.

Anatoli Juditsky, Arkadii S. Nemirovskii, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm, 2008.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, pp. 2305–2313, 2015.

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

Dmitry Kovalev, Aleksandr Beznosikov, Abdurakhmon Sadiev, Michael Persiianov, Peter Richtárik, and Alexander Gasnikov. Optimal algorithms for decentralized stochastic variational inequalities. *arXiv preprint arXiv:2202.02771*, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Zhen-Zhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 907–915. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/liang19b.html.

Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, and Tianbao Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. *arXiv preprint arXiv:1912.11940*, 2019.

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. *arXiv preprint arXiv:1807.02629*, 2018.

George J. Minty. Monotone (nonlinear) operators in Hilbert space. *Duke Mathematical Journal*, 29 (3):341 – 346, 1962. doi: 10.1215/S0012-7094-62-02933-2. URL https://doi.org/10.1215/S0012-7094-62-02933-2.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Aryan Mokhtari, Asuman E Ozdaglar, and Sarath Pattathil. Convergence rate of o(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020.

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*, 15:229–251, 01 2004. doi: 10.1137/S1052623403425629.

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

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

Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, pp. 1–35, 2019.

Wei Peng, Yu-Hong Dai, Hui Zhang, and Lizhi Cheng. Training gans with centripetal acceleration. *Optimization Methods and Software*, 35(5):955–973, 2020.

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.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Tijmen Tieleman, Geoffrey Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

Zhewei Yao, Amir Gholami, Sheng Shen, Kurt Keutzer, and Michael W Mahoney. ADAHESSIAN: An adaptive second order optimizer for machine learning. *arXiv preprint arXiv:2006.00719*, 2020.

Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

A EXPERIMENTS

A.1 IMPLEMENTATION DETAILS

In our implementation of experiments we use the PyTorch[2] framework. For the computation of the FID and IS metrics we utilize the the package *pytorch_gan_metrics*[3] that reproduces the original Tensorflow implementation in PyTorch. We provide all source code of our experiments as a part of supplementary material. We also attach configuration files for each model to run its training with specified hyperparameters.

A.2 ARCHITECTURES

We replicate the experimental setup as in (Miyato et al., 2018; Chavdarova et al., 2019). We use ResNet architectures for the generator G and the discriminator D as described in Table 4.

Table 4: ResNet architectures for CIFAR10 dataset. We use similar architectures to the ones used in (Miyato et al., 2018).
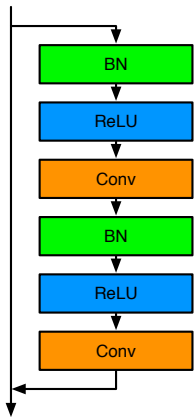
Figure 2: ResBlock architecture. For the discriminator we removed BN layers in ResBlock.

$z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$
dense, $4 \times 4 \times 256$
ResBlock up 256
ResBlock up 256
ResBlock up 256
BN, ReLU, 3×3 conv, 3 Tanh

(a) Generator

RGB image $x \in \mathbb{R}^{32 \times 32 \times 3}$
ResBlock down 128
ResBlock down 128
ResBlock 128
ResBlock 128
ReLU
Global sum pooling
dense $\to 1$

(b) Discriminator

A.3 TRAINING LOSSES

For training we use the hinge version of the non-saturating adversarial loss as in (Miyato et al., 2018):

$$\mathcal{L}_D = \mathbb{E}_{\boldsymbol{x} \sim q_{\mathrm{data}}(\boldsymbol{x})} \left[\min\left(0, -1 + D(\boldsymbol{x})\right) \right] + \mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z})} \left[\min\left(0, -1 - D\left(G(\boldsymbol{z})\right)\right) \right] \quad (16)$$

$$\mathcal{L}_G = -\mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z})} \left[D\left(G(\boldsymbol{z})\right) \right], \quad (17)$$

A.4 HYPERPARAMETERS

Table 5 lists the hyperparameters we used to obtain the reported results on CIFAR-10 dataset. These values were tuned for each method independently. D_steps means the number of the discriminator steps per one generator step. eps denotes the threshold in OASIS algorithm that is used to clip the diagonal elements of the Hessian. $hess\_upd\_each$ means how often we update the diagonal elements of the Hessian.

---

[2]https://pytorch.org
[3]https://pypi.org/project/pytorch-gan-metrics/

Table 5: Hyperparameters we used for CIFAR-10 dataset.

Method	Hyperparameters							
	lr_G	lr_D	β_1	β_2	D_steps	batch size	eps	$hess\_upd\_each$
RMSProp	0.0002	0.0002	–	0.999	5	64	–	–
Adam	0.0002	0.0002	0.0	0.999	5	64	–	–
AdaHessian	0.0002	0.0002	0.0	0.999	5	64	–	20
OASIS	0.0002	0.0002	0.0	0.999	5	64	0.01	20
Extra RMSProp	0.0002	0.0002	–	0.999	5	64	–	–
Extra Adam	0.0002	0.0002	0.0	0.999	5	64	–	–
Extra AdaHessian	0.0002	0.0002	0.0	0.999	5	64	–	20
Extra OASIS	0.0002	0.0002	0.0	0.999	5	64	0.01	20
OASIS-M	0.0002	0.0002	0.1	0.999	5	64	0.01	20
Extra OASIS-M	0.0002	0.0002	0.1	0.999	5	64	0.01	20

A.5 TOTAL AMOUNT OF COMPUTE RESOURCES

We run our experiments on Tesla V100 GPUs. We used approximately 2500 GPU hours to obtain the reported results.

A.6 SAMPLES OF THE TRAINED GENERATORS

We show random samples of the trained generators for each method. We chose the best averaging strategy for each algorithm based on the visual quality. The provided samples are in Figures 3-12.
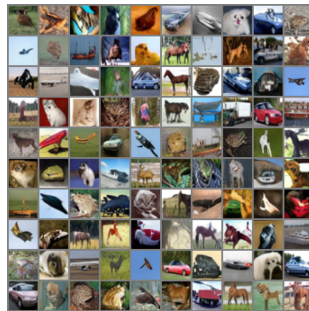
Figure 3: Samples produced by the generator trained by the RMSProp algorithm using uniform average on the weights.
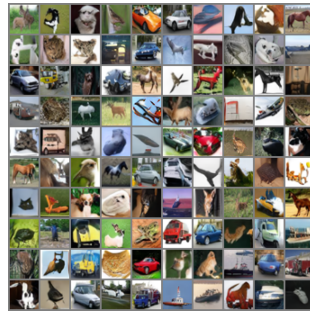
Figure 4: Samples produced by the generator trained by the Adam algorithm using uniform average on the weights.
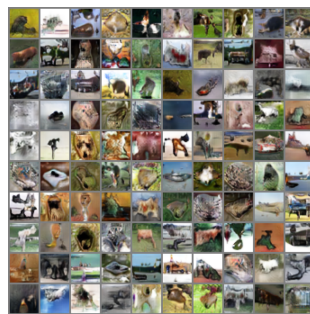
Figure 5: Samples produced by the generator trained by the AdaHessian algorithm using exponential moving average on the weights.
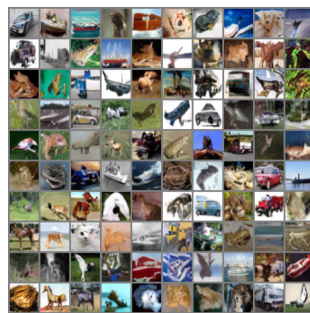

Figure 6: Samples produced by the generator trained by the OASIS algorithm using exponential moving average on the weights.
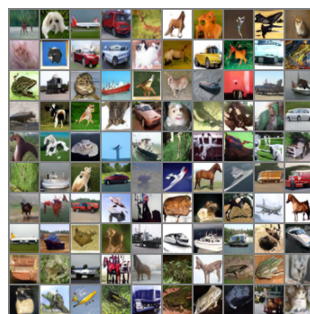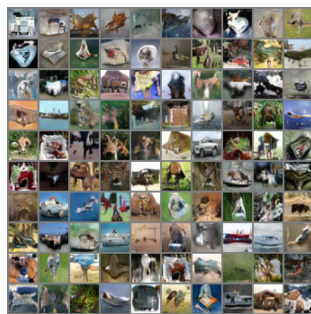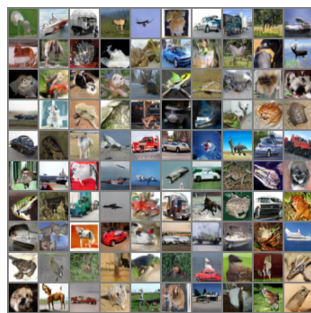

Figure 7: Samples produced by the generator trained by the Extra RMSProp algorithm using uniform average on the weights.

Figure 8: Samples produced by the generator trained by the Extra Adam algorithm using uniform average on the weights.

Figure 9: Samples produced by the generator trained by the Extra AdaHessian algorithm using exponential moving average on the weights.

Figure 10: Samples produced by the generator trained by the Extra OASIS algorithm using exponential moving average on the weights.
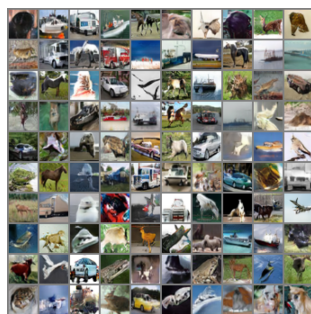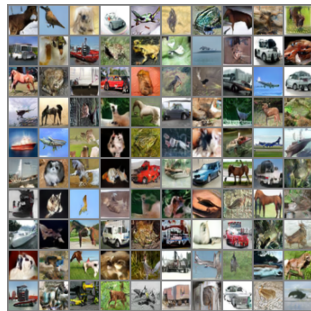

Figure 11: Samples produced by the generator trained by the OASIS-M algorithm using exponential moving average on the weights.

Figure 12: Samples produced by the generator trained by the Extra OASIS-M algorithm using exponential moving average on the weights.

B TECHNICAL LEMMA

Lemma 3. For any $T \in \mathbb{N}$, we get

$$\left(1 - \frac{1}{T}\right)^{\sqrt{T}} \leq 1 - \frac{1}{2\sqrt{T}}. \tag{18}$$

Proof: We start from the inequality that for all $x \in \mathbb{R}$ it holds $1 + x \leq \exp(x)$. Hence, for $x = -\frac{1}{T}$ we have

$$\left(1 - \frac{1}{T}\right)^{\sqrt{T}} \leq \exp\left(-\frac{\sqrt{T}}{T}\right) = \exp\left(-\frac{1}{\sqrt{T}}\right). \tag{19}$$

Using the Taylor expansion of $\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ with $x = -\frac{1}{\sqrt{T}}$, we obtain

$$\exp\left(-\frac{1}{\sqrt{T}}\right) = 1 - \frac{1}{\sqrt{T}} + \frac{1}{2T} - \frac{1}{6T\sqrt{T}} + \sum_{n=4}^{\infty} \frac{(-1)^n}{n! \cdot T^{n/2}}$$

One can note that for $T \geq 1$:

$$\frac{1}{2T} - \frac{1}{6T\sqrt{T}} + \sum_{n=4}^{\infty} \frac{(-1)^n}{n! \cdot T^{n/2}} \leq \frac{1}{2T} \leq \frac{1}{2\sqrt{T}}.$$

Therefore,

$$\exp\left(-\frac{1}{\sqrt{T}}\right) \leq 1 - \frac{1}{\sqrt{T}} + \frac{1}{2\sqrt{T}} = 1 - \frac{1}{2\sqrt{T}}. \tag{20}$$

Combining the inequalities (19) and (20), we have:

$$\left(1 - \frac{1}{T}\right)^{\sqrt{T}} \leq \exp\left(-\frac{1}{\sqrt{T}}\right) \leq 1 - \frac{1}{2\sqrt{T}}.$$

This finishes the proof. $\square$

C MISSING PROOFS

C.1 PROOF OF LEMMA 1

1) It is easy to see that due to the fact that all matrices H_t^x and D_0^x are diagonal, then based on rules (7) and (8), we can conclude that all matrices D_t^x are also diagonal. Again from rules (7), (8) and $\beta_t \in [0; 1]$, we get that all elements of D_t^x matrices are not greater than Γ in absolute value. (9) gives that $\hat{D}_t^x$ matrices are diagonal and their elements from e to Γ. This completes the proof. Similar proof for the variable y. $\square$

2) We start from simple steps:

$$eI \preccurlyeq \hat{D}_{t+1}^x = \hat{D}_t^x + \hat{D}_{t+1}^x - \hat{D}_t^x = \hat{D}_t^x + (\hat{D}_{t+1}^x - \hat{D}_t^x)(\hat{D}_{t+1}^x + \hat{D}_t^x)(\hat{D}_{t+1}^x + \hat{D}_t^x)^{-1}.$$

Since matrices $\hat{D}_{t+1}^x$ and $\hat{D}_t^x$ are diagonal, we have that $\hat{D}_{t+1}^x \hat{D}_t^x = \hat{D}_t^x \hat{D}_{t+1}^x$ and then

$$eI \preccurlyeq \hat{D}_{t+1}^x = \hat{D}_t^x + ((\hat{D}_{t+1}^x)^2 - (\hat{D}_t^x)^2)(\hat{D}_{t+1}^x + \hat{D}_t^x)^{-1}$$
$$\preccurlyeq \hat{D}_t^x + \|(\hat{D}_{t+1}^x)^2 - (\hat{D}_t^x)^2\|_\infty \|(\hat{D}_{t+1}^x + \hat{D}_t^x)^{-1}\|_\infty I.$$

All diagonal elements of $\hat{D}_{t+1}^x$ and $\hat{D}_t^x$ are positive from e to Γ, it means that $\|(\hat{D}_{t+1}^x + \hat{D}_t^x)^{-1}\|_\infty \leq \frac{1}{2e}$. Hence,

$$eI \preccurlyeq \hat{D}_{t+1}^x \preccurlyeq \hat{D}_t^x + \frac{1}{2e}\|(\hat{D}_{t+1}^x)^2 - (\hat{D}_t^x)^2\|_\infty I.$$

One can note the following:

- if $((D_{t+1}^x)^2)_{ii} \geq e^2$ and $((D_t^x)^2)_{ii} \geq e^2$ then $|((\hat{D}_{t+1}^x)^2)_{ii} - ((\hat{D}_t^x)^2)_{ii}| = |((D_{t+1}^x)^2)_{ii} - ((D_t^x)^2)_{ii}|$;

- if $((D_{t+1}^x)^2)_{ii} \geq e^2$ and $((D_t^x)^2)_{ii} < e^2$ then $|((\hat{D}_{t+1}^x)^2)_{ii} - ((\hat{D}_t^x)^2)_{ii}| \leq |((D_{t+1}^x)^2)_{ii} - ((D_t^x)^2)_{ii}|$;

- if $((D_{t+1}^x)^2)_{ii} < e^2$ and $((D_t^x)^2)_{ii} \geq e^2$ then $|((\hat{D}_{t+1}^x)^2)_{ii} - ((\hat{D}_t^x)^2)_{ii}| \leq |((D_{t+1}^x)^2)_{ii} - ((D_t^x)^2)_{ii}|$;

- if $((D_{t+1}^x)^2)_{ii} < e^2$ and $((D_t^x)^2)_{ii} < e^2$ then $|((\hat{D}_{t+1}^x)^2)_{ii} - ((\hat{D}_t^x)^2)_{ii}| = 0 \leq |((D_{t+1}^x)^2)_{ii} - ((D_t^x)^2)_{ii}|$.

Hence,

$$eI \preccurlyeq \hat{D}_{t+1}^x \preccurlyeq \hat{D}_t^x + \frac{1}{2e}\|(D_{t+1}^x)^2 - (D_t^x)^2\|_\infty I$$
$$= \hat{D}_t^x + \frac{1 - \beta_{t+1}}{2e}\|(H_t^x)^2 - (D_t^x)^2\|_\infty I$$

All diagonal elements of $(H_t^x)^2$ and $(D_t^x)^2$ are from 0 to Γ^2, then

$$eI \preccurlyeq \hat{D}_{t+1}^x \preccurlyeq \hat{D}_t^x + \frac{(1 - \beta_{t+1})\Gamma^2}{2e}I.$$

Finally, using that $I \preccurlyeq \frac{1}{e}$ and get

$$eI \preccurlyeq \hat{D}_{t+1}^x \preccurlyeq \left(1 + \frac{(1 - \beta_{t+1})\Gamma^2}{2e^2}\right)\hat{D}_t^x.$$

Similar proof for the variable y. $\square$

3) We start from simple steps:

$$eI \preccurlyeq \hat{D}_{t+1}^x = \hat{D}_t^x + \hat{D}_{t+1}^x - \hat{D}_t^x.$$

Then,

$$eI \preccurlyeq \hat{D}_{t+1}^x \preccurlyeq \hat{D}_t^x + \|\hat{D}_{t+1}^x - \hat{D}_t^x\|_\infty I.$$

One can note the following:

- if $(D_{t+1}^x)_{ii} \geq e$ and $(D_t^x)_{ii} \geq e$ then $|(\hat{D}_{t+1}^x)_{ii} - (\hat{D}_t^x)_{ii}| = |(D_{t+1}^x)_{ii} - (D_t^x)_{ii}|$;

- if $(D_{t+1}^x)_{ii} \geq e$ and $(D_t^x)_{ii} < e$ then $|(\hat{D}_{t+1}^x)_{ii} - (\hat{D}_t^x)_{ii}| \leq |(D_{t+1}^x)_{ii} - (D_t^x)_{ii}|$;

- if $(D_{t+1}^x)_{ii} < e$ and $(D_t^x)_{ii} \geq e$ then $|(\hat{D}_{t+1}^x)_{ii} - (\hat{D}_t^x)_{ii}| \leq |(D_{t+1}^x)_{ii} - (D_t^x)_{ii}|$;

- if $(D_{t+1}^x)_{ii} < e$ and $(D_t^x)_{ii} < e$ then $|(\hat{D}_{t+1}^x)_{ii} - (\hat{D}_t^x)_{ii}| = 0 \leq |(D_{t+1}^x)_{ii} - (D_t^x)_{ii}|$.

Hence,

$$eI \preccurlyeq \hat{D}_{t+1}^x \preccurlyeq \hat{D}_t^x + \|D_{t+1}^x - D_t^x\|_\infty I$$
$$= \hat{D}_t^x + (1 - \beta_{t+1})\|H_t^x - D_t^x\|_\infty I$$

All diagonal elements of H_t^x and D_t^x are from $-\Gamma$ to Γ, then

$$eI \preccurlyeq \hat{D}_{t+1}^x \preccurlyeq \hat{D}_t^x + 2(1 - \beta_{t+1})\Gamma I.$$

Finally, using that $I \preccurlyeq \frac{1}{e}$ and get

$$eI \preccurlyeq \hat{D}_{t+1}^x \preccurlyeq \left(1 + \frac{2(1 - \beta_{t+1})\Gamma}{e}\right)\hat{D}_t^x.$$

Similar proof for the variable y. $\square$

19

C.2 Proof of Lemma 2

1) **OASIS:** $H^x = \text{diag}(q^x)$, where $q^x = (v^x \odot \nabla^2_{xx} f(x,y,\xi) v^x)$ with Rademacher vector v. H^x is diagonal from the definition. Moreover,

$$
\begin{aligned}
\|H^x\|^2_\infty = \|q^x\|^2_\infty &\leq \left(\max_i \left[\sum_{j=1}^{d_x} |(\nabla^2_{xx} f(x,y,\xi))_{ij}| \right] \right)^2 \\
&\leq \max_i \left(\sum_{j=1}^{d_x} |(\nabla^2_{xx} f(x,y,\xi))_{ij}| \right)^2 \\
&\leq \max_i \left[d_x \sum_{j=1}^{d_x} (\nabla^2_{xx} f(x,y,\xi))^2_{ij} \right] \\
&\leq d_x \sum_{i=1}^{d_x} \sum_{j=1}^{d_x} (\nabla^2_{xx} f(x,y,\xi))^2_{ij} \\
&= d_x \|\nabla^2_{xx} f(x,y,\xi)\|^2_2 \leq d_x L^2.
\end{aligned}
$$

Finally, we have $\|H^x\|_\infty \leq \sqrt{d_x} L = \Gamma$. Similar proof for the variable y.

