Stochastic Compositional Minimax Optimization with Provable Convergence Guarantees

Yuyang Deng
Columbia University

Fuli Qiao
Penn State University

Mehrdad Mahdavi
Penn State University

Abstract

Stochastic compositional minimax problems are prevalent in machine learning, yet there exist only limited established findings on the convergence of this class of problems. In this paper, we propose a formal definition of the stochastic compositional minimax problem, which involves optimizing a minimax loss with a compositional structure either in primal, dual, or both primal and dual variables. We introduce a simple yet effective algorithm, stochastically Corrected stOchastic gradient Descent Ascent (CODA), which is a primal-dual type algorithm with compositional correction steps, and establish its convergence rate in the aforementioned three settings. We also propose a variance reduced variant, CODA+, which achieves the best-known rate on nonconvex-strongly-concave and nonconvex-concave compositional minimax problems. This work initiates the theoretical study of the stochastic compositional minimax problem in various settings and may inform modern machine learning scenarios such as domain adaptation or robust model-agnostic meta-learning.

1 Introduction

We consider the following minimax problem

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{x},\mathbf{y}) := h(\mathbf{x}) + f(\mathbf{x},\mathbf{y}) - r(\mathbf{y}), \quad (1)$$

where $f(\mathbf{x},\mathbf{y})$ is a stochastic compositional function. The compositional part can be defined over $\mathbf{x}$, $\mathbf{y}$ or both $\mathbf{x}$ and $\mathbf{y}$, which correspond to the following three scenarios:

- Composition on $\mathbf{x}$: $f(\mathbf{x},\mathbf{y}) := \mathbb{E}_{\zeta\sim\mathcal{S}_f}[f(\mathbb{E}_{\xi\sim\mathcal{S}_g}[g(\mathbf{x};\xi)],\mathbf{y};\zeta)]$
- Composition on $\mathbf{y}$: $f(\mathbf{x},\mathbf{y}) := \mathbb{E}_{\zeta\sim\mathcal{S}_f}[f(\mathbf{x},\mathbb{E}_{\xi\sim\mathcal{S}_g}[g(\mathbf{y};\xi)];\zeta)]$
- Composition on both $\mathbf{x}$ and $\mathbf{y}$: $f(\mathbf{x},\mathbf{y}) := \mathbb{E}_{\zeta\sim\mathcal{S}_f}[f(\mathbb{E}_{\xi\sim\mathcal{S}_g}[g(\mathbf{x},\mathbf{y};\xi)];\zeta)]$,

where $\mathcal{S}_f$ and $\mathcal{S}_g$ capture inner and outer stochasticity in composition, respectively.

Many practical machine learning problems can be cast as an instance of the above optimization problem, such as robust meta learning [Collins et al., 2020], AUC maximization [Zhang et al., 2023], statistical discrepancy minimization [Mansour et al., 2009], multiple source domain adaptation [Konstantinov and Lampert, 2019, Mansour et al., 2021, Deng et al., 2023], and stochastic compositional constrained programming [Yang et al., 2022]. Here we briefly describe several important applications that motivate the proposed compositional minimax problem (1).

AUC Optimization [Yuan et al., 2021] AUC optimization has been a powerful training paradigm that can mitigate the training data imbalance issue in classification problems. Consider a binary classification problem with label $y \in \{-1,+1\}$, and model parameter $\mathbf{w} \in \mathbb{R}^d$. Our goal is to solve the following problem:

$$\min_{\mathbf{w},a,b} \max_{\theta\in\Omega} \Phi(\mathbf{w}-\alpha\nabla L(\mathbf{w}),a,b,\theta) = \frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{w}-\alpha\nabla L(\mathbf{w}),a,b,\theta;\mathbf{x}_i,y_i),$$

where $L(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{w};\mathbf{x}_i,y_i)$. Usually, the objective has a negative quadratic dependency on dual variable θ, which renders this objective a nonconvex-strongly-concave minimax problem.

Robust MAML [Collins et al., 2020] In task robust MAML, our goal is to learn a model that can perform well over all distributions of the observed tasks, hence making it robust to task shifts in the testing time. The objective can be formulated as the follow-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

ing minimax problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d}\max_{\boldsymbol{\alpha}\in\Delta^N}\sum_{i=1}^{N}\alpha(i)L_i(\mathbf{w}-\eta\nabla L_i(\mathbf{w}))$$

where Δ^N denotes simplex, $L_i(\mathbf{w}) := \frac{1}{n_i}\sum_{j=1}^{n_i}\ell(\mathbf{w};\xi_j)$, N is the number of tasks with task i having n_i samples. It is a nonconvex-concave compositional minimax problem, with composition on the primal variable.

Stochastic Minimization with Compositional Constraint [Yang et al., 2022] In many risk management problems, such as conditional value-at-risk (CVaR) [Rockafellar et al., 2000], risk-averse mean-deviation constraint [Cole et al., 2017], or portfolio optimization with high-moment constraint [Harvey et al., 2010], we are tasked with solving the following programming with compositional constraint:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \quad \text{s.t. } g_1(\mathbb{E}[g_2(\mathbf{x};\xi)])\le 0.$$

To solve the above constraint problem, one may consider solving the unconstrained variant:

$$\min_{\mathbf{x}}\max_{\lambda} f(\mathbf{x})+\lambda g_1(\mathbb{E}[g_2(\mathbf{x};\xi)])$$

which turns the original objective into a nonconvex-concave compositional minimax problem, with composition on the primal variable.

Mixture Weights Estimation [Konstantinov and Lampert, 2019] In multiple source domain adaptation, the goal is learn a predictive model for a target domain by finding a mixture weights $\boldsymbol{\alpha}$ to optimally aggregate empirical risks over N data sources, which can be achieved by solving the following minimax problem:

$$\min_{\boldsymbol{\alpha}\in\Delta^N}\max_{\mathbf{w}\in\mathcal{W}}\sum_{i=1}^{N}\alpha(i)f(\mathcal{L}_{\mathcal{T}}(\mathbf{w})-\mathcal{L}_{\mathcal{D}_i}(\mathbf{w}))+\lambda\|\boldsymbol{\alpha}\|_{\mathbf{M}}$$

where $\boldsymbol{\alpha}$ is the mixture weight, $\mathbf{x}$ is the model parameter, $\mathcal{L}_{\mathcal{T}}(\mathbf{w})$ is the empirical risk realized by target domain data, and $\mathcal{L}_{\mathcal{D}_i}(\mathbf{w})$ is the empirical risk realized by ith source domain data. $f(z)$ is the smooth approximation of the absolute value function. $\|\cdot\|_{\mathbf{M}}^2=\boldsymbol{\alpha}^\top\mathbf{M}\boldsymbol{\alpha}$ where $\mathbf{M}=\mathrm{diag}\,(1/m_1,\ldots,1/m_N)$. This is a convex-nonconcave minimax problem with composition on the dual variable, and if we relax the norm of $\boldsymbol{\alpha}$ to the squared norm, it becomes a strongly-convex-nonconcave minimax problem.

Discrepancy Minimization [Mansour et al., 2009, Sriperumbudur et al., 2009] In the domain adaptation and empirical estimation of integral probability metrics (IPMs), the goal is to estimate the statistical discrepancy among two distributions or to manipulate a source distribution to align it well with a target distribution, leading to the following problem:

$$\min_{\mathcal{S}}\sup_{\mathbf{w},\mathbf{w}'\in\mathcal{W}}|\mathcal{L}_{\mathcal{S}}(\mathbf{w},\mathbf{w}')-\mathcal{L}_{\mathcal{T}}(\mathbf{w},\mathbf{w}')|$$

where $\mathcal{L}_{\mathcal{S}}$ and $\mathcal{L}_{\mathcal{T}}$ denote empirical risks over source and target domains, respectively. Here the goal is to find a (empirical) distribution $\mathcal{S}$, such that its discrepancy with target distribution $\mathcal{T}$ is minimized, leading to a nonconvex-nonconcave compositional minimax problem.

Despite the well-established convergence theory on non-compositional minimax, there is few literature focusing on its compositional counterpart. The prior works on compositional minimax [Gao et al., 2021, Liu et al., 2023] only consider the case of function composition on the primal variable, and hence they only establish the convergence rate in the nonconvex-strongly-concave setting. However, a significant portion of ML applications including aforementioned applications do not satisfy this assumption. The function composition can also be on the dual variable, or on both primal and dual variables, which makes the problem convex-nonconcave or even nonconvex-nonconcave. To fill this gap, we propose a simple yet effective algorithm dubbed as stochastically Corrected stOchastic gradient Descent Ascent (CODA). The proposed algorithm is a variant of the celebrated Stochastic Gradient Descent-Ascent (SGDA) for minimax optimization, with a compositional correction step inspired from [Chen et al., 2021a]. Depending on the objective structure, the correction step is either applied on the primal update, dual update, or both updates, which yields three algorithm instances: CODA-Primal, CODA-Dual, and CODA-PD (Primal-Dual), respectively. We establish their convergence in nonconvex-(strongly)-concave, (strongly)-convex-nonconcave and weakly-convex-weakly-concave settings correspondingly. Our results demonstrate that intricate compositional structure inherent to these minimax problems does not increase their asymptotic computational complexity compared to non-compositional versions, aligning with findings in compositional minimization [Chen et al., 2021a].

Contributions The main contributions of the present work are summarized as follows:

- We introduce a general algorithmic framework, dubbed CODA, for stochastic compositional minimax optimization that encompasses three fundamental scenarios of objective composition: composition on the primal variable (Section 2), composition on the dual variable (Section 3), and composition on both variables (Section 4). We provide a rigorous convergence analysis for various CODA variants tailored to these settings.

Objective type	Assumption	Algorithm	Complexity
Composition on $\mathbf{x}$	Nonconvex-strongly-concave	SCGDA [Gao et al., 2021]	$O(\kappa^4\epsilon^{-4})$
		CODA-Primal (Theorem 1)	$O(\kappa^4\epsilon^{-4})$
		†NSTORM [Liu et al., 2024]	$O(\kappa^3\epsilon^{-3})$
		†CODA-Primal+ (Theorem 6)	$O(\kappa^2\epsilon^{-3})$
	Nonconvex-concave	CODA-Primal (Theorem 2)	$O(\epsilon^{-8})$
		CODA-Primal+ (Theorem 7)	$O(\epsilon^{-7})$
Composition on $\mathbf{y}$	Strongly-convex-nonconcave	CODA-Dual (Theorem 3)	$O(\kappa^5\epsilon^{-4})$
	Convex-nonconcave	CODA-Dual (Theorem 4)	$O(\epsilon^{-8})$
Composition on $\mathbf{x}$ and $\mathbf{y}$	Weakly-convex-weakly-concave	CODA-PD (Theorem 5)	$O(\epsilon^{-4})$

Table 1: A summary of prior works and our results on compositional minimax optimization, in different objective settings. †: These algorithms use variance reduction techniques.

- In composition on the primal variable setting, we examine both nonconvex-strongly-concave and nonconvex-concave cases. Remarkably, our analysis demonstrates that the CODA-Primal algorithm exhibits the same convergence rate as the widely acclaimed SGDA algorithm in their respective non-compositional counterparts.
- In composition on the dual variable setting, we show that CODA-Dual algorithm can provably converge in strongly-convex-nonconcave and convex-nonconcave cases. This is also the first work that establishes the convergence guarantee on stochastic (compositional) minimax optimization on the general convex-nonconcave objective.
- In composition on the both variable setting, we study weakly-convex-weakly-concave cases and also provide the convergence guarantee of our CODA-PD algorithm. This work introduces the first compositional minimax algorithm with guaranteed convergence for weakly-convex-weakly-concave objectives.
- We further propose a variance reduced variant, CODA+, which achieves the best-known convergence rates on nonconvex-strongly-concave and nonconvex-concave compositional minimax problems (Section 5).
- Lastly, we conduct experiments across various applications that corroborate the theoretical analysis and demonstrate the effectiveness of the proposed algorithms (Section 6).

2 Composition on Primal Variable

We begin by considering the composition in the primal variable as stated in the following objective:

$$F(\mathbf{x},\mathbf{y}) := h(\mathbf{x}) + \mathbb{E}_{\zeta\sim\mathcal{S}_f}[f(\mathbb{E}_{\xi\sim\mathcal{S}_g}[g(\mathbf{x});\xi],\mathbf{y};\zeta)] - r(\mathbf{y}).$$

Due to the compositional structure, the objective will become nonconvex in $\mathbf{x}$, and hence we consider two settings: nonconvex-strongly-concave and nonconvex-concave. Our algorithm builds on top of classic SGDA algorithm for minimax optimization. However, due to the compositional structure, we cannot have an unbiased estimation of the gradient with respect to $\mathbf{x}$. To see this, notice the fact: $\mathbb{E}_{\xi,\zeta}[\nabla_1 f(g(\mathbf{x};\xi),\mathbf{y};\zeta)\nabla g(\mathbf{x};\xi)] \neq \nabla_1 f(g(\mathbf{x}),\mathbf{y})\nabla g(\mathbf{x})$. Hence, we borrow the technique from compositional minimization problem [Chen et al., 2021a] and maintain an auxiliary variable $\mathbf{z}$ to approximate inner function $g(\mathbf{x})$. Given a mini-batch $\mathcal{M}^t$ with size M, we update $\mathbf{z}$ as:

$$\mathbf{z}^{t+1} = (1-\beta)\left(\mathbf{z}^t + g(\mathbf{x}^t;\mathcal{M}^t) - g(\mathbf{x}^{t-1};\mathcal{M}^t)\right) + \beta g(\mathbf{x}^t;\mathcal{M}^t),$$

where $g(\mathbf{x};\mathcal{M}^t) := \frac{1}{M}\sum_{\xi\in\mathcal{M}^t} g(\mathbf{x};\xi)$. Then, the proposed CODA-Primal algorithm computes the following primal-dual gradients using a mini-batch $\mathcal{B}^t$ with size B:

$$\mathbf{g}_{\mathbf{x}}^t = \frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t} \nabla_1 f(\mathbf{z}^{t+1},\mathbf{y}^t;\zeta)\nabla g(\mathbf{x}^t;\xi) + \nabla h(\mathbf{x}^t),$$

$$\mathbf{g}_{\mathbf{y}}^t = \frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t} \nabla_2 f(\mathbf{z}^{t+1},\mathbf{y}^t;\zeta) - \nabla r(\mathbf{y}^t),$$

and then performs primal-dual updates on $\mathbf{x}$ and $\mathbf{y}$: $\mathbf{x}^{t+1} = \mathbf{x}^t - \eta_{\mathbf{x}}\mathbf{g}_{\mathbf{x}}^t$ and $\mathbf{y}^{t+1} = \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y}^t + \eta_y \mathbf{g}_{\mathbf{y}}^t\right)$ where $\mathcal{P}_{\mathcal{Y}}(\cdot)$ is the projection operator onto a convex set $\mathcal{Y}$.

■ **Nonconvex-Strongly-Concave Setting.** We start by providing the convergence rate of CODA-Primal (Algorithm 1), in the nonconvex-strongly-concave setting. From an algorithmic standpoint, the proposed method is reminiscent of existing methods, but the convergence analysis is more involved due to the minimax, compositional structure, and nonconvexity of the objective. We make the following assumptions on the objective:

Assumption 1. *$F(\mathbf{x},\mathbf{y})$ is μ-strongly-concave in $\mathbf{y}$, $\forall \mathbf{x}\in\mathbb{R}^d$.*

Assumption 2. *Domain $\mathcal{Y}$ has bounded diameter, i.e., $\|\mathbf{y}-\mathbf{y}'\| \le D_{\mathcal{Y}}$ $\forall \mathbf{y},\mathbf{y}'\in\mathcal{Y}$.*

Algorithm 1: CODA-Primal

Input: $(\mathbf{x}^0, \mathbf{y}^0), \mathbf{z}^0$ such that $\mathbb{E}\left\|\mathbf{z}^0 - g(\mathbf{x}^0)\right\|^2 \le \delta$.
for $t = 0, \ldots, T$ **do**
  Sample a mini-batch $\mathcal{M}^t$ of M data from $\mathcal{S}_g$.
  $\mathbf{z}^{t+1} = (1-\beta)\left(\mathbf{z}^t + \frac{1}{M}\sum_{\xi^t\in\mathcal{M}^t}\left(g(\mathbf{x}^t;\xi^t) - g(\mathbf{x}^{t-1};\xi^t)\right)\right) + \beta\frac{1}{M}\sum_{\xi^t\in\mathcal{M}^t} g(\mathbf{x}^t;\xi^t)$.
  Sample a mini-batch $\mathcal{B}^t$ of B data pair (ζ^t, ξ^t) from $\mathcal{S}_g \times \mathcal{S}_f$.
  Compute
  $\mathbf{g}^t_{\mathbf{x}} = \frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t} \nabla_1 f(\mathbf{z}^{t+1}, \mathbf{y}^t; \zeta)\nabla g(\mathbf{x}^t;\xi) + \nabla h(\mathbf{x}^t)$,
  $\mathbf{g}^t_{\mathbf{y}} = \frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t} \nabla_2 f(\mathbf{z}^{t+1}, \mathbf{y}^t; \zeta) - \nabla r(\mathbf{y}^t)$.
  $\mathbf{x}^{t+1} = \mathbf{x}^t - \eta_{\mathbf{x}}\mathbf{g}^t_{\mathbf{x}}$, $\mathbf{y}^{t+1} = \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y}^t + \eta_{\mathbf{y}}\mathbf{g}^t_{\mathbf{y}}\right)$.
end
Output: $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ uniformly sampled from $\{(\mathbf{x}^t, \mathbf{y}^t)\}_{t=1}^T$.

Assumption 3. *We assume:*

1. $f(\mathbf{z}, \mathbf{y})$ *is* L_f *smooth,* $g(\mathbf{x})$ *is* L_g *smooth , and* $F(\mathbf{x}, \mathbf{y})$ *is* L *smooth.*
2. $\mathbb{E}[\|\nabla g(\mathbf{x};\xi)\|^2] \le G_g^2$ *and* $\mathbb{E}[\|\nabla f(\mathbf{z}, \mathbf{y}; \zeta)\|^2] \le G_f^2$.
3. $\mathbb{E}\|g(\mathbf{x};\xi) - g(\mathbf{x})\|^2 \le \sigma^2$,
$\mathbb{E}\|\nabla g(\mathbf{x};\xi) - \nabla g(\mathbf{x})\|^2 \le \sigma^2$,
$\mathbb{E}\|\nabla f(\mathbf{z}, \mathbf{y}; \zeta) - \nabla f(\mathbf{z}, \mathbf{y})\|^2 \le \sigma^2$.

Since the objective is nonconvex, we are unable to show the convergence to the global saddle point. Thus, following the standard machinery in nonconvex-concave analysis [Lin et al., 2019, Thekumparampil et al., 2019, Rafique et al., 2018], we introduce the following primal function to measure the convergence.

Definition 1. *We define primal function* $\Phi(\mathbf{x}) = \max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{x}, \mathbf{y})$ *to facilitate our analysis.*

We consider the convergence rate to the first order stationary point of $\Phi(\mathbf{x})$, as advocated in seminal nonconvex-concave minimax literature [Lin et al., 2019, Rafique et al., 2018, Thekumparampil et al., 2019].

Theorem 1. *Under Assumptions 1, 2, 3, defining* $\kappa := L/\mu$, *for Algorithm 1, if we choose* $\delta = \frac{1}{\kappa}$, $M = B = \Theta\left(\max\left\{\frac{\kappa^2 L\sigma^2}{\epsilon^2}, 1\right\}\right)$, $\beta = \frac{1}{2}$, $\eta_{\mathbf{x}} = \Theta\left(\frac{1}{\kappa^2 L}\right)$, $\eta_{\mathbf{y}} = \Theta\left(\frac{1}{L}\right)$, *then Algorithm 1 guarantees that* $\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla\Phi(\mathbf{x}^t)\|^2 \le \epsilon^2$ *with the gradient complexity bounded by:*

$$O\left(\frac{\kappa^2(\Delta_\Phi + L^2 D_{\mathcal{Y}})}{\epsilon^2}\max\left\{\frac{\kappa^2 L\sigma^2}{\epsilon^2}, 1\right\}\right),$$

where $\Delta_\Phi := \Phi(\mathbf{x}^0) - \min_{\mathbf{x}\in\mathcal{X}}\Phi(\mathbf{x})$.

The proof of Theorem 1 is deferred to Appendix C.1. Here we achieve an $O(\frac{\kappa^4}{\epsilon^4})$ rate, which is the same as the rate of SGDA on non-compositional objective [Lin et al., 2019]. This suggests that we can solve nonconvex-strongly-concave compositional minimax problems as readily as non-compositional cases. The most relevant work is [Gao et al., 2021], where they also achieve an $O(\frac{\kappa^4}{\epsilon^4})$ rate. An even faster rate is attainable through the use of variance reduction techniques, as will be shown and discussed in Section 5. As a final remark, we need to choose a good initialization of $\mathbf{z}^0$ such that it is close to $g(\mathbf{x}^0)$, i.e., $\mathbb{E}\left\|\mathbf{z}^0 - g(\mathbf{x}^0)\right\|^2 \le \frac{1}{\kappa}$. This will take $\kappa\sigma^2$ samples drawn from $\mathcal{S}_g$.

■ **Nonconvex-Concave Setting.** We now turn to the analysis in nonconvex-concave setting.

Assumption 4. $F(\mathbf{x}, \mathbf{y})$ *is concave in* $\mathbf{y}$, $\forall\mathbf{x}\in\mathbb{R}^d$. $h(\mathbf{x})$ *is* G_h *Lipschitz.*

Since F is merely concave in $\mathbf{y}$, primal function $\Phi(\mathbf{x})$ is not smooth anymore, and its gradient is not a viable quantity to measure the convergence. Following the convention [Lin et al., 2019], we consider the gradient of Moreau Envelope of primal function as the convergence measure.

Definition 2 (Moreau Envelope)**.** *A function* $\Phi_\rho(\mathbf{w})$ *is the* ρ*-Moreau envelope of a function* Φ *if* $\Phi_\rho(\mathbf{w}) := \min_{\mathbf{w}'\in\mathcal{W}}\{\Phi(\mathbf{w}') + \frac{1}{2\rho}\|\mathbf{w}' - \mathbf{w}\|^2\}$.

Theorem 2. *Under Assumptions 2, 3, and 4, for Algorithm 1, if we choose* $\delta = O(1)$, $B = \Theta(1)$, $M = \Theta\left(\frac{\sigma^2 L^4 D_{\mathcal{Y}}^2}{\epsilon^4}\right)$, $\beta = \frac{1}{2}$, $\eta_{\mathbf{y}} = \Theta\left(\frac{\epsilon^2}{L\sigma^2}\right)$, $\eta_{\mathbf{x}} = \Theta\left(\frac{\epsilon^6}{LD_{\mathcal{Y}}\sigma^2(G_h + G_f G_g)}\right)$, *then Algorithm 1 guarantees that* $\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left\|\nabla\Phi_{1/2L}(\mathbf{x}^t)\right\|^2 \le \epsilon^2$ *with gradient complexity bounded by*

$$O\left(\max\left\{\frac{L^3\sigma^2 D_{\mathcal{Y}}^2(G_h + G_f G_g)\Delta_{\hat{\Phi}}}{\epsilon^8}, \frac{L\Delta_\Phi}{\epsilon^2}\right\}\right),$$

where $\Delta_\Phi := \Phi(\mathbf{x}^0) - \min_{\mathbf{x}\in\mathcal{X}}\Phi(\mathbf{x})$, $\Delta_{\hat{\Phi}} := \Phi_{1/2L}(\mathbf{x}^0) - \min_{\mathbf{x}\in\mathcal{X}}\Phi_{1/2L}(\mathbf{x})$.

The proof of Theorem 2 is deferred to Appendix C.2. Unlike the proof technique used in compositional minimization or minimax optimization [Chen et al., 2021a, Gao et al., 2021] that only considers potential function with second moment of inner function estimation error $\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\right\|^2$, in nonconvex-concave minimax setting we rely on the potential function with both first and second moment of : $\Psi^t = \Phi_{1/2L}(\mathbf{x}^t) + O(\frac{\eta_{\mathbf{x}}}{(2\beta-\beta^2)}G_g^2 L_f^2)\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\right\|^2 + O(\frac{\eta_{\mathbf{x}} L^2 D_{\mathcal{Y}}}{\beta})\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\right\|$. We recover $O(\epsilon^{-8})$ gradient complexity, the same as SGDA on the non-compositional nonconvex-concave objective. Notice

Algorithm 2: CODA-Dual

Input: $(\mathbf{x}^0, \mathbf{y}^0), \mathbf{z}^0$ such that $\mathbb{E}\left\|\mathbf{z}^0 - g(\mathbf{x}^0)\right\|^2 \le \delta$, and $\{\alpha_t\}_{t=0}^{T-1}$.

for $t = 0, ..., T-1$ **do**

  Sample a mini-batch $\mathcal{M}^t$ of M data from $\mathcal{S}_g$.

  $\mathbf{z}^{t+1} = (1-\beta)\left(\mathbf{z}^t + \frac{1}{M}\sum_{\xi\in\mathcal{M}^t}\left(g(\mathbf{y}^t;\xi) - g(\mathbf{y}^{t-1};\xi)\right)\right) + \beta\frac{1}{M}\sum_{\xi\in\mathcal{M}^t} g(\mathbf{y}^t;\xi)$.

  Sample a mini-batch $\mathcal{B}^t$ of B data pair (ζ, ξ) from $\mathcal{S}_f \times \mathcal{S}_g$,

  Compute $\mathbf{g}_{\mathbf{x}}^t = \frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t}\nabla_1 f(\mathbf{x}^t, \mathbf{z}^{t+1};\zeta) + \nabla h(\mathbf{x}^t) + \alpha_t\mathbf{x}^t$,

  $\mathbf{g}_{\mathbf{y}}^t = \frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t}\nabla_2 f(\mathbf{x}^t, \mathbf{z}^{t+1};\zeta)\nabla g(\mathbf{y}^t;\xi) - \nabla r(\mathbf{y}^t)$.

  $\mathbf{x}^{t+1} = \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}^t - \eta_{\mathbf{x}}\mathbf{g}_{\mathbf{x}}^t\right)$,

  $\mathbf{y}^{t+1} = \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y}^t + \eta_{\mathbf{y}}\mathbf{g}_{\mathbf{y}}^t\right)$.

end

Output: $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ uniformly sampled from $\{(\mathbf{x}^t, \mathbf{y}^t)\}_{t=1}^T$.

that we also need $T \cdot M = O(\epsilon^{-12})$ zeroth order evaluations for maintaining $\mathbf{z}$ variable, but zeroth order evaluation (i.e., inference) is usually considered much cheaper than gradient evaluation, and can be implemented efficiently in practice. This is also the first convergence result for nonconvex-concave compositional minimax optimization.

3 Composition on Dual Variable

In this section, we consider the setting where the composition happens on the dual variable:

$$F(\mathbf{x}, \mathbf{y}) := h(\mathbf{x}) + \mathbb{E}_{\zeta\sim\mathcal{S}_f}[f(\mathbf{x}, \mathbb{E}_{\xi\sim\mathcal{S}_g}[g(\mathbf{y});\xi];\zeta)] - r(\mathbf{y}).$$

In this setting the objective is possibly nonconcave in $\mathbf{y}$. We hence consider strongly-convex-nonconcave and convex-nonconcave settings to establish the convergence rate. The algorithm employed is CODA-Dual, which is akin to CODA-Primal, but instead utilizes a compositional correction step on the dual variable as shown in Algorithm 2. The little difference is that, we have a $\alpha_t\mathbf{x}^t$ term which serves as a regularizer. When the objective is strongly convex in $\mathbf{x}$, we will set α_t to be 0. When it is merely convex in $\mathbf{x}$, we will set α_t to be a small value to make the objective strongly convex. This technique was used in vanilla convex-nonconcave minimax problem [Xu et al., 2023].

■ **Strongly-convex-Nonconcave Setting.** We first present the convergence result of CODA-Dual in the strongly-convex-nonconcave setting by making the following standard assumptions.

Assumption 5. *$F(\mathbf{x}, \mathbf{y})$ is μ-strongly-convex in $\mathbf{x}$, $\forall\mathbf{y}\in\mathcal{Y}$ and L smooth.*

Assumption 6. *We make the following assumptions:*

1. *$f(\mathbf{x}, \mathbf{z})$ is L_f smooth, $g(\mathbf{y})$ is L_g smooth, and $F(\mathbf{x}, \mathbf{y})$ is L smooth.*
2. *$\mathbb{E}[\|\nabla g(\mathbf{y};\xi)\|^2] \le G_g^2$ and $\mathbb{E}[\|\nabla f(\mathbf{x}, \mathbf{z};\zeta)\|^2] \le G_f^2$.*
3. *$\mathbb{E}\|g(\mathbf{y};\xi) - g(\mathbf{y})\|^2 \le \sigma^2$, $\mathbb{E}\|\nabla g(\mathbf{y};\xi) - \nabla g(\mathbf{y})\|^2 \le \sigma^2$, $\mathbb{E}\|\nabla f(\mathbf{x}, \mathbf{z};\zeta) - \nabla f(\mathbf{x}, \mathbf{z})\|^2 \le \sigma^2$.*

Assumption 7. *Domain $\mathcal{X}$ and $\mathcal{Y}$ have bounded diameter, i.e., $\|\mathbf{x} - \mathbf{x}'\| \le D_{\mathcal{X}}$ and $\|\mathbf{y} - \mathbf{y}'\| \le D_{\mathcal{Y}}$, $\forall\mathbf{x}, \mathbf{x}' \in \mathcal{X}, \mathbf{y}, \mathbf{y}' \in \mathcal{Y}$. We use D to denote $D = \max\{D_{\mathcal{X}}, D_{\mathcal{Y}}\}$.*

Assumption 8. *There exists a constant $F_{\max}$ such that $\max_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} F(\mathbf{x}, \mathbf{y}) \le F_{\max}$.*

Next, we consider the following convergence measure:

Definition 3 (Convergence Measure [Xu et al., 2023])**.** *Given two parameters, $\mathbf{x}$ and $\mathbf{y}$, we define the following quantity as a stationary gap $\nabla G(\mathbf{x}, \mathbf{y}) = \begin{pmatrix}\frac{1}{\eta_{\mathbf{x}}}\left(\mathbf{x} - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x} - \eta_{\mathbf{x}}\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y})\right)\right) \\ \frac{1}{\eta_{\mathbf{y}}}\left(\mathbf{y} - \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y} + \eta_{\mathbf{y}}\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y})\right)\right)\end{pmatrix}$.*

Given the nonconcave nature of $F(\mathbf{x}, \mathbf{y})$, we are only able to show the convergence to a stationary point. Definition 3 measures the stationarity given parameter pair $(\mathbf{x}, \mathbf{y})$ by examining how much the parameter will change if we run one step projected gradient descent-ascent on them. The widely employed convergence measure *primal function* $\|\nabla\Phi(\mathbf{x})\|$ does not apply here due to non-concavity.

Theorem 3. *Under Assumptions 5, 6, 7 and 8, if we choose $\beta = 0.1$, $\delta = O(1)$, $M = \Theta\left(\max\left\{\frac{\kappa^3 L\sigma^2}{\epsilon^2}1\right\}\right)$, $B = \Theta\left(\max\left\{\frac{\kappa^2 L_f^2\sigma^2}{\epsilon^2}, 1\right\}\right)$, $\eta_{\mathbf{x}} = \Theta\left(\min\left\{\frac{1}{L^2}, \frac{\mu}{L_f^2}\right\}\right)$, $\eta_{\mathbf{y}} = \Theta\left(\frac{\eta_{\mathbf{x}}}{\kappa^2}\right)$, and $\alpha_t = 0$, then Algorithm 2 guarantees that $\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla G(\mathbf{x}^t, \mathbf{y}^t)\|^2 \le \epsilon^2$ with the gradient complexity bounded by:*

$$O\left(\max\left\{\frac{\kappa^2 L_f^2\sigma^2}{\epsilon^2}, 1\right\} \cdot \frac{\kappa^3 F_{\max}}{\epsilon^2}\right).$$

The proof of Theorem 3 is deferred to Appendix D.1. The heart of our proof is constructing a two-level potential function, together with controlling the iterate difference of auxiliary variable $\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2$, so that we can derive the proper descent inequality. Here we achieve $O(\epsilon^{-4})$ rate, worse than *deterministic* non-compositional setting [Xu et al., 2023] by $O(\epsilon^{-2})$. The most relevant work is [Deng et al., 2023], where they also solve a strongly-convex-nonconcave compositional minimax problem. The main difference is that their objective has a linear coupling of primal and dual variables, while we consider a more general objective.

Reducing SC-NC to NC-SC Problem. One may argue that finding stationary point of a strongly-convex-nonconcave (SC-NC) minimax problem can be reduced to finding that of a nonconvex-strongly-concave (NC-SC) problem. However, in the SC-NC problem, we must consider the gradient norm of $F(\mathbf{x}, \mathbf{y})$ as the convergence measure, while all existing NC-SC compositional minimax algorithm only has convergence guarantee in terms of gradient norm of primal function $\Phi(\mathbf{x})$. It is well-known that ([Lin et al., 2020a], Proposition 4.11), translating an ϵ stationary point in term of $\Phi(\mathbf{x})$ to ϵ stationary point in term of F will need $O(\epsilon^{-2})$ more stochastic gradient evaluations, hence if we simply use the result from Theorem 1, we will need another sub-rountine to translate the stationary point, which is less preferred than our single loop algorithm CODA-Dual.

■ **Convex-Nonconcave Setting.** In this section, we present the convergence result of CODA-Dual when the objective is merely convex in $\mathbf{x}$.

Assumption 9. *$F(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$, $\forall \mathbf{y} \in \mathcal{Y}$.*

Theorem 4. *Under Assumptions 6, 7, 8 and 9, if we choose $\beta = 0.1$, $\delta = O(1)$, $\eta_{\mathbf{x}} = \Theta\left(\frac{1}{L^2}\right)$, $\eta_{\mathbf{y}} = \Theta\left(\frac{\epsilon^2}{L^4 D_{\mathcal{X}}}\right)$, $M = \Theta\left(\max\left\{\frac{L^6 L_f^2 D_{\mathcal{X}}^4 \sigma^2}{\epsilon^4}, 1\right\}\right)$, $B = \Theta\left(\max\left\{\frac{L^3 L_f^2 D_{\mathcal{X}}^2 \sigma^2}{\epsilon^4}, 1\right\}\right)$, and $\alpha_t = \Theta\left(\frac{\epsilon}{L D_{\mathcal{X}}}\right)$, then for Algorithm 2 it is guaranteed that $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left\|\nabla G(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 \leq \epsilon^2$ holds with the gradient complexity of*

$$O\left(\max\left\{\frac{L^3 L_f^2 D_{\mathcal{X}}^2 \sigma^2}{\epsilon^4}, 1\right\} \cdot \frac{F_{\max} D_{\mathcal{X}}^2 L^4}{\epsilon^4}\right).$$

The proof of Theorem 4 is deferred to Appendix D.3. Here we achieve an $O(\epsilon^{-8})$ convergence rate to the stationary point of F in the convex-nonconcave setting. This is also the first convergence rate for stochastic convex-nonconcave minimax optimization.

4 Composition on Primal and Dual Variables

In this section, we consider composition on both variables

$$F(\mathbf{x}, \mathbf{y}) := h(\mathbf{x}) + \mathbb{E}_{\zeta\sim\mathcal{S}_f}\left[f(\mathbb{E}_{\xi\sim\mathcal{S}_g}[g(\mathbf{x}, \mathbf{y}; \xi)]; \zeta)\right] - r(\mathbf{y}).$$

In this setting, we may lose convexity and concavity on primal and dual variables, so the problem becomes nonconvex-nonconcave minimax optimization. In general, if we make no assumptions (except smoothness), the first-order stationarity may not be a viable quantity to control convergence, and a more sophisticated optimality measure is necessary, e.g., greedy adversarial equilibrium [Mangoubi and Vishnoi, 2021]. In this paper, following a branch of nonconvex-nonconcave optimization literature [Liu et al., 2021, Diakonikolas et al., 2021], we make a mild assumption on our nonconvex-nonconcave objective $F(\mathbf{x}, \mathbf{y})$, that there exists a point that is the solution of Minty Variational Inequality induced by $F(\mathbf{x}, \mathbf{y})$.

Assumption 10 (Existence of Solution for Minty Variational Inequality). *There exists a $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{X} \times \mathcal{Y}$, such that the following Minty variational inequality induced by ∇F can hold:*

$$\langle \nabla F(\mathbf{x}, \mathbf{y}),\, (\mathbf{x}, \mathbf{y}) - (\mathbf{x}^*, \mathbf{y}^*)\rangle \geq 0, \forall (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}.$$

Algorithm 3: CODA-SCSC$(F, \mathbf{x}^0, \mathbf{y}^0, T)$

Input: $(\mathbf{x}^0, \mathbf{y}^0), \mathbf{z}^0$ such that $\mathbb{E}\left\|\mathbf{z}^0 - g(\mathbf{x}^0)\right\|^2 \leq \delta$.
for $t = 0, ..., T-1$ **do**
  Sample a mini-batch $\mathcal{M}^t$ of M data from $\mathcal{S}_g$.
  $\mathbf{z}^{t+1} = (1-\beta)\left(\mathbf{z}^t + \frac{1}{M}\sum_{\xi\in\mathcal{M}^t} g(\mathbf{x}^t, \mathbf{y}^t; \xi) - g(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}; \xi)\right) + \beta\frac{1}{M}\sum_{\xi\in\mathcal{M}^t} g(\mathbf{x}^t, \mathbf{y}^t; \xi)$.
  Sample a mini-batch $\mathcal{B}^t$ of B data pair (ζ^t, ξ^t) from $\mathcal{S}_f \times \mathcal{S}_g$.
  $\mathbf{g}_{\mathbf{x}}^t = \frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t} \nabla f(\mathbf{z}^{t+1}; \zeta)\nabla_{\mathbf{x}} g(\mathbf{x}^t, \mathbf{y}^t; \xi) + \nabla h(\mathbf{x}^t) + \frac{2}{\gamma}\left(\mathbf{x}^t - \mathbf{x}^0\right)$,
  $\mathbf{g}_{\mathbf{y}}^t = \frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t} \nabla f(\mathbf{z}^{t+1}; \zeta)\nabla_{\mathbf{y}} g(\mathbf{x}^t, \mathbf{y}^t; \xi) - \nabla r(\mathbf{y}^t) - \frac{2}{\gamma}\left(\mathbf{y}^t - \mathbf{y}^0\right)$.
  $\mathbf{x}^{t+1} = \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}^t - \eta_{\mathbf{x}}\mathbf{g}_{\mathbf{x}}^t\right)$,
  $\mathbf{y}^{t+1} = \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y}^t + \eta_{\mathbf{y}}\mathbf{g}_{\mathbf{y}}^t\right)$.
end
Output: $(\mathbf{x}^T, \mathbf{y}^T)$.

Algorithm 4: CODA-PD

Input: $(\mathbf{x}_0, \mathbf{y}_0)$, γ, weights $\left\{\theta^k\right\}_{k=0}^{K-1}$.
for $k = 0, \ldots, K-1$ **do**
  Construct
  $F_k = F(\mathbf{x}, \mathbf{y}) + \frac{1}{2\gamma}\left(\|\mathbf{x} - \mathbf{x}_k\|^2 - \|\mathbf{y} - \mathbf{y}_k\|^2\right)$
  $(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) = \texttt{CODA-SCSC}(F_k, \mathbf{x}_k, \mathbf{y}_k, T_k)$
end
Output: $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ sampled from $\left\{(\mathbf{x}_k, \mathbf{y}_k)\right\}_{k=0}^{K-1}$ with $\Pr\left((\hat{\mathbf{x}}, \hat{\mathbf{y}}) = (\mathbf{x}_k, \mathbf{y}_k)\right) = \frac{\theta^k}{\sum_{k=0}^{K-1}\theta^k}$.

■ **Weakly-convex-weakly-concave Setting.** Relying on reduction introduced in [Liu et al., 2021], we can break solving a weakly-convex-weakly-concave minimax problem into solving a bunch of strongly monotone variational inequality problems. The algorithm flow is described in Algorithm 4.

Assumption 11. *We make the following assumptions:*

1. $F(\mathbf{x},\mathbf{y})$ *is is* L *smooth,* ρ*-weakly-convex in* $\mathbf{x}$*, and* ρ*-weakly-concave in* $\mathbf{y}$.
2. $f(\mathbf{z})$ *is* L_f *smooth and* $g(\mathbf{x},\mathbf{y})$ *is* L_g *smooth.*
3. $\mathbb{E}[\|\nabla g(\mathbf{x},\mathbf{y};\xi)\|^2] \le G_g^2$ *and* $\mathbb{E}[\|\nabla f(\mathbf{z};\zeta)\|^2] \le G_f^2$.
4. $\mathbb{E}\|g(\mathbf{x},\mathbf{y};\xi) - g(\mathbf{x},\mathbf{y})\|^2 \le \sigma^2$,
$\mathbb{E}\|\nabla g(\mathbf{x},\mathbf{y};\xi) - \nabla g(\mathbf{x},\mathbf{y})\|^2 \le \sigma^2$,
$\mathbb{E}\|\nabla f(\mathbf{z};\zeta) - \nabla f(\mathbf{z})\|^2 \le \sigma^2$.

Definition 4 (Convergence Measure [Liu et al., 2021])**.** *A solution* $\mathbf{w} = (\mathbf{x},\mathbf{y})$ *is* ϵ *stationary point of problem (1) if there exists* $\bar{\mathbf{w}} = (\bar{\mathbf{x}},\bar{\mathbf{y}})$ *such that* $\|\mathbf{w}-\bar{\mathbf{w}}\| \le c\epsilon, dist\left(0, \partial(F(\bar{\mathbf{x}},\bar{\mathbf{y}}) + \mathbf{1}_{\mathcal{X}\times\mathcal{Y}}(\bar{\mathbf{x}},\bar{\mathbf{y}}))\right) \le \epsilon$ *for some* $c > 0$.

Theorem 5. *Under Assumptions 7, 10 and 11, if we choose* $B = \Theta\left(\max\left\{\frac{\sigma^2}{\epsilon^2}, 1\right\}\right), \beta = 1 - \frac{\mu}{16G_g^2 L_f}, \eta_{\mathbf{x}} = \eta_{\mathbf{y}} = \Theta\left(\frac{1}{L^2}\right)$*,* $\delta = O(1)$*,* $\theta^k = (k+1)^\alpha$ *for some* $\alpha \in (0,1)$*,* $\gamma = \frac{1}{\rho}$*, then Algorithm 4 returns and* ϵ *stationary solution* $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ *with the gradient complexity of:*

$$O\left(\left(D + L^2\right)\frac{\rho^2 D L^4 \sigma^2}{\epsilon^4}\log\left(\frac{1}{\epsilon}\right)\right).$$

The proof is deferred to Appendix E. Here we achieved an $\tilde{O}(\epsilon^{-4})$ rate. The most relevant work is [Liu et al., 2021], demonstrating an $\tilde{O}(\epsilon^{-2})$ rate using (deterministic) gradient descent ascent as inner problem solver.

5 Faster Rates via Variance Reduction

In this section, our goal is to explore the possibility of achieving faster rates. If we assume individual or point-wise smoothness of f and g, i.e., smoothness of $f(g(\mathbf{x}),\mathbf{y};\zeta)$ and $g(\mathbf{x};\xi)$, then the natural idea would be employing variance reduction minimax algorithms, such as VR-SAPD [Zhang et al., 2022], combined with our compositional correction steps. Unfortunately, this integration solely does not lead to an improved rate since the main bottleneck comes from the estimation error of inner function $g(\mathbf{x})$, i.e., $\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\right\|^2$. This observation raises the following question: can we accelerate the convergence of estimation error of inner function to entail a faster convergence rate overall? Inspired by Spider [Fang et al., 2018], we propose to use the following **variance reduced inner function estimation** when maintaining $\mathbf{z}$ variable:

$$\mathbf{z}^{t+1} = (1-\beta)(\mathbf{z}^t + g(\mathbf{x}^t;\mathcal{I}_{\mathbf{z}}^t) - g(\mathbf{x}^{t-1};\mathcal{I}_{\mathbf{z}}^t)) + \beta g^{t-1},$$

where $g^t = g^{t-1} + g(\mathbf{x}^t;\mathcal{I}_{\mathbf{z}}^t) - g(\mathbf{x}^{t-1};\mathcal{I}_{\mathbf{z}}^t)$ and every τ iterations, we use a large mini-batch to update g^t: $g^t = g(\mathbf{x}^t;\mathcal{B}_{\mathbf{z}}^t)$.

It can be shown, the gap $\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\right\|^2$ converges faster than standard compositional optimization as in [Chen et al., 2021a]. Equipped with this technique, we propose our CODA+ algorithm depicted in Algorithm 5. We define the following notations for the

Algorithm 5: CODA-Primal+

Input: $(\mathbf{x}^0,\mathbf{y}^0)$, $\mu_{\mathbf{x}}$.
for $k = 0, ..., K-1$ **do**
  Construct $F_k(\mathbf{x},\mathbf{y}) = F(\mathbf{x},\mathbf{y}) + \frac{\mu_{\mathbf{x}}}{2}\|\mathbf{x}-\mathbf{x}_k\|^2$
  $(\mathbf{x}_{k+1},\mathbf{y}_{k+1}) = \texttt{CODA-SCSC+}(F_k, \mathbf{x}_k, \mathbf{y}_k, T_k)$
end

Algorithm 6: CODA-SCSC+$(F, \mathbf{x}^0, \mathbf{y}^0, T)$

Input: $(\mathbf{x}^0,\mathbf{y}^0), \mathbf{z}^0$ such that $\mathbb{E}\left\|\mathbf{z}^0 - g(\mathbf{x}^0)\right\|^2 \le \delta$.
for $t = 0, ..., T-1$ **do**
  if $t \mod \tau = 0$ **then**
    Sample mini-batch $\mathcal{B}_{\mathbf{z}}^t = \{\xi_1, \ldots, \xi_{B_\tau}\}$, and $\mathcal{B}^t = \{(\zeta_1,\xi_1), \ldots, (\zeta_{B_\tau},\xi_{B_\tau})\}$.
    $g^t = g(\mathbf{x}^t;\mathcal{B}_{\mathbf{z}}^t)$, $\mathbf{z}^{t+1} = (1-\beta)(\mathbf{z}^t + g(\mathbf{x}^t;\mathcal{B}_{\mathbf{z}}^t) - g(\mathbf{x}^{t-1};\mathcal{B}_{\mathbf{z}}^t)) + \beta g^t$,
    $\mathbf{q}^t = G_k^{\mathbf{y}}(\mathbf{x}^t,\mathbf{z}^{t+1},\mathbf{y}^t;\mathcal{B}^t)$,
    $\mathbf{g}_{\mathbf{y}}^t = 2\mathbf{q}^t - \mathbf{q}^{t-1}$,
    $\mathbf{y}^{t+1} = \mathbf{y}^t + \eta_{\mathbf{y}}(\mathbf{g}_{\mathbf{y}}^t - \nabla r(\mathbf{y}^t))$
    $\mathbf{g}_{\mathbf{x}}^t = G_k^{\mathbf{x}}(\mathbf{x}^t,\mathbf{z}^{t+1},\mathbf{y}^{t+1};\mathcal{B}^t)$.
  end
  else
    Sample mini-batch $\mathcal{I}_{\mathbf{z}}^t = \{\xi_1, ..., \xi_B\}$, and $\mathcal{I}^t = \{(\zeta_1,\xi_1), \ldots, (\zeta_B,\xi_B)\}$.
    $g^t = g^{t-1} + g(\mathbf{x}^t;\mathcal{I}_{\mathbf{z}}^t) - g(\mathbf{x}^{t-1};\mathcal{I}_{\mathbf{z}}^t)$, $\mathbf{z}^{t+1} = (1-\beta)(\mathbf{z}^t + g(\mathbf{x}^t;\mathcal{I}_{\mathbf{z}}^t) - g(\mathbf{x}^{t-1};\mathcal{I}_{\mathbf{z}}^t)) + \beta g^{t-1}$,
    $\mathbf{q}^t = \mathbf{q}^{t-1} + G_k^{\mathbf{y}}(\mathbf{x}^t,\mathbf{z}^{t+1},\mathbf{y}^t;\mathcal{I}^t) - G_k^{\mathbf{y}}(\mathbf{x}^{t-1},\mathbf{z}^t,\mathbf{y}^{t-1};\mathcal{I}^t)$,
    $\mathbf{g}_{\mathbf{y}}^t = 2\mathbf{q}^t - \mathbf{q}^{t-1}$,
    $\mathbf{y}^{t+1} = \mathbf{y}^t + \eta_{\mathbf{y}}(\mathbf{g}_{\mathbf{y}}^t - \nabla r(\mathbf{y}^t))$
    $\mathbf{g}_{\mathbf{x}}^t = \mathbf{g}_{\mathbf{x}}^{t-1} + G_k^{\mathbf{x}}(\mathbf{x}^t,\mathbf{z}^{t+1},\mathbf{y}^{t+1};\mathcal{I}^t) - G_k^{\mathbf{x}}(\mathbf{x}^{t-1},\mathbf{z}^t,\mathbf{y}^t;\mathcal{I}^t)$.
  end
  $\mathbf{x}^{t+1} = \mathbf{x}^t - \eta_{\mathbf{x}}(\mathbf{g}_{\mathbf{x}}^t + \nabla h(\mathbf{x}^t))$
end
Output: $(\hat{\mathbf{x}},\hat{\mathbf{y}}) = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}^t,\mathbf{y}^t)$

convenience of presenting:

$$G_k^{\mathbf{x}}(\mathbf{x},\mathbf{z},\mathbf{y};\mathcal{I}) = \frac{1}{|\mathcal{I}|}\sum_{(\zeta,\xi)\in\mathcal{I}} \nabla_1 f(\mathbf{z},\mathbf{y};\zeta)\nabla g(\mathbf{x};\xi) + \mu_{\mathbf{x}}(\mathbf{x}-\mathbf{x}^0),$$
$$G_k^{\mathbf{y}}(\mathbf{x},\mathbf{z},\mathbf{y};\mathcal{I}) = \frac{1}{|\mathcal{I}|}\sum_{(\zeta,\xi)\in\mathcal{I}} \nabla_2 f(\mathbf{z},\mathbf{y};\zeta).$$

Roughly speaking, we use large mini-batch every τ iterations to update gradients and inner function estimation variable, and use small mini-batch for the rest of iterations.

■ **Nonconvex-strongly-concave Setting.**

Assumption 12. *We assume:*

1. *$f(\mathbf{z},\mathbf{y};\zeta)$ is L_f smooth, $g(\mathbf{x};\xi)$ is L_g smooth for any $\zeta \in \mathcal{S}_f, \xi \in \mathcal{S}_g$, and $F(\mathbf{x},\mathbf{y})$ is L smooth.*
2. *$\mathbb{E}[\|\nabla g(\mathbf{x};\xi)\|^2] \le G_g^2$ and $\mathbb{E}[\|\nabla f(\mathbf{z},\mathbf{y};\zeta)\|^2] \le G_f^2$.*
3. *$\mathbb{E}\|g(\mathbf{x};\xi)-g(\mathbf{x})\|^2 \le \sigma^2$,*
   $\mathbb{E}\|\nabla g(\mathbf{x};\xi)-\nabla g(\mathbf{x})\|^2 \le \sigma^2$,
   $\mathbb{E}\|\nabla f(\mathbf{z},\mathbf{y};\zeta)-\nabla f(\mathbf{z},\mathbf{y})\|^2 \le \sigma^2$.

The following theorem establishes the convergence:

Theorem 6. *Under Assumptions 1, 2, 12, for Algorithm 5, if we choose $\delta = O(\frac{\epsilon^2}{L})$, $B_\tau = \Theta\left(\frac{L^2\sigma^2}{\mu\epsilon^2}\right)$, $\tau = \sqrt{\frac{B_\tau}{\kappa}}$, $B = \sqrt{\kappa B_\tau}$, and $\eta_{\mathbf{x}} = \Theta\left(\frac{1}{L}\right)$, $\eta_{\mathbf{y}} = \Theta\left(\frac{1}{L}\right)$, $\mu_{\mathbf{x}} = 2L$, then it guarantees that $\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left\|\nabla\Phi_{1/2L}(\mathbf{x}^t)\right\|^2 \le \epsilon^2$ with the gradient complexity bounded by:*

$$O\left(\frac{\kappa^2 L^{1.5}\Delta_\Phi\sigma}{\epsilon^3}\right).$$

The proof of Theorem 6 is deferred to Appendix F.1. Here we achieve $O\left(\frac{\kappa^2L^{1.5}}{\epsilon^3}\right)$ rate for finding the stationary point of Moreau envelope of primal function. Notice that once we find ϵ stationary point of Moreau envelope, there will be an efficient algorithm that takes only $\tilde{O}(\frac{1}{\epsilon})$ stochastic gradient complexity to find ϵ stationary point of primal function, as indicated in [Zhang et al., 2022]. This rate is faster than previous SOTA [Liu et al., 2024], by $O(\kappa)$ factor. Compared to non-compositional minimax optimization, the best-known rate is $O\left(\frac{\kappa^2L}{\epsilon^3}\right)$, so our result almost recovers the current non-compositional minimax SOTA rate. As a last final, we need to choose a good initialization of $\mathbf{z}^0$ such that $\mathbb{E}\left\|\mathbf{z}^0-g(\mathbf{x}^0)\right\|^2 \le O(\epsilon^2/L)$. This will take $O(L\sigma^2/\epsilon^2)$ samples drawn from $\mathcal{S}_g$.

■ **Nonconvex-concave Setting.** In the nonconvex-concave setting, we consider CODA-Primal+ on the following augmented strongly concave objective with strongly concavity parameter $\mu = \frac{\epsilon^2}{LD_{\mathcal{Y}}^2}$:

$$\tilde{F}(\mathbf{x},\mathbf{y}) := F(\mathbf{x},\mathbf{y}) - \frac{\epsilon^2}{LD_{\mathcal{Y}}^2}\|\mathbf{y}-\mathbf{y}_0\|^2. \tag{2}$$

We now establish the convergence rate of the proposed algorithm, formalized in the following theorem.

Theorem 7. *Under Assumptions 2, 4 and 12, if we run Algorithm 5 on $\tilde{F}(\mathbf{x},\mathbf{y})$ defined in (2), and use the same parameter choice as in Theorem 6 with $\mu = \frac{\epsilon^2}{LD_{\mathcal{Y}}^2}$, then it guarantees that $\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left\|\nabla\Phi_{1/2L}(\mathbf{x}^t)\right\|^2 \le \epsilon^2$ with the gradient complexity bounded by $O\left(\frac{L^{5.5}\Delta_\Phi\sigma}{\epsilon^7}\right)$.*

The proof of Theorem 7 is deferred to Appendix F.2. This result improves the original $O(\frac{1}{\epsilon^8})$ rate in Theorem 2. Faster rate is possible, if we can develop an algorithm which converges at the rate of $O(\frac{\kappa}{\epsilon^4})$ in the strongly concave setting, and choose $\mu = \frac{\epsilon^2}{LD_{\mathcal{Y}}^2}$ which will yield $O(\frac{1}{\epsilon^6})$ rate in the merely concave setting. Such an algorithm can be possibly achieved by modifying SAPD [Zhang et al., 2022], combined with our variance-reduced inner function estimation, which we leave as a future work.

6 Experiment

In this section, we conduct experiments to verify the effectiveness of our proposed algorithms. In primal composition setting, we run CODA-Primal on AUC optimization (nonconvex-strongly-concave) and robust MAML tasks (nonconvex-concave). In dual composition setting, we run CODA-Dual on mixture weight estimation problem (strongly-convex-nonconcave). The additional details and results are provided in Appendix B.

Deep AUC Maximization. To evaluate our CODA-Primal algorithm on nonconvex-strongly-concave objective, we consider deep AUC maximization task [Yuan et al., 2021]. We choose four benchmark image classification datasets: CIFAR-10, CIFAR-100, CATvsDog, and STL-10. We generate imbalanced binary versions of these datasets with imbalanced ratios of 1%, 10%, and 30%, following the positive examples' proportion to total training samples, following the work [Yuan et al., 2021]. We use ResNet20 as our model and optimize for 100 epochs, with batch size of 128. For all benchmark datasets, we conduct experiments over three different random seeds, calculating mean values and standard deviations. Fig. 1 illustrates the convergence and AUC scores of CODA-Primal across these four datasets at different imbalance ratios. We also compare CODA with SCGDA [Gao et al., 2021] on testing AUC performance, and the results are shown in Fig. 2.

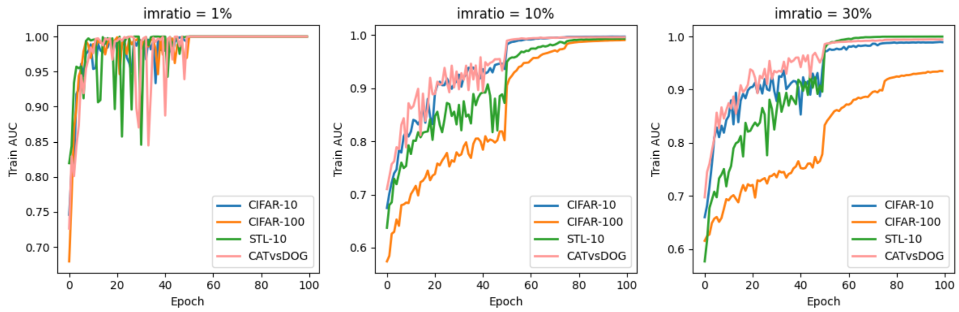

Figure 1: Performance of CODA-Primal on AUC maximization on four datasets with different imbalance ratios

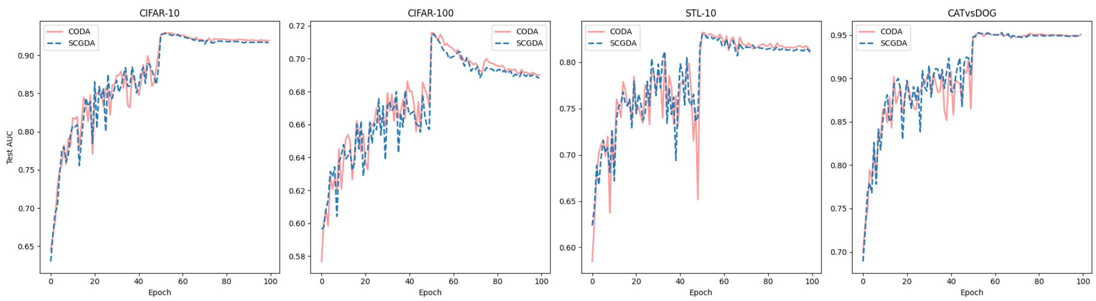

Figure 2: Testing AUC comparison of CODA and SCGDA on four benchmarks under imratio= 10%

(N, K)	**Meta-training Alphabets**	
	Mean(Std. Dev.)	Worst(Std. Dev.)
(5,1)	98.56(0.001)	98.41(0.012)
(20,1)	95.66(0.007)	94.64(0.011)
(N, K)	**Meta-testing Alphabets**	
	Mean(Std. Dev.)	Worst(Std. Dev.)
(5,1)	92.57(0.006)	92.08(0.038)
(20,1)	81.49(0.072)	75.28(0.086)

Table 2: Omniglot N-way, K-shot classification accuracies over three random runs

Task-agnostic Robust MAML. We apply CODA-Primal algorithm to a nonconvex-concave objective. by considering task-agnostic robust MAML [Collins et al., 2020]. We use the Omniglot dataset [Lake et al., 2015], 1623 handwritten characters from 50 alphabets, to learn a task-robust meta-model to enhance performance and adaptability across diverse task distributions. We define tasks as N-way, K-shot classification problems, each involving characters from a single alphabet, and use a 4-layer CNN [Finn et al., 2017]. Our approach involves meta-training on 25 alphabets and meta-testing on 20 different alphabets, following the same data splits in meta-learning evaluations [Triantafillou et al., 2019]. After 60,000 meta-training iterations, we evaluate our model on 5,000 tasks from meta-test alphabets and 5,000 tasks from meta-training alphabets. The detailed accuracies, particularly when $N = 5$ and $N = 20$, are shown in Tab. 2, demonstrating the effectiveness of CODA in nonconvex-concave compositional minimax problem.

Mixture Weights Estimation in Multi-source Domain Adaptation: We address the problem of multi-source domain adaptation via solving a compositional convex-nonconcave minimax problem, as proposed in Deng et al. [2023]. Tab. 3 shows CODA-Dual outperforms in test accuracy, effectively identifying source domains with mixture weight $\boldsymbol{\alpha}$. Results are averaged over three random runs.

