A Projection-free Algorithm for Constrained Stochastic Multi-level Composition Optimization

Tesi Xiao
Department of Statistics
University of California, Davis
texiao@ucdavis.edu

Krishnakumar Balasubramanian
Department of Statistics
University of California, Davis
kbala@ucdavis.edu

Saeed Ghadimi
Department of Management Sciences
University of Waterloo
sghadimi@uwaterloo.ca

Abstract

We propose a projection-free conditional gradient-type algorithm for smooth stochastic multi-level composition optimization, where the objective function is a nested composition of T functions and the constraint set is a closed convex set. Our algorithm assumes access to noisy evaluations of the functions and their gradients, through a stochastic first-order oracle satisfying certain standard unbiasedness and second-moment assumptions. We show that the number of calls to the stochastic first-order oracle and the linear-minimization oracle required by the proposed algorithm, to obtain an ϵ-stationary solution, are of order $\mathcal{O}_T(\epsilon^{-2})$ and $\mathcal{O}_T(\epsilon^{-3})$ respectively, where $\mathcal{O}_T$ hides constants in T. Notably, the dependence of these complexity bounds on ϵ and T are separate in the sense that changing one does not impact the dependence of the bounds on the other. For the case of $T = 1$, we also provide a high-probability convergence result that depends poly-logarithmically on the inverse confidence level. Moreover, our algorithm is parameter-free and does not require any (increasing) order of mini-batches to converge unlike the common practice in the analysis of stochastic conditional gradient-type algorithms.

1 Introduction

We study projection-free algorithms for solving the following stochastic multi-level composition optimization problem

$$\min_{x\in\mathcal{X}} \quad F(x) := f_1 \circ \cdots \circ f_T(x), \tag{1}$$

where $f_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i-1}}, i = 1, \cdots, T$ $(d_0 = 1)$ are continuously differentiable functions, the composite function F is bounded below by $F^\star > -\infty$ and $\mathcal{X} \subset \mathbb{R}^d$ is a closed convex set. We assume that the exact function values and derivatives of f_i's are not available. In particular, we assume that $f_i(y) = \mathbb{E}_{\xi_i}[G_i(y, \xi_i)]$ for some random variable ξ_i. Our goal is to solve the above optimization problem, given access to noisy evaluations of ∇f_i's and f_i's.

There are two main challenges to address in developing efficient projection-free algorithms for solving (1). First, note that denoting the transpose of the Jacobian of f_i by ∇f_i, the gradient of the objective function $F(x)$ in (1), is given by $\nabla F(x) = \nabla f_T(y_T)\nabla f_{T-1}(y_{T-1}) \cdots \nabla f_1(y_1)$, where $y_i = f_{i+1} \circ \cdots \circ f_T(x)$ for $1 \le i < T$, and $y_T = x$. Because of the nested nature of the gradient $\nabla F(x)$, obtaining an unbiased gradient estimator in the stochastic first-order setting, with controlled moments, becomes non-trivial. Using naive stochastic gradient estimators lead to oracle complexities

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

that depend exponentially on T (in terms of the accuracy parameter). Next, even when $T = 1$ in the stochastic setting, projection-free algorithms like the conditional gradient method or its sliding variants invariably require increasing order of mini-batches[1] [28, 40, 24, 39, 52], which make their practical implementation infeasible.

In this work, we propose a projection-free conditional gradient-type algorithm that achieves *level-independent* oracle complexities (i.e., the dependence of the complexities on the target accuracy is independent of T) using only *one sample* of $(\xi_i)_{1\leq i\leq T}$ in each iteration, thereby addressing both of the above challenges. Our algorithm uses moving-average based stochastic estimators of the gradient and function values, also used recently by [19] and [4], along with an inexact conditional gradient method used by [3] (which in turn is inspired by the sliding method by [28]). In order to establish our oracle complexity results, we use a novel merit function based convergence analysis. To the best of our knowledge, such an analysis technique is used for the first time in the context of analyzing stochastic conditional gradient-type algorithms.

1.1 Preliminaries and Main Contributions

We now introduce the technical preliminaries required to state and highlight the main contributions of this work. Throughout this work, $\|\cdot\|$ denotes the Euclidean norm for vectors and the Frobenius norm for matrices. We first describe the set of assumptions on the objective functions and the constraint set.

Assumption 1 (Constraint set)**.** *The set $\mathcal{X} \subset \mathbb{R}^d$ is convex and closed with $\max_{x,y\in\mathcal{X}} \|x - y\| \leq D_{\mathcal{X}}$.*

Assumption 2 (Smoothness)**.** *All functions $f_1, \ldots, f_T$ and their derivatives are Lipschitz continuous with Lipschitz constants L_{f_i} and $L_{\nabla f_i}$, respectively.*

The above assumptions on the constraint set and the objective function are standard in the literature on stochastic optimization, and in particular in the analysis of conditional gradient algorithms and multi-level optimization; see, for example, [28], [50] and [4]. We emphasize here that the above smoothness assumptions are made only on the functions $(f_i)_{1\leq i\leq T}$ and not on the stochastic functions $(G_i)_{1\leq i\leq T}$ (which would be a stronger assumption). Moreover, the Lipschitz continuity of f_i's are implied by the Assumption 1 and assuming the functions are continuously differentiable. However, for sake of illustration, we state both assumptions separately. In addition to these assumptions, we also make unbiasedness and bounded-variance assumptions on the stochastic first-order oracle. Due to its technical nature, we state the precise details later in Section 3 (see Assumption 3).

We next turn our attention to the convergence criterion that we use in this work to evaluate the performance of the proposed algorithm. Gradient-based algorithms iteratively solve sub-problems in the form of

$$\arg\min_{u\in\mathcal{X}} \left\{ \langle g, u\rangle + \frac{\beta}{2}\|u - x\|^2 \right\}, \tag{2}$$

for some $\beta > 0$, $g \in \mathbb{R}^d$ and $x \in \mathcal{X}$. Denoting the optimal solution of the above problem by $P_{\mathcal{X}}(x, g, \beta)$ and noting its optimality condition, one can easily show that

$$-\nabla F(\bar{x}) \in \mathcal{N}_{\mathcal{X}}(\bar{x}) + \mathcal{B}\Big(0, \|g - \nabla F(\bar{x})\|D_{\mathcal{X}} + \beta\|x - P_{\mathcal{X}}(x, g, \beta)\|(D_{\mathcal{X}} + \|\nabla F(\bar{x})\|/\beta)\Big),$$

where $\mathcal{N}_{\mathcal{X}}(\bar{x})$ is the normal cone to $\mathcal{X}$ at $\bar{x}$ and $\mathcal{B}(0, r)$ denotes a ball centered at the origin with radius r. Thus, reducing the radius of the ball in the above relation will result in finding an approximate first-order stationary point of the problem for non-convex constrained minimization problems. Motivated by this fact, we can define the gradient mapping at point $\bar{x} \in \mathcal{X}$ as

$$\mathcal{G}_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x}), \beta) \coloneqq \beta\left(\bar{x} - P_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x}), \beta)\right) = \beta\left(\bar{x} - \Pi_{\mathcal{X}}\left(\bar{x} - \frac{1}{\beta}\nabla F(\bar{x})\right)\right), \tag{3}$$

where $\Pi_{\mathcal{X}}(y)$ denotes the Euclidean projection of the vector y onto the set $\mathcal{X}$. The gradient mapping is a classical measure has been widely used in the literature as a convergence criterion when solving nonconvex constrained problems [35]. It plays an analogous role of the gradient for constrained optimization problems; in fact when the set $\mathcal{X} \equiv \mathbb{R}^d$ the gradient mapping just reduces to $\nabla F(\bar{x})$. It should be emphasized that while the gradient mapping cannot be computed in the stochastic setting, it

[1] We discuss in detail about some recent works that avoid requiring increasing mini-batches, albeit under stronger assumptions, in Section 1.2.

Algorithm	Criterion	# of levels	Batch size	SFO	LMO
SPIFER-SFW [52]	FW-gap	1	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-2})$
1-SFW [54]	FW-gap	1	1	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$
SCFW [1]	FW-gap	2	1	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$
SCGS [39]	GM	1	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$
SGD+ICG [3]	GM	1	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$
LiNASA+ICG (Algorithm 1)	GM	T	1	$\mathcal{O}_T(\epsilon^{-2})$	$\mathcal{O}_T(\epsilon^{-3})$

Table 1: Complexity results for stochastic conditional gradient type algorithms to find an ϵ-stationary solution in the nonconvex setting. FW-Gap and GM stands for Frank-Wolfe Gap (see (4)) and Gradient Mapping (see (3)) respectively. $\mathcal{O}_T$ hides constants in T. Existing one-sample based stochastic conditional gradient algorithms are either (i) not applicable to the case of general $T > 1$, or (ii) require strong assumptions [54], or (iii) are not truly online [1]; see Section 1.2 for detailed discussion. The results in [3] are actually presented for the zeroth-order setting; however the above stated first-order complexities follow immediately.

still serves as a measure of convergence. Our main goal in this work is to find an ϵ-stationary solution to (1), in the sense described below.

Definition 1. *A point $\bar{x} \in \mathcal{X}$ generated by an algorithm for solving* (1) *is called an ϵ-stationary point, if we have $\mathbb{E}[\|\mathcal{G}_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x}), \beta)\|^2] \leq \epsilon$, where the expectation is taken over all the randomness involved in the problem.*

In the literature on conditional gradient methods for the nonconvex setting, the so-called Frank-Wolfe gap is also widely used to provide convergence analysis. The Frank-Wolfe Gap is defined formally as

$$g_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x})) := \max_{y \in \mathcal{X}} \langle \nabla F(\bar{x}), \bar{x} - y \rangle. \tag{4}$$

As pointed out by [3], the gradient mapping criterion and the Frank-Wolfe gap are related to each other in the following sense.

Proposition 1. [3] *Let $g_{\mathcal{X}}(\cdot)$ be the Frank-Wolfe gap defined in* (4) *and $\mathcal{G}_{\mathcal{X}}(\cdot)$ be the gradient mapping defined in* (3)*. Then, we have $\|\mathcal{G}_{\mathcal{X}}(x, \nabla F(x), \beta)\|^2 \leq g_{\mathcal{X}}(x, \nabla F(x)), \forall x \in \mathcal{X}$. Moreover, under Assumption 1, 2, we have $g_{\mathcal{X}}(x, \nabla F(x)) \leq \left[(1/\beta)\prod_{i=1}^{T} L_{f_i} + D_{\mathcal{X}}\right] \|\mathcal{G}_{\mathcal{X}}(x, \nabla F(x), \beta)\|$.*

For stochastic conditional gradient-type algorithms, the oracle complexity is measured in terms of number of calls to the Stochastic First-order Oracle (SFO) and the Linear Minimization Oracle (LMO) used to the solve the sub-problems (that are of the form of minimizing a linear function over the convex feasible set) arising in the algorithm. In this work, we hence measure the number of calls to SFO and LMO required by the proposed algorithm to obtain an ϵ-stationary solution in the sense of Definition 1. We now highlight our **main contributions**:

- We propose a projection-free conditional gradient-type method (Algorithm 1) for solving (1). In Theorem 2, we show that the SFO and LMO complexities of this algorithm, in order to achieve an ϵ-stationary solution in the sense of Definition 1, are of order $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-3})$, respectively.
- The above SFO and LMO complexities are in particular level-independent (i.e., the dependence of the complexities on the target accuracy is independent of T). The proposed algorithm is parameter-free and does not require any mini-batches, making it applicable for the online setting.
- When considering the case of $T \leq 2$, we present a simplified method (Algorithm 3 and 4) to obtain the same oracle complexities. Intriguingly, while this simplified method is still parameter-free for $T = 1$, it is not when $T = 2$ (see Theorem 3 and Remark 3.1). Furthermore, for the case of $T = 1$, we also establish high-probability bounds (see Theorem 5).

A summary of oracle complexities for stochastic conditional gradient-type algorithms is in Table 1.

1.2 Related Work

Conditional Gradient-Type Method. The conditional gradient algorithm [15, 29], has had a renewed interest in the machine learning and optimization communities in the past decade; see [33,

26, 20, 27, 5, 17] for a partial list of recent works. Considering the stochastic convex setup, [22, 24] provided expected oracle complexity results for the stochastic conditional gradient algorithm. The complexities were further improved by a sliding procedure in [28], based on Nesterov's acceleration method. [40, 52, 24] considered variance reduced stochastic conditional gradient algorithms, and provided expected oracle complexities in the non-convex setting. [39] analyzed the sliding algorithm in the non-convex setting and provided results for the gradient mapping criterion. *All of the above works require increasing orders of mini-batches to obtain their oracle complexity results*.

[34] and [54] proposed a stochastic conditional gradient-type algorithm with moving-average gradient estimator for the convex and non-convex setting that uses only one-sample in each iteration. However, [34] and [54] require several restrictive assumptions, which we explain next (focusing on [54] which considers the nonconvex case). Specifically, [54] requires that the stochastic function $G_1(x, \xi_1)$ has uniformly bounded function value, gradient-norm, and Hessian spectral-norm, and the distribution of the random vector ξ_1 has an absolutely continuous density p such that the norm of the gradient of $\log p$ and spectral norm of the Hessian of $\log p$ has finite fourth and second-moments respectively. In contrasts, we do not require such stringent assumptions. Next, all of the above works focus only on the case of $T = 1$. [1] considered stochastic conditional gradient algorithm for solving (1), with $T = 2$. However, [1] also makes stringent assumptions: (i) the stochastic objective functions $G_1(x, \xi_1)$ and $G_2(x, \xi_1)$ themselves have Lipschitz gradients almost surely and (ii) for a given instance of random vectors ξ_1 and ξ_2, one could query the oracle at the current and previous iterations, which makes the algorithm not to be truly online. See Table 1 for a summary.

Stochastic Multi-level Composition Optimization. Compositional optimization problems of the form in (1) have been considered as early as 1970s by [12]. Recently, there has been a renewed interest on this problem. [13] and [10] considered a sample-average approximation approach for solving (1) and established several asymptotic results. For the case of $T = 2$, [48], [49] and [6] proposed and analyzed stochastic gradient descent-type algorithms in the smooth setting. [9] and [11] considered the non-smooth setting and established oracle complexity results. Furthermore, [25] proposed algorithms when the randomness between the two levels are not necessarily independent. For the general case of $T \geq 1$, [50] proposed stochastic gradient descent-type algorithms and established oracle complexities established that depend exponentially on T and are hence sub-optimal. Indeed, large deviation and Central Limit Theorem results established in [13, 10], respectively, show that in the sample-average approximation setting, the $\arg\min$ of the problem in (1) based on n samples, converges at a level-independent rate (i.e., dependence of the convergence rate on the target accuracy is independent of T) to the true minimizer, under suitable regularity conditions.

[19] proposed a single time-scale Nested Averaged Stochastic Approximation (NASA) algorithm and established optimal rates for the cases of $T = 1, 2$. For the general case of $T \geq 1$, [4] proposed a linearized NASA algorithm and established level-independent and optimal convergence rates. Concurrently, [43] considered the case when the function f_i are non-smooth and established asymptotic convergence results. [53] also established non-asymptotic level-independent oracle complexities, however, under stronger assumptions than that in [4]. Firstly, they assumed that for a fixed batch of samples, one could query the oracle on different points, which is not suited for the general online stochastic optimization setup. Next, they assume a much stronger mean-square Lipschitz smoothness assumption on the individual functions f_i and their gradients. Finally, they required mini-batches sizes that depend exponentially on T, which makes their method impractical. Concurrent to [4], level-independent rates were also obtained for *unconstrained* problems by [7], albeit, under the stronger assumption that the stochastic functions $G_i(x, \xi_i)$ are Lipschitz, almost surely. It is also worth mentioning that while some of the above papers considered constrained problems, the algorithms proposed and analyzed in the above works are not projection-free, which is the main focus of this work.