2) **AdaHessian:** $(H^x)^2 = \text{diag}(v^x \odot v^x)$. The proof can be given in a similar way to the previous point.

3) **Adam:** $(H^x)^2 = \text{diag}(\nabla_x f(x,y,\xi) \odot \nabla_x f(x,y,\xi))$. H^x is diagonal from the definition. Moreover, using diagonal structure of H^x and $(H^x)^2$,

$$
\begin{aligned}
\|H^x\|^2_\infty = \|(H^x)^2\|_\infty &= \|\text{diag}(\nabla_x f(x,y,\xi) \odot \nabla_x f(x,y,\xi))\|_\infty \\
&= \|\nabla_x f(x,y,\xi) \odot \nabla_x f(x,y,\xi)\|_\infty \\
&= \|\nabla_x f(x,y,\xi)\|^2_\infty \\
&\leq \|\nabla_x f(x,y,\xi)\|^2 \leq M^2.
\end{aligned}
$$

Finally, we have $\|H^x\|_\infty \leq M = \Gamma$. Similar proof for the variable y.

4) **RMSProp:** $(H^x)^2 = \text{diag}(\nabla_x f(x,y,\xi) \odot \nabla_x f(x,y,\xi))$. The proof can be given in a similar way to the previous point. $\square$

C.3 Proof of Theorem 1

We start with additional notation:

$$
F(z_t) = F(x_t, y_t) = \begin{pmatrix} \nabla_x f(x_t, y_t) \\ -\nabla_y f(x_t, y_t) \end{pmatrix}, \quad g_t = \begin{pmatrix} \nabla_x f(x_t, y_t, \xi_t) \\ -\nabla_y f(x_t, y_t, \xi_t) \end{pmatrix},
$$

$$
\hat{V}_t = \begin{pmatrix} \hat{D}^x_t & 0 \\ 0 & \hat{D}^y_t \end{pmatrix}.
$$

For the matrices $\{\hat{V}_t\}$ we can prove an analogue of Lemma 1. With new notation Algorithm 1 can be rewritten as follows:

$$
\begin{aligned}
z_{t+1/2} &= z_t - \gamma_t (\hat{V}_t)^{-1} g_t \\
z_{t+1} &= z_t - \gamma_t (\hat{V}_t)^{-1} g_{t+1/2}.
\end{aligned}
$$

This is the notation we will use in the proof. Before prove Theorem, we consider the following lemma:

Lemma 4. Let $z, y \in \mathbb{R}^n$ and $D \in \mathcal{S}_{++}^d$. We set $z^+ = z - y$, then for all $u \in \mathbb{R}^n$:

$$\|z^+ - u\|_D^2 = \|z - u\|_D^2 - 2\langle y, z^+ - u\rangle_D - \|z^+ - z\|_D^2.$$

Proof of Lemma: Let us consider the following chain:

$$
\begin{aligned}
\|z^+ - u\|_D^2 &= \|z^+ - z + z - u\|_D^2 \\
&= \|z - u\|_D^2 + 2\langle z^+ - z, z - u\rangle_D + \|z^+ - z\|_D^2 \\
&= \|z - u\|_D^2 + 2\langle z^+ - z, z^+ - u\rangle_D - \|z^+ - z\|_D^2 \\
&= \|z - u\|_D^2 + 2\langle z^+ - (z - y), z^+ - u\rangle_D - 2\langle y, z^+ - u\rangle_D - \|z^+ - z\|_D^2 \\
&= \|z - u\|_D^2 - 2\langle y, z^+ - u\rangle_D - \|z^+ - z\|_D^2.
\end{aligned}
$$

$\square$

Proof of Theorem 1: Applying the previous Lemma with $z^+ = z_{t+1}$, $z = z_t$, $u = z$, $y = \gamma \hat{V}_t^{-1} g_{t+1/2}$ and $D = \hat{V}_t$, we get

$$
\begin{aligned}
\|z_{t+1} - z\|_{\hat{V}_t}^2 &= \|z_t - z\|_{\hat{V}_t}^2 - 2\gamma\langle \hat{V}_t^{-1} g_{t+1/2}, z_{t+1} - z\rangle_{\hat{V}_t} - \|z_{t+1} - z_t\|_{\hat{V}_t}^2 \\
&= \|z_t - z\|_{\hat{V}_t}^2 - 2\gamma\langle g_{t+1/2}, z_{t+1} - z\rangle - \|z_{t+1} - z_t\|_{\hat{V}_t}^2,
\end{aligned}
$$

and with $z^+ = z_{t+1/2}$, $z = z_t$, $u = z_{t+1}$, $y = \gamma V_t^{-1} g_t$, $D = \hat{V}_t$:

$$
\begin{aligned}
\|z_{t+1/2} - z_{t+1}\|_{\hat{V}_t}^2 &= \|z_t - z_{t+1}\|_{\hat{V}_t}^2 - 2\gamma\langle \hat{V}_t^{-1} g_t, z_{t+1/2} - z_{t+1}\rangle_{\hat{V}_t} - \|z_{t+1/2} - z_t\|_{\hat{V}_t}^2 \\
&= \|z_t - z_{t+1}\|_{\hat{V}_t}^2 - 2\gamma\langle g_t, z_{t+1/2} - z_{t+1}\rangle - \|z_{t+1/2} - z_t\|_{\hat{V}_t}^2.
\end{aligned}
$$

Next, we sum up the two previous equalities

$$
\begin{aligned}
\|z_{t+1} - z\|_{\hat{V}_t}^2 + \|z_{t+1/2} - z_{t+1}\|_{\hat{V}_t}^2 = &\|z_t - z\|_{\hat{V}_t}^2 - \|z_{t+1/2} - z_t\|_{\hat{V}_t}^2 \\
&- 2\gamma\langle g_{t+1/2}, z_{t+1} - z\rangle - 2\gamma\langle g_t, z_{t+1/2} - z_{t+1}\rangle.
\end{aligned}
$$

A small rearrangement gives

$$
\begin{aligned}
\|z_{t+1} - z\|_{\hat{V}_t}^2 &+ \|z_{t+1/2} - z_{t+1}\|_{\hat{V}_t}^2 \\
= &\|z_t - z\|_{\hat{V}_t}^2 - \|z_{t+1/2} - z_t\|_{\hat{V}_t}^2 \\
&- 2\gamma\langle g_{t+1/2}, z_{t+1/2} - z\rangle - 2\gamma\langle g_t - g_{t+1/2}, z_{t+1/2} - z_{t+1}\rangle \\
\leq &\|z_t - z\|_{\hat{V}_t}^2 - \|z_{t+1/2} - z_t\|_{\hat{V}_t}^2 - 2\gamma\langle g^{t+1/2}, z_{t+1/2} - z\rangle \\
&+ \gamma^2\|g_{t+1/2} - g_t\|_{\hat{V}_t^{-1}}^2 + \|z_{t+1/2} - z_{t+1}\|_{\hat{V}_t}^2.
\end{aligned}
$$

And then

$$
\begin{aligned}
\|z_{t+1} - z\|_{\hat{V}_t}^2 \leq &\|z_t - z\|_{\hat{V}_t}^2 - \|z_{t+1/2} - z_t\|_{\hat{V}_t}^2 - 2\gamma\langle g^{t+1/2}, z_{t+1/2} - z\rangle \\
&+ \gamma^2\|g_{t+1/2} - g_t\|_{\hat{V}_t^{-1}}^2.
\end{aligned}
$$

Next, we use that $\hat{V}_t^{-1} \preccurlyeq \frac{1}{e}I$ and get

$$
\begin{aligned}
\|z_{t+1} - z\|_{\hat{V}_t}^2 \leq &\|z_t - z\|_{\hat{V}_t}^2 - \|z_{t+1/2} - z_t\|_{\hat{V}_t}^2 - 2\gamma\langle g^{t+1/2}, z_{t+1/2} - z\rangle + \frac{\gamma^2}{e}\|g_{t+1/2} - g_t\|^2 \\
\leq &\|z_t - z\|_{\hat{V}_t}^2 - \|z_{t+1/2} - z_t\|_{\hat{V}_t}^2 - 2\gamma\langle g^{t+1/2}, z_{t+1/2} - z\rangle \\
&+ \frac{3\gamma^2}{e}\|F(z_{t+1/2}) - F(z_t)\|^2 + \frac{3\gamma^2}{e}\|g_{t+1/2} - F(z_{t+1/2})\|^2 \\
&+ \frac{3\gamma^2}{e}\|F(z_t) - g_t\|^2.
\end{aligned}
$$

Smoothness of function f (Assumption 2) gives

$$
\|z_{t+1} - z\|_{\hat{V}_t}^2 \leq \|z_t - z\|_{\hat{V}_t}^2 - \|z_{t+1/2} - z_t\|_{\hat{V}_t}^2 - 2\gamma\langle g^{t+1/2}, z_{t+1/2} - z\rangle
$$

$$+ \frac{3\gamma^2 L^2}{e}\|z_{t+1/2} - z_t\|^2 + \frac{3\gamma^2}{e}\|g_{t+1/2} - F(z_{t+1/2})\|^2$$

$$+ \frac{3\gamma^2}{e}\|F(z_t) - g_t\|^2$$

$$= \|z_t - z\|_{\hat{V}_t}^2 - \|z_{t+1/2} - z_t\|_{\hat{V}_t}^2 - 2\gamma\langle F(z_{t+1/2}), z_{t+1/2} - z\rangle$$

$$+ \frac{3\gamma^2 L^2}{e}\|z_{t+1/2} - z_t\|^2 + \frac{3\gamma^2}{e}\|g_{t+1/2} - F(z_{t+1/2})\|^2$$

$$+ \frac{3\gamma^2}{e}\|F(z_t) - g_t\|^2 + 2\gamma\langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} - z\rangle. \tag{21}$$

Strongly convex–strongly concave case. We substitute $z = z^*$ and take the total expectation of both sides of the equation

$$\mathbb{E}\left[\|z_{t+1} - z^*\|_{\hat{V}_t}^2\right] \leq \mathbb{E}\left[\|z_t - z^*\|_{\hat{V}_t}^2\right] - \mathbb{E}\left[\|z_{t+1/2} - z_t\|_{\hat{V}_t}^2\right] - 2\gamma\mathbb{E}\left[\langle F(z_{t+1/2}), z_{t+1/2} - z^*\rangle\right]$$

$$+ \frac{3\gamma^2 L^2}{e}\mathbb{E}\left[\|z_{t+1/2} - z_t\|^2\right] + \frac{3\gamma^2}{e}\mathbb{E}\left[\|g_{t+1/2} - F(z_{t+1/2})\|^2\right]$$

$$+ \frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_t) - g_t\|^2\right] + 2\gamma\mathbb{E}\left[\langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} - z^*\rangle\right].$$

Next, using the property of the solution z^*: $\langle F(z^*), z_{t+1/2} - z^*\rangle = 0$, and get

$$\mathbb{E}\left[\|z_{t+1} - z^*\|_{\hat{V}_t}^2\right] \leq \mathbb{E}\left[\|z_t - z^*\|_{\hat{V}_t}^2\right] - \mathbb{E}\left[\|z_{t+1/2} - z_t\|_{\hat{V}_t}^2\right]$$

$$- 2\gamma\mathbb{E}\left[\langle F(z_{t+1/2}) - F(z^*), z_{t+1/2} - z^*\rangle\right]$$

$$+ \frac{3\gamma^2 L^2}{e}\mathbb{E}\left[\|z_{t+1/2} - z_t\|^2\right] + \frac{3\gamma^2}{e}\mathbb{E}\left[\|g_{t+1/2} - F(z_{t+1/2})\|^2\right]$$

$$+ \frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_t) - g_t\|^2\right] + 2\gamma\mathbb{E}\left[\langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} - z^*\rangle\right]$$

$$= \mathbb{E}\left[\|z_t - z^*\|_{\hat{V}_t}^2\right] - \mathbb{E}\left[\|z_{t+1/2} - z_t\|_{\hat{V}_t}^2\right]$$

$$- 2\gamma\mathbb{E}\left[\langle F(z_{t+1/2}) - F(z^*), z_{t+1/2} - z^*\rangle\right] + \frac{3\gamma^2 L^2}{e}\mathbb{E}\left[\|z_{t+1/2} - z_t\|^2\right]$$

$$+ \frac{3\gamma^2}{e}\mathbb{E}\left[\mathbb{E}_{t+1/2}\left[\|g_{t+1/2} - F(z_{t+1/2})\|^2\right]\right] + \frac{3\gamma^2}{e}\mathbb{E}\left[\mathbb{E}_t\left[\|F(z_t) - g_t\|^2\right]\right]$$

$$+ 2\gamma\mathbb{E}\left[\langle\mathbb{E}_{t+1/2}\left[F(z_{t+1/2}) - g_{t+1/2}\right], z_{t+1/2} - z^*\rangle\right].$$

Assumptions 1 and 3 on strongly convexity - strongly concavity and on stochastisity give

$$\mathbb{E}\left[\|z_{t+1} - z^*\|_{\hat{V}_t}^2\right] \leq \mathbb{E}\left[\|z_t - z^*\|_{\hat{V}_t}^2\right] - \mathbb{E}\left[\|z_{t+1/2} - z_t\|_{\hat{V}_t}^2\right]$$

$$- 2\gamma\mu\mathbb{E}\left[\|z_{t+1/2} - z^*\|^2\right] + \frac{3\gamma^2 L^2}{e}\mathbb{E}\left[\|z_{t+1/2} - z_t\|^2\right]$$

$$+ \frac{6\gamma^2\sigma^2}{eb}$$

$$\leq \mathbb{E}\left[\|z_t - z^*\|_{\hat{V}_t}^2\right] - \mathbb{E}\left[\|z_{t+1/2} - z_t\|_{\hat{V}_t}^2\right] - \gamma\mu\mathbb{E}\left[\|z_t - z^*\|^2\right]$$

$$+ \left(2\gamma\mu + \frac{3\gamma^2 L^2}{e}\right)\mathbb{E}\left[\|z_{t+1/2} - z_t\|^2\right] + \frac{6\gamma^2\sigma^2}{eb}.$$

With $\frac{1}{\Gamma}\hat{V}_t \preccurlyeq I \preccurlyeq \frac{1}{e}\hat{V}_t$, we get

$$\mathbb{E}\left[\|z_{t+1} - z^*\|_{\hat{V}_t}^2\right] \leq \left(1 - \frac{\gamma\mu}{\Gamma}\right)\mathbb{E}\left[\|z_t - z^*\|_{\hat{V}_t}^2\right]$$

$$- \left(1 - \frac{2\gamma\mu}{e} - \frac{3\gamma^2 L^2}{e^2}\right)\mathbb{E}\left[\|z_{t+1/2} - z_t\|_{\hat{V}_t}^2\right] + \frac{6\gamma^2\sigma^2}{eb}.$$

The choice of $\gamma \leq \min\left\{\frac{e}{4\mu}; \frac{e}{3L}\right\}$ move us to

$$\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_t}\right] \leq \left(1 - \frac{\gamma\mu}{\Gamma}\right)\mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_t}\right] + \frac{6\gamma^2\sigma^2}{eb}.$$

Lemma 1 with $C = \frac{\Gamma^2}{2e^2}$ for (7) and $C = \frac{2\Gamma}{e}$ for (8) gives

$$\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_{t+1}}\right] \leq (1 + (1 - \beta_{t+1})C)\,\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_t}\right].$$

Then we have

$$\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_{t+1}}\right] \leq \left(1 - \frac{\gamma\mu}{\Gamma}\right)(1 + (1 - \beta_{t+1})C)\,\mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_t}\right]$$
$$+ \frac{6\gamma^2\sigma^2}{eb}(1 + (1 - \beta_{t+1})C).$$

This completes the proof of the strongly convex-strongly concave case. $\square$

Convex–concave case. We start from (21) with small rearrangement

$$2\gamma\langle F(z_{t+1/2}), z_{t+1/2} - z\rangle \leq \|z_t - z\|^2_{\hat{V}_t} - \|z_{t+1} - z\|^2_{\hat{V}_t} - \|z_{t+1/2} - z_t\|^2_{\hat{V}_t}$$
$$+ \frac{3\gamma^2 L^2}{e}\|z_{t+1/2} - z_t\|^2 + \frac{3\gamma^2}{e}\|g_{t+1/2} - F(z_{t+1/2})\|^2$$
$$+ \frac{3\gamma^2}{e}\|F(z_t) - g_t\|^2 + 2\gamma\langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} - z\rangle.$$

With $I \preccurlyeq \frac{1}{e}\hat{V}_t$ and Lemma 1 with $C = \frac{\Gamma^2}{2e^2}$ for (7) and $C = \frac{2\Gamma}{e}$ for (8), we get

$$2\gamma\langle F(z_{t+1/2}), z_{t+1/2} - z\rangle \leq \|z_t - z\|^2_{\hat{V}_t} - (1 + (1 - \beta_{t+1})C)^{-1}\|z_{t+1} - z\|^2_{\hat{V}_{t+1}}$$
$$- \left(1 - \frac{3\gamma^2 L^2}{e^2}\right)\|z_{t+1/2} - z_t\|^2_{\hat{V}_t} + \frac{3\gamma^2}{e}\|g_{t+1/2} - F(z_{t+1/2})\|^2$$
$$+ \frac{3\gamma^2}{e}\|F(z_t) - g_t\|^2 + 2\gamma\langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} - z\rangle.$$

Next, we sum over all t from 0 to $T - 1$ and take maximum on $\mathcal{Z}$

$$2\gamma\max_{z\in\mathcal{Z}}\frac{1}{T}\sum_{t=0}^{T-1}\langle F(z_{t+1/2}), z_{t+1/2} - z\rangle$$
$$\leq \max_{z\in\mathcal{Z}}\frac{1}{T}\sum_{t=0}^{T-1}\left[\|z_t - z\|^2_{\hat{V}_t} - (1 + (1 - \beta_{t+1})C)^{-1}\|z_{t+1} - z\|^2_{\hat{V}_{t+1}}\right]$$
$$- \left(1 - \frac{3\gamma^2 L^2}{e^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\left[\|z_{t+1/2} - z_t\|^2_{\hat{V}_t}\right] + \frac{3\gamma^2}{e}\frac{1}{T}\sum_{t=0}^{T-1}\left[\|g_{t+1/2} - F(z_{t+1/2})\|^2\right]$$
$$+ \frac{3\gamma^2}{e}\frac{1}{T}\sum_{t=0}^{T-1}\left[\|F(z_t) - g_t\|^2\right] + \max_{z\in\mathcal{Z}}\frac{1}{T}\sum_{t=0}^{T-1}\left[2\gamma\langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} - z\rangle\right].$$
$$(22)$$