Algorithm	**Group 1**	**Group 2**
Target-only	86.97 (0.01)	69.43 (0.02)
Avg(equal)-weight	33.63 (0.04)	23.53 (0.03)
CODA-Dual	**92.90** (0.02)	**78.03** (0.02)
Algorithm	**Group 3**	**Group 4**
Target-only	90.20 (0.01)	65.30 (0.02)
Avg(equal)-weight	27.40 (0.05)	29.63 (0.02)
CODA-Dual	**96.17** (0.005)	**68.50** (0.02)

Table 3: Test accuracy of algorithms across various target domains. Results where an algorithm outperforms others are bolded. Standard deviations are from three random runs.

7 Conclusion

In this paper, we study the compositional minimax problem, with three different settings: composition on primal variable, on dual variable, and on both variables. We give simple yet effective provable algorithms for three settings. Additionally, we give a variance reduction algorithm which achieves SOTA convergence rates. Extensive experimental results across different applications corroborate our theoretical findings and effectiveness of the proposed methods.

Acknowledgement

This work was partially supported by NSF CAREER Award #2239374 and NSF CNS Award #1956276.

References

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021a.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Tighter analysis of alternating stochastic gradient method for stochastic nested problems. *arXiv preprint arXiv:2106.13781*, 2021b.

Shawn Cole, Xavier Giné, and James Vickery. How Does Risk Management Influence Production Decisions? Evidence from a Field Experiment. *The Review of Financial Studies*, 30(6):1935–1970, 03 2017. ISSN 0893-9454. doi: 10.1093/rfs/hhw080. URL `https://doi.org/10.1093/rfs/hhw080`.

Liam Collins, Aryan Mokhtari, and Sanjay Shakkottai. Task-robust model-agnostic meta-learning. *Advances in Neural Information Processing Systems*, 33:18860–18871, 2020.

Yuyang Deng, Ilja Kuzborskij, and Mehrdad Mahdavi. Mixture weight estimation and model prediction in multi-source multi-target domain adaptation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

Hongchang Gao, Xiaoqian Wang, Lei Luo, and Xinghua Shi. On the convergence of stochastic compositional gradient descent ascent method. In *Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm for general convex-concave saddle point problems. *arXiv preprint arXiv:1803.01401*, 2, 2018.

Yuze Han, Guangzeng Xie, and Zhihua Zhang. Lower complexity bounds of finite-sum optimization problems: The results and construction. *ArXiv*, abs/2103.08280, 2021.

Campbell R Harvey, John C Liechty, Merrill W Liechty, and Peter Müller. Portfolio selection with higher moments. *Quantitative Finance*, 10(5):469–485, 2010.

Weiwei Kong and Renato DC Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *arXiv preprint arXiv:1905.13433*, 2019.

Nikola Konstantinov and Christoph Lampert. Robust learning from untrusted sources. In *International conference on machine learning*, pages 3488–3498. PMLR, 2019.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266): 1332–1338, 2015.

Haochuan Li, Yi Tian, Jingzhao Zhang, and A. Jadbabaie. Complexity lower bounds for nonconvex-strongly-concave min-max optimization. *ArXiv*, abs/2104.08708, 2021.

Tianyi Lin, Chi Jin, and Michael I Jordan. On gradient descent ascent for nonconvex-concave minimax problems. *arXiv preprint arXiv:1906.00331*, 2019.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020a.

Tianyi Lin, Chi Jin, Michael Jordan, et al. Near-optimal algorithms for minimax optimization. *arXiv preprint arXiv:2002.02417*, 2020b.

Jin Liu, Xiaokang Pan, Junwen Duan, Hongdong Li, Youqi Li, and Zhe Qu. Breaking the complexity barrier in compositional minimax optimization. *arXiv preprint arXiv:2308.09604*, 2023.

Jin Liu, Xiaokang Pan, Junwen Duan, Hong-Dong Li, Youqi Li, and Zhe Qu. Faster stochastic variance reduction methods for compositional minimax optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13927–13935, 2024.

Mingrui Liu, Hassan Rafique, Qihang Lin, and Tianbao Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems. *The Journal of Machine Learning Research*, 22(1): 7651–7684, 2021.

Pouria Mahdavinia, Yuyang Deng, Haochuan Li, and Mehrdad Mahdavi. Tight analysis of extra-gradient and optimistic gradient methods for nonconvex minimax problems. *Advances in Neural Information Processing Systems*, 35:31213–31225, 2022.

Oren Mangoubi and Nisheeth K Vishnoi. Greedy adversarial equilibrium: an efficient alternative to nonconvex-nonconcave min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 896–909, 2021.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Yishay Mansour, Mehryar Mohri, Jae Ro, Ananda Theertha Suresh, and Ke Wu. A theory of multiple-source adaptation with limited target labeled data. In *International Conference on Artificial Intelligence and Statistics*, pages 2332–2340. PMLR, 2021.

Maher Nouiehed, Maziar Sanjabi, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *arXiv preprint arXiv:1902.08297*, 2019.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

Benjamin Recht and Stephen J. Wright. *Optimization for Modern Data Analysis*. 2019. Preprint available at `http://eecs.berkeley.edu/~brecht/opt4mlbook`.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Kiran Koshy Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. *arXiv preprint arXiv:1907.01543*, 2019.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*, 2019.

Rasul Tutunov, Minne Li, Alexander I Cowen-Rivers, Jun Wang, and Haitham Bou-Ammar. Compositional adam: An adaptive compositional solver. *arXiv preprint arXiv:2002.03755*, 2020.

Mengdi Wang, Ethan X Fang, and Han Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161:419–449, 2017a.

Mengdi Wang, Ji Liu, and Ethan X Fang. Accelerating stochastic composition optimization. *Journal of Machine Learning Research*, 18(105):1–23, 2017b.

Zi Xu, Huiling Zhang, Yang Xu, and Guanghui Lan. A unified single-loop alternating gradient projection algorithm for nonconvex–concave and convex–nonconcave minimax problems. *Mathematical Programming*, pages 1–72, 2023.

Yan Yan, Yi Xu, Qihang Lin, Wei Liu, and Tianbao Yang. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. *Advances in Neural Information Processing Systems*, 33:5789–5800, 2020.

Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. *arXiv preprint arXiv:2112.05604*, 2021.

Shuoguang Yang, Zhe Zhang, and Ethan X Fang. Stochastic compositional optimization with compositional constraints. *arXiv preprint arXiv:2209.04086*, 2022.

Zhuoning Yuan, Zhishuai Guo, Nitesh Chawla, and Tianbao Yang. Compositional training for end-to-end deep auc maximization. In *International Conference on Learning Representations*, 2021.

Jiawei Zhang, Peijun Xiao, Ruoyu Sun, and Zhiquan Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in neural information processing systems*, 33:7377–7389, 2020.

Siqi Zhang, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He. The complexity of nonconvex-strongly-concave minimax optimization. *arXiv preprint arXiv:2103.15888*, 2021.

Xinwen Zhang, Yihan Zhang, Tianbao Yang, Richard Souvenir, and Hongchang Gao. Federated compositional deep auc maximization. *arXiv preprint arXiv:2304.10101*, 2023.

Xuan Zhang, Necdet Serhat Aybat, and Mert Gurbuzbalaban. Sapd+: An accelerated stochastic method for nonconvex-concave minimax problems. *Advances in Neural Information Processing Systems*, 35:21668–21681, 2022.

Taoli Zheng, Linglingzhi Zhu, Anthony Man-Cho So, José Blanchet, and Jiajin Li. Universal gradient descent ascent method for nonconvex-nonconcave minimax optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Organization The appendix is organized as follows:

- In Appendix A we review and discuss additional related works.
- In Appendix B we will provide more experimental results and setup details.
- In Appendix C we provide proofs for CODA-Primal.
- In Appendix D we provide proofs for CODA-Dual.
- In Appendix E we provide proofs for CODA-PD.
- Finally, in Appendix F we provide proofs for CODA-Primal+.

A Additional Related Works

In this section, we review additional related works and provide further elaboration on their relevance to our approach, highlighting key contributions and differences that set our work apart.

Compositional Optimization Stochastic compositional optimization has wide applications in risk management and machine learning. Wang et al. [2017a] is the pioneering work in this field where they proposed the first compositional optimization algorithm, SCGD, and achieved $O(\epsilon^{-4})$ rate. An accelerated version of SCGD is then proposed by [Wang et al., 2017b], and achieves faster rate $O(\epsilon^{-2.25})$. Tutunov et al. [2020] also proposed an accelerated compositional optimization algorithm and achieved the same rate as [Wang et al., 2017b]. The SOTA rate is achieved by [Chen et al., 2021a] where their rate matches with the non-compositional setting, demonstrating that solving the compositional minimization problem can be as easy as the non-compositional counterpart.

A line of works also study the compositional minimax problem, particularly in the AUC maximization area. Yuan et al. [2021] proposed to optimize a compositional minimax objective, which improves the AUC training results. In their setting, the objective is nonconvex-strongly-concave, and they achieve $O(\epsilon^{-4})$ gradient complexity. Deng et al. [2023] studied the mixture weight estimation problem and cast it as a strongly-convex-nonconcave compositional minimax problem. They also proposed an SGDA-based compositional algorithm and achieved $O(\epsilon^{-4})$ gradient complexity to find the stationary point. Gao et al. [2021] for the first time study the general compositional minimax problem. They consider nonconvex-strongly-concave setting, and achieve $O(\kappa^4\epsilon^{-4})$ rate. Liu et al. [2023] proposed the first variance reduction algorithm for stochastic compositional minimax problem and achieved $O(\kappa^3\epsilon^{-3})$ rate.

Nonconvex/Nonconcave Minimax optimization Nonconvex minimax optimization has received increasing attention over the past decade. For nonconvex-strongly-concave minimax optimization, Lin et al. [2019] proved the first non-asymptotic convergence of GDA to ϵ-stationary point of primal function, with the gradient complexity of $O(\kappa^2\epsilon^{-2})$. Mahdavinia et al. [2022] proved that OGDA and EG method enjoy the same rate in solving nonconvex-strongly-concave problem. Yang et al. [2021] developed a single loop smoothed AGDA, which achieved gradient complexity of $O(\kappa\epsilon^{-2})$. Lin et al. [2020b] proposed an accelerated gradient method based algorithm, and achieved the best known rate $O(\sqrt{\kappa}\epsilon^{-2})$. Zhang et al. [2021] proposed triple loop algorithms which also achieved gradient complexity of $O(\sqrt{\kappa}\epsilon^{-2})$ using catalyst idea, and inexact proximal point method. These two results match the existing lower bound for deterministic setting [Li et al., 2021, Zhang et al., 2021, Han et al., 2021]. For stochastic setting, Lin et al. [2019] proved that SGDA finds ϵ-stationary point of primal function, with the gradient complexity of $O(\kappa^3\epsilon^{-4})$ respectively. Chen et al. [2021b] proposed a double loop algorithm and using this idea they achieved gradient complexity of $O(\kappa^3\epsilon^{-4})$ with fixed batch size. Yang et al. [2021] introduced stochastic smoothed AGDA with gradient complexity of $O(\kappa^2\epsilon^{-4})$ to find stationary point. The significant improvement over SGDA is that stochastic smoothed AGDA uses fixed batch size. Zhang et al. [2022] proposed SAPD+, which achieve the SOTA rates of $O(\kappa\epsilon^{-4})$ and $O(\kappa^2\epsilon^{-3})$ under average smoothness and point-wise smoothness assumptions respectively.

For nonconvex-concave optimization, Rafique et al. [2018] proposed the pioneering work Proximally Guided Stochastic Mirror Descent Method, which achieves $O(\epsilon^{-6})$ gradient complexity to find the stationary point. Nouiehed et al. [2019] presented a double-loop algorithm with $O(\epsilon^{-7})$ rate. Lin et al. [2020a] gave the first proof of GDA and SGDA on nonconvex-concave functions, with $O(\epsilon^{-6})$ rate and $O(\epsilon^{-8})$ rate respectively. Zhang et al. [2020] proposed smoothed-GDA and also achieve $O(\epsilon^{-4})$ rate. Thekumparampil et al. [2019] proposed Proximal Dual Implicit Accelerated Gradient method and achieved the best known rate $O(\epsilon^{-3})$ for nonconvex-concave problem. Kong and Monteiro [2019] proposed an accelerated inexact proximal point method and also achieve $O(\epsilon^{-3})$ rate. Lin et al. [2020b] designed an algorithm which uses accelerated gradient method as a sub-problem solver, with $O(\epsilon^{-3})$ rate. Zhang et al. [2022] proposed a variance reduction algorithm SAPD+, which achieve the $O(\epsilon^{-6})$ convergence rate in stochastic setting.

For convex-nonconcave optimization, Xu et al. [2023] gave the first convergence proof in this setting, where they consider alternating gradient descent ascent. They achieved $O(\epsilon^{-2})$ and $O(\epsilon^{-4})$ rate for strongly-convex-nonconcave and convex-nonconcave setting respectively. Later on Deng et al. [2023] gave a stochastic algorithm for a special case of compositional strongly-convex-nonconcave minimax problem and achieved $O(\epsilon^{-4})$ rate.

For nonconvex-nonconcave optimization, first-order stationary points may not exist if we do not post any further assumptions. Hence, a line of work [Liu et al., 2021, Diakonikolas et al., 2021] made extra assumptions and developed convergent algorithms. Liu et al Liu et al. [2021] assumed there is a solution of the MVI induced by objective, and gave the first provably convergent algorithm framework for nonconvex-nonconcave minimax problem. Diakonikolas et al. [2021] considered the relaxed assumption that only weak MVI solution exists and proposed an EG-based algorithm with convergence guarantee. Zheng et al. [2023] consider nonconvex-nonconcave minimax problem with one-side KL condition, and proposed Doubly Smoothed GDA which enjoys $O(\epsilon^{-4})$ convergence rate. Some work also explores to design new convergence measures in general nonconvex-nonconcave minimax problems. Mangoubi and Vishnoi [2021] proposed a greedy adversarial equilibrium and developed an algorithm with polynomial gradient and hessian complexity to find it.

B Experiments Details

This section provides additional details and experimental results to demonstrate the effectiveness of our proposed algorithms. In the primal composition setting, we apply CODA-Primal to AUC optimization and robust MAML tasks, addressing nonconvex-strongly-concave and nonconvex-concave compositional minimax problems, respectively. For the dual composition setting, we use CODA-Dual for the mixture weight estimation problem, which involves a strongly-convex-nonconcave compositional minimax formulation.

Deep AUC Maximization: In a binary classification task where the label $y \in \{-1, +1\}$ and the model parameter $\mathbf{w} \in \mathbb{R}^d$, the model prediction on sample $\mathbf{x}$ is $f(\mathbf{w}; \mathbf{x}_i)$. The objective is formulated as a minimax problem:

$$\min_{\mathbf{w},a,b} \max_{\theta \in \Omega} \Phi(\mathbf{w} - \alpha \nabla L(\mathbf{w}), a, b, \theta) = \frac{1}{n} \sum_{i=1}^{n} \phi(\mathbf{w} - \alpha \nabla L(\mathbf{w}), a, b, \theta; \mathbf{x}_i, y_i)$$

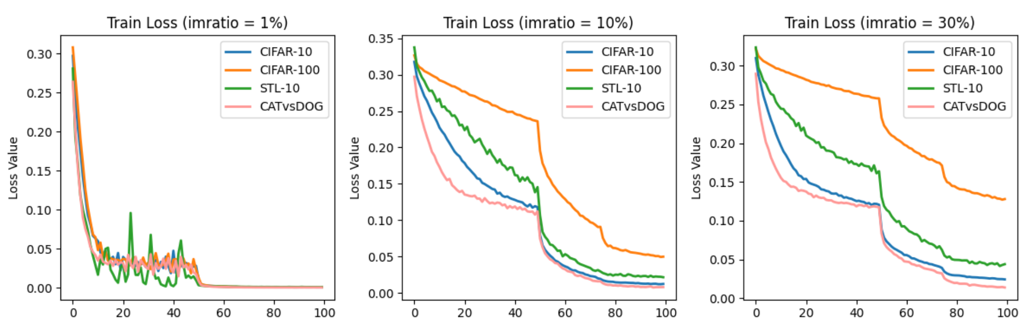

Figure 3: Training loss with different imbalance ratios on four benchmark datasets

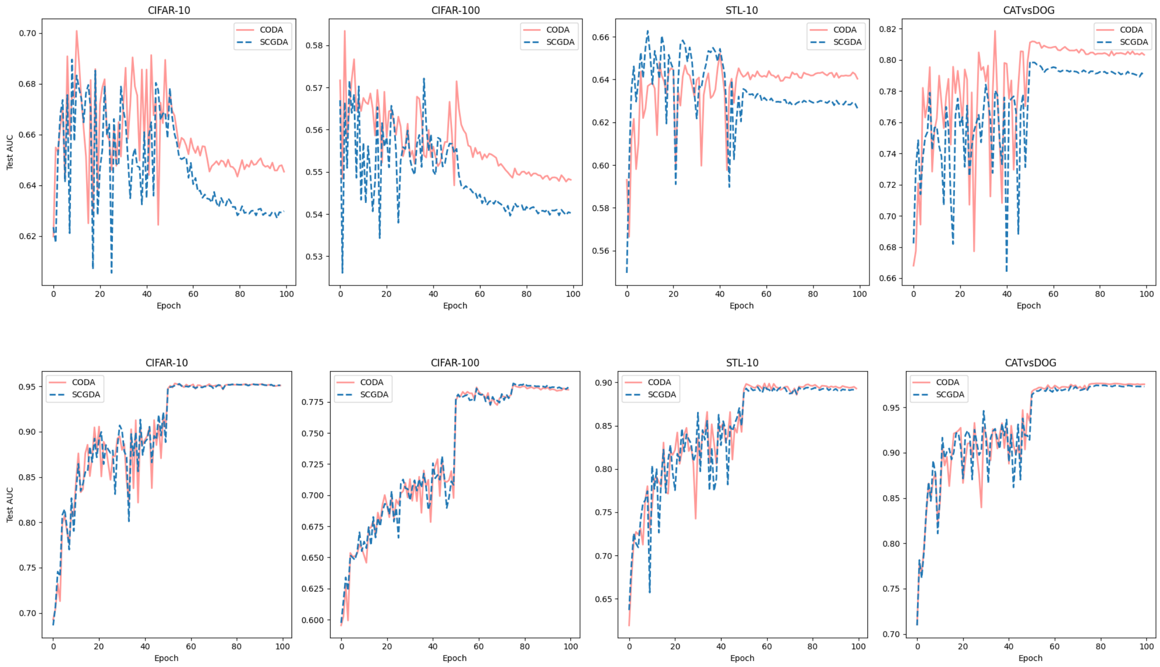

Figure 4: Testing AUC performance comparison of CODA and SCGDA on four benchmarks. The first row is under imratio= 1% and the second row is under imratio= 30%

where $L(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}; \mathbf{x}_i, y_i)$ and

$$\begin{aligned}\phi(\mathbf{w}, a, b, \theta; \mathbf{x}_i, y_i) =& (1-p)(f(\mathbf{w}; \mathbf{x}_i) - a)^2 \mathbb{I}[y_i = 1] + p(f(\mathbf{w}; \mathbf{x}_i) - b)^2 \mathbb{I}[y_i = -1] \\ &+ 2\theta \left(p f(\mathbf{w}; \mathbf{x}_i)\mathbb{I}[y_i = -1] - (1-p) f(\mathbf{w}; \mathbf{x}_i)\mathbb{I}[y_i = 1]\right) \\ &+ 2p f(\mathbf{w}; \mathbf{x}_i)\mathbb{I}[y_i = 1] - 2(1-p) f(\mathbf{w}; \mathbf{x}_i)\mathbb{I}[y_i = -1] - p(1-p)\theta^2\end{aligned}$$

At time t, sample a minibatch $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_B, y_B)$ from the training set.
First, update the auxiliary variable:

$$\begin{aligned}\mathbf{z}^{t+1} =& (1-\beta^t)\left(\mathbf{z}^t + \left(\mathbf{w}^t - \alpha\frac{1}{B}\sum_{i=1}^{B}\nabla\ell(\mathbf{w}^t; \mathbf{x}_i, y_i)\right) - \left(\mathbf{w}^{t-1} - \alpha\frac{1}{B}\sum_{i=1}^{B}\nabla\ell(\mathbf{w}^{t-1}; \mathbf{x}_i, y_i)\right)\right) \\ &+ \beta^t\left(\mathbf{w}^t - \alpha\frac{1}{B}\sum_{i=1}^{B}\nabla\ell(\mathbf{w}^t; \mathbf{x}_i, y_i)\right)\end{aligned}$$

Then, sample another minibatch $(\tilde{\mathbf{x}}_1, \tilde{y}_1), ..., (\tilde{\mathbf{x}}_B, \tilde{y}_B)$ and compute:

$$\begin{aligned}\mathbf{g}^t_{\mathbf{w}} = \frac{1}{B}\sum_{i=1}^{B}\nabla\phi(\mathbf{z}^{t+1}, a^t, b^t, \theta^t; \tilde{\mathbf{x}}_i, \tilde{y}_i),\ g^t_a = \frac{1}{B}\sum_{i=1}^{B}\nabla_a\phi(\mathbf{z}^{t+1}, a^t, b^t, \theta^t; \tilde{\mathbf{x}}_i, \tilde{y}_i), \\ g^t_b = \frac{1}{B}\sum_{i=1}^{B}\nabla_b\phi(\mathbf{z}^{t+1}, a^t, b^t, \theta^t; \tilde{\mathbf{x}}_i, \tilde{y}_i),\ g^t_\theta = \frac{1}{B}\sum_{i=1}^{B}\nabla_\theta\phi(\mathbf{z}^{t+1}, a^t, b^t, \theta^t; \tilde{\mathbf{x}}_i, \tilde{y}_i)\end{aligned}$$

Finally, update the model and $\boldsymbol{\alpha}$:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta_w \mathbf{g}^t_w,\ a^{t+1} = a^t - \eta_a g^t_a,\ b^{t+1} = b^t - \eta_b g^t_b,\ \theta^{t+1} = \mathcal{P}_\Omega\left(\theta^t + \eta_\theta \mathbf{g}^t_\theta\right)$$

Tab. 4 is the dataset description in our experimental setting. For the benchmark datasets, the training and validation splits are set as follows: 19k/1k for CatvsDog, 45k/5k for CIFAR10, 45k/5k for CIFAR100, and 4k/1k

Dataset	Number of samples	Number of classes	Imbalanced Ratio
CIFAR-10	50,000	2 (binary)	1%,10%,30%
CIFAR-100	50,000	2 (binary)	1%,10%,30%
CATvsDOG	20,000	2 (binary)	1%,10%,30%
STL10	5,000	2 (binary)	1%,10%,30%

Table 4: Dataset description for classification tasks

Datasets		AUC score			Accuracy		
	imratio	**1%**	**10%**	**30%**	**1%**	**10%**	**30%**
CIFAR-10	CODA-Primal	**63.6(0.003)**	91.5(0.005)	**95.3(0.001)**	**50.7(0.002)**	79.7(0.006)	**87.7(0.003)**
	PDSCA	62.7(0.009)	**92.3(0.016)**	95.2(0.0003)	50.6(0.0006)	**81.2(0.033)**	86.9(0.0003)
CIFAR-100	CODA-Primal	**55.1(0.001)**	**68.1(0.001)**	**78.6(0.003)**	**50.1(0.0002)**	**57.6(0.003)**	**66.0(0.006)**
	PDSCA	54.3(0.005)	68.5(0.001)	78.6(0.002)	50.1(0.0001)	57.7(0.003)	65.6(0.002)

Table 5: Testing Performance on datasets over three random runs

for STL10. The learning rate is initially 0.1, with a reduction by a factor of 10 at 50% and 75% of total training time. Fig. 3 shows the training loss under different imbalanced ratios, which illustrates the convergence of our CODA-Primal algorithm across the four datasets. The testing AUC scores and accuracy results are detailed in Tab. 5, where we compare our CODA-Primal algorithm with the PDSCA algorithm in the work Yuan et al. [2021]. We also compare CODA with SCGDA Gao et al. [2021] on testing AUC performance, and the results are shown in Fig. 4.

Task-agnostic Robust MAML: In task robust MAML, we aim to train a model that exhibits consistent performance across all observed task distributions, thereby ensuring robustness against task shifts at test time. The objective can be formulated as the following minimax problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \max_{\boldsymbol{\alpha}\in\Delta^N} \sum_{i=1}^{N} \alpha(i)\ell_i^{\text{test}}(\mathbf{w}-\eta\nabla\ell_i^{\text{train}}(\mathbf{w}))$$

where $\ell_i^{\text{train}}(\mathbf{w}) := \frac{1}{|S_i^{\text{train}}|}\sum_{\xi\in S_i^{\text{train}}}\ell(\mathbf{w};\xi)$, $\ell_i^{\text{test}}(\mathbf{w}) := \frac{1}{|S_i^{\text{test}}|}\sum_{\xi\in S_i^{\text{test}}}\ell(\mathbf{w};\xi)$. This constitutes a nonconvex-concave compositional minimax problem involving the primal variable. At time t, sample minibatch $\xi_1^{\text{train}},...,\xi_N^{\text{train}}$ from the training set of tasks $1,...,N$, and similarly for $\xi_1^{\text{test}},...,\xi_N^{\text{test}}$.
First, update the auxiliary variable:

$$\mathbf{z}_i^{t+1} = (1-\beta^t)\left(\mathbf{z}_i^t + (\mathbf{w}^t - \eta\nabla\ell_i(\mathbf{w}^t;\xi_i^{\text{train}})) - (\mathbf{w}^{t-1} - \eta\nabla\ell_i(\mathbf{w}^{t-1};\xi_i^{\text{train}}))\right) + \beta^t(\mathbf{w}^t - \eta\nabla\ell_i(\mathbf{w}^t;\xi_i^{\text{train}}))\forall i\in[N]$$

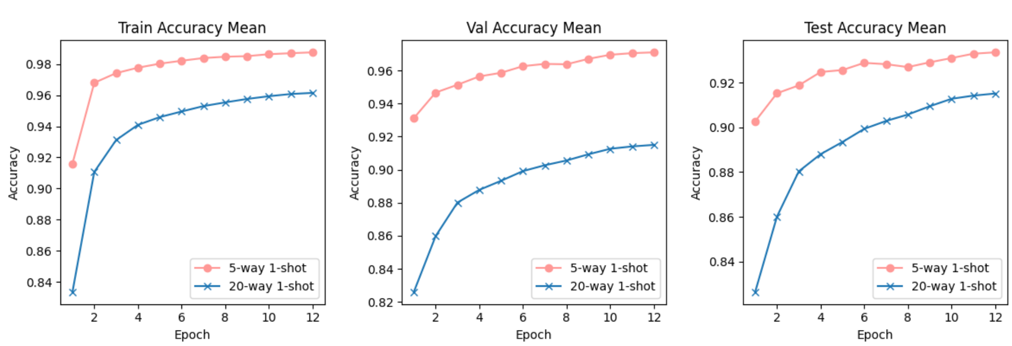

Figure 5: Meta-Training, Meta-Validation, and Meta-Testing accuracy over epochs for different task sizes ($K = 5$ and $K = 20$)

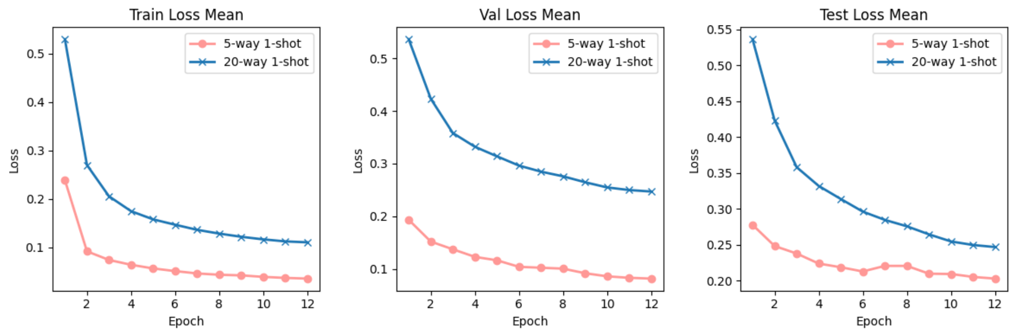

Figure 6: Meta-Training, Meta-Validation, and Meta-Testing losses over epochs for different task sizes ($K = 5$ and $K = 20$)

Next, compute the gradients:

$$\mathbf{g}_\alpha^t = [\ell_1^{\text{test}}(\mathbf{z}_1^{t+1}; \xi_1^{\text{test}}), ..., \ell_N^{\text{test}}(\mathbf{z}_N^{t+1}; \xi_N^{\text{test}})],$$
$$\mathbf{g}_\mathbf{w}^t = \sum_{i=1}^N \alpha^t(i) \nabla \ell_i^{\text{test}}(\mathbf{z}_i^{t+1}; \xi_i^{\text{test}})(\mathbf{I} - \eta \nabla^2 \ell_i^{\text{train}}(\mathbf{w}^t; \xi_i^{\text{train}}))$$

Finally, update the model and the mixture weights $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}^{t+1} = \mathcal{P}_{\Delta^N}\left(\boldsymbol{\alpha}^t + \eta_\alpha \mathbf{g}_{\boldsymbol{\alpha}}^t\right), \mathbf{w}^{t+1} = \mathbf{w}^t - \eta_w \mathbf{g}_w^t$$

It's notable that $(\mathbf{I} - \eta \nabla^2 \ell_i^{\text{train}}(\mathbf{w}^t; \xi_i^{\text{train}}))$ involves the Hessian matrix $\nabla^2 \ell_i^{\text{train}}(\mathbf{w}^t; \xi_i^{\text{train}})$. Inspired by practical implementations of MAML Finn et al. [2017] that simply ignore the second-order term, we use the same trick in our experiments. Our experimental framework involves a range of tasks, each comprising several instances. Contrasting with the standard few-shot image classification including characters from different alphabets, we consider a 5-way 1-shot classification problem involving characters from 5 distinct classes (characters) which are from a specific alphabet. Note that during meta-training, we first sample an alphabet uniformly, then sample an N-way, K-shot problem uniformly from that alphabet. The convergence of our CODA-Primal algorithm for cases with $K = 5$ and $K = 20$ is illustrated in Fig. 6, which shows the meta-training, meta-validation, and meta-testing loss. Fig. 5 compares the accuracies in different settings, which shows that the accuracies in 5-way 1-shot learning are higher than those in 20-way 1-shot learning.

Mixture Weights Estimation in Multi-source Domain Adaptation: We address the problem of multi-source domain adaptation with N source domains $\mathcal{D}_1, ..., \mathcal{D}_N$ and a single target domain $\mathcal{T}$. The objective is to train a model on a mix of multiple source domains to achieve high accuracy on the target domain. Our proposed algorithm consists of two phases: mixing weight estimation and weighted empirical risk minimization.
Phase I: mixing weight estimation We are aimed at solving the following optimization problem:

$$\min_{\alpha \in \Delta_N} \max_{\mathbf{w} \in \mathcal{W}} F(\boldsymbol{\alpha}, \mathbf{w}) = \sum_{j=1}^N \alpha(j) f(\mathcal{L}_\mathcal{T}(\mathbf{w}) - \mathcal{L}_{\mathcal{D}_j}(\mathbf{w})) + C \sum\nolimits_{j=1}^N \frac{\alpha^2(j)}{m_j},$$

where $f(x) = \sqrt{x^2 + c}$ is the approximator of the absolute value and $\mathcal{L}_\mathcal{T}, \mathcal{L}_{\mathcal{D}_1}, ..., \mathcal{L}_{\mathcal{D}_N}$ represent loss functions realized by target domain and source domains (training) data. Here, $m_1, ..., m_N$ denote the number of training data from source $1, ..., N$.

Phase II: Weighted empirical risk minimization Once the mixing weights $\boldsymbol{\alpha}$ is obtained, we perform empirical risk minimization on the mixed domain:

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^N \alpha(i) \mathcal{L}_{\hat{\mathcal{D}}_i}(h).$$

Group	**Classes**	**Domains per Group**	**Samples per Domain**
1	Airplanes, Cars, Birds	5	1500
2	Cats, Deer, Dogs	5	1500
3	Frogs, Horses, Ships, Trucks	5	1500
4	Groups 1 & 2 Combined	5	1500

Table 6: Classes and Samples per Domain for Each CIFAR-10 Group

Subsequently, we evaluate $\hat{h}$ on the target domain's test data and present the accuracy findings. We partition the CIFAR-10 dataset into a non-IID setup consisting of 4 groups, as detailed in Tab. 6. Each group consists of 5 domains (e.g., Group 1 draws data from the first 3 classes), we treat one group as the target and the remaining as sources, totaling 15 source domains with 1500 samples each. Following [Mansour et al., 2021], we compare with two baselines: (i) target-only: training only on limited target domain data, (ii) avg(equal)-weight: equal weight training across all source domains. Tab. 3 shows CODA-Dual outperforms in test accuracy, effectively identifying source domains with mixture weight $\boldsymbol{\alpha}$. Results are averaged over three random runs.

C Proof of Primal Composition Setting

In this section we provide the proof of results in primal compostion setting (Theorem 1 and Theorem 2).

The following Lemma is from Chen et al. [2021a], which bound the tracking error of the stochastic correction algorithm.

Lemma 1 (Tracking Error Chen et al. [2021a]). *For Algorithm 1, under the assumptions of Theorem 1, the following statement holds true:*

$$\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t)\right\|^2 \leq (1-\beta)^2\mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{x}^{t-1})\right\|^2+4(1-\beta)^2G_g^2\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2+2\beta^2\frac{\sigma^2}{B}.$$

C.1 Proof of Nonconvex-strongly-concave Setting

Lemma 2. *For Algorithm 1, under the assumptions of Theorem 1, the following statement holds true:*

$$\begin{aligned}\mathbb{E}[\Phi(\mathbf{x}^{t+1})] \leq \mathbb{E}[\Phi(\mathbf{x}^t)] &- \frac{\eta_{\mathbf{x}}}{2}\mathbb{E}\left\|\nabla\Phi(\mathbf{x}^t)\right\|^2-\left(\frac{\eta_{\mathbf{x}}}{2}-\eta_{\mathbf{x}}^2\kappa L\right)\mathbb{E}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2\\ &+\eta_{\mathbf{x}}G_g^2L_f^2\mathbb{E}\left\|g(\mathbf{x}^t)-\mathbf{z}^{t+1}\right\|^2+\eta_{\mathbf{x}}G_g^2L_f^2\mathbb{E}\left\|\mathbf{y}^*(\mathbf{x}^t)-\mathbf{y}^t\right\|^2+\frac{2\eta_{\mathbf{x}}^2\kappa L(G_f^2\sigma^2+G_g^2\sigma^2)}{B},\end{aligned}$$

where $\bar{\mathbf{g}}_{\mathbf{x}}^t=\nabla_1 f(\mathbf{z}^{t+1},\mathbf{y}^t)\nabla g(\mathbf{x}^t)$.

Proof. According $\kappa L+L$ smoothness [Lemma 4.3 Lin et al. [2019]] of Φ and updating rule for $\mathbf{x}$, we have:

$$\begin{aligned}\Phi(\mathbf{x}^{t+1}) &\leq \Phi(\mathbf{x}^t)+\left\langle\nabla\Phi(\mathbf{x}^t),\,\mathbf{x}^{t+1}-\mathbf{x}^t\right\rangle+\kappa L\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2\\ &=\Phi(\mathbf{x}^t)-\left\langle\nabla\Phi(\mathbf{x}^t),\,\eta_{\mathbf{x}}\mathbf{g}_{\mathbf{x}}^t\right\rangle+\kappa L\eta_{\mathbf{x}}\left\|\mathbf{g}_{\mathbf{x}}^t\right\|^2,\end{aligned}$$

where $\mathbf{g}_{\mathbf{x}}^t=\frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t}\nabla_1 f(\mathbf{z}^{t+1},\mathbf{y}^t;\zeta)\nabla g(\mathbf{x}^t;\xi)$. Notice the fact that

$$\mathbb{E}\left\|\nabla_1 f(\mathbf{z}^{t+1},\mathbf{y}^t;\zeta)\nabla g(\mathbf{x}^t;\xi)-\nabla_1 f(\mathbf{z}^{t+1},\mathbf{y}^t)\nabla g(\mathbf{x}^t)\right\|^2 \leq 2G_f^2\sigma^2+2G_g^2\sigma^2.$$

Taking expectation over the randomness of $\mathcal{B}^t$ yields:

$$\begin{aligned}\mathbb{E}[\Phi(\mathbf{x}^{t+1})] \leq \mathbb{E}[\Phi(\mathbf{x}^t)] &- \mathbb{E}\left\langle\nabla\Phi(\mathbf{x}^t),\,\eta_{\mathbf{x}}\nabla_1 f(\mathbf{z}^{t+1},\mathbf{y}^t)\nabla g(\mathbf{x}^t)\right\rangle\\ &+\eta_{\mathbf{x}}^2\kappa L\mathbb{E}\left\|\nabla_1 f(\mathbf{z}^{t+1},\mathbf{y}^t)\nabla g(\mathbf{x}^t)+\nabla h(\mathbf{x}^t)\right\|^2+\frac{2\eta_{\mathbf{x}}^2\kappa L(G_f^2\sigma^2+G_g^2\sigma^2)}{B}.\end{aligned}$$

Due to the identity $\langle\mathbf{a},\,\mathbf{b}\rangle=\frac{1}{2}\left\|\mathbf{a}\right\|^2+\frac{1}{2}\left\|\mathbf{b}\right\|^2-\frac{1}{2}\left\|\mathbf{a}-\mathbf{b}\right\|^2$, we have:

$$\begin{aligned}\mathbb{E}[\nabla\Phi(\mathbf{x}^{t+1})] \leq \mathbb{E}[\nabla\Phi(\mathbf{x}^t)] &- \frac{\eta_{\mathbf{x}}}{2}\mathbb{E}\left\|\nabla\Phi(\mathbf{x}^t)\right\|^2-\frac{\eta_{\mathbf{x}}}{2}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2\\ &+\frac{\eta_{\mathbf{x}}}{2}\left\|\nabla\Phi(\mathbf{x}^t)-\nabla_1 f(\mathbf{z}^{t+1},\mathbf{y}^t)\nabla g(\mathbf{x}^t)-\nabla h(\mathbf{x}^t)\right\|^2+\eta_{\mathbf{x}}^2\kappa L\mathbb{E}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2+\frac{2\eta_{\mathbf{x}}^2\kappa L(G_f^2\sigma^2+G_g^2\sigma^2)}{B}\end{aligned}$$

Due to [Lemma 4.3 Lin et al. [2019]], we know $\nabla\Phi(\mathbf{x}) = \nabla F(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$, we have:

$$
\begin{aligned}
\mathbb{E}[\nabla\Phi(\mathbf{x}^{t+1})] &\leq \mathbb{E}[\nabla\Phi(\mathbf{x}^t)] - \frac{\eta_{\mathbf{x}}}{2}\mathbb{E}\left\|\nabla\Phi(\mathbf{x}^t)\right\|^2 - \frac{\eta_{\mathbf{x}}}{2}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 \\
&\quad + \frac{\eta_{\mathbf{x}}}{2}\left\|\nabla_{\mathbf{x}} f(g(\mathbf{x}^t), \mathbf{y}^*(\mathbf{x}^t)) - \nabla_1 f(\mathbf{z}^{t+1}, \mathbf{y}^t)\nabla g(\mathbf{x}^t)\right\|^2 + \eta_{\mathbf{x}}^2\kappa L\mathbb{E}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 + \frac{2\eta_{\mathbf{x}}^2\kappa L(G_f^2\sigma^2 + G_g^2\sigma^2)}{B} \\
&\leq \mathbb{E}[\nabla\Phi(\mathbf{x}^t)] - \frac{\eta_{\mathbf{x}}}{2}\mathbb{E}\left\|\nabla\Phi(\mathbf{x}^t)\right\|^2 - \frac{\eta_{\mathbf{x}}}{2}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 + \eta_{\mathbf{x}}^2\kappa L\mathbb{E}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 + \frac{2\eta_{\mathbf{x}}^2\kappa L(G_f^2\sigma^2 + G_g^2\sigma^2)}{B} \\
&\quad + \frac{\eta_{\mathbf{x}}}{2}\left\|\nabla_{\mathbf{x}} f(g(\mathbf{x}^t), \mathbf{y}^*(\mathbf{x}^t)) - \nabla_{\mathbf{x}} f(g(\mathbf{x}^t), \mathbf{y}^t) + \nabla_{\mathbf{x}} f(g(\mathbf{x}^t), \mathbf{y}^t) - \nabla_1 f(\mathbf{z}^{t+1}, \mathbf{y}^t)\nabla g(\mathbf{x}^t)\right\|^2 \\
&\leq \mathbb{E}[\nabla\Phi(\mathbf{x}^t)] - \frac{\eta_{\mathbf{x}}}{2}\mathbb{E}\left\|\nabla\Phi(\mathbf{x}^t)\right\|^2 - \frac{\eta_{\mathbf{x}}}{2}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 + \eta_{\mathbf{x}}^2\kappa L\mathbb{E}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 + \frac{2\eta_{\mathbf{x}}^2\kappa L(G_f^2\sigma^2 + G_g^2\sigma^2)}{B} \\
&\quad + \eta_{\mathbf{x}}\left\|\nabla_1 f(g(\mathbf{x}^t), \mathbf{y}^*(\mathbf{x}^t))\nabla g(\mathbf{x}^t) - \nabla_1 f(g(\mathbf{x}^t), \mathbf{y}^t)\nabla g(\mathbf{x}^t)\right\|^2 \\
&\quad + \eta_{\mathbf{x}}\left\|\nabla_1 f(g(\mathbf{x}^t), \mathbf{y}^t)\nabla g(\mathbf{x}^t) - \nabla_1 f(\mathbf{z}^{t+1}, \mathbf{y}^t)\nabla g(\mathbf{x}^t)\right\|^2 \\
&\leq \mathbb{E}[\nabla\Phi(\mathbf{x}^t)] - \frac{\eta_{\mathbf{x}}}{2}\mathbb{E}\left\|\nabla\Phi(\mathbf{x}^t)\right\|^2 - \left(\frac{\eta_{\mathbf{x}}}{2} - \eta_{\mathbf{x}}^2\kappa L\right)\mathbb{E}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 \\
&\quad + \eta_{\mathbf{x}} G_g^2 L_f^2\mathbb{E}\left\|g(\mathbf{x}^t) - \mathbf{z}^{t+1}\right\|^2 + \eta_{\mathbf{x}} G_g^2 L_f^2\mathbb{E}\left\|\mathbf{y}^*(\mathbf{x}^t) - \mathbf{y}^t\right\|^2 + \frac{2\eta_{\mathbf{x}}^2\kappa L(G_f^2\sigma^2 + G_g^2\sigma^2)}{B}.
\end{aligned}
$$

□

Lemma 3. *For Algorithm 1, under the assumptions of Theorem 1, the following statement holds true:*

$$
\begin{aligned}
\left\|\mathbf{y}^{t+1} - \mathbf{y}^*(\mathbf{x}^{t+1})\right\|^2 \leq \left(1 - \frac{1}{16\kappa}\right)\left\|\mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\|^2 + 2\left(\frac{4\eta_{\mathbf{y}}}{\mu} + 2\eta_{\mathbf{y}}^2\right)L_f^2\left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\right\|^2 \\
+ 16\kappa^3\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 + \frac{2\eta_{\mathbf{y}}^2\sigma^2}{B}.
\end{aligned}
$$

Proof. We notice the following canonical decomposition due to Young's inequality:

$$
\left\|\mathbf{y}^{t+1} - \mathbf{y}^*(\mathbf{x}^{t+1})\right\|^2 \leq \left(1 + \frac{1}{2(2\kappa - 1)}\right)\underbrace{\left\|\mathbf{y}^{t+1} - \mathbf{y}^*(\mathbf{x}^t)\right\|^2}_{T_1} + (1 + 2(2\kappa - 1))\underbrace{\left\|\mathbf{y}^*(\mathbf{x}^t) - \mathbf{y}^*(\mathbf{x}^{t+1})\right\|^2}_{T_2}.
$$

We bound T_1 first. Define $\bar{\mathbf{g}}_{\mathbf{y}}^t = \nabla_{\mathbf{y}} f(\mathbf{z}^{t+1}, \mathbf{y}^t) - \nabla r(\mathbf{y}^t)$. By the updating rule we have:

$$
\begin{aligned}
\mathbb{E}\left\|\mathbf{y}^{t+1} - \mathbf{y}^*(\mathbf{x}^t)\right\|^2 &= \mathbb{E}\left\|\mathcal{P}_{\mathcal{Y}}(\mathbf{y}^t + \eta_{\mathbf{y}}\mathbf{g}_{\mathbf{y}}^t) - \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y}^*(\mathbf{x}^t) + \eta_{\mathbf{y}}\nabla_{\mathbf{y}} F(\mathbf{x}^t, \mathbf{y}^*(\mathbf{x}^t))\right)\right\|^2 \\
&\leq \mathbb{E}\left\|\mathbf{y}^t + \eta_{\mathbf{y}}\bar{\mathbf{g}}_{\mathbf{y}}^t - \mathbf{y}^*(\mathbf{x}^t) - \eta_{\mathbf{y}}\nabla_{\mathbf{y}} F(\mathbf{x}^t, \mathbf{y}^*(\mathbf{x}^t))\right\|^2 + \frac{\eta_{\mathbf{y}}^2\sigma^2}{B} \\
&= \mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\|^2 + 2\eta_{\mathbf{y}}\left\langle\bar{\mathbf{g}}_{\mathbf{y}}^t - \nabla_{\mathbf{y}} F(\mathbf{x}^t, \mathbf{y}^*(\mathbf{x}^t)),\, \mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\rangle + \frac{\eta_{\mathbf{y}}^2\sigma^2}{B} \\
&\quad + \eta_{\mathbf{y}}^2\left\|\bar{\mathbf{g}}_{\mathbf{y}}^t - \nabla_{\mathbf{y}} F(\mathbf{x}^t, \mathbf{y}^*(\mathbf{x}^t))\right\|^2 \\
&= \left\|\mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\|^2 + 2\eta_{\mathbf{y}}\left\langle\nabla_{\mathbf{y}} F(\mathbf{x}^t, \mathbf{y}^t) - \nabla_{\mathbf{y}} F(\mathbf{x}^t, \mathbf{y}^*(\mathbf{x}^t)),\, \mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\rangle \\
&\quad + 2\eta_{\mathbf{y}}\left\langle\nabla_{\mathbf{y}} f(\mathbf{z}^{t+1}, \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t),\, \mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\rangle \\
&\quad + 2\eta_{\mathbf{y}}^2\left\|\nabla_{\mathbf{y}} F(\mathbf{x}^t, \mathbf{y}^t) - \nabla_{\mathbf{y}} F(\mathbf{x}^t, \mathbf{y}^*(\mathbf{x}^t))\right\|^2 + 2\eta_{\mathbf{y}}^2\left\|\bar{\mathbf{g}}_{\mathbf{y}}^t - \nabla_{\mathbf{y}} F(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 + \frac{\eta_{\mathbf{y}}^2\sigma^2}{B} \\
&\leq \left\|\mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\|^2 + (2\eta_{\mathbf{y}} - 2\eta_{\mathbf{y}}^2 L)\left\langle\nabla_{\mathbf{y}} F(\mathbf{x}^t, \mathbf{y}^t) - \nabla_{\mathbf{y}} F(\mathbf{x}^t, \mathbf{y}^*(\mathbf{x}^t)),\, \mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\rangle \\
&\quad + 2\eta_{\mathbf{y}}\left\langle\nabla_{\mathbf{y}} f(\mathbf{z}^{t+1}, \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t),\, \mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\rangle + 2\eta_{\mathbf{y}}^2\left\|\bar{\mathbf{g}}_{\mathbf{y}}^t - \nabla_{\mathbf{y}} F(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 + \frac{\eta_{\mathbf{y}}^2\sigma^2}{B} \\
&\leq \left(1 - \frac{1}{2\kappa}\right)\left\|\mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\|^2 + 2\eta_{\mathbf{y}}\left\langle\nabla_{\mathbf{y}} f(\mathbf{z}^{t+1}, \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t),\, \mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\rangle \\
&\quad + 2\eta_{\mathbf{y}}^2\left\|\nabla_{\mathbf{y}} f(\mathbf{z}^{t+1}, \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t)\right\|^2 + \frac{\eta_{\mathbf{y}}^2\sigma^2}{B},
\end{aligned}
$$

where at first step, we use the fact that $\mathbf{y}^*(\mathbf{x}^t) = \mathcal{P}_{\mathcal{Y}}(\mathbf{y}^*(\mathbf{x}^t) + \eta_{\mathbf{y}} \nabla_{\mathbf{y}} F(\mathbf{x}^t, \mathbf{y}^*(\mathbf{x}^t)))$ for any $\eta_{\mathbf{y}} \geq 0$ [Recht and Wright, 2019, Proposition 7.4]; at fifth step, we use the classic result for smooth concave function: $\frac{1}{L} \left\|\nabla_{\mathbf{y}} F(\mathbf{x}^t, \mathbf{y}^t) - \nabla_{\mathbf{y}} F(\mathbf{x}^t, \mathbf{y}^*(\mathbf{x}^t))\right\|^2 \leq \langle \nabla_{\mathbf{y}} F(\mathbf{x}^t, \mathbf{y}^*(\mathbf{x}^t)) - \nabla_{\mathbf{y}} F(\mathbf{x}^t, \mathbf{y}^t), \mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\rangle$, and at last step we use the μ strongly concavity of $F(\mathbf{x}, \cdot)$, and $\eta_{\mathbf{y}} \leq \frac{1}{2L}$.

For the remaining inner product, we apply Cauchy-schwartz inequality:

$$\begin{aligned}
\left\langle \nabla_{\mathbf{y}} f(\mathbf{z}^{t+1}, \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t), \mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\rangle &\leq \left\|\nabla_{\mathbf{y}} f(\mathbf{z}^{t+1}, \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t)\right\| \left\|\mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\| \\
&\leq \frac{2}{\mu} \left\|\nabla_{\mathbf{y}} f(\mathbf{z}^{t+1}, \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t)\right\|^2 + \frac{\mu}{8} \left\|\mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\|^2 \\
&\leq \frac{2L_f^2}{\mu} \left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\right\|^2 + \frac{\mu}{8} \left\|\mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\|^2.
\end{aligned}$$

Similarly for $\left\|\nabla_{\mathbf{y}} f(\mathbf{z}^{t+1}, \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t)\right\|^2$ we have:

$$\left\|\nabla_{\mathbf{y}} f(\mathbf{z}^{t+1}, \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t)\right\|^2 \leq L_f^2 \left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\right\|^2$$

Putting pieces together yields:

$$\left\|\mathbf{y}^{t+1} - \mathbf{y}^*(\mathbf{x}^t)\right\|^2 \leq \left(1 - \frac{1}{8\kappa}\right) \left\|\mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\|^2 + \left(\frac{4\eta_{\mathbf{y}}}{\mu} + 2\eta_{\mathbf{y}}^2\right) L_f^2 \left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\right\|^2 + \frac{\eta_{\mathbf{y}}^2 \sigma^2}{B}.$$

We then turn to bounding T_2.

$$T_2 = \left\|\mathbf{y}^*(\mathbf{x}^t) - \mathbf{y}^*(\mathbf{x}^{t+1})\right\|^2 \leq \kappa^2 \left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2$$

Putting pieces together will conclude the proof:

$$\begin{aligned}
\left\|\mathbf{y}^{t+1} - \mathbf{y}^*(\mathbf{x}^{t+1})\right\|^2 \leq& \left(1 + \frac{1}{2(8\kappa - 1)}\right) \left(\left(1 - \frac{1}{8\kappa}\right) \left\|\mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\|^2 + \left(\frac{4\eta_{\mathbf{y}}}{\mu} + 2\eta_{\mathbf{y}}^2\right) L_f^2 \left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\right\|^2 + \frac{\eta_{\mathbf{y}}^2 \sigma^2}{B}\right) \\
&+ (1 + 2(8\kappa - 1)) \left\|\mathbf{y}^*(\mathbf{x}^t) - \mathbf{y}^*(\mathbf{x}^{t+1})\right\|^2 \\
\leq& \left(1 - \frac{1}{16\kappa}\right) \left\|\mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\|^2 + \left(1 + \frac{1}{2(8\kappa - 1)}\right) \left(\frac{4\eta_{\mathbf{y}}}{\mu} + 2\eta_{\mathbf{y}}^2\right) L_f^2 \left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\right\|^2 \\
&+ 16\kappa^3 \left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 + \frac{2\eta_{\mathbf{y}}^2 \sigma^2}{B} \\
\leq& \left(1 - \frac{1}{16\kappa}\right) \left\|\mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\|^2 + 2\left(\frac{4\eta_{\mathbf{y}}}{\mu} + 2\eta_{\mathbf{y}}^2\right) L_f^2 \left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\right\|^2 \\
&+ 16\kappa^3 \left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 + \frac{2\eta_{\mathbf{y}}^2 \sigma^2}{B}.
\end{aligned}$$

□

C.1.1 Proof of Theorem 1

Proof Sketch. Before delving into the detailed proof, we first provide a sketch. Unlike the classic nonconvex-strongly-concave analysis, we consider the following potential function:

$$\begin{aligned}
\Psi(\mathbf{x}^{t+1}) =& \mathbb{E}[\Phi(\mathbf{x}^{t+1})] + O\left(\kappa \eta_{\mathbf{x}} G_g^2 L_f^4 \left(\frac{\eta_{\mathbf{y}}}{\mu} + \eta_{\mathbf{y}}^2\right)\right) \mathbb{E}\left\|g(\mathbf{x}^t) - \mathbf{z}^{t+1}\right\|^2 \\
&+ O(\kappa \eta_{\mathbf{x}} G_g^2 L_f^2) \mathbb{E}\left\|\mathbf{y}^*(\mathbf{x}^t) - \mathbf{y}^t\right\|^2 + O(\kappa^4 \eta_{\mathbf{x}} G_g^2 L_f^2) \mathbb{E}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2.
\end{aligned}$$

Then, we derive the convergence of dual gap $\left\|\mathbf{y}^{t+1} - \mathbf{y}^*(\mathbf{x}^{t+1})\right\|^2$:

$$\begin{aligned}
\left\|\mathbf{y}^{t+1} - \mathbf{y}^*(\mathbf{x}^{t+1})\right\|^2 \leq& \left(1 - \frac{1}{16\kappa}\right) \left\|\mathbf{y}^t - \mathbf{y}^*(\mathbf{x}^t)\right\|^2 + 2\left(\frac{4\eta_{\mathbf{y}}}{\mu} + 2\eta_{\mathbf{y}}^2\right) L_f^2 \left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\right\|^2 \\
&+ 16\kappa^3 \left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 + \frac{2\eta_{\mathbf{y}}^2 \sigma^2}{B},
\end{aligned}$$

where $\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t)\right\|^2$ is the error of inner function estimation, and can be bounded using standard result from Chen et al. [2021a]. Combining above ingredients will yield

$$\begin{aligned}\Psi(\mathbf{x}^{t+1}) &\leq \Psi(\mathbf{x}^t) - \frac{\eta_{\mathbf{x}}}{2}\mathbb{E}\left\|\nabla\Phi(\mathbf{x}^t)\right\|^2 \\ &\quad + O\left(\frac{(\kappa^4\eta_{\mathbf{x}}^3 G_g^2 L_f^2 + \eta_{\mathbf{x}}^2\kappa L)(G_f^2+G_g^2)\sigma^2}{B}\right) + O\left(\kappa^2 L\eta_{\mathbf{x}}\frac{\sigma^2}{B}\right) + O\left(\kappa\eta_{\mathbf{x}} G_g^2 L_f^2 \frac{2\eta_{\mathbf{y}}^2\sigma^2}{B}\right).\end{aligned}$$

Unrolling above recursion will yield desired result.

