Noname manuscript No.
(will be inserted by the editor)

Katyusha Acceleration for Convex Finite-Sum Compositional Optimization

Yibo Xu · Yangyang Xu

October 21, 2019

Abstract Structured optimization problems arise in many applications. To efficiently solve these problems, it is important to leverage the structure information in the algorithmic design. This paper focuses on convex problems with a finite-sum compositional structure. Finite-sum problems appear as the sample average approximation of a stochastic optimization problem and also arise in machine learning with a huge amount of training data. One popularly used numerical approach for finite-sum problems is the stochastic gradient method (SGM). However, the additional compositional structure prohibits easy access to unbiased stochastic approximation of the gradient, so directly applying the SGM to a finite-sum compositional optimization problem (COP) is often inefficient.

We design new algorithms for solving strongly-convex and also convex two-level finite-sum COPs. Our design incorporates the Katyusha acceleration technique and adopts the mini-batch sampling from both outer-level and inner-level finite-sum. We first analyze the algorithm for strongly-convex finite-sum COPs. Similar to a few existing works, we obtain linear convergence rate in terms of the expected objective error, and from the convergence rate result, we then establish complexity results of the algorithm to produce an ε-solution. Our complexity results have the same dependence on the number of component functions as existing works. However, due to the use of Katyusha acceleration, our results have better dependence on the condition number κ and improve to $\kappa^{2.5}$ from the best-known κ^3. Finally, we analyze the algorithm for convex finite-sum COPs, which uses as a subroutine the algorithm for strongly-convex finite-sum COPs. Again, we obtain better complexity results than existing works in terms of the dependence on ε, improving to $\varepsilon^{-2.5}$ from the best-known ε^{-3}.

Keywords Finite-sum composition · Katyusha · variance reduction · stochastic approximation

Mathematics Subject Classification (2000) 90C06 · 90C15 · 90C25 · 62L20 · 65C60 · 65Y20

Yibo Xu
Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY
E-mail: xuy24@rpi.edu
Yangyang Xu
Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY
E-mail: xuy21@rpi.edu

1 Introduction

Utilizing structure information of a problem is crucial for designing efficient algorithms, especially when the problem involves a high-dimensional variable and/or a huge amount of data. For example, recent works (e.g., [7, 24]) have shown that on solving a finite-sum problem, the variance-reduced stochastic gradient method, which utilizes the finite-sum structure information, can significantly outperform a deterministic gradient method and a non-variance-reduced stochastic gradient method.

In this paper, we focus on the finite-sum compositional optimization problem (COP):

$$\min_{x\in\mathbb{R}^{N_2}} H(x) \equiv \frac{1}{n_1}\sum_{i=1}^{n_1} F_i\left(\frac{1}{n_2}\sum_{j=1}^{n_2} G_j(x)\right) + h(x), \tag{1.1}$$

where $F_i\colon \mathbb{R}^{N_1}\to\mathbb{R}$ is a differentiable function for each $i=1,\dots,n_1$, $G_j\colon \mathbb{R}^{N_2}\to\mathbb{R}^{N_1}$ is a differentiable map for each $j=1,\dots,n_2$, and h is a simple (but possibly non-differentiable) function. For ease of description, we let $F\colon \mathbb{R}^{N_1}\to\mathbb{R}$ and $G\colon \mathbb{R}^{N_2}\to\mathbb{R}^{N_1}$ respectively denote the average of $\{F_i\}$ and $\{G_j\}$, i.e.,

$$F = \frac{1}{n_1}\sum_{i=1}^{n_1} F_i, \quad G = \frac{1}{n_2}\sum_{j=1}^{n_2} G_j.$$

Also, we let $f\colon \mathbb{R}^{N_2}\to\mathbb{R}$ be the composition of F with G, namely,

$$f = F\circ G. \tag{1.2}$$

The problem (1.1) can be viewed as a sample average approximation (SAA) of a two-level stochastic COP, for which [21] proposes and analyzes a class of stochastic compositional gradient methods. Very recently, [26,29] extend the results to a multiple-level stochastic COP. Although it is possible to extend our method and analysis to a multiple-level finite-sum COP, we will focus on the two-level case because of the applications that we are interested in. Our main goal is to design a gradient-based (also called *first-order*) algorithm for (1.1) and to analyze its complexity to produce a stochastic ε-solution $\bar{x}$, i.e., $\mathbb{E}\big[H(\bar{x}) - H(x^*)\big] \le \varepsilon$, where x^* is a minimizer of H. The complexity is measured by the number of gradient evaluations of each F_i and Jacobian matrix evaluations of each G_j. The method in [21] can certainly be applied to (1.1). However, due to the utilization of the finite-sum structure, our method can have significantly better complexity result.

Below, we give two examples that motivate us to consider problems in the form of (1.1).

Example I: Risk-Averse Learning. Given a set of data $\{(\boldsymbol{a}_i, b_i)\}_{i=1}^n$ sampled from a certain distribution, the (sample) mean-variance minimization [13,14] can be formulated as

$$\min_{\boldsymbol{x}\in X}\ \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{x};\boldsymbol{a}_i,b_i) + \frac{\lambda}{n}\sum_{i=1}^n\left[\ell(\boldsymbol{x};\boldsymbol{a}_i,b_i) - \frac{1}{n}\sum_{j=1}^n \ell(\boldsymbol{x};\boldsymbol{a}_j,b_j)\right]^2, \tag{1.3}$$

where ℓ is the loss function, X is a closed convex set in $\mathbb{R}^N$, and $\lambda>0$ is to balance the trade-off between mean and variance. As shown in [9], define $G_j:\mathbb{R}^N\to\mathbb{R}^{N+1}$ and $F_i:\mathbb{R}^{N+1}\to\mathbb{R}$ by

$$G_j(\boldsymbol{x}) = \big[\boldsymbol{x}; \ell(\boldsymbol{x};\boldsymbol{a}_j,b_j)\big], \quad F_i(\boldsymbol{z},y) = \lambda(\ell(\boldsymbol{z};\boldsymbol{a}_i,b_i)-y)^2 + \ell(\boldsymbol{z};\boldsymbol{a}_i,b_i),$$

for $j = 1, \ldots, n$ and $i = 1, \ldots, n$, and let h be the indicator function on X. Then (1.3) can be rewritten into the form of (1.1) with $n_1 = n_2 = n$.

As an alternative, by expanding the square term, we write (1.3) equivalently into

$$\min_{\boldsymbol{x}\in X} \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{x}; \boldsymbol{a}_i, b_i) + \frac{\lambda}{n}\sum_{i=1}^{n} \left(\ell(\boldsymbol{x}; \boldsymbol{a}_i, b_i)\right)^2 - \lambda \left[\frac{1}{n}\sum_{j=1}^{n} \ell(\boldsymbol{x}; \boldsymbol{a}_j, b_j)\right]^2. \tag{1.4}$$

Define $G_j : \mathbb{R}^N \to \mathbb{R}^{N+1}$ and $F_i : \mathbb{R}^{N+1} \to \mathbb{R}$ by

$$G_j(\boldsymbol{x}) = \big[\boldsymbol{x}; \ell(\boldsymbol{x}; \boldsymbol{a}_j, b_j)\big], \quad F_i(\boldsymbol{z}, y) = \lambda \ell^2(\boldsymbol{z}; \boldsymbol{a}_i, b_i) + \ell(\boldsymbol{z}; \boldsymbol{a}_i, b_i) - \lambda y^2,$$

for $j = 1, \ldots, n$ and $i = 1, \ldots, n$, and also let h be the indicator function on X. Then we can rewrite (1.4) into the form of (1.1) with $n_1 = n_2 = n$.

Example II: finite-sum constrained problems via augmented Lagrangian. Consider a problem with a finite-sum objective and also finite-sum constraints:

$$\min_{x\in X} \frac{1}{m}\sum_{i=1}^{m} f_i(x), \quad \text{s.t.} \quad \frac{1}{n}\sum_{j=1}^{n} g_{jk}(x) \le 0,\ k = 1, \ldots, K, \tag{1.5}$$

where X is a closed convex set in $\mathbb{R}^N$. The Neyman-Pearson classification problem [18] can be formulated as (1.5) with $J = 1$, and the fairness-constrained classification problem [28] can be written in the form of (1.5) with $J = 2$. The augmented Lagrangian method (ALM) is one popular and effective way for solving functional constrained problems. It has been shown [25] that applying an optimal first-order method (FOM) within the ALM framework can yield an overall (near) optimal FOM for nonlinear functional constrained problems. By the classic AL function, the primal subproblem (that is the most expensive step in the ALM) takes the form:

$$\min_{x\in X} \frac{1}{m}\sum_{i=1}^{m} f_i(x) + \sum_{k=1}^{K} \psi_\beta\left(\frac{1}{n}\sum_{j=1}^{n} g_{jk}(x),\, z_k\right), \tag{1.6}$$

where $\beta > 0$ is the augmented penalty parameter, and

$$\psi_\beta(u, v) := \frac{1}{2\beta}\left[\max\{\beta u,\, -v\}\right]^2 + \frac{v}{\beta}\max\{\beta u,\, -v\}.$$

Given $\beta > 0$ and $z \in \mathbb{R}^K$, define $G_j : \mathbb{R}^N \to \mathbb{R}^{N+K}$ and $F_i : \mathbb{R}^{N+K} \to \mathbb{R}$ by

$$G_j(x) = \big[x; g_{j1}(x); \ldots; g_{jK}(x)\big], \quad F_i(x, y) = f_i(x) + \sum_{k=1}^{K} \psi_\beta(y_k, z_k)$$

for $j = 1, \ldots, n$ and $i = 1, \ldots, m$, and also let h be the indicator function on X. Then we can write (1.6) into the form of (1.1) with $n_1 = m$ and $n_2 = n$.

1.1 Related Works

In this subsection, we review existing works on solving problems in the form of (1.1) or its special cases.

Approaches for convex finite-sum problems. When each G_j is the identity map, (1.1) reduces to the traditional finite-sum problem

$$\min_{x\in\mathbb{R}^N} f(x)+h(x) \equiv \frac{1}{n}\sum_{i=1}^{n} F_i(x)+h(x). \tag{1.7}$$

For solving (1.7), one can apply the proximal gradient method (PG) or its accelerated version APG [15]. Each iteration of PG or APG computes the gradient f at a point and performs a proximal mapping of h, and thus their per-iteration complexity is n, in terms of the number of gradient evaluation of component functions $\{F_i\}$. If each F_i has an L-Lipschitz continuous gradient, and f is μ-strongly convex, then PG and APG respectively need $O\left(\kappa\log\frac{1}{\varepsilon}\right)$ and $O\left(\sqrt{\kappa}\log\frac{1}{\varepsilon}\right)$ iterations to produce an ε-solution, where $\kappa=\frac{L}{\mu}$ denotes the condition number. Hence, their total complexities are respectively $O\left(n\kappa\log\frac{1}{\varepsilon}\right)$ and $O\left(n\sqrt{\kappa}\log\frac{1}{\varepsilon}\right)$, which are high as n is large. The stochastic gradient descent (SGD) can be used for the big-n case of (1.7). At each update, it only needs to evaluate the gradient of one or a few randomly sampled functions and can produce a stochastic ε-solution with a complexity of $O\left(\frac{G\kappa}{\mu\varepsilon}\right)$. Here, G is a bound for the second moment of sample gradients. The complexity of SGD could be lower than those of PG and APG if n is large and ε is not too tiny. While PG and APG treat (1.7) as a regular deterministic problem, the SGD simply takes it as a stochastic program. None of them utilize the finite-sum structure. It turns out that a better complexity of $O\left((n+\kappa)\log\frac{1}{\varepsilon}\right)$ can be obtained by utilizing the special structure, through a random sampling together with a variance-reduction (VR) technique (e.g., [5,7,19,24]). Furthermore, the Katyusha acceleration [1] incorporates the VR and the linear coupling technique [3] that is disassembled from Nesterov's acceleration. The Katyusha accelerated method achieves a complexity of $O\left((n+\sqrt{n\kappa})\log\frac{1}{\varepsilon}\right)$, which is lower than $O\left((n+\kappa)\log\frac{1}{\varepsilon}\right)$ if $n\ll\kappa$. Similar results have also been shown in [4,10,20]. They match with the lower complexity bound given in [23] and thus are optimal for solving problems in the form of (1.7).

Approaches for finite-sum COPs. Several methods have been designed specifically for solving problems with finite-sum composition structure. However, their complexity results are generally worse than those obtained for solving problems in the form of (1.7). For example, to produce a stochastic ε-solution, the methods in [6,8] both use the VR technique and bear a complexity of $O\left((n_1+n_2+\kappa^3)\log\frac{1}{\varepsilon}\right)$ if the objective in (1.1) is strongly convex and has condition number κ. This result is significantly worse than the complexity of $O\left((n+\kappa)\log\frac{1}{\varepsilon}\right)$ mentioned previously for solving (1.7), in terms of the dependence on the condition number κ. The worse result is caused by the additional composition structure, which prohibits easy access to unbiased stochastic estimation of ∇f. To see this, note that

$$\nabla f(x)=[\nabla G(x)]^\top\nabla F(G(x))=\left[\frac{1}{n_2}\sum_{j=1}^{n_2}\nabla G_j(x)\right]^\top\nabla F\left(\frac{1}{n_2}\sum_{j=1}^{n_2}G_j(x)\right).$$

Hence, if we can unbiasedly estimate the Jacobian matrix $\nabla G(x)$ and the gradient $\nabla F(G(x))$ independently, then an unbiased estimation of $\nabla f(x)$ can be obtained. However, $\nabla F(\mathbb{E}[\xi])=\mathbb{E}[\nabla F(\xi)]$ does not generally hold for a random vector ξ, and thus though we can easily have an unbiased stochastic approximation of $\nabla G(x)$ and $G(x)$ by randomly sampling from $\{G_j\}$, this way does not guarantee an unbiased estimation of $\nabla F(G(x))$, let alone $\nabla f(x)$.

Complexity results have been established in the literature for problem (1.1) under different scenarios. For example, [6, 8] studied the scenario where f is smooth and strongly convex. Both of them inherit the algorithmic design from [7] and achieved linear convergence. Besides the strongly convex scenario, other cases of (1.1) have also been studied. The case with a smooth and convex f was treated, for example, in [6,11]. The case with a smooth but non-convex f is studied, for example, in [12,22]. The work [12] employs the variance reduction technique while sampling the inner map G and its Jacobian and the outer map F, for which various mini-batch sizes can be taken. Instead of a finite-sum problem, [22] studies a stochastic composition problem. Assuming that the sampling at the inner layer is unbaised and has a bounded variance, [22] provides sublinear guarantees for strongly convex, convex, and nonconvex cases. The work [27] deals with a finite-sum composition problem with an additional linear constraint. It integrates variance reduction with the alternating direction method of multipliers. For comparison, we list complexity results of state-of-the-art methods for the strongly-convex and convex cases in Table 1.

1.2 Our Contributions

The contributions of this paper are mainly on designing new algorithms for solving convex and strongly-convex finite-sum compositional optimization problems and establishing complexity results that appear the best so far. They are summarized as follows.