Then, by $x_T^{avg} = \frac{1}{T}\sum_{t=0}^{T-1}x_{t+1/2}$ and $y_T^{avg} = \frac{1}{T}\sum_{t=0}^{T-1}y_{t+1/2}$, Jensen's inequality and convexity-concavity of f:

$$\text{gap}(z_T^{avg}) \leq \max_{y'\in\mathcal{Y}}f\left(\frac{1}{T}\left(\sum_{t=0}^{T-1}x_{t+1/2}\right), y'\right) - \min_{x'\in\mathcal{X}}f\left(x', \frac{1}{T}\left(\sum_{t=0}^{T-1}y_{t+1/2}\right)\right)$$
$$\leq \max_{y'\in\mathcal{Y}}\frac{1}{T}\sum_{t=0}^{T-1}f(x_{t+1/2}, y') - \min_{x'\in\mathcal{X}}\frac{1}{T}\sum_{t=0}^{T-1}f(x', y_{t+1/2}).$$

Given the fact of linear independence of x' and y':

$$\text{gap}(z_T^{avg}) \leq \max_{(x',y')\in\mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \left(f(x_{t+1/2}, y') - f(x', y_{t+1/2}) \right).$$

Using convexity and concavity of the function f:

$$\text{gap}(z_T^{avg}) \leq \max_{(x',y')\in\mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \left(f(x_{t+1/2}, y') - f(x', y_{t+1/2}) \right)$$

$$= \max_{(x',y')\in\mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \left(f(x_{t+1/2}, y') - f(x_{t+1/2}, y_{t+1/2}) + f(x_{t+1/2}, y_{t+1/2}) - f(x', y_{t+1/2}) \right)$$

$$\leq \max_{(x',y')\in\mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \left(\langle \nabla_y f(x_{t+1/2}, y_{t+1/2}), y' - y_{t+1/2} \rangle + \langle \nabla_x f(x_{t+1/2}, y_{t+1/2}), x_{t+1/2} - x' \rangle \right)$$

$$\leq \max_{z\in\mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \langle F(z_{t+1/2}), z_{t+1/2} - z \rangle. \tag{23}$$

Combining (22) and (23), we get

$$2\gamma\text{gap}(z_T^{avg}) \leq \max_{z\in\mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|z_t - z\|_{\hat{V}_t}^2 - (1 + (1 - \beta_{t+1})C)^{-1} \|z_{t+1} - z\|_{\hat{V}_{t+1}}^2 \right]$$

$$- \left(1 - \frac{3\gamma^2 L^2}{e^2} \right) \frac{1}{T} \sum_{t=0}^{T-1} \left[\|z_{t+1/2} - z_t\|_{\hat{V}_t}^2 \right] + \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|g_{t+1/2} - F(z_{t+1/2})\|^2 \right]$$

$$+ \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|F(z_t) - g_t\|^2 \right] + \max_{z\in\mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \left[2\gamma\langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} - z \rangle \right].$$

The choice of $\gamma \leq \frac{e}{2L}$ gives

$$2\gamma \cdot \text{gap}(z_T^{avg}) \leq \max_{z\in\mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|z_t - z\|_{\hat{V}_t}^2 - (1 + (1 - \beta_{t+1})C)^{-1} \|z_{t+1} - z\|_{\hat{V}_{t+1}}^2 \right]$$

$$+ \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|g_{t+1/2} - F(z_{t+1/2})\|^2 \right] + \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|F(z_t) - g_t\|^2 \right]$$

$$+ \max_{z\in\mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \left[2\gamma\langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} - z \rangle \right]$$

$$\leq \max_{z\in\mathcal{Z}} \left[\frac{\|z_0 - z\|_{\hat{V}_0}^2}{T} + (1 + (1 - \beta_{t+1})C)^{-1} (1 - \beta_{t+1})C \cdot \frac{1}{T} \sum_{t=1}^{T-1} \|z_t - z\|_{\hat{V}_t}^2 \right]$$

$$+ \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|g_{t+1/2} - F(z_{t+1/2})\|^2 \right] + \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|F(z_t) - g_t\|^2 \right]$$

$$+ \frac{1}{T} \sum_{t=0}^{T-1} \left[2\gamma\langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} \rangle \right]$$

$$+ \max_{z\in\mathcal{Z}} \left[2\gamma\langle \frac{1}{T} \sum_{t=0}^{T-1} [F(z_{t+1/2}) - g_{t+1/2}], -z \rangle \right].$$

One can note that $(1 + (1 - \beta_{t+1})C) \geq 1$, and get

$$2\gamma \cdot \text{gap}(z_T^{avg}) \leq \max_{z\in\mathcal{Z}} \left[\frac{\|z_0 - z\|_{\hat{V}_0}^2}{T} + \frac{1}{T} \sum_{t=1}^{T-1} (1 - \beta_{t+1})C\|z_t - z\|_{\hat{V}_t}^2 \right]$$

$$+ \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|g_{t+1/2} - F(z_{t+1/2})\|^2 \right] + \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|F(z_t) - g_t\|^2 \right]$$

$$+ \frac{1}{T} \sum_{t=0}^{T-1} \left[2\gamma \langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} \rangle \right]$$

$$+ \max_{z \in \mathcal{Z}} \left[2\gamma \langle \frac{1}{T} \sum_{t=0}^{T-1} [F(z_{t+1/2}) - g_{t+1/2}], -z \rangle \right].$$

Next, we use $\hat{V}_t \preccurlyeq \Gamma I$ and the fact that iterations and the set $\mathcal{Z}$ are bounded: $\|z_t\|, \|z\| \leq \Omega$

$$2\gamma \cdot \text{gap}(z_T^{avg}) \leq \max_{z \in \mathcal{Z}} \left[\frac{\Gamma\|z_0 - z\|^2}{T} + \frac{1}{T} \sum_{t=1}^{T-1} (1 - \beta_{t+1}) C\Gamma\|z_t - z\|^2 \right]$$

$$+ \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|g_{t+1/2} - F(z_{t+1/2})\|^2 \right] + \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|F(z_t) - g_t\|^2 \right]$$

$$+ \frac{1}{T} \sum_{t=0}^{T-1} \left[2\gamma \langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} \rangle \right]$$

$$+ \max_{z \in \mathcal{Z}} \left[2\gamma \langle \frac{1}{T} \sum_{t=0}^{T-1} [F(z_{t+1/2}) - g_{t+1/2}], -z \rangle \right]$$

$$\leq \frac{4\Gamma\Omega^2}{T} + \frac{1}{T} \sum_{t=1}^{T-1} 4(1 - \beta_{t+1}) C\Gamma\Omega^2$$

$$+ \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|g_{t+1/2} - F(z_{t+1/2})\|^2 \right] + \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|F(z_t) - g_t\|^2 \right]$$

$$+ \frac{1}{T} \sum_{t=0}^{T-1} \left[2\gamma \langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} \rangle \right]$$

$$+ \max_{z \in \mathcal{Z}} \left[2\gamma \langle \frac{1}{T} \sum_{t=0}^{T-1} [F(z_{t+1/2}) - g_{t+1/2}], -z \rangle \right].$$

Next, we take a full expectation and use Assumption 3

$$2\gamma \cdot \mathbb{E}\left[\text{gap}(z_T^{avg})\right] \leq \frac{4\Gamma\Omega^2}{T} + \frac{1}{T} \sum_{t=1}^{T-1} 4(1 - \beta_{t+1}) C\Gamma\Omega^2$$

$$+ \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \left[\mathbb{E}\left[\|g_{t+1/2} - F(z_{t+1/2})\|^2 \right] \right]$$

$$+ \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \left[\mathbb{E}\left[\|F(z_t) - g_t\|^2 \right] \right]$$

$$+ \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[2\gamma \langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} \rangle \right]$$

$$+ \mathbb{E}\left[\max_{z \in \mathcal{Z}} \left[2\gamma \langle \frac{1}{T} \sum_{t=0}^{T-1} [F(z_{t+1/2}) - g_{t+1/2}], -z \rangle \right] \right]$$

$$\leq \frac{4\Gamma\Omega^2}{T} + \frac{1}{T} \sum_{t=1}^{T-1} 4(1 - \beta_{t+1}) C\Gamma\Omega^2 + \frac{6\gamma^2\sigma^2}{eb}$$

$$+ \mathbb{E}\left[\max_{z \in \mathcal{Z}} \left[2\gamma \langle \frac{1}{T} \sum_{t=0}^{T-1} [F(z_{t+1/2}) - g_{t+1/2}], -z \rangle \right] \right]$$

$$\leq \frac{4\Gamma\Omega^2}{T} + \frac{1}{T} \sum_{t=1}^{T-1} 4(1-\beta_{t+1})C\Gamma\Omega^2 + \frac{6\gamma^2\sigma^2}{eb}$$

$$+ \mathbb{E}\left[\max_{z \in \mathcal{Z}} \left[\frac{1}{T} \left\| \gamma \sum_{t=0}^{T-1} F(z_{t+1/2}) - g_{t+1/2} \right\|_{\hat{V}_t^{-1}}^2 + \frac{\|z\|_{\hat{V}_t}^2}{T} \right] \right]$$

$$\leq \frac{4\Gamma\Omega^2}{T} + \frac{1}{T} \sum_{t=1}^{T-1} 4(1-\beta_{t+1})C\Gamma\Omega^2 + \frac{6\gamma^2\sigma^2}{eb}$$

$$+ \mathbb{E}\left[\max_{z \in \mathcal{Z}} \left[\frac{1}{eT} \left\| \gamma \sum_{t=0}^{T-1} F(z_{t+1/2}) - g_{t+1/2} \right\|^2 + \frac{\Gamma\|z\|^2}{T} \right] \right]. \qquad (24)$$

With unbiasedness and independence of all $F(z_{t+1/2}) - g_{t+1/2}$, we get

$$2\gamma \cdot \mathbb{E}\left[\text{gap}(z_T^{avg})\right] \leq \frac{4\Gamma\Omega^2}{T} + \frac{1}{T} \sum_{t=1}^{T-1} 4(1-\beta_{t+1})C\Gamma\Omega^2 + \frac{6\gamma^2\sigma^2}{eb}$$

$$+ \max_{z \in \mathcal{Z}} \frac{\Gamma\|z\|^2}{T} + \frac{\gamma^2}{eT} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| F(z_{t+1/2}) - g_{t+1/2} \right\|^2\right]$$

$$\leq \frac{5\Gamma\Omega^2}{T} + \frac{1}{T} \sum_{t=1}^{T-1} 4(1-\beta_{t+1})C\Gamma\Omega^2 + \frac{7\gamma^2\sigma^2}{eb}.$$

Here we additionally use Assumption 3. Finally,

$$\mathbb{E}\left[\text{gap}(z_T^{avg})\right] \leq \frac{3\Gamma\Omega^2}{\gamma T} + \frac{2C\Gamma\Omega^2}{\gamma T} \sum_{t=2}^{T}(1-\beta_t) + \frac{4\gamma\sigma^2}{eb}.$$

This completes the proof of the convex-concave case. $\square$

Non-convex–non-concave case. We start from (21), substitute $z = z^*$ and take the total expectation of both sides of the equation

$$\mathbb{E}\left[\|z_{t+1} - z^*\|_{\hat{V}_t}^2\right] \leq \mathbb{E}\left[\|z_t - z^*\|_{\hat{V}_t}^2\right] - \mathbb{E}\left[\|z_{t+1/2} - z_t\|_{\hat{V}_t}^2\right] - 2\gamma\mathbb{E}\left[\langle F(z_{t+1/2}), z_{t+1/2} - z^* \rangle\right]$$

$$+ \frac{3\gamma^2 L^2}{e} \mathbb{E}\left[\|z_{t+1/2} - z_t\|^2\right] + \frac{3\gamma^2}{e} \mathbb{E}\left[\|g_{t+1/2} - F(z_{t+1/2})\|^2\right]$$

$$+ \frac{3\gamma^2}{e} \mathbb{E}\left[\|F(z_t) - g_t\|^2\right] + 2\gamma\mathbb{E}\left[\langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} - z^* \rangle\right]$$

$$= \mathbb{E}\left[\|z_t - z^*\|_{\hat{V}_t}^2\right] - \mathbb{E}\left[\|z_{t+1/2} - z_t\|_{\hat{V}_t}^2\right]$$

$$- 2\gamma\mathbb{E}\left[\langle F(z_{t+1/2}), z_{t+1/2} - z^* \rangle\right] + \frac{3\gamma^2 L^2}{e} \mathbb{E}\left[\|z_{t+1/2} - z_t\|^2\right]$$

$$+ \frac{3\gamma^2}{e} \mathbb{E}\left[\mathbb{E}_{t+1/2}\left[\|g_{t+1/2} - F(z_{t+1/2})\|^2\right]\right] + \frac{3\gamma^2}{e} \mathbb{E}\left[\mathbb{E}_t\left[\|F(z_t) - g_t\|^2\right]\right]$$

$$+ 2\gamma\mathbb{E}\left[\langle \mathbb{E}_{t+1/2}\left[F(z_{t+1/2}) - g_{t+1/2}\right], z_{t+1/2} - z^* \rangle\right].$$

Assumptions 1 and 3 on non-convexity–non-concavity and on stochastisity give

$$\mathbb{E}\left[\|z_{t+1} - z^*\|_{\hat{V}_t}^2\right] \leq \mathbb{E}\left[\|z_t - z^*\|_{\hat{V}_t}^2\right] - \mathbb{E}\left[\|z_{t+1/2} - z_t\|_{\hat{V}_t}^2\right]$$

$$+ \frac{3\gamma^2 L^2}{e} \mathbb{E}\left[\|z_{t+1/2} - z_t\|^2\right] + \frac{6\gamma^2\sigma^2}{eb}.$$

With $eI \preccurlyeq \hat{V}_t$, we get

$$\mathbb{E}\left[\|z_{t+1} - z^*\|_{\hat{V}_t}^2\right] \leq \mathbb{E}\left[\|z_t - z^*\|_{\hat{V}_t}^2\right] - \left(1 - \frac{3\gamma^2 L^2}{e^2}\right) \mathbb{E}\left[\|z_{t+1/2} - z_t\|_{\hat{V}_t}^2\right] + \frac{6\gamma^2\sigma^2}{eb}$$

$$= \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_t}\right] - \left(1 - \frac{3\gamma^2 L^2}{e^2}\right)\gamma^2 \mathbb{E}\left[\|g_t\|^2_{\hat{V}_t^{-1}}\right] + \frac{6\gamma^2\sigma^2}{eb}.$$

With $\gamma \leq \frac{e}{3L}$ we get

$$\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_t}\right] \leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_t}\right] - \frac{2\gamma^2}{3}\mathbb{E}\left[\|g_t\|^2_{\hat{V}_t^{-1}}\right] + \frac{6\gamma^2\sigma^2}{eb}.$$

Lemma 1 with $C = \frac{\Gamma^2}{2e^2}$ for (7) and $C = \frac{2\Gamma}{e}$ for (8) gives

$$\begin{aligned}
(1 + (1 - \beta_{t+1})C)^{-1}\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_{t+1}}\right] &\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_t}\right] - \frac{2\gamma^2}{3}\mathbb{E}\left[\|g_t\|^2_{\hat{V}_t^{-1}}\right] + \frac{6\gamma^2\sigma^2}{eb} \\
&\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_t}\right] - \frac{2\gamma^2}{3}\mathbb{E}\left[\|g_t\|^2_{\hat{V}_t^{-1}}\right] + \frac{6\gamma^2\sigma^2}{eb} \\
&\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_t}\right] - \frac{2\gamma^2}{3\Gamma}\mathbb{E}\left[\|g_t\|^2\right] + \frac{6\gamma^2\sigma^2}{eb}.
\end{aligned}$$

In the last step we use that $\frac{1}{\Gamma}I \preccurlyeq \hat{V}_t^{-1}$. Then, using Assumption 3, we have

$$\begin{aligned}
(1 + (1 - \beta_{t+1})C)^{-1}\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_{t+1}}\right] &\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_t}\right] - \frac{\gamma^2}{3\Gamma}\mathbb{E}\left[\|F(z_t)\|^2\right] \\
&\quad + \frac{2\gamma^2}{3\Gamma}\mathbb{E}\left[\|g_t - F(z_t)\|^2\right] + \frac{6\gamma^2\sigma^2}{eb} \\
&\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_t}\right] - \frac{\gamma^2}{3\Gamma}\mathbb{E}\left[\|F(z_t)\|^2\right] \\
&\quad + \frac{7\gamma^2\sigma^2}{eb}.
\end{aligned}$$

Small rearrangement gives

$$\frac{\gamma^2}{3\Gamma}\mathbb{E}\left[\|F(z_t)\|^2\right] \leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_t}\right] - (1 + (1 - \beta_{t+1})C)^{-1}\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_{t+1}}\right] + \frac{7\gamma^2\sigma^2}{eb}.$$

Summing over all t from 0 to $T-1$ and averaging, we get

$$\begin{aligned}
\frac{\gamma^2}{3\Gamma} \cdot \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|F(z_t)\|^2\right] &\leq \frac{\mathbb{E}\left[\|z_0 - z^*\|^2_{\hat{V}_0}\right]}{T} + \frac{7\gamma^2\sigma^2}{eb} \\
&\quad + \frac{1}{T}\sum_{t=1}^{T-1}(1 + (1 - \beta_{t+1})C)^{-1}(1 - \beta_{t+1})C\mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_t}\right] \\
&\leq \frac{\Gamma\|z_0 - z^*\|^2}{T} + \frac{7\gamma^2\sigma^2}{eb} \\
&\quad + \frac{1}{T}\sum_{t=1}^{T-1}(1 + (1 - \beta_{t+1})C)^{-1}(1 - \beta_{t+1})C\Gamma\mathbb{E}\left[\|z_t - z^*\|^2\right].
\end{aligned}$$

Here we use $\hat{V}_t \preccurlyeq \Gamma I$. With the fact that iterations are bounded: $\|z_t\| \leq \Omega$, we obtain

$$\begin{aligned}
\frac{\gamma^2}{3\Gamma} \cdot \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|F(z_t)\|^2\right] &\leq \frac{4\Gamma\Omega^2}{T} + \frac{7\gamma^2\sigma^2}{eb} \\
&\quad + \frac{1}{T}\sum_{t=1}^{T-1}(1 + (1 - \beta_{t+1})C)^{-1} \cdot 4(1 - \beta_{t+1})C\Gamma\Omega^2.
\end{aligned}$$

One can note that $(1 + (1 - \beta_{t+1})C) \geq 1$, and get

$$\frac{\gamma^2}{3\Gamma} \cdot \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|F(z_t)\|^2\right] \leq \frac{4\Gamma\Omega^2}{T} + \frac{7\gamma^2\sigma^2}{eb} + 4C\Gamma\Omega^2 \cdot \frac{1}{T}\sum_{t=1}^{T-1}(1 - \beta_{t+1}).$$

Finally,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|F(z_t)\|^2\right] \leq \frac{12\Gamma^2\Omega^2}{\gamma^2 T} + \frac{21\Gamma\sigma^2}{eb} + \frac{12C\Gamma^2\Omega^2}{\gamma^2}\cdot\frac{1}{T}\sum_{t=1}^{T}(1-\beta_t).$$