Proof. First due to Lemma 2 we have:

$$\begin{aligned}\mathbb{E}[\Phi(\mathbf{x}^{t+1})] &\leq \mathbb{E}[\Phi(\mathbf{x}^t)] - \frac{\eta_{\mathbf{x}}}{2}\mathbb{E}\left\|\nabla\Phi(\mathbf{x}^t)\right\|^2 - \left(\frac{\eta_{\mathbf{x}}}{2} - \eta_{\mathbf{x}}^2\kappa L\right)\mathbb{E}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 \\ &\quad + \eta_{\mathbf{x}} G_g^2 L_f^2 \mathbb{E}\left\|g(\mathbf{x}^t)-\mathbf{z}^{t+1}\right\|^2 + \eta_{\mathbf{x}} G_g^2 L_f^2 \mathbb{E}\left\|\mathbf{y}^*(\mathbf{x}^t)-\mathbf{y}^t\right\|^2 + \frac{2\eta_{\mathbf{x}}^2\kappa L(G_f^2\sigma^2+G_g^2\sigma^2)}{B}.\end{aligned}$$

Define potential function $\Psi(\mathbf{x}^{t+1}) = \mathbb{E}[\Phi(\mathbf{x}^{t+1})] + C_1\mathbb{E}\left\|g(\mathbf{x}^t)-\mathbf{z}^{t+1}\right\|^2 + C_2\mathbb{E}\left\|\mathbf{y}^*(\mathbf{x}^t)-\mathbf{y}^t\right\|^2 + C_3\mathbb{E}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2$, where C_1, C_2 and C_3 are determined later. Hence we have :

$$\begin{aligned}\Psi(\mathbf{x}^{t+1}) &\leq \Psi(\mathbf{x}^{t+1}) - \frac{\eta_{\mathbf{x}}}{2}\mathbb{E}\left\|\nabla\Phi(\mathbf{x}^t)\right\|^2 - \left(\frac{\eta_{\mathbf{x}}}{2} - \eta_{\mathbf{x}}^2\kappa L\right)\mathbb{E}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 \\ &\quad + \eta_{\mathbf{x}} G_g^2 L_f^2 \mathbb{E}\left\|g(\mathbf{x}^t)-\mathbf{z}^{t+1}\right\|^2 + \eta_{\mathbf{x}} G_g^2 L_f^2 \mathbb{E}\left\|\mathbf{y}^*(\mathbf{x}^t)-\mathbf{y}^t\right\|^2 + \frac{2\eta_{\mathbf{x}}^2\kappa L(G_f^2\sigma^2+G_g^2\sigma^2)}{B} \\ &\quad + C_1\mathbb{E}\left\|g(\mathbf{x}^t)-\mathbf{z}^{t+1}\right\|^2 + C_2\mathbb{E}\left\|\mathbf{y}^*(\mathbf{x}^t)-\mathbf{y}^t\right\|^2 + C_3\mathbb{E}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2 \\ &\quad - C_1\mathbb{E}\left\|g(\mathbf{x}^{t-1})-\mathbf{z}^t\right\|^2 - C_2\mathbb{E}\left\|\mathbf{y}^*(\mathbf{x}^{t-1})-\mathbf{y}^{t-1}\right\|^2 - C_3\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2.\end{aligned}$$

Since $\mathbb{E}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2 \leq \eta_{\mathbf{x}}^2\mathbb{E}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 + \eta_{\mathbf{x}}^2\frac{2(G_f^2+G_g^2)\sigma^2}{B}$, we have

$$\begin{aligned}\Psi(\mathbf{x}^{t+1}) &\leq \Psi(\mathbf{x}^{t+1}) - \frac{\eta_{\mathbf{x}}}{2}\mathbb{E}\left\|\nabla\Phi(\mathbf{x}^t)\right\|^2 - \left(\frac{\eta_{\mathbf{x}}}{2} - \eta_{\mathbf{x}}^2\kappa L - C_3\eta_{\mathbf{x}}^2\right)\mathbb{E}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 + \frac{2C_3\eta_{\mathbf{x}}^2(G_f^2\sigma^2+G_g^2\sigma^2)}{B} \\ &\quad + \left(\eta_{\mathbf{x}} G_g^2 L_f^2 + C_1\right)\mathbb{E}\left\|g(\mathbf{x}^t)-\mathbf{z}^{t+1}\right\|^2 + \left(\eta_{\mathbf{x}} G_g^2 L_f^2 + C_2\right)\mathbb{E}\left\|\mathbf{y}^*(\mathbf{x}^t)-\mathbf{y}^t\right\|^2 + \frac{2\eta_{\mathbf{x}}^2\kappa L(G_f^2\sigma^2+G_g^2\sigma^2)}{B} \\ &\quad - C_1\mathbb{E}\left\|g(\mathbf{x}^{t-1})-\mathbf{z}^t\right\|^2 - C_2\mathbb{E}\left\|\mathbf{y}^*(\mathbf{x}^{t-1})-\mathbf{y}^{t-1}\right\|^2 - C_3\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2.\end{aligned}$$

We plug in the bound for $\mathbb{E}\left\|g(\mathbf{x}^t)-\mathbf{z}^{t+1}\right\|^2$ and $\mathbb{E}\left\|\mathbf{y}^*(\mathbf{x}^t)-\mathbf{y}^t\right\|^2$:

$$\begin{aligned}\Psi(\mathbf{x}^{t+1}) &\leq \Psi(\mathbf{x}^t) - \frac{\eta_{\mathbf{x}}}{2}\mathbb{E}\left\|\nabla\Phi(\mathbf{x}^t)\right\|^2 - \left(\frac{\eta_{\mathbf{x}}}{2} - \eta_{\mathbf{x}}^2\kappa L - C_3\eta_{\mathbf{x}}^2\right)\mathbb{E}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 + \frac{(2C_3\eta_{\mathbf{x}}^2 + 2\eta_{\mathbf{x}}^2\kappa L)(G_f^2\sigma^2+G_g^2\sigma^2)}{B} \\ &\quad + \left(\eta_{\mathbf{x}} G_g^2 L_f^2 + C_1\right)\left((1-\beta)^2\mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{x}^{t-1})\right\|^2 + 4(1-\beta)^2 G_g^2\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2 + 2\beta^2\frac{\sigma^2}{B}\right) \\ &\quad + \left(\eta_{\mathbf{x}} G_g^2 L_f^2 + C_2\right) \\ &\quad \times\left(\left(1-\frac{1}{16\kappa}\right)\left\|\mathbf{y}^{t-1}-\mathbf{y}^*(\mathbf{x}^{t-1})\right\|^2 + 2\left(\frac{4\eta_{\mathbf{y}}}{\mu}+2\eta_{\mathbf{y}}^2\right)L_f^2\left\|\mathbf{z}^t-g(\mathbf{x}^{t-1})\right\|^2 + 16\kappa^3\left\|\mathbf{x}^{t-1}-\mathbf{x}^t\right\|^2 + \frac{2\eta_{\mathbf{y}}^2\sigma^2}{B}\right) \\ &\quad - C_1\mathbb{E}\left\|g(\mathbf{x}^{t-1})-\mathbf{z}^t\right\|^2 - C_2\mathbb{E}\left\|\mathbf{y}^*(\mathbf{x}^{t-1})-\mathbf{y}^{t-1}\right\|^2 - C_3\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2 \\ &\leq \Psi(\mathbf{x}^{t+1}) - \frac{\eta_{\mathbf{x}}}{2}\mathbb{E}\left\|\nabla\Phi(\mathbf{x}^t)\right\|^2 - \left(\frac{\eta_{\mathbf{x}}}{4} - C_3\eta_{\mathbf{x}}^2\right)\mathbb{E}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 \\ &\quad + \frac{(2C_3\eta_{\mathbf{x}}^2 + \eta_{\mathbf{x}}^2\kappa L)(G_f^2+G_g^2)\sigma^2}{B} + \left(\eta_{\mathbf{x}} G_g^2 L^2 + C_1\right)2\beta^2\frac{\sigma^2}{B} + \left(\eta_{\mathbf{x}} G_g^2 L_f^2 + C_2\right)\frac{2\eta_{\mathbf{y}}^2\sigma^2}{B} \\ &\quad + \left[\left(\eta_{\mathbf{x}} G_g^2 L_f^2 + C_1\right)(1-\beta)^2 + \left(\eta_{\mathbf{x}} G_g^2 L^2 + C_2\right)2\left(\frac{4\eta_{\mathbf{y}}}{\mu}+2\eta_{\mathbf{y}}^2\right)L_f^2 - C_1\right]\mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{x}^{t-1})\right\|^2 \\ &\quad + \left[\left(\eta_{\mathbf{x}} G_g^2 L_f^2 + C_2\right)\left(1-\frac{1}{16\kappa}\right) - C_2\right]\mathbb{E}\left\|\mathbf{y}^{t-1}-\mathbf{y}^*(\mathbf{x}^{t-1})\right\|^2 + \left[16\kappa^3\left(\eta_{\mathbf{x}} G_g^2 L_f^2 + C_2\right) - C_3\right]\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2,\end{aligned}$$

where at second inequality we use the fact $\eta_{\mathbf{x}} \leq \frac{1}{4\kappa L}$, so $\frac{\eta_{\mathbf{x}}}{2} - \eta_{\mathbf{x}}\kappa L \geq \frac{\eta_{\mathbf{x}}}{4}$. We choose C_1, C_2, C_3 such that the following conditions are satisfied:

$$\begin{aligned}
\frac{\eta_{\mathbf{x}}}{4} - C_3\eta_{\mathbf{x}}^2 &\geq 0,\\
\left(\eta_{\mathbf{x}} G_g^2 L_f^2 + C_1\right)(1-\beta) + \left(\eta_{\mathbf{x}} G_g^2 L_f^2 + C_2\right) 2\left(\frac{4\eta_{\mathbf{y}}}{\mu} + 2\eta_{\mathbf{y}}^2\right) L_f^2 - C_1 &\leq 0,\\
\left(\eta_{\mathbf{x}} G_g^2 L_f^2 + C_2\right)\left(1 - \frac{1}{16\kappa}\right) - C_2 &\leq 0,\\
16\kappa^3\left(\eta_{\mathbf{x}} G_g^2 L_f^2 + C_2\right) - C_3 &\leq 0.
\end{aligned}$$

It suffices to guarantee the following inequality holding:

$$\begin{aligned}
&16\kappa\left(1 - \frac{1}{16\kappa}\right)\eta_{\mathbf{x}} G_g^2 L_f^2 \leq C_2 \leq \frac{1}{16\kappa^3} C_3 - \eta_{\mathbf{x}} G_g^2 L_f^2,\\
&C_1 \geq \frac{1}{\beta}\left(\eta_{\mathbf{x}} G_g^2 L_f^2 + C_2\right) 2\left(\frac{4\eta_{\mathbf{y}}}{\mu} + 2\eta_{\mathbf{y}}^2\right) L_f^2 + \frac{1-\beta}{\beta}\eta_{\mathbf{x}} G_g^2 L_f^2.
\end{aligned}$$

A set of choices that can satisfy above conditions are the following:

$$\begin{aligned}
&\beta = \frac{1}{2}, C_1 = 80\kappa\eta_{\mathbf{x}} G_g^2 L_f^4\left(\frac{4\eta_{\mathbf{y}}}{\mu} + 2\eta_{\mathbf{y}}^2\right) + \eta_{\mathbf{x}} G_g^2 L_f^2,\\
&C_2 = 16\kappa\eta_{\mathbf{x}} G_g^2 L_f^2,\ C_3 = 320\kappa^4\eta_{\mathbf{x}} G_g^2 L_f^2.
\end{aligned}$$

We also need $C_3 = 320\kappa^4\eta_{\mathbf{x}} G_g^2 L_f^2 \geq \frac{1}{4\eta_{\mathbf{x}}}$, which can be satisfied by choosing $\eta_{\mathbf{x}} = \frac{1}{36\kappa^2 L}$.

Hence we have:

$$\begin{aligned}
\Psi(\mathbf{x}^{t+1}) \leq{}& \Psi(\mathbf{x}^t) - \frac{\eta_{\mathbf{x}}}{2}\mathbb{E}\left\|\nabla\Phi(\mathbf{x}^t)\right\|^2\\
&+ \frac{(640\kappa^4\eta_{\mathbf{x}}^3 G_g^2 L_f^2 + \eta_{\mathbf{x}}^2\kappa L)(G_f^2 + G_g^2)\sigma^2}{B} + \left(\eta_{\mathbf{x}} G_g^2 L_f^2 + C_1\right)\frac{\sigma^2}{B} + 17\kappa\eta_{\mathbf{x}} G_g^2 L_f^2\frac{2\eta_{\mathbf{y}}^2\sigma^2}{B}.
\end{aligned}$$

Notice that $L \geq \max\{G_g L_f, L_f\}$, and we choose $\eta_{\mathbf{x}} \leq \frac{1}{L}$, we know $C_1 \leq 320\kappa\eta_{\mathbf{x}}\kappa L^2 + \eta_{\mathbf{x}} L^2 \leq 321\kappa^2 L^2\eta_{\mathbf{x}}$. Summing over $t = 0$ to T and re-arranging terms yield:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi(\mathbf{x}^t)\right\|^2 \leq{}& \frac{2(\Psi(\mathbf{x}^0) - \Psi(\mathbf{x}^T))}{\eta_{\mathbf{x}} T} + \frac{2(640\kappa^4\eta_{\mathbf{x}}^3 G_g^2 L_f^2 + \eta_{\mathbf{x}}^2\kappa L)(G_f^2 + G_g^2)\sigma^2}{\eta_{\mathbf{x}} B} + 642\kappa^2 L^2\frac{\sigma^2}{B} + 17\kappa\frac{\sigma^2}{B}\\
\leq{}& 2\frac{E[\Phi(\mathbf{x}^0)] + 321\kappa^2 L^2\eta_{\mathbf{x}}\mathbb{E}\left\|g(\mathbf{x}^{-1}) - \mathbf{z}^0\right\|^2 + 16\kappa\eta_{\mathbf{x}} L^2\mathbb{E}\left\|\mathbf{y}^*(\mathbf{x}^{-1}) - \mathbf{y}^{-1}\right\|^2 + 320\kappa^4\eta_{\mathbf{x}} L^2\mathbb{E}\left\|\mathbf{x}^0 - \mathbf{x}^{-1}\right\|^2}{\eta_{\mathbf{x}} T}\\
&+ \frac{2(640\kappa^4\eta_{\mathbf{x}}^3 G_g^2 L_f^2 + \eta_{\mathbf{x}}^2\kappa L)(G_f^2 + G_g^2)\sigma^2}{\eta_{\mathbf{x}} B} + 642\kappa^2 L^2\frac{\sigma^2}{B} + 17\kappa\frac{\sigma^2}{B}.
\end{aligned}$$

By convention, we set $\mathbf{x}^{-1} = \mathbf{x}^0$, and with our choice $\eta_{\mathbf{x}} = \frac{1}{36\kappa^2 L}$ we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi(\mathbf{x}^t)\right\|^2 \leq{}& O\left(\frac{E[\Phi(\mathbf{x}^0) - \min_{\mathbf{x}}\Phi(\mathbf{x})] + \kappa\mathbb{E}\left\|g(\mathbf{x}^0) - \mathbf{z}^0\right\|^2 + LD_{\mathcal{Y}}^2}{\eta_{\mathbf{x}} T}\right)\\
&+ O\left(\frac{(\kappa^4\eta_{\mathbf{x}}^3 L^2 + \eta_{\mathbf{x}}^2\kappa L)(G_f^2 + G_g^2)\sigma^2}{\eta_{\mathbf{x}} B} + \kappa^2 L\frac{\sigma^2}{B} + \kappa\frac{\sigma^2}{B}\right).
\end{aligned}$$

Due to our initialization, $\mathbb{E}\left\|g(\mathbf{x}^0) - \mathbf{z}^0\right\|^2 \leq \frac{1}{\kappa}$, and this can be done by $O(\kappa^2)$ zero order sampling. We choose $B = \Theta\left(\max\left\{\frac{\kappa^2 L\sigma^2}{\epsilon^2}, 1\right\}\right)$, $T = O\left(\max\left\{\frac{\kappa^2 L\Delta_\Phi}{\epsilon^2}, \frac{\kappa^2 L^2 D_{\mathcal{Y}}}{\epsilon^2}\right\}\right)$, which yields final gradient complexity of

$$O\left(\frac{\kappa^2(\Delta_\Phi + L^2 D_{\mathcal{Y}})}{\epsilon^2}\max\left\{\frac{\kappa^2 L\sigma^2}{\epsilon^2}, 1\right\}\right),$$

as desired. □

C.2 Proof of Nonconvex-concave setting

Proof Sketch: Before delving into the detailed proof, we first provide a sketch. In nonconvex-concave setting, we consider the following potential function:

$$\Psi^t = \Phi_{1/2L}(\mathbf{x}^t) + C_1\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\right\|^2 + C_2\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\right\| \tag{3}$$

where $C_1 = \frac{\eta_{\mathbf{x}}}{2(2\beta-\beta^2)}G_g^2L_f^2$ and $C_2 = \frac{2\eta_{\mathbf{x}}L^2D_{\mathbf{y}}}{\beta}$. Another key step in nonconvex-concave analysis is to bound primal function gap. During the dynamic of our algorithm, the following statement holds:

$$\begin{aligned}
&\mathbb{E}\left(\Phi(\mathbf{x}^{t-1}) - F(\mathbf{x}^{t-1},\mathbf{y}^{t-1})\right)\\
\leq&(2t-1-2s)\eta_{\mathbf{x}}\sqrt{2G_h^2+2G_f^2G_g^2} + \frac{1}{2\eta_{\mathbf{y}}}(\mathbb{E}\|\mathbf{y}^*(\mathbf{x}^s)-\mathbf{y}^{t-1}\|^2 - \mathbb{E}\|\mathbf{y}^*(\mathbf{x}^s)-\mathbf{y}^t\|^2) + \eta_{\mathbf{y}}\frac{\sigma^2}{M}\\
&+ LD_{\mathcal{Y}}\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\| + \mathbb{E}\left[F(\mathbf{x}^t,\mathbf{y}^t) - F(\mathbf{x}^{t-1},\mathbf{y}^{t-1})\right].
\end{aligned}$$

Then we can derive that our algorithm satisfies the following descent inequality:

$$\begin{aligned}
\mathbb{E}[\Psi^t - \Psi^{t-1}] \leq\ & 2\eta_{\mathbf{x}}L\left(\frac{\mathbb{E}\|\mathbf{y}^*(\mathbf{x}^s)-\mathbf{y}^{t-1}\|^2 - \mathbb{E}\|\mathbf{y}^*(\mathbf{x}^s)-\mathbf{y}^t\|^2}{2\eta_{\mathbf{y}}} + \eta_{\mathbf{y}}\frac{\sigma^2}{M} + \mathbb{E}[F(\mathbf{x}^t,\mathbf{y}^t) - F(\mathbf{x}^{t-1},\mathbf{y}^{t-1})]\right)\\
&-\frac{\eta_{\mathbf{x}}}{8}\mathbb{E}\|\nabla\Phi_{1/2L}(\mathbf{x}^{t-1})\|^2 + 2L\eta_{\mathbf{x}}^2(G_h^2+G_f^2G_g^2) + 4\eta_{\mathbf{x}}L(2t-1-2s)\eta_{\mathbf{x}}(G_h+G_fG_g)\\
&+C_1\left(8(1-\beta)^2G_g^2\eta_{\mathbf{x}}^2(G_h^2+G_f^2G_g^2) + 2\beta^2\frac{\sigma^2}{B}\right) + C_2\sqrt{8(1-\beta)^2G_g^2\eta_{\mathbf{x}}^2(G_h^2+G_f^2G_g^2)+2\beta^2\frac{\sigma^2}{B}}.
\end{aligned}$$

Lemma 4. *For CODA-Primal (Algorithm 1), under same assumptions as in Theorem 2, the following statement holds :*

$$\begin{aligned}
\mathbb{E}[\Phi_{1/2L}(\mathbf{x}^t)] \leq\ & \mathbb{E}[\Phi_{1/2L}(\mathbf{x}^{t-1})] + 2\eta_{\mathbf{x}}L\mathbb{E}\left(\Phi(\mathbf{x}^{t-1}) - F(\mathbf{x}^{t-1},\mathbf{y}^{t-1})\right) - \frac{\eta_{\mathbf{x}}}{8}\mathbb{E}\|\nabla\Phi_{1/2L}(\mathbf{x}^{t-1})\|^2\\
&+ L\eta_{\mathbf{x}}^2G_f^2G_g^2 + \frac{\eta_{\mathbf{x}}}{2}G_g^2L_f^2\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2.
\end{aligned}$$

Proof. Let $\hat{\mathbf{x}}^{t-1} = \arg\min_{\mathbf{x}\in\mathbb{R}^d}\Phi(\mathbf{x}) + L\|\mathbf{x}-\mathbf{x}^{t-1}\|^2$. Notice that:

$$\begin{aligned}
\mathbb{E}[\Phi_{1/2L}(\mathbf{x}^t)] &\leq \mathbb{E}[\Phi_{1/2L}(\hat{\mathbf{x}}^{t-1})] + L\mathbb{E}\|\hat{\mathbf{x}}^{t-1} - \mathbf{x}^t\|^2\\
&\leq \mathbb{E}[\Phi_{1/2L}(\hat{\mathbf{x}}^{t-1})] + L(\mathbb{E}\|\mathbf{x}^{t-1} - \hat{\mathbf{x}}^{t-1}\|^2 + 2\eta_{\mathbf{x}}\langle\mathbf{g}_{\mathbf{x}}^t, \hat{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\rangle + \eta_{\mathbf{x}}^2\mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^t\right\|^2).
\end{aligned}$$

Due to the boundedness of gradients, we have

$$\begin{aligned}
\mathbb{E}[\Phi_{1/2L}(\mathbf{x}^t)] =&\mathbb{E}[\Phi_{1/2L}(\hat{\mathbf{x}}^{t-1})]\\
&+ L(\mathbb{E}\|\mathbf{x}^{t-1} - \hat{\mathbf{x}}^{t-1}\|^2 + 2\eta_{\mathbf{x}}\langle\nabla_1 f(\mathbf{z}^t,\mathbf{y}^{t-1})\nabla g(\mathbf{x}^{t-1}) + \nabla h(\mathbf{x}^{t-1}), \hat{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\rangle + 2\eta_{\mathbf{x}}^2(G_h^2 + G_f^2G_g^2))\\
=&\mathbb{E}[\Phi_{1/2L}(\hat{\mathbf{x}}^{t-1})] + L(\mathbb{E}\|\mathbf{x}^{t-1} - \hat{\mathbf{x}}^{t-1}\|^2 + 2\eta_{\mathbf{x}}\langle\nabla_x F(\mathbf{x}^{t-1},\mathbf{y}^{t-1}), \hat{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\rangle + 2\eta_{\mathbf{x}}^2(G_h^2 + G_f^2G_g^2))\\
&+ 2\eta_{\mathbf{x}}L\langle\nabla_1 f(\mathbf{z}^t,\mathbf{y}^{t-1})\nabla g(\mathbf{x}^{t-1}) - \nabla_1 f(\mathbf{x}^{t-1},\mathbf{y}^{t-1})\nabla g(\mathbf{x}^{t-1}), \hat{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\rangle.
\end{aligned}$$

According to smoothness of $F(\cdot,\mathbf{y})$, we have:

$$\begin{aligned}
\langle\hat{\mathbf{x}}^{t-1} - \mathbf{x}_{t-1}, \nabla_x F(\mathbf{x}^{t-1},\mathbf{y}^{t-1})\rangle &\leq F(\hat{\mathbf{x}}^{t-1},\mathbf{y}^{t-1}) - F(\mathbf{x}^{t-1},\mathbf{y}^{t-1}) + \frac{L}{2}\|\hat{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\|^2\\
&= F(\hat{\mathbf{x}}^{t-1},\mathbf{y}^{t-1}) + L\|\hat{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\|^2 - F(\mathbf{x}^{t-1},\mathbf{y}^{t-1}) - \frac{L}{2}\|\hat{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\|^2\\
&\leq \Phi(\hat{\mathbf{x}}^{t-1}) + L\|\hat{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\|^2 - F(\mathbf{x}^{t-1},\mathbf{y}^{t-1}) - \frac{L}{2}\|\hat{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\|^2\\
&\leq \Phi(\mathbf{x}^{t-1}) - F(\mathbf{x}^{t-1},\mathbf{y}^{t-1}) - \frac{L}{2}\|\hat{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\|^2,
\end{aligned}$$

where last step is due to that $\hat{\mathbf{x}}^{t-1}$ is the minimizer of $\Phi(\cdot) + L\left\|\cdot - \mathbf{x}^{t-1}\right\|^2$. So we have

$$\begin{aligned}
\mathbb{E}[\Phi_{1/2L}(\mathbf{x}^t)] &\leq \mathbb{E}[\Phi_{1/2L}(\hat{\mathbf{x}}^{t-1})] + L\mathbb{E}\|\hat{\mathbf{x}}^{t-1} - \mathbf{x}^t\|^2 \\
&\leq \mathbb{E}[\Phi_{1/2L}(\hat{\mathbf{x}}^{t-1})] + L\mathbb{E}\|\mathbf{x}^{t-1} - \hat{\mathbf{x}}^{t-1}\|^2 \\
&\quad + 2\eta_{\mathbf{x}} L\mathbb{E}\left(\Phi(\mathbf{x}^{t-1}) - f(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}) - \frac{L}{2}\mathbb{E}\|\hat{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\|^2\right) + 2L\eta_{\mathbf{x}}^2(G_h^2 + G_f^2 G_g^2) \\
&\quad + \eta_{\mathbf{x}} L\left(\frac{1}{2L}\mathbb{E}\|\nabla_1 f(\mathbf{z}^t, \mathbf{y}^{t-1})\nabla g(\mathbf{x}^{t-1}) - \nabla_1 f(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\nabla g(\mathbf{x}^{t-1})\|^2 + \frac{L}{2}\mathbb{E}\|\mathbf{x}^{t-1} - \hat{\mathbf{x}}^{t-1}\|^2\right) \\
&\leq \mathbb{E}[\Phi_{1/2L}(\mathbf{x}^{t-1})] + 2\eta_{\mathbf{x}} L\mathbb{E}\left(\Phi(\mathbf{x}^{t-1}) - f(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\right) - \frac{\eta_{\mathbf{x}} L^2}{2}\mathbb{E}\|\hat{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\|^2 \\
&\quad + 2L\eta_{\mathbf{x}}^2(G_h^2 + G_f^2 G_g^2) + \frac{\eta_{\mathbf{x}}}{2}G_g^2 L_f^2\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2.
\end{aligned}$$

Finally using the fact that $\left\|\hat{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\right\| = \left\|\nabla\Phi_{1/2L}(\mathbf{x}^{t-1})\right\|/2L$ concludes the proof. □

Lemma 5. *For CODA-Primal (Algorithm 1), under the same assumptions made in Theorem 2, if we choose $\eta \leq 1/4L$ the following statement holds for any $\mathbf{y} \in \mathcal{Y}$:*

$$\begin{aligned}
\mathbb{E}[F(\mathbf{x}^{t-1}, \mathbf{y}) - F(\mathbf{x}^{t-1}, \mathbf{y}^t)] \leq \frac{1}{2\eta_{\mathbf{y}}}(\mathbb{E}\|\mathbf{y} - \mathbf{y}^{t-1}\|^2 - \mathbb{E}\|\mathbf{y} - \mathbf{y}^t\|^2) - \left(\frac{1}{4\eta_{\mathbf{y}}} - \frac{L}{2}\right)\mathbb{E}\|\mathbf{y}^t - \mathbf{y}^{t-1}\|^2 + \eta_{\mathbf{y}}\sigma^2 \\
+ L_f D_{\mathcal{Y}}\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|.
\end{aligned}$$

Proof. According to updating rule of $\mathbf{y}$:

$$\mathbf{y}^t = \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y}^{t-1} + \eta_{\mathbf{y}}\mathbf{g}_y^{t-1}\right) = \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y}^{t-1} + \eta_{\mathbf{y}}(\nabla_2 f(\mathbf{z}^t, \mathbf{y}^{t-1}; \zeta^{t-1}) - \nabla r(\mathbf{y}^{t-1}))\right).$$

We define

$$\begin{aligned}
\varepsilon_{t-1} &= (\nabla_2 f(\mathbf{z}^t, \mathbf{y}^{t-1}; \zeta^{t-1}) - \nabla r(\mathbf{y}^{t-1})) - \nabla_{\mathbf{y}} F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}) \\
&= \nabla_2 f(\mathbf{z}^t, \mathbf{y}^{t-1}; \zeta^{t-1}) - \nabla_2 f(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}),
\end{aligned}$$

and re-write the updating rule as:

$$\mathbf{y}^t = \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y}^{t-1} + \eta_{\mathbf{y}}\nabla_{\mathbf{y}} F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}) + \eta\varepsilon_{t-1}\right).$$

Due to the property of projection we have:

$$(\mathbf{y} - \mathbf{y}^t)^\top(\mathbf{y}^t - \mathbf{y}^{t-1} - \eta_{\mathbf{y}}\nabla_{\mathbf{y}} F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}) - \eta_{\mathbf{y}}\varepsilon_{t-1}) \geq 0.$$

Using the identity that $\langle \mathbf{a}, \mathbf{b}\rangle = \frac{1}{2}(\|\mathbf{a} + \mathbf{b}\|^2 - \|\mathbf{a}\|^2 - \|\mathbf{b}\|^2)$ we have:

$$\begin{aligned}
0 &\leq \|\mathbf{y} - \mathbf{y}^{t-1} - \eta_{\mathbf{y}}\nabla_{\mathbf{y}} F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}) - \eta_{\mathbf{y}}\varepsilon_{t-1}\|^2 - \|\mathbf{y} - \mathbf{y}^t\|^2 - \|\mathbf{y}^t - \mathbf{y}^{t-1} - \eta_{\mathbf{y}}\nabla_{\mathbf{y}} F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}) - \eta_{\mathbf{y}}\varepsilon_{t-1}\|^2 \\
&= \|\mathbf{y} - \mathbf{y}^{t-1}\|^2 - \|\mathbf{y} - \mathbf{y}^t\|^2 - \|\mathbf{y}^t - \mathbf{y}^{t-1}\|^2 + 2\langle \mathbf{y}^{t-1} - \mathbf{y}, \eta_{\mathbf{y}}\nabla_{\mathbf{y}} F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\rangle \\
&\quad + 2\langle \mathbf{y}^t - \mathbf{y}^{t-1}, \eta_{\mathbf{y}}\nabla_{\mathbf{y}} F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\rangle - 2\eta_{\mathbf{y}}\langle \mathbf{y} - \mathbf{y}^{t-1}, \varepsilon_{t-1}\rangle + 2\eta_{\mathbf{y}}\langle \mathbf{y}^t - \mathbf{y}^{t-1}, \varepsilon_{t-1}\rangle.
\end{aligned}$$

Due to L smoothness and concavity, we have

$$\begin{aligned}
&2\langle \mathbf{y}^{t-1} - \mathbf{y}, \eta_{\mathbf{y}}\nabla_{\mathbf{y}} F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\rangle + 2\langle \mathbf{y}^t - \mathbf{y}^{t-1}, \eta_{\mathbf{y}}\nabla_{\mathbf{y}} F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\rangle \\
\leq &2\eta_{\mathbf{y}}(f(\mathbf{x}^{t-1}, \mathbf{y}^t) - f(\mathbf{x}^{t-1}, \mathbf{y})) + \eta_{\mathbf{y}} L\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2.
\end{aligned}$$

So we have:

$$
\begin{aligned}
f(\mathbf{x}^{t-1}, \mathbf{y}) - f(\mathbf{x}^{t-1}, \mathbf{y}^t) \leq& \frac{1}{2\eta_{\mathbf{y}}}(\|\mathbf{y} - \mathbf{y}^{t-1}\|^2 - \|\mathbf{y} - \mathbf{y}^t\|^2) - \left(\frac{1}{2\eta_{\mathbf{y}}} - \frac{L}{2}\right)\|\mathbf{y}^t - \mathbf{y}^{t-1}\|^2 \\
&- \langle \mathbf{y} - \mathbf{y}^{t-1}, \varepsilon_{t-1}\rangle + \langle \mathbf{y}^t - \mathbf{y}^{t-1}, \varepsilon_{t-1}\rangle \\
\leq& \frac{1}{2\eta_{\mathbf{y}}}(\|\mathbf{y} - \mathbf{y}^{t-1}\|^2 - \|\mathbf{y} - \mathbf{y}^t\|^2) - \left(\frac{1}{2\eta_{\mathbf{y}}} - \frac{L}{2}\right)\|\mathbf{y}^t - \mathbf{y}^{t-1}\|^2 \\
&- \langle \mathbf{y} - \mathbf{y}^{t-1}, \nabla_2 f(\mathbf{z}^t, \mathbf{y}^{t-1}; \zeta^{t-1}) - \nabla_2 f(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\rangle \\
&+ \langle \mathbf{y}^t - \mathbf{y}^{t-1}, \nabla_2 f(\mathbf{z}^t, \mathbf{y}^{t-1}; \zeta^{t-1}) - \nabla_2 f(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\rangle \\
\leq& \frac{1}{2\eta_{\mathbf{y}}}(\|\mathbf{y} - \mathbf{y}^{t-1}\|^2 - \|\mathbf{y} - \mathbf{y}^t\|^2) - \left(\frac{1}{2\eta_{\mathbf{y}}} - \frac{L}{2}\right)\|\mathbf{y}^t - \mathbf{y}^{t-1}\|^2 \\
&- \langle \mathbf{y} - \mathbf{y}^{t-1}, \nabla_2 f(\mathbf{z}^t, \mathbf{y}^{t-1}; \zeta^{t-1}) - \nabla_2 f(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\rangle \\
&+ \langle \mathbf{y}^t - \mathbf{y}^{t-1}, \nabla_2 f(\mathbf{z}^t, \mathbf{y}^{t-1}; \zeta^{t-1}) - \nabla_2 f(\mathbf{z}^t, \mathbf{y}^{t-1})\rangle \\
&+ \langle \mathbf{y}^t - \mathbf{y}^{t-1}, \nabla_2 f(\mathbf{z}^t, \mathbf{y}^{t-1}) - \nabla_2 f(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\rangle.
\end{aligned}
$$

Now we take expectation over the randomness of ζ^{t-1}

$$
\begin{aligned}
\mathbb{E}[f(\mathbf{x}^{t-1}, \mathbf{y}) - f(\mathbf{x}^{t-1}, \mathbf{y}^t)] \leq& \frac{1}{2\eta_{\mathbf{y}}}(\mathbb{E}\|\mathbf{y} - \mathbf{y}^{t-1}\|^2 - \mathbb{E}\|\mathbf{y} - \mathbf{y}^t\|^2) - \left(\frac{1}{2\eta_{\mathbf{y}}} - \frac{L}{2}\right)\mathbb{E}\|\mathbf{y}^t - \mathbf{y}^{t-1}\|^2 \\
&- \mathbb{E}\langle \mathbf{y} - \mathbf{y}^{t-1}, \nabla_2 f(\mathbf{z}^t, \mathbf{y}^{t-1}) - \nabla_2 f(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\rangle \\
&+ \left(\frac{1}{4\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 + \eta_{\mathbf{y}}\mathbb{E}\left\|\nabla_2 f(\mathbf{z}^t, \mathbf{y}^{t-1}; \zeta^{t-1}) - \nabla_2 f(\mathbf{z}^t, \mathbf{y}^{t-1})\right\|^2\right) \\
&+ \mathbb{E}\langle \mathbf{y}^t - \mathbf{y}^{t-1}, \nabla_2 f(\mathbf{z}^t, \mathbf{y}^{t-1}) - \nabla_2 f(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\rangle \\
\leq& \frac{1}{2\eta_{\mathbf{y}}}(\mathbb{E}\|\mathbf{y} - \mathbf{y}^{t-1}\|^2 - \mathbb{E}\|\mathbf{y} - \mathbf{y}^t\|^2) - \left(\frac{1}{4\eta_{\mathbf{y}}} - \frac{L}{2}\right)\mathbb{E}\|\mathbf{y}^t - \mathbf{y}^{t-1}\|^2 + \eta_{\mathbf{y}}\sigma^2 \\
&+ L_f D_{\mathcal{Y}}\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|.
\end{aligned}
$$

□

Lemma 6. *Define potential function:*

$$
\Psi^t = \Phi_{1/2L}(\mathbf{x}^t) + C_1\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\right\|^2 + C_2\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\right\| \tag{4}
$$

where $C_1 = \frac{\eta_{\mathbf{x}}}{2(2\beta-\beta^2)}G_g^2 L_f^2$ *and* $C_2 = \frac{2\eta_{\mathbf{x}}L^2 D_{\mathbf{y}}}{\beta}$. *Then the following statement holds:*

$$
\begin{aligned}
\mathbb{E}[\Psi^t - \Psi^{t-1}] \leq& 2\eta_{\mathbf{x}}L\left(\frac{\mathbb{E}\|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^{t-1}\|^2 - \mathbb{E}\|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^t\|^2}{2\eta_{\mathbf{y}}} + \eta_{\mathbf{y}}\frac{\sigma^2}{M} + \mathbb{E}[F(\mathbf{x}^t, \mathbf{y}^t) - F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})]\right) \\
&- \frac{\eta_{\mathbf{x}}}{8}\mathbb{E}\|\nabla\Phi_{1/2L}(\mathbf{x}^{t-1})\|^2 + 2L\eta_{\mathbf{x}}^2(G_h^2 + G_f^2 G_g^2) + 4\eta_{\mathbf{x}}L(2t - 1 - 2s)\eta_{\mathbf{x}}(G_h + G_f G_g) \\
&+ C_1\left(8(1-\beta)^2 G_g^2\eta_{\mathbf{x}}^2(G_h^2 + G_f^2 G_g^2) + 2\beta^2\frac{\sigma^2}{B}\right) + C_2\sqrt{8(1-\beta)^2 G_g^2\eta_{\mathbf{x}}^2(G_h^2 + G_f^2 G_g^2) + 2\beta^2\frac{\sigma^2}{B}}.
\end{aligned}
$$

Proof. Recalling Lemma 4, we have

$$
\begin{aligned}
\mathbb{E}[\Phi_{1/2L}(\mathbf{x}^t)] \leq& \mathbb{E}[\Phi_{1/2L}(\mathbf{x}^{t-1})] + 2\eta_{\mathbf{x}}L\mathbb{E}\left(\Phi(\mathbf{x}^{t-1}) - F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\right) - \frac{\eta_{\mathbf{x}}}{8}\mathbb{E}\|\nabla\Phi_{1/2L}(\mathbf{x}^{t-1})\|^2 \\
&+ 2L\eta_{\mathbf{x}}^2(G_h^2 + G_f^2 G_g^2) + \frac{\eta_{\mathbf{x}}}{2}G_g^2 L_f^2\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2.
\end{aligned}
$$

Following Lin et al. [2019], we split $\mathbb{E}\left(\Phi(\mathbf{x}^{t-1}) - F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\right)$ as follows:

$$
\begin{aligned}
&\mathbb{E}\left(\Phi(\mathbf{x}^{t-1}) - F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\right) \\
=&\mathbb{E}\Bigg[\underbrace{F(\mathbf{x}^{t-1}, \mathbf{y}^*(\mathbf{x}^{t-1})) - F(\mathbf{x}^{t-1}, \mathbf{y}^*(\mathbf{x}^s))}_{A} + \underbrace{F(\mathbf{x}^{t-1}, \mathbf{y}^*(\mathbf{x}^s)) - F(\mathbf{x}^{t-1}, \mathbf{y}^t)}_{B} + \underbrace{F(\mathbf{x}^{t-1}, \mathbf{y}^t) - F(\mathbf{x}^t, \mathbf{y}^t)}_{C}\Bigg] \\
&+ \mathbb{E}\left[F(\mathbf{x}^t, \mathbf{y}^t) - F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\right]
\end{aligned}
$$

for some iteration s.

For A:

$$
\begin{aligned}
F(\mathbf{x}^{t-1}, \mathbf{y}^*(\mathbf{x}^{t-1})) - F(\mathbf{x}^{t-1}, \mathbf{y}^*(\mathbf{x}^s)) &\leq F(\mathbf{x}^{t-1}, \mathbf{y}^*(\mathbf{x}^{t-1})) - F(\mathbf{x}^s, \mathbf{y}^*(\mathbf{x}^{t-1})) + F(\mathbf{x}^s, \mathbf{y}^*(\mathbf{x}^s)) - F(\mathbf{x}^{t-1}, \mathbf{y}^*(\mathbf{x}^s)) \\
&\leq 2(t-1-s)\eta_{\mathbf{x}}\sqrt{2G_h^2 + 2G_f^2 G_g^2}.
\end{aligned}
$$

For B, we evoke Lemma 5:

$$
\mathbb{E}[F(\mathbf{x}^{t-1}, \mathbf{y}^*(\mathbf{x}^s)) - F(\mathbf{x}^{t-1}, \mathbf{y}^t)] \leq \frac{1}{2\eta_{\mathbf{y}}}(\mathbb{E}\|\mathbf{y} - \mathbf{y}^{t-1}\|^2 - \mathbb{E}\|\mathbf{y} - \mathbf{y}^t\|^2) + \eta_{\mathbf{y}}\frac{\sigma^2}{M} + LD_{\mathcal{Y}}\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|.
$$

For C:

$$
F(\mathbf{x}^{t-1}, \mathbf{y}^t) - F(\mathbf{x}^t, \mathbf{y}^t) \leq \eta_{\mathbf{x}}\sqrt{2G_h^2 + 2G_f^2 G_g^2}.
$$

Putting pieces together yields:

$$
\begin{aligned}
&\mathbb{E}\left(\Phi(\mathbf{x}^{t-1}) - F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\right) \\
\leq&(2t-1-2s)\eta_{\mathbf{x}}\sqrt{2G_h^2 + 2G_f^2 G_g^2} + \frac{1}{2\eta_{\mathbf{y}}}(\mathbb{E}\|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^{t-1}\|^2 - \mathbb{E}\|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^t\|^2) + \eta_{\mathbf{y}}\frac{\sigma^2}{M} + LD_{\mathcal{Y}}\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\| \\
&+ \mathbb{E}\left[F(\mathbf{x}^t, \mathbf{y}^t) - F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\right].
\end{aligned}
$$

Notice that $\sqrt{2G_h^2 + 2G_f^2 G_g^2} \leq 2(G_h + G_f G_g)$ Plugging above bound back to Lemma 4 yields:

$$
\begin{aligned}
&\mathbb{E}[\Phi_{1/2L}(\mathbf{x}^t)] \leq \mathbb{E}[\Phi_{1/2L}(\mathbf{x}^{t-1})] \\
&+ 2\eta_{\mathbf{x}}L\mathbb{E}\Bigg((2t-1-2s)\eta_{\mathbf{x}}2(G_h + G_f G_g) + \frac{\mathbb{E}\|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^{t-1}\|^2 - \mathbb{E}\|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^t\|^2}{2\eta_{\mathbf{y}}} + \eta_{\mathbf{y}}\frac{\sigma^2}{B} + LD_{\mathcal{Y}}\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\| \\
&\qquad + \mathbb{E}\left[F(\mathbf{x}^t, \mathbf{y}^t) - F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\right]\Bigg) \\
&- \frac{\eta_{\mathbf{x}}}{8}\mathbb{E}\|\nabla\Phi_{1/2L}(\mathbf{x}^{t-1})\|^2 + 2L\eta_{\mathbf{x}}^2(G_h^2 + G_f^2 G_g^2) + \frac{\eta_{\mathbf{x}}}{2}G_g^2 L_f^2\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2 \\
&\leq \mathbb{E}[\Phi_{1/2L}(\mathbf{x}^{t-1})] + \frac{\eta_{\mathbf{x}}}{2}G_g^2 L_f^2\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2 + 2\eta_{\mathbf{x}}L^2 D_{\mathcal{Y}}\sqrt{\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2} \\
&+ 2\eta_{\mathbf{x}}L\mathbb{E}\Bigg((2t-1-2s)\eta_{\mathbf{x}}2(G_h + G_f G_g) + \frac{\|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^{t-1}\|^2 - \|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^t\|^2}{2\eta_{\mathbf{y}}} + \eta_{\mathbf{y}}\frac{\sigma^2}{M} + F(\mathbf{x}^t, \mathbf{y}^t) - F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\Bigg) \\
&- \frac{\eta_{\mathbf{x}}}{8}\mathbb{E}\|\nabla\Phi_{1/2L}(\mathbf{x}^{t-1})\|^2 + 2L\eta_{\mathbf{x}}^2(G_h^2 + G_f^2 G_g^2).
\end{aligned}
$$

Recalling the definition of Ψ^t

$$
\begin{aligned}
&\mathbb{E}[\Psi^t - \Psi^{t-1}] = \mathbb{E}[\Phi_{1/2L}(\mathbf{x}^t) - \Phi_{1/2L}(\mathbf{x}^{t-1})] \\
&\quad + C_1\mathbb{E}\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\|^2 + C_2\sqrt{\mathbb{E}\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\|^2} - \left(C_1\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2 + C_2\sqrt{\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2}\right) \\
&\leq \frac{\eta_{\mathbf{x}}}{2} G_g^2 L_f^2 \mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2 + 2\eta_{\mathbf{x}} L^2 D_{\mathcal{Y}}\sqrt{\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2} \\
&+ 2\eta_{\mathbf{x}} L\mathbb{E}\left((2t-1-2s)\eta_{\mathbf{x}} 2(G_h + G_f G_g) + \frac{\|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^{t-1}\|^2 - \|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^t\|^2}{2\eta_{\mathbf{y}}} + \eta_{\mathbf{y}}\frac{\sigma^2}{M} + F(\mathbf{x}^t, \mathbf{y}^t) - F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\right) \\
&\quad - \frac{\eta_{\mathbf{x}}}{8}\mathbb{E}\|\nabla\Phi_{1/2L}(\mathbf{x}^{t-1})\|^2 + 2L\eta_{\mathbf{x}}^2(G_h^2 + G_f^2 G_g^2) \\
&\quad + C_1\mathbb{E}\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\|^2 + C_2\sqrt{\mathbb{E}\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\|^2} - \left(C_1\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2 + C_2\sqrt{\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2}\right) \\
&\leq \left(\frac{\eta_{\mathbf{x}}}{2} G_g^2 L_f^2 - C_1\right)\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2 + \left(2\eta_{\mathbf{x}} L^2 D_{\mathcal{Y}} - C_2\right)\sqrt{\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2} \\
&+ 2\eta_{\mathbf{x}} L\mathbb{E}\left((2t-1-2s)\eta_{\mathbf{x}} 2(G_h + G_f G_g) + \frac{\|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^{t-1}\|^2 - \|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^t\|^2}{2\eta_{\mathbf{y}}} + \eta_{\mathbf{y}}\frac{\sigma^2}{M} + F(\mathbf{x}^t, \mathbf{y}^t) - F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\right) \\
&\quad - \frac{\eta_{\mathbf{x}}}{8}\mathbb{E}\|\nabla\Phi_{1/2L}(\mathbf{x}^{t-1})\|^2 + 2L\eta_{\mathbf{x}}^2(G_h^2 + G_f^2 G_g^2) + C_1\mathbb{E}\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\|^2 + C_2\sqrt{\mathbb{E}\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\|^2}.
\end{aligned}
$$

Evoking Lemma 1 to replace $\mathbb{E}\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\|^2$ yields:

$$
\begin{aligned}
&\mathbb{E}[\Psi^t - \Psi^{t-1}] \leq \left(\frac{\eta_{\mathbf{x}}}{2} G_g^2 L_f^2 - C_1\right)\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2 + \left(2\eta_{\mathbf{x}} L^2 D_{\mathcal{Y}} - C_2\right)\sqrt{\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2} \\
&+ 2\eta_{\mathbf{x}} L\mathbb{E}\left((2t-1-2s)\eta_{\mathbf{x}} 2(G_h + G_f G_g) + \frac{\|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^{t-1}\|^2 - \|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^t\|^2}{2\eta_{\mathbf{y}}} + \eta_{\mathbf{y}}\frac{\sigma^2}{M} + F(\mathbf{x}^t, \mathbf{y}^t) - F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\right) \\
&\quad - \frac{\eta_{\mathbf{x}}}{8}\mathbb{E}\|\nabla\Phi_{1/2L}(\mathbf{x}^{t-1})\|^2 + 2L\eta_{\mathbf{x}}^2(G_h^2 + G_f^2 G_g^2) \\
&\quad + C_1\left((1-\beta)^2\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 + 4(1-\beta)^2 G_g^2\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 + 2\beta^2\frac{\sigma^2}{B}\right) \\
&\quad + C_2\sqrt{(1-\beta)^2\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 + 4(1-\beta)^2 G_g^2\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 + 2\beta^2\frac{\sigma^2}{B}} \\
&\leq \underbrace{\left(\frac{\eta_{\mathbf{x}}}{2} G_g^2 L_f^2 - C_1 + (1-\beta)^2 C_1\right)}_{\clubsuit}\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2 + \underbrace{\left(2\eta_{\mathbf{x}} L^2 D_{\mathcal{Y}} - C_2 + (1-\beta)C_2\right)}_{\heartsuit}\sqrt{\mathbb{E}\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\|^2} \\
&+ 2\eta_{\mathbf{x}} L\mathbb{E}\left((2t-1-2s)\eta_{\mathbf{x}} 2(G_h + G_f G_g) + \frac{\|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^{t-1}\|^2 - \|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^t\|^2}{2\eta_{\mathbf{y}}} + \eta_{\mathbf{y}}\frac{\sigma^2}{B} + F(\mathbf{x}^t, \mathbf{y}^t) - F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\right) \\
&\quad - \frac{\eta_{\mathbf{x}}}{8}\mathbb{E}\|\nabla\Phi_{1/2L}(\mathbf{x}^{t-1})\|^2 + 2L\eta_{\mathbf{x}}^2(G_h^2 + G_f^2 G_g^2) \\
&\quad + C_1\left(4(1-\beta)^2 G_g^2\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 + 2\beta^2\frac{\sigma^2}{B}\right) + C_2\sqrt{4(1-\beta)^2 G_g^2\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 + 2\beta^2\frac{\sigma^2}{B}}
\end{aligned}
$$

Since we choose $C_1 = \frac{\eta_{\mathbf{x}}}{2(2\beta - \beta^2)} G_g^2 L_f^2$ and $C_2 = \frac{2\eta_{\mathbf{x}} L^2 D_{\mathbf{y}}}{\beta}$, we know $\clubsuit = 0$ and $\heartsuit = 0$. Hence we have

$$
\begin{aligned}
&\mathbb{E}[\Psi^t - \Psi^{t-1}] \\
&\leq 2\eta_{\mathbf{x}} L\left(\frac{\mathbb{E}\|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^{t-1}\|^2 - \mathbb{E}\|\mathbf{y}^*(\mathbf{x}^s) - \mathbf{y}^t\|^2}{2\eta_{\mathbf{y}}} + \eta_{\mathbf{y}}\frac{\sigma^2}{B} + \mathbb{E}[F(\mathbf{x}^t, \mathbf{y}^t) - F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})]\right) \\
&\quad - \frac{\eta_{\mathbf{x}}}{8}\mathbb{E}\|\nabla\Phi_{1/2L}(\mathbf{x}^{t-1})\|^2 + 2L\eta_{\mathbf{x}}^2(G_h^2 + G_f^2 G_g^2) + 2\eta_{\mathbf{x}} L(2t-1-2s)\eta_{\mathbf{x}}(G_h + G_f G_g) \\
&\quad + C_1\left(4(1-\beta)^2 G_g^2 2\eta_{\mathbf{x}}^2(G_h^2 + G_f^2 G_g^2) + 2\beta^2\frac{\sigma^2}{B}\right) + C_2\sqrt{4(1-\beta)^2 G_g^2 2\eta_{\mathbf{x}}^2(G_h^2 + G_f^2 G_g^2) + 2\beta^2\frac{\sigma^2}{B}}
\end{aligned}
$$

which concludes the proof. $\square$

C.2.1 Proof of Theorem 2

We first evoke Lemma 6:

$$\begin{aligned}
\mathbb{E}\|\nabla\Phi_{1/2L}(\mathbf{x}^{t-1})\|^2 \leq& \frac{8\mathbb{E}[\Psi^{t-1}-\Psi^t]}{\eta_{\mathbf{x}}}\\
&+16L\left(\frac{\mathbb{E}\|\mathbf{y}^*(\mathbf{x}^s)-\mathbf{y}^{t-1}\|^2-\mathbb{E}\|\mathbf{y}^*(\mathbf{x}^s)-\mathbf{y}^t\|^2}{2\eta_{\mathbf{y}}}+\mathbb{E}[F(\mathbf{x}^t,\mathbf{y}^t)-F(\mathbf{x}^{t-1},\mathbf{y}^{t-1})]\right)\\
&+16L\eta_{\mathbf{x}}(G_h^2+G_f^2G_g^2)+16L(2t-1-2s)\eta_{\mathbf{x}}(G_h+G_fG_g)+16\eta_{\mathbf{y}}L\frac{\sigma^2}{B}\\
&+\frac{8C_1}{\eta_{\mathbf{x}}}\left(8(1-\beta)^2G_g^2\eta_{\mathbf{x}}^2(G_h^2+G_f^2G_g^2)+2\beta^2\frac{\sigma^2}{B}\right)+\frac{8C_2}{\eta_{\mathbf{x}}}\sqrt{8(1-\beta)^2G_g^2\eta_{\mathbf{x}}^2(G_h^2+G_f^2G_g^2)+2\beta^2\frac{\sigma^2}{B}}.
\end{aligned}$$

We sum above inequality for $t=jS$ to $(j+1)S-1$, with $\mathbf{x}^s=\mathbf{x}^{jS}$:

$$\begin{aligned}
&\sum_{t=jS}^{(j+1)S-1}\mathbb{E}\|\nabla\Phi_{1/2L}(\mathbf{x}^t)\|^2\\
\leq&\sum_{t=jS}^{(j+1)S-1}\left[\frac{8\mathbb{E}[\Psi^t-\Psi^{t+1}]}{\eta_{\mathbf{x}}}+16L\left(\frac{\mathbb{E}\|\mathbf{y}^*(\mathbf{x}^{jS})-\mathbf{y}^t\|^2-\mathbb{E}\|\mathbf{y}^*(\mathbf{x}^{jS})-\mathbf{y}^{t+1}\|^2}{2\eta_{\mathbf{y}}}+\mathbb{E}[F(\mathbf{x}^{t+1},\mathbf{y}^{t+1})-F(\mathbf{x}^t,\mathbf{y}^t)]\right)\right]\\
&+16SL\eta_{\mathbf{x}}(G_h^2+G_f^2G_g^2)+\sum_{t=jS}^{(j+1)S-1}16L(2t+1-2jS)\eta_{\mathbf{x}}(G_h+G_fG_g)+16S\eta_{\mathbf{y}}L\frac{\sigma^2}{B}\\
&+8SC_1\left(8(1-\beta)^2G_f^2\eta_{\mathbf{x}}^2(G_h^2+G_f^2G_g^2)+2\beta^2\frac{\sigma^2}{B}\right)+8SC_2\sqrt{8(1-\beta)^2G_f^2\eta_{\mathbf{x}}^2(G_h^2+G_f^2G_g^2)+2\beta^2\frac{\sigma^2}{B}}\\
\leq&\frac{8\mathbb{E}[\Psi^{jS}-\Psi^{(j+1)S)}]}{\eta_{\mathbf{x}}}\\
&+16L\left(\frac{\mathbb{E}\|\mathbf{y}^*(\mathbf{x}^{jS})-\mathbf{y}^{jS}\|^2-\mathbb{E}\|\mathbf{y}^*(\mathbf{x}^{jS})-\mathbf{y}^{(j+1)S}\|^2}{2\eta_{\mathbf{y}}}+\mathbb{E}[F(\mathbf{x}^{(j+1)S},\mathbf{y}^{(j+1)S})-F(\mathbf{x}^{jS},\mathbf{y}^{jS})]\right)\\
&+16SL\eta_{\mathbf{x}}(G_h^2+G_f^2G_g^2)+32S^2L\eta_{\mathbf{x}}(G_h+G_fG_g)+16S\eta_{\mathbf{y}}L\frac{\sigma^2}{B}\\
&+8S\frac{C_1}{\eta_{\mathbf{x}}}\left(8(1-\beta)^2G_f^2\eta_{\mathbf{x}}^2(G_h^2+G_f^2G_g^2)+2\beta^2\frac{\sigma^2}{B}\right)+8S\frac{C_2}{\eta_{\mathbf{x}}}\sqrt{8(1-\beta)^2G_f^2\eta_{\mathbf{x}}^2(G_h^2+G_f^2G_g^2)+2\beta^2\frac{\sigma^2}{B}}.
\end{aligned}$$

Now we further sum over $j=0$ to $(T+1)/S-1$

$$\begin{aligned}
\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla\Phi_{1/2L}(\mathbf{x}^t)\|^2=&\frac{1}{T+1}\sum_{j=0}^{(T+1)/S-1}\sum_{t=jS}^{(j+1)S-1}\mathbb{E}\|\nabla\Phi_{1/2L}(\mathbf{x}^t)\|^2\\
\leq&\frac{1}{T+1}\frac{8\mathbb{E}[\Psi^0-\Psi^T]}{\eta_{\mathbf{x}}}+16L\left(\frac{D_{\mathcal{Y}}^2}{2S\eta_{\mathbf{y}}}+\frac{\mathbb{E}[F(\mathbf{x}^T,\mathbf{y}^T)-F(\mathbf{x}^0,\mathbf{y}^0)]}{T+1}\right)\\
&+16L\eta_{\mathbf{x}}(G_h^2+G_f^2G_g^2)+32SL\eta_{\mathbf{x}}(G_h+G_fG_g)+16\eta_{\mathbf{y}}L\frac{\sigma^2}{B}\\
&+8\frac{C_1}{\eta_{\mathbf{x}}}\left(8(1-\beta)^2G_g^2\eta_{\mathbf{x}}^2(G_h^2+G_f^2G_g^2)+2\beta^2\frac{\sigma^2}{M}\right)+8\frac{C_2}{\eta_{\mathbf{x}}}\sqrt{8(1-\beta)^2G_g^2\eta_{\mathbf{x}}^2(G_h^2+G_f^2G_g^2)+2\beta^2\frac{\sigma^2}{M}},
\end{aligned}$$

where $C_1=\frac{\eta_{\mathbf{x}}}{2(2\beta-\beta^2)}G_g^2L_f^2$ and $C_2=\frac{2\eta_{\mathbf{x}}L^2D_{\mathbf{y}}}{\beta}$. Notice that $\mathbb{E}[F(\mathbf{x}^T,\mathbf{y}^T)-F(\mathbf{x}^0,\mathbf{y}^0)]\leq\mathbb{E}[F(\mathbf{x}^T,\mathbf{y}^T)-F(\mathbf{x}^0,\mathbf{y}^T)+F(\mathbf{x}^0,\mathbf{y}^T)-F(\mathbf{x}^0,\mathbf{y}^0)]\leq T\eta_{\mathbf{x}}(G_h+G_fG_g)+\Delta_\Phi$.

Now, We choose $S=\sqrt{\frac{D_{\mathcal{Y}}^2}{4\eta_{\mathbf{x}}\eta_{\mathbf{y}}(G_h+G_fG_g)}}$, $\eta_{\mathbf{y}}=\Theta\left(\frac{\epsilon^2}{L\sigma^2}\right)$ and $\eta_{\mathbf{x}}=\Theta\left(\frac{\epsilon^6}{LD_{\mathcal{Y}}\sigma^2(G_h+G_fG_g)}\right)$, $B=\Theta(1)$, $M=$

$\Theta\left(\frac{\sigma^2 L^4 D_{\mathcal{Y}}^2}{\epsilon^4}\right)$ and $T = \Theta(\max\left\{\frac{\Delta_{\hat{\Phi}}}{\eta_{\mathbf{x}}\epsilon^2}, \frac{L\Delta_{\Phi}}{\epsilon^2}\right\})$, which yields total gradient complexity of

$$O\left(\max\left\{\frac{L^3\sigma^2 D_{\mathcal{Y}}^2(G_h + G_f G_g)\Delta_{\hat{\Phi}}}{\epsilon^8}, \frac{L\Delta_{\Phi}}{\epsilon^2}\right\}\right).$$

D Proof of Dual Composition Setting

In this section we provide the proof of results in primal compostion setting (Theorem 3 and Theorem 4).

Proof Sketch: Before delving into the detailed proof, we first provide a sketch. We construct the following two-level potential function:

$$\begin{aligned}\hat{F}^{t+1} :=& F(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}) + s^{t+1} - \left(\frac{1}{4\eta_{\mathbf{y}}} + 4L_g^2 G_f^2 G_g^2 \eta_{\mathbf{y}} + \frac{\eta_{\mathbf{x}} L^2}{2} + \frac{96 G_g^2 L_f^2}{\mu^2 \eta_{\mathbf{x}} \beta} + \frac{48 G_g^2 L_f^2 \beta}{\mu^2 \eta_{\mathbf{x}}}\right)\left\|\mathbf{y}^{t+1} - \mathbf{y}^t\right\|^2 \\ &- \left(\frac{1}{8\eta_{\mathbf{y}}} + \frac{48 G_g^2 L_f^2}{\mu^2 \eta_{\mathbf{x}} \beta}\right)\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 + \left(\frac{7}{2\eta_{\mathbf{x}}} + \mu - \frac{\eta_{\mathbf{x}} L^2}{2} - \frac{2L_f^2}{\mu}\right)\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2\end{aligned}$$

where $s^{t+1} := -\frac{2}{\eta_{\mathbf{x}}^2 \mu}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2$ and

$$\tilde{F}^{t+1} := \hat{F}^{t+1} - \mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{y}^t)\right\|^2 - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{4L_f^2}{\mu^2 \eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2.$$

$\mathbb{E}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2$ is the iterate difference of our auxilliary variables, and it can be controlled by

$$\begin{aligned}\mathbb{E}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2 \leq& \left(1 - \frac{\beta}{2}\right)^2 \mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2 + 4\left(1 + \frac{2}{\beta}\right) G_g^2 \left(\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 + \left\|\mathbf{y}^{t-1} - \mathbf{y}^{t-2}\right\|^2\right) \\ &+ 2\left(1 + \frac{2}{\beta}\right)\beta^2 G_g^2 \mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2.\end{aligned}$$

Putting pieces together we can derive the following descent property:

$$\begin{aligned}\mathbb{E}[\tilde{F}^{t+1} - \tilde{F}^t] \geq& \min\left\{\frac{1}{\eta_{\mathbf{y}}}, \frac{1}{\eta_{\mathbf{x}}}\right\} \cdot \mathbb{E}\left\|\nabla G(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 \\ &- O\left(\eta_{\mathbf{y}}(G_g^2 + G_f^2) + \eta_{\mathbf{x}}\right)\frac{\sigma^2}{B} - O\left(\beta^2 L_f^2 G_g^2 \eta_{\mathbf{y}} + \beta^2 \eta_{\mathbf{y}} L_f^2 G_g^2 + \beta^2 \eta_{\mathbf{x}} L_f^2\right)\frac{\sigma^2}{M}.\end{aligned}$$

Unrolling above recursion will give desired rate. The converegcne of merely-convex-nonconcave setting follows similar technique.

D.1 Proof of Strongly-convex-nonconcave setting

In nonconcave setting, besides the Lemma 1, we also need the following bound on tracking error between two iterates.

Lemma 7 (Convergence of Iterates difference)**.** *For Algorithm 2, under the assumptions of Theorem 3, the following statement holds true:*

$$\begin{aligned}\mathbb{E}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2 \leq& \left(1 - \frac{\beta}{2}\right)^2 \mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2 + 4\left(1 + \frac{2}{\beta}\right) G_g^2 \left(\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 + \left\|\mathbf{y}^{t-1} - \mathbf{y}^{t-2}\right\|^2\right) \\ &+ \left(1 + \frac{2}{\beta}\right)\beta^2\left(3G_g^2 \mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 + 6\frac{\sigma^2}{M}\right).\end{aligned}$$

Proof. For the ease of presentation, we define the following two auxiliary variables:

$$g^t = \frac{1}{B}\sum_{\xi \in \mathcal{M}^t} g(\mathbf{y}^t; \xi),\ g^{t \mapsto t-1} = \frac{1}{M}\sum_{\xi \in \mathcal{M}^t}\left(g(\mathbf{y}^t; \xi) - g(\mathbf{y}^{t-1}; \xi)\right).$$

According to updating rule of $\mathbf{z}$, we have:

$$\mathbf{z}^{t+1} - \mathbf{z}^t = (1-\beta)(\mathbf{z}^t - \mathbf{z}^{t-1}) + (1-\beta)(g^{t\mapsto t-1} - g^{t-1\mapsto t-2}) + \beta(g^t - g^{t-1}).$$

Taking expectation w.r.t. $\mathcal{M}^t$, and $\mathcal{M}^{t-1}$ yields:

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2 &= \mathbb{E}\left\|(1-\beta)(\mathbf{z}^t - \mathbf{z}^{t-1}) + (1-\beta)(g^{t\mapsto t-1} - g^{t-1\mapsto t-2}) + \beta(g^t - g^{t-1})\right\|^2 \\
&\le \left(1+\frac{\beta}{2-2\beta}\right)(1-\beta)^2\mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1} + (g^{t\mapsto t-1} - g^{t-1\mapsto t-2})\right\|^2 + \left(1+\frac{2-2\beta}{\beta}\right)\left\|\beta(g^t - g^{t-1})\right\|^2 \\
&\le \left(1-\frac{\beta}{2}\right)(1-\beta)\mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1} + (g^{t\mapsto t-1} - g^{t-1\mapsto t-2})\right\|^2 + \left(1+\frac{2}{\beta}\right)\beta^2\mathbb{E}\left\|g^t - g^{t-1}\right\|^2 \\
&\le \left(1-\frac{\beta}{2}\right)(1-\beta)\left(1+\frac{\beta}{2-2\beta}\right)\mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2 \\
&\quad + \left(1-\frac{\beta}{2}\right)(1-\beta)\left(1+\frac{2-2\beta}{\beta}\right)\mathbb{E}\left\|g^{t\mapsto t-1} - g^{t-1\mapsto t-2}\right\|^2 + \left(1+\frac{2}{\beta}\right)\beta^2\mathbb{E}\left\|g^t - g^{t-1}\right\|^2 \\
&\le \left(1-\frac{\beta}{2}\right)^2\mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2 + \underbrace{\left(1+\frac{2}{\beta}\right)\mathbb{E}\left\|g^{t\mapsto t-1} - g^{t-1\mapsto t-2}\right\|^2}_{T_1} + \underbrace{\left(1+\frac{2}{\beta}\right)\beta^2\mathbb{E}\left\|g^t - g^{t-1}\right\|^2}_{T_2}
\end{aligned}$$

where in the first and third steps we use Young's inequality that $\|\mathbf{a}+\mathbf{b}\|^2 \le (1+a)\|\mathbf{a}\|^2 + (1+\frac{1}{a})\|\mathbf{b}\|^2$. Now we bound T_1 as follows:

$$\begin{aligned}
T_1 \le& 2\left(1+\frac{2}{\beta}\right)\left(\frac{1}{M}\sum_{\xi^t\in\mathcal{M}^t}\mathbb{E}\left\|g(\mathbf{y}^t;\xi^t) - g(\mathbf{y}^{t-1};\xi^t)\right\|^2\right) \\
&+ 2\left(1+\frac{2}{\beta}\right)\left(\frac{1}{M}\sum_{\xi^{t-1}\in\mathcal{M}^{t-1}}\mathbb{E}\left\|g(\mathbf{y}^{t-1};\xi^{t-1}) - g(\mathbf{y}^{t-2};\xi^{t-1})\right\|^2\right) \\
\le& 4\left(1+\frac{2}{\beta}\right)G_g^2\left(\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 + \mathbb{E}\left\|\mathbf{y}^{t-1} - \mathbf{y}^{t-2}\right\|^2\right),
\end{aligned}$$

and for T_2:

$$T_2 \le \left(1+\frac{2}{\beta}\right)\beta^2\left(3G_g^2\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 + 6\frac{\sigma^2}{M}\right).$$

Putting pieces together will conclude the proof. □

The following Lemma establishes the difference of the exact gradient computed on F versus the gradients we actually used in Algorithm 2.