2 Methodology

In this section, we present our projection-free algorithm for solving problem (1). The method generates three random sequences, namely, approximate solutions $\{x^k\}$, average gradients $\{z^k\}$, and average function values $\{u^k\}$, defined on a certain probability space $(\Omega, \mathscr{F}, P)$. We let $\mathscr{F}_k$ to be the σ-algebra generated by $\{x^0, \ldots, x^k, z^0, \ldots, z^k, u_1^0, \ldots, u_1^k, \ldots, u_T^0, \ldots, u_T^k\}$. The overall method is given in Algorithm 1. In (7), the stochastic Jacobians $J_i^{k+1} \in \mathbb{R}^{d_i \times d_{i-1}}$, and the product

Algorithm 1 Linearized NASA with Inexact Conditional Gradient Method (`LiNASA+ICG`)

Input: $x^0 \in \mathcal{X}$, $z^0 = 0 \in \mathbb{R}^d$, $u_i^0 \in \mathbb{R}^{d_i}$, $i = 1, \ldots, T$, $\beta_k > 0$, $t_k > 0$, $\tau_k \in (0, 1]$, $\delta \geq 0$.
for $k = 0, 1, 2, \ldots, N$ **do**
1. Update the solution:

$$\tilde{y}^k = \texttt{ICG}(x^k, z^k, \beta_k, t_k, \delta), \tag{5}$$

$$x^{k+1} = x^k + \tau_k(\tilde{y}^k - x^k), \tag{6}$$

and compute stochastic Jacobians J_i^{k+1}, and function values G_i^{k+1} at u_{i+1}^k for $i = 1, \ldots, T$.
2. Update average gradients z and function value estimates u_i for each level $i = 1, \ldots, T$

$$z^{k+1} = (1 - \tau_k)z^k + \tau_k \prod_{i=1}^{T} J_{T+1-i}^{k+1}, \tag{7}$$

$$u_i^{k+1} = (1 - \tau_k)u_i^k + \tau_k G_i^{k+1} + \langle J_i^{k+1}, u_{i+1}^{k+1} - u_{i+1}^k \rangle. \tag{8}$$

end for
Output: $(x^R, z^R, u_1^R, \cdots, u_T^R)$, where R is uniformly distributed over $\{1, 2, \ldots, N\}$

Algorithm 2 Inexact Conditional Gradient Method (`ICG`)

Input: (x, z, β, M, δ)
Set $w^0 = x$.
for $t = 0, 1, 2, \ldots, M$ **do**
1. Find $v^t \in \mathcal{X}$ with a quantity $\delta \geq 0$ such that

$$\langle z + \beta(w^t - x), v^t \rangle \leq \min_{v \in \mathcal{X}} \langle z + \beta(w^t - x), v \rangle + \frac{\beta D_{\mathcal{X}}^2 \delta}{t + 2}.$$

2. Set $w^{t+1} = (1 - \mu_t)w^t + \mu_t v^t$ with $\mu_t = \min\left\{1, \frac{\langle \beta(x - w^t) - z, v^t - w^t \rangle}{\beta \|v^t - w^t\|^2}\right\}$.

end for
Output: w^M

$\prod_{i=1}^{T} J_{T+1-i}^{k+1}$ is calculated as $J_T^{k+1} J_{T-1}^{k+1} \cdots J_1^{k+1} \in \mathbb{R}^{d_T \times d_1} \equiv \mathbb{R}^{d_T \times 1}$. In (8), we use the notation $\langle \cdot, \cdot \rangle$ to represent both matrix-vector multiplication and vector-vector inner product. There are two aspects of the algorithm that we highlight specifically: (`i`) In addition to estimating the gradient of F, we also estimate a stochastic linear approximation of the inner functions f_i by a moving-average technique. In the multi-level setting we consider, it helps us to avoid the accumulation of bias, when estimating the f_i directly. Linearization techniques were used in the stochastic optimization since the work of [42]. A similar approach was used in [4] in the context of projected-based methods for solving (1). It is also worth mentioning that other linearization techniques have been used in [9] and [11] for estimating the stochastic inner function values for weakly convex two-level composition problems, and (`ii`) The `ICG` method given in Algorithm 2 is essentially applying *deterministic* conditional gradient method with the exact line search for solving the quadratic minimization subproblem in (2) with the estimated gradient z_k in (7). It was also used in [3] as a sub-routine and is motivated by the sliding approach of [28].

3 Main Results

In this section, we present our main result on the oracle complexity of Algorithm 1. Before we proceed, we present our assumptions on the stochastic first-order oracle.

Assumption 3 (Stochastic First-Order Oracle)**.** *Denote $u_{T+1}^k \equiv x^k$. For each k, u_{i+1}^k being the input, the stochastic oracle outputs $G_i^{k+1} \in \mathbb{R}^{d_i}$ and J_i^{k+1} such that given $\mathscr{F}_k$ and for any $i \in \{1, \ldots, T\}$*

(a) $\mathbb{E}[J_i^{k+1} | \mathscr{F}_k] = \nabla f_i(u_{i+1}^k)$, $\mathbb{E}[G_i^{k+1} | \mathscr{F}_k] = f_i(u_{i+1}^k)$,

(b) $\mathbb{E}[\|G_i^{k+1} - f_i(u_{i+1}^k)\|^2|\mathscr{F}_k] \le \sigma_{G_i}^2$, $\mathbb{E}[\|J_i^{k+1} - \nabla f_i(u_{i+1}^k)\|^2|\mathscr{F}_k] \le \sigma_{J_i}^2$,

(c) The outputs of the stochastic oracle at level i, G_i^{k+1} *and* J_i^{k+1}*, are independent. The outputs of the stochastic oracle are independent between levels, i.e.,* $\{G_i^{k+1}\}_{i=1,\dots,T}$ *are independent and so are* $\{J_i^{k+1}\}_{i=1,\dots,T}$.

Parts (a) and (b) in Assumption 3 are standard unbiasedness and bounded variance assumptions on the stochastic gradient, common in the literature. Part (c) is essential to establish the convergence results in the multi-level case. Similar assumptions have been made, for example, in [50] and [4]. We also emphasize that unlike some prior works (see e.g., [54]), Assumption 3 allows the case of endogenous uncertainty, and we do not require the distribution of the random variables $(\xi_i)_{1\le i\le T}$ to be independent of the distribution of the decision variables $(u_i)_{1\le i\le T}$.

Remark. *Under Assumption 2, and 3, we can immediately conclude that* $\mathbb{E}[\|J_i^{k+1}\|^2|\mathscr{F}_k] = \mathbb{E}[\|J_i^{k+1} - \nabla f_i(u_{i+1}^k)\|^2|\mathscr{F}_k] + \|\nabla f_i(u_{i+1}^k)\|^2 \le \sigma_{J_i}^2 + L_{f_i}^2 := \hat{\sigma}_{J_i}^2$. *In the sequel,* $\hat{\sigma}_{J_i}^2$ *will be used to simplify the presentation.*

We start with the merit function used in this work and its connection to the gradient mapping criterion. Our proof leverages the following merit function:

$$W_{\alpha,\gamma}(x,z,u) = F(x) - F^\star - \eta(x,z) + \alpha\|\nabla F(x) - z\|^2 + \sum_{i=1}^{T}\gamma_i\|f_i(u_{i+1}) - u_i\|^2, \quad (9)$$

where $\alpha, \{\gamma_i\}_{1\le i\le T}$ are positive constants and

$$\eta(x,z) = \min_{y\in\mathcal{X}}\left\{H(y;x,z,\beta) := \langle z, y-x\rangle + \frac{\beta}{2}\|y-x\|^2\right\}. \quad (10)$$

Compared to [4], we require the additional term $\|\nabla F(x) - z\|^2$, which turns out to be essential in our proof due to the `ICG` routine. The following proposition relates the merit function above to the gradient mapping.

Proposition 2. *Let* $\mathcal{G}_\mathcal{X}(\cdot)$ *be the gradient mapping defined in* (3) *and* $\eta(\cdot,\cdot)$ *be defined in* (10)*. For any pair of* (x,z) *and* $\beta > 0$*, we have* $\|\mathcal{G}_\mathcal{X}(x,\nabla F(x),\beta)\|^2 \le -4\beta\eta(x,z) + 2\|\nabla F(x) - z\|^2$.

Proof. By expanding the square, and using the properties of projection operation, we have

$$\|\Pi_\mathcal{X}(x - \frac{1}{\beta}z) - x\|^2 + \|\Pi_\mathcal{X}(x - \frac{1}{\beta}z) - (x - \frac{1}{\beta}z)\|^2 \le \|\bar{x} - (x - \frac{1}{\beta}z)\|^2 = \|\frac{1}{\beta}z\|^2.$$

Thus, we have $\eta(x,z) \le -\frac{\beta}{2}\|\Pi_\mathcal{X}(x - \frac{1}{\beta}z) - x\|^2$. The proof is completed immediately by noting that $\|\mathcal{G}(x,\nabla F(x),\beta)\|^2 \le 2\beta^2\|\Pi_\mathcal{X}(x - \frac{1}{\beta}z) - x\|^2 + 2\|\nabla F(x) - z\|^2$. □

We now present out main result on the oracle complexity of Algorithm1.

Theorem 2. *Under Assumption 1, 2, 3, let* $\{x^k, z^k, \{u_i^k\}_{1\le i\le T}\}_{k\ge 0}$ *be the sequence generated by Algorithm 1 with* $N \ge 1$ *and*

$$\beta_k \equiv \beta > 0, \qquad \tau_0 = 1,\ t_0 = 0, \quad \tau_k = \frac{1}{\sqrt{N}},\ t_k = \lceil\sqrt{k}\rceil, \quad \forall k \ge 1, \quad (11)$$

where β *is an arbitrary positive constant. Provided that the merit function* $W_{\alpha,\gamma}(x,z,u)$ *is defined as* (9) *with*

$$\alpha = \frac{\beta}{20L_{\nabla F}^2}, \quad \gamma_1 = \frac{\beta}{2}, \quad \gamma_j = \left(2\alpha + \frac{1}{4\alpha L_{\nabla F}^2}\right)(T-1)C_j^2 + \frac{\beta}{2}, \quad 2 \le j \le T, \quad (12)$$

we have,

$$\mathbb{E}\left[\|\mathcal{G}_\mathcal{X}(x^R,\nabla F(x^R),\beta)\|^2\right] \le \frac{2(\beta + \frac{20L_{\nabla F}^2}{\beta})\left[2W_{\alpha,\gamma}(x^0,z^0,u^0) + \mathcal{B}(\beta,\sigma^2,L,D_\mathcal{X},T,\delta)\right]}{\sqrt{N}}, \quad (13)$$

$$\mathbb{E}\left[\|f_i(u_{i+1}^R) - u_i^R\|^2\right] \leq \frac{2W_{\alpha,\gamma}(x^0, z^0, u^0) + \mathcal{B}(\beta, \sigma^2, L, D_{\mathcal{X}}, T, \delta)}{\beta\sqrt{N}}, \quad 1 \leq i \leq T. \tag{14}$$

where $u_{T+1} = x, \mathcal{B}(\beta, \sigma^2, L, D_{\mathcal{X}}, T, \delta) = 4\hat{\sigma}^2 + 32\beta D_{\mathcal{X}}^2(1+\delta)\left(\frac{3}{5} + \frac{5L_{\nabla F}^2}{\beta^2}\right)$*, and* $\hat{\sigma}^2$ *is a constant depending on the parameters* $(\beta, \sigma^2, L, D_{\mathcal{X}}, T)$ *given in* (42)*. The expectation is taken with respect to all random sequences generated by the method and an independent random integer number* R *uniformly distributed over* $\{1, \ldots, N\}$*. That is to say, the number of calls to SFO and LMO to get an* ϵ*-stationary point is upper bounded by* $\mathcal{O}_T(\epsilon^{-2}), \mathcal{O}_T(\epsilon^{-3})$ *respectively.*

Remark. *The constant* $\mathcal{B}(\beta, \sigma^2, L, D_{\mathcal{X}}, T, \delta)$ *is* $\mathcal{O}(T)$ *given the definition of* $\hat{\sigma}^2$ *and the value of* γ_j *in* (12)*, which further implies that the total number of calls to SFO and LMO of Algorithm 1 for finding an* ϵ*-stationary point of* (1)*, are bounded by* $\mathcal{O}(T^2\epsilon^{-2}) = \mathcal{O}_T(\epsilon^{-2})$ *and* $\mathcal{O}(T^3\epsilon^{-3}) = \mathcal{O}_T(\epsilon^{-3})$ *respectively. Furthermore, it is worth noting that this complexity bound for Algorithm 1 is obtained without any dependence of the parameter* β_k *on Lipschitz constants due to the choice of arbitrary positive constant* β *in* (11)*, and* τ_k, t_k *depend only on the number of iterations* N *and* k *respectively. This makes Algorithm 1 parameter-free and easy to implement.*

Remark. *As discussed in Section 2, the* `ICG` *routine given in Algorithm 2 is a* deterministic *method with the estimated gradient* z_k *in* (7)*. The number of iterations,* t_k*, required to run Algorithm 2 is given by* $t_k = \lceil\sqrt{k}\rceil$*. That is, we require more precise solutions for the* `ICG` *routine, only for later outer iterations. Furthermore, due to the deterministic nature of the* `ICG` *routine, further advances in the analysis of deterministic conditional gradient methods under additional assumptions on the constraint set* $\mathcal{X}$ *(see, for example, [16, 18]) could be leveraged to improve the overall LMO complexity.*

3.1 The special cases of $T = 1$ and $T = 2$

We now discuss several intriguing points regarding the choice of tuning parameter β, for the case of $T = 2$, and the more standard case of $T = 1$. Specifically, the linearization technique used in Algorithm 1 turns out to be not necessary for the case of $T = 2$ and $T = 1$ to obtain similar rates. However, without linearization, the choice of β is dependent on the problem parameters for $T = 2$. Whereas it turns out to be independent of the problem parameters (similar to Algorithm 1 and Theorem 2 which holds for all $T \geq 1$) for $T = 1$. As the outer function value estimates (i.e., u_1^{k+1} sequence) are not required for the convergence analysis, we remove them in Algorithms 3 and 4.

Algorithm 3 NASA with Inexact Conditional Gradient Method (`NASA+ICG`) for $T = 2$

Replace Step 2 of Algorithm 1 with the following:

2'. Update the average gradient z and the function value estimate u_2 respectively as:

$$z^{k+1} = (1 - \tau_k)z^k + \tau_k J_2^{k+1} J_1^{k+1} \quad \text{and} \quad u_2^{k+1} = (1 - \tau_k)u^k + \tau_k G_2^{k+1}$$

Algorithm 4 ASA with Inexact Conditional Gradient Method (`ASA+ICG`) for $T = 1$

Replace Step 2 of Algorithm 1 with the following:

2". Update the average gradient z as: $z^{k+1} = (1 - \tau_k)z^k + \tau_k J_1^{k+1}$.

Theorem 3. *Let Assumptions 1, 2, 3 be satisfied by the optimization problem* (1)*. Let* $\mathcal{C}_1, \mathcal{C}_2$ *and* $\mathcal{C}_3$ *be some constants depending on the parameters* $(\beta, \sigma^2, L, D_{\mathcal{X}}, \delta)$*, as defined in* (54) *and* (62)*. Let* $\tau_0 = 1, t_0 = 0$*,* $\tau_k = \frac{1}{\sqrt{N}}, t_k = \lceil\sqrt{k}\rceil, \forall k \geq 1$*, where* N *is the total number of iterations.*

(a) Let $T = 2$*, and let* $\{x^k, z^k, u_2^k\}_{k\geq 0}$ *be the sequence generated by Algorithm 3 with*

$$\beta_k \equiv \beta \geq 6\rho L_{\nabla F} + (2\rho + \frac{2}{3\rho})L_{\nabla f_1} L_{f_2}^2, \quad \rho > 0. \tag{15}$$

Then, we have $\forall N \geq 1$*,*

$$\mathbb{E}\left[\|\mathcal{G}_{\mathcal{X}}(x^R, \nabla F(x^R), \beta)\|^2\right] \leq \frac{\mathcal{C}_1}{\sqrt{N}}, \quad \mathbb{E}\left[\|f_2(x^R) - u_2^R\|^2\right] \leq \frac{\mathcal{C}_2}{\sqrt{N}}.$$

(b) Let $T = 1$ and let $\{x^k, z^k\}_{k\geq 0}$ be the sequence generated by Algorithm 4 with $\beta_k \equiv \beta > 0$. Then, we have $\forall N \geq 1$,

$$\mathbb{E}\left[\|\mathcal{G}_{\mathcal{X}}(x^R, \nabla F(x^R), \beta)\|^2\right] \leq \frac{\mathcal{C}_3}{\sqrt{N}}.$$

All expectations are taken with respect to all random sequences generated by the respective algorithms and an independent random integer number R uniformly distributed over $\{1, \ldots, N\}$. In both cases, the number of calls to SFO and LMO to get an ϵ-stationary point is upper bounded by $\mathcal{O}(\epsilon^{-2}), \mathcal{O}(\epsilon^{-3})$ respectively.

Remark. *While we can obtain the same complexities without using the linear approximation of the inner function for $T = 2$, it seems necessary to have a parameter-free algorithm as the choice of β in* (15) *depends on the knowledge of the problem parameters. Indeed, the linearization term in* (8) *helps use to better exploit the Lipschitz smoothness of the gradients get an error bound in the order of $\tau_k^2\|d^k\|^2$ for estimating the inner function values. Without this term, we are only able to use the Lipschitz continuity of the inner functions and so the error estimate will increase to the order of $\tau_k\|d^k\|$. Hence, we need to choose a larger $beta$ (as in* (15)*) to reduce $\|d^k\|$ and handle the error term without compromising the complexities. However, this is not the case for $T = 1$ as it can be seen as a two-level problem whose inner function is exactly known (the identity map). In this case, the choice of β is independent of the problem parameters with or without the linearization term.*

3.2 High-Probability Convergence for $T = 1$

In this subsection, we establish an oracle complexity result with high-probability for the case of $T = 1$. We first provide a notion of (ϵ, δ)-stationary point and a related tail assumption on the stochastic first-order oracle below.

Definition 4. *A point $\bar{x} \in \mathcal{X}$ generated by an algorithm for solving* (1) *is called an (ϵ, δ)-stationary point, if we have $\|\mathcal{G}_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x}), \beta)\|^2 \leq \epsilon$ with probability $1 - \delta$.*

Assumption 4. *Let $\Delta^{k+1} = \nabla F(x^k) - J_1^{k+1}$ for $k \geq 0$. For each k, given $\mathscr{F}_k$ we have $\mathbb{E}[\Delta^{k+1}|\mathscr{F}_k] = 0$ and $\|\Delta^{k+1}\|\big|\mathscr{F}_k$ is K-sub-Gaussian.*

The above assumption is commonly used in the literature; see [23, 21, 30, 55]. We also refer to [45] and Appendix E for additional details. The high-probability bound for solving non-convex constrained problems by Algorithm 4 is given below.