- First, we propose a new algorithm for solving strongly convex compositional optimization. Our design incorporates Katyusha acceleration [1] together with a mini-batch sampling technique.
- Second, we conduct the complexity analysis of the new algorithm in three different scenarios. We start from the scenario where the outer finite-sum in (1.1) has a relatively small number n_1 but the inner finite-sum takes a big number n_2. Then we analyze the scenario where both n_1 and n_2 are big, and each $F_i \circ G$ in (1.1) may be non-convex. The third scenario also has big n_1 and n_2 but assumes convexity of each $F_i \circ G$. For all the three scenarios, our complexity results are roughly in the order of $(n_1 + n_2 + \kappa^{2.5}) \log \frac{1}{\varepsilon}$ to produce a stochastic ε-solution. The third scenario can have a result with a better dependence on κ; see Corollary 5.10 below and its remark. Our complexity results are better than existing ones by an order of $\sqrt{\kappa}$. This is due to the incorporation of the Katyusha acceleration in our algorithmic design.
- Thirdly, we propose a new algorithm for solving convex compositional optimization, by applying the optimal black-box reduction technique [2] and our proposed strongly-convex problem solver as a subroutine. For all the three scenarios mentioned above, our complexity results are roughly in the order of $(n_1 + n_2) \log \frac{1}{\varepsilon} + \frac{1}{\varepsilon^{2.5}}$. Compared to existing results, ours are better by an order of $\frac{1}{\sqrt{\varepsilon}}$.

1.3 Notation and organization

Throughout the paper, we use $\|\cdot\|$ to denote the Euclidean norm of a vector and also the spectral norm of a matrix. For any real number a, we use $\lceil a \rceil$ for the least integer that is lower bounded by a, $\lfloor a \rfloor$ for the greatest integer that is upper bounded by a, and for any positive integer n, we use $[n]$ for the set $\{1, \ldots, n\}$. For a differentiable scalar function f, ∇f denotes its gradient, and for a differentiable vector function G, ∇G denotes its Jacobian matrix. $\mathbb{E}$ is used for the full expectation, and a subscript will be added for conditional expectation. We use the big-O, big-Ω, and big-Θ notation with the standard meanings to compare two numbers that both can go to infinity. Specifically, $a = O(b)$ means that there exists a uniform constant

Table 1 A comparison of complexity results amongst several state-of-the-art algorithms for solving problems in the form of (1.1) to produce a stochastic ε-solution; see Definition 1. "$h \neq 0$" is to reflect whether the algorithm handles a proximal term. "conv. outer sum." stands for convex outer summand, meaning whether the convexity of each $F_i \circ G$ is assumed. In that column, "both" indicates that the analysis is done with the assumption and also done without the assumption. In the fourth column, we use κ for the condition number. Although the compared papers have different definitions of κ, they all have μ as the denominator in the fractions.

	$h \neq 0$	conv. outer sum.	f strongly convex	f convex
AGD [16, 17]	yes	both	$(n_1+n_2)\sqrt{\kappa}\log\frac{1}{\varepsilon}$	$(n_1+n_2)\frac{1}{\sqrt{\varepsilon}}$
Compositional-SVRG-2 [8]	no	yes	$(n_1+n_2+\kappa^3)\log\frac{1}{\varepsilon}$	–
VRSC-PG [6]	yes	yes	$(n_1+n_2+\kappa^3)\log\frac{1}{\varepsilon}$	$n_1+n_2+\frac{(n_1+n_2)^{\frac{2}{3}}}{\varepsilon^2}$
SCVRG [11]	yes	no	–	$(n_1+n_2)\log\frac{1}{\varepsilon}+\frac{1}{\varepsilon^3}$
This paper	yes	both	$(n_1+n_2+\kappa^{2.5})\log\frac{1}{\varepsilon}$	$(n_1+n_2)\log\frac{1}{\varepsilon}+\frac{1}{\varepsilon^{2.5}}$

$C > 0$ such that $a \leq C \cdot b$, $a = \Omega(b)$ means that there exists a uniform constant $c > 0$ such that $a \geq c \cdot b$, and $a = \Theta(b)$ means that $a = O(b)$ and $a = \Omega(b)$ both hold.

Definition 1 (stochastic ε-solution)**.** *Given $\varepsilon > 0$, a random vector $\bar{x}$ is called a stochastic ε-solution of (1.1) if $\mathbb{E}\big[H(\bar{x}) - H(x^*)\big] \leq \varepsilon$, where x^* is a minimizer of H.*

Definition 2 (L-smoothness)**.** *A differentiable scalar (resp. vector) function ϕ on a set X is called L-smooth with $L \geq 0$ if its gradient (resp. Jacobian matrix) $\nabla\phi$ is L-Lipschitz continuous, namely,*

$$\|\nabla\phi(x) - \nabla\phi(x')\| \leq L\|x - x'\|, \forall x, x' \in X.$$

Definition 3 (bounded gradient)**.** *A differentiable scalar (resp. vector) function ϕ on a set X has a b-bounded gradient (resp. Jacobian matrix) ∇f, if*

$$\|\nabla f(x)\| \leq b, \forall x \in X.$$

Definition 4 (μ-strong convexity)**.** *A function ϕ on a convex set X is called μ-strongly convex for some $\mu > 0$, if*

$$\phi(x') \geq \phi(x) + \langle\tilde{\nabla}\phi(x), x' - x\rangle + \frac{\mu}{2}\|x' - x\|^2, \forall x, x' \in X,$$

where $\tilde{\nabla}\phi(x)$ stands for a subgradient of ϕ at x.

The rest of the paper is organized as follows. In Section 2, we give the technical assumptions and also the algorithm for the strongly convex case of (1.1). A few lemmas are established in Section 3. In Section 4, we analyze the algorithm for the case of relatively small n_1 and big n_2, and in Section 5, we conduct the analysis for big n_1 and big n_2. Strong convexity is assumed in Sections 4 and 5, and in Section 6, we propose an algorithm for the convex case of (1.1) and give the complexity results. Section 7 concludes the paper.

2 Technical assumptions and proposed algorithm

The following three assumptions are made throughout the analysis for strongly convex cases of (1.1).

Assumption 2.1. *The function f given in (1.2) is convex, and the function h in (1.1) is μ-strongly convex with $\mu > 0$.*

Assumption 2.2. *For every $i \in [n_1]$, F_i is L_F-smooth and has a B_F-bounded gradient, and for every $j \in [n_2]$, G_j is L_G-smooth and has a B_G-bounded Jacobian matrix.*

By this assumption, f must be smooth, and also F is L_F-smooth and has a B_F-bounded gradient. Note that we do not assume the smoothness of h, but the proximal mapping of h needs to be easy to implement our algorithm.

Assumption 2.3. *For every $i \in [n_1]$ and every $j \in [n_2]$, it holds*

$$\left\|[\nabla G_j(x)]^\top \nabla F_i(G(x)) - [\nabla G_j(y)]^\top \nabla F_i(G(y))\right\| \leq L\|x-y\|, \forall\, x,\, y \in \mathbb{R}^{N_2}.$$

This assumption is a conventional one made in the literature. It guarantees the L-smoothness of $f_i := F_i \circ G$ for each $i \in [n_1]$ by the following arguments:

$$\begin{aligned}\|\nabla f_i(x) - \nabla f_i(y)\| &= \left\|[\nabla G(x)]^\top \nabla F_i(G(x)) - [\nabla G(y)]^\top \nabla F_i(G(y))\right\| \\ &= \left\|\frac{1}{n_2}\sum_{j=1}^{n_2}\Big([\nabla G_j(x)]^\top \nabla F_i(G(x)) - [\nabla G_j(y)]^\top \nabla F_i(G(y))\Big)\right\| \\ &\leq \frac{1}{n_2}\sum_{j=1}^{n_2}\left\|[\nabla G_j(x)]^\top \nabla F_i(G(x)) - [\nabla G_j(y)]^\top \nabla F_i(G(y))\right\| \leq L\|x-y\|, \end{aligned} \tag{2.1}$$

and it implies the L-smoothness of f as well by noting:

$$\|\nabla f(x) - \nabla f(y)\| = \left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\big(\nabla f_i(x) - \nabla f_i(y)\big)\right\| \leq \frac{1}{n_1}\sum_{i=1}^{n_1}\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x-y\|.$$

Notice that some model is provided with f being smooth and μ-strongly convex while h is only convex. For this case, one can let $f \leftarrow f - \frac{\mu}{2}\|\cdot\|^2$ and $h \leftarrow h + \frac{\mu}{2}\|\cdot\|^2$ to fit our assumptions. We assume $L \geq \mu$.

Under the above assumptions, we design the SoCK method to solve Problem (1.1), and the pseudocode is given in Algorithm 1. The design incorporates the linear coupling technique, i.e., the x_{k+1} update, which is used in [1], together with the variance-reduction technique that is adopted by the state-of-the-art algorithms, e.g. [6, 8, 11]. By the linear coupling technique with carefully selected momentum weights (τ_1 and τ_2 here), one can achieve the optimal deterministic rates for convex objectives as in [3] and the optimal stochastic rate for a stochastic finite-sum objective as in [1].

Algorithm 1 **S**trongly C**o**nvex **C**ompositional **K**atyusha (SoCK)

Input: $x_0 \in \mathbb{R}^{N_2}$, S, $m \le \frac{L}{2\mu}$, $\theta > 1$, inner minibatch sizes A and B, and outer minibatch size $C \ge 2 + 24\frac{L}{\mu}$;
Let $y_0 = z_0 = \widetilde{x}^0 \leftarrow x_0$;
for $s = 0$ to $S-1$ **do**
  Compute $G(\widetilde{x}^s)$, $\nabla G(\widetilde{x}^s)$ and $\nabla f(\widetilde{x}^s) \leftarrow [\nabla G(\widetilde{x}^s)]^\top \nabla F(G(\widetilde{x}^s))$; ▷ take a snapshot
  for $j = 0$ to $m-1$ **do**
    $k \leftarrow sm + j$;
    $x_{k+1} \leftarrow \tau_1 z_k + \tau_2 \widetilde{x}^s + (1-\tau_1-\tau_2) y_k$; ▷ linear coupling step
    Sample $\mathcal{A}_k$ and $\mathcal{B}_k$ uniformly at random from $[n_2]$ with replacement such that $|\mathcal{A}_k| = A$ and $|\mathcal{B}_k| = B$;
    Let $\widehat{G}_k \leftarrow \frac{1}{A}\sum_{j_k \in \mathcal{A}_k} \big(G_{j_k}(x_{k+1}) - G_{j_k}(\widetilde{x}^s)\big) + G(\widetilde{x}^s)$;
    Let $\nabla \widehat{G}_k \leftarrow \frac{1}{B}\sum_{j_k \in \mathcal{B}_k} \big(\nabla G_{j_k}(x_{k+1}) - \nabla G_{j_k}(\widetilde{x}^s)\big) + \nabla G(\widetilde{x}^s)$;
    Option I: let $\widetilde{\nabla}_{k+1} \leftarrow \Big[\nabla \widehat{G}_k\Big]^\top \nabla F(\widehat{G}_k)$; ▷ batch step for outer function
    Option II: Sample $\mathcal{C}_k$ uniformly at random from $[n_1]$ with $|\mathcal{C}_k| = C$ and let ▷ minibatch step for outer function

$$\widetilde{\nabla}_{k+1} \leftarrow \frac{1}{C}\sum_{i \in \mathcal{C}_k} \left(\Big[\nabla \widehat{G}_k\Big]^\top \nabla F_i(\widehat{G}_k) - [\nabla G(\widetilde{x}^s)]^\top \nabla F_i(G(\widetilde{x}^s)) \right) + \nabla f(\widetilde{x}^s)$$

    Let $z_{k+1} \leftarrow \arg\min_z \langle \widetilde{\nabla}_{k+1}, z\rangle + \frac{1}{2\alpha}\|z - z_k\|^2 + h(z)$; ▷ mirror descent step
    Let $y_{k+1} \leftarrow \arg\min_y \langle \widetilde{\nabla}_{k+1}, y\rangle + \frac{3L}{2}\|y - x_{k+1}\|^2 + h(y)$; ▷ gradient descent step
  end for
  $\widetilde{x}^{s+1} \leftarrow (\sum_{j=0}^{m-1}\theta^j)^{-1} \cdot \sum_{j=0}^{m-1}\theta^j y_{sm+j+1}$; ▷ update to the snapshot point
end for
return $\widetilde{x}^S$ or $x^{\text{out}} \leftarrow \frac{(1-\tau_1-\tau_2+\tau_1/8)y_{Sm} + (\tau_2+\tau_1/8)m\widetilde{x}^S}{1-\tau_1-\tau_2+\tau_1/8+(\tau_2+\tau_1/8)m}$.

The main process in Algorithm 1 follows [1]. It takes a snapshot every m iterations; the x-trajectory takes a linear coupling of the y and z-trajectories and the snapshot trajectory; the z-trajectory performs a mirror descent step over its own query point; the y-trajectory performs a proximal gradient step on the current x-query point, while the gradient direction is sampled on the current x-query point, similar to what is done in the literature. The update to the snapshot query point is an artifact of the analysis, for the sake of telescoping the progress among all snapshot points. The two different settings of the final output are also artifacts from our analysis. Notice that by counting the number of component function/gradient/Jacobian evaluations, the overall complexity of the algorithm is $S\big(n_1 + 2n_2 + m(2A + 2B + n_1)\big)$ if **Option I** is taken and $S\big(n_1 + 2n_2 + 2m(A + B + C)\big)$ if **Option II** is taken. Hence, we will take **Option I** if n_1 is in the same magnitude of $A + B$ and **Option II** only if $n_1 \gg A + B$. Corresponding to the two options, we will conduct the analysis separately in Section 4 and Section 5.

The randomness of the algorithm comes from the uniform samples $\mathcal{A}_k$, $\mathcal{B}_k$ and $\mathcal{C}_k$, for all $0 \le k \le Sm-1$. In our analysis, we use $\mathbb{E}_{k-1}$ for the conditional expectation with the history until the k-th iteration is fixed. More precisely, in Section 4 with **Option I** taken, $\mathbb{E}_{k-1}[\,\cdot\,] = \mathbb{E}\big[\,\cdot \,|\, \{\mathcal{A}_i, \mathcal{B}_i\}_{i=0}^{k-1}\big]$, and in Section 5 with **Option II** taken, $\mathbb{E}_{k-1}[\,\cdot\,] = \mathbb{E}\big[\,\cdot \,|\, \{\mathcal{A}_i, \mathcal{B}_i, \mathcal{C}_i\}_{i=0}^{k-1}\big]$. For ease of notation, we will use the following shorthands in our analysis

$$D_k \equiv \mathbb{E}\big[H(y_k) - H(x^*)\big],\ \widetilde{D}^s \equiv \mathbb{E}\big[H(\widetilde{x}^s) - H(x^*)\big], \tag{2.2}$$

and for readers' convenience, we list some important parameters with their meanings in Table 2.

Table 2 List of parameters and notations

L	smoothness parameter for problem (1.1); see Assumption 2.3
μ	strong-convexity parameter of h in (1.1)
A	number of samples with replacement used for the output estimation of the inner map G
B	number of samples with replacement used for the Jacobian estimation of the inner map G
C	number of samples with replacement used for the gradient estimation of the outer function F
S	number of outer loops
m	number of inner loops
$\widehat{G}_k$	the output estimation of the inner map G at iteration k, implimented variance reduction and mini-batch
$\nabla\widehat{G}_k$	the Jacobian estimation of the inner map G at iteration k, implimented variance reduction and mini-batch
$\widetilde{\nabla}_{k+1}$	the gradient estimation of the composition f at iteration k, implimented variance reduction and mini-batch
x^*	the optimal solution of (1.1)

3 Preparatory lemmas

In this section, we establish a few lemmas about the proposed SoCK method in Algorithm 1. These results hold if either **Option I** or **Option II** is taken, and thus they can be used to show our main convergence rate results in Section 4 and Section 5.

The first lemma is about the progress that the algorithm makes after obtaining y_{k+1}. Its proof follows that of [1, Lemma 3.3].