If we choose $\bar{z}_T$ randomly and uniformly, then $\mathbb{E}\left[\|F(\bar{z}_T)\|^2\right] \leq \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|F(z_t)\|^2\right]$ and

$$\mathbb{E}\left[\|F(\bar{z}_T)\|^2\right] \leq \frac{12\Gamma^2\Omega^2}{\gamma^2 T} + \frac{21\Gamma\sigma^2}{eb} + \frac{12C\Gamma^2\Omega^2}{\gamma^2}\cdot\frac{1}{T}\sum_{t=1}^{T}(1-\beta_t).$$

This completes the proof of the non-convex–non-concave case. $\square$

C.4 PROOF OF COROLLARY 1

Strongly convex–strongly concave case. With $\beta_t \equiv \beta \geq 1 - \frac{\gamma\mu}{2\Gamma C}$, we have (10) in the following form

$$\mathbb{E}\left[R_{t+1}^2\right] \leq \left(1 - \frac{\gamma\mu}{2\Gamma}\right)\mathbb{E}\left[R_t^2\right] + \frac{12\gamma^2\sigma^2}{eb}.$$

Running the recursion, we have

$$\mathbb{E}\left[R_T^2\right] \leq \left(1-\frac{\gamma\mu}{2\Gamma}\right)^T\mathbb{E}\left[R_0^2\right] + \frac{12\gamma^2\sigma^2}{eb}\sum_{t=0}^{T-1}\left(1-\frac{\gamma\mu}{2\Gamma}\right)^t$$

$$\leq \left(1-\frac{\gamma\mu}{2\Gamma}\right)^T\mathbb{E}\left[R_0^2\right] + \frac{12\gamma^2\sigma^2}{eb}\sum_{t=0}^{\infty}\left(1-\frac{\gamma\mu}{2\Gamma}\right)^t$$

$$\leq \left(1-\frac{\gamma\mu}{2\Gamma}\right)^T\mathbb{E}\left[R_0^2\right] + \frac{24\Gamma\gamma\sigma^2}{eb\mu}$$

$$\leq \exp\left(-\frac{\gamma\mu T}{2\Gamma}\right)\mathbb{E}\left[R_0^2\right] + \frac{24\Gamma\gamma\sigma^2}{eb\mu}.$$

Finally, we need tuning of $\gamma \leq \frac{e}{4L}$:

- If $\frac{e}{4L} \geq \frac{2\Gamma\ln\left(\max\{2,eb\mu^2 R_0^2 T/(48\Gamma^2\sigma^2)\}\right)}{\mu T}$ then $\gamma = \frac{2\Gamma\ln\left(\max\{2,eb\mu^2 R_0^2 T/(48\Gamma^2\sigma^2)\}\right)}{\mu T}$ gives

$$\tilde{\mathcal{O}}\left(\exp\left(-\ln\left(\max\{2,eb\mu^2 R_0^2 T/(48\Gamma^2\sigma^2)\}\right)\right)R_0^2 + \frac{\Gamma^2\sigma^2}{e\mu^2 T}\right) = \tilde{\mathcal{O}}\left(\frac{\Gamma^2\sigma^2}{eb\mu^2 T}\right).$$

- If $\frac{e}{4L} \leq \frac{2\Gamma\ln\left(\max\{2,eb\mu^2 R_0^2 T/(48\Gamma^2\sigma^2)\}\right)}{\mu T}$ then $\gamma = \frac{e}{4L}$ gives

$$\tilde{\mathcal{O}}\left(\exp\left(-\frac{e\mu T}{8\Gamma L}\right)R_0^2 + \frac{\Gamma\gamma\sigma^2}{eb\mu}\right) \leq \tilde{\mathcal{O}}\left(\exp\left(-\frac{e\mu T}{8\Gamma L}\right)R_0^2 + \frac{\Gamma^2\sigma^2}{eb\mu^2 T}\right).$$

What in the end gives that

$$\mathbb{E}\left[R_T^2\right] = \tilde{\mathcal{O}}\left(\exp\left(-\frac{e\mu T}{8\Gamma L}\right)R_0^2 + \frac{\Gamma^2\sigma^2}{eb\mu^2 T}\right).$$

This completes the proof of the strongly convex–strongly concave case. $\square$

Convex–concave case. With $\beta_t \equiv \beta \geq 1 - \frac{1}{CT}$, we have (11) in the following form

$$\mathbb{E}\left[\text{gap}(x_T^{avg}, y_T^{avg})\right] \leq \frac{3\Gamma\Omega^2}{\gamma T} + \frac{2\Gamma\Omega^2}{\gamma T} + \frac{4\gamma\sigma^2}{eb}.$$

If $\gamma = \min\left\{\frac{e}{2L}; \frac{\sqrt{\Gamma eb}\Omega}{\sigma\sqrt{T}}\right\}$, then

$$\mathbb{E}\left[\text{gap}(x_T^{avg}, y_T^{avg})\right] = \mathcal{O}\left(\frac{\Gamma L\Omega^2}{eT} + \frac{\sqrt{\Gamma}\sigma\Omega}{\sqrt{ebT}}\right).$$

28

This completes the proof of the convex–concave case. $\square$

Non-convex–non-concave case. With $\beta_t \equiv \beta \geq 1 - \frac{1}{CT}$, we have (12) in the following form

$$\mathbb{E}\left[\|\nabla_x f(\bar{x}_T, \bar{y}_T)\|^2 + \|\nabla_y f(\bar{x}_T, \bar{y}_T)\|^2\right] \leq \frac{24\Gamma^2\Omega^2}{\gamma^2 T} + \frac{21\Gamma\sigma^2}{eb}.$$

The choice of $\gamma = \frac{e}{3L}$ gives

$$\mathbb{E}\left[\|\nabla_x f(\bar{x}_T, \bar{y}_T)\|^2 + \|\nabla_y f(\bar{x}_T, \bar{y}_T)\|^2\right] \leq \frac{240\Gamma^2 L^2\Omega^2}{e^2 T} + \frac{21\Gamma\sigma^2}{eb}.$$

The batch size $b \sim \frac{\Gamma\sigma^2}{e\varepsilon^2}$ completes the proof of the non-convex–non-concave case. $\square$

C.5 Proof of Corollary 2

Strongly convex–strongly concave case. We start from (10) with the small rearrangement

$$\frac{\gamma\mu}{2\Gamma}\mathbb{E}\left[R_t^2\right] \leq \left(1 - \frac{\gamma\mu}{2\Gamma} + (1 - \beta_{t+1})C\right)\mathbb{E}\left[R_t^2\right] - \mathbb{E}\left[R_{t+1}^2\right] + \frac{6\gamma^2\sigma^2}{eb}\left(1 + (1 - \beta_{t+1})C\right).$$

Summing over all t from 0 to $T - 1$ and averaging, we get

$$\frac{\gamma\mu}{2\Gamma T}\mathbb{E}\left[\sum_{t=0}^{T-1}\|z_t - z^*\|_{\hat{V}_t}^2\right] \leq \frac{\mathbb{E}\left[\|z_0 - z^*\|_{\hat{V}_0}^2\right]}{T} + \frac{1}{T}\sum_{t=0}^{T-1}\left((1 - \beta_{t+1})C - \frac{\gamma\mu}{2\Gamma}\right)\mathbb{E}\left[\|z_t - z^*\|_{\hat{V}_t}^2\right]$$
$$+ \frac{6\gamma^2\sigma^2}{eb} \cdot \frac{1}{T}\sum_{t=0}^{T-1}\left(1 + (1 - \beta_{t+1})C\right).$$

Using $eI \preccurlyeq \hat{V}_t \preccurlyeq \Gamma I$ and $\|z_t\| \leq \Omega$, we have

$$\frac{e\gamma\mu}{2\Gamma T}\mathbb{E}\left[\sum_{t=0}^{T-1}\|z_t - z^*\|^2\right] \leq \frac{4\Gamma\Omega^2}{T} + \frac{1}{T}\sum_{t=1}^{T}\left((1 - \beta_t)C - \frac{\gamma\mu}{2\Gamma}\right)\mathbb{E}\left[\|z_{t-1} - z^*\|_{\hat{V}_{t-1}}^2\right]$$
$$+ \frac{6\gamma^2\sigma^2}{eb} \cdot \frac{1}{T}\sum_{t=1}^{T}\left(1 + (1 - \beta_t)C\right).$$

Let us define $\alpha = \left(\frac{2\Gamma C}{\gamma\mu}\right)^2$. Then, $1 - \beta_t = \frac{1-\beta}{1-\beta^{t+1}}$ with $\beta = 1 - \frac{1}{\alpha}$. (18) gives

$$\beta^{\sqrt{\alpha}} \leq 1 - \frac{1}{2\sqrt{\alpha}}.$$

And hence, for all $t \geq \sqrt{\alpha}$

$$1 - \beta_t \leq \frac{2\sqrt{\alpha}}{\alpha} = \frac{2}{\sqrt{\alpha}} = \frac{\gamma\mu}{2\Gamma C}.$$

Then we get

$$\frac{e\gamma\mu}{2\Gamma T}\mathbb{E}\left[\sum_{t=0}^{T-1}\|z_t - z^*\|^2\right] \leq \frac{4\Gamma\Omega^2}{T} + \frac{1}{T}\sum_{t=1}^{\sqrt{\alpha}}\left((1 - \beta_t)C - \frac{\gamma\mu}{2\Gamma}\right)\mathbb{E}\left[\|z_{t-1} - z^*\|_{\hat{V}_{t-1}}^2\right]$$
$$+ \frac{6\gamma^2\sigma^2}{eb} \cdot \frac{1}{T}\sum_{t=1}^{T}\left(1 + (1 - \beta_t)C\right)$$
$$\leq \frac{4\Gamma\Omega^2}{T} + \frac{1}{T}\sum_{t=1}^{\sqrt{\alpha}}4C\Gamma\Omega^2 + \frac{6\gamma^2\sigma^2}{eb} \cdot \frac{1}{T}\sum_{t=1}^{T}\left(1 + (1 - \beta_t)C\right)$$
$$\leq \frac{4\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T} \cdot \frac{2\Gamma C}{\gamma\mu} + \frac{6\gamma^2\sigma^2}{eb} \cdot \frac{1}{T}\sum_{t=1}^{T}\left(1 + (1 - \beta_t)C\right)$$

29

$$\leq \frac{12C^2\Gamma^2\Omega^2}{\gamma\mu T} + \frac{12C\gamma^2\sigma^2}{eb}.$$

Then,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|z_t - z^*\|^2\right] \leq \frac{24C^2\Gamma^3\Omega^2}{e\gamma^2\mu^2 T} + \frac{24\Gamma C\gamma\sigma^2}{e^2\mu b}.$$

Finally, $\gamma = \min\left\{\frac{e}{4L}; \sqrt[3]{\frac{C\Gamma^2\Omega^2 eb}{\mu\sigma^2 T}}\right\}$ gives

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=0}^{T-1}z_t - z^*\right\|^2\right] \leq \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|z_t - z^*\|^2\right] = \mathcal{O}\left(\frac{C^2\Gamma^3 L^2\Omega^2}{e^3\mu^2 T} + \left(\frac{C^4\Gamma^5\sigma^4\Omega^2}{e^5\mu^4 b^2 T}\right)^{1/3}\right).$$

This completes the proof of the strongly convex–strongly concave case. $\square$

Convex–concave case. With $\beta_t = \frac{\beta - \beta^{t+1}}{1 - \beta^{t+1}}$ we get (11) in the following form

$$\mathbb{E}\left[\mathrm{gap}(x_T^{avg}, y_T^{avg})\right] \leq \frac{3\Gamma\Omega^2}{\gamma T} + \frac{4\gamma\sigma^2}{eb} + \frac{2(1-\beta)C\Gamma\Omega^2}{\gamma T}\sum_{t=1}^{T}\frac{1}{1 - \beta^{t+1}}$$

$$\leq \frac{3\Gamma\Omega^2}{\gamma T} + \frac{4\gamma\sigma^2}{eb}$$
$$+ \frac{2(1-\beta)C\Gamma\Omega^2}{\gamma T}\sum_{t=1}^{\sqrt{T}}\frac{1}{1 - \beta} + \frac{2(1-\beta)C\Gamma\Omega^2}{\gamma T}\sum_{t=1}^{T}\frac{1}{1 - \beta^{\sqrt{T}}}$$
$$\leq \frac{3\Gamma\Omega^2}{\gamma T} + \frac{4\gamma\sigma^2}{eb}$$
$$+ \frac{2C\Gamma\Omega^2}{\gamma\sqrt{T}} + \frac{2(1-\beta)C\Gamma\Omega^2}{\gamma T}\sum_{t=1}^{T}\frac{1}{1 - \beta^{\sqrt{T}}}.$$

Next, we substitute $\beta = 1 - \frac{1}{T}$. From (18) we get that $\beta^{\sqrt{T}} \leq 1 - \frac{1}{2\sqrt{T}}$. Then, we get

$$\mathbb{E}\left[\mathrm{gap}(x_T^{avg}, y_T^{avg})\right] \leq \frac{3\Gamma\Omega^2}{\gamma T} + \frac{4\gamma\sigma^2}{eb}$$
$$+ \frac{2C\Gamma\Omega^2}{\gamma\sqrt{T}} + \frac{4(1-\beta)C\Gamma\Omega^2\sqrt{T}}{\gamma}$$
$$\leq \frac{3\Gamma\Omega^2}{\gamma T} + \frac{4\gamma\sigma^2}{eb} + \frac{2C\Gamma\Omega^2}{\gamma\sqrt{T}} + \frac{4C\Gamma\Omega^2}{\gamma\sqrt{T}}.$$

It means that

$$\mathbb{E}\left[\mathrm{gap}(x_T^{avg}, y_T^{avg})\right] = \mathcal{O}\left(\frac{C\Gamma\Omega^2}{\gamma\sqrt{T}} + \frac{\gamma\sigma^2}{eb}\right).$$

With $\gamma = \min\left\{\frac{e}{2L}; \frac{\sqrt{C\Gamma eb}\Omega}{\sigma T^{1/4}}\right\}$

$$\mathbb{E}\left[\mathrm{gap}(x_T^{avg}, y_T^{avg})\right] = \mathcal{O}\left(\frac{C\Gamma L\Omega^2}{e\sqrt{T}} + \frac{\sqrt{C\Gamma eb}\sigma\Omega}{eT^{1/4}}\right).$$

This completes the proof of the convex–concave case. $\square$

Non-convex–non-concave case. With $\beta_t = \frac{\beta - \beta^{t+1}}{1 - \beta^{t+1}}$ we get (12) in the following form

$$\mathbb{E}\left[\|\nabla_x f(\bar{x}_T, \bar{y}_T)\|^2 + \|\nabla_y f(\bar{x}_T, \bar{y}_T)\|^2\right] \leq \frac{12\Gamma^2\Omega^2}{\gamma^2 T} + \frac{21\Gamma\sigma^2}{eb} + \frac{12(1-\beta)C\Gamma^2\Omega^2}{\gamma^2 T}\sum_{t=1}^{T}\frac{1}{1 - \beta^{t+1}}$$

$$\leq \frac{12\Gamma^2\Omega^2}{\gamma^2 T} + \frac{21\Gamma\sigma^2}{eb} + \frac{12(1-\beta)C\Gamma^2\Omega^2}{\gamma^2 T} \sum_{t=1}^{\sqrt{T}} \frac{1}{1-\beta}$$

$$+ \frac{12(1-\beta)C\Gamma^2\Omega^2}{\gamma^2 T} \sum_{t=1}^{T} \frac{1}{1-\beta^{\sqrt{T}}}$$

$$\leq \frac{12\Gamma^2\Omega^2}{\gamma^2 T} + \frac{21\Gamma\sigma^2}{eb} + \frac{12C\Gamma^2\Omega^2}{\gamma^2\sqrt{T}}$$

$$+ \frac{12(1-\beta)C\Gamma^2\Omega^2}{\gamma^2} \cdot \frac{1}{1-\beta^{\sqrt{T}}}.$$

Next, we substitute $\beta = 1 - \frac{1}{T}$. From (18) we get that $\beta^{\sqrt{T}} \leq 1 - \frac{1}{2\sqrt{T}}$. Then, we get

$$\mathbb{E}\left[\|\nabla_x f(\bar{x}_T, \bar{y}_T)\|^2 + \|\nabla_y f(\bar{x}_T, \bar{y}_T)\|^2\right] \leq \frac{12\Gamma^2\Omega^2}{\gamma^2 T} + \frac{21\Gamma\sigma^2}{eb} + \frac{12C\Gamma^2\Omega^2}{\gamma^2\sqrt{T}} + \frac{24C\Gamma^2\Omega^2}{\gamma^2\sqrt{T}}$$

$$\leq \frac{48C\Gamma^2\Omega^2}{\gamma^2\sqrt{T}} + \frac{21\Gamma\sigma^2}{eb}.$$

The batch size $b \sim \frac{\Gamma\sigma^2}{e\varepsilon^2}$ and $\gamma = \frac{e}{3L}$ complete the proof of the non-convex–non-concave case. $\square$

C.6 Proof of Theorem 2

We start the same way as in the proof of Theorem 1. Applying Lemma 4 with $z^+ = z_{t+1}$, $z = z_t$, $u = z$, $y = -\eta(\hat{V}_{t-1/2})^{-1}(w_t - z_t) + \gamma \hat{V}_{t-1/2}^{-1} g_{t+1/2}$ and $D = \hat{V}_{t-1/2}$, we get

$$\|z_{t+1} - z\|_{\hat{V}_{t-1/2}}^2 = \|z_t - z\|_{\hat{V}_{t-1/2}}^2 + 2\eta\langle w_t - z_t, z_{t+1/2} - z\rangle$$
$$- 2\gamma\langle g_{t+1/2}, z_{t+1} - z\rangle - \|z_{t+1} - z_t\|_{\hat{V}_{t-1/2}}^2,$$

and with $z^+ = z_{t+1/2}$, $z = z_t$, $u = z_{t+1}$, $y = \gamma V_{t-1/2}^{-1} g_{t-1/2}$, $D = \hat{V}_{t-1/2}$:

$$\|z_{t+1/2} - z_{t+1}\|_{\hat{V}_{t-1/2}}^2 = \|z_t - z_{t+1}\|_{\hat{V}_{t-1/2}}^2 - 2\gamma\langle g_{t-1/2}, z_{t+1/2} - z_{t+1}\rangle - \|z_{t+1/2} - z_t\|_{\hat{V}_{t-1/2}}^2.$$

Next, we sum up the two previous equalities

$$\|z_{t+1} - z\|_{\hat{V}_{t-1/2}}^2 + \|z_{t+1/2} - z_{t+1}\|_{\hat{V}_{t-1/2}}^2 = \|z_t - z\|_{\hat{V}_{t-1/2}}^2 - \|z_{t+1/2} - z_t\|_{\hat{V}_{t-1/2}}^2$$
$$- 2\gamma\langle g_{t+1/2}, z_{t+1} - z\rangle$$
$$- 2\gamma\langle g_{t-1/2}, z_{t+1/2} - z_{t+1}\rangle$$
$$+ 2\eta\langle w_t - z_t, z_{t+1/2} - z\rangle.$$