Theorem 5. *Let Assumptions 1, 2, 4 be satisfied by the optimization problem* (1) *with $T = 1$. Let $\tau_0 = 1, t_0 = 0$, $\tau_k = \frac{1}{\sqrt{N}}, t_k = \lceil\sqrt{k}\rceil, \forall k \geq 1$, where N is the total number of iterations. Let $T = 1$ and let $\{x^k, z^k\}_{k\geq 0}$ be the sequence generated by Algorithm 4 with $\beta_k \equiv \beta > 0$. Then, we have $\forall N \geq 1, \delta > 0$, with probability at least $1 - \delta$,*

$$\min_{k=1,\ldots,N} \left\|\mathcal{G}_{\mathcal{X}}(x^k, \nabla F(x^k), \beta)\right\|^2 \leq \mathcal{O}\left(\frac{K^2 \log(1/\delta)}{\sqrt{N}}\right)$$

Therefore, the number of calls to SFO and LMO to get an (ϵ, δ)-stationary point is upper bounded by $\mathcal{O}(\epsilon^{-2}\log^2(1/\delta)), \mathcal{O}(\epsilon^{-3}\log^3(1/\delta))$ respectively.

Remark. *To the best of our knowledge, the above result is (i) the first high-probability bound for one-sample stochastic conditional gradient-type algorithm for the case of $T = 1$, and (ii) the first high-probability bound for constrained stochastic optimization algorithms in the non-convex setting; see Appendix J of [32].*

4 Proof Sketch of Main Results

In this section, we only present the proof sketch. The complete proofs are provided in the appendix. For convenience, let $u_{T+1} = x$, and we denote H_k as the function value of the subproblem at step k, y^k as the optimal solution of the subproblem i.e.,

$$H_k(y) := H(y; x^k, z^k, \beta_k), \quad y^k = \arg\min_{y\in\mathcal{X}} H_k(y). \tag{16}$$

Then, the proof of Theorem 2 proceeds via the following steps:

1. We first leverage the merit function $W_k := W_{\alpha,\gamma}(x^k, z^k, u^k)$ defined in (9) with appropriate choices of α, γ for any $\beta > 0$ to obtain

$$W_{k+1} - W_k \leq -\frac{\tau_k}{2}\left(\beta\left[\|d^k\|^2 + \sum_{i=1}^{T}\|f_i(u_{i+1}^k) - u_i^k\|^2\right] + \frac{\beta}{20L_{\nabla F}^2}\|\nabla F(x^k) - z^k\|^2\right) + \mathbf{R}_k + \tau_k\left(\frac{12}{5} + \frac{20L_{\nabla F}^2}{\beta^2}\right)\left(H_k(\tilde{y}^k) - H_k(y^k)\right), \quad \forall k \geq 0$$

   where $\mathbf{R}_k$ is the residual term (see (31)) and $\mathbb{E}[\mathbf{R}_k|\mathscr{F}_k] \leq \hat{\sigma}^2\tau_k^2$, as shown in Proposition 3.

2. Telescoping the above inequality, in Lemma 11 we obtain the following:

$$\sum_{k=1}^{N}\tau_k\left[\beta\left(\|d^k\|^2 + \sum_{i=1}^{T}\|f_i(u_{i+1}^k) - u_i^k\|^2\right) + \frac{\beta}{20L_{\nabla F}^2}\|\nabla F(x^k) - z^k\|^2\right] \leq 2W_0 + 2\sum_{k=0}^{N}\mathbf{R}_k + \left(\frac{24}{5} + \frac{40L_{\nabla F}^2}{\beta^2}\right)\sum_{k=0}^{N}\tau_k\left(H_k(\tilde{y}^k) - H_k(y^k)\right), \quad \forall N \geq 1.$$

3. To further control the error term $H_k(\tilde{y}^k) - H_k(y^k)$ introduced by the `ICG` method, we set t_k, the number of iterations in `ICG` method at step k, to $\lceil\sqrt{k}\rceil$. By Lemma 8, we therefore have

$$H_k(\tilde{y}^k) - H_k(y^k) \leq \frac{2\beta D_{\mathcal{X}}^2(1+\delta)}{t_k + 2} \leq \frac{2\beta D_{\mathcal{X}}^2(1+\delta)}{\sqrt{k}}, \quad \forall k \geq 1.$$

   Also, with the choice of $\tau_k = \frac{1}{\sqrt{N}}$ and $z^0 = 0$, we can conclude that

$$\sum_{k=0}^{N}\tau_k\left(H_k(\tilde{y}^k) - H_k(y^k)\right) \leq \frac{2\beta D_{\mathcal{X}}^2(1+\delta)}{\sqrt{N}}\sum_{k=1}^{N}\frac{1}{\sqrt{k}} \leq 4\beta D_{\mathcal{X}}^2(1+\delta).$$

4. Then, taking expectation of both sides and by the definition of random integer R, we have

$$\mathbb{E}\left[\beta\left(\|d^R\|^2 + \sum_{i=1}^{T}\|f_i(u_{i+1}^R) - u_i^R\|^2\right) + \frac{\beta}{20L_{\nabla F}^2}\|\nabla F(x^R) - z^R\|^2\right] \leq 2W_0 + \mathcal{B},$$

   $\forall N \geq 1$, where $\mathcal{B}$ is a constant depending on the problem parameters $(\beta, \sigma^2, L, D_{\mathcal{X}}, T, \delta)$.

5. As a result, we can obtain (13) and (14) by noting that $\forall k \geq 1$

$$\|\mathcal{G}(x^k, \nabla F(x^k), \beta)\|^2 \leq 2\beta^2\|d^k\|^2 + 2\beta^2\left\|\Pi_{\mathcal{X}}\left(x^k - \frac{1}{\beta}\nabla F(x^k)\right) - \Pi_{\mathcal{X}}\left(x^k - \frac{1}{\beta}z^k\right)\right\|^2 \leq 2\beta^2\|d^k\|^2 + 2\|\nabla F(x^k) - z^k\|^2.$$

   where the second inequality follows the non-expansiveness of the projection operator.

The proofs of Theorems 3 and 5 follow the same argument with appropriate modifications. The high-probability convergence proof of Theorem 5 mainly consists of controlling the tail probability of the residual term $\mathbf{R}_k$ being large.

5 Discussion

In this work, we propose and analyze projection-free conditional gradient-type algorithms for constrained stochastic multi-level composition optimization of the form in (1). We show that the oracle complexity of the proposed algorithms is level-independent in terms of the target accuracy. Furthermore, our algorithm does not require any increasing order of mini-batches under standard unbiasedness and bounded second-moment assumptions on the stochastic first-order oracle, and is parameter-free. Some open questions for future research: (`i`) Considering the one-sample setting, either improving the LMO complexity from $\mathcal{O}(\epsilon^{-3})$ to $\mathcal{O}(\epsilon^{-2})$ for general closed convex constraint sets or establishing lower bounds showing that $\mathcal{O}(\epsilon^{-3})$ is necessary while keeping the SFO in the order of $\mathcal{O}(\epsilon^{-2})$, is extremely interesting; and (`ii`) Providing high-probability bounds for stochastic multi-level composition problems ($T > 1$) and under sub-Gaussian or heavy-tail assumptions (as in [32, 31]) is interesting to explore.

Acknowledgment

The authors are grateful to anonymous reviewers for their constructive comments that greatly improved the presentation of this paper. TX was partially supported by a seed grant from the Center for Data Science and Artificial Intelligence Research, UC Davis and National Science Foundation (NSF) grant CCF-1934568. KB was partially supported by a seed grant from the Center for Data Science and Artificial Intelligence Research, UC Davis and NSF grant DMS-2053918. SG was partially supported by an NSERC Discovery Grant.

References

[1] Zeeshan Akhtar, Amrit Singh Bedi, Srujan Teja Thomdapu, and Ketan Rajawat. Projection-Free Algorithm for Stochastic Bi-level Optimization. *arXiv preprint arXiv:2110.11721*, 2021.

[2] Raul Astudillo and Peter Frazier. Bayesian optimization of function networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[3] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, 22(1):35–76, 2022.

[4] Krishnakumar Balasubramanian, Saeed Ghadimi, and Anthony Nguyen. Stochastic multilevel composition optimization algorithms with level-independent convergence rates. *SIAM Journal on Optimization*, 32(2):519–544, 2022.

[5] Amir Beck and Shimrit Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Mathematical Programming*, 164(1-2):1–27, 2017.

[6] Jose Blanchet, Donald Goldfarb, Garud Iyengar, Fengpei Li, and Chaoxu Zhou. Unbiased simulation for optimizing stochastic function compositions. *arXiv preprint arXiv:1711.07564*, 2017.

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

[8] Weilin Cong, Rana Forsati, Mahmut Kandemir, and Mehrdad Mahdavi. Minimal variance sampling with provable guarantees for fast training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1393–1403, 2020.

[9] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[10] Darinka Dentcheva, Spiridon Penev, and Andrzej Ruszczyński. Statistical estimation of composite risk functionals and risk optimization problems. *Annals of the Institute of Statistical Mathematics*, 69(4):737–760, 2017.

[11] John Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.

[12] Yuri Ermoliev. Methods of stochastic programming. *Nauka, Moscow*, 1976.

[13] Yuri Ermoliev and Vladimir Norkin. Sample average approximation method for compound stochastic optimization problems. *SIAM Journal on Optimization*, 23(4):2231–2263, 2013.

[14] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Generalization of model-agnostic meta-learning algorithms: Recurring and unseen tasks. *Advances in Neural Information Processing Systems*, 34, 2021.

[15] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[16] Dan Garber and Elad Hazan. Faster rates for the frank-wolfe method over strongly-convex sets. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2015.

[17] Dan Garber, Atara Kaplan, and Shoham Sabach. Improved complexities of conditional gradient-type methods with applications to robust matrix recovery problems. *Mathematical Programming*, 186(1):185–208, 2021.

[18] Dan Garber and Noam Wolf. Frank-Wolfe with a nearest extreme point oracle. In *Conference on Learning Theory*, pages 2103–2132. PMLR, 2021.

[19] Saeed Ghadimi, Andrzej Ruszczynski, and Mengdi Wang. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM Journal on Optimization*, 30(1):960–979, 2020.

[20] Zaid Harchaoui, Anatoli Juditsky, and Arkadi Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1-2):75–112, 2015.

[21] Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR, 2019.

[22] Elad Hazan and Satyen Kale. Projection-free online learning. In *29th International Conference on Machine Learning, ICML 2012*, pages 521–528, 2012.

[23] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

[24] Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *International Conference on Machine Learning*, pages 1263–1271, 2016.

[25] Yifan Hu, Siqi Zhang, Xin Chen, and Niao He. Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[26] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR, 2013.

[27] Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *Advances in Neural Information Processing Systems*, pages 496–504, 2015.

[28] Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016.

[29] Evgeny Levitin and Boris Polyak. Constrained minimization methods. *USSR Computational mathematics and mathematical physics*, 6(5):1–50, 1966.

[30] Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive sgd with momentum. *arXiv preprint arXiv:2007.14294*, 2020.

[31] Zhipeng Lou, Wanrong Zhu, and Wei Biao Wu. Beyond sub-gaussian noises: Sharp concentration analysis for stochastic gradient descent. *Journal of Machine Learning Research*, 23:1–22, 2022.

[32] Liam Madden, Emiliano Dall'Anese, and Stephen Becker. High-probability convergence bounds for non-convex stochastic gradient descent. *arXiv preprint arXiv:2006.05610*, 2021.

[33] Athanasios Migdalas. A regularization of the frank—wolfe method and unification of certain nonlinear programming methods. *Mathematical Programming*, 65(1):331–345, 1994.

[34] Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *Journal of machine learning research*, 2020.

[35] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[36] Qi Qi, Zhishuai Guo, Yi Xu, Rong Jin, and Tianbao Yang. An online method for a class of distributionally robust optimization with non-convex objectives. *Advances in Neural Information Processing Systems*, 34, 2021.

[37] Qi Qi, Youzhi Luo, Zhao Xu, Shuiwang Ji, and Tianbao Yang. Stochastic optimization of areas under precision-recall curves with provable convergence. *Advances in Neural Information Processing Systems*, 34, 2021.

[38] Zi-Hao Qiu, Quanqi Hu, Yongjian Zhong, Lijun Zhang, and Tianbao Yang. Large-scale stochastic optimization of ndcg surrogates for deep learning with provable convergence. *arXiv preprint arXiv:2202.12183*, 2022.

[39] Chao Qu, Yan Li, and Huan Xu. Non-convex conditional gradient sliding. In *International Conference on Machine Learning*, pages 4208–4217. PMLR, 2018.

[40] Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic frank-wolfe methods for nonconvex optimization. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1244–1251. IEEE, 2016.

[41] David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric regression*. Number 12. Cambridge university press, 2003.

[42] Andrzej Ruszczyński. A linearization method for nonsmooth stochastic programming problems. *Mathematics of Operations Research*, 12(1):32–49, 1987.

[43] Andrzej Ruszczynski. A stochastic subgradient method for nonsmooth nonconvex multilevel composition optimization. *SIAM Journal on Control and Optimization*, 59(3):2301–2320, 2021.

[44] Andrzej Ruszczyński and Alexander Shapiro. Optimization of convex risk functions. *Mathematics of operations research*, 31(3):433–452, 2006.

[45] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[46] Gang Wang, Georgios B Giannakis, and Yonina C Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 64(2):773–794, 2017.

[47] Guanghui Wang, Ming Yang, Lijun Zhang, and Tianbao Yang. Momentum accelerates the convergence of stochastic AUPRC maximization. In *International Conference on Artificial Intelligence and Statistics*, pages 3753–3771. PMLR, 2022.

[48] Mengdi Wang, Ethan Fang, and Han Liu. Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449, 2017.

[49] Mengdi Wang, Ji Liu, and Ethan Fang. Accelerating stochastic composition optimization. In *Advances in Neural Information Processing Systems*, 2016.

[50] Shuoguang Yang, Mengdi Wang, and Ethan Fang. Multilevel stochastic gradient methods for nested composition optimization. *SIAM Journal on Optimization*, 29(1):616–659, 2019.

[51] Zhuoran Yang, Krishnakumar Balasubramanian, and Han Liu. High-dimensional non-gaussian single index models via thresholded score function estimation. In *International conference on machine learning*, pages 3851–3860. PMLR, 2017.

[52] Alp Yurtsever, Suvrit Sra, and Volkan Cevher. Conditional gradient methods via stochastic path-integrated differential estimator. In *International Conference on Machine Learning*, pages 7282–7291. PMLR, 2019.

[53] Junyu Zhang and Lin Xiao. Multilevel composite stochastic optimization via nested variance reduction. *SIAM Journal on Optimization*, 31(2):1131–1157, 2021.

[54] Mingrui Zhang, Zebang Shen, Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. One-sample Stochastic Frank-Wolfe. In *International Conference on Artificial Intelligence and Statistics*, pages 4012–4023. PMLR, 2020.

[55] Dongruo Zhou, Jinghui Chen, Yuan Cao, Yiqi Tang, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] This work considers a particular projection-free algorithm for a specific class of optimization problems. See also Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This is a theoretical paper which does not have any direct negative societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Assumption 1, 2, 3, 4.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We included a .ipynb file to reproduce the experimental results.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable?x [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Supplementary Materials

The supplementary materials are organized as follows. Appendix A provides motivating examples for stochastic multilevel optimization. Appendix B introduces the essential technical lemmas to complete the proof. We present the whole proofs of Theorem 2 and Theorem 3 in Appendix C and D. Finally, we present the high-probability convergence analysis particularly for the case when $T = 1$ in Appendix E.

A Motivating Examples

Problems of the form in (1) are generalizations of the standard constrained stochastic optimization problem which is obtained when $T = 1$, and arise in several machine learning applications. Some examples include sparse additive modeling in non-parametric statistics [48, Section 4.1], Bayesian optimization [2], model-agnostic meta-learning [7, 14], distributionally robust optimization [36], training graph neural networks [8], reinforcement learning [49, Setion 1.1] and AUPRC maximization [37, 47, 38]. Below, we provide a concrete motivating example from the field of risk-averse stochastic optimization [44].