Lemma 3.1. *If*

$$y_{k+1} = \arg\min_y \langle \widetilde{\nabla}_{k+1}, y - x_{k+1}\rangle + \frac{3L}{2}\|y - x_{k+1}\|^2 + h(y) - h(x_{k+1}),$$

and

$$\mathrm{Prog}(x_{k+1}) \equiv -\min_y \left\{ \langle \widetilde{\nabla}_{k+1}, y - x_{k+1}\rangle + \frac{3L}{2}\|y - x_{k+1}\|^2 + h(y) - h(x_{k+1}) \right\} \geq 0,$$

we have

$$H(x_{k+1}) - H(y_{k+1}) \geq \mathrm{Prog}(x_{k+1}) - \frac{1}{4L}\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2.$$

Proof. We have

$$\begin{aligned}
\mathrm{Prog}(x_{k+1}) =& -\left(\langle \widetilde{\nabla}_{k+1}, y_{k+1} - x_{k+1}\rangle + \frac{3L}{2}\|y_{k+1} - x_{k+1}\|^2 + h(y_{k+1}) - h(x_{k+1}) \right) \\
=& -\left(\langle \nabla f(x_{k+1}), y_{k+1} - x_{k+1}\rangle + \frac{L}{2}\|y_{k+1} - x_{k+1}\|^2 + h(y_{k+1}) - h(x_{k+1}) \right) \\
&+ \left(\langle \nabla f(x_{k+1}) - \widetilde{\nabla}_{k+1}, y_{k+1} - x_{k+1}\rangle - L\|y_{k+1} - x_{k+1}\|^2 \right) \\
\leq& -\left(f(y_{k+1}) - f(x_{k+1}) + h(y_{k+1}) - h(x_{k+1})\right) + \frac{1}{4L}\|\nabla f(x_{k+1}) - \widetilde{\nabla}_{k+1}\|^2.
\end{aligned}$$

The last inequality above uses the L smoothness of f, as well as Young's inequality $\langle a, b\rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$. □

The bound on the variance of the biased sample gradient $\widetilde{\nabla}_{k+1}$ is critical for the convergence result. The next lemma will be used to derive a bound on the variance for the algorithm with either **Option I** or **Option II**.

Lemma 3.2. *Let $\widehat{G}_k$ and $\nabla\widehat{G}_k$ be those in Algorithm 1, and let $g(\cdot)$ be any function on $\mathbb{R}^{N_1}$ that is l-smooth and has b-bounded gradient, then*

$$\mathbb{E}_{k-1}\left[\left\|\left[\nabla\widehat{G}_k\right]^\top\nabla g(\widehat{G}_k)-[\nabla G(x_{k+1})]^\top\nabla g(G(x_{k+1}))\right\|^2\right]\le\left(\frac{2B_G^4l^2}{A}+\frac{2b^2L_G^2}{B}\right)\|\widetilde{x}^s-x_{k+1}\|^2.$$

Proof. First, we observe that

$$\begin{aligned}
&\left\|\left[\nabla\widehat{G}_k\right]^\top\nabla g(\widehat{G}_k)-\left[\nabla G(x_{k+1})\right]^\top\nabla g(G(x_{k+1}))\right\|^2\\
\overset{①}{\le}\ &2\left\|\left[\nabla\widehat{G}_k\right]^\top\nabla g(\widehat{G}_k)-\left[\nabla G(x_{k+1})\right]^\top\nabla g(\widehat{G}_k)\right\|^2+2\left\|\left[\nabla G(x_{k+1})\right]^\top\nabla g(\widehat{G}_k)-\left[\nabla G(x_{k+1})\right]^\top\nabla g(G(x_{k+1}))\right\|^2\\
\overset{②}{\le}\ &2b^2\left\|\nabla\widehat{G}_k-\nabla G(x_{k+1})\right\|^2+2B_G^2\left\|\nabla g(\widehat{G}_k)-\nabla g(G(x_{k+1}))\right\|^2\\
\overset{③}{\le}\ &2b^2\left\|\nabla\widehat{G}_k-\nabla G(x_{k+1})\right\|^2+2B_G^2l^2\left\|\widehat{G}_k-G(x_{k+1})\right\|^2.
\end{aligned}\tag{3.1}$$

Here, ① uses the Young's inequality, ② follows from the boundedness of ∇g and ∇G, and ③ is from the l-smoothness of g.

Notice that $\widehat{G}_k$ and $\nabla\widehat{G}_k$ are respectively unbiased estimators of $G(x_{k+1})$ and $\nabla G(x_{k+1})$. Hence,

$$\begin{aligned}
\mathbb{E}_{k-1}\left[\left\|\widehat{G}_k-G(x_{k+1})\right\|^2\right]&=\mathbb{E}_{k-1}\left[\left\|\frac{1}{A}\sum_{j_k\in\mathcal{A}_k}\left(G_{j_k}(x_{k+1})-G_{j_k}(\widetilde{x}^s)\right)-\left(G(x_{k+1})-G(\widetilde{x}^s)\right)\right\|^2\right]\\
\overset{①}{=}&\frac{1}{A^2}\sum_{j_k\in\mathcal{A}_k}\mathbb{E}_{k-1}\left[\left\|\left(G_{j_k}(x_{k+1})-G_{j_k}(\widetilde{x}^s)\right)-\left(G(x_{k+1})-G(\widetilde{x}^s)\right)\right\|^2\right]\\
\overset{②}{\le}&\frac{1}{A^2}\sum_{j_k\in\mathcal{A}_k}\mathbb{E}_{k-1}\left[\|G_{j_k}(x_{k+1})-G_{j_k}(\widetilde{x}^s)\|^2\right]\overset{③}{\le}\frac{B_G^2}{A^2}\sum_{j_k\in\mathcal{A}_k}\|x_{k+1}-\widetilde{x}^s\|^2=\frac{B_G^2}{A}\|x_{k+1}-\widetilde{x}^s\|^2.
\end{aligned}\tag{3.2}$$

Here, ① comes from the fact that $\left\{\left(G_{j_k}(x_{k+1})-G_{j_k}(\widetilde{x}^s)\right)-\left(G(x_{k+1})-G(\widetilde{x}^s)\right)\right\}$ are conditionally independent with each other, and there expectations all equal 0, ② holds because the variance is bounded by the second moment, and ③ follows from the intermediate value theorem and the boundedness of the Jacobian of each G_{j_k}.

A completely parallel argument gives $\mathbb{E}_{k-1}\left[\left\|\nabla\widehat{G}_k-\nabla G(x_{k+1})\right\|^2\right]\le\frac{L_G^2}{B}\|x_{k+1}-\widetilde{x}^s\|^2$. Plugging this inequality and that in (3.2) into (3.1) leads to the desired result. □

The next lemma is from [1, Lemma 3.5].

Lemma 3.3. *Suppose $h(\cdot)$ is μ-strongly convex. Given $\widetilde{\nabla}_{k+1}$, if*

$$z_{k+1}=\arg\min_z\ \alpha\langle\widetilde{\nabla}_{k+1},z-z_k\rangle+\frac{1}{2}\|z-z_k\|^2+\alpha h(z)-\alpha h(z_k),$$

then it holds for any $u \in \mathbb{R}^{N_2}$ *that*

$$\alpha\langle\widetilde{\nabla}_{k+1}, z_{k+1} - u\rangle + \alpha h(z_{k+1}) - \alpha h(u) \le -\frac{1}{2}\|z_k - z_{k+1}\|^2 + \frac{1}{2}\|z_k - u\|^2 - \frac{1+\alpha\mu}{2}\|z_{k+1} - u\|^2. \tag{3.3}$$

The following lemma serves as a critical step in combining the progress of an entire iteration, enabled by the linear coupling update.

Lemma 3.4. *Let* x_{k+1}, y_{k+1} *and* z_{k+1} *be those given in Algorithm 1. If* $\tau_1 \in (0, \frac{1}{3\alpha L}]$ *and* $\tau_2 \in [0, 1-\tau_1]$ *in the linear coupling step, then for any* $u \in \mathbb{R}^{N_2}$ *and any positive* β*, it holds*

$$\begin{aligned}
&\alpha\big\langle\nabla f(x_{k+1}), z_k - u\big\rangle - \alpha h(u)\\
\le&\frac{\alpha}{\tau_1}\big(f(x_{k+1}) - H(y_{k+1})\big) + \alpha\Big(\frac{1}{4\tau_1 L} + \frac{\beta}{2}\Big)\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2 + \frac{1+\alpha/\beta}{2}\|z_k - u\|^2\\
&- \frac{1+\alpha\mu}{2}\|z_{k+1} - u\|^2 + \frac{\alpha\tau_2}{\tau_1}h(\widetilde{x}^s) + \frac{\alpha(1-\tau_1-\tau_2)}{\tau_1}h(y_k).
\end{aligned} \tag{3.4}$$

Proof. Let $v = \tau_1 z_{k+1} + \tau_2\widetilde{x}^s + (1-\tau_1-\tau_2)y_k$. We have $x_{k+1} - v = \tau_1(z_k - z_{k+1})$, and therefore

$$\begin{aligned}
&\alpha\langle\widetilde{\nabla}_{k+1}, z_k - z_{k+1}\rangle - \frac{1}{2}\|z_k - z_{k+1}\|^2 = \frac{\alpha}{\tau_1}\langle\widetilde{\nabla}_{k+1}, x_{k+1} - v\rangle - \frac{1}{2\tau_1^2}\|x_{k+1} - v\|^2\\
=&\frac{\alpha}{\tau_1}\left(\langle\widetilde{\nabla}_{k+1}, x_{k+1} - v\rangle - \frac{1}{2\alpha\tau_1}\|x_{k+1} - v\|^2 - h(v) + h(x_{k+1})\right) + \frac{\alpha}{\tau_1}\big(h(v) - h(x_{k+1})\big)\\
\overset{①}{\le}&\frac{\alpha}{\tau_1}\left(\langle\widetilde{\nabla}_{k+1}, x_{k+1} - v\rangle - \frac{3L}{2}\|x_{k+1} - v\|^2 - h(v) + h(x_{k+1})\right) + \frac{\alpha}{\tau_1}\big(h(v) - h(x_{k+1})\big)\\
\overset{②}{\le}&\frac{\alpha}{\tau_1}\left(H(x_{k+1}) - H(y_{k+1}) + \frac{1}{4L}\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2\right) + \frac{\alpha}{\tau_1}\big(h(v) - h(x_{k+1})\big)\\
\overset{③}{\le}&\frac{\alpha}{\tau_1}\left(H(x_{k+1}) - H(y_{k+1}) + \frac{1}{4L}\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2\right)\\
&\qquad + \frac{\alpha}{\tau_1}\big(\tau_1 h(z_{k+1}) + \tau_2 h(\widetilde{x}^s) + (1-\tau_1-\tau_2)h(y_k) - h(x_{k+1})\big).
\end{aligned} \tag{3.5}$$

Here ① holds because $\tau_1 \le \frac{1}{3\alpha L}$, ② uses Lemma 3.1, and ③ follows from the convexity of $h(\cdot)$ and the definition of v.

Meanwhile, for any vector u and any positive β,

$$\begin{aligned}
&\alpha\langle\nabla f(x_{k+1}), z_k - u\rangle\\
=&\,\alpha\langle\widetilde{\nabla}_{k+1}, z_k - u\rangle + \alpha\langle\nabla f(x_{k+1}) - \widetilde{\nabla}_{k+1}, z_k - u\rangle\\
\le&\,\alpha\langle\widetilde{\nabla}_{k+1}, z_k - z_{k+1}\rangle + \alpha\langle\widetilde{\nabla}_{k+1}, z_{k+1} - u\rangle + \frac{\alpha}{2}\left[\beta\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2 + \frac{1}{\beta}\|z_k - u\|^2\right],
\end{aligned} \tag{3.6}$$

where the inequality follows from the Young's inequality.

Over (3.6), we substitute in (3.5) and Lemma 3.3 and combine alike terms to get:

$$\begin{aligned}&\alpha\langle\nabla f(x_{k+1}), z_k - u\rangle - \alpha h(u)\\ \le&\frac{\alpha}{\tau_1}\big(H(x_{k+1}) - H(y_{k+1})\big) + \alpha\Big(\frac{1}{4\tau_1 L} + \frac{\beta}{2}\Big)\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2 + \frac{1+\alpha/\beta}{2}\|z_k - u\|^2\\ &- \frac{1+\alpha\mu}{2}\|z_{k+1} - u\|^2 + \frac{\alpha\tau_2}{\tau_1}h(\widetilde{x}^s) + \frac{\alpha(1-\tau_1-\tau_2)}{\tau_1}h(y_k) - \frac{\alpha}{\tau_1}h(x_{k+1})\end{aligned}$$

which finishes the proof upon simplification. □

Lemma 3.5. *Let x_{k+1} be given in the linear coupling step and x^* be the solution of (1.1). Then*

$$\|x_{k+1} - \widetilde{x}^s\|^2 \le 3\left(\tau_1^2\|z_k - x^*\|^2 + \frac{2(1-\tau_2)^2}{\mu}\big(H(\widetilde{x}^s) - H(x^*)\big) + \frac{2(1-\tau_1-\tau_2)^2}{\mu}\big(H(y_k) - H(x^*)\big)\right). \quad (3.7)$$

Proof. By the update formula of x_{k+1} and also the Young's inequality, we have

$$\begin{aligned}\|x_{k+1} - \widetilde{x}^s\|^2 &= \|\tau_1(z_k - x^*) + (1-\tau_2)(x^* - \widetilde{x}^s) + (1-\tau_1-\tau_2)(y_k - x^*)\|^2\\ &\le 3\left(\tau_1^2\|z_k - x^*\|^2 + (1-\tau_2)^2\|x^* - \widetilde{x}^s\|^2 + (1-\tau_1-\tau_2)^2\|y_k - x^*\|^2\right)\end{aligned}$$

Since H is μ-strongly convex, it holds that $H(u) - H(x^*) \ge \frac{\mu}{2}\|u - x^*\|^2$ for any u. Hence, we obtain (3.7) by bounding $\|x^* - \widetilde{x}^s\|^2$ and $\|y_k - x^*\|^2$ by the function values of H. □

Lemma 3.6. *Suppose $\tau_1 \in (0, \frac{1}{3\alpha L}]$ and $\tau_2 \in [0, 1-\tau_1]$ in the linear coupling step of Algorithm 1. Let x^* be the solution of (1.1). Then for any positive number β, it holds*

$$\begin{aligned}0 \le &\frac{\tau_2}{\tau_1}\big(H(\widetilde{x}^s) - H(x^*)\big) + \frac{(1-\tau_1-\tau_2)}{\tau_1}\big(H(y_k) - H(x^*)\big) - \frac{1}{\tau_1}\big(H(y_{k+1}) - H(x^*)\big) \quad (3.8)\\ &+ \left(\frac{1}{4\tau_1 L} + \frac{\beta}{2}\right)\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2 + \frac{1+\alpha/\beta}{2\alpha}\|z_k - x^*\|^2 - \frac{1+\alpha\mu}{2\alpha}\|z_{k+1} - x^*\|^2.\end{aligned}$$

Proof. For any u, we have

$$\begin{aligned}\alpha\big(f(x_{k+1}) - f(u)\big) &\overset{①}{\le} \alpha\big\langle\nabla f(x_{k+1}), x_{k+1} - u\big\rangle = \alpha\langle\nabla f(x_{k+1}), x_{k+1} - z_k\rangle + \alpha\langle\nabla f(x_{k+1}), z_k - u\rangle\\ &\overset{②}{=} \frac{\alpha\tau_2}{\tau_1}\langle\nabla f(x_{k+1}), \widetilde{x}^s - x_{k+1}\rangle + \frac{\alpha(1-\tau_1-\tau_2)}{\tau_1}\langle\nabla f(x_{k+1}), y_k - x_{k+1}\rangle + \alpha\langle\nabla f(x_{k+1}), z_k - u\rangle \quad (3.9)\\ &\overset{③}{\le} \frac{\alpha\tau_2}{\tau_1}\big(f(\widetilde{x}^s) - f(x_{k+1})\big) + \frac{\alpha(1-\tau_1-\tau_2)}{\tau_1}\big(f(y_k) - f(x_{k+1})\big) + \alpha\langle\nabla f(x_{k+1}), z_k - u\rangle. \quad (3.10)\end{aligned}$$

