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

Optimal Algorithms for Convex Nested Stochastic Composite Optimization

Zhe Zhang · Guanghui Lan

the date of receipt and acceptance should be inserted later

Abstract Recently, convex nested stochastic composite optimization (NSCO) has received considerable attention for its application in reinforcement learning and risk-averse optimization. However, In the current literature, there exists a significant gap in the iteration complexities between these NSCO problems and other simpler stochastic composite optimization problems (e.g., sum of smooth and nonsmooth functions) without the nested structure. In this paper, we close the gap by reformulating a class of convex NSCO problems as "min max ... max" saddle point problems under mild assumptions and proposing two primal-dual type algorithms with the optimal $\mathcal{O}\{1/\epsilon^2\}$ (resp., $\mathcal{O}\{1/\epsilon\}$) complexity for solving nested (resp., strongly) convex problems. More specifically, for the often-considered two-layer smooth-nonsmooth problem, we introduce a simple vanilla stochastic sequential dual (SSD) algorithm which can be implemented purely in the primal form. For the multi-layer problem, we propose a general stochastic sequential dual framework. The framework consists of modular dual updates for different types of functions (smooth, smoothable, and non-smooth, etc.), so that it can handle a more general composition of layer functions. Moreover, we present modular convergence proofs to show that the complexity of the general SSD is optimal with respect to nearly all the problem parameters.

1 Introduction

1.1 Motivation

During the past few years, composite optimization has attracted considerable interest due to their importance in applications, e.g., compressed sensing, image processing and machine learning. Many algorithmic studies have been focused on composite optimization of the form $\min_{x\in X} f(x) + g(x)$, where f is a smooth convex function and g is a nonsmooth function with certain special structures. Optimal first-order methods have been developed in [18, 21, 1, 10, 11] for solving these problems under different assumptions about g. In the stochastic setting, Lan [8, 9] present an accelerated stochastic approximation method that can achieve the optimal iteration/sampling complexity when one only has access to stochastic (sub)gradients of the objective function (see, e.g., [5, 4, 3] for extensions). The study of composite optimization has been later expanded to more complex nested composition problems. Specifically, Lewis and Wright [15] developed a globally convergent algorithm for solving $\min_{x\in X} f(g(x))$, where the outer layer function f can be non-smooth, non-convex and extended-valued. Lan [12] also studied the complexity of these problems when f is relatively simple. Wang et. al. [23] are the first to study nested stochastic composite optimization (NSCO) problems when f and g are given as expectation functions. NSCO found wide applications in reinforcement learning [23], meta-learning [2], and risk-averse optimization [19], and thus becomes a more and more important topic in stochastic optimization.

This research was partially supported by the ARO grant W911NF-18-1-0223 and the NSF grant 1953199.

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. (email: jimmy_zhang@gatech.edu).

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. (email: george.lan@isye.gatech.edu).

Address(es) of author(s) should be given

A key challenge in NSCO is the lack of unbiased gradient estimators for its composite objective function. This issue can be illustrated with a simple two-layer problem:

$$\min_{x \in X} \{f(x) := f_1(f_2(x))\}, \tag{1.1}$$

where f_1 and f_2 are accessible only through stochastic first-order oracles, denoted by $\mathcal{SO}_i$, which return unbiased estimators $(f_i(y_i, \xi_i), f_i'(y_i, \xi_i))$ for exact first-order information $(f_i(y_i), f_i'(y_i)))$ upon a query at y_i. Here ξ_i are the random variables associated with $\mathcal{SO}_i$, $i = 1$ or 2. By the chain rule, we know that the gradient of f is $f'(x) = f_1'(f_2(x))f_2'(x)$[①]. Because f_2 is stochastic, only stochastic estimators for the argument to $f_1'(\cdot)$ is available. As such, unless f_1 is affine, obtaining unbiased estimators for $f_1'(f_2(x))$ is impossible.