Lemma 8 (Gradient difference)**.** *For Algorithm 2, under the assumptions of Theorem 3, the following statement holds true:*

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{g}_{\mathbf{y}}^t - \nabla_{\mathbf{y}}F(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 &\le \frac{4(G_f^2+G_g^2)\sigma^2}{B} + 2G_g^2L_f^2\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{y}^t)\right\|^2. \\
\mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^t - \nabla_{\mathbf{x}}F(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 &\le 2L_f^2\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{y}^t)\right\|^2 + 2\frac{\sigma^2}{B}.
\end{aligned}$$

Proof. By the definition of $\mathbf{g}_{\mathbf{y}}^t$, we have

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{g}_{\mathbf{y}}^t - \nabla_{\mathbf{y}} F(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 &= \mathbb{E}\left\|\frac{1}{B}\sum_{(\zeta^t,\xi^t)\in\mathcal{B}^t}\nabla_2 f(\mathbf{x}^t,\mathbf{z}^{t+1};\zeta^t)\nabla g(\mathbf{y}^t;\xi^t) - \nabla_{\mathbf{y}} F(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 \\
&\leq 2\mathbb{E}\left\|\frac{1}{B}\sum_{(\zeta^t,\xi^t)\in\mathcal{B}^t}\nabla_2 f(\mathbf{x}^t,\mathbf{z}^{t+1};\zeta^t)\nabla g(\mathbf{y}^t;\xi^t) - \nabla_2 f(\mathbf{x}^t,\mathbf{z}^{t+1})\nabla g(\mathbf{y}^t)\right\|^2 \\
&\quad + 2\mathbb{E}\left\|\nabla_2 f(\mathbf{x}^t,\mathbf{z}^{t+1})\nabla g(\mathbf{y}^t) - \nabla_2 f(\mathbf{x}^t,g(\mathbf{y}^t))\nabla g(\mathbf{y}^t)\right\|^2 \\
&\leq \frac{4(G_f^2+G_g^2)\sigma^2}{B} + 2G_g^2L_f^2\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{y}^t)\right\|^2,
\end{aligned}$$

where at the last step we use the bounded variance assumption. Similarly by the definition of $\mathbf{g}_{\mathbf{x}}^t$ we have:

$$\mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^t - \nabla_{\mathbf{x}} F(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 = \mathbb{E}\left\|\frac{1}{B}\sum_{(\zeta^t,\xi^t)\in\mathcal{B}^t}\nabla_1 f(\mathbf{x}^t,\mathbf{z}^{t+1};\zeta^t) - \nabla_1 f(\mathbf{x}^t,g(\mathbf{y}^t))\right\|^2 \leq 2L_f^2\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{y}^t)\right\|^2 + 2\frac{\sigma^2}{B}.$$

□

Lemma 9 (Connection between stationary measure and iterates)**.** *For Algorithm 2, under the assumptions of Theorem 3, if we define*

$$\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x},\mathbf{y}) = \frac{1}{\eta_{\mathbf{x}}}\left(\mathbf{x} - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x} - \eta_{\mathbf{x}}\nabla_{\mathbf{x}} F(\mathbf{x},\mathbf{y})\right)\right), \hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x},\mathbf{y}) = \frac{1}{\eta_{\mathbf{y}}}\left(\mathbf{y} - \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y} + \eta_{\mathbf{y}}\nabla_{\mathbf{y}} F(\mathbf{x},\mathbf{y})\right)\right)$$

, then the following statement holds:

$$\begin{aligned}
\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 &\leq \frac{2}{\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t+1}\right\|^2 + 4L_f^2\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{y}^t)\right\|^2 + 4\frac{\sigma^2}{B}, \\
\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 &\leq \frac{2}{\eta_{\mathbf{y}}^2}\mathbb{E}\left\|\mathbf{y}^t-\mathbf{y}^{t+1}\right\|^2 + 4L_f^2G_g^2\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{y}^t)\right\|^2 + \frac{8(G_f^2+G_g^2)\sigma^2}{B}.
\end{aligned}$$

Proof. We begin with proving the first statement. According to updating rule we have

$$\begin{aligned}
\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t,\mathbf{y}^t))\right\|^2 &= \mathbb{E}\left\|\frac{1}{\eta_{\mathbf{x}}}\left(\mathbf{x}^t - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}^t - \eta_{\mathbf{x}}\nabla_{\mathbf{x}} F(\mathbf{x}^t,\mathbf{y}^t)\right)\right)\right\|^2 \\
&\leq \frac{2}{\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathbf{x}^t - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}^t - \eta_{\mathbf{x}}\mathbf{g}_{\mathbf{x}}^t\right)\right\|^2 + \frac{2}{\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathcal{P}_{\mathcal{X}}\left(\mathbf{x}^t - \eta\mathbf{g}_{\mathbf{x}}^t\right) - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}^t - \eta\nabla_{\mathbf{x}} F(\mathbf{x}^t,\mathbf{y}^t)\right)\right\|^2 \\
&\leq \frac{2}{\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t+1}\right\|^2 + 2\mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^t - \nabla_{\mathbf{x}} F(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 \\
&\leq \frac{2}{\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t+1}\right\|^2 + 4\mathbb{E}\left\|\nabla_1 f(\mathbf{x}^t,\mathbf{z}^{t+1};\zeta^t) - \nabla_1 f(\mathbf{x}^t,\mathbf{z}^{t+1})\right\|^2 \\
&\quad + 4\mathbb{E}\left\|\nabla_1 f(\mathbf{x}^t,\mathbf{z}^{t+1}) - \nabla_1 f(\mathbf{x}^t,g(\mathbf{y}^t))\right\|^2 \\
&\leq \frac{2}{\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t+1}\right\|^2 + 4L_f^2\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{y}^t)\right\|^2 + 4\frac{\sigma^2}{B},
\end{aligned}$$

where at last step we apply Lemma 8. Similarly, for the second statement we have:

$$\begin{aligned}
\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 &= \mathbb{E}\left\|\frac{1}{\eta_{\mathbf{y}}}\left(\mathbf{y}^t - \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y}^t + \eta\nabla_{\mathbf{y}}F(\mathbf{x}^t,\mathbf{y}^t)\right)\right)\right\|^2 \\
&\leq \frac{2}{\eta_{\mathbf{y}}^2}\mathbb{E}\left\|\mathbf{y}^t - \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y}^t + \eta_{\mathbf{y}}\mathbf{g}_{\mathbf{y}}^t\right)\right\|^2 + \frac{2}{\eta_{\mathbf{y}}^2}\mathbb{E}\left\|\mathcal{P}_{\mathcal{Y}}\left(\mathbf{y}^t + \eta_{\mathbf{y}}\mathbf{g}_{\mathbf{y}}^t\right) - \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y}^t + \eta_{\mathbf{y}}\nabla_{\mathbf{y}}F(\mathbf{x}^t,\mathbf{y}^t)\right)\right\|^2 \\
&\leq \frac{2}{\eta_{\mathbf{y}}^2}\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t+1}\right\| + 2\mathbb{E}\left\|\mathbf{g}_{\mathbf{y}}^t - \nabla_{\mathbf{y}}F(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 \\
&= \frac{2}{\eta_{\mathbf{y}}^2}\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t+1}\right\| + 2\mathbb{E}\left\|\frac{1}{B}\sum_{(\zeta^t,\xi^t)\in\mathcal{B}^t}\nabla_2 f(\mathbf{x}^t,\mathbf{z}^{t+1};\zeta^t)\nabla g(\mathbf{y}^t;\xi^t) - \nabla_2 f(\mathbf{x}^t,g(\mathbf{y}^t))\nabla g(\mathbf{y}^t)\right\|^2 \\
&\leq \frac{2}{\eta_{\mathbf{y}}^2}\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t+1}\right\| + 4\mathbb{E}\left\|\frac{1}{B}\sum_{(\zeta^t,\xi^t)\in\mathcal{B}^t}\nabla_2 f(\mathbf{x}^t,\mathbf{z}^{t+1};\zeta^t)\nabla g(\mathbf{y}^t;\xi^t) - \nabla_2 f(\mathbf{x}^t,\mathbf{z}^{t+1})\nabla g(\mathbf{y}^t)\right\|^2 \\
&\quad + 4\mathbb{E}\left\|\nabla_2 f(\mathbf{x}^t,\mathbf{z}^{t+1})\nabla g(\mathbf{y}^t) - \nabla_2 f(\mathbf{x}^t,g(\mathbf{y}^t))\nabla g(\mathbf{y}^t)\right\|^2 \\
&\leq \frac{2}{\eta_{\mathbf{y}}^2}\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t+1}\right\|^2 + 4L_f^2G_g^2\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{y}^t)\right\|^2 + \frac{8(G_f^2+G_g^2)\sigma^2}{B}.
\end{aligned}$$

□

The following lemma connects $F(\mathbf{x}^{t+1},\mathbf{y}^{t+1})$ and $F(\mathbf{x}^{t+1},\mathbf{y}^t)$.

Lemma 10 (Dual Variable One Iteration Analysis)**.** *For Algorithm 2, under the assumptions of Theorem 3, the following statement holds true:*

$$\begin{aligned}
\mathbb{E}[F(\mathbf{x}^{t+1},\mathbf{y}^{t+1}) - F(\mathbf{x}^{t+1},\mathbf{y}^t)] \geq& \left(\frac{1}{2\eta_{\mathbf{y}}} - \frac{L}{2} - \frac{L^2\eta_{\mathbf{x}}}{2}\right)\mathbb{E}\|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 \\
&- \left(\eta_{\mathbf{y}}2(G_g^2+G_f^2)\frac{\sigma^2}{B} + 2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M}\right) \\
&- L_f^2G_g^2\eta_{\mathbf{y}}(1-\beta)^2\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 - 4(1-\beta)^2L_f^2G_g^4\eta_{\mathbf{y}}\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2.
\end{aligned}$$

Proof. By the optimality of projection and updating rule of $\mathbf{w}$:

$$\left\langle \eta_{\mathbf{y}}\mathbf{g}_{\mathbf{y}}^t + (\mathbf{y}^t - \mathbf{y}^{t+1}), \mathbf{y}^t - \mathbf{y}^{t+1}\right\rangle \leq 0$$

Re-arranging terms we have:

$$\left\langle \mathbf{g}_{\mathbf{y}}^t, \mathbf{y}^t - \mathbf{y}^{t+1}\right\rangle \leq -\frac{1}{\eta_{\mathbf{y}}}\|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2$$

Adding and subtracting $\nabla_{\mathbf{y}}F(\mathbf{x}^t,\mathbf{y}^t)$ yields:

$$\left\langle \nabla_{\mathbf{y}}F(\mathbf{x}^t,\mathbf{y}^t), \mathbf{y}^t - \mathbf{y}^{t+1}\right\rangle + \left\langle \mathbf{g}_{\mathbf{y}}^t - \nabla_{\mathbf{y}}F(\mathbf{x}^t,\mathbf{y}^t), \mathbf{y}^t - \mathbf{y}^{t+1}\right\rangle \leq -\frac{1}{\eta_{\mathbf{y}}}\|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2$$

Applying Cauchy-Schwartz inequality yields:

$$\frac{1}{\eta_{\mathbf{y}}}\|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2 \leq \left\langle \nabla_{\mathbf{y}}F(\mathbf{x}^t,\mathbf{y}^t), \mathbf{y}^{t+1} - \mathbf{y}^t\right\rangle + \frac{1}{2\eta_{\mathbf{y}}}\left\|\mathbf{y}^{t+1} - \mathbf{y}^t\right\|^2 + \frac{1}{2}\eta_{\mathbf{y}}\left\|\mathbf{g}_{\mathbf{y}}^t - \nabla_{\mathbf{y}}F(\mathbf{x}^t,\mathbf{y}^t)\right\|^2$$

Now, we take expectation over the randomness of $\mathcal{B}^t$, and apply Lemma 8:

$$\frac{1}{2\eta_{\mathbf{y}}}\mathbb{E}\|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2 \leq \mathbb{E}\left\langle \nabla_{\mathbf{y}}F(\mathbf{x}^t,\mathbf{y}^t), \mathbf{y}^{t+1} - \mathbf{y}^t\right\rangle + \eta_{\mathbf{y}}\frac{2(G_f^2+G_g^2)\sigma^2}{B} + \eta_{\mathbf{y}}G_g^2L_f^2\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{y}^t)\right\|^2.$$

Plugging in Lemma 1 yields:

$$\begin{aligned}\frac{1}{2\eta_{\mathbf{y}}}\mathbb{E}\|\mathbf{y}^{t+1}-\mathbf{y}^t\|^2 &\leq \mathbb{E}\left\langle \nabla_{\mathbf{y}} F(\mathbf{x}^t,\mathbf{y}^t), \mathbf{y}^{t+1}-\mathbf{y}^t\right\rangle + +\eta_{\mathbf{y}}\frac{2(G_f^2+G_g^2)\sigma^2}{B}\\ &\quad + \eta_{\mathbf{y}} G_g^2 L_f^2\left((1-\beta)^2\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 + 4(1-\beta)^2 G_g^2\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2 + 2\beta^2\frac{\sigma^2}{M}\right)\end{aligned}$$

On the other hand, since F is L smooth, we have:

$$\begin{aligned}
&F(\mathbf{x}^{t+1},\mathbf{y}^{t+1}) - F(\mathbf{x}^{t+1},\mathbf{y}^t) \geq \left\langle \nabla_{\mathbf{y}} F(\mathbf{x}^{t+1},\mathbf{y}^t),\, \mathbf{y}^{t+1}-\mathbf{y}^t\right\rangle - \frac{L}{2}\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2\\
&\quad = \left\langle \nabla_{\mathbf{y}} F(\mathbf{x}^t,\mathbf{y}^t),\, \mathbf{y}^{t+1}-\mathbf{y}^t\right\rangle - \frac{L}{2}\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2 + \left\langle \nabla_{\mathbf{y}} F(\mathbf{x}^{t+1},\mathbf{y}^t) - \nabla_{\mathbf{y}} F(\mathbf{x}^t,\mathbf{y}^t),\, \mathbf{y}^{t+1}-\mathbf{y}^t\right\rangle\\
&\quad \geq \left\langle \nabla_{\mathbf{y}} F(\mathbf{x}^t,\mathbf{y}^t),\, \mathbf{y}^{t+1}-\mathbf{y}^t\right\rangle - \frac{L}{2}\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2 - \frac{1}{L\sqrt{\eta_{\mathbf{x}}}}\left\|\nabla_{\mathbf{y}} F(\mathbf{x}^{t+1},\mathbf{y}^t) - \nabla_{\mathbf{y}} F(\mathbf{x}^t,\mathbf{y}^t)\right\| L\sqrt{\eta_{\mathbf{x}}}\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|\\
&\quad \geq \left\langle \nabla_{\mathbf{y}} F(\mathbf{x}^t,\mathbf{y}^t),\, \mathbf{y}^{t+1}-\mathbf{y}^t\right\rangle - \frac{L}{2}\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2 - \frac{L^2\eta_{\mathbf{x}}}{2}\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2\\
&\quad \geq \left\langle \nabla_{\mathbf{y}} F(\mathbf{x}^t,\mathbf{y}^t),\, \mathbf{y}^{t+1}-\mathbf{y}^t\right\rangle - \left(\frac{L}{2}+\frac{L^2\eta_{\mathbf{x}}}{2}\right)\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2
\end{aligned}$$

Putting pieces together will conclude the proof:

$$\begin{aligned}\mathbb{E}[F(\mathbf{x}^{t+1},\mathbf{y}^{t+1}) - F(\mathbf{x}^{t+1},\mathbf{y}^t)] &\geq \left(\frac{1}{2\eta_{\mathbf{y}}} - \frac{L}{2} - \frac{L^2\eta_{\mathbf{x}}}{2}\right)\mathbb{E}\|\mathbf{y}^{t+1}-\mathbf{y}^t\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2\\
&\quad - \left(\eta_{\mathbf{y}} 2(G_g^2+G_f^2)\frac{\sigma^2}{B} + 2\beta^2 L_f^2 G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M}\right)\\
&\quad - L_f^2 G_g^2\eta_{\mathbf{y}}(1-\beta)^2\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 - 4(1-\beta)^2 L_f^2 G_g^4\eta_{\mathbf{y}}\mathbb{E}\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2.\end{aligned}$$

□

The following lemma connects $F(\mathbf{x}^{t+1},\mathbf{y}^{t+1})$ and $F(\mathbf{x}^t,\mathbf{y}^t)$.

Lemma 11 (One Iteration Descent Lemma)**.** *For Algorithm 2, under the assumptions of Theorem 3, the following statement holds true:*

$$\begin{aligned}\mathbb{E}[F(\mathbf{x}^{t+1},\mathbf{y}^{t+1}) - F(\mathbf{x}^t,\mathbf{y}^t)] &\geq \left(\frac{1}{2\eta_{\mathbf{y}}} - \frac{L}{2} - \frac{L^2\eta_{\mathbf{x}}}{2}\right)\mathbb{E}\|\mathbf{y}^{t+1}-\mathbf{y}^t\|^2\\
&\quad - \left(4(1-\beta)^2 L_f^2 G_f^2 G_g^2\eta_{\mathbf{y}} + \frac{L^2\eta_{\mathbf{x}}}{2}\right)\mathbb{E}\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2\\
&\quad + \left(\frac{\mu}{2} - \frac{3}{2\eta_{\mathbf{x}}} - \frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}\right)\mathbb{E}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2 + \left(\mu - \frac{1}{2\eta_{\mathbf{x}}} - \frac{\eta_{\mathbf{x}} L^2}{2}\right)\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2\\
&\quad - \left(L_f^2 G_g^2\eta_{\mathbf{y}} + \eta_{\mathbf{x}} L_f^2\right)(1-\beta)^2\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 - \left(2\eta_{\mathbf{y}}(G_g^2+G_f^2)\frac{\sigma^2}{B} + 2\beta^2 L_f^2 G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M}\right).\end{aligned}$$

Proof. According to property of projection we have:

$$\left\langle \mathbf{g}_{\mathbf{x}}^t + \frac{1}{\eta}(\mathbf{x}^{t+1}-\mathbf{x}^t), \mathbf{x}-\mathbf{x}^{t+1}\right\rangle \geq 0 \tag{5}$$

$$\left\langle \mathbf{g}_{\mathbf{x}}^t + \frac{1}{\eta}(\mathbf{x}^{t+1}-\mathbf{x}^t), \mathbf{x}^t-\mathbf{x}^{t+1}\right\rangle \geq 0 \tag{6}$$

$$\left\langle \mathbf{g}_{\mathbf{x}}^{t-1} + \frac{1}{\eta}(\mathbf{x}^t-\mathbf{x}^{t-1}), \mathbf{x}^{t+1}-\mathbf{x}^t\right\rangle \geq 0 \tag{7}$$

Since $F(\cdot, \mathbf{y})$ is strongly convex for any $\mathbf{y} \in \mathcal{Y}$, we have:

$$
\begin{aligned}
F(\mathbf{x}^{t+1}, \mathbf{y}^t) - F(\mathbf{x}^t, \mathbf{y}^t) &\geq \left\langle \nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}^t), \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle + \frac{\mu}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 \\
&\geq \left\langle \nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}^t) - \mathbf{g}_{\mathbf{x}}^{t-1}, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle + \frac{\mu}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 - \frac{1}{\eta_{\mathbf{x}}} \left\langle \mathbf{x}^t - \mathbf{x}^{t-1}, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle \\
&\geq \left\langle \nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}^t) - \nabla_{\mathbf{x}} F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}), \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle + \frac{\mu}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 \\
&\quad + \left\langle \mathbf{g}_{\mathbf{x}}^{t-1} - \nabla_{\mathbf{x}} F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}), \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle - \frac{1}{\eta_{\mathbf{x}}} \left\langle \mathbf{x}^t - \mathbf{x}^{t-1}, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle
\end{aligned}
$$

where at second inequality we use the fact of (7) . For the first dot product term, we observe that:

$$
\begin{aligned}
\left\langle \nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}^t) - \nabla_{\mathbf{x}} F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}), \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle &= \left\langle \nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}^t) - \nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}^{t-1}), \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle \\
&\quad + \left\langle \nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}^{t-1}) - \nabla_{\mathbf{x}} F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}), \Delta \right\rangle \\
&\quad + \left\langle \nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}^{t-1}) - \nabla_{\mathbf{x}} F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}), \mathbf{x}^t - \mathbf{x}^{t-1} \right\rangle \\
&\geq -\frac{L^2 \eta_{\mathbf{x}}}{2}\|\mathbf{y}^t - \mathbf{y}^{t-1}\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 - \frac{\eta_{\mathbf{x}} L^2}{2}\|\mathbf{x}^t - \mathbf{x}^{t-1}\|^2 \\
&\quad - \frac{1}{2\eta_{\mathbf{x}}}\|\Delta\|^2 + \mu\|\mathbf{x}^t - \mathbf{x}^{t-1}\|^2,
\end{aligned} \tag{8}
$$

where $\Delta = (\mathbf{x}^{t+1} - \mathbf{x}^t) - (\mathbf{x}^t - \mathbf{x}^{t-1})$.

For the second dot product, we have:

$$
\begin{aligned}
\mathbb{E} \left\langle \mathbf{g}_{\mathbf{x}}^{t-1} - \nabla_{\mathbf{x}} F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}), \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle &\geq -\eta_{\mathbf{x}}(1-\beta)^2 \mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^{t-1} - \nabla_{\mathbf{x}} F(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\right\|^2 - \frac{1}{4\eta(1-\beta)^2}\mathbb{E}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 \\
&\geq -\eta_{\mathbf{x}}(1-\beta)^2 L_f^2 \mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 - \frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}\mathbb{E}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2,
\end{aligned}
$$

where we plug in Lemma 8 at last step.

For the third dot product, we use the following identity:

$$
\frac{1}{\eta_{\mathbf{x}}} \left\langle \mathbf{x}^t - \mathbf{x}^{t-1}, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle = \frac{1}{\eta_{\mathbf{x}}} \left(\frac{1}{2}\|\mathbf{x}^t - \mathbf{x}^{t-1}\|^2 + \frac{1}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 - \frac{1}{2}\|\Delta\|^2 \right) \tag{9}
$$

By putting pieces together we have:

$$
\begin{aligned}
\mathbb{E}[F(\mathbf{x}^{t+1}, \mathbf{y}^t) - F(\mathbf{x}^t, \mathbf{y}^t)] &\geq -\frac{L^2 \eta_{\mathbf{x}}}{2}\mathbb{E}\|\mathbf{y}^t - \mathbf{y}^{t-1}\|^2 + \left(\frac{\mu}{2} - \frac{1}{\eta_{\mathbf{x}}} - \frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}\right)\mathbb{E}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 \\
&\quad + \left(\mu - \frac{1}{2\eta_{\mathbf{x}}} - \frac{\eta_{\mathbf{x}} L^2}{2}\right)\mathbb{E}\|\mathbf{x}^t - \mathbf{x}^{t-1}\|^2 - \eta_{\mathbf{x}}(1-\beta)^2 L_f^2 \mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2.
\end{aligned} \tag{10}
$$

Recall Lemma 10 gives the following lower bound of $\mathbb{E}[F(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}) - F(\mathbf{x}^{t+1}, \mathbf{y}^t)]$:

$$
\begin{aligned}
\mathbb{E}[F(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}) - F(\mathbf{x}^{t+1}, \mathbf{y}^t)] &\geq \left(\frac{1}{2\eta_{\mathbf{y}}} - \frac{L}{2} - \frac{L^2 \eta_{\mathbf{x}}}{2}\right)\mathbb{E}\|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 \\
&\quad - \left(2\eta_{\mathbf{y}}(G_g^2 + G_f^2)\frac{\sigma^2}{B} + 2\beta^2 L_f^2 G_g^2 \eta_{\mathbf{y}} \frac{\sigma^2}{M}\right) \\
&\quad - L_f^2 G_g^2 \eta_{\mathbf{y}}(1-\beta)^2 \mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 - 4(1-\beta)^2 L_f^2 G_f^2 G_g^2 \eta_{\mathbf{y}} \mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2.
\end{aligned} \tag{11}
$$

Hence, Adding (10) and (11) yields:

$$\begin{aligned}
\mathbb{E}[F(\mathbf{x}^{t+1},\mathbf{y}^{t+1})-F(\mathbf{x}^t,\mathbf{y}^t)] \geq & \left(\frac{1}{2\eta_{\mathbf{y}}}-\frac{L}{2}-\frac{L^2\eta_{\mathbf{x}}}{2}\right)\mathbb{E}\|\mathbf{y}^{t+1}-\mathbf{y}^t\|^2 \\
& -\left(4(1-\beta)^2L_f^2G_f^2G_g^2\eta_{\mathbf{y}}+\frac{L^2\eta_{\mathbf{x}}}{2}\right)\mathbb{E}\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2 \\
& +\left(\frac{\mu}{2}-\frac{3}{2\eta_{\mathbf{x}}}-\frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}\right)\mathbb{E}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2+\left(\mu-\frac{1}{2\eta_{\mathbf{x}}}-\frac{\eta_{\mathbf{x}}L^2}{2}\right)\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2 \\
& -\left(L_f^2G_g^2\eta_{\mathbf{y}}+\eta_{\mathbf{x}}L_f^2\right)(1-\beta)^2\mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{y}^{t-1})\right\|^2-\left(2\eta_{\mathbf{y}}(G_g^2+G_f^2)\frac{\sigma^2}{B}+2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M}\right),
\end{aligned}$$

which concludes the proof.

□

Lemma 12. *For Algorithm 2, under the assumptions of Theorem 3, define the following auxillary function* $\hat{F}^{t+1}$

$$\begin{aligned}
\hat{F}^{t+1} :=& F(\mathbf{x}^{t+1},\mathbf{y}^{t+1})+s^{t+1}-\left(\frac{1}{4\eta_{\mathbf{y}}}+4L_g^2G_f^2G_g^2\eta_{\mathbf{y}}+\frac{\eta_{\mathbf{x}}L^2}{2}+\frac{168G_g^2L_f^2}{\mu^2\eta_{\mathbf{x}}\beta}+\frac{48G_g^2L_f^2\beta}{\mu^2\eta_{\mathbf{x}}}\right)\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2 \\
& -\left(\frac{1}{8\eta_{\mathbf{y}}}+\frac{84G_g^2L_f^2}{\mu^2\eta_{\mathbf{x}}\beta}\right)\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2+\left(\frac{7}{2\eta_{\mathbf{x}}}+\mu-\frac{\eta_{\mathbf{x}}L^2}{2}-\frac{2L_f^2}{\mu}\right)\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2
\end{aligned}$$

where $s^{t+1} := -\frac{2}{\eta_{\mathbf{x}}^2\mu}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2$ *then the following statement holds true:*

$$\begin{aligned}
\mathbb{E}[\hat{F}^{t+1}-\hat{F}^t] \geq & C_1\mathbb{E}\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2+\frac{1}{8\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2+\frac{1}{8\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^{t-1}-\mathbf{y}^{t-2}\right\|^2+C_2\mathbb{E}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2 \\
& -(1-\beta)^2\left(L_g^2G_f^2\eta_{\mathbf{y}}+\eta_{\mathbf{x}}L_f^2\right)\mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{y}^{t-1})\right\|^2 \\
& -\left(2\eta_{\mathbf{y}}(G_g^2+G_f^2)\frac{\sigma^2}{B}+2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M}+\frac{72L_f^2}{\mu^2\eta_{\mathbf{x}}}\beta\frac{\sigma^2}{M}\right)-\left(1-\frac{\beta}{2}\right)^2\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}_j^t-\mathbf{z}^{t-1}\right\|^2.
\end{aligned}$$

where

$$C_1 = \frac{1}{4\eta_{\mathbf{y}}}-\frac{L}{2}-\eta_{\mathbf{x}}L^2-4L_g^2G_f^2G_g^2\eta_{\mathbf{y}}-\frac{168G_g^2L_f^2}{\mu^2\eta_{\mathbf{x}}\beta}-\frac{48G_g^2L_f^2\beta}{\mu^2\eta_{\mathbf{x}}}, \tag{12}$$

$$C_2 = \frac{1}{\eta_{\mathbf{x}}}+\frac{3}{2}\mu-\frac{\eta_{\mathbf{x}}L^2}{2}-\frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}-\frac{2L_f^2}{\mu}. \tag{13}$$

Proof. Define $\Delta^{t+1} := \mathbf{x}^{t+1}-\mathbf{x}^t-(\mathbf{x}^t-\mathbf{x}^{t-1})$. According to (6) and (7):

$$\frac{1}{\eta}\left\langle\Delta^{t+1},\mathbf{x}^t-\mathbf{x}^{t+1}\right\rangle \geq \left\langle\mathbf{g}_{\mathbf{x}}^t-\mathbf{g}_{\mathbf{x}}^{t-1},\,\mathbf{x}^{t+1}-\mathbf{x}^t\right\rangle.$$

If we define $\mathbf{g}_{\mathbf{x}}^t(\mathbf{x},\mathbf{z}) = \nabla_1 f(\mathbf{x},\mathbf{z};\mathcal{B}^t) + \nabla h(\mathbf{x})$.

$$
\begin{aligned}
\frac{1}{\eta}\left\langle \Delta^{t+1}, \mathbf{x}^t - \mathbf{x}^{t+1}\right\rangle &\geq \left\langle \mathbf{g}_{\mathbf{x}}^t - \mathbf{g}_{\mathbf{x}}^t(\mathbf{x}^t,\mathbf{z}^t;\xi^t),\, \mathbf{x}^{t+1} - \mathbf{x}^t\right\rangle \\
&\quad + \left\langle \mathbf{g}_{\mathbf{x}}^t(\mathbf{x}^t,\mathbf{z}^t;\xi^t) - \mathbf{g}_{\mathbf{x}}^{t-1},\, \mathbf{x}^{t+1} - \mathbf{x}^t\right\rangle \\
&\geq \left\langle \mathbf{g}_{\mathbf{x}}^t - \mathbf{g}_{\mathbf{x}}(\mathbf{x}^t,\mathbf{z}^t),\, \mathbf{x}^{t+1} - \mathbf{x}^t\right\rangle \\
&\quad + \left\langle \mathbf{g}_{\mathbf{x}}(\mathbf{x}^t,\mathbf{z}^t) - \mathbf{g}_{\mathbf{x}}^{t-1},\, \mathbf{x}^{t+1} - \mathbf{x}^t - (\mathbf{x}^t - \mathbf{x}^{t-1})\right\rangle \\
&\quad + \left\langle \mathbf{g}_{\mathbf{x}}(\mathbf{x}^t,\mathbf{z}^t) - \mathbf{g}_{\mathbf{x}}^{t-1},\, (\mathbf{x}^t - \mathbf{x}^{t-1})\right\rangle \\
&\geq -\frac{L_f^2}{\mu}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2 - \frac{\mu}{4}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 \\
&\quad - \frac{\eta_{\mathbf{x}} L_f^2}{2}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\left\|\Delta^{t+1}\right\|^2 + \mu\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2
\end{aligned}
$$

where we use the Lipschitzness of $\nabla f(\mathbf{x},\mathbf{z};\xi)$ at the last step. Since $\frac{1}{\eta_{\mathbf{x}}}\left\langle \Delta^{t+1}, \mathbf{x}^t - \mathbf{x}^{t+1}\right\rangle = \frac{1}{2\eta_{\mathbf{x}}}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\left\|\Delta^{t+1}\right\|^2$ we have:

$$
\begin{aligned}
\frac{1}{2\eta_{\mathbf{x}}}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 &\geq -\frac{L_f^2}{\mu}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2 - \frac{\mu}{4}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 \\
&\quad - \frac{\eta_{\mathbf{x}} L_f^2}{2}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 + \mu\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2
\end{aligned}
$$

Multiplying both sides with $\frac{4}{\mu\eta_{\mathbf{x}}}$ yields:

$$
\begin{aligned}
\frac{2}{\mu\eta_{\mathbf{x}}^2}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 - \frac{2}{\mu\eta_{\mathbf{x}}^2}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 &\geq -\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2 - \frac{1}{\eta_{\mathbf{x}}}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 \\
&\quad - \frac{2L_f^2}{\mu}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 + \frac{4}{\eta_{\mathbf{x}}}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2
\end{aligned} \tag{14}
$$

Recall that in Lemma 11 we have:

$$
\begin{aligned}
\mathbb{E}[F(\mathbf{x}^{t+1},\mathbf{y}^{t+1}) - F(\mathbf{x}^t,\mathbf{y}^t)] &\geq \left(\frac{1}{2\eta_{\mathbf{y}}} - \frac{L}{2} - \frac{L^2\eta_{\mathbf{x}}}{2}\right)\mathbb{E}\|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2 \\
&\quad - \left(4(1-\beta)^2 L_f^2 G_f^2 G_g^2 \eta_{\mathbf{y}} + \frac{L^2\eta_{\mathbf{x}}}{2}\right)\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 \\
&\quad + \left(\frac{\mu}{2} - \frac{3}{2\eta_{\mathbf{x}}} - \frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}\right)\mathbb{E}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 + \left(\mu - \frac{1}{2\eta_{\mathbf{x}}} - \frac{\eta_{\mathbf{x}} L^2}{2}\right)\mathbb{E}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 \\
&\quad - \left(L_f^2 G_g^2 \eta_{\mathbf{y}} + \eta_{\mathbf{x}} L_f^2\right)(1-\beta)^2\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 - \left(2\eta_{\mathbf{y}}(G_g^2 + G_f^2)\frac{\sigma^2}{B} + 2\beta^2 L_f^2 G_g^2 \eta_{\mathbf{y}}\frac{\sigma^2}{M}\right).
\end{aligned}
$$

Combining Lemma 11 together with (24) yields:

$$
\begin{aligned}
&\mathbb{E}[F(\mathbf{x}^{t+1},\mathbf{y}^{t+1}) + s^{t+1} - (F(\mathbf{x}^t,\mathbf{y}^t) + s^t)] \\
&\geq \left(\frac{1}{2\eta_{\mathbf{y}}} - \frac{L}{2} - \frac{L^2\eta_{\mathbf{x}}}{2}\right)\mathbb{E}\|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2 - \left(4(1-\beta)^2 L_f^2 G_f^2 G_g^2 \eta_{\mathbf{y}} + \frac{L^2\eta_{\mathbf{x}}}{2}\right)\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 \\
&+ \left(\frac{\mu}{2} - \frac{3}{2\eta_{\mathbf{x}}} - \frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2} - \frac{1}{\eta_{\mathbf{x}}}\right)\mathbb{E}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 + \left(\mu - \frac{1}{2\eta_{\mathbf{x}}} - \frac{\eta_{\mathbf{x}} L^2}{2} + \frac{4}{\eta_{\mathbf{x}}} - \frac{2L_f^2}{\mu}\right)\mathbb{E}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 \\
&- (1-\beta)^2\left(L_f^2 G_g^2 \eta_{\mathbf{y}} + \eta_{\mathbf{x}} L_f^2\right)\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 - \left(2\eta_{\mathbf{y}}(G_g^2 + G_f^2)\frac{\sigma^2}{B} + 2\beta^2 L_f^2 G_g^2 \eta_{\mathbf{y}}\frac{\sigma^2}{M}\right) - \frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2.
\end{aligned}
$$

Now we plug in Lemma 7 to replace $\mathbb{E}\left\|\mathbf{z}^{t+1}-\mathbf{z}^t\right\|^2$ together with the fact that $\frac{1}{\beta}\geq 1$ and get:

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{x}^{t+1},\mathbf{y}^{t+1})+s^{t+1}-(F(\mathbf{x}^t,\mathbf{y}^t)+s^t)] \\
&\quad\geq \left(\frac{1}{2\eta_{\mathbf{y}}}-\frac{L}{2}-\frac{L^2\eta_{\mathbf{x}}}{2}\right)\mathbb{E}\|\mathbf{y}^{t+1}-\mathbf{y}^t\|^2 \\
&\quad - \left(4L_f^2G_f^2G_g^2\eta_{\mathbf{y}}+\frac{L^2\eta_{\mathbf{x}}}{2}+\frac{84G_g^2L_f^2}{\mu^2\eta_{\mathbf{x}}\beta}+\frac{48G_g^2L_f^2\beta}{\mu^2\eta_{\mathbf{x}}}\right)\mathbb{E}\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2-\frac{48G_g^2L_f^2}{\mu^2\eta_{\mathbf{x}}\beta}\mathbb{E}\left\|\mathbf{y}^{t-1}-\mathbf{y}^{t-2}\right\|^2 \\
&\quad + \left(\frac{\mu}{2}-\frac{5}{2\eta_{\mathbf{x}}}-\frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}\right)\mathbb{E}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2+\left(\mu+\frac{7}{2\eta_{\mathbf{x}}}-\frac{\eta_{\mathbf{x}}L^2}{2}-\frac{2L_f^2}{\mu}\right)\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2 \\
&\quad - (1-\beta)^2\left(L_f^2G_g^2\eta_{\mathbf{y}}+\eta_{\mathbf{x}}L_f^2\right)\mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{y}^{t-1})\right\|^2 \\
&\quad - \left(2\eta_{\mathbf{y}}(G_g^2+G_f^2)\frac{\sigma^2}{B}+2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M}\right)-\left(1-\frac{\beta}{2}\right)^2\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^t-\mathbf{z}^{t-1}\right\|^2+\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\cdot 18\beta\frac{\sigma^2}{M}.
\end{aligned}$$

Recall our definition of potential function $\hat{F}^{t+1}$

$$\begin{aligned}
\hat{F}^{t+1} :=& F(\mathbf{x}^{t+1},\mathbf{y}^{t+1})+s^{t+1}-\left(\frac{1}{4\eta_{\mathbf{y}}}+4L_f^2G_f^2G_g^2\eta_{\mathbf{y}}+\frac{\eta_{\mathbf{x}}L^2}{2}+\frac{168G_g^2L_f^2}{\mu^2\eta_{\mathbf{x}}\beta}+\frac{48G_g^2L_f^2\beta}{\mu^2\eta_{\mathbf{x}}}\right)\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2 \\
&-\left(\frac{1}{8\eta_{\mathbf{y}}}+\frac{84G_g^2L_f^2}{\mu^2\eta_{\mathbf{x}}\beta}\right)\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2+\left(\frac{7}{2\eta_{\mathbf{x}}}+\mu-\frac{\eta_{\mathbf{x}}L^2}{2}-\frac{2L_f^2}{\mu}\right)\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2.
\end{aligned}$$

We conclude that:

$$\begin{aligned}
\mathbb{E}[\hat{F}^{t+1}-\hat{F}^t]\geq& \left(\frac{1}{4\eta_{\mathbf{y}}}-\frac{L}{2}-\eta_{\mathbf{x}}L^2-4L_f^2G_f^2G_g^2\eta_{\mathbf{y}}-\frac{168G_g^2L_f^2}{\mu^2\eta_{\mathbf{x}}\beta}-\frac{48G_g^2L_f^2\beta}{\mu^2\eta_{\mathbf{x}}}\right)\mathbb{E}\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2 \\
&+\frac{1}{8\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2+\frac{1}{8\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^{t-1}-\mathbf{y}^{t-2}\right\|^2+\left(\frac{1}{\eta_{\mathbf{x}}}+\frac{3}{2}\mu-\frac{\eta_{\mathbf{x}}L^2}{2}-\frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}-\frac{2L_f^2}{\mu}\right)\mathbb{E}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2 \\
&-(1-\beta)^2\left(L_f^2G_g^2\eta_{\mathbf{y}}+\eta_{\mathbf{x}}L_f^2\right)\mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{y}^{t-1})\right\|^2 \\
&-\left(2\eta_{\mathbf{y}}(G_g^2+G_f^2)\frac{\sigma^2}{B}+2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M}+\frac{72L_f^2}{\mu^2\eta_{\mathbf{x}}}\beta\frac{\sigma^2}{M}\right)-\left(1-\frac{\beta}{2}\right)^2\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^t-\mathbf{z}^{t-1}\right\|^2.
\end{aligned}$$

□

Lemma 13. *Let C_1, C_2 be defined in (12) and (13). If the following conditions hold:*

$$\frac{1}{8\eta_{\mathbf{y}}}-4(2C_1\eta_{\mathbf{y}}^2L_f^2G_g^2+2C_2\eta_{\mathbf{x}}^2L_f^2+1)(1-\beta)^2G_f^2-\frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\left(\frac{12}{\beta}+6\beta\right)G_g^2\geq 0, \tag{15}$$

$$1-\left(2C_1\eta_{\mathbf{y}}^2L_f^2G_g^2+2C_2\eta_{\mathbf{x}}^2L_f^2+L_f^2G_g^2\eta_{\mathbf{y}}+\eta_{\mathbf{x}}L_f^2+1\right)(1-\beta)^2\geq 0, \tag{16}$$

$$\frac{1}{8\eta_{\mathbf{y}}}-\frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\frac{6}{\beta}G_g^2\geq 0, \tag{17}$$

then we have:

$$\begin{aligned}
\mathbb{E}[\tilde{F}^{t+1}-\tilde{F}^t]\geq& \frac{C_1}{2}\eta_{\mathbf{y}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2+\frac{C_2}{2}\eta_{\mathbf{x}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 \\
&-\left(4C_1\eta_{\mathbf{y}}^2(G_g^2+G_f^2)+2C_2\eta_{\mathbf{x}}^2\right)\frac{\sigma^2}{B}-\left(2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}+4\beta^2C_1\eta_{\mathbf{y}}^2L_f^2G_g^2+4\beta^2C_2\eta_{\mathbf{x}}^2L_f^2\right)\frac{\sigma^2}{M}-\frac{72L_f^2}{\mu^2\eta_{\mathbf{x}}}\beta\frac{\sigma^2}{M},
\end{aligned}$$

where

$$\tilde{F}^{t+1} := \hat{F}^{t+1} - \mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{y}^t))\right\|^2 - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2.$$

Proof. According to Lemma 12:

$$\begin{aligned}
\mathbb{E}[\hat{F}^{t+1} - \hat{F}^t] \geq{}& C_1\mathbb{E}\left\|\mathbf{y}^{t+1} - \mathbf{y}^t\right\|^2 + \frac{1}{8\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 + \frac{1}{8\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^{t-1} - \mathbf{y}^{t-2}\right\|^2 + C_2\mathbb{E}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 \\
&- (1-\beta)^2\left(L_f^2G_g^2\eta_{\mathbf{y}} + \eta_{\mathbf{x}}L_f^2\right)\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 \\
&- \left(2\eta_{\mathbf{y}}(G_g^2 + G_f^2)\frac{\sigma^2}{B} + 2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M} + \frac{72L_f^2}{\mu^2\eta_{\mathbf{x}}}\beta\frac{\sigma^2}{M}\right) - \left(1-\frac{\beta}{2}\right)^2\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}_j^t - \mathbf{z}^{t-1}\right\|^2.
\end{aligned}$$

where

$$\begin{aligned}
C_1 &= \frac{1}{4\eta_{\mathbf{y}}} - \frac{L}{2} - \eta_{\mathbf{x}}L^2 - 4L_f^2G_f^2G_g^2\eta_{\mathbf{y}} - \frac{168G_g^2L_f^2}{\mu^2\eta_{\mathbf{x}}\beta} - \frac{48G_g^2L_f^2\beta}{\mu^2\eta_{\mathbf{x}}}, \\
C_2 &= \frac{1}{\eta_{\mathbf{x}}} + \frac{3}{2}\mu - \frac{\eta_{\mathbf{x}}L^2}{2} - \frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2} - \frac{2L_f^2}{\mu}.
\end{aligned}$$

Now we plug in Lemma 9:

$$\begin{aligned}
\mathbb{E}[\hat{F}^{t+1} - \hat{F}^t] \geq{}& C_1\left(\frac{1}{2}\eta_{\mathbf{y}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 - 4\eta_{\mathbf{y}}^2(G_f^2 + G_g^2)\frac{\sigma^2}{B} - 2\eta_{\mathbf{y}}^2G_g^2L_f^2\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t))\right\|^2\right) \\
&+ C_2\left(\frac{1}{2}\eta_{\mathbf{x}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 - 2\frac{\eta_{\mathbf{x}}^2\sigma^2}{B} - 2\eta_{\mathbf{x}}^2L_f^2\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{y}^t))\right\|^2\right) \\
&- (1-\beta)^2\left(L_f^2G_g^2\eta_{\mathbf{y}} + \eta_{\mathbf{x}}L_f^2\right)\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 - \left(2\eta_{\mathbf{y}}(G_g^2 + G_f^2)\frac{\sigma^2}{B} + 2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M} + \frac{72L_f^2}{\mu^2\eta_{\mathbf{x}}}\beta\frac{\sigma^2}{M}\right) \\
&- \left(1-\frac{\beta}{2}\right)^2\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2 + \frac{1}{8\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 + \frac{1}{8\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^{t-1} - \mathbf{y}^{t-2}\right\|^2.
\end{aligned}$$

Plugging in Lemma 1 yields:

$$\begin{aligned}
&\mathbb{E}[\hat{F}^{t+1} - \hat{F}^t] \geq \frac{C_1}{2}\eta_{\mathbf{y}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 + C_2\frac{1}{2}\eta_{\mathbf{x}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 \\
&- 2C_1\eta_{\mathbf{y}}^2L_f^2G_g^2\left((1-\beta)^2\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 + 4(1-\beta)^2G_g^2\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 + 2\beta^2\frac{\sigma^2}{M}\right) \\
&- 2C_2\eta_{\mathbf{x}}^2L_f^2\left((1-\beta)^2\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 + 4(1-\beta)^2G_g^2\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 + 2\beta^2\frac{\sigma^2}{M}\right) \\
&- (1-\beta)^2\left(L_f^2G_g^2\eta_{\mathbf{y}} + \eta_{\mathbf{x}}L_f^2\right)\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 - \left(4C_1\eta_{\mathbf{y}}^2(G_g^2 + G_f^2) + 2C_2\eta_{\mathbf{x}}^2\right)\frac{\sigma^2}{B} - 2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M} - \frac{72L_f^2}{\mu^2\eta_{\mathbf{x}}}\beta\frac{\sigma^2}{M} \\
&- \left(1-\frac{\beta}{2}\right)^2\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2 + \frac{1}{8\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 + \frac{1}{8\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^{t-1} - \mathbf{y}^{t-2}\right\|^2.
\end{aligned}$$

Rearranging the terms yields:

$$
\begin{aligned}
&\mathbb{E}[\hat{F}^{t+1} - \hat{F}^t] \\
&\geq \frac{C_1}{2}\eta_{\mathbf{y}}^2 \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 + \frac{C_2}{2}\eta_{\mathbf{x}}^2 \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 + \left(\frac{1}{8\eta_{\mathbf{y}}} - (2C_1\eta_{\mathbf{y}}^2 L_f^2 G_g^2 + 2C_2\eta_{\mathbf{x}}^2 L_f^2) 4(1-\beta)^2 G_g^2\right)\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 \\
&- \left(2C_1\eta_{\mathbf{y}}^2 L_f^2 G_g^2 + 2C_2\eta_{\mathbf{x}}^2 L_f^2 + L_f^2 G_g^2 \eta_{\mathbf{y}} + \eta_{\mathbf{x}} L_f^2\right)(1-\beta)^2 \mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 + \frac{1}{8\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^{t-1} - \mathbf{y}^{t-2}\right\|^2 \\
&- \left(4C_1\eta_{\mathbf{y}}^2 (G_g^2 + G_f^2) + 2C_2\eta_{\mathbf{x}}^2\right)\frac{\sigma^2}{B} - \left(2\beta^2 L_f^2 G_g^2 \eta_{\mathbf{y}} + 4\beta^2 C_1 \eta_{\mathbf{y}}^2 L_f^2 G_g^2 + 4\beta^2 C_2 \eta_{\mathbf{x}}^2 L_f^2\right)\frac{\sigma^2}{M} - \frac{72L_f^2}{\mu^2\eta_{\mathbf{x}}}\beta\frac{\sigma^2}{M} \\
&- \left(1 - \frac{\beta}{2}\right)^2 \frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2.
\end{aligned}
$$

Recall our definition of potential function $\tilde{F}^{t+1}$:

$$
\tilde{F}^{t+1} := \hat{F}^{t+1} - \mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{y}^t))\right\|^2 - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2.
$$

Hence we have:

$$
\begin{aligned}
&\mathbb{E}[\tilde{F}^{t+1} - \tilde{F}^t] \\
&\geq \frac{C_1}{2}\eta_{\mathbf{y}}^2 \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 + \frac{C_2}{2}\eta_{\mathbf{x}}^2 \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 + \left(\frac{1}{8\eta_{\mathbf{y}}} - (2C_1\eta_{\mathbf{y}}^2 L_f^2 G_g^2 + 2C_2\eta_{\mathbf{x}}^2 L_f^2) 4(1-\beta)^2 G_g^2\right)\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 \\
&- \mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{y}^t))\right\|^2 + \mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 \\
&- \left(2C_1\eta_{\mathbf{y}}^2 L_f^2 G_g^2 + 2C_2\eta_{\mathbf{x}}^2 L_f^2 + L_f^2 G_g^2 \eta_{\mathbf{y}} + \eta_{\mathbf{x}} L_f^2\right)(1-\beta)^2 \mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 + \frac{1}{8\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^{t-1} - \mathbf{y}^{t-2}\right\|^2 \\
&- \left(4C_1\eta_{\mathbf{y}}^2 (G_g^2 + G_f^2) + 2C_2\eta_{\mathbf{x}}^2\right)\frac{\sigma^2}{B} - \left(2\beta^2 L_f^2 G_g^2 \eta_{\mathbf{y}} + 4\beta^2 C_1 \eta_{\mathbf{y}}^2 L_f^2 G_g^2 + 4\beta^2 C_2 \eta_{\mathbf{x}}^2 L_f^2\right)\frac{\sigma^2}{M} - \frac{72L_f^2}{\mu^2\eta_{\mathbf{x}}}\beta\frac{\sigma^2}{M} \\
&- \left(1 - \frac{\beta}{2}\right)^2 \frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2 - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2 + \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2.
\end{aligned}
$$

Now we plug in Lemma 1 and 7:

$$
\begin{aligned}
\mathbb{E}[\tilde{F}^{t+1} - \tilde{F}^t] &\geq \frac{C_1}{2}\eta_{\mathbf{y}}^2 \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 + \frac{C_2}{2}\eta_{\mathbf{x}}^2 \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 \\
&+ \left(\frac{1}{8\eta_{\mathbf{y}}} - 4(2C_1\eta_{\mathbf{y}}^2 L_f^2 G_g^2 + 2C_2\eta_{\mathbf{x}}^2 L_f^2 + 1)(1-\beta)^2 G_g^2 - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\left(\frac{12}{\beta} + 6\beta\right)G_g^2\right)\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 \\
&+ \left(1 - \left(2C_1\eta_{\mathbf{y}}^2 L_f^2 G_g^2 + 2C_2\eta_{\mathbf{x}}^2 L_f^2 + L_f^2 G_g^2 \eta_{\mathbf{y}} + \eta_{\mathbf{x}} L_f^2 + 1\right)(1-\beta)^2\right)\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 \\
&+ \left(\frac{1}{8\eta_{\mathbf{y}}} - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\frac{6}{\beta}G_g^2\right)\mathbb{E}\left\|\mathbf{y}^{t-1} - \mathbf{y}^{t-2}\right\|^2 \\
&- \left(4C_1\eta_{\mathbf{y}}^2 (G_g^2 + G_f^2) + 2C_2\eta_{\mathbf{x}}^2\right)\frac{\sigma^2}{B} - \left(2\beta^2 L_f^2 G_g^2 \eta_{\mathbf{y}} + 4\beta^2 C_1 \eta_{\mathbf{y}}^2 L_f^2 G_g^2 + 4\beta^2 C_2 \eta_{\mathbf{x}}^2 L_f^2\right)\frac{\sigma^2}{M} - \frac{72L_f^2}{\mu^2\eta_{\mathbf{x}}}\beta\frac{\sigma^2}{M} \\
&+ (1 - \frac{\beta}{2})^2 \frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\underbrace{\left(\frac{1}{1-(1-\frac{\beta}{2})^2} - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2} - 1\right)}_{=0}\mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2.
\end{aligned}
$$

By our choice of $\eta_{\mathbf{y}}$, we know that

$$\frac{1}{8\eta_{\mathbf{y}}} - 4(2C_1\eta_{\mathbf{y}}^2 L_f^2 G_g^2 + 2C_2\eta_{\mathbf{x}}^2 L_f^2 + 1)(1-\beta)^2 G_f^2 - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\left(\frac{12}{\beta} + 6\beta\right) G_g^2 \geq 0,$$
$$1 - \left(2C_1\eta_{\mathbf{y}}^2 L_f^2 G_g^2 + 2C_2\eta_{\mathbf{x}}^2 L_f^2 + L_f^2 G_g^2 \eta_{\mathbf{y}} + \eta_{\mathbf{x}} L_f^2 + 1\right)(1-\beta)^2 \geq 0,$$
$$\frac{1}{8\eta_{\mathbf{y}}} - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\frac{6}{\beta} G_g^2 \geq 0.$$

Now we can have the clean bound:

$$\mathbb{E}[\tilde{F}^{t+1} - \tilde{F}^t] \geq \frac{C_1}{2}\eta_{\mathbf{y}}^2 \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 + \frac{C_2}{2}\eta_{\mathbf{x}}^2 \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2$$
$$- \left(4C_1\eta_{\mathbf{y}}^2(G_g^2 + G_f^2) + 2C_2\eta_{\mathbf{x}}^2\right)\frac{\sigma^2}{B} - \left(2\beta^2 L_f^2 G_g^2 \eta_{\mathbf{y}} + 4\beta^2 C_1 \eta_{\mathbf{y}}^2 L_f^2 G_g^2 + 4\beta^2 C_2 \eta_{\mathbf{x}}^2 L_f^2\right)\frac{\sigma^2}{M} - \frac{72L_f^2}{\mu^2\eta_{\mathbf{x}}}\beta\frac{\sigma^2}{M}.$$

□

D.2 Proof of Theorem 3

Evoking Lemma 13 and summing the inequality from $t = 1$ to T yields:

$$\frac{\mathbb{E}[\tilde{F}^T - \tilde{F}^0]}{T} + \left(4C_1\eta_{\mathbf{y}}^2(G_g^2 + G_f^2) + 2C_2\eta_{\mathbf{x}}^2\right)\frac{\sigma^2}{B} + \left(2\beta^2 L_f^2 G_g^2 \eta_{\mathbf{y}} + 4\beta^2 C_1 \eta_{\mathbf{y}}^2 L_f^2 G_g^2 + 4\beta^2 C_2 \eta_{\mathbf{x}}^2 L_f^2\right)\frac{\sigma^2}{M} + \frac{72L_f^2}{\mu^2\eta_{\mathbf{x}}}\beta\frac{\sigma^2}{M}$$
$$\geq \frac{1}{T}\sum_{t=0}^{T-1}\frac{C_1}{2}\eta_{\mathbf{y}}^2 \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 + \frac{C_2}{2}\eta_{\mathbf{x}}^2 \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2$$
$$\geq \min\left\{\frac{C_1}{2}\eta_{\mathbf{y}}^2, \frac{C_2}{2}\eta_{\mathbf{x}}^2\right\}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 + \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2$$

We compute the upper and lower bound of C_1 and C_2. For C_1

$$C_1 = \frac{1}{4\eta_{\mathbf{y}}} - \frac{L}{2} - \eta_{\mathbf{x}} L^2 - 4L_f^2 G_f^2 G_g^2 \eta_{\mathbf{y}} - \frac{168 G_g^2 L_f^2}{\mu^2 \eta_{\mathbf{x}} \beta} - \frac{48 G_g^2 L_f^2 \beta}{\mu^2 \eta_{\mathbf{x}}}$$

The upper bound $C_1 \leq \frac{1}{4\eta_{\mathbf{y}}}$ holds trivially. For lower bound, since we choose

$$\eta_{\mathbf{y}} = \min\left\{\frac{1}{20L}, \frac{1}{40\eta_{\mathbf{x}} L^2}, \frac{1}{10L_f^2 G_g^2 G_f^2}, \frac{\mu^2\eta_{\mathbf{x}}\beta}{3840 G_g^2 L_f^2}\right\}$$

we can conclude that $C_1 \geq \frac{1}{8\eta_{\mathbf{y}}}$.

For C_2:

$$C_2 = \frac{1}{\eta_{\mathbf{x}}} + \frac{3}{2}\mu - \frac{\eta_{\mathbf{x}} L^2}{2} - \frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2} - \frac{2L_f^2}{\mu}.$$

The upper bound $C_2 \leq \frac{1}{\eta_{\mathbf{x}}} + \frac{3\mu}{2}$ holds trivially. For lower bound, since we choose:

$$\eta_{\mathbf{x}} = \min\left\{\frac{1}{\sqrt{6}L^2}, \frac{\mu}{12L_f^2}\right\}, \beta = 0.1 \leq 1 - \frac{\sqrt{3}}{2}$$

it holds that $C_2 \geq \frac{1}{3\eta_{\mathbf{x}}}$.