Here, ① uses the convexity of $f(\cdot)$, ② is by the definition of x_{k+1}, ③ uses the convexity of $f(\cdot)$ twice. Adding (3.4) and (3.10), we have

$$\begin{aligned}&\alpha\big(f(x_{k+1}) - H(u)\big)\\ \le&\frac{\alpha\tau_2}{\tau_1}\big(H(\widetilde{x}^s) - f(x_{k+1})\big) + \frac{\alpha(1-\tau_1-\tau_2)}{\tau_1}\big(H(y_k) - f(x_{k+1})\big) + \frac{\alpha}{\tau_1}\big(f(x_{k+1}) - H(y_{k+1})\big)\\ &+ \alpha\left(\frac{1}{4\tau_1 L} + \frac{\beta}{2}\right)\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2 + \frac{1+\alpha/\beta}{2}\|z_k - u\|^2 - \frac{1+\alpha\mu}{2}\|z_{k+1} - u\|^2.\end{aligned}$$

Setting $u = x^*$ in the above inequality and dividing both sides by α, we obtain the desired result by rearranging terms. □

By the inequality in (3.8) with $\beta = \frac{6}{\mu}$ and also bounding $\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2$, we will obtain the following inequality

$$0 \le \left(\frac{(1-\tau_1-\tau_2)}{\tau_1} + \frac{M}{\mu}2(1-\tau_1-\tau_2)^2\right)(H(y_k) - H(x^*)) - \frac{1}{\tau_1}\big(\mathbb{E}_{k-1}[H(y_{k+1})] - H(x^*)\big) \tag{3.11}$$
$$+ \left(\frac{\tau_2}{\tau_1} + \frac{M}{\mu}2(1-\tau_2)^2\right)(H(\widetilde{x}^s) - H(x^*)) + \left(\frac{1}{2\alpha} + \frac{\mu}{12} + M\tau_1^2\right)\|z_k - x^*\|^2 - \frac{1+\alpha\mu}{2\alpha}\,\mathbb{E}_{k-1}[\|z_{k+1} - x^*\|^2],$$

where M is a positive number to be defined. The proof of (3.11) differs slightly for **Option I** and **Option II** of Algorithm 1 by choosing appropriate M for different scenarios.

The following lemma is a key to establishing our main results.

Lemma 3.7. *Suppose that (3.11) holds for each k and s and that $2M\tau_1^2 < \frac{5\mu}{6}$. Let $\overline{A}, \overline{B}$ and $\overline{C}$ be numbers that satisfy*

$$\overline{A} \ge \frac{2M}{\mu}(1-\tau_1-\tau_2)^2,\ \overline{B} \ge \frac{2M}{\mu}(1-\tau_2)^2,\ \overline{C} = \frac{1}{2\alpha} + \frac{\mu}{12} + M\tau_1^2.$$

If $\theta \in (1, \frac{1+\alpha\mu}{1+\alpha(\mu/6+2M\tau_1^2)}]$, then

$$\left(\frac{(\tau_1+\tau_2-1+1/\theta)}{\tau_1} - \overline{A}\right)\theta\widetilde{D}^{s+1}\sum_{j=0}^{m-1}\theta^j \le \left(\frac{(1-\tau_1-\tau_2)}{\tau_1} + \overline{A}\right)\left(D_{sm} - \theta^m D_{(s+1)m}\right)$$
$$+ \left(\frac{\tau_2}{\tau_1} + \overline{B}\right)\widetilde{D}^s\sum_{j=0}^{m-1}\theta^j + \overline{C}\mathbb{E}\big[\|z_{sm} - x^*\|^2\big] - \overline{C}\theta^m\,\mathbb{E}\left[\|z_{(s+1)m} - x^*\|^2\right], \tag{3.12}$$

where $\widetilde{D}^s$ and D_{sm} are defined in (2.2).

Proof. Taking full expectation over both sides of (3.11) and using the definition of $\widetilde{D}^s$ and D_{sm} in (2.2), we have from the conditions on $\overline{A}, \overline{B}$ and $\overline{C}$ that

$$0 \le \left(\frac{(1-\tau_1-\tau_2)}{\tau_1} + \overline{A}\right)D_k - \frac{1}{\tau_1}D_{k+1} + \left(\frac{\tau_2}{\tau_1} + \overline{B}\right)\widetilde{D}^s + \overline{C}\mathbb{E}\big[\|z_k - x^*\|^2\big] - \overline{C}\theta\,\mathbb{E}\left[\|z_{k+1} - x^*\|^2\right].$$

Multiplying the above inequality by θ^j for each $k = sm + j$ and summing up the resulting inequalities for all $j = 0, \ldots, m-1$, we obtain

$$0 \le \left(\frac{(1-\tau_1-\tau_2)}{\tau_1} + \overline{A}\right)\sum_{j=0}^{m-1}\theta^j D_{sm+j} - \frac{1}{\tau_1}\sum_{j=0}^{m-1}\theta^j D_{sm+j+1} + \left(\frac{\tau_2}{\tau_1} + \overline{B}\right)\widetilde{D}^s\sum_{j=0}^{m-1}\theta^j$$
$$+ \overline{C}\mathbb{E}\big[\|z_{sm} - x^*\|^2\big] - \overline{C}\theta^m\,\mathbb{E}\left[\|z_{(s+1)m} - x^*\|^2\right].$$

which can be rewritten as

$$\left(\frac{(\tau_1+\tau_2-1+1/\theta)}{\tau_1} - \overline{A}\right)\sum_{j=1}^{m}\theta^j D_{sm+j} \le \left(\frac{(1-\tau_1-\tau_2)}{\tau_1} + \overline{A}\right)\left(D_{sm} - \theta^m D_{(s+1)m}\right)$$
$$+ \left(\frac{\tau_2}{\tau_1} + \overline{B}\right)\widetilde{D}^s\sum_{j=0}^{m-1}\theta^j + \overline{C}\mathbb{E}\big[\|z_{sm} - x^*\|^2\big] - \overline{C}\theta^m\,\mathbb{E}\left[\|z_{(s+1)m} - x^*\|^2\right].$$

By the convexity of $H(\cdot)$ and the choice of $\widetilde{x}^{s+1}$, we have $\widetilde{D}^{s+1} \le (\sum_{j=0}^{m-1}\theta^j)^{-1}\cdot\sum_{j=0}^{m-1}\theta^j D_{sm+j+1}$. Therefore, the above inequality implies (3.12), and we complete the proof. □

4 Convergence results of the SoCK method with outer batch step

In this section, we analyze the convergence rate of Algorithm 1 that takes **Option I** and estimate its complexity to produce a stochastic ε-solution of (1.1). More precisely, we assume n_1 is not too big, so we view the outer finite-sum as a single function F.

First, we show that (3.11) holds with an appropriate choice of M and thus (3.12) follows. Then we establish the convergence rate by using (3.12). The next lemma bounds the variance of $\widetilde{\nabla}_{k+1}$, and it follows from Lemma 3.2 with $g = F$.

Lemma 4.1. *Let f be that in (1.2), and let $\widetilde{x}^s$ and x_{k+1} be those given in Algorithm 1 with $\widetilde{\nabla}_{k+1}$ computed by **Option I**. Then*

$$\mathbb{E}_{k-1}\left[\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2\right] \le \left(\frac{2B_G^4L_F^2}{A} + \frac{2B_F^2L_G^2}{B}\right)\|\widetilde{x}^s - x_{k+1}\|^2. \tag{4.1}$$

Plugging (4.1) into (3.8), we are able to show (3.11) and thus (3.12) with an appropriate M.

Lemma 4.2. *Suppose $\tau_1 \in (0, \frac{1}{3\alpha L}]$ and $\tau_2 \in [0, 1-\tau_1]$ in the linear coupling step of Algorithm 1. Let x^* be the solution of (1.1). If $\widetilde{\nabla}_{k+1}$ is computed by **Option I**, then (3.11) holds with*

$$M = 3(\frac{1}{4\tau_1 L} + \frac{3}{\mu})\left(\frac{2B_G^4L_F^2}{A} + \frac{2B_F^2L_G^2}{B}\right). \tag{4.2}$$

Proof. Taking conditional expectation $\mathbb{E}_{k-1}$ on both sides of (3.8) with $\beta = \frac{6}{\mu}$ and plugging (4.1), we immediately have (3.11) by using the choice of M in (4.2). □

The next result is easy to show. Its proof is given in the appendix.

Lemma 4.3. *Let $\overline{A}, \overline{B} \in (0, \frac{1}{8}]$, $\tau_1 \in [\frac{1}{2m}, 1)$, $\tau_2 \le \tau_1$. If $\theta = 1 + \frac{1}{12m}$, then $\frac{(\tau_1+\tau_2-1+1/\theta-\overline{A}\tau_1)\theta}{\tau_2+\overline{B}\tau_1} > \frac{13}{12}$.*

Now we are ready to show our first main convergence rate result.

Theorem 4.4 (convergence rate with **Option I**). *Under Assumptions 2.1, 2.2 and 2.3, let $\{\widetilde{x}^s\}$ be generated from Algorithm 1 with $\widetilde{\nabla}_{k+1}$ computed by **Option I** and with parameters set as follows:*

$$m \leftarrow \left\lceil \frac{1}{2}\sqrt{\frac{L}{\mu}} \right\rceil,\ \tau_1 \leftarrow \frac{1}{2m},\ \tau_2 \leftarrow \frac{1}{2m},\ \theta \leftarrow 1 + \frac{1}{12m},\ \alpha \leftarrow \frac{1}{3\tau_1 L},\ A \leftarrow \frac{720B_G^4L_F^2}{\mu^2},\ B \leftarrow \frac{720B_F^2L_G^2}{\mu^2}. \tag{4.3}$$

Then

$$\mathbb{E}\left[H(\widetilde{x}^S) - H(x^*)\right] \le 11\cdot\left(\frac{12}{13}\right)^S\left(H(x_0) - H(x^*)\right).$$

Proof. First, notice $m^2 \geq \frac{L}{4\mu}$. Hence, $\alpha = \frac{1}{3\tau_1 L} = \frac{2m}{3L} \geq \frac{1}{6m\mu}$, and thus $\alpha\mu \geq \frac{1}{6m}$. Secondly, $m < \frac{1}{2}\sqrt{\frac{L}{\mu}} + 1 \leq \frac{3L}{2\mu}$, so $\alpha\mu < 1$. Thirdly, by the choices of A and B, the M given in (4.2) satisfies

$$M = \frac{\mu^2}{60}\left(\frac{1}{4\tau_1 L} + \frac{3}{\mu}\right) \leq \frac{\mu^2}{60}\left(\frac{3}{4\mu} + \frac{3}{\mu}\right) = \frac{\mu}{16}, \tag{4.4}$$

where the first inequality holds because $\frac{1}{\tau_1} = 2m < \frac{3L}{\mu}$. Therefore, $2M\tau_1^2 = \frac{M}{2m^2} \leq \frac{M}{2} < \frac{\mu}{6}$, and thus

$$\frac{1+\alpha\mu}{1+\alpha(\mu/6+2M\tau_1^2)} > \frac{1+\alpha\mu}{1+\alpha\mu/3} \geq 1 + \frac{\alpha\mu}{2} \geq 1 + \frac{1}{12m} = \theta, \tag{4.5}$$

where we have used $\alpha\mu < 1$ in the second inequality and $\alpha\mu \geq \frac{1}{6m}$ in the third inequality.

Let $\overline{A} = \frac{2M}{\mu}(1-\tau_1-\tau_2)^2$, $\overline{B} = \frac{2M}{\mu}(1-\tau_2)^2$, and $\overline{C} = \frac{1}{2\alpha} + \frac{\mu}{12} + M\tau_1^2$. Then by Lemma 4.2, all conditions required in Lemma 3.7 are satisfied. Hence, we have (3.12).

Since $\frac{2M}{\mu} \leq \frac{1}{8}$ from (4.4), we have $\overline{A} \leq \frac{1}{8}$ and $\overline{B} \leq \frac{1}{8}$, and thus by Lemma 4.3,

$$\frac{(\tau_1+\tau_2-1+1/\theta-\overline{A}\tau_1)\theta}{\tau_2+\overline{B}\tau_1} > \frac{13}{12}. \tag{4.6}$$

In addition, since $\theta = 1 + \frac{1}{12m}$, we have $\theta^m \geq \frac{13}{12}$. Therefore, (3.12) implies

$$\begin{aligned}
&\left(\frac{(1-\tau_1-\tau_2)}{\tau_1} + \overline{A}\right) D_{(s+1)m} + \left(\frac{\tau_2}{\tau_1} + \overline{B}\right)\widetilde{D}^{s+1}\sum_{j=0}^{m-1}\theta^j + \overline{C}\mathbb{E}\left[\|z_{(s+1)m} - x^*\|^2\right] \\
&\qquad \leq \frac{12}{13}\left(\left(\frac{(1-\tau_1-\tau_2)}{\tau_1} + \overline{A}\right) D_{sm} + \left(\frac{\tau_2}{\tau_1} + \overline{B}\right)\widetilde{D}^{s}\sum_{j=0}^{m-1}\theta^j + \overline{C}\mathbb{E}\left[\|z_{sm} - x^*\|^2\right]\right).
\end{aligned} \tag{4.7}$$

Repeatedly using (4.7) for $s = 0$ through $s = S-1$ gives

$$\left(\frac{\tau_2}{\tau_1} + \overline{B}\right)\widetilde{D}^S\sum_{j=0}^{m-1}\theta^j \leq \left(\frac{12}{13}\right)^S\left(\left(\frac{(1-\tau_1-\tau_2)}{\tau_1} + \overline{A}\right) D_0 + \left(\frac{\tau_2}{\tau_1} + \overline{B}\right)\widetilde{D}^0\sum_{j=0}^{m-1}\theta^j + \overline{C}\|z_0 - x^*\|^2\right).$$

Dividing by $\left(\frac{\tau_2}{\tau_1} + \overline{B}\right)\sum_{j=0}^{m-1}\theta^j$ both sides of the above inequality, we have

$$\widetilde{D}^S \leq \left(\frac{12}{13}\right)^S\left(\frac{\frac{(1-\tau_1-\tau_2)}{\tau_1} + \overline{A}}{\left(\frac{\tau_2}{\tau_1} + \overline{B}\right)\sum_{j=0}^{m-1}\theta^j} D_0 + \widetilde{D}^0 + \frac{\overline{C}}{\left(\frac{\tau_2}{\tau_1} + \overline{B}\right)\sum_{j=0}^{m-1}\theta^j}\|z_0 - x^*\|^2\right). \tag{4.8}$$

Notice $\sum_{j=0}^{m-1}\theta^j \geq m$, $\tau_1 = \tau_2 = \frac{1}{2m}$, $\overline{A} \leq \frac{1}{8}$, and $\overline{B} > 0$, and thus we have

$$\frac{\frac{(1-\tau_1-\tau_2)}{\tau_1} + \overline{A}}{\left(\frac{\tau_2}{\tau_1} + \overline{B}\right)\sum_{j=0}^{m-1}\theta^j} \leq 2. \tag{4.9}$$

In addition, recall $M \le \frac{\mu}{16}$ from (4.4). Hence, $\overline{C} = \frac{1}{2\alpha} + \frac{\mu}{12} + M\tau_1^2 = \frac{1}{2\alpha} + \frac{\mu}{12} + \frac{\mu}{4m^2} \le \frac{1}{2\alpha}\left(1 + \frac{\alpha\mu}{3}\right)$, and thus

$$\frac{\overline{C}}{\left(\frac{\tau_2}{\tau_1} + \overline{B}\right)\sum_{j=0}^{m-1}\theta^j} \le \frac{1+\alpha\mu/3}{2m\alpha}. \tag{4.10}$$

Moreover, by the μ-strong convexity of H, it holds $H(x_0) - H(x^*) \ge \frac{\mu}{2}\|x_0 - x^*\|^2 = \frac{\mu}{2}\|z_0 - x^*\|^2$. Plugging this inequality and also (4.9) and (4.10) into (4.8), we obtain by recalling the definition of $\widetilde{D}^s$ and D_k in (2.2) that

$$\mathbb{E}\left[H(\widetilde{x}^S) - H(x^*)\right] \le \left(\frac{12}{13}\right)^S \left(3 + \frac{1+\alpha\mu/3}{m\alpha\mu}\right)\left(H(x_0) - H(x^*)\right).$$

Since $\frac{1}{m\alpha\mu} \le 6$ and $\alpha\mu < 1$ as we showed at the beginning of the proof, it holds $\frac{1+\alpha\mu/3}{m\alpha\mu} \le 8$, which together with the above inequality gives the desired result. □

By Theorem 4.4, we can estimate the complexity of Algorithm 1 in terms of the number of evaluations on ∇F_i, G_j, ∇G_j, and the proximal mapping of h.