A small rearrangement gives

$$\|z_{t+1} - z\|_{\hat{V}_{t-1/2}}^2 + \|z_{t+1/2} - z_{t+1}\|_{\hat{V}_{t-1/2}}^2$$
$$= \|z_t - z\|_{\hat{V}_{t-1/2}}^2 - \|z_{t+1/2} - z_t\|_{\hat{V}_{t-1/2}}^2$$
$$- 2\gamma\langle g_{t+1/2}, z_{t+1/2} - z\rangle - 2\gamma\langle g_{t-1/2} - g_{t+1/2}, z_{t+1/2} - z_{t+1}\rangle$$
$$+ 2\eta\langle w_t - z, z_{t+1/2} - z\rangle + 2\eta\langle z - z_t, z_{t+1/2} - z\rangle$$
$$\leq \|z_t - z\|_{\hat{V}_{t-1/2}}^2 - \|z_{t+1/2} - z_t\|_{\hat{V}_{t-1/2}}^2 - 2\gamma\langle g^{t+1/2}, z_{t+1/2} - z\rangle$$
$$+ \gamma^2\|g_{t+1/2} - g_{t-1/2}\|_{\hat{V}_{t-1/2}^{-1}}^2 + \|z_{t+1/2} - z_{t+1}\|_{\hat{V}_{t-1/2}}^2$$
$$+ \eta\|w_t - z\|^2 + \eta\|z_{t+1/2} - z\|^2 - \eta\|z_{t+1/2} - w_t\|^2$$
$$- \eta\|z - z_t\|^2 - \eta\|z_{t+1/2} - z\|^2 + \eta\|z_{t+1/2} - z_t\|^2$$
$$\leq \|z_t - z\|_{\hat{V}_{t-1/2}}^2 - \|z_{t+1/2} - z_t\|_{\hat{V}_{t-1/2}}^2 - 2\gamma\langle g^{t+1/2}, z_{t+1/2} - z\rangle$$
$$+ \gamma^2\|g_{t+1/2} - g_{t-1/2}\|_{\hat{V}_{t-1/2}^{-1}}^2 + \|z_{t+1/2} - z_{t+1}\|_{\hat{V}_{t-1/2}}^2$$
$$+ \eta\|w_t - z\|^2 - \eta\|z - z_t\|^2 + \eta\|z_{t+1/2} - z_t\|^2 - \eta\|z_{t+1/2} - w_t\|^2.$$

And then,

$$\|z_{t+1} - z\|_{\hat{V}_{t-1/2}}^2 \leq \|z_t - z\|_{\hat{V}_{t-1/2}}^2 - \|z_{t+1/2} - z_t\|_{\hat{V}_{t-1/2}}^2 - 2\gamma\langle g^{t+1/2}, z_{t+1/2} - z\rangle$$
$$+ \gamma^2\|g_{t+1/2} - g_{t-1/2}\|_{\hat{V}_{t-1/2}^{-1}}^2$$
$$+ \eta\|w_t - z\|^2 - \eta\|z - z_t\|^2 + \eta\|z_{t+1/2} - z_t\|^2 - \eta\|z_{t+1/2} - w_t\|^2$$
$$\leq \|z_t - z\|_{\hat{V}_{t-1/2}}^2 - \|z_{t+1/2} - z_t\|_{\hat{V}_{t-1/2}}^2 - 2\gamma\langle g^{t+1/2}, z_{t+1/2} - z\rangle$$
$$+ \frac{\gamma^2}{e}\|g_{t+1/2} - g_{t-1/2}\|^2$$
$$+ \eta\|w_t - z\|^2 - \eta\|z - z_t\|^2 + \eta\|z_{t+1/2} - z_t\|^2 - \eta\|z_{t+1/2} - w_t\|^2$$
$$\leq \|z_t - z\|_{\hat{V}_{t-1/2}}^2 - \|z_{t+1/2} - z_t\|_{\hat{V}_{t-1/2}}^2 - 2\gamma\langle g^{t+1/2}, z_{t+1/2} - z\rangle$$
$$+ \frac{3\gamma^2}{e}\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2 + \frac{3\gamma^2}{e}\|g_{t+1/2} - F(z_{t+1/2})\|^2$$
$$+ \frac{3\gamma^2}{e}\|F(z_{t-1/2}) - g_{t-1/2}\|^2$$

$$+ \eta\|w_t - z\|^2 - \eta\|z - z_t\|^2 + \eta\|z_{t+1/2} - z_t\|^2 - \eta\|z_{t+1/2} - w_t\|^2. \quad (25)$$

Using smoothness of f and the update of Algorithm 2

$$\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2$$
$$\leq L^2\|z_{t+1/2} - z_{t-1/2}\|^2$$
$$\leq 2L^2\|z_{t+1/2} - z_t\|^2 + 2L^2\|z_t - z_{t-1/2}\|^2$$
$$\leq 2L^2\|z_{t+1/2} - z_t\|^2 + \frac{2L^2}{e}\|z_t - z_{t-1/2}\|^2_{\hat{V}_{t-3/2}}$$
$$\leq 2L^2\|z_{t+1/2} - z_t\|^2$$
$$\quad + \frac{2L^2}{e}\|z_{t-1} + \eta(\hat{V}_{t-3/2})^{-1}(w_{t-1} - z_{t-1}) - \gamma(\hat{V}_{t-3/2})^{-1}g_{t-1/2} - z_{t-1} + \gamma(\hat{V}^x_{t-3/2})^{-1}g_{t-3/2}\|^2_{\hat{V}_{t-3/2}}$$
$$\leq 2L^2\|z_{t+1/2} - z_t\|^2$$
$$\quad + \frac{2L^2}{e}\|\eta(w_{t-1} - z_{t-1}) - \gamma g_{t-1/2} + \gamma g_{t-3/2}\|^2_{\hat{V}^{-1}_{t-3/2}}.$$

Hence,

$$\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2 \leq 2L^2\|z_{t+1/2} - z_t\|^2$$
$$+ \frac{2L^2}{e^2}\|\eta(w_{t-1} - z_{t-1}) - \gamma g_{t-1/2} + \gamma g_{t-3/2}\|^2$$
$$\leq 2L^2\|z_{t+1/2} - z_t\|^2 + \frac{4L^2\eta^2}{e^2}\|w_{t-1} - z_{t-1}\|^2$$
$$+ \frac{4L^2\gamma^2}{e^2}\|g_{t-1/2} - g_{t-3/2}\|^2$$
$$\leq 2L^2\|z_{t+1/2} - z_t\|^2 + \frac{4L^2\eta^2}{e^2}\|w_{t-1} - z_{t-1}\|^2$$
$$+ \frac{12L^2\gamma^2}{e^2}\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2$$
$$+ \frac{12L^2\gamma^2}{e^2}\|g_{t-1/2} - F(z_{t-1/2})\|^2$$
$$+ \frac{12L^2\gamma^2}{e^2}\|F(z_{t-3/2}) - g_{t-3/2}\|^2. \quad (26)$$

Combining (25) and (26), we have

$$\|z_{t+1} - z\|^2_{\hat{V}_{t-1/2}} + \frac{3\gamma^2 M}{e}\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2$$
$$\leq \|z_t - z\|^2_{\hat{V}_{t-1/2}} - \|z_{t+1/2} - z_t\|^2_{\hat{V}_{t-1/2}} - 2\gamma\langle g^{t+1/2}, z_{t+1/2} - z\rangle$$
$$+ \frac{3\gamma^2}{e}\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2 + \frac{3\gamma^2}{e}\|g_{t+1/2} - F(z_{t+1/2})\|^2$$
$$+ \frac{3\gamma^2}{e}\|F(z_{t-1/2}) - g_{t-1/2}\|^2$$
$$+ \eta\|w_t - z\|^2 - \eta\|z - z_t\|^2 + \eta\|z_{t+1/2} - z_t\|^2 - \eta\|z_{t+1/2} - w_t\|^2$$
$$+ \frac{6\gamma^2 L^2 M}{e}\|z_{t+1/2} - z_t\|^2 + \frac{12\gamma^2 L^2\eta^2 M}{e^3}\|w_{t-1} - z_{t-1}\|^2$$
$$+ \frac{36\gamma^4 L^2 M}{e^3}\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2$$
$$+ \frac{36\gamma^4 L^2 M}{e^3}\|g_{t-1/2} - F(z_{t-1/2})\|^2 + \frac{36\gamma^4 L^2 M}{e^3}\|F(z_{t-3/2}) - g_{t-3/2}\|^2.$$

Here M is some positive constant, which we will define later. Next, we get

$$\|z_{t+1} - z\|^2_{\hat{V}_{t-1/2}} + \frac{3\gamma^2 M}{e}\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2$$

33

$$\leq \|z_t - z\|^2_{\hat{V}_{t-1/2}} - \|z_{t+1/2} - z_t\|^2_{\hat{V}_{t-1/2}} - 2\gamma\langle g^{t+1/2}, z_{t+1/2} - z\rangle$$

$$+ \frac{3\gamma^2}{e}\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2 + \frac{3\gamma^2}{e}\|g_{t+1/2} - F(z_{t+1/2})\|^2$$

$$+ \frac{3\gamma^2}{e}\|F(z_{t-1/2}) - g_{t-1/2}\|^2$$

$$+ \eta\|w_t - z\|^2 - \eta\|z - z_t\|^2 + \eta\|z_{t+1/2} - z_t\|^2 - \frac{\eta}{2}\|z_t - w_t\|^2$$

$$+ \eta\|z_{t+1/2} - z_t\|^2$$

$$+ \frac{6\gamma^2 L^2 M}{e}\|z_{t+1/2} - z_t\|^2 + \frac{12\gamma^2 L^2 \eta^2 M}{e^3}\|w_{t-1} - z_{t-1}\|^2$$

$$+ \frac{36\gamma^4 L^2 M}{e^3}\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2$$

$$+ \frac{36\gamma^4 L^2 M}{e^3}\|g_{t-1/2} - F(z_{t-1/2})\|^2 + \frac{36\gamma^4 L^2 M}{e^3}\|F(z_{t-3/2}) - g_{t-3/2}\|^2.$$

Small rearrangement gives

$$\|z_{t+1} - z\|^2_{\hat{V}_{t-1/2}} + \frac{3\gamma^2(M-1)}{e}\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2 + \frac{\eta}{2}\|z_t - w_t\|^2$$

$$\leq \|z_t - z\|^2_{\hat{V}_{t-1/2}} - \eta\|z_t - z\|^2 + \eta\|w_t - z\|^2 - 2\gamma\langle g^{t+1/2}, z_{t+1/2} - z\rangle$$

$$+ \frac{12\gamma^2 L^2 M}{(M-1)e^2} \cdot \frac{3\gamma^2(M-1)}{e}\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2 + \frac{12\gamma^2 L^2 \eta^2 M}{e^3}\|z_{t-1} - w_{t-1}\|^2$$

$$+ \frac{6\gamma^2 L^2 M}{e}\|z_{t+1/2} - z_t\|^2 + 2\eta\|z_{t+1/2} - z_t\|^2 - \|z_{t+1/2} - z_t\|^2_{\hat{V}_{t-1/2}}$$

$$+ \frac{3\gamma^2}{e}\|g_{t+1/2} - F(z_{t+1/2})\|^2 + \frac{3\gamma^2}{e}\|F(z_{t-1/2}) - g_{t-1/2}\|^2$$

$$+ \frac{36\gamma^4 L^2 M}{e^3}\|g_{t-1/2} - F(z_{t-1/2})\|^2 + \frac{36\gamma^4 L^2 M}{e^3}\|F(z_{t-3/2}) - g_{t-3/2}\|^2.$$

With $M = 2$, $\gamma \leq \frac{e}{10L}$ and $\eta \leq ep \leq \frac{e}{4}$, we have $\frac{12\gamma^2 L^2 M}{(M-1)e^2} \leq \frac{1}{2}$ and $\frac{12\gamma^2 L^2 \eta M}{e^3} \leq \frac{1}{4}$. Then,

$$\|z_{t+1} - z\|^2_{\hat{V}_{t-1/2}} + \frac{3\gamma^2}{e}\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2 + \frac{\eta}{2}\|z_t - w_t\|^2$$

$$\leq \|z_t - z\|^2_{\hat{V}_{t-1/2}} - \eta\|z_t - z\|^2 + \eta\|w_t - z\|^2 - 2\gamma\langle g^{t+1/2}, z_{t+1/2} - z\rangle$$

$$+ \frac{1}{2} \cdot \frac{3\gamma^2}{e}\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2 + \frac{1}{2} \cdot \frac{\eta}{2}\|z_{t-1} - w_{t-1}\|^2$$

$$+ \frac{e}{4}\|z_{t+1/2} - z_t\|^2 + \frac{e}{2}\|z_{t+1/2} - z_t\|^2 - \|z_{t+1/2} - z_t\|^2_{\hat{V}_{t-1/2}}$$

$$+ \frac{3\gamma^2}{e}\|g_{t+1/2} - F(z_{t+1/2})\|^2 + \frac{3\gamma^2}{e}\|F(z_{t-1/2}) - g_{t-1/2}\|^2$$

$$+ \frac{3\gamma^2 L^2}{e}\|g_{t-1/2} - F(z_{t-1/2})\|^2 + \frac{3\gamma^2 L^2}{e}\|F(z_{t-3/2}) - g_{t-3/2}\|^2.$$

Using $I \preccurlyeq \frac{1}{e}\hat{V}_{t-1/2}$, we obtain

$$\|z_{t+1} - z\|^2_{\hat{V}_{t-1/2}} + \frac{3\gamma^2}{e}\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2 + \frac{\eta}{2}\|z_t - w_t\|^2$$

$$\leq \|z_t - z\|^2_{\hat{V}_{t-1/2}} - \eta\|z_t - z\|^2 + \eta\|w_t - z\|^2 - 2\gamma\langle g^{t+1/2}, z_{t+1/2} - z\rangle$$

$$+ \frac{1}{2} \cdot \frac{3\gamma^2}{e}\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2 + \frac{1}{2} \cdot \frac{\eta}{2}\|z_{t-1} - w_{t-1}\|^2$$

$$+ \frac{1}{4}\|z_{t+1/2} - z_t\|^2_{\hat{V}_{t-1/2}} + \frac{1}{2}\|z_{t+1/2} - z_t\|^2_{\hat{V}_{t-1/2}} - \|z_{t+1/2} - z_t\|^2_{\hat{V}_{t-1/2}}$$

$$+ \frac{3\gamma^2}{e}\|g_{t+1/2} - F(z_{t+1/2})\|^2 + \frac{3\gamma^2}{e}\|F(z_{t-1/2}) - g_{t-1/2}\|^2$$

$$+ \frac{3\gamma^2}{e}\|g_{t-1/2} - F(z_{t-1/2})\|^2 + \frac{3\gamma^2}{e}\|F(z_{t-3/2}) - g_{t-3/2}\|^2$$

$$\leq \|z_t - z\|^2_{\hat{V}_{t-1/2}} - \eta\|z_t - z\|^2 + \eta\|w_t - z\|^2 - 2\gamma\langle g^{t+1/2}, z_{t+1/2} - z\rangle$$

$$+ \frac{1}{2}\cdot\frac{3\gamma^2}{e}\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2 + \frac{1}{2}\cdot\frac{\eta}{2}\|z_{t-1} - w_{t-1}\|^2$$

$$- \frac{1}{4}\|z_{t+1/2} - z_t\|^2_{\hat{V}_{t-1/2}} + \frac{3\gamma^2}{e}\|g_{t+1/2} - F(z_{t+1/2})\|^2$$

$$+ \frac{6\gamma^2}{e}\|F(z_{t-1/2}) - g_{t-1/2}\|^2 + \frac{3\gamma^2}{e}\|F(z_{t-3/2}) - g_{t-3/2}\|^2. \tag{27}$$

Strongly convex–strongly concave case. We substitute $z = z^*$ and take the total expectation of both sides of the equation

$$\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_{t-1/2}}\right] + \frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2\right] + \frac{\eta}{2}\mathbb{E}\left[\|z_t - w_t\|^2\right]$$

$$\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right] - \eta\mathbb{E}\left[\|z_t - z^*\|^2\right] + \eta\mathbb{E}\left[\|w_t - z^*\|^2\right]$$

$$- 2\gamma\mathbb{E}\left[\langle\mathbb{E}_{t+1/2}\left[g^{t+1/2}\right], z_{t+1/2} - z^*\rangle\right]$$

$$+ \frac{1}{2}\cdot\frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2}\cdot\frac{\eta}{2}\mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]$$

$$- \frac{1}{4}\mathbb{E}\left[\|z_{t+1/2} - z_t\|^2_{\hat{V}_{t-1/2}}\right] + \frac{3\gamma^2}{e}\mathbb{E}\left[\|g_{t+1/2} - F(z_{t+1/2})\|^2\right]$$

$$+ \frac{6\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - g_{t-1/2}\|^2\right] + \frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-3/2}) - g_{t-3/2}\|^2\right].$$

Next, using the property of the solution z^*: $\langle F(z^*), z_{t+1/2} - z^*\rangle = 0$, and get

$$\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_{t-1/2}}\right] + \frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2\right] + \frac{\eta}{2}\mathbb{E}\left[\|z_t - w_t\|^2\right]$$

$$\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right] - \eta\mathbb{E}\left[\|z_t - z^*\|^2\right] + \eta\mathbb{E}\left[\|w_t - z^*\|^2\right]$$

$$- 2\gamma\mathbb{E}\left[\langle F(z_{t+1/2}) - z^*, z_{t+1/2} - z^*\rangle\right]$$

$$+ \frac{1}{2}\cdot\frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2}\cdot\frac{\eta}{2}\mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]$$

$$- \frac{1}{4}\mathbb{E}\left[\|z_{t+1/2} - z_t\|^2_{\hat{V}_{t-1/2}}\right] + \frac{3\gamma^2}{e}\mathbb{E}\left[\|g_{t+1/2} - F(z_{t+1/2})\|^2\right]$$

$$+ \frac{6\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - g_{t-1/2}\|^2\right] + \frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-3/2}) - g_{t-3/2}\|^2\right].$$

Assumptions 1 and 3 on strongly convexity - strongly concavity and on stochastisity give

$$\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_{t-1/2}}\right] + \frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2\right] + \frac{\eta}{2}\mathbb{E}\left[\|z_t - w_t\|^2\right]$$

$$\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right] - \eta\mathbb{E}\left[\|z_t - z^*\|^2\right] + \eta\mathbb{E}\left[\|w_t - z^*\|^2\right]$$

$$- 2\gamma\mu\mathbb{E}\left[\|z_{t+1/2} - z^*\|^2\right]$$

$$+ \frac{1}{2}\cdot\frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2}\cdot\frac{\eta}{2}\mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]$$

$$- \frac{1}{4}\mathbb{E}\left[\|z_{t+1/2} - z_t\|^2_{\hat{V}_{t-1/2}}\right] + \frac{12\gamma^2\sigma^2}{eb}$$

$$\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right] - \eta\mathbb{E}\left[\|z_t - z^*\|^2\right] + \eta\mathbb{E}\left[\|w_t - z^*\|^2\right]$$

$$- \gamma\mu\mathbb{E}\left[\|z_t - z^*\|^2\right] + 2\gamma\mu\mathbb{E}\left[\|z_{t+1/2} - z_t\|^2\right]$$

$$+ \frac{1}{2} \cdot \frac{3\gamma^2}{e} \mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2} \cdot \frac{\eta}{2} \mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]$$

$$- \frac{1}{4} \mathbb{E}\left[\|z_{t+1/2} - z_t\|^2_{\hat{V}_{t-1/2}}\right] + \frac{12\gamma^2\sigma^2}{eb}.$$