Since $\frac{1}{8\eta_{\mathbf{y}}} \leq C_1 \leq \frac{1}{4\eta_{\mathbf{y}}}$ and $\frac{1}{3\eta_{\mathbf{x}}} \leq C_2 \leq \frac{1}{\eta_{\mathbf{x}}} + \frac{3\mu}{2} \leq \frac{2}{\eta_{\mathbf{x}}}$, we have:

$$\frac{\mathbb{E}[\tilde{F}^T - \tilde{F}^0]}{T} + \left(\eta_{\mathbf{y}}(G_g^2 + G_f^2) + 4\eta_{\mathbf{x}}\right)\frac{\sigma^2}{B} + \left(2\beta^2 L_f^2 G_g^2 \eta_{\mathbf{y}} + \beta^2 \eta_{\mathbf{y}} L_f^2 G_g^2 + 8\beta^2 \eta_{\mathbf{x}} L_f^2\right)\frac{\sigma^2}{M} + \frac{72L_f^2}{\mu^2\eta_{\mathbf{x}}}\beta\frac{\sigma^2}{M}$$
$$\geq \min\left\{\frac{1}{16}\eta_{\mathbf{y}}, \frac{1}{6}\eta_{\mathbf{x}}\right\}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 + \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2$$

We then need to verify the conditions 15, 16 and 17 in Lemma 13 can hold under our choice of $\eta_{\mathbf{x}}$, $\eta_{\mathbf{y}}$ and β. To guarantee (15) holding, we need:

$$\eta_{\mathbf{y}} \leq \min\left\{\frac{1}{64G_f^2(\eta_x L_f^2 G_g^2 + 4\eta_x L_f^2 + 1)}, 2\times 10^{-4}\times\frac{\mu^2\eta_{\mathbf{x}}}{L_f^2 G_g^2}\right\}.$$

To guarantee condition (16) holding, we need

$$\eta_{\mathbf{x}} \leq \frac{23}{1700L_f^2}, \eta_{\mathbf{y}} \leq \frac{23}{500L_f^2 G_g^2}.$$

To guarantee condition (17) holding, we need:

$$\eta_{\mathbf{y}} \leq \frac{\mu^2\eta_{\mathbf{x}}}{14400L_f^2 G_g^2}.$$

Next we examine how large the $\mathbb{E}[\tilde{F}^T - \tilde{F}^0]$ is. By definition of potential function, we have:

$$\begin{aligned}
\mathbb{E}[\tilde{F}^T - \tilde{F}^0] &= \hat{F}^T - \mathbb{E}\left\|\mathbf{z}^T - g(\mathbf{y}^{T-1})\right\|^2 - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2 \\
&\quad - \left(\hat{F}^0 - \mathbb{E}\left\|\mathbf{z}^0 - g(\mathbf{y}^{-1})\right\|^2 - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^0 - \mathbf{z}^{-1}\right\|^2\right) \\
&\leq \hat{F}^T - \hat{F}^0 + \mathbb{E}\left\|\mathbf{z}^0 - g(\mathbf{y}^{-1})\right\|^2 + \frac{(1-\frac{\beta}{2})^2}{\beta - \frac{\beta^2}{4}}\frac{4L_f^2}{\mu^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}_j^0 - \mathbf{z}_j^{-1}\right\|^2.
\end{aligned}$$

By convention $\mathbf{x}^{-1} = \mathbf{x}^0$, $\mathbf{y}^0 = \mathbf{y}^{-1}$, and our choice $\mathbb{E}\left\|\mathbf{z}_j^0 - g(\mathbf{y}^0)\right\|^2 \leq O(1)$, , we have

$$\mathbb{E}[\tilde{F}^T - \tilde{F}^0] \leq \hat{F}^T - \hat{F}^0 + O(1).$$

Next we examine how large the $\mathbb{E}[\hat{F}^T - \hat{F}^0]$ is.

$$\begin{aligned}
\mathbb{E}[\hat{F}^T - \hat{F}^0] &= F(\mathbf{x}^T,\mathbf{y}^T) + s^T - \left(\frac{1}{4\eta_{\mathbf{y}}} + 4L_f^2 G_f^2 G_g^2\eta_{\mathbf{y}} + \frac{\eta_{\mathbf{x}}L^2}{2} + \frac{168G_g^2L_f^2}{\mu^2\eta_{\mathbf{x}}\beta} + \frac{48G_g^2L_f^2\beta}{\mu^2\eta_{\mathbf{x}}}\right)\left\|\mathbf{y}^T - \mathbf{y}^{T-1}\right\|^2 \\
&\quad - \left(\frac{1}{8\eta_{\mathbf{y}}} + \frac{84G_g^2L_f^2}{\mu^2\eta_{\mathbf{x}}\beta}\right)\left\|\mathbf{y}^{T-1} - \mathbf{y}^{T-2}\right\|^2 + \left(\frac{7}{2\eta_{\mathbf{x}}} + \mu - \frac{\eta_{\mathbf{x}}L^2}{2} - \frac{2L_f^2}{\mu}\right)\left\|\mathbf{x}^T - \mathbf{x}^{T-1}\right\|^2 \\
&\quad - F(\mathbf{x}^0,\mathbf{y}^0) - s^0 + \left(\frac{1}{4\eta_{\mathbf{y}}} + 4L_f^2G_f^2G_g^2\eta_{\mathbf{y}} + \frac{\eta_{\mathbf{x}}L^2}{2} + \frac{168G_g^2L_f^2}{\mu^2\eta_{\mathbf{x}}\beta} + \frac{48G_g^2L_f^2\beta}{\mu^2\eta_{\mathbf{x}}}\right)\left\|\mathbf{y}^0 - \mathbf{y}^{-1}\right\|^2 \\
&\quad + \left(\frac{1}{8\eta_{\mathbf{y}}} + \frac{84G_g^2}{\mu^2\eta_{\mathbf{x}}\beta}\right)\left\|\mathbf{y}^{-1} - \mathbf{y}^{-2}\right\|^2 - \left(\frac{7}{2\eta_{\mathbf{x}}} + \mu - \frac{\eta_{\mathbf{x}}L^2}{2} - \frac{2L_f^2}{\mu}\right)\left\|\mathbf{x}^0 - \mathbf{x}^{-1}\right\|^2 \\
&\leq F_{\max} + \left(\frac{7}{2\eta_{\mathbf{x}}} + \mu - \frac{\eta_{\mathbf{x}}L^2}{2} - \frac{2L_f^2}{\mu}\right)\left\|\mathbf{x}^T - \mathbf{x}^{T-1}\right\|^2,
\end{aligned}$$

where we used the convention $\mathbf{x}^{-1} = \mathbf{x}^0$ and $\mathbf{y}^{-1} = \mathbf{y}^{-2}$. Notice that $\mathbb{E}\left\|\mathbf{x}^T - \mathbf{x}^{T-1}\right\|^2 \leq \eta_{\mathbf{x}}^2 \mathbb{E}\left\|\nabla_1 f(\mathbf{x}^{T-1}, \mathbf{z}^T; \mathcal{B}^{T-1}) + \nabla h(\mathbf{x}^{T-1})\right\|^2 \leq 2\eta_{\mathbf{x}}^2(G_f^2 + G_h^2)$, we have $\mathbb{E}[\hat{F}^T - \hat{F}^0] \leq F_{\max} + 7\eta_{\mathbf{x}}(G_f^2 + G_h^2) + 2\mu\eta_{\mathbf{x}}^2(G_f^2 + G_h^2)$.

Using the fact that $\beta - \frac{\beta^2}{2} \geq \frac{1}{2}\beta$ when $\beta \leq 2$ yield:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\left\|\nabla G(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 &\leq \max\left\{\frac{16}{\eta_{\mathbf{y}}}, \frac{6}{\eta_{\mathbf{x}}}\right\} \cdot \frac{F_{\max} + 7\eta_{\mathbf{x}}(G_f^2 + G_h^2) + 2\mu\eta_{\mathbf{x}}^2(G_f^2 + G_h^2) + O(1)}{T} \\
&\quad + \max\left\{\frac{16}{\eta_{\mathbf{y}}}, \frac{6}{\eta_{\mathbf{x}}}\right\} \cdot \left(\left(\eta_{\mathbf{y}}(G_g^2 + G_f^2) + 4\eta_{\mathbf{x}}\right)\frac{\sigma^2}{B} + \left(2\beta^2 L_f^2 G_g^2 \eta_{\mathbf{y}} + \beta^2 \eta_{\mathbf{y}} L_f^2 G_g^2 + 8\beta^2 \eta_{\mathbf{x}} L_f^2\right)\frac{\sigma^2}{M} + \frac{72 L_f^2}{\mu^2 \eta_{\mathbf{x}}}\beta\frac{\sigma^2}{M}\right) \\
&\leq O\left(\frac{F_{\max} + \eta_{\mathbf{x}}(G_f^2 + G_h^2) + \mu\eta_{\mathbf{x}}^2(G_f^2 + G_h^2)}{\eta_{\mathbf{y}} T}\right) + O\left(G_g^2 + G_f^2 + L_f^2 G_g^2 + \frac{\eta_{\mathbf{x}} L_f^2}{\eta_{\mathbf{y}}} + \frac{\kappa^2}{\eta_{\mathbf{x}}}\right)\max\left\{\frac{\sigma^2}{B}, \frac{\sigma^2}{M}\right\}.
\end{aligned}
$$

To guarantee RHS is less than ϵ^2, we need: $T = O\left(\frac{F_{\max}}{\eta_{\mathbf{y}}\epsilon^2}\right)$, and

$$
\begin{aligned}
M &= \Theta\left(\max\left\{\frac{\kappa^3 L \sigma^2}{\epsilon^2} 1\right\}\right), B = \Theta\left(\max\left\{\frac{\kappa^2 L_f^2 \sigma^2}{\epsilon^2}, 1\right\}\right), \beta = 0.1, \\
\eta_{\mathbf{x}} &= \Theta\left(\min\left\{\frac{1}{L^2}, \frac{\mu}{L_f^2}\right\}\right), \eta_{\mathbf{y}} = \Theta\left(\min\left\{\frac{1}{L}, \frac{1}{\eta_{\mathbf{x}} L^2}, \frac{1}{L_f^2 G_g^2 G_f^2}, \frac{\mu^2 \eta_{\mathbf{x}}}{G_g^2 L_f^2}\right\}\right)
\end{aligned}
$$

which yields the total gradient complexity:

$$
O\left(\max\left\{\frac{\kappa^2 L_f^2 \sigma^2}{\epsilon^2}, 1\right\} \cdot \frac{\kappa^3 F_{\max}}{\epsilon^2}\right).
$$

D.3 Proof of Convex-nonconcave Setting

Lemma 14 (Connection between stationary measure and iterates)**.** *For Algorithm 2, under assumptions of Theorem 4, if we define*

$$
\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t, \mathbf{y}) = \frac{1}{\eta_{\mathbf{x}}}\left(\mathbf{x}^t - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}^t - \eta_{\mathbf{x}}\nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}) - \eta_{\mathbf{x}}\alpha_t \mathbf{x}^t\right)\right), \hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}, \mathbf{y}) = \frac{1}{\eta_{\mathbf{y}}}\left(\mathbf{y} - \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y} + \eta_{\mathbf{y}}\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y})\right)\right)
$$

, then the following statement holds:

$$
\begin{aligned}
\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 &\leq \frac{2}{\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathbf{x}^t - \mathbf{x}^{t+1}\right\|^2 + 8L_f^2\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{y}^t)\right\|^2 + 8\frac{\sigma^2}{B}, \\
\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 &\leq \frac{2}{\eta_{\mathbf{y}}^2}\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t+1}\right\|^2 + 4L_f^2 G_g^2\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{y}^t)\right\|^2 + \frac{8(G_f^2 + G_g^2)\sigma^2}{B}.
\end{aligned}
$$

Proof. We begin with proving the first statement. According to updating rule we have

$$\begin{aligned}
\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t,\mathbf{y}^t))\right\|^2 &= \mathbb{E}\left\|\frac{1}{\eta_{\mathbf{x}}}\left(\mathbf{x}^t - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}^t - \eta_{\mathbf{x}}\nabla_{\mathbf{x}}F(\mathbf{x}^t,\mathbf{y}^t) - \eta_{\mathbf{x}}\alpha_t\mathbf{x}^t\right)\right)\right\|^2 \\
&\leq \frac{2}{\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathbf{x}^t - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}^t - \eta_{\mathbf{x}}\mathbf{g}_{\mathbf{x}}^t\right)\right\|^2 + \frac{2}{\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathcal{P}_{\mathcal{X}}\left(\mathbf{x}^t - \eta\mathbf{g}_{\mathbf{x}}^t\right) - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}^t - \eta\nabla_{\mathbf{x}}F(\mathbf{x}^t,\mathbf{y}^t) - \eta_{\mathbf{x}}\alpha_t\mathbf{x}^t\right)\right\|^2 \\
&\leq \frac{2}{\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathbf{x}^t - \mathbf{x}^{t+1}\right\|^2 + 2\mathbb{E}\left\|\frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t}\nabla_1 f(\mathbf{x}^t,\mathbf{z}^{t+1};\zeta) - \nabla_{\mathbf{x}} f(\mathbf{x}^t,\mathbf{z}^{t+1})\right\|^2 \\
&\leq \frac{2}{\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathbf{x}^t - \mathbf{x}^{t+1}\right\|^2 + 8\mathbb{E}\left\|\frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t}\nabla_1 f(\mathbf{x}^t,\mathbf{z}^{t+1};\zeta) - \nabla_1 f(\mathbf{x}^t,\mathbf{z}^{t+1})\right\|^2 \\
&\quad + 8\mathbb{E}\left\|\nabla_1 f(\mathbf{x}^t,\mathbf{z}^{t+1}) - \nabla_1 f(\mathbf{x}^t,g(\mathbf{y}^t))\right\|^2 \\
&\leq \frac{2}{\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathbf{x}^t - \mathbf{x}^{t+1}\right\|^2 + 8L_f^2\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{y}^t)\right\|^2 + 8\frac{\sigma^2}{B}
\end{aligned}$$

where at last step we apply Lemma 8. Similarly, for the second statement we have:

$$\begin{aligned}
\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 &= \mathbb{E}\left\|\frac{1}{\eta_{\mathbf{y}}}\left(\mathbf{y}^t - \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y}^t + \eta\nabla_{\mathbf{y}}F(\mathbf{x}^t,\mathbf{y}^t)\right)\right)\right\|^2 \\
&\leq \frac{2}{\eta_{\mathbf{y}}^2}\mathbb{E}\left\|\mathbf{y}^t - \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y}^t + \eta_{\mathbf{y}}\mathbf{g}_{\mathbf{y}}^t\right)\right\|^2 + \frac{2}{\eta_{\mathbf{y}}^2}\mathbb{E}\left\|\mathcal{P}_{\mathcal{Y}}\left(\mathbf{y}^t + \eta_{\mathbf{y}}\mathbf{g}_{\mathbf{y}}^t\right) - \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y}^t + \eta_{\mathbf{y}}\nabla_{\mathbf{y}}F(\mathbf{x}^t,\mathbf{y}^t)\right)\right\|^2 \\
&\leq \frac{2}{\eta_{\mathbf{y}}^2}\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t+1}\right\| + 2\mathbb{E}\left\|\mathbf{g}_{\mathbf{y}}^t - \nabla_{\mathbf{y}}F(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 \\
&= \frac{2}{\eta_{\mathbf{y}}^2}\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t+1}\right\| + 2\mathbb{E}\left\|\frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t}\nabla_2 f(\mathbf{x}^t,\mathbf{z}^{t+1};\zeta)\nabla g(\mathbf{y}^t;\xi) - \nabla_2 f(\mathbf{x}^t,g(\mathbf{y}^t))\nabla g(\mathbf{y}^t)\right\|^2 \\
&= \frac{2}{\eta_{\mathbf{y}}^2}\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t+1}\right\| + 4\mathbb{E}\left\|\frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t}\nabla_2 f(\mathbf{x}^t,\mathbf{z}^{t+1};\zeta)\nabla g(\mathbf{y}^t;\xi) - \nabla_2 f(\mathbf{x}^t,\mathbf{z}^{t+1})\nabla g(\mathbf{y}^t)\right\|^2 \\
&\quad + 4\mathbb{E}\left\|\nabla_2 f(\mathbf{x}^t,\mathbf{z}^{t+1})\nabla g(\mathbf{y}^t) - \nabla_2 f(\mathbf{x}^t,g(\mathbf{y}^t))\nabla g(\mathbf{y}^t)\right\|^2 \\
&\leq \frac{2}{\eta_{\mathbf{y}}^2}\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t+1}\right\|^2 + 4L_f^2G_g^2\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{y}^t)\right\|^2 + \frac{8(G_f^2 + G_g^2)\sigma^2}{B}.
\end{aligned}$$

□

Proposition 1. *For Algorithm 2, under assumptions of Theorem 4, the following statement holds:*

$$\begin{aligned}
&\left\langle\nabla F_{t-1}(\mathbf{x}^t,\mathbf{y}^{t-1}) - \nabla F_{t-1}(\mathbf{x}^{t-1},\mathbf{y}^{t-1}),\, \mathbf{x}^t - \mathbf{x}^{t-1}\right\rangle \\
&\geq \frac{1}{2(L' + \alpha_{t-1})}\left\|\nabla F_{t-1}(\mathbf{x}^t,\mathbf{y}^{t-1}) - \nabla F_{t-1}(\mathbf{x}^{t-1},\mathbf{y}^{t-1})\right\|^2 + \frac{\alpha_{t-1}L'}{2(L' + \alpha_{t-1})}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2.
\end{aligned}$$

Proof.

$$
\begin{aligned}
&\left\langle \nabla F_{t-1}(\mathbf{x}^t, \mathbf{y}^{t-1}) - \nabla F_{t-1}(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}),\, \mathbf{x}^t - \mathbf{x}^{t-1} \right\rangle \\
=& \frac{\frac{L'}{2}}{L' + \alpha_{t-1}} \underbrace{\left\langle \nabla F_{t-1}(\mathbf{x}^t, \mathbf{y}^{t-1}) - \nabla F_{t-1}(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}),\, \mathbf{x}^t - \mathbf{x}^{t-1} \right\rangle}_{A} \\
&+ \frac{\frac{L'}{2} + \alpha_{t-1}}{L' + \alpha_{t-1}} \underbrace{\left\langle \nabla F_{t-1}(\mathbf{x}^t, \mathbf{y}^{t-1}) - \nabla F_{t-1}(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}),\, \mathbf{x}^t - \mathbf{x}^{t-1} \right\rangle}_{B} \\
\geq& \frac{1}{2(L' + \alpha_{t-1})} \left\| \nabla F_{t-1}(\mathbf{x}^t, \mathbf{y}^{t-1}) - \nabla F_{t-1}(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}) \right\|^2 + \frac{\alpha_{t-1} L'}{2(L' + \alpha_{t-1})} \left\| \mathbf{x}^t - \mathbf{x}^{t-1} \right\|^2
\end{aligned}
$$

where the last inequality is due to
(1) $F_{t-1}(\mathbf{x}, \mathbf{y}^{t-1})$ is L' smooth and hence $A \geq \frac{1}{L'} \left\| \nabla F_{t-1}(\mathbf{x}^t, \mathbf{y}^{t-1}) - \nabla F_{t-1}(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}) \right\|^2$; and
(2) $F_{t-1}(\mathbf{x}, \mathbf{y}^{t-1})$ is α_{t-1} strongly convex and $B \geq \alpha_{t-1} \left\| \mathbf{x}^t - \mathbf{x}^{t-1} \right\|^2$,

∎

Lemma 15. *For Algorithm 2, under assumptions of Theorem 4, then the following statement holds:*

$$
\begin{aligned}
\mathbb{E}[F(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}) - F(\mathbf{x}^t, \mathbf{y}^t)] \geq& -\left(\frac{L^2 \eta_{\mathbf{x}}}{2} + 4(1-\beta)^2 L_f^2 G_g^4 \eta_{\mathbf{y}} \right) \mathbb{E}\|\mathbf{y}^t - \mathbf{y}^{t-1}\|^2 + \left(\frac{1}{2\eta_{\mathbf{y}}} - \frac{L}{2} - \frac{L^2 \eta_{\mathbf{x}}}{2} \right) \mathbb{E}\|\mathbf{y}^{t+1} - \mathbf{y}^t\|^2 \\
&+ \left(\frac{\alpha_{t-1} L'}{2(L' + \alpha_{t-1})} - \frac{1}{2\eta_{\mathbf{x}}} \right) \mathbb{E} \left\| \mathbf{x}^t - \mathbf{x}^{t-1} \right\|^2 + \left(\frac{\alpha_{t-1}}{2} - \frac{3}{2\eta_{\mathbf{x}}} - \frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2} \right) \mathbb{E}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 \\
&- \frac{1}{2}\alpha_{t-1}(\mathbb{E} \left\| \mathbf{x}^{t+1} \right\|^2 - \mathbb{E} \left\| \mathbf{x}^t \right\|^2) - \left(2\eta_{\mathbf{x}} L_f^2 + L_f^2 G_g^2 \eta_{\mathbf{y}} \right) (1-\beta)^2 \mathbb{E} \left\| \mathbf{z}^t - g(\mathbf{y}^{t-1}) \right\|^2 \\
&- 2\eta_{\mathbf{x}}(1-\beta)^2 \frac{\sigma^2}{B} - \left(2\eta_{\mathbf{y}}(G_g^2 + G_f^2) \frac{\sigma^2}{B} + 2\beta^2 L_f^2 G_g^2 \eta_{\mathbf{y}} \frac{\sigma^2}{M} \right).
\end{aligned}
$$

Proof. According to updating rule:

$$\left\langle \mathbf{g}_{\mathbf{x}}^t + \frac{1}{\eta_{\mathbf{x}}}(\mathbf{x}^{t+1} - \mathbf{x}^t), \mathbf{x} - \mathbf{x}^{t+1} \right\rangle \geq 0 \tag{18}$$

$$\left\langle \mathbf{g}_{\mathbf{x}}^t + \frac{1}{\eta_{\mathbf{x}}}(\mathbf{x}^{t+1} - \mathbf{x}^t), \mathbf{x}^t - \mathbf{x}^{t+1} \right\rangle \geq 0 \tag{19}$$

$$\left\langle \mathbf{g}_{\mathbf{x}}^{t-1} + \frac{1}{\eta_{\mathbf{x}}}(\mathbf{x}^t - \mathbf{x}^{t-1}), \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle \geq 0. \tag{20}$$

Due to the α_t-strong convexity of $F_t(\cdot, \mathbf{y})$, together with (20) we have

$$
\begin{aligned}
F_t(\mathbf{x}^{t+1}, \mathbf{y}^t) - F_t(\mathbf{x}^t, \mathbf{y}^t) \geq& \left\langle \nabla F_t(\mathbf{x}^t, \mathbf{y}^t),\, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle + \frac{1}{2}\alpha_t \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|^2 \\
\geq& \left\langle \nabla F_t(\mathbf{x}^t, \mathbf{y}^t) - \mathbf{g}_{\mathbf{x}}^{t-1},\, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle - \frac{1}{\eta_{\mathbf{x}}} \left\langle \mathbf{x}^t - \mathbf{x}^{t-1},\, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle + \frac{1}{2}\alpha_t \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|^2 \\
=& \underbrace{\left\langle \nabla F_t(\mathbf{x}^t, \mathbf{y}^t) - \nabla F_{t-1}(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}),\, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle}_{A} - \underbrace{\frac{1}{\eta_{\mathbf{x}}} \left\langle \mathbf{x}^t - \mathbf{x}^{t-1},\, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle}_{B} \\
&+ \underbrace{\left\langle \nabla F_{t-1}(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}) - \mathbf{g}_{\mathbf{x}}^{t-1},\, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle}_{C} + \frac{1}{2}\alpha_t \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|^2
\end{aligned}
$$

Similar to the proof of Lemma 11, for A, we observe that:

$$
\begin{aligned}
\left\langle \nabla_{\mathbf{x}} F_t(\mathbf{x}^t, \mathbf{y}^t) - \nabla_{\mathbf{x}} F_{t-1}(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}), \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle &= \underbrace{\left\langle \nabla_{\mathbf{x}} F_t(\mathbf{x}^t, \mathbf{y}^t) - \nabla_{\mathbf{x}} F_{t-1}(\mathbf{x}^t, \mathbf{y}^{t-1}), \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle}_{①} \\
&+ \underbrace{\left\langle \nabla_{\mathbf{x}} F_{t-1}(\mathbf{x}^t, \mathbf{y}^{t-1}) - \nabla_{\mathbf{x}} F_{t-1}(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}), \Delta \right\rangle}_{②} \\
&+ \underbrace{\left\langle \nabla_{\mathbf{x}} F_{t-1}(\mathbf{x}^t, \mathbf{y}^{t-1}) - \nabla_{\mathbf{x}} F_{t-1}(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}), \mathbf{x}^t - \mathbf{x}^{t-1} \right\rangle}_{③}
\end{aligned}
$$

where $\Delta = (\mathbf{x}^{t+1} - \mathbf{x}^t) - (\mathbf{x}^t - \mathbf{x}^{t-1})$. For ①, recall the definition of F_t:

$$
\begin{aligned}
① &= \left\langle \nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}^t) - \nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}^{t-1}) + \alpha_t \mathbf{x}^t - \alpha_{t-1}\mathbf{x}^t,\, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle \\
&= \left\langle \nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}^t) - \nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}^{t-1}),\, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle + \left\langle \alpha_t \mathbf{x}^t - \alpha_{t-1}\mathbf{x}^t,\, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle \\
&\geq -\frac{L^2 \eta_{\mathbf{x}}}{2}\|\mathbf{y}^t - \mathbf{y}^{t-1}\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 + (\alpha_t - \alpha_{t-1})\left\langle \mathbf{x}^t,\, \mathbf{x}^{t+1} \right\rangle - (\alpha_t - \alpha_{t-1})\left\|\mathbf{x}^t\right\|^2 \\
&= -\frac{L^2 \eta_{\mathbf{x}}}{2}\|\mathbf{y}^t - \mathbf{y}^{t-1}\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 - \frac{1}{2}(\alpha_{t-1} - \alpha_t)\left(\left\|\mathbf{x}^{t+1}\right\|^2 - \left\|\mathbf{x}^t\right\|^2\right) + \frac{1}{2}(\alpha_{t-1} - \alpha_t)\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2.
\end{aligned}
$$

For ②, we use Cauchy-Schwartz:

$$
② \geq -\frac{\eta_{\mathbf{x}}}{2}\|\nabla F_{t-1}(\mathbf{x}^t, \mathbf{y}^{t-1}) - \nabla F_{t-1}(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\|\Delta\|^2.
$$

For ③, we use Proposition 1:

$$
\begin{aligned}
&\left\langle \nabla F_{t-1}(\mathbf{x}^t, \mathbf{y}^{t-1}) - \nabla F_{t-1}(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}),\, \mathbf{x}^t - \mathbf{x}^{t-1} \right\rangle \\
&\geq \frac{1}{2(L' + \alpha_{t-1})}\left\|\nabla F_{t-1}(\mathbf{x}^t, \mathbf{y}^{t-1}) - \nabla F_{t-1}(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\right\|^2 + \frac{\alpha_{t-1} L'}{2(L' + \alpha_{t-1})}\|\mathbf{x}^t - \mathbf{x}^{t-1}\|^2
\end{aligned}
$$

Since $\frac{1}{2(L'+\alpha_{t-1})} - \frac{\eta_{\mathbf{x}}}{2} \geq 0$, putting pieces together yields:

$$
\begin{aligned}
A \geq &-\frac{L^2 \eta_{\mathbf{x}}}{2}\|\mathbf{y}^t - \mathbf{y}^{t-1}\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 - \frac{1}{2}(\alpha_{t-1} - \alpha_t)\left(\left\|\mathbf{x}^{t+1}\right\|^2 - \left\|\mathbf{x}^t\right\|^2\right) + \frac{1}{2}(\alpha_{t-1} - \alpha_t)\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 \\
&- \frac{1}{2\eta_{\mathbf{x}}}\|\Delta\|^2.
\end{aligned}
$$

For B, we have:

$$
\frac{1}{\eta_{\mathbf{x}}}\left\langle \mathbf{x}^t - \mathbf{x}^{t-1},\, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle = \frac{1}{2\eta_{\mathbf{x}}}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 + \frac{1}{2\eta_{\mathbf{x}}}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\left\|\Delta\right\|^2.
$$

For C, we have:

$$
\begin{aligned}
\mathbb{E}\left\langle \mathbf{g}_{\mathbf{x}}^{t-1} - \nabla_{\mathbf{x}} F_{t-1}(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}), \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle &\geq -\eta_{\mathbf{x}}(1-\beta)^2 \mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^{t-1} - \nabla_{\mathbf{x}} F_{t-1}(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\right\|^2 \\
&\quad - \frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}\mathbb{E}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 \\
&\geq -2\eta_{\mathbf{x}}(1-\beta)^2 L_f^2 \mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 - 2\eta_{\mathbf{x}}(1-\beta)^2\frac{\sigma^2}{B} \\
&\quad - \frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}\mathbb{E}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2,
\end{aligned}
$$

where we plug in Lemma 8 at last step.

Putting lower bound of A, B and C together yields:

$$F_t(\mathbf{x}^{t+1},\mathbf{y}^t) - F_t(\mathbf{x}^t,\mathbf{y}^t) \geq -\frac{L^2\eta_{\mathbf{x}}}{2}\|\mathbf{y}^t-\mathbf{y}^{t-1}\|^2 + \left(\frac{1}{2}\alpha_{t-1} - \frac{1}{\eta_{\mathbf{x}}} - \frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}\right)\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2$$
$$\left(\frac{\alpha_{t-1}L'}{2(L'+\alpha_{t-1})} - \frac{1}{2\eta_{\mathbf{x}}}\right)\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2 - \frac{1}{2}\left(\alpha_{t-1}-\alpha_t\right)\left(\left\|\mathbf{x}^{t+1}\right\|^2 - \left\|\mathbf{x}^t\right\|^2\right)$$
$$- 2\eta_{\mathbf{x}}(1-\beta)^2L_f^2\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 - 2\eta_{\mathbf{x}}(1-\beta)^2\frac{\sigma^2}{B}.$$

Since $F_t(\mathbf{x}^{t+1},\mathbf{y}^t) - F_t(\mathbf{x}^t,\mathbf{y}^t) = F(\mathbf{x}^{t+1},\mathbf{y}^t) - F(\mathbf{x}^t,\mathbf{y}^t) + \frac{1}{2}\alpha_t(\left\|\mathbf{x}^{t+1}\right\|^2 - \left\|\mathbf{x}^t\right\|^2)$, and also recall in Lemma 10,

$$\mathbb{E}[F(\mathbf{x}^{t+1},\mathbf{y}^{t+1}) - F(\mathbf{x}^{t+1},\mathbf{y}^t)] \geq \left(\frac{1}{2\eta_{\mathbf{y}}} - \frac{L}{2} - \frac{L^2\eta_{\mathbf{x}}}{2}\right)\mathbb{E}\|\mathbf{y}^{t+1}-\mathbf{y}^t\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2$$
$$- \left(\eta_{\mathbf{y}}2(G_g^2+G_f^2)\frac{\sigma^2}{B} + 2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M}\right)$$
$$- L_f^2G_g^2\eta_{\mathbf{y}}(1-\beta)^2\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 - 4(1-\beta)^2L_f^2G_g^4\eta_{\mathbf{y}}\mathbb{E}\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2,$$

we can conclude the proof:

$$\mathbb{E}[F(\mathbf{x}^{t+1},\mathbf{y}^{t+1}) - F(\mathbf{x}^t,\mathbf{y}^t)] \geq -\left(\frac{L^2\eta_{\mathbf{x}}}{2} + 4(1-\beta)^2L_f^2G_g^4\eta_{\mathbf{y}}\right)\mathbb{E}\|\mathbf{y}^t-\mathbf{y}^{t-1}\|^2 + \left(\frac{1}{2\eta_{\mathbf{y}}} - \frac{L}{2} - \frac{L^2\eta_{\mathbf{x}}}{2}\right)\mathbb{E}\|\mathbf{y}^{t+1}-\mathbf{y}^t\|^2$$
$$+ \left(\frac{\alpha_{t-1}L'}{2(L'+\alpha_{t-1})} - \frac{1}{2\eta_{\mathbf{x}}}\right)\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2 + \left(\frac{\alpha_{t-1}}{2} - \frac{3}{2\eta_{\mathbf{x}}} - \frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}\right)\mathbb{E}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2$$
$$- \frac{1}{2}\alpha_{t-1}(\mathbb{E}\left\|\mathbf{x}^{t+1}\right\|^2 - \mathbb{E}\left\|\mathbf{x}^t\right\|^2) - \left(2\eta_{\mathbf{x}}L_f^2 + L_f^2G_g^2\eta_{\mathbf{y}}\right)(1-\beta)^2\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2$$
$$- 2\eta_{\mathbf{x}}(1-\beta)^2\frac{\sigma^2}{B} - \left(2\eta_{\mathbf{y}}(G_g^2+G_f^2)\frac{\sigma^2}{B} + 2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M}\right).$$

□

Lemma 16. *For Algorithm 2, under assumptions of Theorem 4, define the following auxiliary function*

$$\hat{F}^{t+1} := F(\mathbf{x}^{t+1},\mathbf{y}^{t+1}) + s^{t+1} - \left(\frac{1}{4\eta_{\mathbf{y}}} + 4L_f^2G_g^4\eta_{\mathbf{y}} + \frac{\eta_{\mathbf{x}}L^2}{2} + \frac{768G_g^2L_f^2}{\alpha_{t+1}^2\eta_{\mathbf{x}}\beta} + \frac{576G_g^2L_f^2\beta}{\alpha_{t+1}^2\eta_{\mathbf{x}}}\right)\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2$$
$$- \left(\frac{1}{8\eta_{\mathbf{y}}} + \frac{768G_g^2L_f^2}{\alpha_{t+1}^2\eta_{\mathbf{x}}\beta}\right)\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2 + \frac{7}{2\eta_{\mathbf{x}}}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2 + \frac{\alpha_t}{2}\left\|\mathbf{x}^{t+1}\right\|^2,$$

where $s^{t+1} := -\frac{8}{\eta_{\mathbf{x}}^2\alpha_{t+1}}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2$ *then the following statement holds true:*

$$\mathbb{E}[\hat{F}^{t+1} - \hat{F}^t] \geq C_1^t\mathbb{E}\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2 + \left(\frac{1}{8\eta_{\mathbf{y}}} - \frac{768G_g^2L_f^2}{\alpha_{t+1}^2\eta_{\mathbf{x}}\beta}\right)\mathbb{E}\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2 + \frac{1}{8\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^{t-1}-\mathbf{y}^{t-2}\right\|^2$$
$$+ C_2^t\mathbb{E}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2 - (1-\beta)^2\left(L_f^2G_g^2\eta_{\mathbf{y}} + 2\eta_{\mathbf{x}}L_f^2\right)\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 + \frac{1}{2}\left(\alpha_t-\alpha_{t-1}\right)\left\|\mathbf{x}^{t+1}\right\|^2$$
$$- \left(1-\frac{\beta}{2}\right)^2\frac{64L_f^2}{\alpha_t^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^t-\mathbf{z}^{t-1}\right\|^2 + \frac{8}{\eta_{\mathbf{x}}}\left(\frac{\alpha_t}{\alpha_{t+1}} - \frac{\alpha_{t-1}}{\alpha_t}\right)\left\|\mathbf{x}^{t+1}\right\|^2$$
$$- 2\eta_{\mathbf{x}}(1-\beta)^2\frac{\sigma^2}{B} - \left(2\eta_{\mathbf{y}}(G_g^2+G_f^2)\frac{\sigma^2}{B} + 2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M} + \frac{1152L_f^2\beta}{\alpha_t^2\eta_{\mathbf{x}}}\frac{\sigma^2}{M}\right),$$

where

$$C_1^t = \frac{1}{4\eta_{\mathbf{y}}} - \frac{L}{2} - \eta_{\mathbf{x}}L^2 - 4L_f^2G_g^4\eta_{\mathbf{y}} - \frac{768G_g^2L_f^2}{\alpha_{t+1}^2\eta_{\mathbf{x}}\beta} - \frac{576G_g^2L_f^2\beta}{\alpha_{t+1}^2\eta_{\mathbf{x}}}, \tag{21}$$
$$C_2^t = \left(\frac{1}{\eta_{\mathbf{x}}} + \frac{\alpha_{t-1}}{2} - \frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}\right). \tag{22}$$

Proof. Define $\Delta^{t+1} := \mathbf{x}^{t+1} - \mathbf{x}^t - (\mathbf{x}^t - \mathbf{x}^{t-1})$. According to (6) and (7):

$$\frac{1}{\eta}\left\langle \Delta^{t+1}, \mathbf{x}^t - \mathbf{x}^{t+1}\right\rangle \geq \left\langle \mathbf{g}_{\mathbf{x}}^t - \mathbf{g}_{\mathbf{x}}^{t-1}, \mathbf{x}^{t+1} - \mathbf{x}^t\right\rangle.$$

If we define $\mathbf{g}_{\mathbf{x}}^t(\mathbf{x}, \mathbf{z}) = \nabla_1 f(\mathbf{x}, \mathbf{z}; \zeta^t) + \nabla h(\mathbf{x}) + \alpha_t \mathbf{x}$.

$$\begin{aligned}
\frac{1}{\eta}\left\langle \Delta^{t+1}, \mathbf{x}^t - \mathbf{x}^{t+1}\right\rangle &\geq \left\langle \mathbf{g}_{\mathbf{x}}^t - \mathbf{g}_{\mathbf{x}}^t(\mathbf{x}^t, \mathbf{z}^t), \mathbf{x}^{t+1} - \mathbf{x}^t\right\rangle \\
&\quad + \left\langle \mathbf{g}_{\mathbf{x}}^t(\mathbf{x}^t, \mathbf{z}^t) - \mathbf{g}_{\mathbf{x}}^{t-1}, \mathbf{x}^{t+1} - \mathbf{x}^t\right\rangle \\
&\geq \left\langle \mathbf{g}_{\mathbf{x}}^t - \mathbf{g}_{\mathbf{x}}^{t-1}(\mathbf{x}^t, \mathbf{z}^t), \mathbf{x}^{t+1} - \mathbf{x}^t\right\rangle \\
&\quad + \left\langle \mathbf{g}_{\mathbf{x}}^{t-1}(\mathbf{x}^t, \mathbf{z}^t) - \mathbf{g}_{\mathbf{x}}^{t-1}, \mathbf{x}^{t+1} - \mathbf{x}^t - (\mathbf{x}^t - \mathbf{x}^{t-1})\right\rangle \\
&\quad + \left\langle \mathbf{g}_{\mathbf{x}}^{t-1}(\mathbf{x}^t, \mathbf{z}^t) - \mathbf{g}_{\mathbf{x}}^{t-1}, (\mathbf{x}^t - \mathbf{x}^{t-1})\right\rangle \\
&\geq -\frac{L_f^2}{2c_t}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2 - \frac{c_t}{2}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 + \frac{\alpha_t - \alpha_{t-1}}{2}\left(\left\|\mathbf{x}^{t+1}\right\|^2 - \left\|\mathbf{x}^t\right\|^2\right) \\
&\quad - \frac{\eta_{\mathbf{x}}}{2}\left\|\mathbf{g}_{\mathbf{x}}^{t-1}(\mathbf{x}^t, \mathbf{z}^t) - \mathbf{g}_{\mathbf{x}}^{t-1}\right\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\left\|\Delta^{t+1}\right\|^2 \\
&\quad + \frac{1}{2(L' + \alpha_{t-1})}\left\|\mathbf{g}_{\mathbf{x}}^{t-1}(\mathbf{x}^t, \mathbf{z}^t) - \mathbf{g}_{\mathbf{x}}^{t-1}\right\|^2 + \frac{\alpha_{t-1} L'}{2(L' + \alpha_{t-1})}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2
\end{aligned}$$

where we use Proposition 1 at the last step. Since $\frac{1}{\eta_{\mathbf{x}}}\left\langle \Delta^{t+1}, \mathbf{x}^t - \mathbf{x}^{t+1}\right\rangle = \frac{1}{2\eta_{\mathbf{x}}}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\left\|\Delta^{t+1}\right\|^2$, and $\frac{1}{2(L'+\alpha_{t-1})} - \frac{\eta_{\mathbf{x}}}{2} \geq 0$, and $L' \geq \alpha_{t-1}$ we have:

$$\begin{aligned}
\frac{1}{2\eta_{\mathbf{x}}}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 &\geq -\frac{L_f^2}{2c_t}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2 - \frac{c_t}{2}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 + \frac{\alpha_t - \alpha_{t-1}}{2}\left(\left\|\mathbf{x}^{t+1}\right\|^2 - \left\|\mathbf{x}^t\right\|^2\right) \\
&\quad + \frac{\alpha_{t-1}}{4}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2.
\end{aligned}$$

Rearranging terms yields:

$$\begin{aligned}
\frac{1}{2\eta_{\mathbf{x}}}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 + \frac{\alpha_t - \alpha_{t-1}}{2}\left\|\mathbf{x}^t\right\|^2 - \frac{1}{2\eta_{\mathbf{x}}}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 - \frac{\alpha_t - \alpha_{t-1}}{2}\left\|\mathbf{x}^{t+1}\right\|^2 \\
\geq -\frac{L_f^2}{2c_t}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2 - \frac{c_t}{2}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 + \frac{\alpha_t}{4}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2,
\end{aligned}$$

where we use the fact $\alpha_{t-1} \geq \alpha_t$.

Multiplying both sides with $\frac{16}{\alpha_t \eta_{\mathbf{x}}}$ yields:

$$\frac{8}{\alpha_t \eta_{\mathbf{x}}^2}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 + \frac{8}{\eta_{\mathbf{x}}}\left(1 - \frac{\alpha_{t-1}}{\alpha_t}\right)\left\|\mathbf{x}^t\right\|^2 - \frac{8}{\alpha_t \eta_{\mathbf{x}}^2}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 - \frac{8}{\eta_{\mathbf{x}}}\left(1 - \frac{\alpha_{t-1}}{\alpha_t}\right)\left\|\mathbf{x}^{t+1}\right\|^2 \tag{23}$$

$$\geq -\frac{8L_f^2}{\alpha_t c_t \eta_{\mathbf{x}}}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2 - \frac{8c_t}{\alpha_t \eta_{\mathbf{x}}}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 + \frac{4}{\eta_{\mathbf{x}}}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2. \tag{24}$$

Recall the definition of s^t, and choose $c_t = \frac{\alpha_t}{8}$:

$$s^{t+1} - s^t \geq -\frac{64L_f^2}{\alpha_t^2 \eta_{\mathbf{x}}}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2 - \frac{1}{\eta_{\mathbf{x}}}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 + \frac{4}{\eta_{\mathbf{x}}}\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 + \frac{8}{\eta_{\mathbf{x}}}\left(\frac{\alpha_t}{\alpha_{t+1}} - \frac{\alpha_{t-1}}{\alpha_t}\right)\left\|\mathbf{x}^{t+1}\right\|^2.$$

Recall that Lemma 15 gives:

$$\begin{aligned}\mathbb{E}[F(\mathbf{x}^{t+1},\mathbf{y}^{t+1})-F(\mathbf{x}^{t},\mathbf{y}^{t})] \geq & -\left(\frac{L^2\eta_{\mathbf{x}}}{2}+4(1-\beta)^2L_f^2G_g^4\eta_{\mathbf{y}}\right)\mathbb{E}\|\mathbf{y}^t-\mathbf{y}^{t-1}\|^2+\left(\frac{1}{2\eta_{\mathbf{y}}}-\frac{L}{2}-\frac{L^2\eta_{\mathbf{x}}}{2}\right)\mathbb{E}\|\mathbf{y}^{t+1}-\mathbf{y}^t\|^2\\ &+\left(\frac{\alpha_{t-1}L'}{2(L'+\alpha_{t-1})}-\frac{1}{2\eta_{\mathbf{x}}}\right)\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2+\left(\frac{\alpha_{t-1}}{2}-\frac{3}{2\eta_{\mathbf{x}}}-\frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}\right)\mathbb{E}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2\\ &-\frac{1}{2}\alpha_{t-1}(\mathbb{E}\left\|\mathbf{x}^{t+1}\right\|^2-\mathbb{E}\left\|\mathbf{x}^{t}\right\|^2)-\left(2\eta_{\mathbf{x}}L_f^2+L_f^2G_g^2\eta_{\mathbf{y}}\right)(1-\beta)^2\mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{y}^{t-1})\right\|^2\\ &-2\eta_{\mathbf{x}}(1-\beta)^2\frac{\sigma^2}{B}-\left(2\eta_{\mathbf{y}}(G_g^2+G_f^2)\frac{\sigma^2}{B}+2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M}\right).\end{aligned}$$

Evoking Lemma 15 together with (24) yields:

$$\begin{aligned}&\mathbb{E}[F(\mathbf{x}^{t+1},\mathbf{y}^{t+1})+s^{t+1}-(F(\mathbf{x}^{t},\mathbf{y}^{t})+s^t)]\\ &\geq\left(\frac{1}{2\eta_{\mathbf{y}}}-\frac{L}{2}-\frac{L^2\eta_{\mathbf{x}}}{2}\right)\mathbb{E}\|\mathbf{y}^{t+1}-\mathbf{y}^t\|^2-\left(4(1-\beta)^2L_f^2G_g^4\eta_{\mathbf{y}}+\frac{L^2\eta_{\mathbf{x}}}{2}\right)\mathbb{E}\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2\\ &+\left(\frac{\alpha_{t-1}}{2}-\frac{3}{2\eta_{\mathbf{x}}}-\frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}-\frac{1}{\eta_{\mathbf{x}}}\right)\mathbb{E}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2+\left(\frac{4}{\eta_{\mathbf{x}}}+\frac{\alpha_{t-1}L'}{2(L'+\alpha_{t-1})}-\frac{1}{2\eta_{\mathbf{x}}}\right)\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2\\ &-(1-\beta)^2\left(L_f^2G_g^2\eta_{\mathbf{y}}+2\eta_{\mathbf{x}}L_f^2\right)\mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{y}^{t-1})\right\|^2-\frac{64L_f^2}{\alpha_t^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^{t+1}-\mathbf{z}^t\right\|^2\\ &-\frac{1}{2}\alpha_{t-1}(\left\|\mathbf{x}^{t+1}\right\|^2-\left\|\mathbf{x}^{t}\right\|^2)+\frac{8}{\eta_{\mathbf{x}}}\left(\frac{\alpha_t}{\alpha_{t+1}}-\frac{\alpha_{t-1}}{\alpha_t}\right)\left\|\mathbf{x}^{t+1}\right\|^2\\ &-2\eta_{\mathbf{x}}(1-\beta)^2\frac{\sigma^2}{B}-\left(2\eta_{\mathbf{y}}(G_g^2+G_f^2)\frac{\sigma^2}{B}+2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M}\right).\end{aligned}$$

Now we plug in Lemma 7 to replace $\mathbb{E}\left\|\mathbf{z}^{t+1}-\mathbf{z}^t\right\|^2$ together with the fact that $\frac{1}{\beta}\geq 1$ and get:

$$\begin{aligned}&\mathbb{E}[F(\mathbf{x}^{t+1},\mathbf{y}^{t+1})+s^{t+1}-(F(\mathbf{x}^{t},\mathbf{y}^{t})+s^t)]\\ &\geq\left(\frac{1}{2\eta_{\mathbf{y}}}-\frac{L}{2}-\frac{L^2\eta_{\mathbf{x}}}{2}\right)\mathbb{E}\|\mathbf{y}^{t+1}-\mathbf{y}^t\|^2-\left(4(1-\beta)^2L_f^2G_g^4\eta_{\mathbf{y}}+\frac{L^2\eta_{\mathbf{x}}}{2}+\frac{768L_f^2G_g^2}{\alpha_t^2\eta_{\mathbf{x}}\beta}+\frac{576L_f^2G_g^2\beta}{\alpha_t^2\eta_{\mathbf{x}}}\right)\mathbb{E}\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2\\ &+\left(\frac{\alpha_{t-1}}{2}-\frac{5}{2\eta_{\mathbf{x}}}-\frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}\right)\mathbb{E}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2+\frac{7}{2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2\\ &-(1-\beta)^2\left(L_f^2G_g^2\eta_{\mathbf{y}}+2\eta_{\mathbf{x}}L_f^2\right)\mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{y}^{t-1})\right\|^2-\frac{64L_f^2}{\alpha_t^2\eta_{\mathbf{x}}}\left(1-\frac{\beta}{2}\right)^2\mathbb{E}\left\|\mathbf{z}^t-\mathbf{z}^{t-1}\right\|^2-\frac{768L_f^2G_g^2}{\alpha_t^2\eta_{\mathbf{x}}\beta}\left\|\mathbf{y}^{t-1}-\mathbf{y}^{t-2}\right\|^2\\ &-\frac{1}{2}\alpha_{t-1}(\left\|\mathbf{x}^{t+1}\right\|^2-\left\|\mathbf{x}^{t}\right\|^2)+\frac{8}{\eta_{\mathbf{x}}}\left(\frac{\alpha_t}{\alpha_{t+1}}-\frac{\alpha_{t-1}}{\alpha_t}\right)\left\|\mathbf{x}^{t+1}\right\|^2\\ &-2\eta_{\mathbf{x}}(1-\beta)^2\frac{\sigma^2}{B}-\left(2\eta_{\mathbf{y}}(G_g^2+G_f^2)\frac{\sigma^2}{B}+2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M}+\frac{64L_f^2}{\alpha_t^2\eta_{\mathbf{x}}}\cdot 18\beta\frac{\sigma^2}{M}\right).\end{aligned}$$

Recall our definition of potential function $\hat{F}^{t+1}$

$$\begin{aligned}\hat{F}^{t+1}:=&F(\mathbf{x}^{t+1},\mathbf{y}^{t+1})+s^{t+1}-\left(\frac{1}{4\eta_{\mathbf{y}}}+4L_f^2G_g^4\eta_{\mathbf{y}}+\frac{\eta_{\mathbf{x}}L^2}{2}+\frac{768G_g^2L_f^2}{\alpha_{t+1}^2\eta_{\mathbf{x}}\beta}+\frac{576G_g^2L_f^2\beta}{\alpha_{t+1}^2\eta_{\mathbf{x}}}\right)\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2\\ &-\left(\frac{1}{8\eta_{\mathbf{y}}}+\frac{768G_g^2L_f^2}{\alpha_{t+1}^2\eta_{\mathbf{x}}\beta}\right)\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2+\frac{7}{2\eta_{\mathbf{x}}}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2+\frac{\alpha_t}{2}\left\|\mathbf{x}^{t+1}\right\|^2.\end{aligned}$$

We conclude that:

$$\begin{aligned}
\mathbb{E}[\hat{F}^{t+1}-\hat{F}^{t}] \geq &\left(\frac{1}{4\eta_{\mathbf{y}}}-\frac{L}{2}-\eta_{\mathbf{x}}L^2-4L_f^2G_g^4\eta_{\mathbf{y}}-\frac{768G_g^2L_f^2}{\alpha_{t+1}^2\eta_{\mathbf{x}}\beta}-\frac{576G_g^2L_f^2\beta}{\alpha_{t+1}^2\eta_{\mathbf{x}}}\right)\mathbb{E}\left\|\mathbf{y}^{t+1}-\mathbf{y}^{t}\right\|^2\\
&+\left(\frac{1}{8\eta_{\mathbf{y}}}-\frac{768G_g^2L_f^2}{\alpha_{t+1}^2\eta_{\mathbf{x}}\beta}\right)\mathbb{E}\left\|\mathbf{y}^{t}-\mathbf{y}^{t-1}\right\|^2+\frac{1}{8\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^{t-1}-\mathbf{y}^{t-2}\right\|^2+\left(\frac{1}{\eta_{\mathbf{x}}}+\frac{\alpha_{t-1}}{2}-\frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}\right)\mathbb{E}\left\|\mathbf{x}^{t+1}-\mathbf{x}^{t}\right\|^2\\
&-(1-\beta)^2\left(L_f^2G_g^2\eta_{\mathbf{y}}+2\eta_{\mathbf{x}}L_f^2\right)\mathbb{E}\left\|\mathbf{z}^{t}-g(\mathbf{y}^{t-1})\right\|^2+\frac{1}{2}\left(\alpha_t-\alpha_{t-1}\right)\left\|\mathbf{x}^{t+1}\right\|^2\\
&-\left(1-\frac{\beta}{2}\right)^2\frac{64L_f^2}{\alpha_t^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^{t}-\mathbf{z}^{t-1}\right\|^2+\frac{8}{\eta_{\mathbf{x}}}\left(\frac{\alpha_t}{\alpha_{t+1}}-\frac{\alpha_{t-1}}{\alpha_t}\right)\left\|\mathbf{x}^{t+1}\right\|^2\\
&-2\eta_{\mathbf{x}}(1-\beta)^2\frac{\sigma^2}{B}-\left(2\eta_{\mathbf{y}}(G_g^2+G_f^2)\frac{\sigma^2}{B}+2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M}+\frac{1152L_f^2\beta}{\alpha_t^2\eta_{\mathbf{x}}}\frac{\sigma^2}{M}\right).
\end{aligned}$$

□

Lemma 17. *Let C_1, C_2 be defined in (21) and (22). If the following conditions hold:*

$$\frac{1}{8\eta_{\mathbf{y}}}-\frac{768G_g^2L_f^2}{\alpha_{t+1}^2\eta_{\mathbf{x}}\beta}-16(1-\beta)^2G_g^2(C_1^t\eta_{\mathbf{y}}L_f^2G_g^2+C_2^t\eta_{\mathbf{x}}^2L_f^2)-\frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{64L_f^2}{\alpha_t^2\eta_{\mathbf{x}}}\left(\frac{12}{\beta}+6\beta\right)G_g^2\geq 0, \tag{25}$$

$$1-(1-\beta)^2\left(L_f^2G_g^2\eta_{\mathbf{y}}+2\eta_{\mathbf{x}}L_f^2+C_1^t4\eta_{\mathbf{y}}^2L_f^2G_g^2+4C_2^t\eta_{\mathbf{x}}^2L_f^2+1\right)\geq 0, \tag{26}$$

$$\frac{1}{8\eta_{\mathbf{y}}}-\frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{64L_f^2}{\alpha_T^2\eta_{\mathbf{x}}}\frac{6}{\beta}G_g^2\geq 0. \tag{27}$$

then for Algorithm 2, under assumptions of Theorem 4, then the following statement holds:

$$\begin{aligned}
\mathbb{E}[\tilde{F}^{t+1}-\tilde{F}^{t}]\geq &\frac{C_1^t}{2}\eta_{\mathbf{y}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2+\frac{C_2^t}{2}\eta_{\mathbf{x}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2-2C_2^t\eta_{\mathbf{x}}^2\alpha_t^2\left\|\mathbf{x}^t\right\|^2\\
&+\frac{1}{2}\left(\alpha_t-\alpha_{t-1}\right)\left\|\mathbf{x}^{t+1}\right\|^2+\frac{8}{\eta_{\mathbf{x}}}\left(\frac{\alpha_t}{\alpha_{t+1}}-\frac{\alpha_{t-1}}{\alpha_t}\right)\left\|\mathbf{x}^{t+1}\right\|^2\\
&-\left(2\eta_{\mathbf{x}}+4C_2^t\eta_{\mathbf{x}}^2+4C_1^t\eta_{\mathbf{y}}^2(G_f^2+G_g^2)+2\eta_{\mathbf{y}}(G_g^2+G_f^2)\right)\frac{\sigma^2}{B}\\
&-\left(2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}+4C_1^t\beta^2\eta_{\mathbf{y}}^2L_f^2G_g^2+8C_2^t\beta^2\eta_{\mathbf{x}}^2L_f^2+\frac{1152L_f^2\beta}{\alpha_t^2\eta_{\mathbf{x}}}\right)\frac{\sigma^2}{M},
\end{aligned}$$

where

$$\tilde{F}^{t+1}:=\hat{F}^{t+1}-\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{y}^t)\right\|^2-\frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{64L_f^2}{\alpha_T^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^{t+1}-\mathbf{z}^t\right\|^2.$$

Proof. According to Lemma 16:

$$\begin{aligned}
\mathbb{E}[\hat{F}^{t+1}-\hat{F}^{t}]\geq &C_1^t\mathbb{E}\left\|\mathbf{y}^{t+1}-\mathbf{y}^{t}\right\|^2+\left(\frac{1}{8\eta_{\mathbf{y}}}-\frac{768G_g^2L_f^2}{\alpha_{t+1}^2\eta_{\mathbf{x}}\beta}\right)\mathbb{E}\left\|\mathbf{y}^{t}-\mathbf{y}^{t-1}\right\|^2+\frac{1}{8\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^{t-1}-\mathbf{y}^{t-2}\right\|^2\\
&+C_2^t\mathbb{E}\left\|\mathbf{x}^{t+1}-\mathbf{x}^{t}\right\|^2-(1-\beta)^2\left(L_f^2G_g^2\eta_{\mathbf{y}}+2\eta_{\mathbf{x}}L_f^2\right)\mathbb{E}\left\|\mathbf{z}^{t}-g(\mathbf{y}^{t-1})\right\|^2+\frac{1}{2}\left(\alpha_t-\alpha_{t-1}\right)\left\|\mathbf{x}^{t+1}\right\|^2\\
&-\left(1-\frac{\beta}{2}\right)^2\frac{64L_f^2}{\alpha_t^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^{t}-\mathbf{z}^{t-1}\right\|^2+\frac{8}{\eta_{\mathbf{x}}}\left(\frac{\alpha_t}{\alpha_{t+1}}-\frac{\alpha_{t-1}}{\alpha_t}\right)\left\|\mathbf{x}^{t+1}\right\|^2\\
&-2\eta_{\mathbf{x}}(1-\beta)^2\frac{\sigma^2}{B}-\left(2\eta_{\mathbf{y}}(G_g^2+G_f^2)\frac{\sigma^2}{B}+2\beta^2L_f^2G_g^2\eta_{\mathbf{y}}\frac{\sigma^2}{M}+\frac{1152L_f^2\beta}{\alpha_t^2\eta_{\mathbf{x}}}\frac{\sigma^2}{M}\right),
\end{aligned}$$

where

$$C_1^t = \frac{1}{4\eta_{\mathbf{y}}} - \frac{L}{2} - \eta_{\mathbf{x}} L^2 - 4L_f^2 G_g^4 \eta_{\mathbf{y}} - \frac{768 G_g^2 L_f^2}{\alpha_{t+1}^2 \eta_{\mathbf{x}} \beta} - \frac{576 G_g^2 L_f^2 \beta}{\alpha_{t+1}^2 \eta_{\mathbf{x}}}, \tag{28}$$

$$C_2^t = \left(\frac{1}{\eta_{\mathbf{x}}} + \frac{\alpha_{t-1}}{2} - \frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2} \right). \tag{29}$$

Now we plug in Lemma 14:

$$\begin{aligned}
\mathbb{E}[\hat{F}^{t+1} - \hat{F}^t] \geq & C_1^t \left(\frac{1}{2}\eta_{\mathbf{y}}^2 \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 - 4\eta_{\mathbf{y}}^2 \frac{(G_f^2 + G_g^2)\sigma^2}{B} - 2\eta_{\mathbf{y}}^2 G_g^2 L_f^2 \mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{y}^t))\right\|^2 \right) \\
& + C_2^t \left(\frac{1}{2}\eta_{\mathbf{x}}^2 \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 - 4\eta_{\mathbf{x}}^2 L_f^2 \mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{y}^t)\right\|^2 - 4\eta_{\mathbf{x}}^2 \frac{\sigma^2}{B} \right) \\
& - (1-\beta)^2 \left(L_f^2 G_g^2 \eta_{\mathbf{y}} + 2\eta_{\mathbf{x}} L_f^2 \right) \mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 \\
& + \frac{1}{2}\left(\alpha_t - \alpha_{t-1}\right) \mathbb{E}\left\|\mathbf{x}^{t+1}\right\|^2 + \frac{8}{\eta_{\mathbf{x}}} \left(\frac{\alpha_t}{\alpha_{t+1}} - \frac{\alpha_{t-1}}{\alpha_t} \right) \mathbb{E}\left\|\mathbf{x}^{t+1}\right\|^2 \\
& - \left(1 - \frac{\beta}{2}\right)^2 \frac{64 L_f^2}{\alpha_t^2 \eta_{\mathbf{x}}} \mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2 + \left(\frac{1}{8\eta_{\mathbf{y}}} - \frac{768 G_g^2 L_f^2}{\alpha_{t+1}^2 \eta_{\mathbf{x}} \beta} \right) \mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 + \frac{1}{8\eta_{\mathbf{y}}} \mathbb{E}\left\|\mathbf{y}^{t-1} - \mathbf{y}^{t-2}\right\|^2 \\
& - 2\eta_{\mathbf{x}}(1-\beta)^2 \frac{\sigma^2}{B} - \left(2\eta_{\mathbf{y}}(G_g^2 + G_f^2)\frac{\sigma^2}{B} + 2\beta^2 L_f^2 G_g^2 \eta_{\mathbf{y}} \frac{\sigma^2}{M} + \frac{1152 L_f^2 \beta}{\alpha_t^2 \eta_{\mathbf{x}}} \frac{\sigma^2}{M} \right).
\end{aligned}$$

Plugging in Lemma 1 yields:

$$\begin{aligned}
\mathbb{E}[\hat{F}^{t+1} - \hat{F}^t] \geq & \frac{C_1^t}{2} \eta_{\mathbf{y}}^2 \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 + \frac{C_2^t}{2} \eta_{\mathbf{x}}^2 \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 \\
& - 2C_1^t \eta_{\mathbf{y}}^2 L_f^2 G_g^2 \left((1-\beta)^2 \mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 + 4(1-\beta)^2 G_g^2 \mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 + 2\beta^2 \frac{\sigma^2}{M} \right) \\
& - 4C_2^t \eta_{\mathbf{x}}^2 L_f^2 \left((1-\beta)^2 \mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 + 4(1-\beta)^2 G_g^2 \mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 + 2\beta^2 \frac{\sigma^2}{M} \right) \\
& - (1-\beta)^2 \left(L_f^2 G_g^2 \eta_{\mathbf{y}} + 2\eta_{\mathbf{x}} L_f^2 \right) \mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{y}^{t-1})\right\|^2 \\
& + \frac{1}{2}\left(\alpha_t - \alpha_{t-1}\right) \left\|\mathbf{x}^{t+1}\right\|^2 + \frac{8}{\eta_{\mathbf{x}}} \left(\frac{\alpha_t}{\alpha_{t+1}} - \frac{\alpha_{t-1}}{\alpha_t} \right) \left\|\mathbf{x}^{t+1}\right\|^2 \\
& - \left(1 - \frac{\beta}{2}\right)^2 \frac{64 L_f^2}{\alpha_t^2 \eta_{\mathbf{x}}} \mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2 + \left(\frac{1}{8\eta_{\mathbf{y}}} - \frac{768 G_g^2 L_f^2}{\alpha_{t+1}^2 \eta_{\mathbf{x}} \beta} \right) \mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2 + \frac{1}{8\eta_{\mathbf{y}}} \mathbb{E}\left\|\mathbf{y}^{t-1} - \mathbf{y}^{t-2}\right\|^2 \\
& - 2\eta_{\mathbf{x}}(1-\beta)^2 \frac{\sigma^2}{B} - \left(2\eta_{\mathbf{y}}(G_g^2 + G_f^2)\frac{\sigma^2}{B} + 2\beta^2 L_f^2 G_g^2 \eta_{\mathbf{y}} \frac{\sigma^2}{M} + \frac{1152 L_f^2 \beta}{\alpha_t^2 \eta_{\mathbf{x}}} \frac{\sigma^2}{M} \right) - 4C_1 \eta_{\mathbf{y}}^2 \frac{(G_f^2 + G_g^2)\sigma^2}{B} - 4C_2 \eta_{\mathbf{x}}^2 \frac{\sigma^2}{B}.
\end{aligned}$$