Corollary 4.5 (complexity result with **Option I**). *Given $\varepsilon > 0$, under the same assumptions as in Theorem 4.4, the complexity of Algorithm 1 to obtain a stochastic ε-solution is*

$$O\left(\left(n_1 + n_2 + \sqrt{\frac{L}{\mu}}\left(\frac{\max\{B_G^4 L_F^2, B_F^2 L_G^2\}}{\mu^2} + n_1\right)\right)\log\frac{H(x_0) - H(x^*)}{\varepsilon}\right). \tag{4.11}$$

Proof. To produce one snapshot point $\widetilde{x}^s$, the complexity is $O\big(n_1 + n_2 + m(A + B + n_1)\big) = O\Big(n_1 + n_2 + \sqrt{\frac{L}{\mu}}\big(\frac{\max\{B_G^4 L_F^2, B_F^2 L_G^2\}}{\mu^2} + n_1\big)\Big)$. In addition, by Theorem 4.4, to have $\mathbb{E}\left[H(\widetilde{x}^S) - H(x^*)\right] \le \varepsilon$, it suffices to have $S = O\left(\log\frac{H(x_0)-H(x^*)}{\varepsilon}\right)$. Hence, the total complexity is $O\big(S(n_1 + n_2 + m(A + B + n_1))\big)$, which is the desired result in (4.11). □

Remark 1 From the complexity result in (4.11), we can easily see that when $n_1 = O\left(\frac{\max\{B_G^4 L_F^2, B_F^2 L_G^2\}}{\mu^2}\right)$, the result will not become better even if we replace $\nabla F(\widehat{G}_k)$ in **Option I** by its stochastic approximation with a mini-batch sampling. Hence, in this case of relatively small n_1, we should always take **Option I** in Algorithm 1.

5 Convergence results of the SoCK method with outer minibatch step

In this section, we assume that both of n_1 and n_2 are very big, and we analyze the convergence rate of Algorithm 1 that takes **Option II** and estimate its complexity to produce a stochastic ε-solution of (1.1). Depending on whether the following assumption is satisfied, we choose the output as either $\widetilde{x}^S$ or x^{out} that is defined in the last line of Algorithm 1.

Assumption 5.1. *For each $i \in [n_1]$, $f_i(x) = F_i(G(x)) = F_i\left(\frac{1}{n_2}\sum_{j=1}^{n_2} G_j(x)\right)$ is convex.*

5.1 Analysis without Assumption 5.1

In this subsection, we analyze the algorithm without assuming Assumption 5.1. Our analysis shares a similar flow as that in the previous section. We first bound $\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2$ and show that (3.11) holds with an appropriate choice of M.

Lemma 5.1. *If Assumptions 2.2 and 2.3 are satisfied, and $\widetilde{\nabla}_{k+1}$ is computed by **Option II** in Algorithm 1, then*

$$\mathbb{E}_{k-1}\left[\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2\right] \leq \left(\frac{4B_G^4 L_F^2}{A} + \frac{4B_F^2 L_G^2}{B} + \frac{2L^2}{C}\right)\|\widetilde{x}^s - x_{k+1}\|^2. \tag{5.1}$$

Proof. Define

$$U_{k+1} = \frac{1}{C}\sum_{i\in\mathcal{C}_k}\left(\left[\nabla G(x_{k+1})\right]^\top \nabla F_i(G(x_{k+1})) - \left[\nabla G(\widetilde{x}^s)\right]^\top \nabla F_i(G(\widetilde{x}^s))\right) + \nabla f(\widetilde{x}^s). \tag{5.2}$$

By the Young's inequality, it holds that

$$\mathbb{E}_{k-1}\left[\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2\right] \leq 2\,\mathbb{E}_{k-1}\left[\|\widetilde{\nabla}_{k+1} - U_{k+1}\|^2\right] + 2\,\mathbb{E}_{k-1}\left[\|U_{k+1} - \nabla f(x_{k+1})\|^2\right]. \tag{5.3}$$

By the definition of $\widetilde{\nabla}_{k+1}$ in **Option II** of Algorithm 1 and the definition of U_{k+1} in (5.2), we have

$$\begin{aligned}\mathbb{E}_{k-1}\left[\|\widetilde{\nabla}_{k+1} - U_{k+1}\|^2\right] &= \frac{1}{C^2}\,\mathbb{E}_{k-1}\left[\left\|\sum_{i\in\mathcal{C}_k}\left(\left[\nabla \widehat{G}_k\right]^\top \nabla F_i(\widehat{G}_k) - \left[\nabla G(x_{k+1})\right]^\top \nabla F_i\big(G(x_{k+1})\big)\right)\right\|^2\right] \\ &\leq \frac{1}{C}\sum_{i\in\mathcal{C}_k}\mathbb{E}_{k-1}\left[\left\|\left[\nabla \widehat{G}_k\right]^\top \nabla F_i(\widehat{G}_k) - \left[\nabla G(x_{k+1})\right]^\top \nabla F_i\big(G(x_{k+1})\big)\right\|^2\right].\end{aligned}$$

Applying Lemma 3.2 with $g = F_i$ for each $i \in \mathcal{C}_k$ to the right-hand-side (r.h.s.) of the above inequality gives

$$\mathbb{E}_{k-1}\left[\|\widetilde{\nabla}_{k+1} - U_{k+1}\|^2\right] \leq 2\left(\frac{B_G^4 L_F^2}{A} + \frac{B_F^2 L_G^2}{B}\right)\|\widetilde{x}^s - x_{k+1}\|^2. \tag{5.4}$$

For the second term of the r.h.s. of (5.3), we have

$$\begin{aligned}&\mathbb{E}_{k-1}\left[\|U_{k+1} - \nabla f(x_{k+1})\|^2\right] \\ &= \frac{1}{C^2}\,\mathbb{E}_{k-1}\left[\left\|\sum_{i\in\mathcal{C}_k}\left(\left[\nabla G(x_{k+1})\right]^\top \nabla F_i(G(x_{k+1})) - \left[\nabla G(\widetilde{x}^s)\right]^\top \nabla F_i(G(\widetilde{x}^s)) - \nabla f(x_{k+1}) + \nabla f(\widetilde{x}^s)\right)\right\|^2\right] \\ &= \frac{1}{C^2}\,\mathbb{E}_{k-1}\left[\sum_{i\in\mathcal{C}_k}\left\|\left[\nabla G(x_{k+1})\right]^\top \nabla F_i(G(x_{k+1})) - \left[\nabla G(\widetilde{x}^s)\right]^\top \nabla F_i(G(\widetilde{x}^s)) - \nabla f(x_{k+1}) + \nabla f(\widetilde{x}^s)\right\|^2\right],\end{aligned} \tag{5.5}$$

where the second equality holds because the summands are conditionally independent and each has a zero mean. Since the variance of a random vector is bounded by its second moment, we have for each $i \in \mathcal{C}_k$ that

$$\begin{aligned}
&\mathbb{E}_{k-1}\left[\left\|\left[\nabla G(x_{k+1})\right]^\top \nabla F_i(G(x_{k+1})) - \left[\nabla G(\widetilde{x}^s)\right]^\top \nabla F_i(G(\widetilde{x}^s)) - \nabla f(x_{k+1}) + \nabla f(\widetilde{x}^s)\right\|^2\right] \\
\le\ &\mathbb{E}_{k-1}\left[\left\|\left[\nabla G(x_{k+1})\right]^\top \nabla F_i(G(x_{k+1})) - \left[\nabla G(\widetilde{x}^s)\right]^\top \nabla F_i(G(\widetilde{x}^s))\right\|^2\right] \qquad (5.6)\\
\le\ &L^2\|\widetilde{x}^s - x_{k+1}\|^2, \qquad (5.7)
\end{aligned}$$

where the second inequality follows from (2.1). Substituting (5.7) into (5.5) yields

$$\mathbb{E}_{k-1}\left[\|U_{k+1} - \nabla f(x_{k+1})\|^2\right] \le \frac{L^2}{C}\|\widetilde{x}^s - x_{k+1}\|^2. \qquad (5.8)$$

We obtain the desired result by plugging (5.4) and (5.8) into (5.3). □

Plugging (5.1) into (3.8), we are able to show (3.11) and thus (3.12) with an appropriate M.

Lemma 5.2. *Suppose $\tau_1 \in (0, \frac{1}{3\alpha L}]$ and $\tau_2 \in [0, 1-\tau_1]$ in the linear coupling step of Algorithm 1. Let x^* be the solution of (1.1). If $\widetilde{\nabla}_{k+1}$ is computed by **Option II**, then (3.11) holds with*

$$M = 3\Big(\frac{1}{4\tau_1 L} + \frac{3}{\mu}\Big)\left(\frac{4B_G^4 L_F^2}{A} + \frac{4B_F^2 L_G^2}{B} + \frac{2L^2}{C}\right). \qquad (5.9)$$

Proof. Taking conditional expectation $\mathbb{E}_{k-1}$ on both sides of (3.8) with $\beta = \frac{6}{\mu}$ and plugging (5.1), we immediately have (3.11) by using the choice of M in (5.9). □

We are now to show our second main result.

Theorem 5.3 (convergence result for **Option II** without Assumption 5.1)**.** *Under Assumptions 2.1, 2.2 and 2.3, let $\{\widetilde{x}^s\}$ be generated from Algorithm 1 with $\widetilde{\nabla}_{k+1}$ computed by **Option II** and with parameters set as follows:*

$$m \leftarrow \left\lceil \frac{1}{2}\sqrt{\frac{L}{\mu}} \right\rceil,\ \tau_1 \leftarrow \frac{1}{2m},\ \tau_2 \leftarrow \frac{1}{2m},\ \theta \leftarrow 1 + \frac{1}{12m},\ \alpha \leftarrow \frac{1}{3\tau_1 L}, \qquad (5.10a)$$

$$A \leftarrow \frac{2160 B_G^4 L_F^2}{\mu^2},\ B \leftarrow \frac{2160 B_F^2 L_G^2}{\mu^2},\ C \leftarrow \frac{1080 L^2}{\mu^2}. \qquad (5.10b)$$

Then

$$\mathbb{E}\left[H(\widetilde{x}^S) - H(x^*)\right] \le 11 \cdot \left(\frac{12}{13}\right)^S \big(H(x_0) - H(x^*)\big).$$

Proof. Notice that the parameters given in (5.10) are the same as those in (4.3) except for A, B and C, and also notice that the choices of A, B and C only affect the value of M. Plugging into (5.9) the values of A, B, C given in (5.10b), we can easily verify that $M \le \frac{\mu}{16}$. Now following the same arguments in the proof of Theorem 4.4, we obtain the desired result. □

By Theorem 5.3, we can estimate the complexity of Algorithm 1 in terms of the number of evaluations on ∇F_i, G_j, ∇G_j, and the proximal mapping of h. Its proof follows that of Corollary 4.5, and we omit it.

Corollary 5.4. *Given $\varepsilon > 0$, under the same assumptions as in Theorem 5.3, the complexity of Algorithm 1 to produce a stochastic ε-solution is*

$$O\left(\left(n_1+n_2+\sqrt{\frac{L}{\mu}}\frac{\max\{B_G^4L_F^2,\, B_F^2L_G^2,\, L^2\}}{\mu^2}\right)\log\frac{H(x_0)-H(x^*)}{\varepsilon}\right).$$

5.2 Analysis with Assumption 5.1

In this subsection, we assume Assumption 5.1 and establish a better result in certain regimes. Again, we first bound $\|\widetilde{\nabla}_{k+1}-\nabla f(x_{k+1})\|^2$ and show that (3.11) holds with an appropriate choice of M.

Lemma 5.5. *If Assumptions 2.2, 2.3 and 5.1 are satisfied, and $\widetilde{\nabla}_{k+1}$ is computed by **Option II** in Algorithm 1, then*

$$\mathbb{E}_{k-1}[\|\widetilde{\nabla}_{k+1}-\nabla f(x_{k+1})\|^2] \le 4\left(\frac{B_G^4L_F^2}{A}+\frac{B_F^2L_G^2}{B}\right)\|\widetilde{x}^s-x_{k+1}\|^2 + \frac{4L}{C}\big[f(\widetilde{x}^s)-f(x_{k+1})-\langle\nabla f(x_{k+1}),\widetilde{x}^s-x_{k+1}\rangle\big]. \quad (5.11)$$

Proof. Plugging (5.6) into (5.5) and using the definition of $f_i = F_i\circ G$, we have

$$\mathbb{E}_{k-1}\left[\|U_{k+1}-\nabla f(x_{k+1})\|^2\right] \le \frac{1}{C^2}\mathbb{E}_{k-1}\left[\sum_{i\in\mathcal{C}_k}\|\nabla f_i(x_{k+1})-\nabla f_i(\widetilde{x}^s)\|^2\right]. \quad (5.12)$$

Since each f_i is L-smooth according to Assumption 2.3 and convex according to Assumption 5.1, we apply [16, Theorem 2.1.5] to have

$$\|\nabla f_i(\widetilde{x}^s)-\nabla f_i(x_{k+1})\|^2 \le 2L\big[f_i(\widetilde{x}^s)-f_i(x_{k+1})-\langle\nabla f_i(x_{k+1}),\widetilde{x}^s-x_{k+1}\rangle\big],$$

and hence for each $i\in\mathcal{C}_k$,

$$\mathbb{E}_{k-1}\left[\|\nabla f_i(\widetilde{x}^s)-\nabla f_i(x_{k+1})\|^2\right] \le 2L\big[f(\widetilde{x}^s)-f(x_{k+1})-\langle\nabla f(x_{k+1}),\widetilde{x}^s-x_{k+1}\rangle\big].$$

Substituting the above inequality into (5.12) gives

$$\mathbb{E}_{k-1}\left[\|U_{k+1}-\nabla f(x_{k+1})\|^2\right] \le \frac{2L}{C}\big[f(\widetilde{x}^s)-f(x_{k+1})-\langle\nabla f(x_{k+1}),\widetilde{x}^s-x_{k+1}\rangle\big]. \quad (5.13)$$

Now plugging (5.4) and (5.13) into (5.3), we obtain the desired result. □

The next lemma shows that with appropriate τ_1,τ_2 and M, (3.11) holds.