Using $I \preccurlyeq \frac{1}{e}\hat{V}_{t-1/2}$, we obtain

$$\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_{t-1/2}}\right] + \frac{3\gamma^2}{e} \mathbb{E}\left[\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2\right] + \frac{\eta}{2} \mathbb{E}\left[\|z_t - w_t\|^2\right]$$

$$\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right] - (\eta + \gamma\mu)\mathbb{E}\left[\|z_t - z^*\|^2\right] + \eta\mathbb{E}\left[\|w_t - z^*\|^2\right]$$

$$+ \frac{1}{2} \cdot \frac{3\gamma^2}{e} \mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2} \cdot \frac{\eta}{2} \mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]$$

$$- \left(\frac{1}{4} - \frac{2\gamma\mu}{e}\right) \mathbb{E}\left[\|z_{t+1/2} - z_t\|^2_{\hat{V}_{t-1/2}}\right] + \frac{12\gamma^2\sigma^2}{eb}.$$

With $\gamma \leq \frac{e}{8\mu}$,

$$\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_{t-1/2}}\right] + \frac{3\gamma^2}{e} \mathbb{E}\left[\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2\right] + \frac{\eta}{2} \mathbb{E}\left[\|z_t - w_t\|^2\right]$$

$$\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right] - (\eta + \gamma\mu)\mathbb{E}\left[\|z_t - z^*\|^2\right] + \eta\mathbb{E}\left[\|w_t - z^*\|^2\right]$$

$$+ \frac{1}{2} \cdot \frac{3\gamma^2}{e} \mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2} \cdot \frac{\eta}{2} \mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]$$

$$+ \frac{12\gamma^2\sigma^2}{eb}.$$

The update of w_{t+1} gives

$$\frac{\eta + \gamma\mu/2}{p} \mathbb{E}\left[\|w_{t+1} - z^*\|^2\right] = \frac{\eta + \gamma\mu/2}{p} \mathbb{E}\left[\mathbb{E}_{w_{t+1}}\left[\|w_{t+1} - z^*\|^2\right]\right]$$

$$= \left(\eta + \frac{\gamma\mu}{2}\right)\mathbb{E}\left[\|z_t - z^*\|^2\right] + \frac{(1-p)(\eta + \gamma\mu/2)}{p}\mathbb{E}\left[\|w_t - z^*\|^2\right].$$

Connecting with the previous equation, we get

$$\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_{t-1/2}}\right] + \frac{\eta + \gamma\mu/2}{p} \mathbb{E}\left[\|w_{t+1} - z^*\|^2\right]$$

$$+ \frac{3\gamma^2}{e} \mathbb{E}\left[\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2\right] + \frac{\eta}{2} \mathbb{E}\left[\|z_t - w_t\|^2\right]$$

$$\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right] - \frac{\gamma\mu}{2}\mathbb{E}\left[\|z_t - z^*\|^2\right]$$

$$+ (1-p) \cdot \frac{\eta + \gamma\mu/2}{p}\mathbb{E}\left[\|w_t - z^*\|^2\right] + \eta\mathbb{E}\left[\|w_t - z^*\|^2\right]$$

$$+ \frac{1}{2} \cdot \frac{3\gamma^2}{e} \mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2} \cdot \frac{\eta}{2} \mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]$$

$$+ \frac{12\gamma^2\sigma^2}{eb}$$

$$\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right] - \frac{\gamma\mu}{2}\mathbb{E}\left[\|z_t - z^*\|^2\right]$$

$$+ \left(1 - \frac{p\gamma\mu}{2\eta + \gamma\mu}\right) \cdot \frac{\eta + \gamma\mu/2}{p}\mathbb{E}\left[\|w_t - z^*\|^2\right]$$

$$+ \frac{1}{2} \cdot \frac{3\gamma^2}{e} \mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2} \cdot \frac{\eta}{2} \mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]$$

$$+ \frac{12\gamma^2\sigma^2}{eb}.$$

Using $\frac{1}{\Gamma}\hat{V}_{t-1/2} \preccurlyeq I$, we get

$$
\mathbb{E}\left[\|z_{t+1} - z^*\|_{\hat{V}_{t-1/2}}^2\right] + \frac{\eta + \gamma\mu/2}{p}\mathbb{E}\left[\|w_{t+1} - z^*\|^2\right]
$$

$$
+ \frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2\right] + \frac{\eta}{2}\mathbb{E}\left[\|z_t - w_t\|^2\right]
$$

$$
\leq \left(1 - \frac{\gamma\mu}{2\Gamma}\right)\mathbb{E}\left[\|z_t - z^*\|_{\hat{V}_{t-1/2}}^2\right] + \left(1 - \frac{p\gamma\mu}{2\eta + \gamma\mu}\right)\frac{\eta + \gamma\mu/2}{p}\mathbb{E}\left[\|w_t - z^*\|^2\right]
$$

$$
+ \frac{1}{2}\cdot\frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2}\cdot\frac{\eta}{2}\mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]
$$

$$
+ \frac{12\gamma^2\sigma^2}{eb}.
$$

Lemma 1 with $C = \frac{\Gamma^2}{2e^2}$ for (7), $C = \frac{2\Gamma}{e}$ for (8) and $B_{t+1} = (1 + (1 - \beta_{t+1})C)$ gives

$$
\frac{1}{B_{t+1}}\mathbb{E}\left[\|z_{t+1} - z^*\|_{\hat{V}_{t+1/2}}^2\right] + \frac{\eta + \gamma\mu/2}{p}\mathbb{E}\left[\|w_{t+1} - z^*\|^2\right]
$$

$$
+ \frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2\right] + \frac{\eta}{2}\mathbb{E}\left[\|z_t - w_t\|^2\right]
$$

$$
\leq \mathbb{E}\left[\|z_{t+1} - z^*\|_{\hat{V}_{t-1/2}}^2\right] + \frac{\eta + \gamma\mu/2}{p}\mathbb{E}\left[\|w_{t+1} - z^*\|^2\right]
$$

$$
+ \frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2\right] + \frac{\eta}{2}\mathbb{E}\left[\|z_t - w_t\|^2\right]
$$

$$
\leq \left(1 - \frac{\gamma\mu}{2\Gamma}\right)\mathbb{E}\left[\|z_t - z^*\|_{\hat{V}_{t-1/2}}^2\right] + \left(1 - \frac{p\gamma\mu}{2\eta + \gamma\mu}\right)\frac{\eta + \gamma\mu/2}{p}\mathbb{E}\left[\|w_t - z^*\|^2\right]
$$

$$
+ \frac{1}{2}\cdot\frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2}\cdot\frac{\eta}{2}\mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]
$$

$$
+ \frac{12\gamma^2\sigma^2}{eb}.
$$

And then,

$$
\mathbb{E}\left[\|z_{t+1} - z^*\|_{\hat{V}_{t+1/2}}^2\right] + \frac{(\eta + \gamma\mu/2)B_{t+1}}{p}\mathbb{E}\left[\|w_{t+1} - z^*\|^2\right]
$$

$$
+ \frac{3\gamma^2 B_{t+1}}{e}\mathbb{E}\left[\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2\right] + \frac{\eta B_{t+1}}{2}\mathbb{E}\left[\|z_t - w_t\|^2\right]
$$

$$
\leq \left(1 - \frac{\gamma\mu}{2\Gamma}\right)B_{t+1}\mathbb{E}\left[\|z_t - z^*\|_{\hat{V}_{t-1/2}}^2\right] + \left(1 - \frac{p\gamma\mu}{2\eta + \gamma\mu}\right)\frac{(\eta + \gamma\mu/2)B_{t+1}}{p}\mathbb{E}\left[\|w_t - z^*\|^2\right]
$$

$$
+ \frac{1}{2}\cdot\frac{3\gamma^2 B_{t+1}}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2}\cdot\frac{\eta B_{t+1}}{2}\mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]
$$

$$
+ \frac{12\gamma^2\sigma^2 B_{t+1}}{eb}.
$$

One can note that in Section 3 we use $\beta_t \equiv \beta$ or $\beta_t = \frac{\beta - \beta^t}{1 - \beta^t}$. For such β_t, it holds that $B_{t+1} \leq B_t$. Then,

$$
\mathbb{E}\left[\|z_{t+1} - z^*\|_{\hat{V}_{t+1/2}}^2\right] + \frac{(\eta + \gamma\mu/2)B_{t+1}}{p}\mathbb{E}\left[\|w_{t+1} - z^*\|^2\right]
$$

$$
+ \frac{3\gamma^2 B_{t+1}}{e}\mathbb{E}\left[\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2\right] + \frac{\eta B_{t+1}}{2}\mathbb{E}\left[\|z_t - w_t\|^2\right]
$$

$$
\leq \left(1 - \frac{\gamma\mu}{2\Gamma}\right)B_t\mathbb{E}\left[\|z_t - z^*\|_{\hat{V}_{t-1/2}}^2\right] + \left(1 - \frac{p\gamma\mu}{2\eta + \gamma\mu}\right)\frac{(\eta + \gamma\mu/2)B_t}{p}\mathbb{E}\left[\|w_t - z^*\|^2\right]
$$

$$
+ \frac{1}{2}\cdot\frac{3\gamma^2 B_t}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2}\cdot\frac{\eta B_t}{2}\mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]
$$

$$+ \frac{12\gamma^2\sigma^2 B_{t+1}}{eb}.$$

The notation

$$\Psi_{t+1} = \|z_{t+1} - z^*\|^2_{\hat{V}_{t+1/2}} + \frac{(\eta + \gamma\mu/2)B_{t+1}}{p}\|w_{t+1} - z^*\|^2$$

$$+ \frac{3\gamma^2 B_{t+1}}{e}\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2 + \frac{\eta B_{t+1}}{2}\|z_t - w_t\|^2$$

gives

$$\mathbb{E}\left[\Psi_{t+1}\right] \le \max\left[\left(1 - \frac{\gamma\mu}{2\Gamma}\right)(1 + (1 - \beta_{t+1})C)\,;\,\left(1 - \frac{1}{\frac{2\eta}{\gamma\mu p} + \frac{1}{p}}\right)\,;\,\frac{1}{2}\right]\cdot\mathbb{E}\left[\Psi_t\right] + \frac{12\gamma^2\sigma^2 B_{t+1}}{eb}.$$

This completes the proof of the strongly convex–strongly concave case. $\square$

Convex–concave case. We start from (27) with small rearrangement

$$2\gamma\langle F(z_{t+1/2}), z_{t+1/2} - z\rangle \le \|z_t - z\|^2_{\hat{V}_{t-1/2}} - \|z_{t+1} - z\|^2_{\hat{V}_{t-1/2}}$$

$$+ \frac{\eta}{p}\|w_t - z\|^2 - \frac{\eta}{p}\|w_{t+1} - z\|^2$$

$$+ \frac{\eta}{p}\|w_{t+1} - z\|^2 - \eta\|z_t - z\|^2 - (1-p)\frac{\eta}{p}\|w_t - z\|^2$$

$$+ \frac{1}{2}\cdot\frac{3\gamma^2}{e}\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2 - \frac{3\gamma^2}{e}\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2$$

$$+ \frac{1}{2}\cdot\frac{\eta}{2}\|z_{t-1} - w_{t-1}\|^2 - \frac{\eta}{2}\|z_t - w_t\|^2$$

$$+ 2\gamma\langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} - z\rangle$$

$$+ \frac{3\gamma^2}{e}\|g_{t+1/2} - F(z_{t+1/2})\|^2$$

$$+ \frac{6\gamma^2}{e}\|F(z_{t-1/2}) - g_{t-1/2}\|^2 + \frac{3\gamma^2}{e}\|F(z_{t-3/2}) - g_{t-3/2}\|^2.$$

With $I \preccurlyeq \frac{1}{e}\hat{V}_t$ and Lemma 1 with $C = \frac{\Gamma^2}{2e^2}$ for (7) and $C = \frac{2\Gamma}{e}$ for (8), we get

$$2\gamma\langle F(z_{t+1/2}), z_{t+1/2} - z\rangle \le \|z_t - z\|^2_{\hat{V}_{t-1/2}} - (1 + (1 - \beta_{t+1})C)^{-1}\|z_{t+1} - z\|^2_{\hat{V}_{t+1/2}}$$

$$+ \frac{\eta}{p}\|w_t - z\|^2 - \frac{\eta}{p}\|w_{t+1} - z\|^2$$

$$+ \frac{\eta}{p}\|w_{t+1} - z\|^2 - \eta\|z_t - z\|^2 - (1-p)\frac{\eta}{p}\|w_t - z\|^2$$

$$+ \frac{1}{2}\cdot\frac{3\gamma^2}{e}\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2 - \frac{3\gamma^2}{e}\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2$$

$$+ \frac{1}{2}\cdot\frac{\eta}{2}\|z_{t-1} - w_{t-1}\|^2 - \frac{\eta}{2}\|z_t - w_t\|^2$$

$$+ 2\gamma\langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} - z\rangle$$

$$+ \frac{3\gamma^2}{e}\|g_{t+1/2} - F(z_{t+1/2})\|^2$$

$$+ \frac{6\gamma^2}{e}\|F(z_{t-1/2}) - g_{t-1/2}\|^2 + \frac{3\gamma^2}{e}\|F(z_{t-3/2}) - g_{t-3/2}\|^2.$$

Next, we sum over all t from 0 to $T - 1$

$$2\gamma\cdot\frac{1}{T}\sum_{t=0}^{T-1}\langle F(z_{t+1/2}), z_{t+1/2} - z\rangle$$

$$\le \frac{\|z_0 - z\|^2_{\hat{V}_{-1/2}}}{T} + \frac{1}{T}\sum_{t=1}^{T-1}(1 - \beta_{t+1})C\|z_t - z\|^2_{\hat{V}_{t-1/2}}$$

$$+ \frac{\eta \|w_0 - z\|^2}{pT} + \frac{3\gamma^2}{2eT} \|F(z_{-1/2}) - F(z_{-3/2})\|^2 + \frac{\eta}{4T} \|z_{-1} - w_{-1}\|^2$$

$$- \frac{\eta}{4T} \sum_{t=1}^{T-1} \|z_t - w_t\|^2$$

$$+ \frac{\eta}{p} \cdot \frac{1}{T} \sum_{t=0}^{T-1} \left[\|w_{t+1} - z\|^2 - p\|z_t - z\|^2 - (1-p)\|w_t - z\|^2 \right]$$

$$+ 2\gamma \cdot \frac{1}{T} \sum_{t=0}^{T-1} \langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} - z \rangle$$

$$+ \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \|g_{t+1/2} - F(z_{t+1/2})\|^2$$

$$+ \frac{6\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \|F(z_{t-1/2}) - g_{t-1/2}\|^2 + \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \|F(z_{t-3/2}) - g_{t-3/2}\|^2. \quad (28)$$

The fact $\hat{V}_t \preccurlyeq \Gamma I$ gives

$$2\gamma \cdot \frac{1}{T} \sum_{t=0}^{T-1} \langle F(z_{t+1/2}), z_{t+1/2} - z \rangle$$

$$\leq \frac{\Gamma \|z_0 - z\|^2}{T} + \frac{\Gamma}{T} \sum_{t=1}^{T-1} (1 - \beta_{t+1}) C \|z_t - z\|^2$$

$$+ \frac{\eta \|w_0 - z\|^2}{pT} + \frac{3\gamma^2}{2eT} \|F(z_{-1/2}) - F(z_{-3/2})\|^2 + \frac{\eta}{4T} \|z_{-1} - w_{-1}\|^2$$

$$+ \frac{\eta}{p} \cdot \frac{1}{T} \sum_{t=0}^{T-1} \left[\|w_{t+1} - z\|^2 - p\|z_t - z\|^2 - (1-p)\|w_t - z\|^2 \right]$$

$$+ 2\gamma \cdot \frac{1}{T} \sum_{t=0}^{T-1} \langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} - z \rangle$$

$$+ \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \|g_{t+1/2} - F(z_{t+1/2})\|^2$$

$$+ \frac{6\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \|F(z_{t-1/2}) - g_{t-1/2}\|^2 + \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \|F(z_{t-3/2}) - g_{t-3/2}\|^2.$$

Taking maximum on $\mathcal{Z}$, using (23) and taking the full expectation, we obtain

$$2\gamma \mathbb{E}\left[\text{gap}(z_T^{avg})\right] \leq \max_{z \in \mathcal{Z}} \frac{\Gamma \|z_0 - z\|^2}{T} + \mathbb{E}\left[\max_{z \in \mathcal{Z}} \frac{\Gamma}{T} \sum_{t=1}^{T-1} (1 - \beta_{t+1}) C \|z_t - z\|^2 \right]$$

$$+ \max_{z \in \mathcal{Z}} \frac{\eta \|w_0 - z\|^2}{pT} + \frac{3\gamma^2}{2eT} \|F(z_{-1/2}) - F(z_{-3/2})\|^2 + \frac{\eta}{4T} \|z_{-1} - w_{-1}\|^2$$

$$+ \frac{\eta}{p} \cdot \mathbb{E}\left[\max_{z \in \mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|w_{t+1} - z\|^2 - p\|z_t - z\|^2 - (1-p)\|w_t - z\|^2 \right] \right]$$

$$+ 2\gamma \cdot \mathbb{E}\left[\max_{z \in \mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} - z \rangle \right]$$

$$+ \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|g_{t+1/2} - F(z_{t+1/2})\|^2 \right]$$

$$+ \frac{6\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|F(z_{t-1/2}) - g_{t-1/2}\|^2\right] + \frac{3\gamma^2}{e} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|F(z_{t-3/2}) - g_{t-3/2}\|^2\right].$$

Next, we use Assumption 3 and the fact that iterations and the set $\mathcal{Z}$ are bounded: $\|z_t\|, \|z\| \leq \Omega$

$$2\gamma\mathbb{E}\left[\text{gap}(z_T^{avg})\right] \leq \frac{4\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T} \sum_{t=1}^{T-1}(1 - \beta_{t+1}) + \frac{4\eta\Omega^2}{pT} + \frac{6\gamma^2 L^2 \Omega^2}{eT} + \frac{\eta\Omega^2}{T} + \frac{12\gamma^2\sigma^2}{eb}$$

$$+ \frac{\eta}{p} \cdot \mathbb{E}\left[\max_{z\in\mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|w_{t+1} - z\|^2 - p\|z_t - z\|^2 - (1-p)\|w_t - z\|^2\right]\right]$$

$$+ 2\gamma \cdot \mathbb{E}\left[\max_{z\in\mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} - z\rangle\right].$$