One possible way to address this issue is to approximate $f_2(x^t)$ with some u^t and use $f_1'(u^t, \xi_1^t)f_2'(x^t, \xi_2^t)$ as a proxy for $f'(x^t)$. Indeed, Wang et. al. [22] proposed to track $f_2(x^t)$ with a moving average approximation

$$u^t := (1 - \beta^{t-1})u^{t-1} + \beta^{t-1} f_2(x^t, \xi_2^t), \tag{1.2}$$

and developed a stochastic compositional gradient descent (SCGD) algorithm which iteratively moves along the quasi-gradient direction $f_1'(u^t, \xi_1^t)f_2'(x^t, \xi_2^t)$. By assuming f_1 to be smooth, they were able to establish the sampling complexities of SCGD for finding an $\epsilon-$optimal solution if f is convex, or finding an ϵ-stationary point if f_2 is smooth. However these complexities are worse than those for one-layer problem and simpler stochastic composite optimization without the nested structure (e.g., f_1 is affine). Therefore, an interesting research question is whether we can solve NSCO with the same complexity as if the nested structure does not exist.

On the one hand, the answer appears to be positive for some non-convex NSCO problems. More specifically, in one-layer problems, we know that it takes $\mathcal{O}\{1/\epsilon^2\}$ [6] queries to the stochastic oracle to find an ϵ-stationary solution $\bar{x}$, i.e., $\mathbb{E}[\|\nabla f(\bar{x})\|^2] \leq \epsilon$. Recently, by using the same moving average approximation (1.2) and a specially-designed potential function, Ghadimi et. al. [7] developed an $\mathcal{O}\{1/\epsilon^2\}$ algorithm for the two-layer problem and Ruszczynski [19] extended it further to solve the multi-layer problem with the same $\mathcal{O}\{1/\epsilon^2\}$ complexity. Moreover, under some stronger smoothness assumptions, a variance reduction algorithm proposed in [25] can improve the iteration complexity further to $\mathcal{O}\{1/\epsilon^{1.5}\}$.

On the other hand, the answer to this question is still unclear for convex NSCO problems. In order to find an ϵ-optimal solution ($\mathbb{E}[f(\bar{x}) - f_*] \leq \epsilon$) for one-layer problems, we know from [16] that the lower complexity bounds are $\mathcal{O}\{1/\epsilon^2\}$ if f is non-smooth or $\mathcal{O}\{1/\epsilon\}$ if f is also strongly convex. However, the results for NSCO in the literature fail to match them. To the best of our knowledge, only a few finite time convergence bounds have been developeed for non-smooth convex NSCO problems in [22,24,23]. These works use moving averages similar to (1.2) to track function values of the inner layers and apply multi-timescale schemes to ensure their faster convergence. However, as shown in Table 1 and 2, there exists signifiant gaps between these complexity results and the lower bounds mentioned above. For two-layer problems, if the innermost layer function is non-smooth, the complexity is $\mathcal{O}\{1/\epsilon^4\}$. Even with an additional smoothness assumption for f_2, the $\mathcal{O}\{1/\epsilon^{2.25}\}$ complexity for convex problems and the $\mathcal{O}\{1/\epsilon^{1.25}\}$ complexity for strongly convex problems still falls short of the one-layer result [9,5]. For multi-layer problems, the complexities are even worse as they exponentially depend on the number of layers k. Furthermore, these algorithms all require outer layer functions, i.e., f_1 for two-layer problems and $f_1, ..., f_{k-1}$ for k-layer problems, to be smooth. This assumption may preclude some important applications of convex NSCO, see the example of risk averse optimization in Section 6.

Table 1: Two-Layer Iteration Complexity

Problem	Type	In the Literature	SSD Framework
Convex	Nonsmooth$^{(1)}$	$\mathcal{O}\{1/\epsilon^4\}$ [22]	$O\{1/\epsilon^2\}$[*]
	Smooth$^{(2)}$	$\mathcal{O}\{1/\epsilon^{2.25}\}$ [23]	
Strongly Convex	Nonsmooth$^{(1)}$	N.A.	$O\{1/\epsilon\}$ [*]$^{(3)}$
	Smooth$^{(2)}$	$\mathcal{O}\{1/\epsilon^{1.25}\}$ [23]	

$^{(1)}$ Results in the literature assume a non-smooth innermost layer and smooth outer layers. The SSD framework allows the outer layers to be smooth, smoothable or semi-smooth [*].
$^{(2)}$ All layers are smooth.
$^{(3)}$ We need all outer layers to be smooth to obtain the $\mathcal{O}\{1/\epsilon\}$ complexity.