The mean-deviation risk-averse optimization is given by the following optimization problem

$$\max_{x\in\mathcal{X}} \left\{ \mathbb{E}[U(x,\xi)] - \rho\mathbb{E}\left[\{\mathbb{E}[U(x,\xi)] - U(x,\xi)\}^2\right]^{1/2} \right\}.$$

As noted by [50], [43] and [4], the above problem is a stochastic 3-level composition optimization problem with

$$f_3(x) := (\mathbb{E}[U(x,\xi)], x) \qquad f_2((y,x)) := (y, \mathbb{E}[\{y - U(x,\xi)\}^2]) \qquad f_1((y,z)) := y - \sqrt{z+\delta}.$$

Here, $\delta > 0$ is added to make the square root function smooth. In particular, we consider a semi-parametric data generating process given by a sparse single-index model of the form $b = g(\langle a, x^*\rangle) + \zeta$, where $g : \mathbb{R} \to \mathbb{R}$ is called the link function, $x^* \in \mathbb{R}^d$ is assumed to be a sparse vector and $\langle\cdot,\cdot\rangle$ represents the Euclidean inner-product between two vectors. Such single-index models are widely used in statistics, machine learning and economics [41]. A standard choices of the link function g is the square function, in which case, the model is also called as the sparse phase retrieval model [46]. Here, a is the input data which is assumed to be independent of the noise ζ. In this case, $\xi := (a, b)$ and the if we consider the squared-loss, then $U(x,\xi) := (b - (\langle a, x\rangle)^2)^2$ and is non-convex in x. The goal is to estimate the sparse index vector x^* in a risk-averse manner, as they are well-known to provide stable solutions [50]. To encourage sparsity, the set $\mathcal{X}$ is the ℓ_1 ball [26].

B Technical Lemmas

Lemma 6. (Smoothness of Composite Functions [4]) *Assume that Assumption 2 holds.*

a) Define $F_i(x) = f_i \circ f_{i+1} \circ \cdots \circ f_T(x)$. Under , the gradient of F_i is Lipschitz continuous with the constant

$$L_{\nabla F_i} = \sum_{j=i}^{T} \left[L_{\nabla f_j} \prod_{l=i}^{j-1} L_{f_l} \prod_{l=j+1}^{T} L_{f_l}^2 \right].$$

b) Define

$$\begin{aligned} &R_1 = L_{\nabla f_1} L_{f_2} \cdots L_{f_T}, \qquad R_j = L_{f_1} \cdots L_{f_{j-1}} L_{\nabla f_j} L_{f_{j+1}} \cdots L_{f_T} / L_{f_j}, \quad 2 \le j \le T-1, \\ &C_2 = R_1, \qquad C_j = \sum_{i=1}^{j-2} R_i \left(\prod_{l=i+1}^{j-1} L_{f_l} \right), \quad 3 \le j \le T \end{aligned} \tag{17}$$

and let $u_{T+1} = x$. Then, for $T \geq 2$, we have

$$\left\| \nabla F(x) - \prod_{i=1}^{T} \nabla f_{T+1-i}(u_{T+2-i}) \right\| \leq \sum_{j=2}^{T} C_j \| f_j(u_{j+1}) - u_j \|. \tag{18}$$

Lemma 7. (Smoothness of $\eta(\cdot,\cdot)$ [19]) *For fixed $\beta > 0$ and, $\eta(x,z)$ defined in* (10)*, the gradient of $\eta(x,z)$ w.r.t. (x,z) is Lipschitz continuous with the constant $L_{\nabla\eta} = 2\sqrt{(1+\beta)^2 + \left(1+\frac{1}{2\beta}\right)^2}$.*

Lemma 8. (Convergnece of `ICG` [26]) *Let $\tilde{y}^k$ be the vector output by Algorithm 2 at step k, and y^k be the optimal solution of the subproblem 16, then under Assumption 1*

$$\frac{\beta}{2}\|\tilde{y}^k - y^k\|^2 \leq H_k(\tilde{y}^k) - H_k(y^k) \leq \frac{2\beta D_{\mathcal{X}}^2(1+\delta)}{t_k + 2}$$

where δ defined in Algorithm 2 is the quality of the linear minimization procedure.

Proof of Lemma 8. The result is obtained by applying Theorem 1 in [26] to H_k and noting that the curvature constant $C_{H_k} = \beta D_{\mathcal{X}}^2, \forall k \geq 0$. □

C Proof of Theorem 2

To establish the rate of convergence for Algorithm 1 in Theorem 2, we first present Lemma 9 and Lemma 10 regarding the basic recursion on the errors in estimating the inner function values and the order of $\mathbb{E}[\|u_i^{k+1} - u_i^k\|^2 | \mathscr{F}_k]$. The proofs follow [4] with minor modifications. We present the complete proofs below for the reader's convenience.

Lemma 9. *Let $\{x^k\}_{k\geq 0}$ and $\{u_i^k\}_{k\geq 0}$ be generated by Algorithm 1 and $u_{T+1} = x$. Define, $1 \leq i \leq T$,*

$$\begin{aligned} \Delta_{G_i}^{k+1} &:= f_i(u_{i+1}^k) - G_i^{k+1}, \quad \Delta_{J_i}^{k+1} := \nabla f_i(u_{i+1}^k) - J_i^{k+1}, \\ e_i^k &:= f_i(u_{i+1}^{k+1}) - f_i(u_{i+1}^k) - \langle \nabla f_i(u_{i+1}^k), u_{i+1}^{k+1} - u_{i+1}^k \rangle. \end{aligned} \tag{19}$$

Under Assumption 2, we have, for $1 \leq i \leq T$,

$$\begin{aligned} \|f_i(u_{i+1}^{k+1}) - u_i^{k+1}\|^2 \leq (1-\tau_k)\|f_i(u_{i+1}^k) - u_i^k\|^2 + \tau_k^2 \|\Delta_{G_i}^{k+1}\|^2 + \dot{r}_i^{k+1} \\ + \left[4L_{f_i}^2 + L_{\nabla f_i}\|f_i(u_{i+1}^k) - u_i^k\| + \|\Delta_{J_i}^{k+1}\|^2\right] \|u_{i+1}^{k+1} - u_{i+1}^k\|^2, \end{aligned} \tag{20}$$

and

$$\|u_i^{k+1} - u_i^k\|^2 \leq \tau_k^2 \left[2\|f_i(u_{i+1}^k) - u_i^k\|^2 + \|\Delta_{G_i}^{k+1}\|^2\right] + 2\|J_i^{k+1}\|^2 \|u_{i+1}^{k+1} - u_{i+1}^k\|^2 + \ddot{r}_i^{k+1} \tag{21}$$

where

$$\begin{aligned} \dot{r}_i^{k+1} &:= 2\tau_k \langle \Delta_{G_i}^{k+1}, e_i^k + (1-\tau_k)(f_i(u_{i+1}^k) - u_i^k) + {\Delta_{J_i}^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k) \rangle \\ &\quad + 2\langle {\Delta_{J_i}^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k), e_i^k + (1-\tau_k)(f_i(u_{i+1}^k) - u_i^k) \rangle, \\ \ddot{r}_i^{k+1} &:= \tau_k \langle -\Delta_{G_i}^{k+1}, \tau_k(f_i(u_{i+1}^k) - u_i^k) + {J_i^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k) \rangle. \end{aligned} \tag{22}$$

Proof. We first prove part (20). By the definitions in (19), (22), for any $1 \leq i \leq T$, we have

$$\begin{aligned} &\|f_i(u_{i+1}^{k+1}) - u_i^{k+1}\|^2 \\ =&\|e_i^k + f_i(u_{i+1}^k) + \nabla f_i(u_{i+1}^k)^\top (u_{i+1}^{k+1} - u_{i+1}^k) - (1-\tau_k)u_i^k - \tau_k G_i^{k+1} - {J_i^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k)\|^2 \\ =&\|e_i^k + {\Delta_{J_i}^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k) + (1-\tau_k)(f_i(u_{i+1}^k) - u_i^k) + \tau_k \Delta_{G_i}^{k+1}\|^2 \\ =&\|{\Delta_{J_i}^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k)\|^2 + \|e_i^k + (1-\tau_k)(f_i(u_{i+1}^k) - u_i^k)\|^2 + \tau_k^2 \|\Delta_{G_i}^{k+1}\|^2 + \dot{r}_i^{k+1} \\ \leq&\|e_i^k + (1-\tau_k)(f_i(u_{i+1}^k) - u_i^k)\|^2 + \tau_k^2 \|\Delta_{G_i}^{k+1}\|^2 + \|\Delta_{J_i}^{k+1}\|^2 \|u_{i+1}^{k+1} - u_{i+1}^k\|^2 + \dot{r}_i^{k+1} \\ \leq&(1-\tau_k)\|f_i(u_{i+1}^k) - u_i^k\|^2 + \|e_i^k\|^2 + 2(1-\tau_k)\|e_i^k\| \|f_i(u_{i+1}^k) - u_i^k\| + \tau_k^2 \|\Delta_{G_i}^{k+1}\|^2 \\ &\quad + \|\Delta_{J_i}^{k+1}\|^2 \|u_{i+1}^{k+1} - u_{i+1}^k\|^2 + \dot{r}_i^{k+1}. \end{aligned}$$

Furthermore, with Assumption 2, we have

$$\|e_i^k\| \le \frac{L_{\nabla f_i}}{2}\|u_{i+1}^{k+1} - u_{i+1}^k\|^2, \qquad \|e_i^k\|^2 \le 4L_{f_i}^2\|u_{i+1}^{k+1} - u_{i+1}^k\|^2, \tag{23}$$

which leads to (20). To show (21), with the update rule given by (8) and the definitions in (19), we have, for $1 \le i \le T$,

$$\begin{aligned}
&\|u_i^{k+1} - u_i^k\|^2 \\
=&\|\tau_k(G_i^{k+1} - u_i^k) + \langle J_i^{k+1}, u_{i+1}^{k+1} - u_{i+1}^k\rangle\|^2 \\
=&\tau_k^2\|G_i^{k+1} - u_i^k\|^2 + \|J_i^{k+1\top}(u_{i+1}^{k+1} - u_{i+1}^k)\|^2 + 2\tau_k\langle G_i^{k+1} - u_i^k, J_i^{k+1\top}(u_{i+1}^{k+1} - u_{i+1}^k)\rangle \\
=&\tau_k^2\|G_i^{k+1} - u_i^k\|^2 + \|J_i^{k+1\top}(u_{i+1}^{k+1} - u_{i+1}^k)\|^2 + 2\tau_k\langle f_i(u_{i+1}^k) - u_i^k, J_i^{k+1\top}(u_{i+1}^{k+1} - u_{i+1}^k)\rangle \\
&\quad + 2\tau_k\langle -\Delta_{G_i}^{k+1}, J_i^{k+1\top}(u_{i+1}^{k+1} - u_{i+1}^k)\rangle \\
\le&\tau_k^2\|G_i^{k+1} - u_i^k\|^2 + 2\|J_i^{k+1}\|^2\|u_{i+1}^{k+1} - u_{i+1}^k\|^2 + \tau_k^2\|f_i(u_{i+1}^k) - u_i^k\|^2 \\
&\quad + 2\tau_k\langle -\Delta_{G_i}^{k+1}, J_i^{k+1\top}(u_{i+1}^{k+1} - u_{i+1}^k)\rangle \\
=&2\tau_k^2\|f_i(u_{i+1}^k) - u_i^k\|^2 + \tau_k^2\|\Delta_{G_i}^{k+1}\|^2 + 2\|J_i^{k+1}\|^2\|u_{i+1}^{k+1} - u_{i+1}^k\|^2 \\
&\quad + 2\tau_k\langle -\Delta_{G_i}^{k+1}, \tau_k(f_i(u_{i+1}^k) - u_i^k) + J_i^{k+1\top}(u_{i+1}^{k+1} - u_{i+1}^k)\rangle.
\end{aligned}$$

where the inequality comes from the fact that $\|J_i^{k+1\top}(u_{i+1}^{k+1} - u_{i+1}^k)\|^2 \le \|J_i^{k+1}\|^2\|u_{i+1}^{k+1} - u_{i+1}^k\|^2$ and $2\tau_k\langle f_i(u_{i+1}^k) - u_i^k, J_i^{k+1\top}(u_{i+1}^{k+1} - u_{i+1}^k)\rangle \le \|J_i^{k+1\top}(u_{i+1}^{k+1} - u_{i+1}^k)\|^2 + \tau_k^2\|f_i(u_{i+1}^k) - u_i^k\|^2$. □

Lemma 10. *Let $u_{T+1} = x$. Under Assumption 2, 3, and with the choice of $\tau_0 = 1$, we have, for $1 \le i \le T$ and $k \ge 0$,*

$$\mathbb{E}[\|f_i(u_{i+1}^{k+1}) - u_i^{k+1}\|^2|\mathscr{F}_k] \le \sigma_{G_i}^2 + (4L_{f_i}^2 + \sigma_{J_i}^2)c_{i+1}, \tag{24}$$

$$\mathbb{E}[\|u_i^{k+1} - u_i^k\|^2|\mathscr{F}_k] \le c_i\tau_k^2, \tag{25}$$

where

$$c_i := 3\sigma_{G_i}^2 + 2(4L_{f_i}^2 + \sigma_{J_i}^2 + \hat{\sigma}_{J_i}^2)c_{i+1}, \qquad c_{T+1} = D_{\mathcal{X}}^2. \tag{26}$$

Proof. By the update rule given in (8) and the definitions in (19), for $1 \le i \le T$ and $k \ge 0$, we have

$$f_i(u_{i+1}^{k+1}) - u_i^{k+1} = (1 - \tau_k)(f_i(u_{i+1}^k) - u_i^k) + \mathbf{D}_{k,i},$$

where $\mathbf{D}_{k,i} := e_i^k + \tau_k\Delta_{G_i}^{k+1} + \Delta_{J_i}^{k+1\top}(u_{i+1}^{k+1} - u_{i+1}^k)$. With the convexity of $\|\cdot\|^2$, we can further obtain

$$\|f_i(u_{i+1}^{k+1}) - u_i^{k+1}\|^2 \le (1 - \tau_k)\|f_i(u_{i+1}^k) - u_i^k\|^2 + \frac{1}{\tau_k}\|\mathbf{D}_{k,i}\|^2, \quad \forall k \ge 0. \tag{27}$$

Moreover, under Assumption 3, we have, for $1 \le i \le T$ and $k \ge 0$,

$$\begin{aligned}
\mathbb{E}[\|\mathbf{D}_{k,i}\|^2|\mathscr{F}_k] &= \mathbb{E}[\|e_i^k\|^2|\mathscr{F}_k] + \tau_k^2\mathbb{E}[\|\Delta_{G_i}^{k+1}\|^2|\mathscr{F}_k] + \mathbb{E}[\|\Delta_{J_i}^{k+1\top}(u_{i+1}^{k+1} - u_{i+1}^k)\|^2|\mathscr{F}_k] \\
&\le \tau_k^2\mathbb{E}[\|\Delta_{G_i}^{k+1}\|^2|\mathscr{F}_k] + \left(4L_{f_i}^2 + \mathbb{E}[\|\Delta_{J_i}^{k+1}\|^2|\mathscr{F}_k]\right)\mathbb{E}[\|u_{i+1}^{k+1} - u_{i+1}^k\|^2|\mathscr{F}_k] \\
&\le \tau_k^2\sigma_{G_i}^2 + \left(4L_{f_i}^2 + \sigma_{J_i}^2\right)\mathbb{E}[\|u_{i+1}^{k+1} - u_{i+1}^k\|^2|\mathscr{F}_k].
\end{aligned} \tag{28}$$

where the second inequality follows from (23). Setting $i = T$ in the inequality above and noting that $u_{T+1}^k = x^k$, we have

$$\mathbb{E}[\|\mathbf{D}_{k,T}\|^2|\mathscr{F}_k] \le \tau_k^2\left[\sigma_{G_T}^2 + (4L_{f_T}^2 + \sigma_{J_T}^2)D_{\mathcal{X}}^2\right], \quad \forall k \ge 0.$$

Thus, with the choice of $\tau_0 = 1$, we obtain

$$\mathbb{E}[\|f_T(x^k) - u_T^k\|^2|\mathscr{F}_k] \le \sigma_{G_T}^2 + (4L_{f_T}^2 + \sigma_{J_T}^2)D_{\mathcal{X}}^2, \quad \forall k \ge 1.$$

Taking expectation of both sides of (21) conditioning on $\mathscr{F}_k$, and under Assumption 3, we obtain

$$\mathbb{E}[\|u_i^{k+1}-u_i^k\|^2|\mathscr{F}_k] \le \tau_k^2 \mathbb{E}\left[2\|f_i(x^k)-u_i^k\|^2 + \|\Delta_{G_i}^{k+1}\|^2 + \frac{2}{\tau_k^2}\|J_i^{k+1}\|^2\|u_{i+1}^{k+1}-u_{i+1}^k\|^2 \middle| \mathscr{F}_k\right]. \tag{29}$$

Setting $i=T$ in the inequality above, we have

$$\mathbb{E}[\|u_T^{k+1}-u_T^k\|^2|\mathscr{F}_k] \le \tau_k^2\left[3\sigma_{G_T}^2 + 2(4L_{f_T}^2+\sigma_{J_T}^2+\hat{\sigma}_{J_T}^2)D_{\mathcal{X}}^2\right], \quad \forall k\ge 1.$$

This completes the proof of (24) and (25) when $i=T$. We now use backward induction to complete the proof. By the above result, the base case of $i=T$ holds. Assume that (25) hold when $i=j$ for some $1<j\le T$, i.e., $\mathbb{E}[\|u_j^{k+1}-u_j^k\|^2|\mathscr{F}_k]\le c_j\tau_k^2, \forall k\ge 0$. Then, setting $i=j-1$ in (28), we obtain

$$\mathbb{E}[\|\mathbf{D}_{k,j-1}\|^2|\mathscr{F}_k] \le \tau_k^2\left[\sigma_{G_{j-1}}^2 + (4L_{f_{j-1}}^2+\sigma_{J_{j-1}}^2)c_j\right], \quad \forall k\ge 0.$$

Furthermore, with (27) and the choice of $\tau_0=1$, we have

$$\mathbb{E}[\|f_{j-1}(u_j^{k+1})-u_{j-1}^{k+1}\|^2|\mathscr{F}_k] \le \sigma_{G_{j-1}}^2 + (4L_{f_{j-1}}^2+\sigma_{J_{j-1}}^2)c_j, \quad \forall k\ge 0.$$

which together with (29), imply that

$$\mathbb{E}[\|u_{j-1}^{k+1}-u_{j-1}^k\|^2|\mathscr{F}_k] \le c_{j-1}\tau_k^2, \quad \forall k\ge 0.$$

□

We now leverage the merit function defined in (9) and provide a basic inequality for establishing convergence analysis of Algorithm 1 in Lemma 11. In Proposition 3, we show the boundedness of the term $\mathbf{R}_k$ appearing on the right hand side of (30) in expectation. These two results form the crucial steps in establishing the convergence analysis of Algorithm 1.