Lemma 5.6. *Suppose $\tau_1\in(0,\frac{1}{3\alpha L}]$ and let $C>0$ be chosen such that $\tau_2=(\frac{1}{L}+\frac{12\tau_1}{\mu})\frac{L}{C}\in[0,1-\tau_1]$. Let x^* be the solution of (1.1). If $\widetilde{\nabla}_{k+1}$ is computed by **Option II**, then (3.11) holds with*

$$M = 12\Big(\frac{1}{4\tau_1 L}+\frac{3}{\mu}\Big)\left(\frac{B_G^4L_F^2}{A}+\frac{B_F^2L_G^2}{B}\right). \quad (5.14)$$

Proof. We start from (3.9) and use the convexity of f to have

$$\alpha(f(x_{k+1})-f(u)) \le \frac{\alpha\tau_2}{\tau_1}\langle \nabla f(x_{k+1}), \widetilde{x}^s - x_{k+1}\rangle + \frac{\alpha(1-\tau_1-\tau_2)}{\tau_1}(f(y_k)-f(x_{k+1})) + \alpha\langle \nabla f(x_{k+1}), z_k - u\rangle.$$

Adding (3.4) to the above inequality gives

$$\begin{aligned}
&\alpha(f(x_{k+1})-H(u))\\
\le&\frac{\alpha\tau_2}{\tau_1}\langle \nabla f(x_{k+1}), \widetilde{x}^s - x_{k+1}\rangle + \frac{\alpha(1-\tau_1-\tau_2)}{\tau_1}\big(H(y_k)-f(x_{k+1})\big) + \frac{\alpha}{\tau_1}\big(f(x_{k+1})-H(y_{k+1})\big)\\
&+\alpha\Big(\frac{1}{4\tau_1 L}+\frac{\beta}{2}\Big)\|\widetilde{\nabla}_{k+1}-\nabla f(x_{k+1})\|^2 + \frac{1+\alpha/\beta}{2}\|z_k-u\|^2 - \frac{1+\alpha\mu}{2}\|z_{k+1}-u\|^2 + \frac{\alpha\tau_2}{\tau_1}h(\widetilde{x}^s),
\end{aligned}$$

where β is an arbitrary positive number. Taking conditional expectation $\mathbb{E}_{k-1}$ on both sides of the above inequality and also plugging (5.11), we have

$$\begin{aligned}
&\alpha\big(f(x_{k+1})-H(u)\big)\\
\le&\frac{\alpha\tau_2}{\tau_1}\langle \nabla f(x_{k+1}), \widetilde{x}^s - x_{k+1}\rangle + \frac{\alpha(1-\tau_1-\tau_2)}{\tau_1}\big(H(y_k)-f(x_{k+1})\big) + \frac{\alpha}{\tau_1}\big(f(x_{k+1})-\mathbb{E}_{k-1}[H(y_{k+1})]\big)\\
&+4\alpha\Big(\frac{1}{4\tau_1 L}+\frac{\beta}{2}\Big)\left(\Big(\frac{B_G^4L_F^2}{A}+\frac{B_F^2L_G^2}{B}\Big)\|\widetilde{x}^s-x_{k+1}\|^2 + \frac{L}{C}\left[f(\widetilde{x}^s)-f(x_{k+1})-\langle\nabla f(x_{k+1}),\widetilde{x}^s-x_{k+1}\rangle\right]\right)\\
&+\frac{1+\alpha/\beta}{2}\|z_k-u\|^2 - \frac{1+\alpha\mu}{2}\,\mathbb{E}_{k-1}\left[\|z_{k+1}-u\|^2\right] + \frac{\alpha\tau_2}{\tau_1}h(\widetilde{x}^s).
\end{aligned}$$

Let $\beta=\frac{6}{\mu}$. Since $\frac{\tau_2}{\tau_1}=(\frac{1}{\tau_1 L}+\frac{12}{\mu})\frac{L}{C}$, it follows from the above inequality that

$$\begin{aligned}
&\alpha\big(f(x_{k+1})-H(u)\big)\\
\le&\frac{\alpha\tau_2}{\tau_1}\big(H(\widetilde{x}^s)-f(x_{k+1})\big) + \frac{\alpha(1-\tau_1-\tau_2)}{\tau_1}\big(H(y_k)-f(x_{k+1})\big) + \frac{\alpha}{\tau_1}\big(f(x_{k+1})-\mathbb{E}_{k-1}[H(y_{k+1})]\big)\\
&+\alpha\Big(\frac{1}{\tau_1 L}+\frac{12}{\mu}\Big)\left(\frac{B_G^4L_F^2}{A}+\frac{B_F^2L_G^2}{B}\right)\|\widetilde{x}^s-x_{k+1}\|^2 + \frac{1+\alpha\mu/6}{2}\|z_k-u\|^2 - \frac{1+\alpha\mu}{2}\,\mathbb{E}_{k-1}[\|z_{k+1}-u\|^2].
\end{aligned}$$

Now setting $u=x^*$, dividing both sides by α, and rearranging terms of the above inequality, we have

$$\begin{aligned}
0\le&\frac{\tau_2}{\tau_1}\big(H(\widetilde{x}^s)-H(x^*)\big) + \frac{(1-\tau_1-\tau_2)}{\tau_1}\big(H(y_k)-H(x^*)\big) - \frac{1}{\tau_1}\left(\mathbb{E}_{k-1}\left[H(y_{k+1})\right]-H(x^*)\right)\\
&+\Big(\frac{1}{\tau_1 L}+\frac{12}{\mu}\Big)\left(\frac{B_G^4L_F^2}{A}+\frac{B_F^2L_G^2}{B}\right)\|\widetilde{x}^s-x_{k+1}\|^2 + \frac{1+\alpha\mu/6}{2\alpha}\|z_k-x^*\|^2 - \frac{1+\alpha\mu}{2\alpha}\,\mathbb{E}_{k-1}\left[\|z_{k+1}-x^*\|^2\right].
\end{aligned}$$

Finally, we substitute (3.7) into the above and use the definition of M in (5.14) to obtain (3.11). □

By Lemmas 5.6 and 3.7, we have the key inequality (3.12) by choosing appropriate parameters. Throughout the rest of this section, we let

$$\tau = \frac{1}{C}\left(1+\frac{12L}{\mu}\right). \tag{5.15}$$

Lemma 5.7. *For any integer $m\ge 1$ and any $t\in[0,\frac{1}{2m}]$, we have $\frac{1}{2}((1+t)^{m-1}-1)\le(m-1)t$.*

Proof. When $m = 1$, the result holds trivially. For $m \geq 2$, it is equivalent to showing

$$\phi(t) := (m-1)\log(1+t) - \log\big(2(m-1)t+1\big) \leq 0,\ \forall t \in [0, \frac{1}{2m}].$$

It is easy to see $\phi(0) = 0$ and also not difficult to verify $\phi'(t) < 0$, $\forall t \in [0, \frac{1}{2m}]$. Hence, $\phi(t) \leq 0$ for any $t \in [0, \frac{1}{2m}]$, and we complete the proof. □

Lemma 5.8. *If* $0 < \theta - 1 \leq \frac{\alpha\mu}{2} \leq \frac{1}{12m}$, $0 \leq \tau_2 \leq \tau$, $3\alpha\mu \leq \tau_1$, $\tau\alpha\mu \leq \frac{\tau_1}{6m}$, *and* $\overline{A} = \overline{B} = \frac{1}{8}$, *then* $\frac{(\tau_1+\tau_2-1+1/\theta)}{\tau_1} - \overline{A} \geq \left(\frac{\tau_2}{\tau_1} + \overline{B}\right)\theta^{m-1}$.

We are now ready to show our third main result.

Theorem 5.9 (convergence result for **Option II** with Assumption 5.1)**.** *Under Assumptions 2.1, 2.2, 2.3 and 5.1, let* $\{\widetilde{x}^s\}$ *be generated from Algorithm 1 with* $\widetilde{\nabla}_{k+1}$ *computed by* ***Option II*** *and with* $m \leq \frac{L}{2\mu}$, $C \geq 2\big(1 + \frac{12L}{\mu}\big)$ *and other parameters set as follows:*

$$\tau_1 = \begin{cases} \min\left\{\sqrt{\frac{2m\mu}{L}}\tau,\ \tau\right\}, & \text{if } m \geq \frac{1}{2\tau} \\ \min\left\{\sqrt{\frac{\mu}{L}}, \frac{1}{2m}\right\}, & \text{otherwise,} \end{cases} \tag{5.16a}$$

$$\alpha = \frac{1}{3\tau_1 L},\ \tau_2 = \frac{1}{C}\big(1 + \frac{12\tau_1 L}{\mu}\big),\ A = \frac{1248B_G^4L_F^2}{\mu^2},\ B = \frac{1248B_F^2L_G^2}{\mu^2}, \tag{5.16b}$$

$$\theta = 1 + \min\left\{\frac{\alpha\mu}{2}, \frac{1}{12m}\right\} \tag{5.16c}$$

Then $\mathbb{E}\big[H(x^{\text{out}}) - H(x^*)\big] \leq \eta\big[H(x_0) - H(x^*)\big]$ *with*

$$\eta = \begin{cases} \big(1 + \sqrt{\frac{\mu}{72mL\tau}}\big)^{-Sm}\left(3 + 20\sqrt{\frac{L\tau}{m\mu}}\right) & \text{if } \frac{\tau}{2} \geq \frac{m\mu}{L} \text{ and } \tau \geq \frac{1}{2m}, \\ 7\left(1 + \sqrt{\frac{\mu}{36L}}\right)^{-Sm} & \text{if } \sqrt{\frac{\mu}{L}} \leq \frac{1}{2m} \text{ and } \tau < \frac{1}{2m}, \\ 30\left(\frac{12}{13}\right)^{S} & \text{otherwise.} \end{cases} \tag{5.17}$$

Proof. Since $C \geq 2\big(1 + \frac{12L}{\mu}\big)$, it holds $\tau \leq \frac{1}{2}$ from (5.15), and thus $\tau_1 \leq 1$ from (5.16a) and $\tau_2 \leq \tau$ from (5.16b). In addition, if $m \geq \frac{1}{2\tau}$, then $\sqrt{\frac{L}{2m\mu\tau}} \leq \sqrt{\frac{L}{\mu}}$, and also $\frac{1}{\tau} \leq 2m \leq \frac{L}{\mu}$ from the condition $m \leq \frac{L}{2\mu}$. Hence, $\frac{1}{\tau_1} = \max\left\{\sqrt{\frac{L}{2m\mu\tau}}, \frac{1}{\tau}\right\} \leq \frac{L}{\mu}$. On the other hand, if $m < \frac{1}{2\tau}$, then $\frac{1}{\tau_1} = \max\left\{\sqrt{\frac{L}{\mu}}, 2m\right\} \leq \frac{L}{\mu}$. Therefore, it always holds that $\frac{1}{\tau_1} \leq \frac{L}{\mu}$, and thus $\alpha\mu = \frac{\mu}{3\tau_1 L} \leq \frac{1}{3}$. Furthermore, with the choices of A and B, the M given in (5.14) satisfies

$$M = \frac{\mu^2}{52}\Big(\frac{1}{4\tau_1 L} + \frac{3}{\mu}\Big) \leq \frac{\mu^2}{52}\Big(\frac{1}{4L}\frac{L}{\mu} + \frac{3}{\mu}\Big) = \frac{\mu}{16}. \tag{5.18}$$

Finally, notice $M\tau_1^2 \le M \le \frac{\mu}{16}$, and by the same arguments in (4.5), we have

$$\theta < \frac{1+\alpha\mu}{1+\alpha(\mu/6+2M\tau_1^2)}. \tag{5.19}$$

We now choose

$$\overline{A} = \overline{B} = \frac{1}{8}, \text{ and } \overline{C} = \frac{1}{2\alpha} + \frac{\mu}{12} + M\tau_1^2 \le \frac{1}{2\alpha}\big(1+\frac{\alpha\mu}{3}\big) \le \frac{5}{9\alpha}, \tag{5.20}$$

where we have used $\alpha\mu \le \frac{1}{3}$ in the last inequality. It follows

$$\frac{\tau_1\overline{C}}{(\tau_2+\tau_1/8)m} \le \frac{8\overline{C}}{m} \le \frac{40}{9m\alpha} \tag{5.21}$$

Since $\frac{2M}{\mu} \le \frac{1}{8}$ from (5.18), by Lemma 5.6, all the conditions required by Lemma 3.7 are satisfied, and thus we have (3.12).

By the convexity of H, $\sum_{j=0}^{m-1}\theta^j \ge m$, and also the definition of D_k and $\widetilde{D}^s$ in (2.2), it holds

$$\mathbb{E}\big[H(\widetilde{x}^{\text{out}}) - H(x^*)\big] \le \frac{(1-\tau_1-\tau_2+\tau_1/8)D_{Sm} + (\tau_2+\tau_1/8)\widetilde{D}^S\sum_{j=0}^{m-1}\theta^j}{1-\tau_1-\tau_2+\tau_1/8+(\tau_2+\tau_1/8)m}. \tag{5.22}$$

The rest of the proof is conducted under four cases.

Case 1. Suppose $\tau \ge \frac{1}{2m}$ and $\frac{m\mu}{L} \le \frac{\tau}{2}$.

In this case, we have $\tau_1 = \sqrt{\frac{2m\mu\tau}{L}} \le \frac{1}{2}$ and $\alpha = \frac{1}{\sqrt{18mL\mu\tau}}$. Hence, $\alpha\mu = \frac{1}{m}\sqrt{\frac{m\mu}{18L\tau}} \le \frac{1}{6m}$, and $\theta = 1+\frac{\alpha\mu}{2}$. It is straightforward to check $\tau_1 = 6m\alpha\mu\tau$. In addition, from $\tau \ge \frac{1}{2m}$, it follows that $\tau_1 \ge 3\alpha\mu$. Hence, all conditions in Lemma 5.8 are satisfied. Therefore, $\frac{(\tau_1+\tau_2-1+1/\theta)}{\tau_1} - \overline{A} \ge \left(\frac{\tau_2}{\tau_1}+\overline{B}\right)\theta^{m-1}$, and thus (3.12) implies

$$\begin{aligned}
&\left(\frac{(1-\tau_1-\tau_2)}{\tau_1}+\overline{A}\right)D_{(s+1)m} + \left(\frac{\tau_2}{\tau_1}+\overline{B}\right)\widetilde{D}^{s+1}\sum_{j=0}^{m-1}\theta^j + \overline{C}\mathbb{E}\big[\|z_{(s+1)m}-x^*\|^2\big] \\
&\qquad \le \theta^{-m}\left(\left(\frac{(1-\tau_1-\tau_2)}{\tau_1}+\overline{A}\right)D_{sm} + \left(\frac{\tau_2}{\tau_1}+\overline{B}\right)\widetilde{D}^{s}\sum_{j=0}^{m-1}\theta^j + \overline{C}\mathbb{E}\big[\|z_{sm}-x^*\|^2\big]\right).
\end{aligned} \tag{5.23}$$

Repeatedly using the above inequality for $s=0$ through $s=S-1$ and plugging $\overline{A}=\overline{B}=\frac{1}{8}$ gives

$$\begin{aligned}
&\left(\frac{(1-\tau_1-\tau_2)}{\tau_1}+\frac{1}{8}\right)D_{Sm} + \left(\frac{\tau_2}{\tau_1}+\frac{1}{8}\right)\widetilde{D}^{S}\sum_{j=0}^{m-1}\theta^j \\
&\qquad \le \theta^{-Sm}\left(\left(\frac{(1-\tau_1-\tau_2)}{\tau_1}+\frac{1}{8}\right)D_0 + \left(\frac{\tau_2}{\tau_1}+\frac{1}{8}\right)\widetilde{D}^{0}\sum_{j=0}^{m-1}\theta^j + \overline{C}\|z_0-x^*\|^2\right).
\end{aligned} \tag{5.24}$$