One can estimate $\mathbb{E}\left[\max_{z\in\mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} - z\rangle\right]$ the same way as (24)

$$2\gamma\mathbb{E}\left[\text{gap}(z_T^{avg})\right]$$

$$\leq \frac{4\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T} \sum_{t=1}^{T-1}(1 - \beta_{t+1}) + \frac{4\eta\Omega^2}{pT} + \frac{6\gamma^2 L^2 \Omega^2}{eT} + \frac{\eta\Omega^2}{T} + \frac{12\gamma^2\sigma^2}{eb}$$

$$+ \frac{\eta}{p} \cdot \mathbb{E}\left[\max_{z\in\mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|w_{t+1} - z\|^2 - p\|z_t - z\|^2 - (1-p)\|w_t - z\|^2\right]\right]$$

$$+ \frac{\Gamma\Omega^2}{T} + \frac{\gamma^2\sigma^2}{eb}$$

$$\leq \frac{5\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T} \sum_{t=1}^{T-1}(1 - \beta_{t+1}) + \frac{4\eta\Omega^2}{pT} + \frac{6\gamma^2 L^2 \Omega^2}{eT} + \frac{\eta\Omega^2}{T} + \frac{13\gamma^2\sigma^2}{eb}$$

$$+ \frac{\eta}{p} \cdot \mathbb{E}\left[\max_{z\in\mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \left[\|w_{t+1} - z\|^2 - p\|z_t - z\|^2 - (1-p)\|w_t - z\|^2\right]\right]$$

$$= \frac{5\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T} \sum_{t=1}^{T-1}(1 - \beta_{t+1}) + \frac{4\eta\Omega^2}{pT} + \frac{6\gamma^2 L^2 \Omega^2}{eT} + \frac{\eta\Omega^2}{T} + \frac{13\gamma^2\sigma^2}{eb}$$

$$+ \frac{\eta}{p} \cdot \mathbb{E}\left[\max_{z\in\mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \left[-2\langle pz_t + (1-p)w_t - w_{t+1}, z\rangle - p\|z_t\|^2 - (1-p)\|w_t\|^2 + \|w_{t+1}\|^2\right]\right]$$

$$= \frac{5\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T} \sum_{t=1}^{T-1}(1 - \beta_{t+1}) + \frac{4\eta\Omega^2}{pT} + \frac{6\gamma^2 L^2 \Omega^2}{eT} + \frac{\eta\Omega^2}{T} + \frac{13\gamma^2\sigma^2}{eb}$$

$$+ \frac{\eta}{p} \cdot \mathbb{E}\left[\max_{z\in\mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \left[-2\langle pz_t + (1-p)w_t - w_{t+1}, z\rangle\right]\right]$$

$$+ \frac{\eta}{p} \cdot \mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} \left[-p\|z_t\|^2 - (1-p)\|w_t\|^2 + \|w_{t+1}\|^2\right]\right].$$

One can note that by definition w_{t+1}: $\mathbb{E}\left[p\|z_t\|^2 + (1-p)\|w_t\|^2 - \|w_{t+1}\|^2\right] = 0$

$$2\gamma\mathbb{E}\left[\text{gap}(z_T^{avg})\right] \leq \frac{5\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T} \sum_{t=1}^{T-1}(1 - \beta_{t+1}) + \frac{4\eta\Omega^2}{pT} + \frac{6\gamma^2 L^2 \Omega^2}{eT} + \frac{\eta\Omega^2}{T} + \frac{13\gamma^2\sigma^2}{eb}$$

$$+ \frac{\eta}{p} \cdot \mathbb{E}\left[\max_{z\in\mathcal{Z}} \frac{1}{T} \sum_{t=0}^{T-1} \left[-2\langle pz_t + (1-p)w_t - w_{t+1}, z\rangle\right]\right].$$

With $\gamma \leq \frac{e}{10L}$ and $\eta \leq ep \leq e \leq \Gamma$, we get

$$
2\gamma\mathbb{E}\left[\text{gap}(z_T^{avg})\right] \leq \frac{11\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T}\sum_{t=1}^{T-1}(1-\beta_{t+1}) + \frac{13\gamma^2\sigma^2}{eb}
$$

$$
+ \frac{2\eta}{pT} \cdot \mathbb{E}\left[\max_{z\in\mathcal{Z}}\left[\left\langle\sum_{t=0}^{T-1}(w_{t+1}-pz_t-(1-p)w_t), z\right\rangle\right]\right]
$$

$$
\leq \frac{11\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T}\sum_{t=1}^{T-1}(1-\beta_{t+1}) + \frac{13\gamma^2\sigma^2}{eb}
$$

$$
+ \frac{2\eta}{pT} \cdot \mathbb{E}\left[\max_{z\in\mathcal{Z}}\left[\left\|\sum_{t=0}^{T-1}(w_{t+1}-pz_t-(1-p)w_t)\right\|^2 + \|z\|^2\right]\right].
$$

Again with the fact that $\mathbb{E}\left[pz_t + (1-p)w_t - w_{t+1}\right] = 0$.

$$
2\gamma\mathbb{E}\left[\text{gap}(z_T^{avg})\right] \leq \frac{11\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T}\sum_{t=1}^{T-1}(1-\beta_{t+1}) + \frac{13\gamma^2\sigma^2}{eb} + \frac{2\eta}{pT}\max_{z\in\mathcal{Z}}\|z\|^2
$$

$$
+ \frac{2\eta}{pT} \cdot \mathbb{E}\left[\sum_{t=0}^{T-1}\|w_{t+1}-pz_t-(1-p)w_t\|^2\right]
$$

$$
\leq \frac{13\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T}\sum_{t=1}^{T-1}(1-\beta_{t+1}) + \frac{13\gamma^2\sigma^2}{eb}
$$

$$
+ \frac{2\eta}{pT} \cdot \mathbb{E}\left[\sum_{t=0}^{T-1}\|w_{t+1}-pz_t-(1-p)w_t\|^2\right]
$$

$$
= \frac{13\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T}\sum_{t=1}^{T-1}(1-\beta_{t+1}) + \frac{13\gamma^2\sigma^2}{eb}
$$

$$
+ \frac{2\eta}{pT} \cdot \mathbb{E}\left[\sum_{t=0}^{T-1}\mathbb{E}_{w_{t+1}}\left[\left\|w_{t+1}-\mathbb{E}_{w_{t+1}}[w_{t+1}]\right\|^2\right]\right]
$$

$$
= \frac{13\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T}\sum_{t=1}^{T-1}(1-\beta_{t+1}) + \frac{13\gamma^2\sigma^2}{eb}
$$

$$
+ \frac{2\eta}{pT} \cdot \mathbb{E}\left[\sum_{t=0}^{T-1}\mathbb{E}_{w_{t+1}}\left[\|w_{t+1}\|^2\right] - \left\|\mathbb{E}_{w_{t+1}}[w_{t+1}]\right\|^2\right]
$$

$$
= \frac{13\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T}\sum_{t=1}^{T-1}(1-\beta_{t+1}) + \frac{13\gamma^2\sigma^2}{eb}
$$

$$
+ \frac{2\eta}{pT} \cdot \mathbb{E}\left[\sum_{t=0}^{T-1}\left[(1-p)\|w_t\|^2 + p\|z_t\|^2\right] - \|pz_t + (1-p)w_t\|^2\right]
$$

$$
= \frac{13\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T}\sum_{t=1}^{T-1}(1-\beta_{t+1}) + \frac{13\gamma^2\sigma^2}{eb}
$$

$$
+ \frac{2\eta}{pT} \cdot \mathbb{E}\left[\sum_{t=0}^{T-1}(1-p)p\|z_t - w_t\|^2\right]
$$

$$
\leq \frac{13\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T}\sum_{t=1}^{T-1}(1-\beta_{t+1}) + \frac{13\gamma^2\sigma^2}{eb}
$$

$$+ 8 \cdot \mathbb{E}\left[\frac{\eta}{4T}\sum_{t=0}^{T-1}\|z_t - w_t\|^2\right]. \tag{29}$$

It remains to estimate $\mathbb{E}\left[\frac{\eta}{4T}\sum_{t=0}^{T-1}\|z_t - w_t\|^2\right]$. For this, we substitute $z = z^*$ and take the expectation

$$2\gamma \cdot \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\langle F(z_{t+1/2}), z_{t+1/2} - z^*\rangle\right]$$

$$\leq \frac{\mathbb{E}\left[\|z_0 - z^*\|^2_{\hat{V}_{-1/2}}\right]}{T} + \frac{1}{T}\sum_{t=1}^{T-1}(1 - \beta_{t+1})C\mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right]$$

$$+ \frac{\eta\mathbb{E}\left[\|w_0 - z^*\|^2\right]}{pT} + \frac{3\gamma^2}{2eT}\|F(z_{-1/2}) - F(z_{-3/2})\|^2 + \frac{\eta}{4T}\|z_{-1} - w_{-1}\|^2$$

$$- \mathbb{E}\left[\frac{\eta}{4T}\sum_{t=1}^{T-1}\|z_t - w_t\|^2\right]$$

$$+ \frac{\eta}{p}\cdot\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|w_{t+1} - z^*\|^2 - p\|z_t - z^*\|^2 - (1-p)\|w_t - z^*\|^2\right]$$

$$+ 2\gamma\cdot\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\langle F(z_{t+1/2}) - g_{t+1/2}, z_{t+1/2} - z^*\rangle\right]$$

$$+ \frac{3\gamma^2}{e}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|g_{t+1/2} - F(z_{t+1/2})\|^2\right]$$

$$+ \frac{6\gamma^2}{e}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|F(z_{t-1/2}) - g_{t-1/2}\|^2\right] + \frac{3\gamma^2}{e}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|F(z_{t-3/2}) - g_{t-3/2}\|^2\right].$$

Assumptions 1 (convexity), 3 and the fact: $\mathbb{E}\left[p\|z_t\|^2 + (1-p)\|w_t\|^2 - \|w_{t+1}\|^2\right] = 0$, give

$$\mathbb{E}\left[\frac{\eta}{4T}\sum_{t=1}^{T-1}\|z_t - w_t\|^2\right] \leq \frac{\mathbb{E}\left[\|z_0 - z^*\|^2_{\hat{V}_{-1/2}}\right]}{T} + \frac{1}{T}\sum_{t=1}^{T-1}(1 - \beta_{t+1})C\mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right]$$

$$+ \frac{\eta\mathbb{E}\left[\|w_0 - z^*\|^2\right]}{pT} + \frac{3\gamma^2}{2eT}\|F(z_{-1/2}) - F(z_{-3/2})\|^2 + \frac{\eta}{4T}\|z_{-1} - w_{-1}\|^2$$

$$+ \frac{12\gamma^2\sigma^2}{eb}$$

$$\leq \frac{4\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T}\sum_{t=1}^{T-1}(1 - \beta_{t+1}) + \frac{4\eta\Omega^2}{pT} + \frac{3\gamma^2L^2\Omega^2}{2eT} + \frac{\eta\Omega^2}{T} + \frac{12\gamma^2\sigma^2}{eb}. \tag{30}$$

Combining (29) and (30), we get

$$2\gamma\mathbb{E}\left[\text{gap}(z_T^{avg})\right] \leq \frac{13\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T}\sum_{t=1}^{T-1}(1 - \beta_{t+1}) + \frac{13\gamma^2\sigma^2}{eb}$$

$$+ 8\cdot\left[\frac{4\Gamma\Omega^2}{T} + \frac{4C\Gamma\Omega^2}{T}\sum_{t=1}^{T-1}(1 - \beta_{t+1}) + \frac{4\eta\Omega^2}{pT} + \frac{3\gamma^2L^2\Omega^2}{2eT} + \frac{\eta\Omega^2}{T} + \frac{12\gamma^2\sigma^2}{eb}\right]$$

$$\leq \frac{100\Gamma\Omega^2}{T} + \frac{36C\Gamma\Omega^2}{T}\sum_{t=1}^{T-1}(1 - \beta_{t+1}) + \frac{110\gamma^2\sigma^2}{eb}.$$

Finally, we have

$$\mathbb{E}\left[\text{gap}(z_T^{avg})\right] \leq \frac{50\Gamma\Omega^2}{\gamma T} + \frac{18C\Gamma\Omega^2}{\gamma T}\sum_{t=1}^{T}(1 - \beta_t) + \frac{55\gamma\sigma^2}{eb}.$$

42

This completes the proof of the convex–concave case. $\square$

Non-convex–non-concave case. We start from (27), substitute $z = z^*$ and take the total expectation of both sides of the equation

$$\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_{t-1/2}}\right] + \frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2\right] + \frac{\eta}{2}\mathbb{E}\left[\|z_t - w_t\|^2\right]$$

$$\leq \mathbb{E}\left[\|z_t - z\|^2_{\hat{V}_{t-1/2}}\right] - \eta\mathbb{E}\left[\|z_t - z^*\|^2\right] + \eta\mathbb{E}\left[\|w_t - z^*\|^2\right]$$

$$- 2\gamma\mathbb{E}\left[\langle F(z^{t+1/2}), z_{t+1/2} - z^*\rangle\right]$$

$$+ \frac{1}{2}\cdot\frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2}\cdot\frac{\eta}{2}\mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]$$

$$- \frac{1}{4}\mathbb{E}\left[\|z_{t+1/2} - z_t\|^2_{\hat{V}_{t-1/2}}\right] + \frac{3\gamma^2}{e}\mathbb{E}\left[\|g_{t+1/2} - F(z_{t+1/2})\|^2\right]$$

$$+ \frac{6\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - g_{t-1/2}\|^2\right] + \frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-3/2}) - g_{t-3/2}\|^2\right].$$

Assumptions 1 and 3 on non-convexity–non-concavity and on stochastisity give

$$\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_{t-1/2}}\right] + \frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2\right] + \frac{\eta}{2}\mathbb{E}\left[\|z_t - w_t\|^2\right]$$

$$\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right] - \eta\mathbb{E}\left[\|z_t - z^*\|^2\right] + \eta\mathbb{E}\left[\|w_t - z^*\|^2\right]$$

$$+ \frac{1}{2}\cdot\frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2}\cdot\frac{\eta}{2}\mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]$$

$$- \frac{\gamma^2}{4}\mathbb{E}\left[\|g_{t-1/2}\|^2_{\hat{V}^{-1}_{t-1/2}}\right] + \frac{12\gamma^2\sigma^2}{eb}$$

$$\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right] - \eta\mathbb{E}\left[\|z_t - z^*\|^2\right] + \eta\mathbb{E}\left[\|w_t - z^*\|^2\right]$$

$$+ \frac{1}{2}\cdot\frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2}\cdot\frac{\eta}{2}\mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]$$

$$- \frac{\gamma^2}{4\Gamma}\mathbb{E}\left[\|g_{t-1/2}\|^2\right] + \frac{12\gamma^2\sigma^2}{eb}$$

$$\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right] - \eta\mathbb{E}\left[\|z_t - z^*\|^2\right] + \eta\mathbb{E}\left[\|w_t - z^*\|^2\right]$$

$$+ \frac{1}{2}\cdot\frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2}\cdot\frac{\eta}{2}\mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]$$

$$- \frac{\gamma^2}{8\Gamma}\mathbb{E}\left[\|F(z_{t-1/2})\|^2\right] + \frac{\gamma^2}{4\Gamma}\mathbb{E}\left[\|g_{t-1/2} - F(z_{t-1/2})\|^2\right] + \frac{12\gamma^2\sigma^2}{eb}$$

$$\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right] - \eta\mathbb{E}\left[\|z_t - z^*\|^2\right] + \eta\mathbb{E}\left[\|w_t - z^*\|^2\right]$$

$$+ \frac{1}{2}\cdot\frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2}\cdot\frac{\eta}{2}\mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]$$

$$- \frac{\gamma^2}{8\Gamma}\mathbb{E}\left[\|F(z_{t-1/2})\|^2\right] + \frac{13\gamma^2\sigma^2}{eb}.$$

Here we additionally use $\frac{1}{\Gamma}I \preccurlyeq \hat{V}^{-1}_{t-1/2}$. Lemma 1 with $C = \frac{\Gamma^2}{2e^2}$ for (7), $C = \frac{2\Gamma}{e}$ for (8) and $B_{t+1} = (1 + (1 - \beta_{t+1})C)$ gives

$$\frac{1}{B_{t+1}}\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_{t+1/2}}\right] + \frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2\right] + \frac{\eta}{2}\mathbb{E}\left[\|z_t - w_t\|^2\right]$$

$$\leq \mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right] - \eta\mathbb{E}\left[\|z_t - z^*\|^2\right] + \eta\mathbb{E}\left[\|w_t - z^*\|^2\right]$$

$$+ \frac{1}{2}\cdot\frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] + \frac{1}{2}\cdot\frac{\eta}{2}\mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right]$$

$$-\frac{\gamma^2}{8\Gamma}\mathbb{E}\left[\|F(z_{t-1/2})\|^2\right] + \frac{13\gamma^2\sigma^2}{eb}.$$

Next, after small rearrangement, we obtain

$$
\begin{aligned}
\frac{\gamma^2}{8\Gamma}\mathbb{E}\left[\|F(z_{t-1/2})\|^2\right] \leq &\mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right] - \frac{1}{B_{t+1}}\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_{t+1/2}}\right] \\
&- \eta\mathbb{E}\left[\|z_t - z^*\|^2\right] + \eta\mathbb{E}\left[\|w_t - z^*\|^2\right] \\
&+ \frac{1}{2}\cdot\frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] - \frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2\right] \\
&+ \frac{1}{2}\cdot\frac{\eta}{2}\mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right] - \frac{\eta}{2}\mathbb{E}\left[\|z_t - w_t\|^2\right] \\
&+ \frac{13\gamma^2\sigma^2}{eb}.
\end{aligned}
$$

The update of w_{t+1} gives

$$
\begin{aligned}
\frac{\eta}{p}\mathbb{E}\left[\|w_{t+1} - z^*\|^2\right] &= \frac{\eta}{p}\mathbb{E}\left[\mathbb{E}_{w_{t+1}}\left[\|w_{t+1} - z^*\|^2\right]\right] \\
&= \eta\mathbb{E}\left[\|z_t - z^*\|^2\right] + \frac{(1-p)\eta}{p}\mathbb{E}\left[\|w_t - z^*\|^2\right].
\end{aligned}
$$

Connecting with the previous equation, we get

$$
\begin{aligned}
\frac{\gamma^2}{8\Gamma}\mathbb{E}\left[\|F(z_{t-1/2})\|^2\right] \leq &\mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right] - \frac{1}{B_{t+1}}\mathbb{E}\left[\|z_{t+1} - z^*\|^2_{\hat{V}_{t+1/2}}\right] \\
&+ \frac{\eta}{p}\mathbb{E}\left[\|w_t - z^*\|^2\right] - \frac{\eta}{p}\mathbb{E}\left[\|w_{t+1} - z^*\|^2\right] \\
&+ \frac{1}{2}\cdot\frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t-1/2}) - F(z_{t-3/2})\|^2\right] - \frac{3\gamma^2}{e}\mathbb{E}\left[\|F(z_{t+1/2}) - F(z_{t-1/2})\|^2\right] \\
&+ \frac{1}{2}\cdot\frac{\eta}{2}\mathbb{E}\left[\|z_{t-1} - w_{t-1}\|^2\right] - \frac{\eta}{2}\mathbb{E}\left[\|z_t - w_t\|^2\right] \\
&+ \frac{13\gamma^2\sigma^2}{eb}.
\end{aligned}
$$

Summing over all t from 0 to $T-1$ and averaging, we get

$$
\begin{aligned}
\frac{\gamma^2}{8\Gamma}\cdot\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|F(z_{t-1/2})\|^2\right] \leq &\frac{\mathbb{E}\left[\|z_0 - z^*\|^2_{\hat{V}_{-1/2}}\right]}{T} + \frac{1}{T}\sum_{t=1}^{T-1}\frac{B_t - 1}{B_t}\mathbb{E}\left[\|z_t - z^*\|^2_{\hat{V}_{t-1/2}}\right] \\
&+ \frac{\eta}{p}\mathbb{E}\left[\|w_0 - z^*\|^2\right] + \frac{3\gamma^2}{2e}\mathbb{E}\left[\|F(z_{-1/2}) - F(z_{-3/2})\|^2\right] \\
&+ \frac{\eta}{4}\mathbb{E}\left[\|z_{-1} - w_{-1}\|^2\right] + \frac{13\gamma^2\sigma^2}{eb}.
\end{aligned}
$$