Lemma 11. *Let $\{x^k,z^k,u^k\}_{k\ge 0}$ be the sequence generated by Algorithm 1, the merit function $W_{\alpha,\gamma}(\cdot,\cdot,\cdot)$ be defined in* (9) *with positive constants $\{\alpha,\{\gamma_i\}_{1\le i\le T}\}$, and $u_{T+1}=x$. Under Assumption 2, for any $\beta>0$, let*

$$\beta_k\equiv\beta, \quad \alpha=\frac{\beta}{20L_{\nabla F}^2}, \quad \gamma_1=\frac{\beta}{2}, \quad \gamma_j=\left(2\alpha+\frac{1}{4\alpha L_{\nabla F}^2}\right)(T-1)C_j^2+\frac{\beta}{2}, \quad 2\le j\le T,$$

where C_j's are defined in (17)*. Then, $\forall N\ge 0$*

$$\begin{aligned}&\sum_{k=0}^N \tau_k\left(\beta\left[\|d^k\|^2+\sum_{i=1}^T\|f_i(u_{i+1}^k)-u_i^k\|^2\right]+\frac{\beta}{20L_{\nabla F}^2}\|\nabla F(x^k)-z^k\|^2\right)\\ &\qquad\le 2W_0+2\sum_{k=0}^N\mathbf{R}_k+\left(\frac{24}{5}+\frac{40L_{\nabla F}^2}{\beta^2}\right)\sum_{k=0}^N\tau_k\left(H_k(\tilde{y}^k)-H_k(y^k)\right),\end{aligned} \tag{30}$$

where $d^k:=y^k-x^k$, $H_k(\cdot), y^k$ are defined in (16)*, and*

$$\begin{aligned}\mathbf{R}_k :=& \sum_{i=1}^T\gamma_i\left[4L_{f_i}^2+L_{\nabla f_i}\|f_i(u_{i+1}^k)-u_i^k\|+\|\Delta_{J_i}^{k+1}\|^2\right]\|u_{i+1}^{k+1}-u_{i+1}^k\|^2\\ &+\tau_k^2\left[\frac{L_{\nabla F}+L_{\nabla\eta}}{2}D_{\mathcal{X}}^2+\sum_{i=1}^T\gamma_i\|\Delta_{G_i}^{k+1}\|^2+\alpha\|\Delta^{k+1}\|^2\right]\\ &+\tau_k\left[\langle d^k,\Delta^{k+1}\rangle+\sum_{i=1}^T\gamma_i\dot{r}_i^{k+1}+2\alpha\,\ddot{\cdot}^{k+1}\right]+\frac{L_{\nabla\eta}}{2}\|z^{k+1}-z^k\|^2,\end{aligned} \tag{31}$$

$$\Delta^{k+1} := \prod_{i=1}^T\nabla f_{T+1-i}(u_{T+2-i}^k)-\prod_{i=1}^T J_{T-i+1}^{k+1},$$

$$\begin{aligned}\ddot{\cdot}^{k+1} :=& \langle\Delta^{k+1},(1-\tau_k)[\nabla F(x^k)-z^k]+\tau_k[\nabla F(x^k)-\prod_{i=1}^T\nabla f_{T+1-i}(u_{T+2-i}^k)]\\ &+\nabla F(x^{k+1})-\nabla F(x^k)\rangle,\end{aligned}$$

$\Delta_{G_i}^{k+1}$, $\Delta_{J_i}^{k+1}$ are defined in (19)*, and $\dot{r}_i^{k+1}$ is defined in* (22)*.*

Proof. We first bound $F(x^{k+1}) - F(x^k)$. By the Lipschitzness of ∇F (Lemma 6), we have

$$\begin{aligned}
&F(x^{k+1}) - F(x^k) \leq \langle \nabla F(x^k), x^{k+1} - x^k \rangle + \frac{L_{\nabla F}\tau_k^2}{2}\|\tilde{d}^k\|^2 \\
&= \tau_k \langle \nabla F(x^k), d^k \rangle + \tau_k \langle \nabla F(x^k) - z^k, \tilde{y}^k - y^k \rangle - \tau_k \langle \beta d^k, \tilde{y}^k - y^k \rangle \\
&\quad + \tau_k \langle z^k + \beta d^k, \tilde{y}^k - y^k \rangle + \frac{L_{\nabla F}\tau_k^2}{2}\|\tilde{d}^k\|^2 \\
&\leq \tau_k \langle \nabla F(x^k), d^k \rangle + \tau_k \|\nabla F(x^k) - z^k\| \|\tilde{y}^k - y^k\| + \tau_k \langle z^k + \beta d^k, \tilde{y}^k - y^k \rangle \\
&\quad + \tau_k \beta \|d^k\| \|\tilde{y}^k - y^k\| + \frac{L_{\nabla F}\tau_k^2}{2}\|\tilde{d}^k\|^2.
\end{aligned} \tag{32}$$

We then provide a bound for $\eta(x^k, z^k) - \eta(x^{k+1}, z^{k+1})$. By the lipschitzness of $\nabla \eta$ (Lemma 7) with the partial gradients of $\nabla \eta$ given by

$$\nabla_x \eta(x^k, z^k) = -z^k - \beta d^k, \quad \nabla_z \eta(x^k, z^k) = d^k,$$

we have

$$\begin{aligned}
&\eta(x^k, z^k) - \eta(x^{k+1}, z^{k+1}) \\
&\leq \left\langle z^k + \beta d^k, x^{k+1} - x^k \right\rangle - \left\langle d^k, z^{k+1} - z^k \right\rangle + \frac{L_{\nabla \eta}}{2}\left[\|x^{k+1} - x^k\|^2 + \|z^{k+1} - z^k\|^2\right] \\
&= \tau_k \left\langle 2z^k + \beta d^k, d^k \right\rangle + \tau_k \left\langle z^k + \beta d^k, \tilde{d}^k - d^k \right\rangle - \tau_k \left\langle d^k, \prod_{i=1}^{T} J_{T-i+1}^{k+1} \right\rangle \\
&\quad + \frac{L_{\nabla \eta}}{2}\left[\tau_k^2 \|\tilde{d}^k\|^2 + \|z^{k+1} - z^k\|^2\right],
\end{aligned} \tag{33}$$

where the second equality comes from (6) and (7). Due to the optimality condition of in the definition of y^k, we have $\left\langle z^k + \beta d^k, x - y^k \right\rangle \geq 0$ for all $x \in \mathcal{X}$, which together with the choice of $x = x^k$ implies that

$$\langle z^k, d^k \rangle + \beta \|d^k\|^2 \leq 0. \tag{34}$$

Thus, combining (33) with (34), we obtain

$$\begin{aligned}
\eta(x^k, z^k) - \eta(x^{k+1}, z^{k+1}) \leq -\beta \tau_k \|d^k\|^2 + \tau_k \left\langle z^k + \beta d^k, \tilde{y}^k - y^k \right\rangle - \tau_k \left\langle d^k, \prod_{i=1}^{T} J_{T-i+1}^{k+1} \right\rangle \\
+ \frac{L_{\nabla \eta}}{2}\left[\tau_k^2 \|\tilde{d}^k\|^2 + \|z^{k+1} - z^k\|^2\right].
\end{aligned} \tag{35}$$

In addition, by Lemma 18, we have

$$\langle d^k, \nabla F(x^k) - \prod_{i=1}^{T} \nabla f_{T+1-i}(u_{T+2-i}^k) \rangle \leq \sum_{j=2}^{T} C_j \|d^k\| \|f_j(u_{j+1}^k) - u_j^k\|. \tag{36}$$

Then combing (32), (35), (36), we have

$$\begin{aligned}
&[F(x^{k+1}) - \eta(x^{k+1}, z^{k+1})] - [F(x^k) - \eta(x^k, z^k)] \\
&\leq \tau_k \Bigg\{ -\beta \|d^k\|^2 + \sum_{j=2}^{T} C_j \|d^k\| \|f_j(u_{j+1}^k) - u_j^k\| + \langle d^k, \Delta^{k+1} \rangle + 2\langle z^k + \beta d^k, \tilde{y}^k - y^k \rangle \\
&\quad + \left[\beta \|d^k\| + \|\nabla F(x^k) - z^k\|\right] \|\tilde{y}^k - y^k\| \Bigg\} + \frac{L_{\nabla F} + L_{\nabla \eta}}{2}\tau_k^2 \|\tilde{d}^k\|^2 + \frac{L_{\nabla \eta}}{2}\|z^{k+1} - z^k\|^2.
\end{aligned} \tag{37}$$

Furthermore, defining

$$\varkappa^k := \nabla F(x^k) - \prod_{i=1}^{T} \nabla f_{T+1-i}(u_{T+2-i}^k), \qquad \bar{\varkappa}^k := \frac{\nabla F(x^{k+1}) - \nabla F(x^k)}{\tau_k},$$

and by the update rule given by (7), we have

$$\begin{aligned}
&\|\nabla F(x^{k+1}) - z^{k+1}\|^2 \\
&= \|(1-\tau_k)[\nabla F(x^k) - z^k] + \tau_k[\varkappa^k + \bar{\varkappa}^k + \Delta^{k+1}]\|^2 \\
&= \|(1-\tau_k)[\nabla F(x^k) - z^k] + \tau_k[\varkappa^k + \bar{\varkappa}^k]\|^2 + \tau_k^2\|\Delta^{k+1}\|^2 + 2\tau_k\, \ddot{r}^{k+1} \\
&\le (1-\tau_k)\|\nabla F(x^k) - z^k\|^2 + 2\tau_k\left[\|\varkappa^k\|^2 + \|\bar{\varkappa}^k\|^2\right] + \tau_k^2\|\Delta^{k+1}\|^2 + 2\tau_k\, \ddot{r}^{k+1} \\
&\le (1-\tau_k)\|\nabla F(x^k) - z^k\|^2 + \tau_k^2\|\Delta^{k+1}\|^2 \\
&+ 2\tau_k\left[(T-1)\sum_{j=2}^{T} C_j^2\|f_j(u_{j+1}) - u_j\|^2 + 2L_{\nabla F}^2(\|d^k\|^2 + \|\tilde{y}^k - y^k\|^2) + \ddot{r}^{k+1}\right].
\end{aligned} \tag{38}$$

where $\ddot{r}^{k+1} := \langle \Delta^{k+1}, (1-\tau_k)[\nabla F(x^k) - z^k] + \tau_k[\varkappa^k + \bar{\varkappa}^k]\rangle$ and the last inequality comes from two fact that $\|\bar{\varkappa}^k\|^2 \le 2L_{\nabla F}^2(\|d^k\|^2 + \|\tilde{y}^k - y^k\|^2)$ and

$$\|\varkappa^k\|^2 = \left\|\nabla F(x^k) - \prod_{i=1}^{T}\nabla f_{T+1-i}(u_{T+2-i}^k)\right\|^2 \le (T-1)\sum_{j=2}^{T} C_j^2\|f_j(u_{j+1}) - u_j\|^2.$$

The above upper bound for the term $\|\varkappa^k\|^2$ is obtained by leveraging Lemma 18 and the fact that $(\sum_{i=1}^n a_i) \le n\sum_{i=1}^n a_i^2$ for non-negative sequence $(a_i)_{1\le i\le n}$.

Moreover, by Lemma 9, we have, for $1 \le i \le T$,

$$\begin{aligned}
\|f_i(u_{i+1}^{k+1}) - u_i^{k+1}\|^2 - \|f_i(u_{i+1}^k) - u_i^k\|^2 &\le -\tau_k\|f_i(u_{i+1}^k) - u_i^k\|^2 + \tau_k^2\|\Delta_{G_i}^{k+1}\|^2 + \dot{r}_i^{k+1} \\
&+ \left[4L_{f_i}^2 + L_{\nabla f_i}\|f_i(u_{i+1}^k) - u_i^k\| + \|\Delta_{J_i}^{k+1}\|^2\right]\|u_{i+1}^{k+1} - u_{i+1}^k\|^2,
\end{aligned} \tag{39}$$

Finally, multiplying both sides of (39) by γ_i for $i = 1, \ldots, T$ and both sides of (38) by α, adding them to (37), rearranging the terms, and noting that $\langle z^k + \beta d^k, \tilde{y}^k - y^k\rangle = H_k(\tilde{y}^k) - H_k(y^k) - (\beta/2)\|\tilde{y}^k - y^k\|^2$ due to the quadratic structure of H_k and $\|\tilde{d}^k\|^2 \le D_{\mathcal{X}}^2$, we obtain

$$W_{k+1} - W_k \le \tau_k\mathbf{A}_k + \mathbf{R}_k \tag{40}$$

where $\mathbf{R}_k$ is defined in (31) and

$$\begin{aligned}
\mathbf{A}_k := &\left(-\beta + 4\alpha L_{\nabla F}^2\right)\|d^k\|^2 + \sum_{j=2}^{T}\left(-\gamma_j + 2\alpha(T-1)C_j^2\right)\|f_j(u_{j+1}^k) - u_j^k\|^2 \\
&- \gamma_1\|f_1(u_2^k) - u_1^k\|^2 - \alpha\|\nabla F(x^k) - z^k\|^2 + \sum_{j=2}^{T} C_j\|d^k\|\|f_j(u_{j+1}) - u_j\| \\
&+ \left(\beta\|d^k\| + \|\nabla F(x^k) - z^k\|\right)\|\tilde{y}^k - y^k\| + \left(4\alpha L_{\nabla F}^2 - \beta\right)\|\tilde{y}^k - y^k\|^2 \\
&+ 2\left(H_k(\tilde{y}^k) - H_k(y^k)\right).
\end{aligned}$$

We can further provide a simplified upper bound for $\mathbf{A}_k$. By Young's inequality, we have

$$\begin{aligned}
\beta\|d^k\|\|\tilde{y}^k - y^k\| &\le \frac{\beta}{4}\|d^k\|^2 + \beta\|\tilde{y}^k - y^k\|^2, \\
\|\nabla F(x^k) - z^k\|\|\tilde{y}^k - y^k\| &\le \frac{\alpha}{2}\|\nabla F(x^k) - z^k\|^2 + \frac{1}{2\alpha}\|\tilde{y}^k - y^k\|^2 \\
C_j\|d^k\|\|f_j(u_{j+1}) - u_j\| &\le \frac{\alpha L_{\nabla F}^2}{T-1}\|d^k\|^2 + \frac{(T-1)C_j^2}{4\alpha L_{\nabla F}^2}\|f_j(u_{j+1}) - u_j\|^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbf{A}_k \le &\left(-\frac{3\beta}{4} + 5\alpha L_{\nabla F}^2\right)\|d^k\|^2 - \gamma_1\|f_1(u_2^k) - u_1^k\|^2 - \frac{\alpha}{2}\|\nabla F(x^k) - z^k\|^2 \\
&+ \sum_{j=2}^{T}\left(-\gamma_j + \left(2\alpha + \frac{1}{4\alpha L_{\nabla F}^2}\right)(T-1)C_j^2\right)\|f_j(u_{j+1}^k) - u_j^k\|^2 \\
&+ \left(4\alpha L_{\nabla F}^2 + \frac{1}{2\alpha}\right)\|\tilde{y}^k - y^k\|^2 + 2\left(H_k(\tilde{y}^k) - H_k(y^k)\right)
\end{aligned}$$

For any $\beta > 0$, let

$$\alpha = \frac{\beta}{20L_{\nabla F}^2}, \qquad \gamma_1 = \frac{\beta}{2}, \qquad \gamma_j = \left(2\alpha + \frac{1}{4\alpha L_{\nabla F}^2}\right)(T-1)C_j^2 + \frac{\beta}{2}, \quad 2 \le j \le T$$

Then, we have

$$\begin{aligned}\mathbf{A}_k \le -\frac{\beta}{2}\left(\|d^k\|^2 + \sum_{i=1}^{T}\|f_i(u_{i+1}^k) - u_i^k\|^2\right) - \frac{\beta}{40L_{\nabla F}^2}\|\nabla F(x^k) - z^k\|^2 \\ + \left(\frac{12}{5} + \frac{20L_{\nabla F}^2}{\beta^2}\right)\left(H_k(\tilde{y}^k) - H_k(y^k)\right).\end{aligned} \tag{41}$$