Rearranging the terms yields:

$$\begin{aligned}
\mathbb{E}[\hat{F}^{t+1}-\hat{F}^{t}] \geq & \frac{C_1^t}{2}\eta_{\mathbf{y}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 + C_2^t\frac{1}{2}\eta_{\mathbf{x}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 + \frac{1}{8\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^{t-1}-\mathbf{y}^{t-2}\right\|^2 \\
& + \left(\frac{1}{8\eta_{\mathbf{y}}} - \frac{768G_g^2L_f^2}{\alpha_{t+1}^2\eta_{\mathbf{x}}\beta} - 16(1-\beta)^2G_g^2(C_1^t\eta_{\mathbf{y}}L_f^2G_g^2 + C_2^t\eta_{\mathbf{x}}^2L_f^2)\right)\mathbb{E}\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2 \\
& - (1-\beta)^2\left(L_f^2G_g^2\eta_{\mathbf{y}} + 2\eta_{\mathbf{x}}L_f^2 + C_1^t4\eta_{\mathbf{y}}^2L_f^2G_g^2 + 4C_2^t\eta_{\mathbf{x}}^2L_f^2\right)\mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{y}^{t-1})\right\|^2 \\
& + \frac{1}{2}(\alpha_t-\alpha_{t-1})\left\|\mathbf{x}^{t+1}\right\|^2 + \frac{8}{\eta_{\mathbf{x}}}\left(\frac{\alpha_t}{\alpha_{t+1}} - \frac{\alpha_{t-1}}{\alpha_t}\right)\left\|\mathbf{x}^{t+1}\right\|^2 - \left(1-\frac{\beta}{2}\right)^2\frac{64L_f^2}{\alpha_t^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^t-\mathbf{z}^{t-1}\right\|^2 \\
& - \left(2\eta_{\mathbf{x}} + 4C_2^t\eta_{\mathbf{x}}^2 + 4C_1^t\eta_{\mathbf{y}}^2(G_f^2+G_g^2) + 2\eta_{\mathbf{y}}(G_g^2+G_f^2)\right)\frac{\sigma^2}{B} \\
& - \left(2\beta^2L_f^2G_g^2\eta_{\mathbf{y}} + 4C_1^t\beta^2\eta_{\mathbf{y}}^2L_f^2G_g^2 + 8C_2^t\beta^2\eta_{\mathbf{x}}^2L_f^2 + \frac{1152L_f^2\beta}{\alpha_t^2\eta_{\mathbf{x}}}\right)\frac{\sigma^2}{M}.
\end{aligned}$$

Recall our definition of potential function $\tilde{F}^{t+1}$: $\tilde{F}^{t+1} := \hat{F}^{t+1} - \mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{y}^t))\right\|^2 - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{64L_f^2}{\alpha_T^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^{t+1}-\mathbf{z}^t\right\|^2$. Hence we have:

$$\begin{aligned}
\mathbb{E}[\hat{F}^{t+1}-\hat{F}^{t}] \geq & \frac{C_1^t}{2}\eta_{\mathbf{y}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 + C_2^t\frac{1}{2}\eta_{\mathbf{x}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 + \frac{1}{8\eta_{\mathbf{y}}}\mathbb{E}\left\|\mathbf{y}^{t-1}-\mathbf{y}^{t-2}\right\|^2 \\
& + \left(\frac{1}{8\eta_{\mathbf{y}}} - \frac{768G_g^2L_f^2}{\alpha_{t+1}^2\eta_{\mathbf{x}}\beta} - 16(1-\beta)^2G_g^2(C_1^t\eta_{\mathbf{y}}L_f^2G_g^2 + C_2^t\eta_{\mathbf{x}}^2L_f^2)\right)\mathbb{E}\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2 \\
& - (1-\beta)^2\left(L_f^2G_g^2\eta_{\mathbf{y}} + 2\eta_{\mathbf{x}}L_f^2 + C_1^t4\eta_{\mathbf{y}}^2L_f^2G_g^2 + 4C_2^t\eta_{\mathbf{x}}^2L_f^2\right)\mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{y}^{t-1})\right\|^2 \\
& + \frac{1}{2}(\alpha_t-\alpha_{t-1})\left\|\mathbf{x}^{t+1}\right\|^2 + \frac{8}{\eta_{\mathbf{x}}}\left(\frac{\alpha_t}{\alpha_{t+1}} - \frac{\alpha_{t-1}}{\alpha_t}\right)\left\|\mathbf{x}^{t+1}\right\|^2 - \left(1-\frac{\beta}{2}\right)^2\frac{64L_f^2}{\alpha_T^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^t-\mathbf{z}^{t-1}\right\|^2 \\
& - \mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{y}^t))\right\|^2 + \mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{y}^{t-1})\right\|^2 \\
& - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{64L_f^2}{\alpha_T^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^{t+1}-\mathbf{z}^t\right\|^2 + \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{64L_f^2}{\alpha_T^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^t-\mathbf{z}^{t-1}\right\|^2 \\
& - \left(2\eta_{\mathbf{x}} + 4C_2^t\eta_{\mathbf{x}}^2 + 4C_1^t\eta_{\mathbf{y}}^2(G_f^2+G_g^2) + 2\eta_{\mathbf{y}}(G_g^2+G_f^2)\right)\frac{\sigma^2}{B} - \left(2\beta^2L_f^2G_g^2\eta_{\mathbf{y}} + 4C_1^t\beta^2\eta_{\mathbf{y}}^2L_f^2G_g^2 + 8C_2^t\beta^2\eta_{\mathbf{x}}^2L_f^2 + \frac{1152L_f^2\beta}{\alpha_t^2\eta_{\mathbf{x}}}\right)\frac{\sigma^2}{M}.
\end{aligned}$$

Now we plug in Lemma 1 and 7:

$$\begin{aligned}
\mathbb{E}[\tilde{F}^{t+1}-\tilde{F}^{t}] \geq & \frac{C_1^t}{2}\eta_{\mathbf{y}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 + \frac{C_2^t}{2}\eta_{\mathbf{x}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 \\
& + \left(\frac{1}{8\eta_{\mathbf{y}}} - \frac{768G_g^2L_f^2}{\alpha_{t+1}^2\eta_{\mathbf{x}}\beta} - 16(1-\beta)^2G_g^2(C_1^t\eta_{\mathbf{y}}L_f^2G_g^2 + C_2^t\eta_{\mathbf{x}}^2L_f^2) - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{64L_f^2}{\alpha_t^2\eta_{\mathbf{x}}}\left(\frac{12}{\beta}+6\beta\right)G_g^2\right)\mathbb{E}\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2 \\
& + \left(1-(1-\beta)^2\left(L_f^2G_g^2\eta_{\mathbf{y}} + 2\eta_{\mathbf{x}}L_f^2 + C_1^t4\eta_{\mathbf{y}}^2L_f^2G_g^2 + 4C_2^t\eta_{\mathbf{x}}^2L_f^2 + 1\right)\right)\mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{y}^{t-1})\right\|^2 \\
& + \left(\frac{1}{8\eta_{\mathbf{y}}} - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{64L_f^2}{\alpha_T^2\eta_{\mathbf{x}}}\frac{6}{\beta}G_g^2\right)\mathbb{E}\left\|\mathbf{y}^{t-1}-\mathbf{y}^{t-2}\right\|^2 \\
& + (1-\frac{\beta}{2})^2\frac{64L_f^2}{\alpha_T^2\eta_{\mathbf{x}}}\underbrace{\left(\frac{1}{1-(1-\frac{\beta}{2})^2} - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2} - 1\right)}_{=0}\mathbb{E}\left\|\mathbf{z}^t-\mathbf{z}^{t-1}\right\|^2 \\
& + \frac{1}{2}(\alpha_t-\alpha_{t-1})\left\|\mathbf{x}^{t+1}\right\|^2 + \frac{8}{\eta_{\mathbf{x}}}\left(\frac{\alpha_t}{\alpha_{t+1}} - \frac{\alpha_{t-1}}{\alpha_t}\right)\left\|\mathbf{x}^{t+1}\right\|^2 \\
& - \left(2\eta_{\mathbf{x}} + 4C_2^t\eta_{\mathbf{x}}^2 + 4C_1^t\eta_{\mathbf{y}}^2(G_f^2+G_g^2) + 2\eta_{\mathbf{y}}(G_g^2+G_f^2)\right)\frac{\sigma^2}{B} - \left(2\beta^2L_f^2G_g^2\eta_{\mathbf{y}} + 4C_1^t\beta^2\eta_{\mathbf{y}}^2L_f^2G_g^2 + 8C_2^t\beta^2\eta_{\mathbf{x}}^2L_f^2 + \frac{1152L_f^2\beta}{\alpha_t^2\eta_{\mathbf{x}}}\right)\frac{\sigma^2}{M}.
\end{aligned}$$

By our choice of $\eta_{\mathbf{y}}$, we know that

$$\frac{1}{8\eta_{\mathbf{y}}} - \frac{768G_g^2L_f^2}{\alpha_{t+1}^2\eta_{\mathbf{x}}\beta} - 16(1-\beta)^2G_g^2(C_1^t\eta_{\mathbf{y}}L_f^2G_g^2 + C_2^t\eta_{\mathbf{x}}^2L_f^2) - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{64L_f^2}{\alpha_t^2\eta_{\mathbf{x}}}\left(\frac{12}{\beta}+6\beta\right)G_g^2 \geq 0,$$
$$1-(1-\beta)^2\left(L_f^2G_g^2\eta_{\mathbf{y}} + 2\eta_{\mathbf{x}}L_f^2 + C_1^t4\eta_{\mathbf{y}}^2L_f^2G_g^2 + 4C_2^t\eta_{\mathbf{x}}^2L_f^2 + 1\right) \geq 0,$$
$$\frac{1}{8\eta_{\mathbf{y}}} - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{64L_f^2}{\alpha_T^2\eta_{\mathbf{x}}}\frac{6}{\beta}G_g^2 \geq 0$$

Now we can have the clean bound:

$$\begin{aligned}\mathbb{E}[\tilde{F}^{t+1} - \tilde{F}^t] \geq{}& \frac{C_1^t}{2}\eta_{\mathbf{y}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 + \frac{C_2^t}{2}\eta_{\mathbf{x}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2\\ &+\frac{1}{2}\left(\alpha_t-\alpha_{t-1}\right)\left\|\mathbf{x}^{t+1}\right\|^2 + \frac{8}{\eta_{\mathbf{x}}}\left(\frac{\alpha_t}{\alpha_{t+1}} - \frac{\alpha_{t-1}}{\alpha_t}\right)\left\|\mathbf{x}^{t+1}\right\|^2\\ &-\left(2\eta_{\mathbf{x}} + 4C_2^t\eta_{\mathbf{x}}^2 + 4C_1^t\eta_{\mathbf{y}}^2(G_f^2+G_g^2) + 2\eta_{\mathbf{y}}(G_g^2+G_f^2)\right)\frac{\sigma^2}{B}\\ &-\left(2\beta^2L_f^2G_g^2\eta_{\mathbf{y}} + 4C_1^t\beta^2\eta_{\mathbf{y}}^2L_f^2G_g^2 + 8C_2^t\beta^2\eta_{\mathbf{x}}^2L_f^2 + \frac{1152L_f^2\beta}{\alpha_t^2\eta_{\mathbf{x}}}\right)\frac{\sigma^2}{M}.\end{aligned}$$

□

D.3.1 Proof Theorem 4

Evoking Lemma 17 and re-arranging terms yields:

$$\begin{aligned}&\frac{C_1^t}{2}\eta_{\mathbf{y}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 + \frac{C_2^t}{2}\eta_{\mathbf{x}}^2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 \leq \mathbb{E}[\tilde{F}^{t+1}-\tilde{F}^t] - \frac{1}{2}\left(\alpha_t-\alpha_{t-1}\right)\left\|\mathbf{x}^{t+1}\right\|^2\\ &\quad-\frac{8}{\eta_{\mathbf{x}}}\left(\frac{\alpha_t}{\alpha_{t+1}} - \frac{\alpha_{t-1}}{\alpha_t}\right)\left\|\mathbf{x}^{t+1}\right\|^2\\ &\quad+\left(2\eta_{\mathbf{x}} + 4C_2^t\eta_{\mathbf{x}}^2 + 4C_1^t\eta_{\mathbf{y}}^2(G_f^2+G_g^2) + 2\eta_{\mathbf{y}}(G_g^2+G_f^2)\right)\frac{\sigma^2}{B}\\ &\quad+\left(2\beta^2L_f^2G_g^2\eta_{\mathbf{y}} + 4C_1^t\beta^2\eta_{\mathbf{y}}^2L_f^2G_g^2 + 8C_2^t\beta^2\eta_{\mathbf{x}}^2L_f^2 + \frac{1152L_f^2\beta}{\alpha_t^2\eta_{\mathbf{x}}}\right)\frac{\sigma^2}{M}.\end{aligned}$$

We compute the upper and lower bound of C_1^t and C_2^t. For C_1^t

$$C_1^t = \frac{1}{4\eta_{\mathbf{y}}} - \frac{L}{2} - \eta_{\mathbf{x}}L^2 - 4L_f^2G_g^4\eta_{\mathbf{y}} - \frac{768G_g^2L_f^2}{\alpha_{t+1}^2\eta_{\mathbf{x}}\beta} - \frac{576G_g^2L_f^2\beta}{\alpha_{t+1}^2\eta_{\mathbf{x}}}.$$

The upper bound $C_1^t \leq \frac{1}{4\eta_{\mathbf{y}}}$ holds trivially. For lower bound, since we choose

$$\eta_{\mathbf{y}} \leq c\cdot\min\left\{\frac{1}{L}, \frac{1}{\eta_{\mathbf{x}}L^2}, \frac{1}{L_fG_g^2}, \frac{\alpha_t^2\eta_{\mathbf{x}}\beta}{G_g^2L_f^2}\right\},$$

for some $c \geq 0$, we can choose sufficiently small c such that $C_1^t \geq \frac{1}{8\eta_{\mathbf{y}}}$.

For C_2^t:

$$C_2^t = \frac{1}{\eta_{\mathbf{x}}} + \frac{\alpha_{t-1}}{2} - \frac{1}{4\eta_{\mathbf{x}}(1-\beta)^2}.$$

The upper bound $C_2^t \leq \frac{1}{\eta_{\mathbf{x}}} + \frac{\alpha_t}{2}$ holds trivially. For lower bound, since we choose:

$$\beta = 0.1 \leq 1 - \frac{\sqrt{2}}{2}$$

it holds that $C_2^t \geq \frac{1}{2\eta_{\mathbf{x}}}$.

Since $\frac{1}{8\eta_{\mathbf{y}}} \leq C_1^t \leq \frac{1}{4\eta_{\mathbf{y}}}$ and $\frac{1}{2\eta_{\mathbf{x}}} \leq C_2^t \leq \frac{1}{\eta_{\mathbf{x}}} + \frac{\alpha_t}{2} \leq \frac{2}{\eta_{\mathbf{x}}}$, we have:

$$\begin{aligned}
\frac{\eta_{\mathbf{y}}}{16}\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 + \frac{\eta_{\mathbf{x}}}{4}\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 &\leq \mathbb{E}[\tilde{F}^{t+1} - \tilde{F}^t] - \frac{1}{2}\left(\alpha_t - \alpha_{t-1}\right)\left\|\mathbf{x}^{t+1}\right\|^2 \\
&\quad - \frac{8}{\eta_{\mathbf{x}}}\left(\frac{\alpha_t}{\alpha_{t+1}} - \frac{\alpha_{t-1}}{\alpha_t}\right)\left\|\mathbf{x}^{t+1}\right\|^2 \\
&\quad + \left(2\eta_{\mathbf{x}} + 8\eta_{\mathbf{x}} + \eta_{\mathbf{y}}(G_f^2 + G_g^2) + 2\eta_{\mathbf{y}}(G_g^2 + G_f^2)\right)\frac{\sigma^2}{B} + \left(2\beta^2 L_f^2 G_g^2 \eta_{\mathbf{y}} + \eta_{\mathbf{y}} L_f^2 G_g^2 + 16\eta_{\mathbf{x}} L_f^2 + \frac{1152 L_f^2 \beta}{\alpha_t^2 \eta_{\mathbf{x}}}\right)\frac{\sigma^2}{M}.
\end{aligned}$$

Summing the above inequality from $t = 0$ to $T - 1$ yields:

$$\begin{aligned}
&\frac{\mathbb{E}[\tilde{F}^T - \tilde{F}^0]}{T} + \frac{\left(\alpha_{T-1} - \alpha_0\right) D_{\mathcal{X}}^2}{2T} - \frac{8\left(\frac{\alpha_{T-1}}{\alpha_T} - \frac{\alpha_0}{\alpha_1}\right) D_{\mathcal{X}}^2}{T} + 4\eta_{\mathbf{x}}\frac{1}{T}\sum_{t=0}^{T-1}\alpha_t^2 D_{\mathcal{X}}^2 \\
&\qquad + O\left(\left(\eta_{\mathbf{x}} + \eta_{\mathbf{y}}(G_g^2 + G_f^2)\right)\frac{\sigma^2}{B} + \left(L_f^2 G_g^2 \eta_{\mathbf{y}} + \eta_{\mathbf{x}} L_f^2 + \frac{L_f^2 \beta}{\alpha_t^2 \eta_{\mathbf{x}}}\right)\frac{\sigma^2}{M}\right) \\
&\qquad \geq \min\left\{\frac{1}{16}\eta_{\mathbf{y}}, \frac{1}{4}\eta_{\mathbf{x}}\right\}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 + \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t,\mathbf{y}^t)\right\|^2\right).
\end{aligned}$$

We then need to verify the conditions 25, 26 and 27 in Lemma 17 can hold under our choice of $\eta_{\mathbf{x}}$, $\eta_{\mathbf{y}}$ and β. To guarantee (25) holding, we need:

$$\eta_{\mathbf{y}} \leq \Theta\left(\min\left\{\frac{1}{\eta_x L_f^2 G_g^2}, \frac{1}{L_f G_g^2}, \frac{\alpha_t^2 \eta_{\mathbf{x}}}{L_f^2 G_g^2}\right\}\right), \forall\ 0 \leq t \leq T-1.$$

To guarantee condition (26) holding, we need

$$\eta_{\mathbf{x}} \leq \Theta\left(\frac{1}{L_f^2}\right), \eta_{\mathbf{y}} \leq \Theta\left(\frac{1}{L_f^2 G_g^2}\right).$$

To guarantee condition (27) holding, we need:

$$\eta_{\mathbf{y}} \leq \Theta\left(\frac{\alpha_T^2 \eta_{\mathbf{x}}}{L_f^2 G_g^2}\right).$$

Next we examine how large the $\mathbb{E}[\tilde{F}^T - \tilde{F}^0]$ is. By definition of potential function, we have:

$$\begin{aligned}
\mathbb{E}[\tilde{F}^T - \tilde{F}^0] &= \hat{F}^T - \mathbb{E}\left\|\mathbf{z}^T - g(\mathbf{y}^{T-1})\right\|^2 - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{64L_f^2}{\alpha_t^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2 \\
&\quad - \left(\hat{F}^0 - \mathbb{E}\left\|\mathbf{z}^0 - g(\mathbf{y}^{-1})\right\|^2 - \frac{(1-\frac{\beta}{2})^2}{1-(1-\frac{\beta}{2})^2}\frac{64L_f^2}{\alpha_t^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}^0 - \mathbf{z}^{-1}\right\|^2\right) \\
&\leq \hat{F}^T - \hat{F}^0 + \mathbb{E}\left\|\mathbf{z}^0 - g(\mathbf{y}^{-1})\right\|^2 + \frac{(1-\frac{\beta}{2})^2}{\beta - \frac{\beta^2}{4}}\frac{64L_f^2}{\alpha_t^2\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{z}_j^0 - \mathbf{z}_j^{-1}\right\|^2
\end{aligned}$$

By convention $\mathbf{y}^0 = \mathbf{y}^{-1}$, $\mathbf{z}^0 = \mathbf{z}^{-1}$, and our choice $\mathbb{E}\left\|\mathbf{z}^0 - g(\mathbf{y}^0)\right\|^2 \leq O(1)$, we have

$$\mathbb{E}[\tilde{F}^T - \tilde{F}^0] \leq \hat{F}^T - \hat{F}^0 + O(1).$$

Next we examine how large the $\mathbb{E}[\hat{F}^T - \hat{F}^0]$ is.

$$
\begin{aligned}
\mathbb{E}[\hat{F}^T - \hat{F}^0] = {} & F(\mathbf{x}^T, \mathbf{y}^T) + s^T - \left(\frac{1}{4\eta_{\mathbf{y}}} + 4L_f^2 G_g^4 \eta_{\mathbf{y}} + \frac{\eta_{\mathbf{x}} L^2}{2} + \frac{768 G_g^2 L_f^2}{\alpha_T^2 \eta_{\mathbf{x}} \beta} + \frac{576 G_g^2 L_f^2 \beta}{\alpha_T^2 \eta_{\mathbf{x}}}\right) \left\|\mathbf{y}^T - \mathbf{y}^{T-1}\right\|^2 \\
& - \left(\frac{1}{8\eta_{\mathbf{y}}} + \frac{768 G_g^2 L_f^2}{a_T^2 \eta_{\mathbf{x}} \beta}\right) \left\|\mathbf{y}^{T-1} - \mathbf{y}^{T-2}\right\|^2 + \frac{7}{2\eta_{\mathbf{x}}} \left\|\mathbf{x}^T - \mathbf{x}^{T-1}\right\|^2 + \frac{\alpha_{T-1}}{2} \left\|\mathbf{x}^T\right\|^2 \\
& - F(\mathbf{x}^0, \mathbf{y}^0) - s^0 + \left(\frac{1}{4\eta_{\mathbf{y}}} + 4L_f^2 G_g^4 \eta_{\mathbf{y}} + \frac{\eta_{\mathbf{x}} L^2}{2} + \frac{768 G_g^2 L_f^2}{a_0^2 \eta_{\mathbf{x}} \beta} + \frac{576 G_g^2 L_f^2 \beta}{a_0^2 \eta_{\mathbf{x}}}\right) \left\|\mathbf{y}^0 - \mathbf{y}^{-1}\right\|^2 \\
& + \left(\frac{1}{8\eta_{\mathbf{y}}} + \frac{768 G_g^2}{a_0^2 \eta_{\mathbf{x}} \beta}\right) \left\|\mathbf{y}^{-1} - \mathbf{y}^{-2}\right\|^2 - \frac{7}{2\eta_{\mathbf{x}}} \left\|\mathbf{x}^0 - \mathbf{x}^{-1}\right\|^2 - \frac{\alpha_0}{2} \left\|\mathbf{x}^0\right\|^2 \\
\leq {} & F_{\max} + \frac{7}{2\eta_{\mathbf{x}}} \left\|\mathbf{x}^T - \mathbf{x}^{T-1}\right\|^2,
\end{aligned}
$$

where we used the convention $\mathbf{x}^{-1} = \mathbf{x}^0$ and $\mathbf{y}^{-1} = \mathbf{y}^{-2}$. Notice that $\mathbb{E}\left\|\mathbf{x}^T - \mathbf{x}^{T-1}\right\|^2 \leq \eta_{\mathbf{x}}^2 \mathbb{E}\left\|\frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^{T-1}} \nabla_1 f(\mathbf{x}^{T-1}, \mathbf{z}^T; \zeta) + \nabla h(\mathbf{x}^{T-1})\right\|^2 \leq 2\eta_{\mathbf{x}}^2 (G_f^2 + G_h^2)$, we have $\mathbb{E}[\hat{F}^T - \hat{F}^0] \leq F_{\max} + 7\eta_{\mathbf{x}}(G_f^2 + G_h^2)$.

Due to our choice that $\eta_{\mathbf{y}} \leq 8\eta_{\mathbf{x}}$, $\alpha_0 = ... = \alpha_T = \alpha$ we have:

$$
\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 + \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2\right) \leq O\left(\frac{F_{\max} + \eta_{\mathbf{x}}(G_f^2 + G_h^2)}{\eta_{\mathbf{y}} T}\right) + O\left(\left(\frac{\eta_{\mathbf{x}}}{\eta_{\mathbf{y}}} G_f^2 + G_g^2\right)\frac{\sigma^2}{B} + \left(L_f^2 G_g^2 + \frac{\eta_{\mathbf{x}}}{\eta_{\mathbf{y}}} L_f^2 + \frac{L_f^2 \beta}{\alpha^2 \eta_{\mathbf{x}} \eta_{\mathbf{y}}}\right)\frac{\sigma^2}{M}\right).
$$

By the definition of stationary measure, we have

$$
\begin{aligned}
\mathbb{E}\left\|\nabla G(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 &= \frac{1}{\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathbf{x}^t - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}^t - \eta\nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}^t)\right)\right\|^2 + \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 \\
&= \frac{2}{\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathbf{x}^t - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}^t - \eta_{\mathbf{x}}\nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}^t) - \eta_{\mathbf{x}}\alpha\mathbf{x}^t\right)\right\|^2 \\
&\quad + \frac{2}{\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathcal{P}_{\mathcal{X}}\left(\mathbf{x}^t - \eta_{\mathbf{x}}\nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}^t)\right) - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}^t - \eta_{\mathbf{x}}\nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}^t) - \eta_{\mathbf{x}}\alpha\mathbf{x}^t\right)\right\|^2 + \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 \\
&= \frac{2}{\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathbf{x}^t - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}^t - \eta_{\mathbf{x}}\nabla_{\mathbf{x}} F(\mathbf{x}^t, \mathbf{y}^t) - \eta_{\mathbf{x}}\alpha\mathbf{x}^t\right)\right\|^2 + \frac{2}{\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\alpha\mathbf{x}^t\right\|^2 + \mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 \\
&\leq 2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{x}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 + 2\mathbb{E}\left\|\hat{\mathbf{g}}_{\mathbf{y}}(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 + 2\mathbb{E}\left\|\alpha\mathbf{x}^t\right\|^2.
\end{aligned}
$$

To guarantee $\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\left\|\nabla G(\mathbf{x}^t, \mathbf{y}^t)\right\|^2$ is less than ϵ^2, we need: $T = O\left(\frac{F_{\max}}{\eta_{\mathbf{y}}\epsilon^2}\right)$, $\alpha = \Theta\left(\frac{\epsilon}{\mathcal{D}_{\mathcal{X}}}\right)$ and

$$
\begin{aligned}
M &= \Theta\left(\max\left\{\frac{L^6 L_f^2 D_{\mathcal{X}}^4 \sigma^2}{\epsilon^4}, 1\right\}\right), B = \Theta\left(\max\left\{\frac{L^3 L_f^2 D_{\mathcal{X}}^2 \sigma^2}{\epsilon^4}, 1\right\}\right), \\
\eta_{\mathbf{y}} &= \Theta\left(\min\left\{\frac{1}{\eta_x L_f^2 G_g^2}, \frac{1}{L_f G_g^2}, \frac{\alpha^2 \eta_{\mathbf{x}}}{L_f^2 G_g^2}\right\}\right), \eta_{\mathbf{x}} = \Theta\left(\frac{1}{L^2}\right),
\end{aligned}
$$

which yields the total gradient complexity:

$$
O\left(\max\left\{\frac{L^3 L_f^2 D_{\mathcal{X}}^2 \sigma^2}{\epsilon^4}, 1\right\} \cdot \frac{F_{\max} D_{\mathcal{X}}^2 L^4}{\epsilon^4}\right).
$$

E Proof of Both Sides Composition

In this section we provide the proof of results in primal and dual composition setting (Theorem 5).

We first define some notations. Let $\mathbf{w} = [\mathbf{x}, \mathbf{y}]$ be stacked variable. $\mathbf{g}_{\mathbf{x}}(\mathbf{w}) := \nabla f(g(\mathbf{x}, \mathbf{y}))\nabla_{\mathbf{x}} g(\mathbf{x}, \mathbf{y}) + \nabla h(\mathbf{x}) + \frac{2}{\gamma}\left(\mathbf{x} - \mathbf{x}^0\right)$, $\mathbf{g}_{\mathbf{y}}(\mathbf{w}) := \nabla f(g(\mathbf{x}, \mathbf{y}))\nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y}) - \nabla r(\mathbf{y}) - \frac{2}{\gamma}\left(\mathbf{y} - \mathbf{y}^0\right)$. Hence we define stacked variable $\mathbf{g}(\mathbf{w}) := [\mathbf{g}_{\mathbf{x}}(\mathbf{w}), \mathbf{g}_{\mathbf{y}}(\mathbf{w})]$. Similarly, we define stacked variable $\mathbf{g}^t := [\mathbf{g}^t_{\mathbf{x}}, \mathbf{g}^t_{\mathbf{y}}]$ for the gradients we used in updating rule.

Proposition 2. *For Algorithm 3, under the assumptions of Theorem 5, the following statement holds true:*

$$\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 \leq (1-\beta)^2 \mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})\right\|^2 + 4(1-\beta)^2 G_g^2 \left(\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 + \left\|\mathbf{y}^t - \mathbf{y}^{t-1}\right\|^2\right) + 2\beta^2 \frac{\sigma^2}{B}.$$

Relying on reduction introduced in Liu et al. [2021], we can break solving a weakly-convex-weakly-concave minimax problem into solving a bunch of strongly monotone variational inequality problems. Hence we first derive the following convergence of CODA-SCSC on solving Strongly Monotone VI, which is given by the following Lemma.

Lemma 18. *Assume that $F_k(\mathbf{x}, \mathbf{y})$ is μ strongly convex in $\mathbf{x}$ and μ strongly concave in $\mathbf{y}$, and L smooth, and $\max\{D_{\mathcal{X}}, D_{\mathcal{Y}}\} \leq D$, then if we run Algorithm 3 on $F_k(\mathbf{x}, \mathbf{y})$ with $\eta_{\mathbf{x}} = \eta_{\mathbf{y}} = \eta$, the solution $\hat{\mathbf{w}} = (\hat{\mathbf{x}}, \hat{\mathbf{y}})$ returned by algorithm guarantees that $\max_{\mathbf{w}\in\mathcal{W}} \mathbb{E}[\nabla F_k(\hat{\mathbf{w}})^\top(\hat{\mathbf{w}} - \mathbf{w})] \leq \epsilon$ with gradient complexity $O\left(\left(D + \frac{1}{\eta}\right)\frac{L^2\sigma^2}{\mu^2\eta\epsilon}\ln\left(\frac{D+1/\eta}{\epsilon}\right)\right)$.*

To prove above Lemma, we need the following technical Lemmas.

Lemma 19 (Convergence of Iterates difference)**.** *For Algorithm 3, under the assumptions of Theorem 5, the following statement holds true:*

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2 \leq{}& \left(1 - \frac{\beta}{2}\right)^2 \mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2 + 4\left(1 + \frac{2}{\beta}\right) G_g^2 \left(\left\|\mathbf{w}^t - \mathbf{w}^{t-1}\right\|^2 + \left\|\mathbf{w}^{t-1} - \mathbf{w}^{t-2}\right\|^2\right) \\
&+ 2\left(1 + \frac{2}{\beta}\right)\beta^2 G_g^2 \mathbb{E}\left\|\mathbf{w}^t - \mathbf{w}^{t-1}\right\|^2.
\end{aligned}$$

Proof. For the ease of presentation, we define the following two auxiliary variables:

$$g^t = g(\mathbf{x}^t, \mathbf{y}^t; \mathcal{M}^t), g^{t\mapsto t-1} = g(\mathbf{x}^t, \mathbf{y}^t; \mathcal{M}^t) - g(\mathbf{x}^{t-1}, \mathbf{y}^{t-1}; \mathcal{M}^t).$$

According to updating rule of $\mathbf{z}$, we have:

$$\mathbf{z}^{t+1} - \mathbf{z}^t = (1-\beta)(\mathbf{z}^t - \mathbf{z}^{t-1}) + (1-\beta)(g^{t\mapsto t-1} - g^{t-1\mapsto t-2}) + \beta(g^t - g^{t-1}).$$

Taking expectation w.r.t. $\mathcal{M}^t$, $\mathcal{M}^{t-1}$ yields:

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|^2 &= \mathbb{E}\left\|(1-\beta)(\mathbf{z}^t - \mathbf{z}^{t-1}) + (1-\beta)(g^{t\mapsto t-1} - g^{t-1\mapsto t-2}) + \beta(g^t - g^{t-1})\right\|^2 \\
&\overset{(1)}{\leq} \left(1 + \frac{\beta}{2-2\beta}\right)(1-\beta)^2 \mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1} + (g^{t\mapsto t-1} - g^{t-1\mapsto t-2})\right\|^2 + \left(1 + \frac{2-2\beta}{\beta}\right)\left\|\beta(g^t - g^{t-1})\right\|^2 \\
&\overset{(2)}{\leq} \left(1 - \frac{\beta}{2}\right)(1-\beta)\mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1} + (g^{t\mapsto t-1} - g^{t-1\mapsto t-2})\right\|^2 + \left(1 + \frac{2}{\beta}\right)\beta^2 \mathbb{E}\left\|g^t - g^{t-1}\right\|^2 \\
&\overset{(3)}{\leq} \left(1 - \frac{\beta}{2}\right)(1-\beta)\left(1 + \frac{\beta}{2-2\beta}\right)\mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2 \\
&\quad + \left(1 - \frac{\beta}{2}\right)(1-\beta)\left(1 + \frac{2-2\beta}{\beta}\right)\mathbb{E}\left\|g^{t\mapsto t-1} - g^{t-1\mapsto t-2}\right\|^2 + \left(1 + \frac{2}{\beta}\right)\beta^2 \mathbb{E}\left\|g^t - g^{t-1}\right\|^2 \\
&\overset{(4)}{\leq} \left(1 - \frac{\beta}{2}\right)^2 \mathbb{E}\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2 + \underbrace{\left(1 + \frac{2}{\beta}\right)\mathbb{E}\left\|g^{t\mapsto t-1} - g^{t-1\mapsto t-2}\right\|^2}_{T_1} + \underbrace{\left(1 + \frac{2}{\beta}\right)\beta^2 \mathbb{E}\left\|g^t - g^{t-1}\right\|^2}_{T_2},
\end{aligned}$$

where in (1) and (3) we use Young's inequality that $\|\mathbf{a} + \mathbf{b}\|^2 \leq (1+a)\|\mathbf{a}\|^2 + (1 + \frac{1}{a})\|\mathbf{b}\|^2$. Now we bound T_1

as follows:

$$\begin{aligned}
T_1 \leq & 2\left(1+\frac{2}{\beta}\right)\left(\mathbb{E}\left\|g(\mathbf{x}^t,\mathbf{y}^t;\mathcal{M}^t)-g(\mathbf{x}^{t-1},\mathbf{y}^{t-1};\mathcal{M}^t)\right\|^2\right) \\
& + 2\left(1+\frac{2}{\beta}\right)\left(\mathbb{E}\left\|g(\mathbf{x}^{t-1},\mathbf{y}^{t-1};\mathcal{M}^{t-1})-g(\mathbf{x}^{t-2},\mathbf{y}^{t-2};\mathcal{M}^{t-1})\right\|^2\right) \\
\leq & 4\left(1+\frac{2}{\beta}\right)G_g^2\left(\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2+\mathbb{E}\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2+\mathbb{E}\left\|\mathbf{x}^{t-1}-\mathbf{x}^{t-2}\right\|^2+\mathbb{E}\left\|\mathbf{y}^{t-1}-\mathbf{y}^{t-2}\right\|^2\right),
\end{aligned}$$

and for T_2:

$$T_2 \leq \left(1+\frac{2}{\beta}\right)\beta^2 G_g^2\left(\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2+\mathbb{E}\left\|\mathbf{y}^t-\mathbf{y}^{t-1}\right\|^2\right).$$

Putting pieces together will conclude the proof. □

Lemma 20. *Let $\mathbf{w}^*$ be the solution of $MVI(F^\gamma,\mathcal{X}\times\mathcal{Y})$. For Algorithm 3, under the assumptions of Theorem 5, the following statement holds true:*

$$A_{t+1} \leq \left(1-\frac{\eta\mu}{4}\right)A_t+\left(8\eta^2+32C_2\frac{\eta}{\mu}\right)\frac{(G_f^2+G_g^2)\sigma^2}{B}+\left(2\beta^2C_1+C_2\frac{8\eta}{\mu}G_g^2L_f^2\beta^2\right)\frac{\sigma^2}{M},$$

where $A_t=\mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2+C_1\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2+C_2\mathbb{E}\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2$, $C_1=220\eta G_g^2L_f^2\max\left\{1,\frac{1}{\mu},G_g^2\right\}$, *and* $C_2=\frac{\mu}{16\eta G_g^2L_f^2}C_1=14\mu\max\left\{1,\frac{1}{\mu},G_g^2\right\}$.

Proof. Recall that Algorithm 3 is optimizing a $\mu:=\frac{1}{\gamma}-\rho$ strongly convex strongly concave function $F^\gamma(\mathbf{x},\mathbf{y})=F(\mathbf{x},\mathbf{y})+\frac{1}{2\gamma}\left\|\mathbf{x}-\mathbf{x}^0\right\|^2-\frac{1}{2\gamma}\left\|\mathbf{y}-\mathbf{y}^0\right\|^2$, and it is $\tilde{L}:=L+\frac{1}{\gamma}$ smooth. According to updating rule we have:

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2 \leq & \mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^{t-1}-\eta(\mathbf{g}^t-\mathbf{g}^{t-1})\right\|^2 \\
\leq & \left(1+\frac{1}{2(\frac{1}{2\mu\eta}-1)}\right)\mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^{t-1}-\eta(\nabla F^\gamma(\mathbf{x}^t,\mathbf{y}^t)-\nabla F^\gamma(\mathbf{x}^{t-1},\mathbf{y}^{t-1}))\right\|^2 \\
& + \left(1+2(\frac{1}{2\mu\eta}-1)\right)\eta^2\mathbb{E}\left\|\mathbf{g}^t-\nabla F^\gamma(\mathbf{x}^t,\mathbf{y}^t)-(\nabla F^\gamma(\mathbf{x}^{t-1},\mathbf{y}^{t-1})-\mathbf{g}^{t-1})\right\|^2 \\
\leq & \left(1+\frac{1}{2(\frac{1}{2\mu\eta}-1)}\right)(1-2\eta\mu)\,\mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^{t-1}\right\|^2+\mathbb{E}\left\|\eta(\nabla F^\gamma(\mathbf{x}^t,\mathbf{y}^t)-\nabla F^\gamma(\mathbf{x}^{t-1},\mathbf{y}^{t-1}))\right\|^2 \\
& + \frac{1}{\mu\eta}\eta^2\mathbb{E}\left\|\mathbf{g}^t-\nabla F^\gamma(\mathbf{x}^t,\mathbf{y}^t)-(\nabla F^\gamma(\mathbf{x}^{t-1},\mathbf{y}^{t-1})-\mathbf{g}^{t-1})\right\|^2 \\
\leq & \left(1-\eta\mu+\eta^2\tilde{L}^2\right)\mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^{t-1}\right\|^2 \\
& + \frac{4\eta}{\mu}G_g^2L_f^2\left(\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2+\mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{x}^{t-1},\mathbf{y}^{t-1})\right\|^2\right)+32\frac{\eta(G_g^2+G_f^2)\sigma^2}{\mu B},
\end{aligned}$$

where the last step is due to

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{g}^t-\bar{\mathbf{g}}^t\right\|^2 = & \mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^t-\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2+\mathbb{E}\left\|\mathbf{g}_{\mathbf{y}}^t-\bar{\mathbf{g}}_{\mathbf{y}}^t\right\|^2 \\
= & \underbrace{\mathbb{E}\left\|\frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t}\nabla f(\mathbf{z}^{t+1};\zeta)\nabla_{\mathbf{x}}g(\mathbf{x}^t,\mathbf{y}^t;\xi)-\nabla f(\mathbf{z}^{t+1})\nabla_{\mathbf{x}}g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2}_{A} \\
& + \underbrace{\mathbb{E}\left\|\frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t}\nabla f(\mathbf{z}^{t+1};\zeta)\nabla_{\mathbf{y}}g(\mathbf{x}^t,\mathbf{y}^t;\xi)-\nabla f(\mathbf{z}^{t+1})\nabla_{\mathbf{y}}g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2}_{B}.
\end{aligned}$$

For A, we have

$$
\begin{aligned}
A \leq & 2\mathbb{E}\left\|\frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t}\nabla f(\mathbf{z}^{t+1};\zeta)\nabla_{\mathbf{x}} g(\mathbf{x}^t,\mathbf{y}^t;\xi) - \frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t}\nabla f(\mathbf{z}^{t+1};\zeta)\nabla_{\mathbf{x}} g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 \\
& + 2\mathbb{E}\left\|\frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t}\nabla f(\mathbf{z}^{t+1};\zeta)\nabla_{\mathbf{x}} g(\mathbf{x}^t,\mathbf{y}^t) - \nabla f(\mathbf{z}^{t+1})\nabla_{\mathbf{x}} g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 \\
\leq & 2\frac{1}{B^2}\sum_{(\zeta,\xi)\in\mathcal{B}^t}\mathbb{E}\left\|\nabla f(\mathbf{z}^{t+1};\zeta)\nabla_{\mathbf{x}} g(\mathbf{x}^t,\mathbf{y}^t;\xi) - \nabla f(\mathbf{z}^{t+1};\zeta)\nabla_{\mathbf{x}} g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 \\
& + 2\frac{1}{B^2}\sum_{(\zeta,\xi)\in\mathcal{B}^t}\mathbb{E}\left\|\nabla f(\mathbf{z}^{t+1};\zeta)\nabla_{\mathbf{x}} g(\mathbf{x}^t,\mathbf{y}^t) - \nabla f(\mathbf{z}^{t+1})\nabla_{\mathbf{x}} g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 \\
= & 2\frac{(G_f^2+G_g^2)\sigma^2}{B}.
\end{aligned}
$$

Similarly we know——- $B \leq 2\frac{(G_f^2+G_g^2)\sigma^2}{B}$. Now we examine the convergence to optimal point:

$$
\begin{aligned}
\mathbb{E}\left\|\mathbf{w}^{t+1}-\mathbf{w}^*\right\|^2 &\leq \mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^*-\eta(\mathbf{g}^t-\mathbf{g}(\mathbf{w}^*))\right\|^2 \\
&= \mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2+\eta^2\mathbb{E}\left\|\mathbf{g}^t-\mathbf{g}(\mathbf{w}^*)\right\|^2 - 2\eta\mathbb{E}\left\langle \mathbf{g}^t-\mathbf{g}(\mathbf{w}^*),\, \mathbf{w}^t-\mathbf{w}^*\right\rangle \\
&= \mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2+\eta^2\mathbb{E}\left\|\mathbf{g}^t-\mathbf{g}(\mathbf{w}^*)\right\|^2 - 2\eta\mathbb{E}\left\langle \mathbf{g}(\mathbf{w}^t)-\mathbf{g}(\mathbf{w}^*),\, \mathbf{w}^t-\mathbf{w}^*\right\rangle \\
&\quad - 2\eta\mathbb{E}\left\langle \mathbf{g}^t-\mathbf{g}(\mathbf{w}^t),\, \mathbf{w}^t-\mathbf{w}^*\right\rangle.
\end{aligned}
$$

Due to μ strong monotoncity we have:

$$
\begin{aligned}
\mathbb{E}\left\|\mathbf{w}^{t+1}-\mathbf{w}^*\right\|^2 &\leq (1-2\eta\mu)\mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2+\eta^2\mathbb{E}\left\|\bar{\mathbf{g}}^t-\mathbf{g}(\mathbf{w}^*)\right\|^2 + 8\eta^2\frac{(G_f^2+G_g^2)\sigma^2}{B} \\
&\quad - 2\eta\mathbb{E}\left\langle \mathbf{g}^t-\mathbf{g}(\mathbf{w}^t),\, \mathbf{w}^t-\mathbf{w}^*\right\rangle \\
&\leq (1-2\eta\mu)\mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2+\eta^2\mathbb{E}\left\|\bar{\mathbf{g}}^t-\mathbf{g}(\mathbf{w}^*)\right\|^2 + 8\eta^2\frac{(G_f^2+G_g^2)\sigma^2}{B} \\
&\quad + \left(\frac{\eta}{\mu}\left\|\bar{\mathbf{g}}^t-\mathbf{g}(\mathbf{w}^t)\right\|^2+\eta\mu\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2\right) \\
&\leq (1-2\eta\mu+2\eta^2\tilde{L}^2)\mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2+2\eta^2\mathbb{E}\left\|\bar{\mathbf{g}}^t-\mathbf{g}(\mathbf{w}^t)\right\|^2 + 8\eta^2\frac{(G_f^2+G_g^2)\sigma^2}{B} \\
&\quad + \left(\frac{\eta}{\mu}\left\|\bar{\mathbf{g}}^t-\mathbf{g}(\mathbf{w}^t)\right\|^2+\eta\mu\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2\right) \\
&\leq (1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)\mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2+\left(2\eta^2G_g^2L_f^2+\frac{\eta}{\mu}G_g^2L_f^2\right)\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 \\
&\quad + 8\eta^2\frac{(G_f^2+G_g^2)\sigma^2}{B}.
\end{aligned}
$$

Now we add $C_1\mathbb{E}\left\|\mathbf{z}^{t+2}-g(\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2$ on both side, and $C_1(1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2$

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}^{t+1}-\mathbf{w}^*\right\|^2 + C_1\mathbb{E}\left\|\mathbf{z}^{t+2}-g(\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2 \\
&\leq (1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)(\mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2+C_1\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2) \\
&\quad + \left(2\eta^2G_g^2L_f^2+\frac{\eta}{\mu}G_g^2L_f^2\right)\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 + 8\eta^2\frac{(G_f^2+G_g^2)\sigma^2}{B} \\
&\quad + C_1\mathbb{E}\left\|\mathbf{z}^{t+2}-g(\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2 - C_1(1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2.
\end{aligned}
$$

We apply Proposition 2 to replace $\mathbb{E}\left\|\mathbf{z}^{t+2}-g(\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2$:

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}^{t+1}-\mathbf{w}^*\right\|^2+C_1\mathbb{E}\left\|\mathbf{z}^{t+2}-g(\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2\\
&\leq (1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)(\mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2+C_1\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2)+8\eta^2\frac{(G_f^2+G_g^2)\sigma^2}{B}\\
&\quad+\left(2\eta^2G_g^2L_f^2+\frac{\eta}{\mu}G_g^2L_f^2\right)\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2\\
&\quad+C_1\left((1-\beta)^2\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2+4(1-\beta)^2G_g^2\mathbb{E}\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2+2\beta^2\frac{\sigma^2}{M}\right)\\
&\quad-C_1(1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2\\
&\leq (1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)(\mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2+C_1\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2)+8\eta^2\frac{(G_f^2+G_g^2)\sigma^2}{B}+2\beta^2C_1\frac{\sigma^2}{M}\\
&\quad+4(1-\beta)^2C_1G_g^2\mathbb{E}\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2\\
&\quad+\left(2\eta^2G_g^2L_f^2+\frac{\eta}{\mu}G_g^2L_f^2+(1-\beta)^2C_1-(1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)C_1\right)\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2.
\end{aligned}
$$

Adding $C_2\mathbb{E}\left\|\mathbf{w}^{t+2}-\mathbf{w}^{t+1}\right\|^2$ and $(1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)C_2\mathbb{E}\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2$ on both sides yields:

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}^{t+1}-\mathbf{w}^*\right\|^2+C_1\mathbb{E}\left\|\mathbf{z}^{t+2}-g(\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2+C_2\mathbb{E}\left\|\mathbf{w}^{t+2}-\mathbf{w}^{t+1}\right\|^2\\
&\leq (1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)\left(\mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2+C_1\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2+C_2\mathbb{E}\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2\right)+8\eta^2\frac{(G_f^2+G_g^2)\sigma^2}{B}+2\beta^2C_1\frac{\sigma^2}{M}\\
&\quad+4(1-\beta)^2C_1G_g^2\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2+\left(2\eta^2G_g^2L_f^2+\frac{\eta}{\mu}G_g^2L_f^2+(1-\beta)^2C_1-(1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)C_1\right)\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2\\
&\quad+C_2\mathbb{E}\left\|\mathbf{w}^{t+2}-\mathbf{w}^{t+1}\right\|^2-(1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)C_2\mathbb{E}\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2.
\end{aligned}
$$

We now plug in the bound for $\mathbb{E}\left\|\mathbf{w}^{t+2}-\mathbf{w}^{t+1}\right\|^2$ and define $A_t=\mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2+C_1\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2+C_2\mathbb{E}\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2$:

$$
\begin{aligned}
&A_{t+1}\leq (1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)A_t+8\eta^2\frac{(G_f^2+G_g^2)\sigma^2}{B}+2\beta^2C_1\frac{\sigma^2}{M}\\
&+\left(2\eta^2G_g^2L_f^2+\frac{\eta}{\mu}G_g^2L_f^2+(1-\beta)^2C_1-(1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)C_1\right)\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2\\
&+\left(4(1-\beta)^2C_1G_g^2-(1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)C_2\right)\mathbb{E}\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2\\
&+C_2\\
&\times\left(\left(1-\eta\mu+\eta^2\tilde{L}^2\right)\mathbb{E}\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2+\frac{4\eta}{\mu}G_g^2L_f^2\left(\mathbb{E}\left\|\mathbf{z}^{t+2}-g(\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2+\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2\right)+32\frac{\eta(G_g^2+G_f^2)\sigma^2}{\mu B}\right)\\
&=(1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)A_t+8\eta^2\frac{(G_f^2+G_g^2)\sigma^2}{B}+2\beta^2C_1\frac{\sigma^2}{M}+32\frac{C_2\eta(G_g^2+G_f^2)\sigma^2}{\mu B}\\
&+\left(2\eta^2G_g^2L_f^2+\frac{\eta}{\mu}G_g^2L_f^2+(1-\beta)^2C_1-(1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)C_1\right)\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2\\
&+\left(4(1-\beta)^2C_1G_g^2+C_2\left(1-\eta\mu+\eta^2\tilde{L}^2\right)-(1-\frac{\eta\mu}{2}+2\eta^2\tilde{L}^2)C_2\right)\mathbb{E}\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2\\
&+C_2\frac{4\eta}{\mu}G_g^2L_f^2\left(\mathbb{E}\left\|\mathbf{z}^{t+2}-g(\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2+\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2\right).
\end{aligned}
$$

Again applying the bound for $\mathbb{E}\left\|\mathbf{z}^{t+2} - g(\mathbf{x}^{t+1}, \mathbf{y}^{t+1})\right\|^2$ yields:

$$\begin{aligned}
A_{t+1} &\leq (1 - \frac{\eta\mu}{2} + 2\eta^2\tilde{L}^2)A_t + 8\eta^2\frac{(G_f^2 + G_g^2)\sigma^2}{B} + 2\beta^2 C_1\frac{\sigma^2}{M} + 32\frac{C_2\eta(G_g^2 + G_f^2)\sigma^2}{\mu B} \\
&+ \left(2\eta^2 G_g^2 L_f^2 + \frac{\eta}{\mu}G_g^2 L_f^2 + (1-\beta)^2 C_1 - (1 - \frac{\eta\mu}{2} + 2\eta^2\tilde{L}^2)C_1\right)\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 \\
&+ \left(4(1-\beta)^2 C_1 G_g^2 + C_2\left(1 - \eta\mu + \eta^2\tilde{L}^2\right) - (1 - \frac{\eta\mu}{2} + 2\eta^2\tilde{L}^2)C_2\right)\mathbb{E}\left\|\mathbf{w}^{t+1} - \mathbf{w}^t\right\|^2 \\
&+ C_2\frac{4\eta}{\mu}G_g^2 L_f^2\left((1-\beta)^2\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 + 4(1-\beta)^2 G_g^2\mathbb{E}\left\|\mathbf{w}^{t+1} - \mathbf{w}^t\right\|^2 + 2\beta^2\frac{\sigma^2}{M} + \mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t, \mathbf{y}^t)\right\|^2\right) \\
&= (1 - \frac{\eta\mu}{2} + 2\eta^2\tilde{L}^2)A_t + 8\eta^2\frac{(G_f^2 + G_g^2)\sigma^2}{B} + 2\beta^2 C_1\frac{\sigma^2}{M} + C_2\frac{8\eta}{\mu}G_g^2 L_f^2\beta^2\frac{\sigma^2}{M} + 32\frac{C_2\eta(G_g^2 + G_f^2)\sigma^2}{\mu B} \\
&+ \left(2\eta^2 G_g^2 L_f^2 + \frac{\eta}{\mu}G_g^2 L_f^2 + (1-\beta)^2 C_1 - (1 - \frac{\eta\mu}{2} + 2\eta^2\tilde{L}^2)C_1 + C_2\frac{4\eta}{\mu}G_g^2 L_f^2\left(1 + (1-\beta)^2\right)\right)\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 \\
&+ \left(4(1-\beta)^2 C_1 G_g^2 - \left(\frac{\eta\mu}{2} - \frac{16\eta}{\mu}G_g^4 L_f^2(1-\beta)^2\right)C_2\right)\mathbb{E}\left\|\mathbf{w}^{t+1} - \mathbf{w}^t\right\|^2.
\end{aligned}$$

We choose $\eta \leq \min\left\{\frac{\mu^3}{512}, \frac{\mu}{8\tilde{L}^2}\right\}$ and $(1-\beta)^2 \leq \frac{\mu^2}{256 G_g^4 L_f^2}$, we know

$$\begin{aligned}
A_{t+1} &\leq (1 - \frac{\eta\mu}{2} + 2\eta^2\tilde{L}^2)A_t + 8\eta^2\frac{(G_f^2 + G_g^2)\sigma^2}{B} + 2\beta^2 C_1\frac{\sigma^2}{M} + C_2\frac{8\eta}{\mu}G_g^2 L_f^2\beta^2\frac{\sigma^2}{M} + 32\frac{C_2\eta(G_g^2 + G_f^2)\sigma^2}{\mu B} \\
&+ \underbrace{\left(2\eta^2 G_g^2 L_f^2 + \frac{\eta}{\mu}G_g^2 L_f^2 + (1-\beta)^2 C_1 - (1 - \frac{\eta\mu}{4})C_1 + C_2\frac{8\eta}{\mu}G_g^2 L_f^2\right)}_{T_1}\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t, \mathbf{y}^t)\right\|^2 \\
&+ \underbrace{\left(\frac{\mu^2}{64 G_g^2 L_f^2}C_1 - \frac{\eta\mu}{4}C_2\right)}_{T_2}\mathbb{E}\left\|\mathbf{w}^{t+1} - \mathbf{w}^t\right\|^2.
\end{aligned}$$

Since we choose $C_2 = \frac{\mu}{16\eta G_g^2 L_f^2}C_1$, we know $T_2 \leq 0$, together with $\eta \leq \frac{\mu^3}{512}$ which yields:

$$\begin{aligned}
A_{t+1} &\leq (1 - \frac{\eta\mu}{4})A_t + \left(8\eta^2 + 32C_2\frac{\eta}{\mu}\right)\frac{(G_f^2 + G_g^2)\sigma^2}{B} + \left(2\beta^2 C_1 + C_2\frac{8\eta}{\mu}G_g^2 L_f^2\beta^2\right)\frac{\sigma^2}{M} \\
&+ \underbrace{\left(2\eta^2 G_g^2 L_f^2 + \frac{\eta}{\mu}G_g^2 L_f^2 + (1-\beta)^2 C_1 - (\frac{1}{2} - \frac{\eta\mu}{4})C_1\right)}_{T_1}\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t, \mathbf{y}^t)\right\|^2.
\end{aligned}$$

To ensure $T_1 \leq 0$, we need

$$C_1 \leq \frac{2\eta^2 G_g^2 L_f^2 + \frac{\eta}{\mu}G_g^2 L_f^2}{\frac{1}{2} - \frac{\eta\mu}{4} - (1-\beta)^2},$$

It can be satisfied If we choose

$$(1-\beta)^2 \leq \frac{1}{2} - \frac{\eta\mu}{4} - \frac{\mu}{220\max\left\{1, \frac{1}{\mu}, G_f^2\right\}}$$

$$C_1 = 220\eta G_g^2 L_f^2 \max\left\{1, \frac{1}{\mu}, G_g^2\right\}$$

and we know $T_1 \leq 0$, which concludes the proof.