Table 2: k-Layer Iteration Complexity

In the Literature	SSD Framework
$\mathcal{O}\{1/\epsilon^{2k}\}$[24]	$O\{k^3/\epsilon^2\}$[*]
$\mathcal{O}\{1/\epsilon^{(7+k)/4}\}$[24]	
N.A.	$O\{k^2/\epsilon\}$[*]$^{(3)}$
$\mathcal{O}\{1/\epsilon^{(3+k)/4}\}$[24]	

① We expresss the gradient of $f_i : \mathbb{R}^m \to \mathbb{R}^n$ as a $R^{n \times m}$ matrix such that the directional derivative $f_i'(x; d) = f_i'(x)d$. Such a representation helps to simplify notations in deriving the gradient of a composed function.

1.2 Our Contributions

In this paper, we provide a positive answer to this question for solving a special but broad class of convex NSCO problems, by closing the gap in sampling complexities between these convex NSCO problems and one-layer stochastic optimization problems. More specifically, we consider the following multi-layer problem with a simple (strongly) convex regularization term $u(x)$:

$$\min_{x\in X}\{f(x) := f_1 \circ f_2 \circ \ldots \circ f_k(x) + u(x)\}, \tag{1.3}$$

and make the following two assumptions about its structure.

Assumption 1 *Every layer function f_i is convex such that it can be reformulated using the Fenchel conjugate f_i^**[②]

$$f_i(y_i) = \max_{\pi_i \in \Pi_i} \pi_i y_i - f_i^*(\pi_i), \tag{1.4}$$

where Π_i is the domain of f_i^*.

The composite function $f(x) := f_1 \circ f_2 \circ \ldots \circ f_k(x) + u(x)$ satisfies the following *compositional monotonicity assumption.*

Assumption 2 *For every non-linear layer function (i.e., $f_i(y_i)$ is not affine with respect to y_i), $\pi_{1:i-1} := \pi_1\pi_2\ldots\pi_{i-1}$ is always component-wise non-negative for any $\pi_1 \in \Pi_1$, $\pi_2 \in \Pi_2$, $\ldots$,$\pi_{i-1} \in \Pi_{i-1}$.*

Note that these two assumptions together form a sufficient condition for the convexity of $f_1 \circ \ldots \circ f_k$, and hence are stronger than the overall convexity assumption in [23,22,24].

These stronger assumptions allow us to rewrite every layer function in (1.3) using biconjugation to obtain a saddle point reformulation:

$$\min_{x\in X} \max_{\pi_{1:k}\in \Pi_{1:k}} \{\mathcal{L}(x;\pi_{1:k}) := \mathcal{L}_1(x;\pi_{1:k}) + u(x)\}, \tag{1.5}$$

where the nested Lagrangian function is defined as

$$\mathcal{L}_i(x;\pi_{i:k}) := \begin{cases} x & \text{if } i = k+1, \\ \pi_i \mathcal{L}_{i+1}(x;\pi_{i+1:k}) - f_i^*(\pi_i) & \text{if } 0 \le i \le k. \end{cases} \tag{1.6}$$

Then by designing algorithms for this reformulated problem, we are able to improve the previous results for solving NSCO in four aspects.

Firstly, we introduce a simple stochastic sequential dual algorithm, called vanilla SSD, to solve the smooth-nonsmooth two-layer problem considered in [22] (but with Assumption 1 and 2). Notice that, other than stochasticity, the saddle point reformulation (1.5) for the two-layer problem is similar to the "$\min\max\max$" problem considered in [26]. Under the deterministic setting, Zhang et. al. [26] presented an efficient first-order method which performs proximal updates for π_2 and then for π_1, before an proximal update for x in each iteration. In order to handle the stochastic layer functions, we extend the primal-dual interpretation in [14] to show that unbiased estimators for the dual proximal update π_i can be obtained with a single query to $\mathcal{SO}_i$. These two innovations allow us to achieve the $\mathcal{O}\{1/\epsilon^2\}$ and $\mathcal{O}\{1/\epsilon\}$ sampling complexities for solving convex and strongly convex two-layer NSCO problems, respectively.