As a result of (40) and (41), we can further obtain

$$\begin{aligned}&\tau_k\left(\beta\left[\|d^k\|^2 + \sum_{i=1}^{T}\|f_i(u_{i+1}^k) - u_i^k\|^2\right] + \frac{\beta}{20L_{\nabla F}^2}\|\nabla F(x^k) - z^k\|^2\right) \\ &\le 2W_k - 2W_{k+1} + 2\mathbf{R}_k + \tau_k\left(\frac{24}{5} + \frac{40L_{\nabla F}^2}{\beta^2}\right)\left(H_k(\tilde{y}^k) - H_k(y^k)\right),\end{aligned}$$

which immediately implies (30) by telescoping. □

Proposition 3. *Let* $\mathbf{R}_k$ *be defined in* (31) *and* $\tau_0 = 1$. *Then, under Assumption 3, we have*

$$\mathbb{E}[\mathbf{R}_k|\mathscr{F}_k] \le \hat{\sigma}^2\tau_k^2, \quad \forall k \ge 1,$$

where

$$\begin{aligned}\hat{\sigma}^2 :=& \sum_{i=1}^{T}\gamma_i\left(\left[4L_{\nabla f_i}^2 + L_{\nabla f_i}\sqrt{\sigma_{G_i}^2 + (4L_{f_i}^2 + \sigma_{J_i}^2)c_{i+1}} + \sigma_{J_i}^2\right]c_{i+1} + \sigma_{G_i}^2\right) \\ &+ (\alpha + 2L_\eta)\prod_{i=1}^{T}\hat{\sigma}_{J_i}^2 + \frac{L_{\nabla F} + L_\eta}{2}D_{\mathcal{X}}^2.\end{aligned} \tag{42}$$

Proof. Note that under Assumption 3, we have, for $1 \le i \le T$,

$$\begin{aligned}&\mathbb{E}[\Delta^{k+1}|\mathscr{F}_k] = 0, \quad \mathbb{E}[\dot{r}_i^{k+1}|\mathscr{F}_k] = 0, \quad \mathbb{E}[\ddot{r}^{k+1}|\mathscr{F}_k] = 0, \\ &\mathbb{E}[\|\Delta_{G_i}^{k+1}\|^2|\mathscr{F}_k] \le \sigma_{G_i}^2, \quad \mathbb{E}[\|\Delta_{J_i}^{k+1}\|^2|\mathscr{F}_k] \le \sigma_{J_i}^2,\end{aligned}$$

and

$$\mathbb{E}[\|\Delta^{k+1}\|^2|\mathscr{F}_k] \le \mathbb{E}\left[\left\|\prod_{i=1}^{T}J_{T-i+1}^{k+1}\right\|^2\middle|\mathscr{F}_k\right] \le \prod_{i=1}^{T}\mathbb{E}\left[\left\|J_{T-i+1}^{k+1}\right\|^2\middle|\mathscr{F}_k\right] \le \prod_{i=1}^{T}\hat{\sigma}_{J_i}^2.$$

In addition, by Lemma 9 and Hölder's inequality. we have $\mathbb{E}[\|u_i^{k+1} - u_i^k\|^2|\mathscr{F}_k] \le c_i\tau_k^2$ and

$$\begin{aligned}&\mathbb{E}[\|f_i(u_{i+1}^{k+1}) - u_i^{k+1}\|\|u_i^{k+1} - u_i^k\|^2|\mathscr{F}_k] \\ &\le \mathbb{E}[\|f_i(u_{i+1}^{k+1}) - u_i^{k+1}\||\mathscr{F}_k]\mathbb{E}[\|u_i^{k+1} - u_i^k\|^2|\mathscr{F}_k] \\ &\le \left(\mathbb{E}[\|f_i(u_{i+1}^{k+1}) - u_i^{k+1}\||\mathscr{F}_k]\right)^{\frac{1}{2}}\mathbb{E}[\|u_i^{k+1} - u_i^k\|^2|\mathscr{F}_k] \\ &\le c_i\sqrt{\sigma_{G_i}^2 + (4L_{f_i}^2 + \sigma_{J_i}^2)c_{i+1}}\,\tau_k^2.\end{aligned}$$

Lastly, from eq.(28) of Proposition 2.1 in [4], we have for any $k \ge 1$,

$$\mathbb{E}[\|z^{k+1} - z^k\|^2|\mathscr{F}_k] \le 4\tau_k^2\prod_{i=1}^{T}\hat{\sigma}_{J_i}^2.$$

The proof is completed by combing all above observations with the expression of $\mathbf{R}_k$ in (31). □

Proof of Theorem 2. We now present the proof of Theorem 2. Note that by Lemma 11 and given values of α, γ in (12), we obtain

$$\sum_{k=1}^{N} \tau_k \left[\beta \left(\|d^k\|^2 + \sum_{i=1}^{T} \|f_i(u_{i+1}^k) - u_i^k\|^2 \right) + \frac{\beta}{20L_{\nabla F}^2} \|\nabla F(x^k) - z^k\|^2 \right]$$
$$\leq 2W_{\alpha,\gamma}(x^0, z^0, u^0) + 2\sum_{k=0}^{N} \mathbf{R}_k + \left(\frac{24}{5} + \frac{40L_{\nabla F}^2}{\beta^2} \right) \sum_{k=0}^{N} \tau_k \left(H_k(\tilde{y}^k) - H_k(y^k) \right),$$

Taking expectation of both sides and noting that $\mathbb{E}[\mathbf{R}_k | \mathscr{F}_k] \leq \hat{\sigma}^2 \tau_k^2$ by Proposition 3, we have

$$\begin{aligned} &\sum_{k=1}^{N} \tau_k \mathbb{E} \left[\rho \left(\|d^k\|^2 + \sum_{i=1}^{T} \|f_i(u_{i+1}^k) - u_i^k\|^2 \right) + \alpha \|\nabla F(x^k) - z^k\|^2 \middle| \mathscr{F}_{k-1} \right] \\ &\leq 2W_{\alpha,\gamma}(x^0, z^0, u^0) + 2\hat{\sigma}^2 \sum_{k=0}^{N} \tau_k^2 + \left(\frac{24}{5} + \frac{40L_{\nabla F}^2}{\beta^2} \right) \sum_{k=0}^{N} \tau_k \left(H_k(\tilde{y}^k) - H_k(y^k) \right). \end{aligned} \tag{43}$$

Then, setting τ_k, t_k to be values in (11) and noting that by Lemma 8, we have

$$H_k(\tilde{y}^k) - H_k(y^k) \leq \frac{2\beta D_{\mathcal{X}}^2(1+\delta)}{t_k + 2} \leq \frac{2\beta D_{\mathcal{X}}^2(1+\delta)}{\sqrt{k}}, \quad \forall k \geq 1.$$

Also, with the choice of $z^0 = 0$, we have $y^0 = \tilde{y}^0 = x^0$. Thus, we can conclude that

$$\sum_{k=0}^{N} \tau_k \left(H_k(\tilde{y}^k) - H_k(y^k) \right) \leq \frac{2\beta D_{\mathcal{X}}^2(1+\delta)}{\sqrt{N}} \sum_{k=1}^{N} \frac{1}{\sqrt{k}} \leq 4\beta D_{\mathcal{X}}^2(1+\delta).$$

which together with (43) immediately imply that $\forall N \geq 1$,

$$\begin{aligned} \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \mathbb{E} \left[\beta \left(\|d^k\|^2 + \sum_{j=1}^{T} \|f_j(u_{j+1}^k) - u_j^k\|^2 \right) + \frac{\beta}{20L_{\nabla F}^2} \|\nabla F(x^k) - z^k\|^2 \middle| \mathscr{F}_{k-1} \right] \\ \leq 2W_{\alpha,\gamma}(x^0, z^0, u^0) + \mathcal{B}(\beta, \sigma^2, L, D_{\mathcal{X}}, T, \delta). \end{aligned}$$

where

$$\mathcal{B}(\beta, \sigma^2, L, D_{\mathcal{X}}, T, \delta) = 4\hat{\sigma}^2 + 32\beta D_{\mathcal{X}}^2(1+\delta) \left(\frac{3}{5} + \frac{5L_{\nabla F}^2}{\beta^2} \right),$$

and $\hat{\sigma}^2$ is given in (42). As a result, we can obtain (13) and (14) by the definition of random integer R and

$$\begin{aligned} \|\mathcal{G}(x^k, \nabla F(x^k), \beta)\|^2 &\leq 2\beta^2 \|d^k\|^2 + 2\beta^2 \left\| \Pi_{\mathcal{X}} \left(x^k - \frac{1}{\beta} \nabla F(x^k) \right) - \Pi_{\mathcal{X}} \left(x^k - \frac{1}{\beta} z^k \right) \right\|^2 \\ &\leq 2\beta^2 \|d^k\|^2 + 2\|\nabla F(x^k) - z^k\|^2. \end{aligned}$$

□

D Proofs for Section 3.1

D.1 Proof of Theorem 3 for $T = 2$

To show the rate of convergence for Algorithm 3, we simplify the merit function in the analysis of the multi-level problems and leverage the following function:

$$W_{\alpha,\gamma}(x^k, z^k, u^k) = F(x^k) - F^\star - \eta(x^k, z^k) + \alpha \|\nabla F(x^k) - z^k\|^2 + \gamma \|f_2(x^k) - u_2^k\|^2, \tag{44}$$

where α, γ are positive constants, $\eta(\cdot, \cdot)$ is defined in (10). We now present the analogue of Lemma 11 for Algorithm 3. The proof follows similar steps as that proof of Lemma 11 with slight modifications, and hence we will skip some arguments already presented before.

Lemma 12. *Let $\{x^k, z^k, u_2^k\}_{k\geq 0}$ be the sequence generated by Algorithm 3 and the merit function $W_{\alpha,\gamma}(\cdot,\cdot,\cdot)$ be defined in* (44) *with*

$$\alpha = \frac{\rho}{L_{\nabla F}}, \qquad \gamma = 3\rho L_{\nabla f_1}, \qquad \rho > 0.$$

Under Assumptions 2 with $T = 2$, setting $\beta_k \equiv \beta \geq 6\rho L_{\nabla F} + (2\rho + \frac{2}{3\rho})L_{\nabla f_1} L_{f_2}^2$, we have $\forall N \geq 0$

$$\begin{aligned}
&\rho \sum_{k=0}^{N} \tau_k \left(L_{\nabla F} \|d^k\|^2 + L_{\nabla f_1} \|f_2(x^k) - u_2^k\|^2 + \frac{1}{L_{\nabla F}} \|\nabla F(x^k) - z^k\|^2 \right) \\
&\leq 2W_0 + 2\sum_{k=0}^{N} \mathbf{R}_k + \left(4 + \frac{2(8\rho + 1/\rho)L_{\nabla F} + 24\rho L_{\nabla f_1} L_{f_2}^2}{\beta} \right) \sum_{k=0}^{N} \tau_k \left(H_k(\tilde{y}^k) - H_k(y^k) \right)
\end{aligned}$$

where $d^k = y^k - x^k$, $H_k(\cdot), y^k$ are defined in (16)*, and*

$$\begin{aligned}
\mathbf{R}_k :=& \tau_k^2 \left[\frac{L_{\nabla F} + L_{\nabla \eta}}{2} D_{\mathcal{X}}^2 + \gamma \|\Delta_{G_2}^{k+1}\|^2 + \alpha \|\Delta^{k+1}\|^2 \right] + \frac{L_\eta}{2} \|z^{k+1} - z^k\|^2 \\
&+ \tau_k \langle d^k, \Delta^{k+1} \rangle + \gamma \dot{r}^{k+1} + \alpha \ddot{r}^{k+1}, \\
\Delta^{k+1} :=& \nabla f_2(x^k) \nabla f_1(u_2^k) - J_2^{k+1} J_1^{k+1}, \quad \Delta_{G_2}^{k+1} := f_2(x^k) - G_2^{k+1} \\
\dot{r}^{k+1} :=& 2\tau_k \langle \Delta_{G_2}^{k+1}, f_2(x^{k+1}) - f_2(x^k) + (1 - \tau_k)(f_2(x^k) - u_2^k) \rangle, \\
\ddot{r}^{k+1} :=& 2\tau_k \langle \Delta^{k+1}, (1 - \tau_k)[\nabla F(x^k) - z^k] + \tau_k [\nabla F(x^k) - \nabla f_2(x^k) \nabla f_1(u^k)] \\
&+ \nabla F(x^{k+1}) - \nabla F(x^k) \rangle.
\end{aligned} \tag{45}$$

Proof of Lemma 12. 1. By the Lipschitzness of ∇F (Lemma 6), we have

$$\begin{aligned}
F(x^{k+1}) - F(x^k) \leq& \tau_k \langle \nabla F(x^k), d^k \rangle + \tau_k \|\nabla F(x^k) - z^k\| \|\tilde{y}^k - y^k\| \\
&+ \tau_k \beta \|d^k\| \|\tilde{y}^k - y^k\| + \tau_k \langle z^k + \beta d^k, \tilde{y}^k - y^k \rangle + \frac{L_{\nabla F} \tau_k^2 \|\tilde{d}^k\|^2}{2}.
\end{aligned} \tag{46}$$

2. Also, by the Lipschitzness of $\nabla \eta$ (Lemma 7) and the optimality condition of in the definition of y^k, we have

$$\begin{aligned}
\eta(x^k, z^k) - \eta(x^{k+1}, z^{k+1}) \leq& -\beta \tau_k \|d^k\|^2 + \tau_k \langle z^k + \beta d^k, \tilde{y}^k - y^k \rangle \\
&- \tau_k \langle d^k, \nabla f_2(x^k) \nabla f_1(u_2^k) \rangle + \tau_k \langle d^k, \Delta^{k+1} \rangle + \frac{L_{\nabla \eta}}{2} \left[\tau_k^2 \|\tilde{d}^k\|^2 + \|z^{k+1} - z^k\|^2 \right].
\end{aligned} \tag{47}$$

3. In addition, by the Lipschitzness of f_2 and ∇f_1, we have

$$\begin{aligned}
\langle d^k, \nabla F(x^k) - \nabla f_2(x^k) \nabla f_1(u_2^k) \rangle &= \langle d^k, \nabla f_2(x^k)^\top \left[\nabla f_1(f_2(x^k)) - \nabla f_1(u_2^k) \right] \rangle \\
&\leq L_{\nabla f_1} L_{f_2} \|d^k\| \|f_2(x^k) - u_2^k\|.
\end{aligned} \tag{48}$$

4. Moreover, by the update rule, we have

$$\begin{aligned}
\|f_2(x^{k+1}) - u_2^{k+1}\|^2 &= \|f_2(x^{k+1}) - f_2(x^k) + (1 - \tau_k)[f_2(x^k) - u_2^k] + \tau_k \Delta_{G_2}^{k+1}\|^2 \\
&= \|(1 - \tau_k)[f_2(x^k) - u_2^k] + f_2(x^{k+1}) - f_2(x^k)\|^2 + \tau_k^2 \|\Delta_{G_2}^{k+1}\|^2 + \dot{r}^{k+1} \\
&\leq (1 - \tau_k) \|f_2(x^k) - u_2^k\|^2 + 2\tau_k L_{f_2}^2 (\|d^k\|^2 + \|\tilde{y}^k - y^k\|^2) + \tau_k^2 \|\Delta_{G_2}^{k+1}\|^2 + \dot{r}^{k+1}
\end{aligned} \tag{49}$$

where $\dot{r}^{k+1} := 2\tau_k \langle \Delta_{G_2}^{k+1}, f_2(x^{k+1}) - f_2(x^k) + (1 - \tau_k)(f_2(x^k) - u_2^k) \rangle$ and the last inequality follows Jensen's inequality for the convex function $\|\cdot\|^2$ as well as

$$\left\| \frac{1}{\tau_k} \left[f_2(x^{k+1}) - f_2(x^k) \right] \right\|^2 \leq L_{f_2}^2 \|\tilde{d}^k\|^2 \leq 2L_{f_2}^2 (\|d^k\|^2 + \|\tilde{y}^k - y^k\|^2).$$

5. Defining

$$e^k := \frac{1}{\tau_k} \left[\nabla F(x^{k+1}) - \nabla F(x^k) \right] + \nabla F(x^k) - \nabla f_2(x^k) \nabla f_1(u^k),$$

and by the update rule, we have

$$\begin{aligned}\|\nabla F(x^{k+1}) - z^{k+1}\|^2 &= \|(1-\tau_k)[\nabla F(x^k) - z^k] + \tau_k[e^k + \Delta^{k+1}]\|^2 \\ &= \|(1-\tau_k)[\nabla F(x^k) - z^k] + \tau_k e^k\|^2 + \tau_k^2\|\Delta^{k+1}\|^2 + \ddot{r}^{k+1} \\ &\leq (1-\tau_k)\|\nabla F(x^k) - z^k\|^2 + \tau_k\|e^k\|^2 + \tau_k^2\|\Delta^{k+1}\|^2 + \ddot{r}^{k+1}\end{aligned} \tag{50}$$

where $\ddot{r}^{k+1} := 2\tau_k\langle\Delta^k, (1-\tau_k)[\nabla F(x^k) - z^k] + \tau_k e^k\rangle$. We can further upper bound the term $\|e^k\|^2$ by

$$\begin{aligned}\|e^k\|^2 &\leq 2L_{\nabla F}^2\|\tilde{d}^k\|^2 + 2L_{\nabla f_1}^2 L_{f_2}^2\|f_2(x^k) - u^k\|^2 \\ &\leq 4L_{\nabla F}^2(\|d^k\|^2 + \|\tilde{y}^k - y^k\|^2) + 2L_{\nabla f_1}^2 L_{f_2}^2\|f_2(x^k) - u^k\|^2\end{aligned} \tag{51}$$