The inequality in (5.24) together with (5.22) implies

$$\mathbb{E}\big[H(\widetilde{x}^{\rm out}) - H(x^*)\big] \le \theta^{-Sm}\frac{(1-\tau_1-\tau_2+\tau_1/8)D_0 + (\tau_2+\tau_1/8)\widetilde{D}^0\sum_{j=0}^{m-1}\theta^j + \tau_1\overline{C}\|z_0-x^*\|^2}{1-\tau_1-\tau_2+\tau_1/8+(\tau_2+\tau_1/8)m}. \quad (5.25)$$

Notice that

$$\sum_{j=0}^{m-1}\theta^j = \frac{\theta^m-1}{\theta-1} = \frac{(1+\alpha\mu/2)^m-1}{\alpha\mu/2} \le \frac{2m(\alpha\mu/2)}{\alpha\mu/2}$$

where to obtain the inequality, we have used $\frac{\alpha\mu}{2} \le \frac{1}{12m} \le \frac{1}{2(m+1)}$ and Lemma 5.7 with $t = \frac{\alpha\mu}{2}$. Hence, we have from (5.25) and $z_0 = x_0$ that

$$\mathbb{E}\big[H(\widetilde{x}^{\rm out}) - H(x^*)\big] \le \theta^{-Sm}\left(D_0 + 2\widetilde{D}^0 + \frac{\tau_1\overline{C}}{1-\tau_1-\tau_2+\tau_1/8+(\tau_2+\tau_1/8)m}\|x_0-x^*\|^2\right) \quad (5.26)$$

From (5.21) and $\alpha = \frac{1}{\sqrt{18mL\mu\tau}}$, it follows

$$\frac{\tau_1\overline{C}}{1-\tau_1-\tau_2+\tau_1/8+(\tau_2+\tau_1/8)m} \le \frac{\tau_1\overline{C}}{(\tau_2+\tau_1/8)m} \le \frac{40\sqrt{18mL\mu\tau}}{9m} \le 20\sqrt{\frac{L\mu\tau}{m}}.$$

Therefore, by the μ-strong convexity of H, we have from (5.26), the definition of D_0 and $\widetilde{D}^0$ in (2.2), and $\theta = 1+\frac{\alpha\mu}{2}$ that

$$\mathbb{E}\big[H(\widetilde{x}^{\rm out}) - H(x^*)\big] \le \left(1+\frac{\alpha\mu}{2}\right)^{-Sm}\left(3+20\sqrt{\frac{L\tau}{m\mu}}\right)\big[H(x_0)-H(x^*)\big]. \quad (5.27)$$

Since $\alpha = \frac{1}{\sqrt{18mL\mu\tau}}$, the above inequality gives the first one in (5.17).

Case 2. Suppose $\tau < \frac{1}{2m}$ and $\sqrt{\frac{\mu}{L}} \le \frac{1}{2m}$.

In this case, $\tau_1 = \sqrt{\frac{\mu}{L}}$ and $\alpha = \frac{1}{3\sqrt{L\mu}}$. Hence, $\alpha\mu = \frac{1}{3}\sqrt{\frac{\mu}{L}} \le \frac{1}{6m}$, and thus $\theta = 1+\frac{\alpha\mu}{2}$. Therefore, all conditions required by Lemma 5.8 are satisfied, so we have $\frac{(\tau_1+\tau_2-1+1/\theta)}{\tau_1} - \overline{A} \ge \left(\frac{\tau_2}{\tau_1}+\overline{B}\right)\theta^{m-1}$. Then similar to **Case 1**, we have (5.23) and thus (5.26). Notice $\tau_1 \le \frac{1}{2}$ and thus $1-\tau_1-\tau_2+m\tau_2 \ge \frac{1}{2}$ to have

$$\frac{\tau_1\overline{C}}{1-\tau_1-\tau_2+\tau_1/8+(\tau_2+\tau_1/8)m} \le 2\tau_1\overline{C} \overset{(5.20)}{\le} \frac{10\tau_1}{9\alpha} = \frac{10\mu}{3} \le 4\mu.$$

Now using the μ-strong convexity and the definition of D_0 and $\widetilde{D}^0$, we obtain from (5.26) that

$$\mathbb{E}\big[H(\widetilde{x}^{\rm out}) - H(x^*)\big] \le 7\left(1+\frac{\alpha\mu}{2}\right)^{-Sm}\big[H(x_0)-H(x^*)\big]. \quad (5.28)$$

Since $\alpha = \frac{1}{3\sqrt{L\mu}}$, the above inequality gives the second one in (5.17).

Case 3. Suppose $\tau \ge \frac{1}{2m}$ and $\frac{m\mu}{L} > \frac{\tau}{2}$.

In this case, we have $\tau_1 = \tau$, $\alpha = \frac{1}{3\tau L} > \frac{1}{6m\mu}$. Hence, by the parameter setting in (5.16), it follows $\alpha\mu > \frac{1}{6m}$, $2m\tau_1 \geq 1$, $\frac{\tau_2}{\tau_1} \leq 1$, and $\theta = 1 + \frac{1}{12m}$, and thus $\theta^m \geq \frac{13}{12}$. By Lemma 4.3, we have $\frac{(\tau_1+\tau_2-1+1/\theta-\overline{A}\tau_1)\theta}{\tau_2+\overline{B}\tau_1} > \frac{13}{12}$. Therefore, we have (4.7) (3.12) through the same arguments above (4.7).

Notice the only difference between (4.7) and (5.23) is that the former inequality has coefficient $\frac{12}{13}$ while the latter has θ^{-m}. In addition, $\sum_{j=0}^{m-1}\theta^j = \frac{\theta^m-1}{\theta-1} \leq 2m$ still holds from Lemma 5.7. Therefore, similar to (5.26), we have

$$\mathbb{E}\big[H(\widetilde{x}^{\rm out}) - H(x^*)\big] \leq \Big(\frac{12}{13}\Big)^S \left(D_0 + 2\widetilde{D}^0 + \frac{\tau_1\overline{C}}{(\tau_2+\tau_1/8)m}\|x_0 - x^*\|^2\right) \tag{5.29}$$

By (5.21) and $\alpha > \frac{1}{6m\mu}$, it holds $\frac{\tau_1\overline{C}}{(\tau_2+\tau_1/8)m} \leq \frac{80\mu}{3} < 27\mu$. Now using the μ-strong convexity and the definition of D_0 and $\widetilde{D}^0$, we obtain from (5.29) that

$$\mathbb{E}\big[H(\widetilde{x}^{\rm out}) - H(x^*)\big] \leq 30\Big(\frac{12}{13}\Big)^S\big[H(x_0) - H(x^*)\big]. \tag{5.30}$$

Case 4. Suppose $\tau < \frac{1}{2m}$ and $\sqrt{\frac{\mu}{L}} > \frac{1}{2m}$.

In this case, $\tau_1 = \frac{1}{2m}$, $\alpha = \frac{2m}{3L} > \frac{1}{6m\mu}$, and $\frac{\tau_2}{\tau_1} \leq \frac{\tau_2}{\tau} \leq 1$. Hence, $\theta = 1 + \frac{1}{12m}$ from the observation $\alpha\mu > \frac{1}{6m}$, and thus $\theta^m \geq \frac{13}{12}$. Therefore (3.12) implies (4.7). Telescope the inequalities (4.7) over all outer loops $s = 0, \ldots, S-1$, Therefore, we obtain (5.30) following the same arguments as those in **Case 3**. □

Remark 5.1. *We make a few remarks about Theorem 5.9 and its proof.*

1. *The four cases in the proof correspond to the four possible values of τ_1 in (5.16a). The convergence rate results of **Cases 3** and **4** are the same and thus combined.*
2. *The overall complexity of Alogrithm 1 that takes **Option II** is $O\big(S(n_1+n_2+m(A+B+C))\big)$. Given $\varepsilon > 0$, to produce a stochastic ε-solution, we can have an explicit upper bound of S about other parameters by the convergence rate results. Hiding constant terms, we have:*

$$S = \begin{cases} O\left(\sqrt{\frac{L\tau}{m\mu}}\log\sqrt{\frac{L\tau}{m\mu}}\frac{H(x_0)-H(x^*)}{\varepsilon}\right), & \text{if } \frac{\tau}{2} \geq \frac{m\mu}{L} \text{ and } \tau \geq \frac{1}{2m}, \\ O\left(\frac{1}{m}\sqrt{\frac{L}{\mu}}\log\frac{H(x_0)-H(x^*)}{\varepsilon}\right), & \text{if } \sqrt{\frac{\mu}{L}} \leq \frac{1}{2m} \text{ and } \tau < \frac{1}{2m}, \\ O\left(\log\frac{H(x_0)-H(x^*)}{\varepsilon}\right), & \text{otherwise.} \end{cases} \tag{5.31}$$

However, notice that in Theorem 5.9, m and C (or equivalently τ) are only specified within a range, and thus to fully determine the complexity, we need to set m and C. This will be done in Corollary 5.10 below by discussing two different scenarios based on the comparison of n_1+n_2 to $\frac{L}{\mu}$.

Below, we discuss the three scenarios in Theorem 5.9 and give the complexity result of Algorithm 1 to produce a stochastic ε-solution. We show that in terms of the order, scenario 2 (i.e., $m \leq \frac{1}{2}\sqrt{\frac{L}{\mu}}$ and $\tau \leq \frac{1}{2m}$) is always the best choice.

Corollary 5.10 (complexity result of Algorithm 1 with Assumption 5.1)**.** *Under Assumptions 2.1, 2.2, 2.3 and 5.1, given $\varepsilon > 0$, Algorithm 1 can produce a stochastic ε-solution with the overall complexity:*

$$O\left(\left(n_1+n_2+\sqrt{(n_1+n_2)\frac{L^2}{\mu^2}}+\sqrt{\frac{L}{\mu}}\frac{\max\{B_G^4L_F^2,\, B_F^2L_G^2\}}{\mu^2}\right)\log\frac{H(x_0)-H(x^*)}{\varepsilon}\right), \tag{5.32}$$

which is achieved by setting $\tau=\frac{1}{2m}$ with $m=\Theta\left(\sqrt{\frac{n_1+n_2}{L/\mu}}\right)$ if $n_1+n_2=O\left(\frac{L^2}{\mu^2}\right)$ and $m=\Theta\big(\sqrt{L/\mu}\big)$ if $n_1+n_2=\Omega\left(\frac{L^2}{\mu^2}\right)$, and setting all other parameters according to (5.16).

Proof. We start from scenario 2 in Theorem 5.9, i.e., $m\le\frac{1}{2}\sqrt{\frac{L}{\mu}}$ and $\tau\le\frac{1}{2m}$. By (5.31) and (5.15), we have the overall complexity: $O\left(\sqrt{\frac{L}{\mu}}\left(\frac{n_1+n_2}{m}+A+B+\frac{1}{\tau}(1+\frac{12L}{\mu})\right)\log\frac{H(x_0)-H(x^*)}{\varepsilon}\right)$. Minimizing the complexity over $\tau\le\frac{1}{2m}$ gives

$$O\left(\sqrt{\frac{L}{\mu}}\Big(\frac{n_1+n_2}{m}+2m(1+\tfrac{12L}{\mu})+A+B\Big)\log\frac{H(x_0)-H(x^*)}{\varepsilon}\right). \tag{5.33}$$

The minimum value of $\frac{n_1+n_2}{m}+2m(1+\frac{12L}{\mu})$ about $m\le\frac{1}{2}\sqrt{\frac{L}{\mu}}$ is reached at $m=\sqrt{\frac{n_1+n_2}{2(1+\frac{12L}{\mu})}}$ if this m is no greater than $\frac{1}{2}\sqrt{\frac{L}{\mu}}$ and otherwise at $m=\frac{1}{2}\sqrt{\frac{L}{\mu}}$. Therefore, the best complexity result of Algorithm 1 under scenario 2 is

$$\begin{cases} O\left(\sqrt{\frac{L}{\mu}}\left(\sqrt{(n_1+n_2)\frac{L}{\mu}}+A+B\right)\log\frac{H(x_0)-H(x^*)}{\varepsilon}\right), & \text{if } n_1+n_2=O\big(\frac{L^2}{\mu^2}\big),\\ O\left(\sqrt{\frac{L}{\mu}}\left(\frac{n_1+n_2}{\sqrt{L/\mu}}+A+B\right)\log\frac{H(x_0)-H(x^*)}{\varepsilon}\right), & \text{if } n_1+n_2=\Omega\big(\frac{L^2}{\mu^2}\big). \end{cases} \tag{5.34}$$

Second, we discuss scenario 1 in Theorem 5.9, i.e., $\tau\ge 2m\frac{\mu}{L}$ and $\tau\ge\frac{1}{2m}$. By (5.31) and (5.15), we have the overall complexity: $O\left(\sqrt{\frac{L\tau}{m\mu}}\Big(n_1+n_2+m\big(A+B+\frac{1}{\tau}(1+\frac{12L}{\mu})\big)\Big)\log\Big(\sqrt{\frac{L\tau}{m\mu}}\frac{H(x_0)-H(x^*)}{\varepsilon}\Big)\right)$. Since $\tau\ge\frac{1}{2m}$, it holds $m\sqrt{\frac{L\tau}{m\mu}}\ge\sqrt{\frac{L}{2\mu}}$. In addition, $\sqrt{\frac{L\tau}{m\mu}}\ge\sqrt{\frac{1}{2}}$. Hence, in this scenario, the best complexity is at least in the order of

$$\sqrt{\frac{L}{\mu}}\Big(\frac{n_1+n_2}{\sqrt{m/\tau}}+\sqrt{\frac{m}{\tau}}(1+\tfrac{12L}{\mu})+\frac{1}{\sqrt{2}}(A+B)\Big)\log\frac{H(x_0)-H(x^*)}{\varepsilon}.$$

Comparing the above to the bound in (5.33) and noting $\sqrt{\frac{m}{\tau}}\le\sqrt{\frac{L}{2\mu}}$, we have that the minimum of the above complexity is the same as that given in (5.34). Hence, scenario 1 will never give better overall complexity (in terms of the order) than scenario 2.

Finally, we discuss scenario 3 in Theorem 5.9, i.e., $\tau\ge\frac{1}{2m}$ and $\frac{m\mu}{L}\ge\frac{\tau}{2}$, or $\tau\le\frac{1}{2m}$ and $\sqrt{\frac{\mu}{L}}\ge\frac{1}{2m}$. Notice that if $\tau\ge\frac{1}{2m}$ and $\frac{m\mu}{L}\ge\frac{\tau}{2}$, then it must hold $\sqrt{\frac{\mu}{L}}\ge\frac{1}{2m}$, and also if $\tau\le\frac{1}{2m}$ and $\sqrt{\frac{\mu}{L}}\ge\frac{1}{2m}$, then

it holds $\frac{m\mu}{L} \geq \frac{\tau}{2}$. Hence, this scenario is simply $\sqrt{\frac{\mu}{L}} \geq \frac{1}{2m}$ and $\frac{m\mu}{L} \geq \frac{\tau}{2}$. By (5.31) and (5.15), we have the overall complexity: $O\left(\left(n_1+n_2+m\big(A+B+\frac{1}{\tau}(1+\frac{12L}{\mu})\big)\right)\log\frac{H(x_0)-H(x^*)}{\varepsilon}\right)$. It is easy to have

$$\min_{m,\tau}\left\{m\big(A+B+\tfrac{1}{\tau}(1+\tfrac{12L}{\mu}\big): \sqrt{\tfrac{\mu}{L}}\geq\tfrac{1}{2m},\ \tfrac{m\mu}{L}\geq\tfrac{\tau}{2}\right\} = \tfrac{\sqrt{L}}{2\sqrt{\mu}}(A+B)+\tfrac{L}{2\mu}(1+\tfrac{12L}{\mu}),$$

where the minimum is reached at $m=\frac{\sqrt{L}}{2\sqrt{\mu}}$ and $\tau=\frac{2m\mu}{L}$. Therefore, the best complexity result of Algorithm 1 under scenario 3 is

$$O\left(\left(n_1+n_2+\tfrac{\sqrt{L}}{2\sqrt{\mu}}(A+B)+\tfrac{L}{2\mu}(1+\tfrac{12L}{\mu})\right)\log\frac{H(x_0)-H(x^*)}{\varepsilon}\right), \tag{5.35}$$

which is no better than that in (5.34).