Secondly, we propose a general stochastic sequential dual (SSD) algorithmic framework to solve multi-layer NSCO problems. In each iteration, the SSD framework performs proximal updates for all dual variables $\pi_k, ..., \pi_1$ sequentially before an x proximal update. In contrast to the usual smoothness assumption for all outer layer functions [22,23,24], the SSD framework can handle a rather general composition of different types of layer functions, including smooth, smoothable and non-smooth functions. We show that SSD can achieve the optimal $\mathcal{O}\{1/\epsilon^2\}$ (resp., $\mathcal{O}\{1/\epsilon\}$) sampling complexity for solving convex (resp., strongly convex) NSCO problems. Moreover, by incorporating momentum extrapolations in the dual updates, SSD exhibits optimal iteration complexities for solving deterministic multilayer problems as well. For example, if all the layer functions are smooth and deterministic, then the iteration complexities are given by $\mathcal{O}\{1/\sqrt{\epsilon}\}$ or $\mathcal{O}\{\log(1/\epsilon)\}$ depending on whether the problem is convex or strongly convex.

Thirdly, the SSD framework is modular. By decomposing the optimality criterion of the entire saddle point problem (1.5) into relatively decoupled individual optimality criterion for each variable $\pi_1, ..., \pi_k, x$, we are able to design modular dual proximal updates for each type of layer function (e.g., smooth, smoothable, and non-smooth).

[②] Note that the conjugate of the vector function $f_i(y_i) := [f_{i,1}, f_{i,2}, \ldots, f_{i,n_{i-1}}]$ is defined component-wise, i.e, $f_i^*(\pi_i) := [f_{i,1}^*(\pi_{i,1}), f_{i,2}^*(\pi_{i,2}), \ldots, f_{i,m_{i-1}}^*(\pi_{i,m_{i-1}})]$ where $f_{i,j}^*$ is the Fenchel conjugate of $f_{i,j}$.

These modular dual updates are almost independent of other composing layer functions, and so deploying the SSD framework for a new nested composite problem is easy. The user only needs to plug the corresponding modular dual updates for each layer function into the SSD framework. Moreover, the detailed iteration complexities of the SSD framework also have an easy modular interpretation. For deterministic problems, the complexities of the SSD framework are given by the sum of optimal iteration complexities for solving some simplified one-layer problems,

$$\min_{x \in X} b_i^\top f_i(A_i x + c_i) + u(x)/k, \ \forall i, \tag{1.7}$$

where b_i and (A_i, c_i) are some linearizion of $f_{1:i-1}$ and $f_{i+1:k}$[③]. For stochastic convex problems, the sampling complexity nearly matches (other than a factor of k) the sum of sampling complexities for solving simplified stochastic one-layer problems in (1.7), where (A_i, c_i) are accessible only through stochastic oracles. In stochastic strongly convex setting, the sampling complexity for finding an ϵ-close solution, i.e., $\mathbb{E}[\|x_N - x^*\|^2] \le \epsilon$, also admits a similar interpretation.

Fourthly, we illustrate the SSD framework by applying it to two interesting applications; one involving minimizing the mean-upper-semideviation risk of order 1 for a two-stage stochastic linear program and the other one involving minimizing the maximum loss associated with a system of stochastic composite functions. We show that the SSD framework is able to achieve the optimal $\mathcal{O}\{1/\epsilon^2\}$ iteration complexity for both of them.

The rest of the paper is organized as follows. First, Section 2 introduces the simple vanilla SSD algorithm for the two-layer problem. Next, Section 3 introduces the full-fledged SSD framework, followed by Sections 4 and 5 which develop the modular dual updates and establishes the convergence properties, respectively. Finally we provide two applications in Section 6 and make some concluding remarks in Section 7.