6. By combing (46), (47), (48), (49), (50), (51), rearranging the terms, and noting that $\langle z^k + \beta d^k, \tilde{y}^k - y^k\rangle = H_k(\tilde{y}^k) - H_k(y^k) - (\beta/2)\|\tilde{y}^k - y^k\|^2$ and $\|\tilde{d}^k\| \leq D_{\mathcal{X}}$, we obtain

$$W_{k+1} - W_k \leq \tau_k \mathbf{A}_k + \mathbf{R}_k \tag{52}$$

where $\mathbf{R}_k$ is defined in (45) and

$$\begin{aligned}\mathbf{A}_k := &\left(-\beta + 4\alpha L_{\nabla F}^2 + 2\gamma L_{f_2}^2\right)\|d^k\|^2 + \left(-\gamma + 2\alpha L_{\nabla f_1}^2 L_{f_2}^2\right)\|f_2(x^k) - u_2^k\|^2 \\ &+ L_{\nabla f_1} L_{f_2}\|d^k\|\|f_2(x^k) - u_2^k\| - \alpha\|\nabla F(x^k) - z^k\|^2 \\ &+ \left(\beta\|d^k\| + \|\nabla F(x^k) - z^k\|\right)\|\tilde{y}^k - y^k\| \\ &+ \left(4\alpha L_{\nabla F}^2 + 2\gamma L_{f_2}^2 - \beta\right)\|\tilde{y}^k - y^k\|^2 + 2\left(H_k(\tilde{y}^k) - H_k(y^k)\right).\end{aligned}$$

We then provide a simplified upper bound for $\mathbf{A}_k$. By the Young's inequality, we have

$$\begin{aligned}\beta\|d^k\|\|\tilde{y}^k - y^k\| &\leq \frac{\beta}{4}\|d^k\|^2 + \beta\|\tilde{y}^k - y^k\|^2, \\ \|\nabla F(x^k) - z^k\|\|\tilde{y}^k - y^k\| &\leq \frac{\alpha}{2}\|\nabla F(x^k) - z^k\|^2 + \frac{1}{2\alpha}\|\tilde{y}^k - y^k\|^2.\end{aligned}$$

In addition, we reparametrize $\alpha = \frac{\rho}{L_{\nabla F}}$. Noting that by Lemma 6 with $T = 2$

$$\frac{L_{\nabla f_1}^2 L_{f_2}^2}{L_{\nabla F}} = \frac{L_{\nabla f_1}^2 L_{f_2}^2}{L_{\nabla f_1} L_{f_2}^2 + L_{f_1} L_{\nabla f_2}} \leq L_{\nabla f_1},$$

we therefore have

$$\begin{aligned}\mathbf{A}_k \leq &\left(-\frac{3\beta}{4} + 4\rho L_{\nabla F} + 2\gamma L_{f_2}^2\right)\|d^k\|^2 + \left(-\gamma + 2\rho L_{\nabla f_1}\right)\|f_2(x^k) - u_2^k\|^2 \\ &+ L_{\nabla f_1} L_{f_2}\|d^k\|\|f_2(x^k) - u_2^k\| - \frac{\rho}{2L_{\nabla F}}\|\nabla F(x^k) - z^k\|^2 \\ &+ \left(4\rho L_{\nabla F} + 2\gamma L_{f_2}^2 + \frac{L_{\nabla F}}{2\rho}\right)\|\tilde{y}^k - y^k\|^2 + 2\left(H_k(\tilde{y}^k) - H_k(y^k)\right)\end{aligned}$$

Then, setting $\gamma = 3\rho L_{\nabla f_1}$ and $\beta \geq 6\rho L_{\nabla F} + (2\rho + \frac{2}{3\rho})L_{\nabla f_1} L_{f_2}^2$, we can obtain

$$\begin{aligned}&\left(-\frac{3\beta}{4} + 4\rho L_{\nabla F} + 2\gamma L_{f_2}^2\right)\|d^k\|^2 + \left(-\gamma + 2\rho L_{\nabla f_1}\right)\|f_2(x^k) - u_2^k\|^2 \\ &\quad + L_{\nabla f_1} L_{f_2}\|d^k\|\|f_2(x^k) - u_2^k\| \leq -\frac{\rho L_{\nabla F}}{2}\|d^k\|^2 - \frac{\rho L_{\nabla f_1}}{2}\|f_2(x^k) - u_2^k\|^2\end{aligned}$$

Also, we have $(\beta/2)\|\tilde{y}^k - y^k\|^2 \leq H_k(\tilde{y}^k) - H_k(y^k)$. Therefore, we can further bound $\mathbf{A}_k$ by

$$\begin{aligned}\mathbf{A}_k \leq &-\frac{\rho L_{\nabla F}}{2}\|d^k\|^2 - \frac{\rho L_{\nabla f_1}}{2}\|g(x^k) - u^k\|^2 - \frac{\rho}{2L_{\nabla F}}\|\nabla F(x^k) - z^k\|^2 \\ &+ \left(2 + \frac{(8\rho + 1/\rho)L_{\nabla F} + 12\rho L_{\nabla f_1} L_{f_2}^2}{\beta}\right)\left(H_k(\tilde{y}^k) - H_k(y^k)\right).\end{aligned} \tag{53}$$

Telescoping (52) together with (53), we get

$$\rho\sum_{k=0}^{N}\tau_k\bigg(L_{\nabla F}\|d^k\|^2+L_{\nabla f_1}\|f_2(x^k)-u_2^k\|^2+\frac{1}{L_{\nabla F}}\|\nabla F(x^k)-z^k\|^2\bigg)$$
$$\leq 2W_0+2\sum_{k=0}^{N}\mathbf{R}_k+\bigg(4+\frac{2(8\rho+1/\rho)L_{\nabla F}+24\rho L_{\nabla f_1}L_{f_2}^2}{\beta}\bigg)\sum_{k=0}^{N}\tau_k\left(H_k(\tilde{y}^k)-H_k(y^k)\right)$$

□

Proof of Theorem 3, part (a). The proof follows the same arguments in the proof of Theorem 2. Note that by Lemma 12 and given values of α,γ in (12), we obtain

$$\rho\sum_{k=1}^{N}\tau_k\left[L_{\nabla F}\|d^k\|^2+L_{\nabla f_1}\|f_2(x^k)-u_2^k\|^2+\frac{1}{L_{\nabla F}}\|\nabla F(x^k)-z^k\|^2\right]\leq 2W_{\alpha,\gamma}(x^0,z^0,u^0)$$
$$+2\sum_{k=0}^{N}\mathbf{R}_k+\bigg(4+\frac{2(8\rho+1/\rho)L_{\nabla F}+24\rho L_{\nabla f_1}L_{f_2}^2}{\beta}\bigg)\sum_{k=0}^{N}\tau_k\left(H_k(\tilde{y}^k)-H_k(y^k)\right).$$

Noting that

$$\mathbb{E}[\mathbf{R}_k|\mathscr{F}_k]=\tau_k^2\left[\frac{L_{\nabla F}+L_{\nabla\eta}}{2}D_{\mathcal{X}}^2+\gamma\sigma_{G_2}^2+(\alpha+2L_\eta)\hat{\sigma}_{J_1}^2\hat{\sigma}_{J_2}^2\right]:=\tau_k^2\hat{\sigma}^2,$$

and taking expectation of both sides, we can complete the proof with the same arguments in the proof of Theorem 2. The constants $\mathcal{C}_1$ and $\mathcal{C}_2$ turn out to be

$$\begin{aligned}\mathcal{C}_1=4\left(\frac{\beta^2}{\rho L_{\nabla F}}+\frac{L_{\nabla F}}{\rho}\right)\bigg\{&W_{\alpha,\gamma}(x^0,z^0,u^0)+\hat{\sigma}^2\\&+4D_{\mathcal{X}}^2(1+\delta)\left[2\beta+(8\rho+\frac{1}{\rho})L_{\nabla F}+12\rho L_{\nabla f_1}L_{f_2}^2)\right]\bigg\},\\ \mathcal{C}_2=\frac{2}{\rho L_{\nabla f_1}}\bigg\{&W_{\alpha,\gamma}(x^0,z^0,u^0)+\hat{\sigma}^2\\&+4D_{\mathcal{X}}^2(1+\delta)\left[2\beta+(8\rho+\frac{1}{\rho})L_{\nabla F}+12\rho L_{\nabla f_1}L_{f_2}^2)\right]\bigg\}.\end{aligned}\tag{54}$$

□

D.2 Proof of Theorem 3 for $T=1$

To show the rate of convergence for Algorithm 4, we leverage the following merit function:

$$W_\alpha(x^k,z^k,u^k)=F(x^k)-F^\star-\eta(x^k,z^k)+\alpha\|\nabla F(x^k)-z^k\|^2,\tag{55}$$

where $\alpha>0$, $\eta(\cdot,\cdot)$ is defined in (10).

Lemma 13. *Let $\{x^k,z^k\}_{k\geq 0}$ be the sequence generated by Algorithm 4 with $\beta_k\equiv\beta>0$ and the merit function $W_\alpha(\cdot,\cdot)$ be defined in* (55) *with $\alpha=\frac{\beta}{4L_{\nabla F}^2}$. Under Assumptions 2 with $T=1$, we have $\forall N\geq 0$*

$$\beta\sum_{k=0}^{N}\tau_k\bigg(\|d^k\|^2+\frac{1}{2L_{\nabla F}^2}\|\nabla F(x^k)-z^k\|^2\bigg)$$
$$\leq 4W_\alpha(x^0,u^0)+4\sum_{k=0}^{N}\mathbf{R}_k+\bigg(12+\frac{16L_{\nabla F}^2}{\beta^2}\bigg)\sum_{k=0}^{N}\tau_k\left(H_k(\tilde{y}^k)-H_k(y^k)\right)$$

where $d^k:=y^k-x^k$, $H_k(\cdot),y^k$ are defined in (16)*, $\Delta^{k+1}:=\nabla F(x^k)-J_1^{k+1}$, and*

$$\begin{aligned}\mathbf{R}_k:=&\tau_k^2\left[\frac{L_{\nabla F}+L_{\nabla\eta}}{2}D_{\mathcal{X}}^2+\alpha\|\Delta^{k+1}\|^2\right]+\frac{L_\eta}{2}\|z^{k+1}-z^k\|^2\\&+\tau_k\langle d^k,\Delta^{k+1}\rangle+\alpha r^{k+1},\\ r^{k+1}:=&2\tau_k\langle\Delta^{k+1},(1-\tau_k)[\nabla F(x^k)-z^k]+\nabla F(x^{k+1})-\nabla F(x^k)\rangle.\end{aligned}\tag{56}$$

Proof. The proof is a essentially a simplified version of the proof of Lemma 12. Hence, we skip some arguments already presented earlier.

1. By the Lipschitzness of ∇F, we have

$$\begin{aligned} F(x^{k+1}) - F(x^k) &\leq \tau_k \langle \nabla F(x^k), d^k \rangle + \tau_k \|\nabla F(x^k) - z^k\| \|\tilde{y}^k - y^k\| \\ &\quad + \tau_k \beta \|d^k\| \|\tilde{y}^k - y^k\| + \tau_k \langle z^k + \beta d^k, \tilde{y}^k - y^k \rangle + \frac{L_{\nabla F} \tau_k^2 \|\tilde{d}^k\|^2}{2}. \end{aligned} \tag{57}$$

2. Also, by the lipschitzness of $\nabla \eta$ (Lemma 7) and the optimality condition of in the definition of y^k, we have

$$\begin{aligned} \eta(x^k, z^k) - \eta(x^{k+1}, z^{k+1}) &\leq -\beta \tau_k \|d^k\|^2 + \tau_k \big\langle z^k + \beta d^k, \tilde{y}^k - y^k \big\rangle \\ &\quad - \tau_k \big\langle d^k, \nabla F(x^k) \big\rangle + \tau_k \big\langle d^k, \Delta^{k+1} \big\rangle + \frac{L_{\nabla \eta}}{2} \left[\tau_k^2 \|\tilde{d}^k\|^2 + \|z^{k+1} - z^k\|^2 \right]. \end{aligned} \tag{58}$$

3. By the update rule, we have

$$\begin{aligned} &\|\nabla F(x^{k+1}) - z^{k+1}\|^2 = \|(1-\tau_k)[\nabla F(x^k) - z^k] + \nabla F(x^{k+1}) - \nabla F(x^k) + \tau_k \Delta^{k+1}\|^2 \\ &= \|(1-\tau_k)[\nabla F(x^k) - z^k] + \nabla F(x^{k+1}) - \nabla F(x^k)\|^2 + \tau_k^2 \|\Delta^{k+1}\|^2 + r^{k+1} \\ &\leq (1-\tau_k)\|\nabla F(x^k) - z^k\|^2 + \frac{1}{\tau_k}\|\nabla F(x^{k+1}) - \nabla F(x^k)\|^2 + \tau_k^2 \|\Delta^{k+1}\|^2 + r^{k+1} \\ &\leq (1-\tau_k)\|\nabla F(x^k) - z^k\|^2 + \tau_k L_{\nabla F}^2 \|\tilde{d}^k\|^2 + \tau_k^2 \|\Delta^{k+1}\|^2 + r^{k+1} \\ &\leq (1-\tau_k)\|\nabla F(x^k) - z^k\|^2 + 2\tau_k L_{\nabla F}^2 (\|d^k\|^2 + \|\tilde{y}^k - y^k\|^2) + \tau_k^2 \|\Delta^{k+1}\|^2 + r^{k+1} \end{aligned} \tag{59}$$

where $r^{k+1} := 2\tau_k \langle \Delta^k, (1-\tau_k)[\nabla F(x^k) - z^k] + \nabla F(x^{k+1}) - \nabla F(x^k) \rangle$.

4. By combing (57), (58) (59), rearranging the terms, and noting that $\langle z^k + \beta d^k, \tilde{y}^k - y^k \rangle = H_k(\tilde{y}^k) - H_k(y^k) - (\beta/2)\|\tilde{y}^k - y^k\|^2$ and $\|\tilde{d}^k\| \leq D_{\mathcal{X}}$, we obtain

$$W_{k+1} - W_k \leq \tau_k \mathbf{A}_k + \mathbf{R}_k \tag{60}$$

where $\mathbf{R}_k$ is defined in (56) and

$$\begin{aligned} \mathbf{A}_k := &\left(-\beta + 2\alpha L_{\nabla F}^2\right) \|d^k\|^2 - \alpha \|\nabla F(x^k) - z^k\|^2 + \left(\beta \|d^k\| + \|\nabla F(x^k) - z^k\|\right) \|\tilde{y}^k - y^k\| \\ &+ \left(2\alpha L_{\nabla F}^2 - \beta\right) \|\tilde{y}^k - y^k\|^2 + 2\left(H_k(\tilde{y}^k) - H_k(y^k)\right). \end{aligned}$$

We then provide a simplified upper bound for $\mathbf{A}_k$. By the Young's inequality, we have

$$\begin{aligned} \beta \|d^k\| \|\tilde{y}^k - y^k\| &\leq \frac{\beta}{4}\|d^k\|^2 + \beta \|\tilde{y}^k - y^k\|^2, \\ \|\nabla F(x^k) - z^k\| \|\tilde{y}^k - y^k\| &\leq \frac{\alpha}{2}\|\nabla F(x^k) - z^k\|^2 + \frac{1}{2\alpha}\|\tilde{y}^k - y^k\|^2. \end{aligned}$$

In addition, setting $\alpha = \frac{\beta}{4L_{\nabla F}^2}$ and noting $(\beta/2)\|\tilde{y}^k - y^k\|^2 \leq H_k(\tilde{y}^k) - H_k(y^k)$, we have

$$\mathbf{A}_k \leq -\frac{\beta}{4}\|d^k\|^2 - \frac{\beta}{8L_{\nabla F}^2}\|\nabla F(x^k) - z^k\|^2 + \left(3 + \frac{4L_{\nabla F}^2}{\beta^2}\right)\left(H_k(\tilde{y}^k) - H_k(y^k)\right) \tag{61}$$

Telescoping (60) together with (61), we get

$$\begin{aligned} &\beta \sum_{k=0}^{N} \tau_k \left(\|d^k\|^2 + \frac{1}{2L_{\nabla F}^2}\|\nabla F(x^k) - z^k\|^2 \right) \\ &\leq 4W_\alpha(x^0, u^0) + 4\sum_{k=0}^{N} \mathbf{R}_k + \left(12 + \frac{16L_{\nabla F}^2}{\beta^2}\right) \sum_{k=0}^{N} \tau_k \left(H_k(\tilde{y}^k) - H_k(y^k)\right) \end{aligned}$$

□

Proof of Theorem 3, part (b). Given Lemma 13, the proof follows the same arguments as in the proof of Theorem 2. The constant $\mathcal{C}_3$ turns out to be

$$\mathcal{C}_3 = 8\left(\beta + \frac{2L_{\nabla F}^2}{\beta}\right)\left\{W_\alpha(x^0,u^0) + D_{\mathcal{X}}^2\left[(1+\delta)\left(12\beta + \frac{16L_{\nabla F}^2}{\beta}\right) + \frac{L_{\nabla F}+L_{\nabla \eta}}{2}\right]\right. \\ \left. +\alpha\sigma_{J_1}^2 + 2L_\eta\hat{\sigma}_{J_1}^2\right\}. \quad (62)$$

□

E High-Probability Convergence for $T = 1$

E.1 Preliminaries

We provide a short review of sub-gaussian and sub-exponential random variables for completeness.