$\square$

Proposition 3 (Three points). *If $\tilde{\mathbf{w}} = \mathcal{P}_{\mathcal{W}}(\mathbf{v} - \eta\mathbf{g})$, then the following statements holds:*

$$\langle\eta\mathbf{g}, \tilde{\mathbf{w}}\rangle + \frac{1}{2}\left\|\mathbf{v} - \tilde{\mathbf{w}}\right\|^2 + \frac{1}{2}\left\|\mathbf{w} - \tilde{\mathbf{w}}\right\|^2 \leq \langle\eta\mathbf{g}, \mathbf{w}\rangle + \frac{1}{2}\left\|\mathbf{w} - \mathbf{v}\right\|^2, \forall\mathbf{w} \in \mathcal{W}.$$

Lemma 21. *Define* $\mathbf{u}^{t+1} = \mathcal{P}_{\mathcal{W}}(\mathbf{w}^t - \eta \mathbf{g}^{t+1})$. *Then for Algorithm 3, under the assumptions of Theorem 5, the following statement holds:*

$$\begin{aligned}\frac{1}{8}\mathbb{E}\left\|\mathbf{w}^t - \mathbf{w}^{t+1}\right\|^2 \leq &\frac{1}{2}\mathbb{E}\left\|\mathbf{w}^* - \mathbf{w}^t\right\|^2 + \left(\frac{2\eta}{\mu} + 72\eta\right) G_g^2 L_f^2 \beta^2 \frac{\sigma^2}{M} + \frac{48\eta(G_g^2 + G_f^2)\sigma^2}{B} \\ &+ \left(72\eta G_g^2 L_f^2 + 144\eta G_g^2 G_f^2 L_f^2 + \frac{\eta}{\mu} G_g^2 L_f^2\right)\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t, \mathbf{y}^t)\right\|^2.\end{aligned}$$

Proof. In Proposition 3, we set $\tilde{\mathbf{w}} = \mathbf{w}^{t+1}$, $\mathbf{v} = \mathbf{w}^t$ and $\mathbf{w} = \mathbf{u}^{t+1}$

$$\begin{aligned}&\left\langle \eta \mathbf{g}^t, \mathbf{w}^{t+1}\right\rangle + \frac{1}{2}\left\|\mathbf{w}^t - \mathbf{w}^{t+1}\right\|^2 + \frac{1}{2}\left\|\mathbf{u}^{t+1} - \mathbf{w}^{t+1}\right\|^2 \leq \left\langle \eta \mathbf{g}^t, \mathbf{u}^{t+1}\right\rangle + \frac{1}{2}\left\|\mathbf{u}^{t+1} - \mathbf{w}^t\right\|^2 \\ &\Longleftrightarrow \\ &\left\langle \eta \mathbf{g}^t, \mathbf{w}^{t+1} - \mathbf{u}^{t+1}\right\rangle + \frac{1}{2}\left\|\mathbf{w}^t - \mathbf{w}^{t+1}\right\|^2 + \frac{1}{2}\left\|\mathbf{u}^{t+1} - \mathbf{w}^{t+1}\right\|^2 \leq \frac{1}{2}\left\|\mathbf{u}^{t+1} - \mathbf{w}^t\right\|^2. \qquad (30)\end{aligned}$$

Again in Proposition 3, we set $\tilde{\mathbf{w}} = \mathbf{u}^{t+1}$, $\mathbf{v} = \mathbf{w}^t$, $\mathbf{g} = \mathbf{g}^{t+1}$

$$\begin{aligned}&\left\langle \eta \mathbf{g}^{t+1}, \mathbf{u}^{t+1}\right\rangle + \frac{1}{2}\left\|\mathbf{w}^t - \mathbf{u}^{t+1}\right\|^2 + \frac{1}{2}\left\|\mathbf{w} - \mathbf{u}^{t+1}\right\|^2 \leq \left\langle \eta \mathbf{g}^{t+1}, \mathbf{w}\right\rangle + \frac{1}{2}\left\|\mathbf{w} - \mathbf{w}^t\right\|^2 \\ &\Longleftrightarrow \\ &\left\langle \eta \mathbf{g}^{t+1}, \mathbf{u}^{t+1} - \mathbf{w}\right\rangle + \frac{1}{2}\left\|\mathbf{w}^t - \mathbf{u}^{t+1}\right\|^2 + \frac{1}{2}\left\|\mathbf{w} - \mathbf{u}^{t+1}\right\|^2 \leq \frac{1}{2}\left\|\mathbf{w} - \mathbf{w}^t\right\|^2 \\ &\Longleftrightarrow \\ &\left\langle \eta \mathbf{g}^{t+1}, \mathbf{w}^{t+1} - \mathbf{w}\right\rangle - \left\langle \eta \mathbf{g}^{t+1}, \mathbf{w}^{t+1} - \mathbf{u}^{t+1}\right\rangle + \frac{1}{2}\left\|\mathbf{w}^t - \mathbf{u}^{t+1}\right\|^2 + \frac{1}{2}\left\|\mathbf{w} - \mathbf{u}^{t+1}\right\|^2 \leq \frac{1}{2}\left\|\mathbf{w} - \mathbf{w}^t\right\|^2. \qquad (31)\end{aligned}$$

Adding (30) and (31) yields:

$$\begin{aligned}&\eta\left\langle \mathbf{g}^t - \mathbf{g}^{t+1}, \mathbf{w}^{t+1} - \mathbf{u}^{t+1}\right\rangle + \left\langle \eta \mathbf{g}^{t+1}, \mathbf{w}^{t+1} - \mathbf{w}\right\rangle \\ &\qquad + \frac{1}{2}\left\|\mathbf{w}^t - \mathbf{w}^{t+1}\right\|^2 + \frac{1}{2}\left\|\mathbf{u}^{t+1} - \mathbf{w}^{t+1}\right\|^2 + \frac{1}{2}\left\|\mathbf{w} - \mathbf{u}^{t+1}\right\|^2 \leq \frac{1}{2}\left\|\mathbf{w} - \mathbf{w}^t\right\|^2 \\ &\Longleftrightarrow \\ &\eta\left\langle \mathbf{g}^t - \mathbf{g}^{t+1}, \mathbf{w}^{t+1} - \mathbf{u}^{t+1}\right\rangle + \left\langle \eta \mathbf{g}(\mathbf{w}^{t+1}), \mathbf{w}^{t+1} - \mathbf{w}\right\rangle + \eta\left\langle \mathbf{g}^{t+1} - \mathbf{g}(\mathbf{w}^{t+1}), \mathbf{w}^{t+1} - \mathbf{w}\right\rangle \\ &\qquad + \frac{1}{2}\left\|\mathbf{w}^t - \mathbf{w}^{t+1}\right\|^2 + \frac{1}{2}\left\|\mathbf{u}^{t+1} - \mathbf{w}^{t+1}\right\|^2 + \frac{1}{2}\left\|\mathbf{w} - \mathbf{u}^{t+1}\right\|^2 \leq \frac{1}{2}\left\|\mathbf{w} - \mathbf{w}^t\right\|^2.\end{aligned}$$

Setting $\mathbf{w} = \mathbf{w}^*$, according to $\frac{1}{\gamma}$-strongly monotone of $\mathbf{g}(\cdot)$, we have

$$\begin{aligned}&\eta\left\langle \mathbf{g}^t - \mathbf{g}^{t+1}, \mathbf{w}^{t+1} - \mathbf{u}^{t+1}\right\rangle + \eta\mu\left\|\mathbf{w}^{t+1} - \mathbf{w}^*\right\|^2 + \eta\left\langle \mathbf{g}^{t+1} - \mathbf{g}(\mathbf{w}^{t+1}), \mathbf{w}^{t+1} - \mathbf{w}^*\right\rangle \\ &\qquad + \frac{1}{2}\left\|\mathbf{w}^t - \mathbf{w}^{t+1}\right\|^2 + \frac{1}{2}\left\|\mathbf{u}^{t+1} - \mathbf{w}^{t+1}\right\|^2 + \frac{1}{2}\left\|\mathbf{w}^* - \mathbf{u}^{t+1}\right\|^2 \leq \frac{1}{2}\left\|\mathbf{w}^* - \mathbf{w}^t\right\|^2.\end{aligned}$$

For the two inner products, we apply Cauchy-Schwartz:

$$\begin{aligned}\eta\mathbb{E}\left\langle \mathbf{g}^{t+1} - \mathbf{g}(\mathbf{w}^{t+1}), \mathbf{w}^{t+1} - \mathbf{w}^*\right\rangle &= \eta\mathbb{E}\left\langle \bar{\mathbf{g}}^{t+1} - \mathbf{g}(\mathbf{w}^{t+1}), \mathbf{w}^{t+1} - \mathbf{w}^*\right\rangle \\ &\geq -\frac{1}{2}\eta\left(\frac{1}{\mu}\mathbb{E}\left\|\bar{\mathbf{g}}^{t+1} - \mathbf{g}(\mathbf{w}^{t+1})\right\|^2 + \mu\mathbb{E}\left\|\mathbf{w}^{t+1} - \mathbf{w}^*\right\|^2\right) \\ &\geq -\frac{1}{2}\eta\left(\frac{1}{\mu}2G_g^2 L_f^2\mathbb{E}\left\|\mathbf{z}^{t+2} - g(\mathbf{x}^{t+1}, \mathbf{y}^{t+1})\right\|^2 + \mu\mathbb{E}\left\|\mathbf{w}^{t+1} - \mathbf{w}^*\right\|^2\right),\end{aligned}$$

and

$$\begin{aligned}\eta\left\langle \mathbf{g}^t - \mathbf{g}^{t+1}, \mathbf{w}^{t+1} - \mathbf{u}^{t+1}\right\rangle &\geq -\frac{1}{2}\eta\left(\left\|\mathbf{g}^t - \mathbf{g}^{t+1}\right\|^2 + \left\|\mathbf{w}^{t+1} - \mathbf{u}^{t+1}\right\|^2\right) \\ &= -\frac{1}{2}\eta\left(\left\|\mathbf{g}_{\mathbf{x}}^t - \mathbf{g}_{\mathbf{x}}^{t+1}\right\|^2 + \left\|\mathbf{g}_{\mathbf{y}}^t - \mathbf{g}_{\mathbf{y}}^{t+1}\right\|^2 + \left\|\mathbf{w}^{t+1} - \mathbf{u}^{t+1}\right\|^2\right).\end{aligned}$$

For $\left\|\mathbf{g}_{\mathbf{x}}^t - \mathbf{g}_{\mathbf{x}}^{t+1}\right\|^2$ we have:

$$\begin{aligned}
\left\|\mathbf{g}_{\mathbf{x}}^t - \mathbf{g}_{\mathbf{x}}^{t+1}\right\|^2 \leq & 6\underbrace{\mathbb{E}\left\|\nabla f(\mathbf{z}^{t+1})\nabla_{\mathbf{x}} g(\mathbf{x}^t,\mathbf{y}^t) - \nabla f(\mathbf{z}^{t+2})\nabla_{\mathbf{x}} g(\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2}_{A} + \frac{4}{\gamma^2}\mathbb{E}\left\|\mathbf{x}^t - \mathbf{x}^{t+1}\right\|^2 \\
& + 6\underbrace{\mathbb{E}\left\|\nabla f(\mathbf{z}^{t+1})\nabla_{\mathbf{x}} g(\mathbf{x}^t,\mathbf{y}^t) - \frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^t}\nabla f(\mathbf{z}^{t+1};\zeta)\nabla_{\mathbf{x}} g(\mathbf{x}^t,\mathbf{y}^t;\xi)\right\|^2}_{B} \\
& + 6\underbrace{\mathbb{E}\left\|\nabla f(\mathbf{z}^{t+2})\nabla_{\mathbf{x}} g(\mathbf{x}^{t+1},\mathbf{y}^{t+1}) - \frac{1}{B}\sum_{(\zeta,\xi)\in\mathcal{B}^{t+1}}\nabla f(\mathbf{z}^{t+2};\zeta)\nabla_{\mathbf{x}} g(\mathbf{x}^{t+1},\mathbf{y}^{t+1};\xi)\right\|^2}_{C}.
\end{aligned}$$

We bound above terms as follows. For A:

$$A \leq 2G_g^2 L_f^2 \mathbb{E}\left\|\mathbf{z}^{t+1} - \mathbf{z}^{t+2}\right\|^2 + 2G_f^2 L_g^2\left\|(\mathbf{x}^t,\mathbf{y}^t) - (\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2$$

For B and C:

$$B, C \leq \frac{2(G_g^2 + G_f^2)\sigma^2}{B}$$

Putting pieces together we know

$$\begin{aligned}
\left\|\mathbf{g}_{\mathbf{x}}^t - \mathbf{g}_{\mathbf{x}}^{t+1}\right\|^2 \leq & 12G_g^2 L_f^2 \mathbb{E}\left\|\mathbf{z}^{t+1} - \mathbf{z}^{t+2}\right\|^2 + 12G_f^2 L_g^2\left\|(\mathbf{x}^t,\mathbf{y}^t) - (\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2 + \frac{4}{\gamma^2}\mathbb{E}\left\|\mathbf{x}^t - \mathbf{x}^{t+1}\right\|^2 \\
& + \frac{48(G_g^2 + G_f^2)\sigma^2}{B}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\left\|\mathbf{g}_{\mathbf{y}}^t - \mathbf{g}_{\mathbf{y}}^{t+1}\right\|^2 \leq & 12G_g^2 L_f^2 \mathbb{E}\left\|\mathbf{z}^{t+1} - \mathbf{z}^{t+2}\right\|^2 + 12G_f^2 L_g^2\left\|(\mathbf{x}^t,\mathbf{y}^t) - (\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2 + \frac{4}{\gamma^2}\mathbb{E}\left\|\mathbf{y}^t - \mathbf{y}^{t+1}\right\|^2 \\
& + \frac{48(G_g^2 + G_f^2)\sigma^2}{B}.
\end{aligned}$$

Hence

$$\eta\left\langle \mathbf{g}^t - \mathbf{g}^{t+1}, \mathbf{w}^{t+1} - \mathbf{u}^{t+1}\right\rangle \geq -\frac{1}{2}\eta\left(24G_g^2 L_f^2 \mathbb{E}\left\|\mathbf{z}^{t+1} - \mathbf{z}^{t+2}\right\|^2 + \left(24G_f^2 L_g^2 + \frac{4}{\gamma^2}\right)\left\|\mathbf{w}^t - \mathbf{w}^{t+1}\right\|^2 + \frac{96(G_g^2 + G_f^2)\sigma^2}{B}\right).$$

Plugging back yields:

$$\begin{aligned}
& -\frac{1}{2}\eta\left(24G_g^2 L_f^2 \mathbb{E}\left\|\mathbf{z}^{t+1} - \mathbf{z}^{t+2}\right\|^2 + \left(24G_f^2 L_g^2 + \frac{4}{\gamma^2}\right)\left\|\mathbf{w}^t - \mathbf{w}^{t+1}\right\|^2 + \frac{96(G_g^2 + G_f^2)\sigma^2}{B}\right) \\
& -\frac{1}{2}\eta\left(\frac{1}{\mu}2G_g^2 L_f^2 \mathbb{E}\left\|\mathbf{z}^{t+2} - g(\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2 + \frac{1}{\gamma}\mathbb{E}\left\|\mathbf{w}^{t+1} - \mathbf{w}^*\right\|^2\right) \\
& + \eta\frac{1}{\gamma}\left\|\mathbf{w}^{t+1} - \mathbf{w}^*\right\| + \frac{1}{2}\left\|\mathbf{w}^t - \mathbf{w}^{t+1}\right\|^2 + \frac{1}{2}\left\|\mathbf{u}^{t+1} - \mathbf{w}^{t+1}\right\|^2 + \frac{1}{2}\left\|\mathbf{w}^* - \mathbf{u}^{t+1}\right\|^2 \leq \frac{1}{2}\left\|\mathbf{w}^* - \mathbf{w}^t\right\|^2 \\
& \Longleftrightarrow \\
& -12\eta G_g^2 L_f^2 \mathbb{E}\left\|\mathbf{z}^{t+1} - \mathbf{z}^{t+2}\right\|^2 - \frac{\eta}{\mu}G_g^2 L_f^2 \mathbb{E}\left\|\mathbf{z}^{t+2} - g(\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2 + \frac{\mu\eta}{2}\left\|\mathbf{w}^{t+1} - \mathbf{w}^*\right\|^2 - \frac{48\eta(G_g^2 + G_f^2)\sigma^2}{B} \\
& + \frac{1}{2}\left(1 - \eta\left(24G_f^2 L_g^2 + \frac{4}{\gamma^2}\right)\right)\left\|\mathbf{w}^t - \mathbf{w}^{t+1}\right\|^2 + \left(\frac{1}{2} - \frac{\eta}{2}\right)\left\|\mathbf{u}^{t+1} - \mathbf{w}^{t+1}\right\|^2 + \frac{1}{2}\left\|\mathbf{w}^* - \mathbf{u}^{t+1}\right\|^2 \leq \frac{1}{2}\left\|\mathbf{w}^* - \mathbf{w}^t\right\|^2.
\end{aligned}$$

By our choice $\eta \leq \frac{1}{2}\left(24G_f^2L_g^2 + \frac{4}{\gamma^2}\right)^{-1}$, we have $\left(1-\eta\left(24G_f^2L_g^2+\frac{4}{\gamma^2}\right)\right) \geq \frac{1}{2}$, which yields:

$$
\begin{aligned}
\frac{1}{4}\left\|\mathbf{w}^t-\mathbf{w}^{t+1}\right\|^2 &\leq \frac{1}{2}\left\|\mathbf{w}^*-\mathbf{w}^t\right\|^2 + 12\eta G_g^2L_f^2\mathbb{E}\left\|\mathbf{z}^{t+1}-\mathbf{z}^{t+2}\right\|^2 + \frac{\eta}{\mu}G_g^2L_f^2\mathbb{E}\left\|\mathbf{z}^{t+2}-g(\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2 + \frac{48\eta(G_g^2+G_f^2)\sigma^2}{B} \\
&\leq \frac{1}{2}\left\|\mathbf{w}^*-\mathbf{w}^t\right\|^2 + \frac{48\eta(G_g^2+G_f^2)\sigma^2}{B} \\
&+ 12\eta G_g^2L_f^2\left(3\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 + 3\mathbb{E}\left\|g(\mathbf{x}^t,\mathbf{y}^t)-g(\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2 + 3\mathbb{E}\left\|g(\mathbf{x}^{t+1},\mathbf{y}^{t+1})-\mathbf{z}^{t+2}\right\|^2\right) \\
&+ \frac{\eta}{\mu}G_g^2L_f^2\left((1-\beta)^2\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 + 4(1-\beta)^2G_g^2\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2 + 2\beta^2\frac{\sigma^2}{M}\right) \\
&\leq \frac{1}{2}\left\|\mathbf{w}^*-\mathbf{w}^t\right\|^2 + \left(36\eta G_g^4L_f^2 + 4(1-\beta)^2\frac{\eta}{\mu}G_g^4L_f^2\right)\left\|\mathbf{w}^t-\mathbf{w}^{t+1}\right\|^2 + 2\frac{\eta}{\mu}G_g^2L_f^2\beta^2\frac{\sigma^2}{M} + \frac{48\eta(G_g^2+G_f^2)\sigma^2}{B} \\
&+ \left(36\eta G_g^2L_f^2 + (1-\beta)^2\frac{\eta}{\mu}G_g^2L_f^2\right)\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 + 36\eta G_g^2L_f^2\mathbb{E}\left\|\mathbf{z}^{t+2}-g(\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2.
\end{aligned}
$$

Plugging bound for $\mathbb{E}\left\|\mathbf{z}^{t+2}-g(\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2$ yields:

$$
\begin{aligned}
\frac{1}{4}\left\|\mathbf{w}^t-\mathbf{w}^{t+1}\right\|^2 &\leq \frac{1}{2}\left\|\mathbf{w}^*-\mathbf{w}^t\right\|^2 + \left(36\eta G_g^4L_f^2 + 4(1-\beta)^2\frac{\eta}{\mu}G_g^4L_f^2\right)\left\|\mathbf{w}^t-\mathbf{w}^{t+1}\right\|^2 + 2\frac{\eta}{\mu}G_g^2L_f^2\beta^2\frac{\sigma^2}{M} + \frac{48\eta(G_g^2+G_f^2)\sigma^2}{B} \\
&+ \left(36\eta G_g^2L_f^2 + (1-\beta)^2\frac{\eta}{\mu}G_g^2L_f^2\right)\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 \\
&+ 36\eta G_g^2L_f^2\left((1-\beta)^2\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 + 4(1-\beta)^2G_g^2\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2 + 2\beta^2\frac{\sigma^2}{M}\right) \\
&= \frac{1}{2}\left\|\mathbf{w}^*-\mathbf{w}^t\right\|^2 + \left(36\eta G_g^4L_f^2 + 4(1-\beta)^2\frac{\eta}{\mu}G_g^4L_f^2\right)\left\|\mathbf{w}^t-\mathbf{w}^{t+1}\right\|^2 + \left(2\frac{\eta}{\mu}+72\eta\right)G_g^2L_f^2\beta^2\frac{\sigma^2}{M} \\
&+ \left(72\eta G_g^2L_f^2 + 144(1-\beta)^2\eta G_g^4L_f^2 + (1-\beta)^2\frac{\eta}{\mu}G_g^2L_f^2\right)\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 + \frac{48\eta(G_g^2+G_f^2)\sigma^2}{B}.
\end{aligned}
$$

Since we choose $\eta \leq \min\left\{\frac{1}{576G_g^4L_f^2}, \frac{\mu}{64G_g^4L_f^2}\right\}$, we know $36\eta G_g^4L_f^2 + 4(1-\beta)^2\frac{\eta}{\mu}G_g^4L_f^2 \leq \frac{1}{8}$, which yields:

$$
\begin{aligned}
\frac{1}{8}\mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^{t+1}\right\|^2 &\leq \frac{1}{2}\mathbb{E}\left\|\mathbf{w}^*-\mathbf{w}^t\right\|^2 + \left(\frac{2\eta}{\mu}+72\eta\right)G_g^2L_f^2\beta^2\frac{\sigma^2}{M} + \frac{48\eta(G_g^2+G_f^2)\sigma^2}{B} \\
&+ \left(72\eta G_g^2L_f^2 + 144\eta G_g^4L_f^2 + \frac{\eta}{\mu}G_g^2L_f^2\right)\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2.
\end{aligned}
$$

□

E.1 Proof of Lemma 18

Proof. First according to [Eq.(23) of Liu et al. [2021]],

$$
\max_{\mathbf{w}\in\mathcal{W}} \left\langle \nabla F_k(\mathbf{w}^{t+1}),\, \mathbf{w}^{t+1}-\mathbf{w}\right\rangle \leq (D+\frac{1}{\eta})\mathbb{E}\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2. \tag{32}
$$

Revoking Lemma 21

$$
\begin{aligned}
\frac{1}{8}\left\|\mathbf{w}^t-\mathbf{w}^{t+1}\right\|^2 &\leq \frac{1}{2}\left\|\mathbf{w}^*-\mathbf{w}^t\right\|^2 + \left(\frac{2\eta}{\mu}+72\eta\right)G_g^2L_f^2\beta^2\frac{\sigma^2}{M} + \frac{48\eta(G_g^2+G_f^2)\sigma^2}{B} \\
&+ \left(72\eta G_g^2L_f^2 + 144\eta G_g^4L_f^2 + \frac{\eta}{\mu}G_g^2L_f^2\right)\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2 \\
&\leq \frac{1}{2}\mathbb{E}\left\|\mathbf{w}^*-\mathbf{w}^t\right\|^2 + \left(\frac{2\eta}{\mu}+72\eta\right)G_g^2L_f^2\beta^2\frac{\sigma^2}{M} + \frac{48\eta(G_g^2+G_f^2)\sigma^2}{B} \\
&+ 220\eta G_g^2L_f^2\max\left\{1,\frac{1}{\mu},G_g^2\right\}\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2.
\end{aligned} \tag{33}
$$

To bound $\mathbb{E}\left\|\mathbf{w}^*-\mathbf{w}^t\right\|^2$, we revoke Lemma 20:

$$\begin{aligned}
&\left\|\mathbf{w}^{t+1}-\mathbf{w}^*\right\|^2+C_1\mathbb{E}\left\|\mathbf{z}^{t+2}-g(\mathbf{x}^{t+1},\mathbf{y}^{t+1})\right\|^2+C_2\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2\\
&\leq\left(1-\frac{\eta\mu}{2}\right)\left(\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2+C_1\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t,\mathbf{y}^t)\right\|^2+C_2\left\|\mathbf{w}^t-\mathbf{w}^{t-1}\right\|^2\right)\\
&\quad+\left(8\eta^2+32C_2\frac{\eta}{\mu}\right)\frac{(G_f^2+G_g^2)\sigma^2}{B}+\left(2\beta^2C_1+C_2\frac{8\eta}{\mu}G_g^2L_f^2\beta^2\right)\frac{\sigma^2}{M}\\
&\leq\left(1-\frac{\eta\mu}{2}\right)^t\left(\left\|\mathbf{w}^0-\mathbf{w}^*\right\|^2+C_1\mathbb{E}\left\|\mathbf{z}^1-g(\mathbf{x}^0,\mathbf{y}^0)\right\|^2+C_2\left\|\mathbf{w}^0-\mathbf{w}^{-1}\right\|^2\right)\\
&\quad+\frac{2}{\eta\mu}\left(\left(8\eta^2+32C_2\frac{\eta}{\mu}\right)\frac{(G_f^2+G_g^2)\sigma^2}{B}+\left(2C_1+C_2\frac{8\eta}{\mu}G_g^2L_f^2\right)\frac{\sigma^2}{M}\right).
\end{aligned}$$

where $C_1=220\eta G_g^2L_f^2\max\left\{1,\frac{1}{\mu},G_g^2\right\}$, and $C_2=14\mu\max\left\{1,\frac{1}{\mu},G_g^2\right\}$. Plugging above bound in (33) yields:

$$\begin{aligned}
\frac{1}{8}\left\|\mathbf{w}^t-\mathbf{w}^{t+1}\right\|^2&\leq\left(1-\frac{\eta\mu}{2}\right)^t\left(\left\|\mathbf{w}^0-\mathbf{w}^*\right\|^2+C_1\mathbb{E}\left\|\mathbf{z}^1-g(\mathbf{x}^0,\mathbf{y}^0)\right\|^2\right)\\
&\quad+\frac{2}{\eta\mu}\left(\left(8\eta^2+32C_2\frac{\eta}{\mu}\right)\frac{(G_f^2+G_g^2)\sigma^2}{B}+\left(2C_1+C_2\frac{8\eta}{\mu}G_g^2L_f^2\right)\frac{\sigma^2}{M}\right)\\
&\quad+\left(\frac{2\eta}{\mu}+72\eta\right)G_g^2L_f^2\beta^2\frac{\sigma^2}{M}+\frac{48\eta(G_g^2+G_f^2)\sigma^2}{B}.
\end{aligned}$$

Combining with (32) yields:

$$\begin{aligned}
\max_{\mathbf{w}\in\mathcal{W}}\left\langle\nabla F_k(\mathbf{w}^{t+1}),\,\mathbf{w}^{t+1}-\mathbf{w}\right\rangle&\leq\left(D+\frac{1}{\eta}\right)\left(1-\frac{\eta\mu}{2}\right)^t\left(\left\|\mathbf{w}^0-\mathbf{w}^*\right\|^2+C_1\mathbb{E}\left\|\mathbf{z}^1-g(\mathbf{x}^0,\mathbf{y}^0)\right\|^2\right)\\
&\quad+\left(D+\frac{1}{\eta}\right)\frac{2}{\eta\mu}\left(\left(8\eta^2+32C_2\frac{\eta}{\mu}\right)\frac{(G_f^2+G_g^2)\sigma^2}{B}+\left(2C_1+C_2\frac{8\eta}{\mu}G_g^2L_f^2\right)\frac{\sigma^2}{M}\right)\\
&\quad+\left(D+\frac{1}{\eta}\right)\left(\frac{2\eta}{\mu}+72\eta\right)G_g^2L_f^2\beta^2\frac{\sigma^2}{M}+\left(D+\frac{1}{\eta}\right)\frac{48\eta(G_g^2+G_f^2)\sigma^2}{B}.
\end{aligned}$$

To ensure RHS is less than ϵ, we need $t\geq\Omega(\frac{1}{\eta\mu}\ln(\frac{D+1/\eta}{\epsilon}))$ and $M=B\geq\Omega\left((D+\frac{1}{\eta})\frac{G_g^2L_f^2\sigma^2}{\mu\epsilon}\max\left\{1,\frac{1}{\mu},G_g^2\right\}\right)$, which yields the gradient complexity of

$$O\left(\left(D+\frac{1}{\eta}\right)\frac{L^2\sigma^2}{\mu^2\eta\epsilon}\log\left(\frac{D+1/\eta}{\epsilon}\right)\right).$$

□

E.2 Proof of Theorem 5

Proof. Let $\bar{\mathbf{w}}$ be the solution of SVI induced by $F_{\mathbf{w}}^{\gamma}:=F(\mathbf{x}',\mathbf{y}')+\frac{1}{2\gamma}(\mathbf{w}'-\mathbf{w})$. Let $\mathbf{w}_{k^*}=(\mathbf{x}_{k^*},\mathbf{y}_{k^*})$ be the solution returned by Algorithm 4. According to [Eq.(21) Liu et al. [2021]], if Algorithm 3 returns a $\tilde{\epsilon}$-accurate solution, then we have:

$$\mathbb{E}\left\|\mathbf{w}_{k^*}-\bar{\mathbf{w}}_{k^*}\right\|^2\leq\frac{2D\theta_K}{\sum_{k=0}^{K}\theta_k}+\frac{4c\tilde{\epsilon}}{\rho}.$$

Since we choose $\theta_k=(k+1)^\alpha$, we have

$$\mathbb{E}\left\|\mathbf{w}_{k^*}-\bar{\mathbf{w}}_{k^*}\right\|^2\leq\frac{2D(\alpha+1)}{K}+\frac{4c\tilde{\epsilon}}{\rho}.$$

According to [Lemma 6 Liu et al. [2021]], we know that if $\|\mathbf{w}-\bar{\mathbf{w}}\|\leq\gamma\epsilon$, then

$$dist\left(0,\partial(F(\bar{\mathbf{x}},\bar{\mathbf{y}})+\mathbf{1}_{\mathcal{X}\times\mathcal{Y}}(\bar{\mathbf{x}},\bar{\mathbf{y}}))\right)\leq\epsilon$$

To ensure $\mathbb{E}\left\|\mathbf{w}_{k^*}-\bar{\mathbf{w}}_{k^*}\right\|^2$ is less than $\gamma^2\epsilon^2$ where $\gamma=\frac{1}{2\rho}$, we need $K \geq \frac{16\rho^2 D(\alpha+1)}{\epsilon^2}$, and $\tilde{\epsilon} \leq \frac{\epsilon^2}{32\rho c}$, together by our choice $\eta=\Theta\left(\frac{1}{L^2}\right)$ which yields the gradient complexity of

$$O\left(\left(D+L^2\right)\frac{\rho^2 D L^4\sigma^2}{\epsilon^4}\log\left(\frac{1}{\epsilon}\right)\right).$$

□

F Proof of Varaince Reduction Algorithm

In this Section we will provide proofs for Theorem 6 and 7.

In the proof we will use the following notations:

$$\begin{aligned}
&F_k(\mathbf{x}^t,\mathbf{y}^t)=F(\mathbf{x}^t,\mathbf{y}^t)+\frac{\mu_x+L}{2}\left\|\mathbf{x}^t-\mathbf{x}_k\right\|^2,\\
&\bar{\mathbf{g}}_{\mathbf{x}}^t=\nabla_1 f(\mathbf{z}^{t+1},\mathbf{y}^{t+1})\nabla g(\mathbf{x}^t)+(\mu_x+L)(\mathbf{x}^t-\mathbf{x}^0),\\
&\bar{\mathbf{g}}_{\mathbf{y}}^t=\nabla_{\mathbf{y}} f(\mathbf{z}^{t+1},\mathbf{y}^{t+1}),
\end{aligned}$$

For notational convenience, we define $\tilde{\sigma}^2=\max\left\{\sigma^2,2(G_f^2+G_g^2)\sigma^2\right\}$.

The following Lemma shows that the tracking error for inner function is reduced via our variance reduction technique.

Lemma 22 (Tracking error with variance reduction)**.** *For Algorithm 5, under assumptions of Theorem 6, the following statement holds:*

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t)\right\|^2 &\leq (1-\beta)\,\mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{x}^{t-1})\right\|^2\\
&\quad+\left(1+\frac{1}{\beta}\right)\left(6(1-\beta)^2G_g^2\mathbb{E}\left\|\mathbf{x}^{t-1}-\mathbf{x}^t\right\|^2+3\beta^2\frac{\sigma^2}{B_\tau}+\frac{G_g^2}{B}\sum_{t'=t_p}^{t-1}\mathbb{E}\left\|\mathbf{x}^{t'}-\mathbf{x}^{t'-1}\right\|^2\right).
\end{aligned}$$

where t_p is the last iteration t such that $t \mod \tau=0$.

Proof. Recall the updating rule:

$$\mathbf{z}^{t+1}=(1-\beta)(\mathbf{z}^t+g(\mathbf{x}^t;\mathcal{I}_{\mathbf{z}}^t)-g(\mathbf{x}^{t-1};\mathcal{I}_{\mathbf{z}}^t))+\beta g^{t-1},$$

where $g^t=g^{t-1}+g(\mathbf{x}^t;\mathcal{I}_{\mathbf{z}}^t)-g(\mathbf{x}^{t-1};\mathcal{I}_{\mathbf{z}}^t)$ and every τ iterations, we use large mini-batch to update g^t: $g^t=g(\mathbf{x}^t;\mathcal{B}_{\mathbf{z}}^t)$. We hence have:

$$\begin{aligned}
\mathbb{E}\left\|g^t-g(\mathbf{x}^t)\right\|^2 &= \mathbb{E}\left\|g^{t-1}-g(\mathbf{x}^{t-1})+\frac{1}{B}\sum_{\xi\in\mathcal{I}_{\mathbf{z}}^t}g(\mathbf{x}^t;\xi)-g(\mathbf{x}^{t-1};\xi)+g(\mathbf{x}^{t-1})-g(\mathbf{x}^t)\right\|^2\\
&=\mathbb{E}\left\|g^{t-1}-g(\mathbf{x}^{t-1})\right\|^2+\mathbb{E}\left\|\frac{1}{B}\sum_{\xi\in\mathcal{I}_{\mathbf{z}}^t}g(\mathbf{x}^t;\xi)-g(\mathbf{x}^{t-1};\xi)+g(\mathbf{x}^{t-1})-g(\mathbf{x}^t)\right\|^2\\
&\leq \mathbb{E}\left\|g^{t-1}-g(\mathbf{x}^{t-1})\right\|^2+\frac{1}{B^2}\sum_{\xi\in\mathcal{I}_{\mathbf{z}}^t}\mathbb{E}\left\|g(\mathbf{x}^t;\xi^t)-g(\mathbf{x}^{t-1};\xi^t)\right\|^2\\
&\leq \mathbb{E}\left\|g^{t-1}-g(\mathbf{x}^{t-1})\right\|^2+\frac{G_g^2}{B}\mathbb{E}\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2\\
&\leq \mathbb{E}\left\|g^{t_p}-g(\mathbf{x}^{t_p})\right\|^2+\frac{G_g^2}{B}\sum_{t'=t_p}^{t-1}\mathbb{E}\left\|\mathbf{x}^{t'}-\mathbf{x}^{t'-1}\right\|^2\\
&\leq \frac{\sigma^2}{B_\tau}+\frac{G_g^2}{B}\sum_{t'=t_p}^{t-1}\mathbb{E}\left\|\mathbf{x}^{t'}-\mathbf{x}^{t'-1}\right\|^2.
\end{aligned}$$

Again by the updating rule we have:

$$\mathbf{z}^{t+1} - g(\mathbf{x}^t) = (1-\beta)(\mathbf{z}^t - g(\mathbf{x}^{t-1})) + (1-\beta)(g(\mathbf{x}^{t-1}) - g(\mathbf{x}^t)) + \beta(g^t - g(\mathbf{x}^t)) + (1-\beta)(g(\mathbf{x}^t;\xi^t) - g(\mathbf{x}^{t-1};\xi^t)).$$

Taking expected norm on both sides and applying Young's inequality yields:

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\right\|^2 &\leq \underbrace{(1+\beta)(1-\beta)^2}_{\leq 1-\beta}\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\right\|^2 \\
&\quad + (1+\frac{1}{\beta})\mathbb{E}\left\|(1-\beta)(g(\mathbf{x}^{t-1}) - g(\mathbf{x}^t)) + \beta(g^t - g(\mathbf{x}^t)) + (1-\beta)(g(\mathbf{x}^t;\xi^t) - g(\mathbf{x}^{t-1};\xi^t))\right\|^2 \\
&\leq (1-\beta)\,\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\right\|^2 \\
&\quad + (1+\frac{1}{\beta})\left(6(1-\beta)^2 G_g^2\mathbb{E}\left\|\mathbf{x}^{t-1} - \mathbf{x}^t\right\|^2 + 3\beta^2\mathbb{E}\left\|g^t - g(\mathbf{x}^t)\right\|^2\right) \\
&\leq (1-\beta)\,\mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\right\|^2 \\
&\quad + (1+\frac{1}{\beta})\left(6(1-\beta)^2 G_g^2\mathbb{E}\left\|\mathbf{x}^{t-1} - \mathbf{x}^t\right\|^2 + 3\beta^2\left(\frac{\sigma^2}{B_\tau}\right) + \frac{G_g^2}{B}\sum_{t'=t_p}^{t-1}\mathbb{E}\left\|\mathbf{x}^{t'} - \mathbf{x}^{t'-1}\right\|^2\right).
\end{aligned}$$

□

Lemma 23 (Gradient bounds)**.** *For Algorithm 5, under assumptions of Theorem 6, the following statement holds:*

$$\mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^t - \bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 \leq \frac{\tilde{\sigma}^2}{B_\tau} + \frac{1}{B}\sum_{t'=t_p}^{t-1}\left(2G_g^2L^2\left\|\mathbf{z}^{t'+1} - \mathbf{z}^{t'+2}\right\|^2 + 2G_g^2L^2\left\|\mathbf{y}^{t'+1} - \mathbf{y}^{t'+2}\right\|^2 + 2G_f^2L_g^2\left\|\mathbf{x}^{t'} - \mathbf{x}^{t'+1}\right\|^2\right)$$

$$\mathbb{E}\left\|\mathbf{q}^t - \bar{\mathbf{g}}_{\mathbf{y}}^t\right\|^2 \leq \frac{\tilde{\sigma}^2}{B_\tau} + \frac{1}{B}\sum_{t'=t_p}^{t} L_f^2\left(\mathbb{E}\left\|\mathbf{z}^{t'} - \mathbf{z}^{t'+1}\right\|^2 + \mathbb{E}\left\|\mathbf{y}^{t'} - \mathbf{y}^{t'+1}\right\|^2\right).$$

Proof. By updating rule in Algorithm 5 we have:

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^t - \bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 &= \mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^{t-1} + G_{\mathbf{x}}^k(\mathbf{x}^t, \mathbf{z}^{t+1}, \mathbf{y}^{t+1}; \mathcal{B}_{\mathbf{x}}^t) - G_{\mathbf{x}}^k(\mathbf{x}^{t-1}, \mathbf{z}^t, \mathbf{y}^t; \mathcal{B}_{\mathbf{x}}^t) - \bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 \\
&= \mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^{t-1} - \bar{\mathbf{g}}_{\mathbf{x}}^{t-1} + \bar{\mathbf{g}}_{\mathbf{x}}^{t-1} + G_{\mathbf{x}}^k(\mathbf{x}^t, \mathbf{z}^{t+1}, \mathbf{y}^{t+1}; \mathcal{B}_{\mathbf{x}}^t) - G_{\mathbf{x}}^k(\mathbf{x}^{t-1}, \mathbf{z}^t, \mathbf{y}^t; \mathcal{B}_{\mathbf{x}}^t) - \bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 \\
&= \mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^{t-1} - \bar{\mathbf{g}}_{\mathbf{x}}^{t-1}\right\|^2 + \mathbb{E}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^{t-1} - G_{\mathbf{x}}^k(\mathbf{x}^{t-1}, \mathbf{z}^t, \mathbf{y}^t; \mathcal{B}_{\mathbf{x}}^t) - (\bar{\mathbf{g}}_{\mathbf{x}}^t - G_{\mathbf{x}}^k(\mathbf{x}^t, \mathbf{z}^{t+1}, \mathbf{y}^{t+1}; \mathcal{B}_{\mathbf{x}}^t))\right\|^2,
\end{aligned}$$

where the last step is due to $E[\bar{\mathbf{g}}_{\mathbf{x}}^{t-1} - G_{\mathbf{x}}^k(\mathbf{x}^{t-1}, \mathbf{z}^t, \mathbf{y}^t; \mathcal{B}_{\mathbf{x}}^t) - (\bar{\mathbf{g}}_{\mathbf{x}}^t - G_{\mathbf{x}}^k(\mathbf{x}^t, \mathbf{z}^{t+1}, \mathbf{y}^{t+1}; \mathcal{B}_{\mathbf{x}}^t))] = 0$. We now turn to bounding the variance term.

$$\begin{aligned}
&\mathbb{E}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^{t-1} - G_{\mathbf{x}}^k(\mathbf{x}^{t-1}, \mathbf{z}^t, \mathbf{y}^t; \mathcal{B}_{\mathbf{x}}^t) - (\bar{\mathbf{g}}_{\mathbf{x}}^t - G_{\mathbf{x}}^k(\mathbf{x}^t, \mathbf{z}^{t+1}, \mathbf{y}^{t+1}; \mathcal{B}_{\mathbf{x}}^t))\right\|^2 \\
=&\frac{1}{B^2}\sum_{i=1}^{B}\mathbb{E}\left\|\bar{\mathbf{g}}_{\mathbf{x}}^{t-1} - G_{\mathbf{x}}^k(\mathbf{x}^{t-1}, \mathbf{z}^t, \mathbf{y}^t; \zeta_i, \xi_i) - (\bar{\mathbf{g}}_{\mathbf{x}}^t - G_{\mathbf{x}}^k(\mathbf{x}^t, \mathbf{z}^{t+1}, \mathbf{y}^{t+1}; \zeta_i, \xi_i))\right\|^2 \\
\leq&\frac{1}{B^2}\sum_{i=1}^{B}\mathbb{E}\left\|G_{\mathbf{x}}^k(\mathbf{x}^{t-1}, \mathbf{z}^t, \mathbf{y}^t; \zeta_i, \xi_i) - G_{\mathbf{x}}^k(\mathbf{x}^t, \mathbf{z}^{t+1}, \mathbf{y}^{t+1}; \zeta_i, \xi_i)\right\|^2 \\
\leq&\frac{1}{B^2}\sum_{i=1}^{B}\left\|\nabla_1 f(\mathbf{z}^t, \mathbf{y}^t; \zeta_i)\nabla g(\mathbf{x}^{t-1}; \xi_i) - \nabla_1 f(\mathbf{z}^{t+1}, \mathbf{y}^{t+1}; \zeta_i)\nabla g(\mathbf{x}^t; \xi_i)\right\|^2.
\end{aligned}$$

By simple variance decomposition we have:

$$\left\|\nabla_1 f(\mathbf{z}^t, \mathbf{y}^t; \zeta_i)\nabla g(\mathbf{x}^{t-1}; \xi_i) - \nabla_1 f(\mathbf{z}^{t+1}, \mathbf{y}^{t+1}; \zeta_i)\nabla g(\mathbf{x}^t; \xi_i)\right\|^2 \leq 2G_g^2L^2\left\|(\mathbf{z}^t, \mathbf{y}^t) - (\mathbf{z}^{t+1}, \mathbf{y}^{t+1})\right\|^2 + 2G_f^2L_g^2\left\|\mathbf{x}^{t-1} - \mathbf{x}^t\right\|^2.$$

Hence

$$\begin{aligned}\mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^t-\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 &\le \mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^{t-1}-\bar{\mathbf{g}}_{\mathbf{x}}^{t-1}\right\|^2+\frac{1}{B}\left(2G_g^2L^2\left\|\mathbf{z}^t-\mathbf{z}^{t+1}\right\|^2+2G_g^2L^2\left\|\mathbf{y}^t-\mathbf{y}^{t+1}\right\|^2+2G_f^2L_g^2\left\|\mathbf{x}^{t-1}-\mathbf{x}^t\right\|^2\right)\\ &\le \mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^{t_p}-\bar{\mathbf{g}}_{\mathbf{x}}^{t_p}\right\|^2+\frac{1}{B}\sum_{t'=t_p}^{t-1}\left(2G_g^2L^2\left\|\mathbf{z}^{t'+1}-\mathbf{z}^{t'+2}\right\|^2+2G_g^2L^2\left\|\mathbf{y}^{t'+1}-\mathbf{y}^{t'+2}\right\|^2+2G_f^2L_g^2\left\|\mathbf{x}^{t'}-\mathbf{x}^{t'+1}\right\|^2\right),\end{aligned}$$

where t_p denotes the last iteration such that $t \mod \tau = 0$. From the updating rule we know $\mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^{t_p}-\bar{\mathbf{g}}_{\mathbf{x}}^{t_p}\right\|^2 \le \frac{\tilde{\sigma}^2}{B_\tau}$. Similarly,

$$\begin{aligned}\mathbb{E}\left\|\mathbf{q}^{t+1}-\bar{\mathbf{g}}_{\mathbf{y}}^{t+1}\right\|^2 &= \mathbb{E}\left\|\mathbf{q}^t-\bar{\mathbf{g}}_{\mathbf{y}}^t\right\|^2+\mathbb{E}\left\|G_k(\mathbf{x}^{t+1},\mathbf{z}^{t+2},\mathbf{y}^{t+1};\mathcal{B}_{\mathbf{y}}^t)-\bar{\mathbf{g}}_{\mathbf{y}}^{t+1}-(G_k(\mathbf{x}^t,\mathbf{z}^{t+1},\mathbf{y}^t;\mathcal{B}_{\mathbf{y}}^t)-\bar{\mathbf{g}}_{\mathbf{y}}^t)\right\|^2\\ &\le \mathbb{E}\left\|\mathbf{q}^t-\bar{\mathbf{g}}_{\mathbf{y}}^t\right\|^2+\frac{1}{B^2}\sum_{i=1}^{B}\mathbb{E}\left\|\nabla_{\mathbf{y}}f(\mathbf{z}^t,\mathbf{y}^t;\zeta_i)-\nabla_{\mathbf{y}}f(\mathbf{z}^{t+1},\mathbf{y}^{t+1};\zeta_i)\right\|^2\\ &\le \mathbb{E}\left\|\mathbf{q}^t-\bar{\mathbf{g}}_{\mathbf{y}}^t\right\|^2+\frac{1}{B}L^2\mathbb{E}\left\|(\mathbf{z}^t,\mathbf{y}^t)-(\mathbf{z}^{t+1},\mathbf{y}^{t+1})\right\|^2\\ &\le \mathbb{E}\left\|\mathbf{q}^{t_p}-\bar{\mathbf{g}}_{\mathbf{y}}^{t_p}\right\|^2+\frac{1}{B}\sum_{t'=t_p}^{t}L_f^2\left(\mathbb{E}\left\|\mathbf{z}^{t'}-\mathbf{z}^{t'+1}\right\|^2+\mathbb{E}\left\|\mathbf{y}^{t'}-\mathbf{y}^{t'+1}\right\|^2\right)\\ &\le \frac{\tilde{\sigma}^2}{B_\tau}+\frac{1}{B}\sum_{t'=t_p}^{t}L_f^2\left(\mathbb{E}\left\|\mathbf{z}^{t'}-\mathbf{z}^{t'+1}\right\|^2+\mathbb{E}\left\|\mathbf{y}^{t'}-\mathbf{y}^{t'+1}\right\|^2\right).\end{aligned}$$

□

Lemma 24. *For Algorithm 5, under assumptions of Theorem 6, the following statement holds:*

$$\begin{aligned}F_k(\mathbf{x}^{t+1},\mathbf{y})-&F_k(\mathbf{x},\mathbf{y}^{t+1})\\ \le&-\left\langle\nabla_{\mathbf{y}}f(g(\mathbf{x}^{t+1}),\mathbf{y}^{t+1})-\nabla_{\mathbf{y}}f(g(\mathbf{x}^t),\mathbf{y}^t),\,\mathbf{y}^{t+1}-\mathbf{y}\right\rangle\\ &+\left\langle\nabla_{\mathbf{y}}f(g(\mathbf{x}^t),\mathbf{y}^t)-\nabla_{\mathbf{y}}f(g(\mathbf{x}^{t-1}),\mathbf{y}^{t-1}),\,\mathbf{y}^t-\mathbf{y}\right\rangle\\ &+\left\langle\nabla_{\mathbf{y}}f(g(\mathbf{x}^t),\mathbf{y}^t)-\nabla_{\mathbf{y}}f(g(\mathbf{x}^{t-1}),\mathbf{y}^{t-1}),\,\mathbf{y}^{t+1}-\mathbf{y}^t\right\rangle\\ &+\frac{L_f'}{2}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2+\frac{1}{2\eta_{\mathbf{y}}}\left(\left\|\mathbf{y}-\mathbf{y}^t\right\|^2-\left\|\mathbf{y}-\mathbf{y}^{t+1}\right\|^2-\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2\right)\\ &+\frac{1}{2\eta_{\mathbf{x}}}\left(\left\|\mathbf{x}-\mathbf{x}^t\right\|^2-\left\|\mathbf{x}-\mathbf{x}^{t+1}\right\|^2-\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2\right)-\frac{\mu_x}{2}\left\|\mathbf{x}-\mathbf{x}^t\right\|^2+\frac{\mu_x}{2}\left\|\mathbf{x}^{t+1}-\mathbf{x}\right\|^2\\ &+\frac{1}{\mu_x}\left\|\mathbf{g}_{\mathbf{x}}^t-\bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2+\left(\frac{L_f^2G_g^2}{\mu_x}+\frac{8L_f^2}{\mu_y}\right)\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t)\right\|^2+\frac{2L_f^2}{\mu_y}\left\|\mathbf{z}^t-g(\mathbf{x}^{t-1})\right\|^2\\ &+\frac{8}{\mu_y}\left\|\mathbf{q}^t-\nabla_{\mathbf{y}}f(\mathbf{z}^{t+1},\mathbf{y}^t)\right\|^2+\frac{2}{\mu_y}\left\|\mathbf{q}^{t-1}-\nabla_{\mathbf{y}}f(\mathbf{z}^t,\mathbf{y}^{t-1})\right\|^2,\end{aligned}$$

where $L_f' := L_f+\mu_x+L$.

Proof. Define $\phi(\mathbf{x}^t,\mathbf{z}^{t+1},\mathbf{y}^{t+1})=h(\mathbf{x}^t)+\nabla_1 f(\mathbf{z}^{t+1},\mathbf{y}^{t+1})g(\mathbf{x}^t),\ f_k(\mathbf{x},\mathbf{y}):=f(g(\mathbf{x}^t),\mathbf{y}^{t+1})+\frac{\mu_x}{2}\left\|\mathbf{x}-\mathbf{x}_k\right\|^2$. According to updating rule we have:

$$\mathbf{x}^{t+1}=\mathbf{x}^t-\eta_{\mathbf{x}}\mathbf{g}_{\mathbf{x}}^t=\arg\min_{\mathbf{x}} h(\mathbf{x})+\phi(\mathbf{x})-\phi(\mathbf{x}^t)-\left\langle\nabla_{\mathbf{x}}\phi(\mathbf{x}^t,\mathbf{z}^{t+1},\mathbf{y}^{t+1}),\,\mathbf{x}-\mathbf{x}^t\right\rangle.$$

Then according to [Lemma 7.1 Hamedani and Aybat [2018]] we have:

$$\begin{aligned}h(\mathbf{x}^{t+1})+\left\langle\mathbf{g}_{\mathbf{x}}^t,\,\mathbf{x}^{t+1}-\mathbf{x}\right\rangle &\le h(\mathbf{x})+\frac{1}{2\eta_{\mathbf{x}}}\left(\left\|\mathbf{x}-\mathbf{x}^t\right\|^2-\left\|\mathbf{x}-\mathbf{x}^{t+1}\right\|^2-\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2\right),\\ r(\mathbf{y}^{t+1})-\left\langle\mathbf{g}_{\mathbf{y}}^t,\,\mathbf{y}^{t+1}-\mathbf{y}\right\rangle &\le r(\mathbf{y})+\frac{1}{2\eta_{\mathbf{y}}}\left(\left\|\mathbf{y}-\mathbf{y}^t\right\|^2-\left\|\mathbf{y}-\mathbf{y}^{t+1}\right\|^2-\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2\right).\end{aligned}$$

Re-arranging terms yields:

$$h(\mathbf{x}^{t+1})+\left\langle\nabla_{\mathbf{x}} f_k(\mathbf{x}^t,\mathbf{y}^{t+1}),\, \mathbf{x}^{t+1}-\mathbf{x}\right\rangle \leq h(\mathbf{x}) + \frac{1}{2\eta_{\mathbf{x}}}\left(\left\|\mathbf{x}-\mathbf{x}^t\right\|^2 - \left\|\mathbf{x}-\mathbf{x}^{t+1}\right\|^2 - \left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2\right) + \varepsilon_{\mathbf{x}}^{t+1}, \quad (34)$$

$$\begin{aligned} r(\mathbf{y}^{t+1}) - &\left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^t),\mathbf{y}^t) + \theta(\nabla_{\mathbf{y}} f(g(\mathbf{x}^t),\mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}),\mathbf{y}^{t-1})),\, \mathbf{y}^{t+1}-\mathbf{y}\right\rangle \\ &\leq r(\mathbf{y}) + \frac{1}{2\eta_{\mathbf{y}}}\left(\left\|\mathbf{y}-\mathbf{y}^t\right\|^2 - \left\|\mathbf{y}-\mathbf{y}^{t+1}\right\|^2 - \left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2\right) + \varepsilon_{\mathbf{y}}^{t+1}, \end{aligned} \quad (35)$$

where

$$\begin{aligned} \varepsilon_{\mathbf{x}}^{t+1} &= \left\langle \mathbf{g}_{\mathbf{x}}^t - \nabla_{\mathbf{x}} f(g(\mathbf{x}^t),\mathbf{y}^{t+1}),\, \mathbf{x}^{t+1}-\mathbf{x}\right\rangle, \\ \varepsilon_{\mathbf{y}}^{t+1} &= \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^t),\mathbf{y}^t) + \theta(\nabla_{\mathbf{y}} f(g(\mathbf{x}^t),\mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}),\mathbf{y}^{t-1})) - \mathbf{g}_{\mathbf{y}}^t,\, \mathbf{y}^{t+1}-\mathbf{y}\right\rangle. \end{aligned}$$

The rest of the proof follows [Lemma 13 Zhang et al. [2022]]. Using μ_x-convexity of f_k w.r.t. $\mathbf{x}$, we can lower bound the inner product in (34) as

$$\begin{aligned} \left\langle\nabla_{\mathbf{x}} f_k(\mathbf{x}^t,\mathbf{y}^{t+1}),\, \mathbf{x}^{t+1}-\mathbf{x}\right\rangle &= \left\langle\nabla_{\mathbf{x}} f_k(\mathbf{x}^t,\mathbf{y}^{t+1}),\, \mathbf{x}^{t}-\mathbf{x}\right\rangle + \left\langle\nabla_{\mathbf{x}} f_k(\mathbf{x}^t,\mathbf{y}^{t+1}),\, \mathbf{x}^{t+1}-\mathbf{x}^t\right\rangle \\ &\geq f_k(\mathbf{x}^t,\mathbf{y}^{t+1}) - f_k(\mathbf{x},\mathbf{y}^{t+1}) + \frac{\mu_x}{2}\left\|\mathbf{x}^t-\mathbf{x}\right\|^2 + \left\langle\nabla_{\mathbf{x}} f_k(\mathbf{x}^t,\mathbf{y}^{t+1}),\, \mathbf{x}^{t+1}-\mathbf{x}^t\right\rangle. \end{aligned}$$

Putting pieces together yields:

$$\begin{aligned} h(\mathbf{x}^{t+1}) + f_k(\mathbf{x}^t,\mathbf{y}^{t+1}) - &f_k(\mathbf{x},\mathbf{y}^{t+1}) + r(\mathbf{y}^{t+1}) - r(\mathbf{y}) \\ &\leq -\frac{\mu_x}{2}\left\|\mathbf{x}^t-\mathbf{x}\right\|^2 - \left\langle\nabla_{\mathbf{x}} f_k(\mathbf{x}^t,\mathbf{y}^{t+1}),\, \mathbf{x}^{t+1}-\mathbf{x}^t\right\rangle \\ &\quad + \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^t),\mathbf{y}^t) + \theta(\nabla_{\mathbf{y}} f(g(\mathbf{x}^t),\mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}),\mathbf{y}^{t-1})),\, \mathbf{y}^{t+1}-\mathbf{y}\right\rangle \\ &\quad + h(\mathbf{x}) + \frac{1}{2\eta_{\mathbf{x}}}\left(\left\|\mathbf{x}-\mathbf{x}^t\right\|^2 - \left\|\mathbf{x}-\mathbf{x}^{t+1}\right\|^2 - \left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2\right) + \varepsilon_{\mathbf{x}}^{t+1} \\ &\quad + \frac{1}{2\eta_{\mathbf{y}}}\left(\left\|\mathbf{y}-\mathbf{y}^t\right\|^2 - \left\|\mathbf{y}-\mathbf{y}^{t+1}\right\|^2 - \left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2\right) + \varepsilon_{\mathbf{y}}^{t+1}. \end{aligned}$$

Adding $f_k(\mathbf{x}^{t+1},\mathbf{y}^{t+1})$ on both sides yields:

$$\begin{aligned} f_k(\mathbf{x}^{t+1},\mathbf{y}^{t+1}) + h(\mathbf{x}^{t+1}) - &f_k(\mathbf{x},\mathbf{y}^{t+1}) + r(\mathbf{y}^{t+1}) - r(\mathbf{y}) - h(\mathbf{x}) \\ &\leq f_k(\mathbf{x}^{t+1},\mathbf{y}^{t+1}) - f_k(\mathbf{x}^t,\mathbf{y}^{t+1}) - \left\langle\nabla_{\mathbf{x}} f_k(\mathbf{x}^t,\mathbf{y}^{t+1}),\, \mathbf{x}^{t+1}-\mathbf{x}^t\right\rangle - \frac{\mu_x}{2}\left\|\mathbf{x}^t-\mathbf{x}\right\|^2 \\ &\quad + \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^t),\mathbf{y}^t) + (\nabla_{\mathbf{y}} f(g(\mathbf{x}^t),\mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}),\mathbf{y}^{t-1})),\, \mathbf{y}^{t+1}-\mathbf{y}\right\rangle \\ &\quad + \frac{1}{2\eta_{\mathbf{x}}}\left(\left\|\mathbf{x}-\mathbf{x}^t\right\|^2 - \left\|\mathbf{x}-\mathbf{x}^{t+1}\right\|^2 - \left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2\right) + \varepsilon_{\mathbf{x}}^{t+1} \\ &\quad + \frac{1}{2\eta_{\mathbf{y}}}\left(\left\|\mathbf{y}-\mathbf{y}^t\right\|^2 - \left\|\mathbf{y}-\mathbf{y}^{t+1}\right\|^2 - \left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2\right) + \varepsilon_{\mathbf{y}}^{t+1}. \end{aligned}$$

Due to $L'_f := L_f + \mu_x + L$ smoothness of f_k, we know $f_k(\mathbf{x}^{t+1},\mathbf{y}^{t+1}) - f_k(\mathbf{x}^t,\mathbf{y}^{t+1}) - \left\langle\nabla_{\mathbf{x}} f_k(\mathbf{x}^t,\mathbf{y}^{t+1}),\, \mathbf{x}^{t+1}-\mathbf{x}^t\right\rangle \leq \frac{L'_f}{2}\left\|\mathbf{x}^t-\mathbf{x}\right\|^2$. Hence we have

$$\begin{aligned} f_k(\mathbf{x}^{t+1},\mathbf{y}^{t+1}) + h(\mathbf{x}^{t+1}) - &f_k(\mathbf{x},\mathbf{y}^{t+1}) + r(\mathbf{y}^{t+1}) - r(\mathbf{y}) - h(\mathbf{x}) \\ &\leq \frac{L'_f}{2}\left\|\mathbf{x}^t-\mathbf{x}\right\|^2 - \frac{\mu_x}{2}\left\|\mathbf{x}^t-\mathbf{x}\right\|^2 \\ &\quad + \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^t),\mathbf{y}^t) + (\nabla_{\mathbf{y}} f(g(\mathbf{x}^t),\mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}),\mathbf{y}^{t-1})),\, \mathbf{y}^{t+1}-\mathbf{y}\right\rangle \\ &\quad + \frac{1}{2\eta_{\mathbf{x}}}\left(\left\|\mathbf{x}-\mathbf{x}^t\right\|^2 - \left\|\mathbf{x}-\mathbf{x}^{t+1}\right\|^2 - \left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2\right) + \varepsilon_{\mathbf{x}}^{t+1} \\ &\quad + \frac{1}{2\eta_{\mathbf{y}}}\left(\left\|\mathbf{y}-\mathbf{y}^t\right\|^2 - \left\|\mathbf{y}-\mathbf{y}^{t+1}\right\|^2 - \left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2\right) + \varepsilon_{\mathbf{y}}^{t+1}. \end{aligned}$$

Adding $f_k(\mathbf{x}^{t+1}, \mathbf{y})$ on both sides yields:

$$\begin{aligned}
&\underbrace{f_k(\mathbf{x}^{t+1}, \mathbf{y}) + h(\mathbf{x}^{t+1}) - r(\mathbf{y}) - (h(\mathbf{x}) + f_k(\mathbf{x}, \mathbf{y}^{t+1}) - r(\mathbf{y}^{t+1}))}_{=F_k(\mathbf{x}^{t+1},\mathbf{y})-F_k(\mathbf{x},\mathbf{y}^{t+1})} \\
\leq & f_k(\mathbf{x}^{t+1}, \mathbf{y}) - f_k(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}) + \frac{L'_f}{2}\left\|\mathbf{x}^t - \mathbf{x}\right\|^2 - \frac{\mu_x}{2}\left\|\mathbf{x}^t - \mathbf{x}\right\|^2 \\
&+ \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t) + (\nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}), \mathbf{y}^{t-1})),\, \mathbf{y}^{t+1} - \mathbf{y} \right\rangle \\
&+ \frac{1}{2\eta_{\mathbf{x}}}\left(\left\|\mathbf{x} - \mathbf{x}^t\right\|^2 - \left\|\mathbf{x} - \mathbf{x}^{t+1}\right\|^2 - \left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2\right) + \varepsilon_{\mathbf{x}}^{t+1} \\
&+ \frac{1}{2\eta_{\mathbf{y}}}\left(\left\|\mathbf{y} - \mathbf{y}^t\right\|^2 - \left\|\mathbf{y} - \mathbf{y}^{t+1}\right\|^2 - \left\|\mathbf{y}^{t+1} - \mathbf{y}^t\right\|^2\right) + \varepsilon_{\mathbf{y}}^{t+1}.
\end{aligned}$$

Due to μ_y strong concavity of f_k w.r.t. $\mathbf{y}$, we have $f_k(\mathbf{x}^{t+1}, \mathbf{y}) - f_k(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}) \leq \left\langle \nabla_{\mathbf{y}} f_k(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}),\, \mathbf{y} - \mathbf{y}^{t+1} \right\rangle - \frac{\mu_y}{2}\left\|\mathbf{y} - \mathbf{y}^{t+1}\right\|^2$. So we have

$$\begin{aligned}
F_k(\mathbf{x}^{t+1}, \mathbf{y}) - F_k(\mathbf{x}, \mathbf{y}^{t+1}) \leq & \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t+1}), \mathbf{y}^{t+1}),\, \mathbf{y} - \mathbf{y}^{t+1} \right\rangle - \frac{\mu_y}{2}\left\|\mathbf{y} - \mathbf{y}^{t+1}\right\|^2 + \frac{L'_f}{2}\left\|\mathbf{x}^t - \mathbf{x}\right\|^2 - \frac{\mu_x}{2}\left\|\mathbf{x}^t - \mathbf{x}\right\|^2 \\
&+ \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t) + (\nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}), \mathbf{y}^{t-1})),\, \mathbf{y}^{t+1} - \mathbf{y} \right\rangle \\
&+ \frac{1}{2\eta_{\mathbf{x}}}\left(\left\|\mathbf{x} - \mathbf{x}^t\right\|^2 - \left\|\mathbf{x} - \mathbf{x}^{t+1}\right\|^2 - \left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2\right) + \varepsilon_{\mathbf{x}}^{t+1} \\
&+ \frac{1}{2\eta_{\mathbf{y}}}\left(\left\|\mathbf{y} - \mathbf{y}^t\right\|^2 - \left\|\mathbf{y} - \mathbf{y}^{t+1}\right\|^2 - \left\|\mathbf{y}^{t+1} - \mathbf{y}^t\right\|^2\right) + \varepsilon_{\mathbf{y}}^{t+1}.
\end{aligned}$$