With the fact that iterations are bounded: $\|z_t\| \leq \Omega$, we obtain

$$
\begin{aligned}
\frac{\gamma^2}{8\Gamma}\cdot\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|F(z_{t-1/2})\|^2\right] \leq &\frac{4\Gamma\Omega^2}{T} + \frac{8\eta\Omega^2}{pT} + \frac{12\gamma^2 L^2\Omega^2}{2eT} + \frac{13\gamma^2\sigma^2}{eb} \\
&+ \frac{4\Gamma\Omega^2}{T}\sum_{t=1}^{T-1}\frac{(1-\beta_t)C}{1+(1-\beta_t)C} \\
\leq &\frac{4\Gamma\Omega^2}{T} + \frac{8\eta\Omega^2}{pT} + \frac{12\gamma^2 L^2\Omega^2}{2eT} + \frac{13\gamma^2\sigma^2}{eb} \\
&+ \frac{4\Gamma C\Omega^2}{T}\sum_{t=1}^{T-1}(1-\beta_t).
\end{aligned}
$$

Here we use $(1 + (1 - \beta_t)C) \geq 1$. With $\eta \leq ep$, we get

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|F(z_{t-1/2})\|^2\right] \leq \frac{96\Gamma^2\Omega^2}{\gamma^2 T} + \frac{48\Gamma L^2\Omega^2}{eT} + \frac{32\Gamma^2 C\Omega^2}{\gamma^2 T}\sum_{t=1}^{T-1}(1-\beta_t) + \frac{105\Gamma\sigma^2}{eb}.$$

If we choose $\bar{z}_{T-1/2}$ randomly and uniformly, then $\mathbb{E}\left[\|F(\bar{z}_{T-1/2})\|^2\right] \leq \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|F(z_{t-1/2})\|^2\right]$ and

$$\mathbb{E}\left[\|F(\bar{z}_{T-1/2})\|^2\right] \leq \frac{96\Gamma^2\Omega^2}{\gamma^2 T} + \frac{48\Gamma L^2\Omega^2}{eT} + \frac{32\Gamma^2 C\Omega^2}{\gamma^2 T}\sum_{t=1}^{T-1}(1-\beta_t) + \frac{105\Gamma\sigma^2}{eb}.$$

This completes the proof of the non-convex–non-concave case. $\square$

C.7 Proof of Corollary 3

Strongly convex–strongly concave case. With $\beta_t \equiv \beta \geq 1 - \frac{\gamma\mu}{4\Gamma C}$, we have (13) in the following form

$$\mathbb{E}\left[\Psi_{t+1}\right] \leq \max\left[\left(1 - \frac{\gamma\mu}{4\Gamma}\right); \left(1 - \frac{1}{\frac{2\eta}{\gamma\mu p} + \frac{1}{p}}\right)\right] \cdot \mathbb{E}\left[\Psi_t\right] + \frac{24\gamma^2\sigma^2}{eb}.$$

Running the recursion, we have

$$\mathbb{E}\left[\Psi_T\right] \leq \max\left[\left(1 - \frac{\gamma\mu}{4\Gamma}\right)^T; \left(1 - \frac{1}{\frac{2\eta}{\gamma\mu p} + \frac{1}{p}}\right)^T\right] \cdot \Psi_0$$

$$+ \frac{12\gamma^2\sigma^2}{eb}\sum_{t=0}^{T-1}\max\left[\left(1 - \frac{\gamma\mu}{4\Gamma}\right)^t; \left(1 - \frac{1}{\frac{2\eta}{\gamma\mu p} + \frac{1}{p}}\right)^t\right]$$

$$\leq \max\left[\exp\left(-\frac{\gamma\mu T}{4\Gamma}\right); \exp\left(-\frac{1}{\frac{2\eta}{\gamma\mu p} + \frac{1}{p}}\right)\right] \cdot \Psi_0 + \frac{48\Gamma\gamma\sigma^2}{eb\mu} + \frac{48\eta\gamma\sigma^2}{ep\mu b} + \frac{24\gamma^2\sigma^2}{epb}.$$

Finally, we need tuning of $\gamma \leq \frac{e}{10L}$:

- If $\frac{e}{10L} \geq \frac{4\Gamma\ln\left(\max\{2, eb\mu^2\Psi_0 T/(48\Gamma^2\sigma^2)\}\right)}{\mu T}$ then $\gamma = \frac{2\Gamma\ln\left(\max\{2, eb\mu^2 R_0^2 T/(48\Gamma^2\sigma^2)\}\right)}{\mu T}$ gives

$$\tilde{\mathcal{O}}\left(\exp\left(-\ln\left(\max\{2, eb\mu^2 R_0^2 T/(48\Gamma^2\sigma^2)\}\right)\right)R_0^2 + \frac{\Gamma^2\sigma^2}{e\mu^2 T}\right) = \tilde{\mathcal{O}}\left(\frac{\Gamma^2\sigma^2}{eb\mu^2 T}\right).$$

- If $\frac{e}{10L} \leq \frac{4\Gamma\ln\left(\max\{2, eb\mu^2 R_0^2 T/(48\Gamma^2\sigma^2)\}\right)}{\mu T}$ then $\gamma = \frac{e}{10L}$ gives

$$\tilde{\mathcal{O}}\left(\exp\left(-\frac{e\mu T}{40\Gamma L}\right)\Psi_0 + \exp\left(-\frac{ep\mu T}{40\eta L}\right)\Psi_0 + \exp\left(-p\right)\Psi_0 + \frac{\Gamma\gamma\sigma^2}{eb\mu}\right)$$

$$\leq \tilde{\mathcal{O}}\left(\exp\left(-\frac{e\mu T}{40\Gamma L}\right)\Psi_0 + \exp\left(-\frac{ep\mu T}{40\eta L}\right)\Psi_0 + \exp\left(-p\right)\Psi_0 + \frac{\Gamma^2\sigma^2}{eb\mu^2 T}\right).$$

What in the end with $\eta \leq ep \leq e$ gives that

$$\mathbb{E}\left[\Psi_T\right] = \tilde{\mathcal{O}}\left(\exp\left(-\frac{e\mu T}{40\Gamma L}\right)\Psi_0 + \exp\left(-\frac{\mu T}{40L}\right)\Psi_0 + \exp\left(-p\right)\Psi_0 + \frac{\Gamma^2\sigma^2}{eb\mu^2 T}\right).$$

This completes the proof of the strongly convex–strongly concave case. $\square$

Convex–concave case. With $\beta_t \equiv \beta \geq 1 - \frac{1}{CT}$, we have (14) in the following form

$$\mathbb{E}\left[\text{gap}(x_T^{avg}, y_T^{avg})\right] \leq \frac{68\Gamma\Omega^2}{\gamma T} + \frac{55\gamma\sigma^2}{e}.$$

If $\gamma = \min\left\{\frac{e}{10L}; \frac{\sqrt{\Gamma eb}\Omega}{\sigma\sqrt{T}}\right\}$, then

$$\mathbb{E}\left[\text{gap}(x_T^{avg}, y_T^{avg})\right] = \mathcal{O}\left(\frac{\Gamma L\Omega^2}{eT} + \frac{\sqrt{\Gamma}\sigma\Omega}{\sqrt{ebT}}\right).$$

This completes the proof of the convex–concave case. □

Non-convex–non-concave case. With $\beta_t \equiv \beta \geq 1 - \frac{1}{CT}$, we have (15) in the following form

$$\mathbb{E}[\|\nabla_x f(\bar{x}_{T-1/2}, \bar{y}_{T-1/2})\|^2 + \|\nabla_y f(\bar{x}_{T-1/2}, \bar{y}_{T-1/2})\|^2]$$
$$\leq \frac{128\Gamma^2\Omega^2}{\gamma^2 T} + \frac{48\Gamma L^2\Omega^2}{eT} + \frac{105\Gamma\sigma^2}{eb}.$$

The choice of $\gamma = \frac{e}{10L}$ gives

$$\mathbb{E}[\|\nabla_x f(\bar{x}_{T-1/2}, \bar{y}_{T-1/2})\|^2 + \|\nabla_y f(\bar{x}_{T-1/2}, \bar{y}_{T-1/2})\|^2] \leq \frac{120^2\Gamma^2 L^2\Omega^2}{e^2 T} + \frac{105\Gamma\sigma^2}{eb}.$$

The batch size $b \sim \frac{\Gamma\sigma^2}{e\varepsilon^2}$ completes the proof of the non-convex–non-concave case. □

C.8 Proof of Corollary 4

Strongly convex–strongly concave case. We start from (13) with small rearrangement

$$\min\left[\frac{\gamma\mu}{4\Gamma}; \frac{1}{\frac{4\eta}{\gamma\mu p} + \frac{2}{p}}\right]\mathbb{E}\left[\Psi_t\right] \leq \max\left[\left(1 - \frac{\gamma\mu}{4\Gamma} + (1 - \beta_{t+1})C\right); \left(1 - \frac{1}{\frac{4\eta}{\gamma\mu p} + \frac{2}{p}}\right)\right] \cdot \mathbb{E}\left[\Psi_t\right]$$
$$- \mathbb{E}\left[\Psi_{t+1}\right] + \frac{12\gamma^2\sigma^2\left(1 + (1 - \beta_{t+1})C\right)}{eb};$$

Summing over all t from 0 to $T - 1$ and averaging, we get

$$\min\left[\frac{\gamma\mu}{4\Gamma}; \frac{1}{\frac{4\eta}{\gamma\mu p} + \frac{2}{p}}\right]\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\Psi_t\right]$$
$$\leq \frac{\mathbb{E}\left[\Psi_0\right]}{T} + \frac{1}{T}\sum_{t=0}^{T-1}\left((1 - \beta_{t+1})C - \frac{\gamma\mu}{4\Gamma}\right)\mathbb{E}\left[\Psi_t\right]$$
$$+ \frac{12\gamma^2\sigma^2}{eb} \cdot \frac{1}{T}\sum_{t=0}^{T-1}\left(1 + (1 - \beta_{t+1})C\right).$$

Using $eI \preccurlyeq \hat{V}_t \preccurlyeq \Gamma I$ and $\|z_t\| \leq \Omega$, we have

$$\min\left[\frac{e\gamma\mu}{4\Gamma}; \frac{e}{\frac{4\eta}{\gamma\mu p} + \frac{2}{p}}\right]\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|z_t - z^*\|^2\right] \leq \frac{20\Gamma\Omega^2}{T} + \frac{1}{T}\sum_{t=1}^{T}\left((1 - \beta_t)C - \frac{\gamma\mu}{4\Gamma}\right)\mathbb{E}\left[\Psi_t\right]$$
$$+ \frac{12\gamma^2\sigma^2}{eb} \cdot \frac{1}{T}\sum_{t=1}^{T}\left(1 + (1 - \beta_t)C\right).$$

Let us define $\alpha = \left(\frac{4\Gamma C}{\gamma\mu}\right)^2$. Then, $1 - \beta_t = \frac{1-\beta}{1-\beta^{t+1}}$ with $\beta = 1 - \frac{1}{\alpha}$. (18) gives

$$\beta^{\sqrt{\alpha}} \leq 1 - \frac{1}{2\sqrt{\alpha}}.$$

And hence, for all $t \geq \sqrt{\alpha}$

$$1 - \beta_t \leq \frac{2\sqrt{\alpha}}{\alpha} = \frac{2}{\sqrt{\alpha}} = \frac{\gamma\mu}{4\Gamma C}.$$

Then we get

$$\min\left[\frac{e\gamma\mu}{4\Gamma};\frac{e}{\frac{4\eta}{\gamma\mu p}+\frac{2}{p}}\right]\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|z_t-z^*\|^2\right]$$

$$\leq\frac{20\Gamma\Omega^2}{T}+\frac{1}{T}\sum_{t=1}^{\sqrt{\alpha}}\left((1-\beta_t)C-\frac{\gamma\mu}{4\Gamma}\right)\mathbb{E}\left[\Psi_t\right]$$

$$+\frac{12\gamma^2\sigma^2}{eb}\cdot\frac{1}{T}\sum_{t=1}^{T}(1+(1-\beta_t)C)$$

$$\leq\frac{20\Gamma\Omega^2}{T}+\frac{1}{T}\sum_{t=1}^{\sqrt{\alpha}}20C\Gamma\Omega^2+\frac{12\gamma^2\sigma^2}{eb}\cdot\frac{1}{T}\sum_{t=1}^{T}(1+(1-\beta_t)C)$$

$$\leq\frac{20\Gamma\Omega^2}{T}+\frac{20C\Gamma\Omega^2}{T}\cdot\frac{4\Gamma C}{\gamma\mu}+\frac{12\gamma^2\sigma^2}{eb}\cdot\frac{1}{T}\sum_{t=1}^{T}(1+(1-\beta_t)C)$$

$$\leq\frac{100C^2\Gamma^2\Omega^2}{\gamma\mu T}+\frac{24C\gamma^2\sigma^2}{eb}.$$

Then,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|z_t-z^*\|^2\right]=\mathcal{O}\left(\frac{C^2\Gamma^3\Omega^2}{e\gamma^2\mu^2 T}+\frac{C\Gamma\gamma\sigma^2}{e^2\mu b}+\frac{C^2\Gamma^2\Omega^2}{\gamma^2\mu^2 T}+\frac{C\gamma\sigma^2}{e\mu b}+\frac{C^2\Gamma^2\Omega^2}{ep\gamma\mu T}+\frac{C\gamma^2\sigma^2}{e^2 pb}\right).$$

Finally, $\gamma=\min\left\{\frac{e}{10L};\sqrt[3]{\frac{C\Gamma^2\Omega^2 eb}{\mu\sigma^2 T}}\right\}$

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=0}^{T-1}z_t-z^*\right\|^2\right]\leq\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|z_t-z^*\|^2\right]$$

$$=\mathcal{O}\left(\frac{C^2\Gamma^3 L^2\Omega^2}{e^3\mu^2 T}+\frac{C^2\Gamma^2 L^2\Omega^2}{e^2\mu^2 T}+\frac{C^2\Gamma^2 L\Omega^2}{e^2 p\mu T}+\left(\frac{C^4\Gamma^5\sigma^4\Omega^2}{e^5\mu^4 b^2 T}\right)^{1/3}\right.$$

$$\left.+\left(\frac{C^4\Gamma^2\sigma^4\Omega^2}{e^2\mu^4 b^2 T}\right)^{1/3}+\left(\frac{C^5\Gamma^4\sigma^2\Omega^4}{e^4\mu^2 p^3 bT^2}\right)^{1/3}\right).$$

This completes the proof of the strongly convex–strongly concave case. $\square$

Convex–concave case. With $\beta_t=\frac{\beta-\beta^{t+1}}{1-\beta^{t+1}}$ we get (14) in the following form

$$\mathbb{E}\left[\text{gap}(x_T^{avg},y_T^{avg})\right]\leq\frac{50\Gamma\Omega^2}{\gamma T}+\frac{55\gamma\sigma^2}{eb}$$

$$+\frac{18(1-\beta)C\Gamma\Omega^2}{\gamma T}\sum_{t=1}^{\sqrt{T}}\frac{1}{1-\beta}+\frac{18C\Gamma\Omega^2}{\gamma T}\sum_{t=1}^{T}\frac{1}{1-\beta^{\sqrt{T}}}$$

$$\leq\frac{50\Gamma\Omega^2}{\gamma T}+\frac{55\gamma\sigma^2}{eb}$$

$$+\frac{18C\Gamma\Omega^2}{\gamma\sqrt{T}}+\frac{18(1-\beta)C\Gamma\Omega^2}{\gamma T}\sum_{t=1}^{T}\frac{1}{1-\beta^{\sqrt{T}}}.$$

Next, we substitute $\beta=1-\frac{1}{T}$. From (18) we get that $\beta^{\sqrt{T}}\leq 1-\frac{1}{2\sqrt{T}}$. Then, we get

$$\mathbb{E}\left[\text{gap}(x_T^{avg},y_T^{avg})\right]\leq\frac{50\Gamma\Omega^2}{\gamma T}+\frac{55\gamma\sigma^2}{eb}+\frac{54C\Gamma\Omega^2}{\gamma\sqrt{T}}.$$

It means that

$$\mathbb{E}\left[\text{gap}(x_T^{avg}, y_T^{avg})\right] = \mathcal{O}\left(\frac{C\Gamma\Omega^2}{\gamma\sqrt{T}} + \frac{\gamma\sigma^2}{eb}\right).$$

With $\gamma = \min\left\{\frac{e}{10L}; \frac{\sqrt{C\Gamma eb}\Omega}{\sigma T^{1/4}}\right\}$

$$\mathbb{E}\left[\text{gap}(x_T^{avg}, y_T^{avg})\right] = \mathcal{O}\left(\frac{C\Gamma L\Omega^2}{e\sqrt{T}} + \frac{\sqrt{C\Gamma eb}\sigma\Omega}{eT^{1/4}}\right).$$

This completes the proof of the convex–concave case. $\square$

Non-convex–non-concave case. With $\beta_t = \frac{\beta - \beta^{t+1}}{1 - \beta^{t+1}}$ we get (15) in the following form

$$\mathbb{E}[\|\nabla_x f(\bar{x}_{T-1/2}, \bar{y}_{T-1/2})\|^2 + \|\nabla_y f(\bar{x}_{T-1/2}, \bar{y}_{T-1/2})\|^2]$$

$$\leq \frac{96\Gamma^2\Omega^2}{\gamma^2 T} + \frac{48\Gamma L^2\Omega^2}{eT} + \frac{32(1-\beta)\Gamma^2 C\Omega^2}{\gamma^2 T}\sum_{t=1}^{T}\frac{1}{1-\beta^t} + \frac{105\Gamma\sigma^2}{eb}$$

$$\leq \frac{96\Gamma^2\Omega^2}{\gamma^2 T} + \frac{48\Gamma L^2\Omega^2}{eT} + \frac{105\Gamma\sigma^2}{eb}$$

$$+ \frac{32(1-\beta)\Gamma^2 C\Omega^2}{\gamma^2 T}\sum_{t=1}^{\sqrt{T}}\frac{1}{1-\beta} + \frac{32(1-\beta)\Gamma^2 C\Omega^2}{\gamma^2 T}\sum_{t=1}^{T}\frac{1}{1-\beta^{\sqrt{T}}}.$$

Next, we substitute $\beta = 1 - \frac{1}{T}$. From (18) we get that $\beta^{\sqrt{T}} \leq 1 - \frac{1}{2\sqrt{T}}$. Then, we get

$$\mathbb{E}[\|\nabla_x f(\bar{x}_{T-1/2}, \bar{y}_{T-1/2})\|^2 + \|\nabla_y f(\bar{x}_{T-1/2}, \bar{y}_{T-1/2})\|^2]$$

$$\leq \frac{96\Gamma^2\Omega^2}{\gamma^2 T} + \frac{48\Gamma L^2\Omega^2}{eT} + \frac{105\Gamma\sigma^2}{eb} + \frac{96\Gamma^2 C\Omega^2}{\gamma^2\sqrt{T}}.$$

The batch size $b \sim \frac{\Gamma\sigma^2}{e\varepsilon^2}$ and $\gamma = \frac{e}{10L}$ complete the proof of the non-convex–non-concave case. $\square$