Definition 14. (Sub-gaussian and Sub-exponential)

(a) A random variable X is K-sub-gaussian if there exists $K > 0$ such that $\mathbb{E}[\exp(X^2/K^2)] \le 2$. The sub-gaussian norm of X, denoted $\|X\|_{\psi_2}$, is defined to be the smallest K. That is to say,

$$\|X\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}[\exp(X^2/t^2)] \le 2\right\}.$$

(b) A random variable X is K-sub-exponential if there exists $K > 0$ such that $\mathbb{E}[\exp(|X|/K)] \le 2$. The sub-exponential norm of X, denoted $\|X\|_{\psi_1}$, is defined to be the smallest K. That is to say,

$$\|X\|_{\psi_1} = \inf\left\{t > 0 : \mathbb{E}[\exp(|X|/t)] \le 2\right\}.$$

The above characterization is based on the so-called orlicz norm of a random variable. There are equivalent definitions of sub-gaussian and sub-exponential random variables. We refer readers to Proposition 2.5.2 and Proposition 2.7.1 in [45]. In particular, we will also use another definition of sub-gaussian random variables based on the moment generating function given below.

Lemma 15. (Sub-gaussian M.G.F. [45]) *If a random variable X is K-sub-gaussian with $\mathbb{E}[X] = 0$, then $\mathbb{E}[\exp(\lambda X)] \le \exp(c\lambda^2K^2)\ \forall\lambda \in \mathbb{R}$, where c is an absolute constant.*

In the high probability results we show for the special case with $T = 1$, we handle the tail probability for two terms involving the mean-zero noise with sub-gaussian norm, $\|\Delta^{k+1}\|^2$ and $\langle\Delta^{k+1}, \Lambda^k\rangle$, where (Δ^k) and (Λ^k) are adapted to $(\mathscr{F}_k)$. Our proof leverages the following two lemmas to control the probability of these two terms being too large.

Lemma 16. (Sub-exponential is sub-gaussian squared [45]) *A random variable X is sub-gaussian if and only if X^2 is sub-exponential. Moreover, $\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$.*

Lemma 17. (Generalized Freedman-type Inequality [21]) *Let $(\Omega, \mathscr{F}, (\mathscr{F}_i), P)$ be a filtered probability space, (X_i) and (K_i) be adapted to $(\mathscr{F}_i)$, and $n \in \mathbb{N}$. Suppose for all $i \in [n]$, $K_{i-1} \ge 0$, $\mathbb{E}[X_i|\mathscr{F}_{i-1}] = 0$, and $\mathbb{E}\left[\exp(\lambda X_i)|\mathscr{F}_{i-1}\right] \le \exp(\lambda^2K_i^2)$. Then for any $t, b \ge 0, a > 0$,*

$$\Pr\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^k X_i \ge t \text{ and } 2\sum_{i=1}^k K_{i-1}^2 \le a\sum_{i=1}^k X_i + b\right\}\right) \le \exp\left(-\frac{t}{4a+8b/t}\right). \quad (63)$$

E.2 Proof of Theorem 5

We start with presenting the lemma below which leverages inequalities in Appendix E to show a high-probability upper bound for terms involving in the previous analysis.

Lemma 18. *Under the conditions of Lemma 13 and Assumption 4, for any $\delta_1, \delta_2, \delta_3, a > 0$, we have*

(a) with probability at least $1 - \delta_1$, $\sum_{k=0}^N \tau_k^2\|\Delta^{k+1}\|^2 \le K^2\log(2/\delta_1)\sum_{k=0}^N \tau_k^2$;

(b) with probability at least $1-\delta_2$*,*

$$\sum_{k=0}^{N}\tau_k^2\sum_{i=0}^{k-1}\alpha_{i,k}\|\Delta^{i+1}\|^2 \le K^2\log(2/\delta_2)\sum_{k=0}^{N}\tau_k^2,$$

where $\alpha_{i,k}>0$ *and* $\sum_{i=0}^{k-1}\alpha_{i,k}=1$*;*

(c) with probability at least $1-\delta_3$*,*

$$\sum_{k=0}^{N}\langle\Delta^{k+1},2\alpha\tau_k(1-\tau_k)[\nabla F(x^k)-z^k]+2\alpha\tau_k(\nabla F(x^{k+1})-\nabla F(x^k))+\tau_k d^k\rangle$$
$$\le 4a\log(1/\delta_3)+\frac{\beta^2K^2}{aL_{\nabla F}^4}\sum_{k=0}^{N}\tau_k^2(1-\tau_k)\left\|\nabla F(x^k)-z^k\right\|^2+\frac{K^2}{a}\sum_{k=0}^{N}\tau_k^2(4+\frac{\beta^2\tau_k}{L_{\nabla F}^2})\left\|d^k\right\|^2.$$

Proof of Lemma 18. We first show (a). Using the law of total expectation, we have $\mathbb{E}\left[\exp\left(\frac{\|\tau_k\Delta^{k+1}\|^2}{\tau_k^2K^2}\right)\right]\le 2$, which implies that $\|\tau_k\Delta^{k+1}\|^2$ is $\tau_k^2K^2$-sub-exponential. Thus, we have with probability at least $1-\delta_1$,

$$\sum_{k=0}^{N}\tau_k^2\|\Delta^{k+1}\|^2\le K^2\log(2/\delta_1)\sum_{k=0}^{N}\tau_k^2. \tag{64}$$

We then show (b). Let $Z_k=\tau_k^2\left\{\sum_{i=0}^{k-1}\alpha_{i,k}\left\|\Delta^{i+1}\right\|^2\right\}\forall k\ge 0$. Note that for all $k\ge 0$, $\|\Delta^{k+1}\|^2$ is K^2-sub-exponential, which further implies that the sub-exponential norm of Z_k $(k>0)$ satisfies $\|Z_k\|_{\psi_1}\le\tau_k^2K^2$. Therefore, we have for any $\delta_2>0$, with probability at least $1-\delta_2$,

$$\sum_{k=0}^{N}Z_k\le K^2\log(2/\delta_2)\sum_{k=0}^{N}\tau_k^2. \tag{65}$$

To prove (c), we apply Lemma 15 and Lemma 17 with

$$X_i=\left\langle\Delta^{k+1},2\alpha\tau_k\left\{(1-\tau_k)[\nabla F(x^k)-z^k]+\nabla F(x^{k+1})-\nabla F(x^k)\right\}+\tau_k d^k\right\rangle,$$
$$K_i=\sqrt{c}K\left\|2\alpha\tau_k\left\{(1-\tau_k)[\nabla F(x^k)-z^k]+\nabla F(x^{k+1})-\nabla F(x^k)\right\}+\tau_k d^k\right\|,$$
$$b=0,t=4a\log(1/\delta_3).$$

Noting that $\alpha=\frac{\beta}{4L_{\nabla F}^2}$, we obtain that for all $a>0$ with probability at least $1-\delta_3$, $\sum_{i=0}^{N}X_i\le 4a\log(1/\delta_3)$ and

$$\sum_{i=0}^{N}X_i\le\frac{2cK^2}{a}\sum_{k=0}^{N}\left\|2\alpha\tau_k\left\{(1-\tau_k)[\nabla F(x^k)-z^k]+\nabla F(x^{k+1})-\nabla F(x^k)\right\}+\tau_k d^k\right\|^2$$
$$\le\frac{4cK^2}{a}\sum_{k=0}^{N}\tau_k^2\left\{4\alpha^2\left\|(1-\tau_k)[\nabla F(x^k)-z^k]+\nabla F(x^{k+1})-\nabla F(x^k)\right\|^2+\left\|d^k\right\|^2\right\}$$
$$\le\frac{4cK^2}{a}\sum_{k=0}^{N}\tau_k^2\left\{4\alpha^2(1-\tau_k)\left\|\nabla F(x^k)-z^k\right\|^2+(1+4\alpha^2L_{\nabla F}^2\tau_k)\left\|d^k\right\|^2\right\}$$
$$=\frac{c\beta^2K^2}{aL_{\nabla F}^4}\sum_{k=0}^{N}\tau_k^2(1-\tau_k)\left\|\nabla F(x^k)-z^k\right\|^2+\frac{cK^2}{a}\sum_{k=0}^{N}\tau_k^2(4+\frac{\beta^2\tau_k}{L_{\nabla F}^2})\left\|d^k\right\|^2,$$

where the third inequality comes from the convexity of $\|\cdot\|^2$ and the Lipschitzness of ∇F. □

Provided with the above lemma and Lemma 13, we now present the complete proof of Theorem 5.

Proof of Theorem 5. Given the update rule of $\{z^k\}$ and the fact that $\tau_0 = 1$, we can obtain

$$z^k = \sum_{i=0}^{k-1} \alpha_{i,k} J_1^{i+1}, \quad \alpha_{i,k} = \frac{\tau_i}{\Gamma_{i+1}} \Gamma_k \;\; 1 \le i \le k, \quad \sum_{i=0}^{k-1} \alpha_{i,k} = 1 \;\; k \ge 1$$

where $\Gamma_k = \Gamma_1 \prod_{i=1}^{k-1}(1-\tau_i)$ and $\Gamma_1 = 1$. Thus,

$$\begin{aligned}
\left\|z^{k+1} - z^k\right\|^2 = \tau_k^2 \left\|J_1^{k+1} - z^k\right\|^2 &\le 2\tau_k^2 \left\{ \left\|J_1^{k+1}\right\|^2 + \left\|\sum_{i=0}^{k-1} \alpha_{i,k} J_1^{i+1}\right\|^2 \right\} \\
&\le 2\tau_k^2 \left\{ \left\|J_1^{k+1}\right\|^2 + \sum_{i=0}^{k-1} \alpha_{i,k} \left\|J_1^{i+1}\right\|^2 \right\} \\
&\le 4\tau_k^2 \left\{ \left\|\Delta^{k+1}\right\|^2 + \left\|\nabla F(x^k)\right\|^2 + \sum_{i=0}^{k-1} \alpha_{i,k} \left[\left\|\Delta^{i+1}\right\|^2 + \|\nabla F(x^i)\|^2 \right] \right\} \\
&\le 4\tau_k^2 \left\{ \left\|\Delta^{k+1}\right\|^2 + \sum_{i=0}^{k-1} \alpha_{i,k} \left\|\Delta^{i+1}\right\|^2 + 2L_F^2 \right\}
\end{aligned}$$

where the second inequality comes from the convexity of $\|\cdot\|^2$. Therefore, we have

$$\sum_{k=0}^{N} \|z^{k+1} - z^k\|^2 \le 4\sum_{k=0}^{N} \tau_k^2 \|\Delta^{k+1}\|^2 + 4\sum_{k=0}^{N} \tau_k^2 \sum_{i=0}^{k-1} \alpha_{i,k} \|\Delta^{i+1}\|^2 + 8L_F^2 \sum_{k=0}^{N} \tau_k^2$$

Applying Lemma 18 with $\delta_1 = \delta_2 = \delta_3 = \delta/3$ and $a = \frac{16c\beta K^2}{L_{\nabla F}^2}$ together with Lemma 13, we have with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{k=0}^{N} \mathbf{R}_k =& \sum_{k=0}^{N} \langle \Delta^{k+1}, 2\alpha\tau_k(1-\tau_k)[\nabla F(x^k) - z^k] + 2\alpha\tau_k(\nabla F(x^{k+1}) - \nabla F(x^k)) + \tau_k d^k \rangle \\
&+ (\alpha + 2L_\eta) \sum_{k=0}^{N} \tau_k^2 \|\Delta^{k+1}\|^2 + 2L_\eta \sum_{k=0}^{N} \tau_k^2 \sum_{i=0}^{k-1} \alpha_{i,k} \|\Delta^{i+1}\|^2 \\
&+ \left[\frac{L_{\nabla F} + L_{\nabla \eta}}{2} D_{\mathcal{X}}^2 + 4L_\eta L_F^2 \right] \sum_{k=0}^{N} \tau_k^2 \\
\le& \frac{64\beta K^2}{L_{\nabla F}^2} \log(3/\delta) + \frac{\beta}{16L_{\nabla F}^2} \sum_{k=0}^{N} \tau_k^2 (1-\tau_k) \left\|\nabla F(x^k) - z^k\right\|^2 + (\frac{L_{\nabla F}^2}{4\beta} + \frac{\beta}{16}) \sum_{k=0}^{N} \tau_k^2 \left\|d^k\right\|^2 \\
&+ \left[(\frac{\beta}{4L_{\nabla F}^2} + 4L_\eta) K^2 \log(6/\delta) + \frac{L_{\nabla F} + L_{\nabla \eta}}{2} D_{\mathcal{X}}^2 + 4L_\eta L_F^2 \right] \sum_{k=0}^{N} \tau_k^2
\end{aligned}$$

Thus, noting that $\|d^k\|^2 \le D_{\mathcal{X}}^2 \; \forall k \ge 0$, we have with probability at least $1-\delta$,

$$\begin{aligned}
&\beta \sum_{k=0}^{N} \tau_k \left(\|d^k\|^2 + \frac{1}{4L_{\nabla F}^2} \|\nabla F(x^k) - z^k\|^2 \right) \\
&\le 4W_\alpha(x^0, u^0) + \left(12 + \frac{16L_{\nabla F}^2}{\beta^2} \right) \sum_{k=0}^{N} \tau_k \left(H_k(\tilde{y}^k) - H_k(y^k) \right) + \frac{256\beta K^2}{L_{\nabla F}^2} \log(3/\delta) \\
&+ \left[\left(\frac{\beta}{L_{\nabla F}^2} + 16L_\eta \right) K^2 \log(6/\delta) + \left(\frac{L_{\nabla F}^2}{\beta} + \frac{\beta}{4} + 2L_{\nabla F} + 2L_{\nabla \eta} \right) D_{\mathcal{X}}^2 + 16L_\eta L_F^2 \right] \sum_{k=0}^{N} \tau_k^2
\end{aligned}$$

Following the same arguments as in the proof of Theorem 2, we have with probability at least $1-\delta$,

$$\min_{k=1,\ldots,N} V(x^k, z^k) \le \mathcal{O}\left(\frac{K^2 \log(1/\delta)}{\sqrt{N}} \right)$$

□

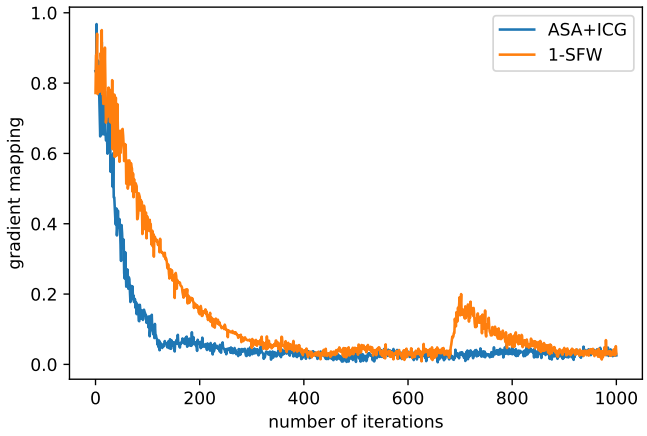

Figure 1: `ASA+ICG` vs. `1-SFW`

F Numerical Experiments for $T = 1$

To demonstrate the effectiveness and efficiency of proposed algorithms compared to 1-SFW [54] for $T = 1$, we consider the following matrix-valued single-index model [51] with low-rank constraints:

$$y = |\langle A, B^\star \rangle_F|^2 + \epsilon, \quad \text{rank}(B^\star) \leq s,$$

where $A, B \in \mathbb{R}^{m \times n}$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product, and s is some positive integer strictly less than m and n. To recover a low-rank matrix B, one can optimize the mean squared loss with nuclear norm constraint, in which the Frank-Wolfe update is much cheaper than the projection operator especially with large-scale matrices [26]. Formally, our problem can be written as

$$\min \; F(B) = \mathbb{E}_{A,\epsilon} \left[(y - |\langle A, B \rangle_F|^2)^2 \right] \quad \text{s.t. } \|B\|_\star \leq s.$$

We evaluate the performance of `ASA+ICG` (Algorithm 4) and `1-SFW` on a toy example where $B^\star = vv^\top / \|vv^\top\|_\star$ is a 4 by 4 rank-1 matrix. The matrix A is generated as $A = I + E$ where $E_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.3)$. The stepsize parameter $\beta = 1$ for `ASA+ICG`, and all the parameters in `1-SFW` is set according to Theorem 2 in [54]. As the exact gradient of F is unavailable, we estimate the gradient mapping by using averaged stochastic gradients. In Figure 1, we plot the value of gradient mapping versus the number of iterations, which demonstrates the superior of our proposed method for $T = 1$ in the one-sample setting.