Therefore, scenario 2 is always the best choice. Recalling the setting of A and B in (5.16b), we obtain the desired result in (5.32) by writing the two bounds in (5.34) into a unified one. □

6 Treating non-strongly convex compositional optimization

In this section, we study convex (but may not be strongly-convex) finite-sum compositional optimization in the form of (1.1), namely, instead of Assumption 2.1, we make the following assumption:

Assumption 6.1. *The function f given in (1.2) is convex, and the function h in (1.1) is convex.*

The non-strongly convex case has been studied in [11]. The algorithmic design in [11] is built on the algorithms for strongly convex models. They modify earlier algorithms for strongly convex COPs in a way that their step size ranges in a fixed magnitude that is free from κ, and their number of inner iterations exponentially increases as the outer loop proceeds. This way, the final outer loop dominates the total computational cost. We adopt a similar trick to find an approximate solution of a convex model by solving a sequence of slightly perturbed but strongly convex problems. More precisely, we follow [2] and implement the black-box reductions; see Algorithm 2, where H_t^* denotes the optimal value of the problem in the t-th outer loop.

Algorithm 2 **N**on-strongly C**o**nvex **C**ompositional **K**atyusha (NoCK)

Input: $x_0\in\mathbb{R}^{N_2}$, initial strong-convexity constant μ_0, and outer loop number T
for $t=0,1,\ldots,T-1$ **do**
Apply Algorithm 1 with starting point x_t to $\min_x\left\{H_t(x):=H(x)+\frac{\mu_t}{2}\|x-x_0\|^2\right\}$ and find x_{t+1} such that

$$\mathbb{E}\big[H_t(x_{t+1})-H_t^*\,\big|\,x_t\big]\leq\frac{H_t(x_t)-H_t^*}{4}. \tag{6.1}$$

Let $\mu_{t+1}=\mu_t/2$.
end for
Return x_T

The following result is from [2, Theorem 3.1] and can be proved in the same way.

Lemma 6.1. *Let x^* be an optimal solution of (1.1) and H^* be the optimal objective value. Suppose $H(x_0) - H^* \leq D_H$ and $\|x_0 - x^*\|^2 \leq D_x$. Set $\mu_0 = D_H/D_x$ and $T = \log_2(D_H/\varepsilon)$. Then $\mathbb{E}[H(x_T) - H^*] \leq \varepsilon$ and the total complexity is $\sum_{t=0}^{T-1} \text{Time}(t)$, where $\text{Time}(t)$ denotes the complexity to produce x_{t+1}.*

Remark 6.1. *The setting of μ_0 and T requires an estimate of D_H and D_x. This can be done if the domain of h is bounded and we know the bound. Otherwise, one can simply tune μ_0 as explained in [2].*

Applying Corollaries 4.5, 5.4, and 5.10 together with Lemma 6.1, we can easily have the complexity result of Algorithm 2 to produce a stochastic ε-solution of (1.1) when it is non-strongly convex.

Theorem 6.2 (case of relatively small n_1)**.** *Let $\varepsilon > 0$. Suppose Algorithm 1 with **Option I** is applied as the subroutine in Algorithm 2. Then under Assumptions 6.1, 2.2 and 2.3, Algorithm 2 with μ_0 and T set to those in Lemma 6.1 can produce a stochastic ε-solution of (1.1) with overall complexity*

$$\sum_{t=0}^{T-1} \text{Time}(t) = O\left((n_1+n_2)\log\frac{D_H}{\varepsilon} + \frac{D_x^{2.5}\sqrt{L}\cdot\max\{B_G^4L_F^2,\, B_F^2L_G^2\}}{\varepsilon^{2.5}} + \frac{\sqrt{D_xL}n_1}{\sqrt{\varepsilon}}\right). \tag{6.2}$$

Proof. From Lemma 6.1, it follows that x_T is a stochastic ε-solution of (1.1), and thus we only need to show the overall complexity $\sum_{t=0}^{T-1} \text{Time}(t)$. By Corollary 4.5 and $\mu_t = \frac{\mu_0}{2^t}$, the complexity to produce x_{t+1} that satisfies (6.1) is

$$\text{Time}(t) = O\left(n_1+n_2+2^{2.5t}\sqrt{\frac{L}{\mu_0}}\frac{\max\{B_G^4L_F^2,\, B_F^2L_G^2\}}{\mu_0^2} + 2^{0.5t}n_1\sqrt{\frac{L}{\mu_0}}\right).$$

Hence, the total complexity is

$$\sum_{t=0}^{T-1} \text{Time}(t) = O\left((n_1+n_2)T + \frac{2^{2.5T}-1}{2^{2.5}-1}\sqrt{\frac{L}{\mu_0}}\frac{\max\{B_G^4L_F^2,\, B_F^2L_G^2\}}{\mu_0^2} + \frac{2^{0.5T}-1}{2^{0.5}-1}n_1\sqrt{\frac{L}{\mu_0}}\right).$$

Since $\mu_0 = \frac{D_H}{D_x}$ and $T = \log_2\frac{D_H}{\varepsilon}$, the above equation implies the desired result. □

Remark 2 From the result in (6.2), we see that if $n_1 = O\left(\frac{D_x^2\max\{B_G^4L_F^2,\, B_F^2L_G^2\}}{\varepsilon^2}\right)$, we should always take **Option I** while applying Algorithm 1, since even we do mini-batch sampling, the complexity result will not become better.

Theorem 6.3 (case of big n_1 without Assumption 5.1)**.** *Let $\varepsilon > 0$, and let Algorithm 1 with **Option II** be applied as the subroutine in Algorithm 2. Then under Assumptions 6.1, 2.2 and 2.3, Algorithm 2 with μ_0 and T set to those in Lemma 6.1 can produce a stochastic ε-solution of (1.1) with overall complexity*

$$\sum_{t=0}^{T-1} \text{Time}(t) = O\left((n_1+n_2)\log\frac{D_H}{\varepsilon} + \frac{D_x^{2.5}\sqrt{L}\cdot\max\{B_G^4L_F^2,\, B_F^2L_G^2, L^2\}}{\varepsilon^{2.5}}\right).$$

Proof. First, notice again that x_T is a stochastic ε-solution of (1.1). By Corollary 5.4 and $\mu_t = \frac{\mu_0}{2^t}$, the complexity to produce x_{t+1} that satisfies (6.1) is

$$\text{Time}(t) = O\left(n_1+n_2+2^{2.5t}\sqrt{\frac{L}{\mu_0}}\frac{\max\{B_G^4L_F^2,\, B_F^2L_G^2,\, L^2\}}{\mu_0^2}\right)$$

Now summing $\text{Time}(t)$ over t and following the proof of Theorem 6.2, we obtain the desired result. □

Theorem 6.4 (case of big n_1 with Assumption 5.1)**.** *Let $\varepsilon > 0$, and let Algorithm 1 with **Option II** be applied as the subroutine in Algorithm 2. Then under Assumptions 6.1, 2.2, 2.3, and 5.1, Algorithm 2 with μ_0 and T set to those in Lemma 6.1 can produce a stochastic ε-solution of (1.1) with overall complexity*

$$\sum_{t=0}^{T-1} \text{Time}(t) = O\left((n_1+n_2)\log\frac{D_H}{\varepsilon} + \frac{D_x}{\varepsilon}L\sqrt{n_1+n_2} + \frac{D_x^{2.5}\sqrt{L}\cdot\max\{B_G^4L_F^2,\, B_F^2L_G^2\}}{\varepsilon^{2.5}}\right).$$

Proof. Again, notice that x_T is a stochastic ε-solution of (1.1). By Corollary 5.10, the complexity to produce x_{t+1} that satisfies (6.1) is

$$\text{Time}(t) = O\left(n_1+n_2+\sqrt{(n_1+n_2)\frac{L^2}{\mu_t^2}} + \sqrt{\frac{L}{\mu_t}}\frac{\max\{B_G^4L_F^2,\, B_F^2L_G^2\}}{\mu_t^2}\right).$$

Plugging in $\mu_t = \mu_0/2^t$ and summing the above over $t = 0$ through $T-1$ give the desired result. □

7 Conclusions

We have proposed an algorithm for solving the strongly convex case of the finite-sum compositional problem (1.1). To produce a stochastic ε-solution, the proposed algorithm generally needs $O\left((n_1+n_2+\kappa^{2.5})\log\frac{1}{\varepsilon}\right)$ evaluations of component function/gradient/Jaconian, where κ denotes the condition number. For convex cases of (1.1), we proposed an algorithm that approximately solves a sequence of strongly convex perturbed problems. The complexity result is generally $O\left((n_1+n_2)\log\frac{1}{\varepsilon}+\varepsilon^{-2.5}\right)$. For both strongly-convex and convex cases, our complexity results are better than the best-known existing ones.

A Missing Proofs

Proof of Lemma 4.3. The following chain of inequalities holds:

$$\begin{aligned}
\frac{(\tau_1+\tau_2-1+1/\theta-\overline{A}\tau_1)\theta}{\tau_2+\overline{B}\tau_1} &= \frac{(1+\frac{\tau_2}{\tau_1}-\overline{A})\tau_1\theta+1-\theta}{(\frac{\tau_2}{\tau_1}+\overline{B})\tau_1} \overset{①}{=} \frac{1+\frac{\tau_2}{\tau_1}-\overline{A}}{\frac{\tau_2}{\tau_1}+\overline{B}}\theta - \frac{1}{12m\tau_1(\frac{\tau_2}{\tau_1}+\overline{B})}\\
\overset{②}{\geq} 1+\frac{1-\overline{A}-\overline{B}}{\frac{\tau_2}{\tau_1}+\overline{B}} - \frac{1}{6(\frac{\tau_2}{\tau_1}+\overline{B})} &= 1+\frac{1-\overline{A}-\overline{B}-\frac{1}{6}}{\frac{\tau_2}{\tau_1}+\overline{B}} \overset{③}{\geq} 1+\frac{1-\overline{A}-\overline{B}-\frac{1}{6}}{1+\overline{B}} \geq 1+\frac{1-\frac{1}{4}-\frac{1}{6}}{1+\frac{1}{8}} > \frac{13}{12}.
\end{aligned}$$

Here, ① holds because $\theta = 1+\frac{1}{12m}$, ② follows from $\theta \geq 1$ and $2m\tau_1 \geq 1$, and ③ uses $\frac{\tau_2}{\tau_1} \leq 1$. □

Proof of Lemma 5.8. The following chain of inequalities holds:

$$\begin{aligned}
&(\tau_2+\overline{B}\tau_1)(\theta^{m-1}-1)+\overline{B}\tau_1+1-1/\theta \overset{①}{\leq} (\tau+\overline{B}\tau_1)(\theta^{m-1}-1)+\overline{B}\tau_1+1-1/\theta\\
\overset{②}{\leq}&(\tau+\overline{B}\tau_1)2(m-1)(\theta-1)+\overline{B}\tau_1+\theta-1 \overset{③}{\leq} (\tau+\overline{B}\tau_1)2(m-1)\alpha\mu+\overline{B}\tau_1+\alpha\mu\\
=&2\tau\alpha\mu(m-1)+2\alpha\mu(m-1)\overline{B}\tau_1+\overline{B}\tau_1+\alpha\mu \overset{④}{\leq} \frac{2(m-1)}{6m}\tau_1+\frac{2(m-1)}{6m}\overline{B}\tau_1+\overline{B}\tau_1+\frac{1}{3}\tau_1\\
\leq&\left(\frac{1}{3}+\frac{1}{3}\cdot\frac{1}{8}+\frac{1}{8}+\frac{1}{3}\right)\tau_1 \leq (1-\overline{A})\tau_1,
\end{aligned}$$

Here, ① holds because $\tau_2 \leq \tau$; ② is from Lemma 5.7 and $\theta > 1$; ③ uses $\theta - 1 \leq \alpha\mu$; ④ follows from $\tau\alpha\mu \leq \frac{\tau_1}{6m}$, $\alpha\mu \leq \frac{1}{6m}$, and $3\alpha\mu \leq \tau_1$. Rearranging term leads to the desired result. □

References

1. Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017. 4, 5, 7, 8, 9, 10
2. Z. Allen-Zhu and E. Hazan. Optimal black-box reductions between optimization objectives. In *Advances in Neural Information Processing Systems*, pages 1614–1622, 2016. 5, 26, 27
3. Z. Allen-Zhu and L. Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014. 4, 7
4. A. Defazio. A simple practical accelerated method for finite sums. In *Advances in neural information processing systems*, pages 676–684, 2016. 4
5. A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014. 4
6. Z. Huo, B. Gu, J. Liu, and H. Huang. Accelerated method for stochastic composition optimization with nonsmooth regularization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 4, 5, 6, 7
7. R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013. 2, 4, 5
8. X. Lian, M. Wang, and J. Liu. Finite-sum composition optimization via variance reduced gradient descent. *arXiv preprint arXiv:1610.04674*, 2016. 4, 5, 6, 7
9. X. Lian, M. Wang, and J. Liu. Finite-sum composition optimization via variance reduced gradient descent. In *Artificial Intelligence and Statistics*, pages 1159–1167, 2017. 2
10. H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In *Advances in neural information processing systems*, pages 3384–3392, 2015. 4
11. T. Lin, C. Fan, M. Wang, and M. I. Jordan. Improved sample complexity for stochastic compositional variance reduced gradient. *arXiv preprint arXiv:1806.00458*, 2018. 5, 6, 7, 26
12. L. Liu, J. Liu, and D. Tao. Variance reduced methods for non-convex composition optimization. *arXiv preprint arXiv:1711.04416*, 2017. 5
13. H. Markowitz. Portfolio selection [reprint of J. Finance **7** (1952), no. 1, 77–91]. In *Financial risk measurement and management*, volume 267 of *Internat. Lib. Crit. Writ. Econ.*, pages 197–211. Edward Elgar, Cheltenham, 2012. 2
14. H. M. Markowitz. *Mean-variance analysis in portfolio choice and capital markets*. Basil Blackwell, Oxford, 1987. 2
15. Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013. 4
16. Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013. 6, 19
17. Y. E. Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983. 6
18. P. Rigollet and X. Tong. Neyman-pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research*, 12(Oct):2831–2855, 2011. 3
19. M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017. 4
20. F. Shang, L. Jiao, K. Zhou, J. Cheng, Y. Ren, and Y. Jin. Asvrg: Accelerated proximal svrg. *arXiv preprint arXiv:1810.03105*, 2018. 4
21. M. Wang, E. X. Fang, and H. Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449, 2017. 2
22. M. Wang, J. Liu, and X. Fang. Accelerating stochastic composition optimization. *Journal of Machine Learning Research*, 18:1–23, 2017. 5
23. B. E. Woodworth and N. Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in neural information processing systems*, pages 3639–3647, 2016. 4
24. L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014. 2, 4
25. Y. Xu. Iteration complexity of inexact augmented lagrangian methods for constrained convex programming. *Mathematical Programming, Series A (online first)*, pages 1–46, 2019. 3
26. S. Yang, M. Wang, and E. X. Fang. Multilevel stochastic gradient methods for nested composition optimization. *SIAM Journal on Optimization*, 29(1):616–659, 2019. 2
27. Y. Yu and L. Huang. Fast stochastic variance reduced admm for stochastic composition optimization. *arXiv preprint arXiv:1705.04138*, 2017. 5
28. M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015. 3
29. J. Zhang and L. Xiao. Multi-level composite stochastic optimization via nested variance reduction. *arXiv preprint arXiv:1908.11468*, 2019. 2