1.3 Notations& Terminology

We use the following notations throughout the paper.

- $\|\cdot\|$ denotes the l_2 (operator) norm unless specified otherwise. The feasible region X is convex and compact with $\mathcal{D}_X^2 = \max_{x_1, x_2 \in X} \frac{1}{2}\|x_1 - x_2\|^2 < \infty$. We use x^* to denote some optimal solution (its existence is guaranteed by the compactness of X and the continuity of all layer functions).
- Every layer function f_i is defined on $\mathbb{R}^{n_{i+1}}$ and it is closed, proper and convex. We call a layer function *deterministic* if there exists a first-order oracle $\mathcal{O}_i$, which, when queried at any $y_i \in Y_i$, returns $(f_i(y_i), f_i'(y_i))$ for some $f_i'(y_i) \in \partial f_i(y_i)$. We assume that the selection of subgradient is consistent, that is, $f_i'(y_i)$ is a fixed element of $\partial f_i(y_i)$. We call it *stochastic* if there exists a stochastic first-order oracle $\mathcal{SO}_i$, which, when queried at some $y_i \in Y_i$, returns a pair of stochastic estimates $(f_i(y_i, \xi_i), f_i'(y_i, \xi_i))$ with

  $$\mathbb{E}[f_i(y_i, \xi_i)] = f_i(y_i), \ \mathbb{E}[f_i'(y_i, \xi_i)] = f_i'(y_i), \mathrm{Var}(f_i(y_i, \xi_i)) \le \sigma_{f_i}^2, \text{ and } \mathrm{Var}(f_i'(y_i, \xi_i)) \le \sigma_{\pi_i}^2.$$

  Moreover, results returned by different queries to $\mathcal{SO}_i$ are independent, and all $\{\mathcal{SO}_i\}$ are independent. Notice that $\mathcal{O}_i$ is a special case of $\mathcal{SO}_i$ with σ_{f_i} and σ_{π_i} being zero, so we use the $\mathcal{SO}_i$ notation when referring to both of them.
- The stochastic sub-gradients satisfies $\mathbb{E}[\|f_i'(y_i, \xi_i)\|^2] \le M_{\Pi_i}^2 < \infty \, \forall y_i$. Note that the Jensen's inequality implies that every f_i is also M_{Π_i}-Lipschitz continuous.
- The Fenchel conjugate of a convex function $g(x)$ is defined as $g^*(y) := \max_x \langle x, y \rangle - g(x)$. The Bregman's distance (or prox-function) associated with a convex function g is defined as $D_g(x, y) = g(y) - g(x) - \langle g'(x), y - x \rangle$. An important relationship for Bregman distance functions of conjugate functions is

  $$D_g(x, y) = D_g^*(g'(y), g'(x)). \tag{1.8}$$

- We say a layer function f_i is L_{f_i}*-smooth* if it satisfies $\|f_i'(y_i^1) - f_i(y_i^2)\| \le L_{f_i}\|y_i^1 - y_i^2\|, \ \forall y_i^1, y_i^2 \in \mathbb{R}^{n_{i+1}}$. We call a function $g(y)$ α*-strongly convex with respect to Bregman distance function* V if it satisfies $g(y^1) - g(y^2) - \langle g'(y^2), y^1 - y^2 \rangle \ge V(y^1, y^2), \ \forall y^1, y^2 \in Y$, and simply call it α*-strongly convex* if the Bregman distance function is $\frac{1}{2}\|x_2 - x_1\|^2$. Moreover, we call a function $g(y)$ *simple* if the following problem can be computed efficiently for any $\pi \in R^n$, $\bar{y} \in Y$ and $\eta \ge 0$: $\min_{y \in Y} \langle \pi, y \rangle + g(y) + \eta\|y - \bar{y}\|^2$.

③ We use the $i:j$ notation as a shorthand for all layer indices from i to j, i.e., $\{i, i+1, ..., j\}$, and if $j = k$, we may simply write it as $i:$. Here $f_{i:j}$ means $f_i \circ f_{i+1} \circ ... \circ f_j$.