Notice the following identity:

$$\begin{aligned}
&\left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t+1}), \mathbf{y}^{t+1}),\, \mathbf{y} - \mathbf{y}^{t+1} \right\rangle + \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t) + (\nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}), \mathbf{y}^{t-1})),\, \mathbf{y}^{t+1} - \mathbf{y} \right\rangle \\
=& \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t+1}), \mathbf{y}^{t+1}) + (\nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}), \mathbf{y}^{t-1})),\, \mathbf{y}^{t+1} - \mathbf{y} \right\rangle \\
=& - \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t+1}), \mathbf{y}^{t+1}) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t),\, \mathbf{y}^{t+1} - \mathbf{y} \right\rangle + \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}), \mathbf{y}^{t-1}),\, \mathbf{y}^t - \mathbf{y} \right\rangle \\
&+ \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}), \mathbf{y}^{t-1}),\, \mathbf{y}^{t+1} - \mathbf{y}^t \right\rangle.
\end{aligned}$$

We thus arrive at

$$\begin{aligned}
&F_k(\mathbf{x}^{t+1}, \mathbf{y}) - F_k(\mathbf{x}, \mathbf{y}^{t+1}) \\
&\qquad \leq - \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t+1}), \mathbf{y}^{t+1}) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t),\, \mathbf{y}^{t+1} - \mathbf{y} \right\rangle + \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}), \mathbf{y}^{t-1}),\, \mathbf{y}^t - \mathbf{y} \right\rangle \\
&\qquad + \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}), \mathbf{y}^{t-1}),\, \mathbf{y}^{t+1} - \mathbf{y}^t \right\rangle \\
&\qquad + \frac{L'_f}{2}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 + \frac{1}{2\eta_{\mathbf{y}}}\left(\left\|\mathbf{y} - \mathbf{y}^t\right\|^2 - \left\|\mathbf{y} - \mathbf{y}^{t+1}\right\|^2 - \left\|\mathbf{y}^{t+1} - \mathbf{y}^t\right\|^2\right) - \frac{\mu_y}{2}\left\|\mathbf{y} - \mathbf{y}^{t+1}\right\|^2 \\
&\qquad + \frac{1}{2\eta_{\mathbf{x}}}\left(\left\|\mathbf{x} - \mathbf{x}^t\right\|^2 - \left\|\mathbf{x} - \mathbf{x}^{t+1}\right\|^2 - \left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2\right) - \frac{\mu_x}{2}\left\|\mathbf{x} - \mathbf{x}^t\right\|^2 + \varepsilon_{\mathbf{x}}^{t+1} + \varepsilon_{\mathbf{y}}^{t+1}.
\end{aligned}$$

Now it remains to bound $\varepsilon_{\mathbf{x}}^{t+1}$ and $\varepsilon_{\mathbf{y}}^{t+1}$. For $\varepsilon_{\mathbf{x}}^{t+1}$:

$$\varepsilon_{\mathbf{x}}^{t+1} \leq \frac{1}{2}\left(\frac{1}{\mu_x}\left\|\mathbf{g}_{\mathbf{x}}^t - \nabla_{\mathbf{x}} f(g(\mathbf{x}^t), \mathbf{y}^{t+1})\right\|^2 + \mu_x\left\|\mathbf{x}^{t+1} - \mathbf{x}\right\|^2\right).$$

Notice that

$$\begin{aligned}
\left\|\mathbf{g}_{\mathbf{x}}^t - \nabla_{\mathbf{x}} f(g(\mathbf{x}^t), \mathbf{y}^{t+1})\right\|^2 &\leq 2\left\|\mathbf{g}_{\mathbf{x}}^t - \bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 + 2\left\|\bar{\mathbf{g}}_{\mathbf{x}}^t - \nabla_{\mathbf{x}} f(g(\mathbf{x}^t), \mathbf{y}^{t+1})\right\|^2 \\
&\leq 2\left\|\mathbf{g}_{\mathbf{x}}^t - \bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 + 2\left\|\nabla f(\mathbf{z}^{t+1}, \mathbf{y}^{t+1})\nabla g(\mathbf{x}^t) - \nabla f(g(\mathbf{x}^t), \mathbf{y}^{t+1})\nabla g(\mathbf{x}^t)\right\|^2 \\
&\leq 2\left\|\mathbf{g}_{\mathbf{x}}^t - \bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 + 2L_f^2 G_g^2\left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\right\|^2.
\end{aligned}$$

Hence we have

$$\varepsilon_{\mathbf{x}}^{t+1} \leq \frac{1}{\mu_x}\left\|\mathbf{g}_{\mathbf{x}}^t - \bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 + \frac{L_f^2 G_g^2}{\mu_x}\left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\right\|^2 + \frac{\mu_x}{2}\left\|\mathbf{x}^{t+1} - \mathbf{x}\right\|^2.$$

Similarly we have for $\varepsilon_{\mathbf{y}}^{t+1}$

$$\begin{aligned}
\varepsilon_{\mathbf{y}}^{t+1} &\leq \frac{1}{2}\left(\frac{1}{\mu_y}\left\|2\mathbf{q}^t - \mathbf{q}^{t-1} - \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t) - (\nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}), \mathbf{y}^{t-1}))\right\|^2 + \mu_y\left\|\mathbf{y}^{t+1} - \mathbf{y}\right\|^2\right)\\
&\leq \frac{\mu_y}{2}\left\|\mathbf{y}^{t+1} - \mathbf{y}\right\|^2 + \frac{4}{\mu_y}\left\|\mathbf{q}^t - \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t)\right\|^2 + \frac{1}{\mu_y}\left\|\mathbf{q}^{t-1} - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}), \mathbf{y}^{t-1})\right\|^2\\
&\leq \frac{\mu_y}{2}\left\|\mathbf{y}^{t+1} - \mathbf{y}\right\|^2 + \frac{8}{\mu_y}\left\|\mathbf{q}^t - \nabla_{\mathbf{y}} f(\mathbf{z}^{t+1}, \mathbf{y}^t)\right\|^2 + \frac{2}{\mu_y}\left\|\mathbf{q}^{t-1} - \nabla_{\mathbf{y}} f(\mathbf{z}^t, \mathbf{y}^{t-1})\right\|^2\\
&\quad + \frac{8}{\mu_y}L_f^2\left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\right\|^2 + \frac{2}{\mu_y}L_f^2\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\right\|^2.
\end{aligned}$$

Putting pieces together will conclude the proof.

$$\begin{aligned}
F_k(\mathbf{x}^{t+1}, \mathbf{y}) - F_k(\mathbf{x}, \mathbf{y}^{t+1})&\\
\leq &-\left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t+1}), \mathbf{y}^{t+1}) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t),\, \mathbf{y}^{t+1} - \mathbf{y}\right\rangle\\
&+ \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}), \mathbf{y}^{t-1}),\, \mathbf{y}^t - \mathbf{y}\right\rangle\\
&+ \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^t), \mathbf{y}^t) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}), \mathbf{y}^{t-1}),\, \mathbf{y}^{t+1} - \mathbf{y}^t\right\rangle\\
&+ \frac{L_f'}{2}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2 + \frac{1}{2\eta_{\mathbf{y}}}\left(\left\|\mathbf{y} - \mathbf{y}^t\right\|^2 - \left\|\mathbf{y} - \mathbf{y}^{t+1}\right\|^2 - \left\|\mathbf{y}^{t+1} - \mathbf{y}^t\right\|^2\right)\\
&+ \frac{1}{2\eta_{\mathbf{x}}}\left(\left\|\mathbf{x} - \mathbf{x}^t\right\|^2 - \left\|\mathbf{x} - \mathbf{x}^{t+1}\right\|^2 - \left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|^2\right) - \frac{\mu_x}{2}\left\|\mathbf{x} - \mathbf{x}^t\right\|^2 + \frac{\mu_x}{2}\left\|\mathbf{x}^{t+1} - \mathbf{x}\right\|^2\\
&+ \frac{1}{\mu_x}\left\|\mathbf{g}_{\mathbf{x}}^t - \bar{\mathbf{g}}_{\mathbf{x}}^t\right\|^2 + \left(\frac{L_f^2 G_g^2}{\mu_x} + \frac{8L_f^2}{\mu_y}\right)\left\|\mathbf{z}^{t+1} - g(\mathbf{x}^t)\right\|^2 + \frac{2L_f^2}{\mu_y}\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\right\|^2\\
&+ \frac{8}{\mu_y}\left\|\mathbf{q}^t - \nabla_{\mathbf{y}} f(\mathbf{z}^{t+1}, \mathbf{y}^t)\right\|^2 + \frac{2}{\mu_y}\left\|\mathbf{q}^{t-1} - \nabla_{\mathbf{y}} f(\mathbf{z}^t, \mathbf{y}^{t-1})\right\|^2.
\end{aligned}$$

□

Lemma 25 (Primal-dual gap convergence)**.** *For Algorithm 5, under assumptions of Theorem 6, the following statement holds:*

$$\begin{aligned}
\mathbb{E}[F_k(\mathbf{x}_{k+1}, \mathbf{y}^*(\mathbf{x}_{k+1})) - F_k(\mathbf{x}^*(\mathbf{y}_{k+1}), \mathbf{y}_{k+1})] \leq & \frac{\mathbb{E}\left\|\mathbf{y}_k^*(\mathbf{x}_{k+1}) - \mathbf{y}_k\right\|^2}{2\eta_{\mathbf{y}} T} + \frac{\mathbb{E}\left\|\mathbf{x}_k^*(\mathbf{y}_{k+1}) - \mathbf{x}_k\right\|^2}{4\eta_{\mathbf{x}} T} + \frac{C\mathbb{E}\left\|\mathbf{z}^0 - g(\mathbf{x}^0)\right\|^2}{T}\\
&+ \frac{\tilde{\sigma}^2}{\mu_x B_\tau} + \frac{10\tilde{\sigma}^2}{\mu_y B_\tau} + \left(\frac{24G_g^2 L_f^2 \tau}{\mu_x B} + \frac{120\tau}{\mu_y B} + \frac{20L_f^2}{\mu_y} + \frac{2L_f^2 G_g^2}{\mu_x}\right)\cdot 9\frac{\tilde{\sigma}^2}{B_\tau},
\end{aligned}$$

where $C = \frac{12G_g^2 L^2 \tau}{\mu_x B} + \frac{60\tau}{\mu_y B} + \frac{12L_f^2}{\mu_y} + \frac{L_f^2 G_g^2}{\mu_x}$, $\mathbf{x}_k^*(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathcal{X}} F_k(\mathbf{x}, \mathbf{y})$ *and* $\mathbf{y}_k^*(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathcal{Y}} F_k(\mathbf{x}, \mathbf{y})$.

Proof. Due to Lemma 24 we have

$$
\begin{aligned}
F_k(\mathbf{x}^{t+1},\mathbf{y}) - F_k(\mathbf{x},\mathbf{y}^{t+1})& \\
\leq & - \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t+1}),\mathbf{y}^{t+1}) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t}),\mathbf{y}^{t}),\, \mathbf{y}^{t+1}-\mathbf{y}\right\rangle \\
& + \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t}),\mathbf{y}^{t}) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}),\mathbf{y}^{t-1}),\, \mathbf{y}^{t}-\mathbf{y}\right\rangle \\
& + \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t}),\mathbf{y}^{t}) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}),\mathbf{y}^{t-1}),\, \mathbf{y}^{t+1}-\mathbf{y}^{t}\right\rangle \\
& + \frac{L_f'}{2}\left\|\mathbf{x}^{t+1}-\mathbf{x}^{t}\right\|^2 + \frac{1}{2\eta_{\mathbf{y}}}\left(\left\|\mathbf{y}-\mathbf{y}^{t}\right\|^2 - \left\|\mathbf{y}-\mathbf{y}^{t+1}\right\|^2 - \left\|\mathbf{y}^{t+1}-\mathbf{y}^{t}\right\|^2\right) \\
& + \frac{1}{2\eta_{\mathbf{x}}}\left(\left\|\mathbf{x}-\mathbf{x}^{t}\right\|^2 - \left\|\mathbf{x}-\mathbf{x}^{t+1}\right\|^2 - \left\|\mathbf{x}^{t+1}-\mathbf{x}^{t}\right\|^2\right) - \frac{\mu_x}{2}\left\|\mathbf{x}-\mathbf{x}^{t}\right\|^2 + \frac{\mu_x}{2}\left\|\mathbf{x}^{t+1}-\mathbf{x}\right\|^2 \\
& + \frac{1}{\mu_x}\left\|\mathbf{g}_{\mathbf{x}}^{t}-\bar{\mathbf{g}}_{\mathbf{x}}^{t}\right\|^2 + \left(\frac{L_f^2G_g^2}{\mu_x}+\frac{8L_f^2}{\mu_y}\right)\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^{t})\right\|^2 + \frac{2L_f^2}{\mu_y}\left\|\mathbf{z}^{t}-g(\mathbf{x}^{t-1})\right\|^2 \\
& + \frac{8}{\mu_y}\left\|\mathbf{q}^{t}-\nabla_{\mathbf{y}} f(\mathbf{z}^{t+1},\mathbf{y}^{t})\right\|^2 + \frac{2}{\mu_y}\left\|\mathbf{q}^{t-1}-\nabla_{\mathbf{y}} f(\mathbf{z}^{t},\mathbf{y}^{t-1})\right\|^2.
\end{aligned}
$$

Define $\mathcal{F}_k^{t+1} = F_k(\mathbf{x}^{t+1},\mathbf{y}) - F_k(\mathbf{x},\mathbf{y}^{t+1}) + C\left(\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^{t})\right\|^2 - \mathbb{E}\left\|\mathbf{z}^{t}-g(\mathbf{x}^{t-1})\right\|^2\right)$, and $\Gamma^t = \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t}),\mathbf{y}^{t}) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}),\mathbf{y}^{t-1}),\, \mathbf{y}^{t}-\mathbf{y}\right\rangle$. Adding $C\left(\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^{t})\right\|^2 - \mathbb{E}\left\|\mathbf{z}^{t}-g(\mathbf{x}^{t-1})\right\|^2\right)$ on both side yields:

$$
\begin{aligned}
\mathcal{F}_k^{t+1} \leq & -\Gamma^{t+1} + \Gamma^{t} + \left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t}),\mathbf{y}^{t}) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}),\mathbf{y}^{t-1}),\, \mathbf{y}^{t+1}-\mathbf{y}^{t}\right\rangle \\
& + \frac{L_f'}{2}\left\|\mathbf{x}^{t+1}-\mathbf{x}^{t}\right\|^2 + \frac{1}{2\eta_{\mathbf{y}}}\left(\left\|\mathbf{y}-\mathbf{y}^{t}\right\|^2 - \left\|\mathbf{y}-\mathbf{y}^{t+1}\right\|^2 - \left\|\mathbf{y}^{t+1}-\mathbf{y}^{t}\right\|^2\right) \\
& + \frac{1}{2\eta_{\mathbf{x}}}\left(\left\|\mathbf{x}-\mathbf{x}^{t}\right\|^2 - \left\|\mathbf{x}-\mathbf{x}^{t+1}\right\|^2 - \left\|\mathbf{x}^{t+1}-\mathbf{x}^{t}\right\|^2\right) - \frac{\mu_x}{2}\left\|\mathbf{x}-\mathbf{x}^{t}\right\|^2 + \frac{\mu_x}{2}\left\|\mathbf{x}^{t+1}-\mathbf{x}\right\|^2 \\
& + \frac{1}{\mu_x}\left\|\mathbf{g}_{\mathbf{x}}^{t}-\bar{\mathbf{g}}_{\mathbf{x}}^{t}\right\|^2 + \left(\frac{L_f^2G_g^2}{\mu_x}+\frac{8L_f^2}{\mu_y}\right)\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^{t})\right\|^2 + \frac{2L_f^2}{\mu_y}\left\|\mathbf{z}^{t}-g(\mathbf{x}^{t-1})\right\|^2 \\
& + \frac{8}{\mu_y}\left\|\mathbf{q}^{t}-\nabla_{\mathbf{y}} f(\mathbf{z}^{t+1},\mathbf{y}^{t})\right\|^2 + \frac{2}{\mu_y}\left\|\mathbf{q}^{t-1}-\nabla_{\mathbf{y}} f(\mathbf{z}^{t},\mathbf{y}^{t-1})\right\|^2 \\
& + C\left(\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^{t})\right\|^2 - \mathbb{E}\left\|\mathbf{z}^{t}-g(\mathbf{x}^{t-1})\right\|^2\right).
\end{aligned}
$$

Notice that by Cauchy-schwartz and AM-GM inequality we have:

$$
\begin{aligned}
\left\langle \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t}),\mathbf{y}^{t}) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}),\mathbf{y}^{t-1}),\, \mathbf{y}^{t+1}-\mathbf{y}^{t}\right\rangle &\leq \eta_{\mathbf{y}}\left\|\nabla_{\mathbf{y}} f(g(\mathbf{x}^{t}),\mathbf{y}^{t}) - \nabla_{\mathbf{y}} f(g(\mathbf{x}^{t-1}),\mathbf{y}^{t-1})\right\|^2 + \frac{1}{4\eta_{\mathbf{y}}}\left\|\mathbf{y}^{t+1}-\mathbf{y}^{t}\right\|^2 \\
&\leq \eta_{\mathbf{y}}L_f^2\left\|(g(\mathbf{x}^{t}),\mathbf{y}^{t}) - (g(\mathbf{x}^{t-1}),\mathbf{y}^{t-1})\right\|^2 + \frac{1}{4\eta_{\mathbf{y}}}\left\|\mathbf{y}^{t+1}-\mathbf{y}^{t}\right\|^2 \\
&\leq \eta_{\mathbf{y}}L_f^2\left(G_g^2\left\|\mathbf{x}^{t-1}-\mathbf{x}^{t}\right\|^2 + \left\|\mathbf{y}^{t-1}-\mathbf{y}^{t}\right\|^2\right) + \frac{1}{4\eta_{\mathbf{y}}}\left\|\mathbf{y}^{t+1}-\mathbf{y}^{t}\right\|^2.
\end{aligned}
$$

Plugging back yields:

$$
\begin{aligned}
\mathcal{F}_k^{t+1} \leq & -\Gamma^{t+1} + \Gamma^{t} + \frac{1}{2\eta_{\mathbf{y}}}\left(\left\|\mathbf{y}-\mathbf{y}^{t}\right\|^2 - \left\|\mathbf{y}-\mathbf{y}^{t+1}\right\|^2\right) - \frac{1}{4\eta_{\mathbf{y}}}\left\|\mathbf{y}^{t+1}-\mathbf{y}^{t}\right\|^2 + \eta_{\mathbf{y}}L_f^2\left\|\mathbf{y}^{t-1}-\mathbf{y}^{t}\right\|^2 \\
& + \left(\frac{1}{2\eta_{\mathbf{x}}}-\frac{\mu_x}{2}\right)\left(\left\|\mathbf{x}-\mathbf{x}^{t}\right\|^2 - \left\|\mathbf{x}-\mathbf{x}^{t+1}\right\|^2\right) - \left(\frac{1}{2\eta_{\mathbf{x}}}-\frac{L_f'}{2}\right)\left\|\mathbf{x}^{t+1}-\mathbf{x}^{t}\right\|^2 + \eta_{\mathbf{y}}L_f^2G_g^2\left\|\mathbf{x}^{t-1}-\mathbf{x}^{t}\right\|^2 \\
& + \frac{1}{\mu_x}\left\|\mathbf{g}_{\mathbf{x}}^{t}-\bar{\mathbf{g}}_{\mathbf{x}}^{t}\right\|^2 + \frac{8}{\mu_y}\left\|\mathbf{q}^{t}-\bar{\mathbf{g}}_{\mathbf{y}}^{t}\right\|^2 + \frac{2}{\mu_y}\left\|\mathbf{q}^{t-1}-\bar{\mathbf{g}}_{\mathbf{y}}^{t-1}\right\|^2 \\
& + \left(\frac{8L_f^2}{\mu_y}+\frac{L_f^2G_g^2}{\mu_x}\right)\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^{t})\right\|^2 + \frac{2L_f^2}{\mu_y}\left\|\mathbf{z}^{t}-g(\mathbf{x}^{t-1})\right\|^2 \\
& + C\left(\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^{t})\right\|^2 - \mathbb{E}\left\|\mathbf{z}^{t}-g(\mathbf{x}^{t-1})\right\|^2\right).
\end{aligned}
$$

Plugging in the bound from Lemma 23 for $\|\mathbf{g}_\mathbf{x}^t - \bar{\mathbf{g}}_\mathbf{x}^t\|^2$ and $\left\|\mathbf{q}^t - \bar{\mathbf{g}}_\mathbf{y}^t\right\|^2$ yields:

$$\begin{aligned}
\mathcal{F}_k^{t+1} \leq & -\Gamma^{t+1} + \Gamma^t + \frac{1}{2\eta_\mathbf{y}}\left(\left\|\mathbf{y}-\mathbf{y}^t\right\|^2 - \left\|\mathbf{y}-\mathbf{y}^{t+1}\right\|^2\right) - \frac{1}{4\eta_\mathbf{y}}\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2 + \eta_\mathbf{y} L_f^2\left\|\mathbf{y}^{t-1}-\mathbf{y}^t\right\|^2 \\
& + \left(\frac{1}{2\eta_\mathbf{x}} - \frac{\mu_x}{2}\right)\left(\left\|\mathbf{x}-\mathbf{x}^t\right\|^2 - \left\|\mathbf{x}-\mathbf{x}^{t+1}\right\|^2\right) - \left(\frac{1}{2\eta_\mathbf{x}} - \frac{L_f'}{2}\right)\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2 + \eta_\mathbf{y} L_f^2 G_g^2\left\|\mathbf{x}^{t-1}-\mathbf{x}^t\right\|^2 \\
& + \frac{\tilde{\sigma}^2}{\mu_x B_\tau} + \frac{1}{\mu_x B}\sum_{t'=t_p}^{t-1}\left(2G_g^2 L_f^2\left\|\mathbf{z}^{t'+1}-\mathbf{z}^{t'+2}\right\|^2 + 2G_g^2 L_f^2\left\|\mathbf{y}^{t'+1}-\mathbf{y}^{t'+2}\right\|^2 + 2G_f^2 L_g^2\left\|\mathbf{x}^{t'}-\mathbf{x}^{t'+1}\right\|^2\right) \\
& + \frac{10\tilde{\sigma}^2}{\mu_y B_\tau} + \frac{8}{\mu_y B}\sum_{t'=t_p}^{t-1} L_f^2\left(\mathbb{E}\left\|\mathbf{z}^{t'}-\mathbf{z}^{t'+1}\right\|^2 + \mathbb{E}\left\|\mathbf{y}^{t'}-\mathbf{y}^{t'+1}\right\|^2\right) \\
& + \frac{2}{\mu_y B}\sum_{t'=t_p}^{t-2} L_f^2\left(\mathbb{E}\left\|\mathbf{z}^{t'}-\mathbf{z}^{t'+1}\right\|^2 + \mathbb{E}\left\|\mathbf{y}^{t'}-\mathbf{y}^{t'+1}\right\|^2\right) \\
& + \left(\frac{8L_f^2}{\mu_y} + \frac{L_f^2 G_g^2}{\mu_x}\right)\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t)\right\|^2 + \frac{2L_f^2}{\mu_y}\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\right\|^2 + C\left(\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t)\right\|^2 - \mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\right\|^2\right).
\end{aligned}$$

Since $\sum_{t'=t_p}^{t-2} L^2\left(\mathbb{E}\left\|\mathbf{z}^{t'}-\mathbf{z}^{t'+1}\right\|^2 + \mathbb{E}\left\|\mathbf{y}^{t'}-\mathbf{y}^{t'+1}\right\|^2\right) \leq \sum_{t'=t_p}^{t-1} L^2\left(\mathbb{E}\left\|\mathbf{z}^{t'}-\mathbf{z}^{t'+1}\right\|^2 + \mathbb{E}\left\|\mathbf{y}^{t'}-\mathbf{y}^{t'+1}\right\|^2\right)$, and $\left\|\mathbf{z}^t-\mathbf{z}^{t+1}\right\|^2 \leq 3\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\right\|^2 + 3G_g^2\left\|\mathbf{x}^{t-1}-\mathbf{x}^t\right\|^2 + 3\left\|g(\mathbf{x}^t)-\mathbf{z}^{t+1}\right\|^2$, we have

$$\begin{aligned}
\mathcal{F}_k^{t+1} \leq & -\Gamma^{t+1} + \Gamma^t + \frac{1}{2\eta_\mathbf{y}}\left(\left\|\mathbf{y}-\mathbf{y}^t\right\|^2 - \left\|\mathbf{y}-\mathbf{y}^{t+1}\right\|^2\right) - \frac{1}{4\eta_\mathbf{y}}\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2 + \eta_\mathbf{y} L_f^2\left\|\mathbf{y}^{t-1}-\mathbf{y}^t\right\|^2 \\
& + \left(\frac{1}{2\eta_\mathbf{x}} - \frac{\mu_x}{2}\right)\left(\left\|\mathbf{x}-\mathbf{x}^t\right\|^2 - \left\|\mathbf{x}-\mathbf{x}^{t+1}\right\|^2\right) - \left(\frac{1}{2\eta_\mathbf{x}} - \frac{L_f'}{2}\right)\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2 + \eta_\mathbf{y} L_f^2 G_g^2\left\|\mathbf{x}^{t-1}-\mathbf{x}^t\right\|^2 \\
& + \frac{\tilde{\sigma}^2}{\mu_x B_\tau} + \frac{2G_g^2 L_f^2}{\mu_x B}\sum_{t'=t_p}^{t-1}\left(3\left\|\mathbf{z}^{t'+1}-g(\mathbf{x}^{t'})\right\|^2 + 3G_g^2\left\|\mathbf{x}^{t'}-\mathbf{x}^{t'+1}\right\|^2 + 3\left\|g(\mathbf{x}^{t'+1})-\mathbf{z}^{t'+2}\right\|^2\right) \\
& + \frac{1}{\mu_x B}\sum_{t'=t_p}^{t-1}\left(2G_g^2 L_f^2\left\|\mathbf{y}^{t'+1}-\mathbf{y}^{t'+2}\right\|^2 + 2G_f^2 L_g^2\left\|\mathbf{x}^{t'}-\mathbf{x}^{t'+1}\right\|^2\right) \\
& + \frac{10\tilde{\sigma}^2}{\mu_y B_\tau} + \frac{10}{\mu_y B}\sum_{t'=t_p}^{t-1} L_f^2\left(3\left\|\mathbf{z}^{t'}-g(\mathbf{x}^{t'-1})\right\|^2 + 3G_g^2\left\|\mathbf{x}^{t'-1}-\mathbf{x}^{t'}\right\|^2 + 3\left\|g(\mathbf{x}^{t'})-\mathbf{z}^{t'+1}\right\|^2 + \mathbb{E}\left\|\mathbf{y}^{t'}-\mathbf{y}^{t'+1}\right\|^2\right) \\
& + \left(\frac{8L_f^2}{\mu_y} + \frac{L_f^2 G^2}{\mu_x}\right)\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t)\right\|^2 + \frac{2L_f^2}{\mu_y}\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\right\|^2 + C\left(\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t)\right\|^2 - \mathbb{E}\left\|\mathbf{z}^t - g(\mathbf{x}^{t-1})\right\|^2\right).
\end{aligned}$$

Summing from $t = 0$ to $T - 1$ yields

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{F}_k^{t+1} \leq{}& \Gamma^0 + \frac{1}{2\eta_{\mathbf{y}}T}\left(\left\|\mathbf{y}-\mathbf{y}^0\right\|^2 - \left\|\mathbf{y}-\mathbf{y}^T\right\|^2\right) + \left(\frac{1}{2\eta_{\mathbf{x}}T} - \frac{\mu_x}{2T}\right)\left(\left\|\mathbf{x}-\mathbf{x}^0\right\|^2 - \left\|\mathbf{x}-\mathbf{x}^T\right\|^2\right) \\
&- \frac{1}{4\eta_{\mathbf{y}}}\frac{1}{T}\sum_{t=0}^{T-1}\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2 + \frac{1}{T}\sum_{t=0}^{T-1}\eta_{\mathbf{y}}L_f^2\left\|\mathbf{y}^{t-1}-\mathbf{y}^t\right\|^2 \\
&- \left(\frac{1}{2\eta_{\mathbf{x}}} - \frac{L_f'}{2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2 + \eta_{\mathbf{y}}L_f^2G_g^2\frac{1}{T}\sum_{t=0}^{T-1}\left\|\mathbf{x}^{t-1}-\mathbf{x}^t\right\|^2 + \frac{\tilde{\sigma}^2}{\mu_x B_\tau} + \frac{10\tilde{\sigma}^2}{\mu_y B_\tau} \\
&+ \frac{2G_g^2L^2}{\mu_x B}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{t'=c(t)}^{t-1}\left(3\left\|\mathbf{z}^{t'+1}-g(\mathbf{x}^{t'})\right\|^2 + 3G_g^2\left\|\mathbf{x}^{t'}-\mathbf{x}^{t'+1}\right\|^2 + 3\left\|g(\mathbf{x}^{t'+1})-\mathbf{z}^{t'+2}\right\|^2\right) \\
&+ \frac{1}{\mu_x B}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{t'=c(t)}^{t-1}\left(2G_g^2L^2\left\|\mathbf{y}^{t'+1}-\mathbf{y}^{t'+2}\right\|^2 + 2G_f^2L_g^2\left\|\mathbf{x}^{t'}-\mathbf{x}^{t'+1}\right\|^2\right) \\
&+ \frac{10}{\mu_y B}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{t'=c(t)}^{t-1}L_f^2\left(3\left\|\mathbf{z}^{t'}-g(\mathbf{x}^{t'-1})\right\|^2 + 3G_g^2\left\|\mathbf{x}^{t'-1}-\mathbf{x}^{t'}\right\|^2 + 3\left\|g(\mathbf{x}^{t'})-\mathbf{z}^{t'+1}\right\|^2 + \mathbb{E}\left\|\mathbf{y}^{t'}-\mathbf{y}^{t'+1}\right\|^2\right) \\
&+ \left(\frac{8L_f^2}{\mu_y} + \frac{L_f^2G_g^2}{\mu_x}\right)\frac{1}{T}\sum_{t=0}^{T-1}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t)\right\|^2 + \frac{2L_f^2}{\mu_y}\frac{1}{T}\sum_{t=0}^{T-1}\left\|\mathbf{z}^t-g(\mathbf{x}^{t-1})\right\|^2 \\
&+ C\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t)\right\|^2 - \mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{x}^{t-1})\right\|^2\right).
\end{aligned}
$$

Re-arranging terms yields:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{F}_k^{t+1} \leq{}& \Gamma^0 + \frac{1}{2\eta_{\mathbf{y}}T}\left(\left\|\mathbf{y}-\mathbf{y}^0\right\|^2 - \left\|\mathbf{y}-\mathbf{y}^T\right\|^2\right) + \left(\frac{1}{2\eta_{\mathbf{x}}T} - \frac{\mu_x}{2T}\right)\left(\left\|\mathbf{x}-\mathbf{x}^0\right\|^2 - \left\|\mathbf{x}-\mathbf{x}^T\right\|^2\right) + \frac{\tilde{\sigma}^2}{\mu_x B_\tau} + \frac{10\tilde{\sigma}^2}{\mu_y B_\tau} \\
&- \left(\frac{1}{4\eta_{\mathbf{y}}} - \eta_{\mathbf{y}}L_f^2 - \frac{2G_g^2L_f^2\tau}{\mu_x B} - \frac{10\tau L^2}{\mu_y B}\right)\frac{1}{T}\sum_{t=0}^{T-1}\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2 \\
&- \left(\frac{1}{2\eta_{\mathbf{x}}} - \frac{L_f'}{2} - \eta_{\mathbf{y}}L_f^2G_g^2 - \frac{6G_g^4L_f^2\tau}{\mu_x B} - \frac{2G_f^2L_g^2\tau}{\mu_x B} - \frac{30G_g^2L_f^2\tau}{\mu_y B}\right)\frac{1}{T}\sum_{t=0}^{T-1}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2 \\
&+ \left(\frac{12G_g^2L_f^2\tau}{\mu_x B} + \frac{60\tau}{\mu_y B} + \frac{8L_f^2}{\mu_y} + \frac{L_f^2G_g^2}{\mu_x} + C\right)\frac{1}{T}\sum_{t=0}^{T-1}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t)\right\|^2 \\
&- \left(C - \frac{2L_f^2}{\mu_y}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{z}^t-g(\mathbf{x}^{t-1})\right\|^2.
\end{aligned}
$$

Plugging the bound for $\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t)\right\|^2$ yields:

$$
\begin{aligned}
&\left(\frac{12G_g^2L^2\tau}{\mu_x B}+\frac{60\tau}{\mu_y B}+\frac{8L_f^2}{\mu_y}+\frac{L_f^2G_g^2}{\mu_x}+C\right)\frac{1}{T}\sum_{t=0}^{T-1}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t)\right\|^2-\left(C-\frac{2L^2}{\mu_y}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{z}^{t}-g(\mathbf{x}^{t-1})\right\|^2\\
&\leq\left(\frac{12G_g^2L^2\tau}{\mu_x B}+\frac{60\tau}{\mu_y B}+\frac{8L_f^2}{\mu_y}+\frac{L_f^2G_g^2}{\mu_x}+C\right)\frac{1}{T}\sum_{t=0}^{T-1}\\
&\quad\times\left((1-\beta)\,\mathbb{E}\left\|\mathbf{z}^{t}-g(\mathbf{x}^{t-1})\right\|^2+\left(1+\frac{1}{\beta}\right)\left(6(1-\beta)^2G_g^2\mathbb{E}\left\|\mathbf{x}^{t-1}-\mathbf{x}^t\right\|^2+3\beta^2\frac{\sigma^2}{B_\tau}+\frac{G_g^2}{B}\sum_{t'=t_p}^{t-1}\mathbb{E}\left\|\mathbf{x}^{t'}-\mathbf{x}^{t'-1}\right\|^2\right)\right)\\
&\quad-\left(C-\frac{2L^2}{\mu_y}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{z}^{t}-g(\mathbf{x}^{t-1})\right\|^2\\
&\leq\left(\frac{12G_g^2L^2\tau}{\mu_x B}+\frac{60\tau}{\mu_y B}+\frac{8L_f^2}{\mu_y}+\frac{L_f^2G_g^2}{\mu_x}+C\right)\frac{1}{T}\sum_{t=0}^{T-1}(1-\beta)\,\mathbb{E}\left\|\mathbf{z}^{t}-g(\mathbf{x}^{t-1})\right\|^2\\
&\quad\left(\frac{12G_g^2L^2\tau}{\mu_x B}+\frac{60\tau}{\mu_y B}+\frac{8L_f^2}{\mu_y}+\frac{L_f^2G_g^2}{\mu_x}+C\right)\left(1+\frac{1}{\beta}\right)\left(6G_g^2+\frac{\tau G_g^2}{B}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{x}^{t}-\mathbf{x}^{t-1}\right\|^2\\
&\quad+\left(\frac{12G_g^2L^2\tau}{\mu_x B}+\frac{60\tau}{\mu_y B}+\frac{8L_f^2}{\mu_y}+\frac{L_f^2G_g^2}{\mu_x}+C\right)\left(1+\frac{1}{\beta}\right)3\frac{\sigma^2}{B_\tau}-\left(C-\frac{2L_f^2}{\mu_y}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{z}^{t}-g(\mathbf{x}^{t-1})\right\|^2.
\end{aligned}
$$

We choose $\beta=\frac{1}{2}$, and $C=\frac{12G_g^2L^2\tau}{\mu_x B}+\frac{60\tau}{\mu_y B}+\frac{12L_f^2}{\mu_y}+\frac{L_f^2G_g^2}{\mu_x}$, so we have:

$$
\left(\frac{12G_g^2L^2\tau}{\mu_x B}+\frac{60\tau}{\mu_y B}+\frac{8L_f^2}{\mu_y}+\frac{L_f^2G_g^2}{\mu_x}+C\right)(1-\beta)-\left(C-\frac{2L_f^2}{\mu_y}\right)=0.
$$

Define $\Delta_{\mathbf{x}}=\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2$ and $\Delta_{\mathbf{y}}=\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{y}^{t+1}-\mathbf{y}^t\right\|^2$, and we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{F}_k^{t+1}\leq\Gamma^0+\frac{1}{2\eta_{\mathbf{y}}T}\left(\left\|\mathbf{y}-\mathbf{y}^0\right\|^2-\left\|\mathbf{y}-\mathbf{y}^T\right\|^2\right)+\left(\frac{1}{2\eta_{\mathbf{x}}T}-\frac{\mu_x}{2T}\right)\left(\left\|\mathbf{x}-\mathbf{x}^0\right\|^2-\left\|\mathbf{x}-\mathbf{x}^T\right\|^2\right)+\frac{\tilde{\sigma}^2}{\mu_x B_\tau}+\frac{10\tilde{\sigma}^2}{\mu_y B_\tau}\\
&-\left(\frac{1}{4\eta_{\mathbf{y}}}-\eta_{\mathbf{y}}L_f^2-\frac{2G_g^2L_f^2\tau}{\mu_x B}-\frac{10\tau L_f^2}{\mu_y B}\right)\Delta_{\mathbf{y}}\\
&-\left(\frac{1}{2\eta_{\mathbf{x}}}-\frac{L_f'}{2}-\eta_{\mathbf{y}}L_f^2G_g^2-\frac{6G_g^4L^2\tau}{\mu_x B}-\frac{2G_f^2L_g^2\tau}{\mu_x B}-\frac{30G_g^2L_f^2\tau}{\mu_y B}-\left(\frac{72G_g^2L_f^2\tau}{\mu_x B}+\frac{360\tau}{\mu_y B}+\frac{60L_f^2}{\mu_y}+\frac{6L_f^2G_g^2}{\mu_x}\right)(6+\frac{\tau}{B})G_g^2\right)\Delta_{\mathbf{x}}\\
&+\left(\frac{24G_g^2L^2\tau}{\mu_x B}+\frac{120\tau}{\mu_y B}+\frac{20L_f^2}{\mu_y}+\frac{2L_f^2G_g^2}{\mu_x}\right)9\frac{\tilde{\sigma}^2}{B_\tau}.
\end{aligned}
$$

We choose $\eta_{\mathbf{y}}\leq\Theta\left(\min\left\{\frac{1}{L_f}\right\},\frac{\mu_y B}{\tau G_g^2L_f^2}\right)$ and $\eta_{\mathbf{x}}\leq\Theta\left(\min\left\{\frac{\mu_y B}{G_g^2L_f^2\tau},\frac{1}{L_f'},\frac{\rho B^2}{G_g^4L_f^2\tau^2},\frac{\mu_y B^2}{\tau^2}\right\}\right)$ which yields:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{F}_k^{t+1}\leq&\Gamma^0+\frac{1}{2\eta_{\mathbf{y}}T}\left(\left\|\mathbf{y}-\mathbf{y}^0\right\|^2-\left\|\mathbf{y}-\mathbf{y}^T\right\|^2\right)+\left(\frac{1}{2\eta_{\mathbf{x}}T}-\frac{\mu_x}{2T}\right)\left(\left\|\mathbf{x}-\mathbf{x}^0\right\|^2-\left\|\mathbf{x}-\mathbf{x}^T\right\|^2\right)\\
&+\frac{\tilde{\sigma}^2}{\mu_x B_\tau}+\frac{10\tilde{\sigma}^2}{\mu_y B_\tau}+\left(\frac{24G_g^2L^2\tau}{\mu_x B}+\frac{120\tau}{\mu_y B}+\frac{20L_f^2}{\mu_y}+\frac{2L_f^2G_g^2}{\mu_x}\right)9\frac{\tilde{\sigma}^2}{B_\tau}.
\end{aligned}
$$

Notice that

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{F}_k^{t+1} &= \frac{1}{T}\sum_{t=0}^{T-1}F_k(\mathbf{x}^{t+1},\mathbf{x})-F_k(\mathbf{x},\mathbf{y}^{t+1})+C(\mathbb{E}\left\|\mathbf{z}^{t+1}-g(\mathbf{x}^t)\right\|^2-\mathbb{E}\left\|\mathbf{z}^{t}-g(\mathbf{x}^{t-1})\right\|^2)\\
&= \frac{1}{T}\sum_{t=0}^{T-1}F_k(\mathbf{x}^{t+1},\mathbf{x})-F_k(\mathbf{x},\mathbf{y}^{t+1})+\frac{1}{T}C(\mathbb{E}\left\|\mathbf{z}^{T}-g(\mathbf{x}^{T-1})\right\|^2-\mathbb{E}\left\|\mathbf{z}^{0}-g(\mathbf{x}^{-1})\right\|^2)\\
&\geq \frac{1}{T}\sum_{t=0}^{T-1}F_k(\mathbf{x}^{t+1},\mathbf{x})-F_k(\mathbf{x},\mathbf{y}^{t+1})-\frac{1}{T}C\mathbb{E}\left\|\mathbf{z}^{0}-g(\mathbf{x}^{-1})\right\|^2.
\end{aligned}
$$

Since we choose $\eta_{\mathbf{x}}\leq\frac{1}{2L}$ and by convention $\mathbf{x}^{-1}=\mathbf{x}^0$, we have

$$
\begin{aligned}
F_k(\mathbf{x}_{k+1},\mathbf{y})-F_k(\mathbf{x},\mathbf{y}_{k+1}) \leq&\frac{1}{T}\sum_{t=0}^{T-1}F_k(\mathbf{x}^{t+1},\mathbf{y})-F_k(\mathbf{x},\mathbf{y}^{t+1})+O\left(\frac{C\mathbb{E}\left\|\mathbf{z}^{0}-g(\mathbf{x}^{0})\right\|^2}{T}\right)\\
\leq&\frac{\mathbb{E}\left\|\mathbf{y}-\mathbf{y}^0\right\|^2}{2\eta_{\mathbf{y}}T}+\frac{\mathbb{E}\left\|\mathbf{x}-\mathbf{x}^0\right\|^2}{4\eta_{\mathbf{x}}T}+O\left(\frac{C\mathbb{E}\left\|\mathbf{z}^{0}-g(\mathbf{x}^{0})\right\|^2}{T}\right)+\frac{\tilde{\sigma}^2}{\mu_xB_\tau}+\frac{10\tilde{\sigma}^2}{\mu_yB_\tau}\\
&+\left(\frac{24G_g^2L_f^2\tau}{\mu_xB}+\frac{120\tau}{\mu_yB}+\frac{20L_f^2}{\mu_y}+\frac{2L_f^2G_g^2}{\mu_x}\right)9\frac{\tilde{\sigma}^2}{B_\tau}.
\end{aligned}
$$

□

Lemma 26. *For Algorithm 5, under assumptions of Theorem 6, the following statement holds:*

$$
\begin{aligned}
&(Q-1)\mathrm{Gap}^k(\mathbf{x}_{k+1},\mathbf{y}_{k+1})\leq\mathrm{Gap}^k(\mathbf{x}_k,\mathbf{y}_k)\\
&\quad+Q\left(\frac{\tilde{\sigma}^2}{\mu_xB_\tau}+\frac{10\tilde{\sigma}^2}{\mu_yB_\tau}+O\left(\frac{C\mathbb{E}\left\|\mathbf{z}^{0}-g(\mathbf{x}^{0})\right\|^2}{T}\right)+\left(\frac{24G_g^2L_f^2\tau}{\mu_xB}+\frac{120\tau}{\mu_yB}+\frac{20L_f^2}{\mu_y}+\frac{2L_f^2G_g^2}{\mu_x}\right)9\frac{\tilde{\sigma}^2}{B_\tau}\right),
\end{aligned}
$$

where $Q=\frac{\min\left\{\frac{\mu_x}{4},\frac{\mu_y}{4}\right\}}{\max\left\{\frac{1}{2\eta_{\mathbf{y}}T},\frac{1}{4\eta_{\mathbf{x}}T}\right\}}$.

Proof. Due to [Lemma. 1 Yan et al. [2020]] we have

$$
\frac{\mu_x}{4}\left\|\mathbf{x}_k^*(\mathbf{y})-\mathbf{x}'\right\|^2+\frac{\mu_y}{4}\left\|\mathbf{y}_k^*(\mathbf{x})-\mathbf{y}'\right\|^2\leq\mathrm{Gap}^k(\mathbf{x},\mathbf{y})+\mathrm{Gap}^k(\mathbf{x}',\mathbf{y}').
$$

We set $\mathbf{x}=\mathbf{x}_{k+1},\mathbf{y}=\mathbf{y}_{k+1}$, $\mathbf{x}'=\mathbf{x}_k,\mathbf{y}'=\mathbf{y}_k$, which yields

$$
\frac{\mu_x}{4}\left\|\mathbf{x}_k^*(\mathbf{y}_{k+1})-\mathbf{x}_k\right\|^2+\frac{\mu_y}{4}\left\|\mathbf{y}_k^*(\mathbf{x}_{k+1})-\mathbf{y}_k\right\|^2\leq\mathrm{Gap}^k(\mathbf{x}_{k+1},\mathbf{y}_{k+1})+\mathrm{Gap}^k(\mathbf{x}_k,\mathbf{y}_k). \tag{36}
$$

Due to Lemma 25 we have

$$
\begin{aligned}
\mathrm{Gap}^k(\mathbf{x}_{k+1},\mathbf{y}_{k+1})\leq&\max\left\{\frac{1}{2\eta_{\mathbf{y}}T},\frac{1}{4\eta_{\mathbf{x}}T}\right\}\left(\mathbb{E}\left\|\mathbf{y}_k^*(\mathbf{x}_{k+1})-\mathbf{y}_k\right\|^2+\mathbb{E}\left\|\mathbf{x}_k^*(\mathbf{y}_{k+1})-\mathbf{x}_k\right\|^2\right)+\frac{C\mathbb{E}\left\|\mathbf{z}^{0}-g(\mathbf{x}^{0})\right\|^2}{T}\\
&+\frac{\tilde{\sigma}^2}{\mu_xB_\tau}+\frac{10\tilde{\sigma}^2}{\mu_yB_\tau}+\left(\frac{24G_g^2L_f^2\tau}{\mu_xB}+\frac{120\tau}{\mu_yB}+\frac{20L_f^2}{\mu_y}+\frac{2L_f^2G_g^2}{\mu_x}\right)9\frac{\tilde{\sigma}^2}{B_\tau}.
\end{aligned} \tag{37}
$$

Combining (36) and (37) yields:

$$
\begin{aligned}
&Q\left(\mathrm{Gap}^k(\mathbf{x}_{k+1},\mathbf{y}_{k+1})-\frac{\sigma^2}{\mu_xB_\tau}-\frac{10\sigma^2}{\mu_yB_\tau}-\frac{C\mathbb{E}\left\|\mathbf{z}^{0}-g(\mathbf{x}^{0})\right\|^2}{T}-\left(\frac{24G_g^2L_f^2\tau}{\mu_xB}+\frac{120\tau}{\mu_yB}+\frac{20L_f^2}{\mu_y}+\frac{2L_f^2G_g^2}{\mu_x}\right)9\frac{\tilde{\sigma}^2}{B_\tau}\right)\\
&\leq\mathrm{Gap}^k(\mathbf{x}_{k+1},\mathbf{y}_{k+1})+\mathrm{Gap}^k(\mathbf{x}_k,\mathbf{y}_k)\\
\Longleftrightarrow\ &(Q-1)\,\mathrm{Gap}^k(\mathbf{x}_{k+1},\mathbf{y}_{k+1})\leq\mathrm{Gap}^k(\mathbf{x}_k,\mathbf{y}_k)\\
&\quad+Q\left(\frac{\sigma^2}{\mu_xB_\tau}+\frac{10\sigma^2}{\mu_yB_\tau}+\frac{C\mathbb{E}\left\|\mathbf{z}^{0}-g(\mathbf{x}^{0})\right\|^2}{T}+\left(\frac{24G_g^2L_f^2\tau}{\mu_xB}+\frac{120\tau}{\mu_yB}+\frac{20L_f^2}{\mu_y}+\frac{2L_f^2G_g^2}{\mu_x}\right)9\frac{\tilde{\sigma}^2}{B_\tau}\right).
\end{aligned}
$$

□

Lemma 27 (Lemma 8 Yan et al. [2020]). *If $F(\mathbf{x},\mathbf{y})$ is ρ weakly convex in $\mathbf{x}$, and let $F_k(\mathbf{x},\mathbf{y}) = F(\mathbf{x},\mathbf{y}) + \frac{\mu_x+\rho}{2}\|\mathbf{x}-\mathbf{x}_k\|^2$. Then the following statements hold true for any $\alpha_1,\alpha_2 \in (0,1)$:*

$$
\begin{aligned}
F_k(\mathbf{x}_{k+1},\mathbf{y}) - F_k(\mathbf{x},\mathbf{y}_{k+1}) &\geq \left(1 - \frac{\mu_x+\rho}{\rho}\left(\frac{1}{\alpha_1}-1\right)\right)\left(F_{k+1}(\mathbf{x}_{k+1},\mathbf{y}) - F_{k+1}(\mathbf{x},\mathbf{y}_{k+1})\right) \\
&\quad - \frac{\mu_x+\rho}{2}\frac{\alpha_1}{1-\alpha_1}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 \qquad (38)\\
F_k(\mathbf{x}_{k+1},\mathbf{y}) - F_k(\mathbf{x},\mathbf{y}_{k+1}) &\geq \Phi(\mathbf{x}_{k+1}) - \Phi(\mathbf{x}_k) + \frac{\mu_x+\rho}{2}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 \qquad (39)\\
F_k(\mathbf{x}_{k+1},\mathbf{y}) - F_k(\mathbf{x},\mathbf{y}_{k+1}) &\geq \frac{\rho\alpha_2}{2}\|\mathbf{x}_k-\mathbf{x}_k^*\|^2 - \frac{\rho\alpha_2}{2(1-\alpha_2)}\|\mathbf{x}_k-\mathbf{x}_{k+1}\|^2. \qquad (40)
\end{aligned}
$$

Lemma 28. *For Algorithm 5, under assumptions of Theorem 6, the following statement holds:*

$$
\begin{aligned}
(Q-1)\frac{a_3\rho\alpha_2}{2}\|\mathbf{x}_k-\mathbf{x}_k^*\|^2 &\leq \mathrm{Gap}^k(\mathbf{x}_k,\mathbf{y}_k) - (Q-1)a_1\left(1-\frac{\mu_x+\rho}{\rho}\left(\frac{1}{\alpha_1}-1\right)\right)\mathrm{Gap}^{k+1}(\mathbf{x}_{k+1},\mathbf{y}_{k+1}) \\
&\quad - (Q-1)a_2\left(\Phi(\mathbf{x}_{k+1})-\Phi(\mathbf{x}_k)\right) \\
&\quad + Q\left(\frac{\sigma^2}{\mu_x B_\tau} + \frac{10\sigma^2}{\mu_y B_\tau} + \frac{C\mathbb{E}\left\|\mathbf{z}^0 - g(\mathbf{x}^0)\right\|^2}{T} + \left(\frac{24G_g^2L_f^2\tau}{\mu_x B} + \frac{120\tau}{\mu_y B} + \frac{20L_f^2}{\mu_y} + \frac{2L_f^2G_g^2}{\mu_x}\right)9\frac{\tilde{\sigma}^2}{B_\tau}\right),
\end{aligned}
$$

where a_1, a_2, a_3 are such that

$$
\begin{aligned}
a_1+a_2+a_3 &= 1,\\
a_1\frac{(\mu_x+\rho)\alpha_1}{2(1-\alpha_1)} - a_2\frac{\mu_x+\rho}{2} + a_3\frac{\rho\alpha_2}{2(1-\alpha_2)} &\leq 0,\\
(Q-1)a_1\left(1-\frac{\mu_x+\rho}{\rho}\left(\frac{1}{\alpha_1}-1\right)\right) &\geq 1.
\end{aligned}
$$

Proof. Due to Lemma 27, $a_1\cdot(38) + a_2\cdot(39) + a_3\cdot(40)$ yields:

$$
\begin{aligned}
\frac{a_3\rho\alpha_2}{2}\|\mathbf{x}_k-\mathbf{x}_k^*\|^2 &\leq \mathrm{Gap}^k(\mathbf{x}_{k+1},\mathbf{y}_{k+1}) - a_1\left(1-\frac{\mu_x+\rho}{\rho}\left(\frac{1}{\alpha_1}-1\right)\right)\mathrm{Gap}^{k+1}(\mathbf{x}_{k+1},\mathbf{y}_{k+1}) \\
&\quad - a_2\left(\Phi(\mathbf{x}_{k+1})-\Phi(\mathbf{x}_k)\right) + \left(a_1\frac{\mu_x+\rho}{2}\frac{\alpha_1}{1-\alpha_1} + a_3\frac{\rho\alpha_2}{2(1-\alpha_2)} - a_2\frac{\mu_x+\rho}{2}\right)\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2,
\end{aligned}
$$

Multiply both sides with $Q-1$, plugging result from Lemma 26, then the proof immediately follows the condition $a_1\frac{(\mu_x+\rho)\alpha_1}{2(1-\alpha_1)} - a_2\frac{\mu_x+\rho}{2} + a_3\frac{\rho\alpha_2}{2(1-\alpha_2)} \leq 0$ and $(Q-1)a_1\left(1-\frac{\mu_x+\rho}{\rho}\left(\frac{1}{\alpha_1}-1\right)\right) \geq 1$. □

F.1 Proof of Theorem 6

Proof. First in Lemma 28, we set $\mu_x = \rho = L$, and we choose $\eta_\mathbf{x} \leq \frac{1}{2}\eta_\mathbf{y}$, together with assumption that $L \geq \mu_y$, we know $Q = \eta_\mathbf{x} T\mu_y$. We set:

$$
a_1 = \frac{1}{12}, a_2 = \frac{10}{12}, a_3 = \frac{1}{12}, \alpha_1 = \frac{3}{4}, \alpha_2 = \frac{3}{4}, \eta_\mathbf{x} T\mu_y \geq 37 \qquad (41)
$$

, the condition in Lemma 28 can hold, and we have:

$$
\begin{aligned}
(\eta_\mathbf{x} T\mu_y - 1)\frac{3L}{96}\|\mathbf{x}_k-\mathbf{x}_k^*\|^2 &\leq \mathrm{Gap}^k(\mathbf{x}_k,\mathbf{y}_k) - \mathrm{Gap}^{k+1}(\mathbf{x}_{k+1},\mathbf{y}_{k+1}) \\
&\quad - (\eta_\mathbf{x} T\mu_y - 1)\frac{5}{6}\left(\Phi(\mathbf{x}_{k+1})-\Phi(\mathbf{x}_k)\right) \\
&\quad + \eta_\mathbf{x} T\mu_y\left(\frac{\tilde{\sigma}^2}{\mu_x B_\tau} + \frac{10\tilde{\sigma}^2}{\mu_y B_\tau} + \left(\frac{24G_g^2L_f^2\tau}{\mu_x B} + \frac{120\tau}{\mu_y B} + \frac{20L_f^2}{\mu_y} + \frac{2L_f^2G_g^2}{\mu_x}\right)9\frac{\tilde{\sigma}^2}{B_\tau}\right),
\end{aligned}
$$

Due to the fact $4L^2\left\|\mathbf{x}_k-\mathbf{x}_k^*\right\|^2=\left\|\Phi_{1/2L}(\mathbf{x}_k)\right\|^2$, and summing over $k=0$ to $K-1$ we have

$$
\begin{aligned}
&(\eta_{\mathbf{x}}T\mu_y-1)\frac{1}{K}\sum_{k=0}^{K-1}\frac{3L}{384L^2}\left\|\nabla\Phi_{1/2L}(\mathbf{x}_k)\right\|^2 \le \frac{\mathrm{Gap}^k(\mathbf{x}_0,\mathbf{y}_0)}{K}+(\eta_{\mathbf{x}}T\mu_y-1)\frac{5(\Phi(\mathbf{x}_0)-\min_{\mathbf{x}\in\mathcal{X}}\Phi(\mathbf{x}))}{6K}\\
&\quad+\eta_{\mathbf{x}}T\mu_y O\left(\frac{C\mathbb{E}\left\|\mathbf{z}^0-g(\mathbf{x}^0)\right\|^2}{T}+\frac{\tilde{\sigma}^2}{\mu_x B_\tau}+\frac{\tilde{\sigma}^2}{\mu_y B_\tau}+\left(\frac{G_g^2L_f^2\tau}{\mu_x B}+\frac{\tau}{\mu_y B}+\frac{L_f^2}{\mu_y}+\frac{L_f^2G_g^2}{\mu_x}\right)\frac{\tilde{\sigma}^2}{B_\tau}\right),\\
\Longleftrightarrow &\frac{1}{K}\sum_{k=0}^{K-1}\left\|\nabla\Phi_{1/2L}(\mathbf{x}_k)\right\|^2 \le \frac{384L\mathrm{Gap}^k(\mathbf{x}_0,\mathbf{y}_0)}{3(\eta_{\mathbf{x}}T\mu_y-1)K}+\frac{1920L\Delta_\Phi}{18K}\\
&\quad+O\left(\frac{C\mathbb{E}\left\|\mathbf{z}^0-g(\mathbf{x}^0)\right\|^2}{T}+\frac{L\tilde{\sigma}^2}{\mu_x B_\tau}+\frac{L\tilde{\sigma}^2}{\mu_y B_\tau}+\left(\frac{G_g^2L_f^2\tau}{\mu_x B}+\frac{\tau}{\mu_y B}+\frac{L_f^2}{\mu_y}+\frac{L_f^2G_g^2}{\mu_x}\right)\frac{L\tilde{\sigma}^2}{B_\tau}\right).
\end{aligned}
$$

Recall that by our initialization method, $\mathbb{E}\left\|\mathbf{z}^0-g(\mathbf{x}^0)\right\|^2\le\delta$. To ensure RHS is less than ϵ^2, we choose $K=\Theta\left(\frac{L\Delta_\Phi}{\epsilon^2}\right)$, $B_\tau=\Theta\left(\frac{L^2\sigma^2}{\mu_y\epsilon^2}\right)$, $\tau=\sqrt{\frac{B_\tau}{\kappa}}$ and $B=\sqrt{\kappa B_\tau}$.

It remains to determine value of T. According to condition 41, we need $T\ge\frac{6}{\mu_y\eta_{\mathbf{x}}}$. Since

$$
\begin{aligned}
\eta_{\mathbf{x}} &\le \Theta\left(\min\left\{\frac{\mu_y B}{G_g^2L_f^2\tau},\frac{1}{L_f'},\frac{\rho B^2}{G_g^4L_f^2\tau^2},\frac{\mu_y B^2}{\tau^2}\right\}\right)\\
&\le \Theta\left(\min\left\{\frac{\mu_y\kappa}{G_g^2L_f^2},\frac{1}{L},\frac{\rho\kappa^2}{G_g^4L_f^2},\mu_y\kappa^2\right\}\right)=\Theta\left(\frac{1}{L}\right),
\end{aligned}
$$

we know that $T\ge\Omega(\kappa)$. And hence we should choose $\delta\le O(\kappa\epsilon^2/C)=O(\epsilon^2/L)$, since $C=\frac{12G_g^2L^2\tau}{\mu_x B}+\frac{60\tau}{\mu_y B}+\frac{12L_f^2}{\mu_y}+\frac{L_f^2G_g^2}{\mu_x}=O(\kappa L)$. The final gradient complexity is

$$
\frac{TKB_\tau}{\tau}+TKB=O\left(\frac{\kappa^2\sqrt{L}\Delta_\Phi\sigma}{\epsilon^3}\right).
$$

□

F.2 Proof of Theorem 7

Let $\tilde{\Phi}(\mathbf{x})=\max_{\mathbf{y}\in\mathcal{Y}}\tilde{F}(\mathbf{x},\mathbf{y})$, and $\tilde{F}(\mathbf{x},\mathbf{y})$ is defined in (2) . According to Zhang et al. [2022] Lemma 27, a $\epsilon/2\sqrt{6}$ stationary point of Moreau Envelope of $\tilde{\Phi}(\mathbf{x})$ is ϵ stationary point of that of $\Phi(\mathbf{x})$. Then the complexity result follows immediately by setting $\mu=\frac{\epsilon^2}{LD_{\mathcal{Y}}^2}$ in Theorem 6.