2 Vanilla SSD for Two-layer Problems

We develop in this section a vanilla SSD algorithm for solving two-layer problems. Such a simple set-up provides a gentle but instructive preview to the general multilayer SSD algorithm without heavy notations and technicalities.

2.1 Saddle-point Reformulation

Consider

$$\min_{x \in X} \{f(x) := f_1(f_2(x)) + u(x)\}, \tag{2.1}$$

where $\{f_i\}$ are convex and M_{Π_i}-Lipschitz continuous and $u(x)$ is simple and strongly convex with modulus $\alpha \geq 0$. We impose the following outer-layer smoothness assumption similar to [22].

Assumption 3 *f_1 is L_{f_1}-Lipschitz smooth with $\mathcal{D}^2_{\Pi_1} := \max_{\pi_1, \bar{\pi}_1 \in dom(f_i^*)} D_{f_i^*}(\pi_1, \bar{\pi}_1) < \infty$, i.e., the radius of the dual variable π_1 with respect to $D_{f_1^*}$ is finite.*

Note that we only need f_2 (f_1) to be Lipschitz continous on X (resp. f_2 image of X), so the preceding assumptions can be easily satisfied for a compact X.

Since both f_1 and f_2 are convex, applying bi-conjugations to them lead to the following *composition Lagrangians*,

$$\begin{aligned} \mathcal{L}_2(x; \pi_2) &= \pi_2 x - f_2^*(\pi_2), \\ \mathcal{L}_1(x; \pi_1, \pi_2) &= \pi_1 \mathcal{L}_2(x; \pi_2) - f_1^*(\pi_1), \end{aligned} \tag{2.2}$$

where $\pi_1 \in \mathbb{R}^{1 \times m_1}$ and $\pi_2 \in \mathbb{R}^{m_1 \times n}$ are the dual variables for f_1 and f_2, and f_1^* and f_2^* are their Fenchel conjugates. Using $\mathcal{L}_1(x; \pi_1, \pi_2)$, we can reformulate (2.1) as a saddle point problem given by

$$\min_{x \in X} \max_{\pi_2 \in \Pi_2} \max_{\pi_1 \in \Pi_1} \{\mathcal{L}(x; \pi_1, \pi_2) = \mathcal{L}_1(x; \pi_1, \pi_2) + u(x)\}, \tag{2.3}$$

where Π_1 and Π_2 denote the domains of f_1^* and f_2^*, respectively.

Proposition 1 below describes some basic duality relationship between (2.1) and (2.3).

Proposition 1 *Let f and $\mathcal{L}$ be defined in* (2.1) *and* (2.3)*. Then the following relations holds for all $x \in X$.*

a) Weak duality: $f(x) \geq \mathcal{L}(x; \pi_1, \pi_2) \quad \forall (\pi_1, \pi_2) \in \Pi_1 \times \Pi_2$.
b) Strong duality: $f(x) = \mathcal{L}(x; \pi_1^*, \pi_2^*)$ for any $\pi_2^* \in \partial f_2(x)$ and $\pi_1^* \in \partial f_1(f_2(x))$.

Proof Let $x \in X$ be given. For the strong duality, it follows from the definitions of π_1^* and π_2^* that

$$\begin{aligned} f_2(x) &= \mathcal{L}_2(x; \pi_2^*), \\ f_1(f_2(x)) &= \mathcal{L}_1(x; \pi_1^*, \pi_2^*), \end{aligned}$$

thus we get $f(x) = \mathcal{L}_1(x; \pi_1^*, \pi_2^*) + u(x)$. For the weak duality, the following decomposition holds for any feasible (π_1, π_2):

$$\begin{aligned} \mathcal{L}(x; \pi_1^*, \pi_2^*) - \mathcal{L}(x; \pi_1, \pi_2) &= \mathcal{L}(x; \pi_1^*, \pi_2^*) - \mathcal{L}(x; \pi_1, \pi_2^*) + \mathcal{L}(x; \pi_1, \pi_2^*) - \mathcal{L}(x; \pi_1, \pi_2) \\ &= \underbrace{\mathcal{L}_1(x; \pi_1^*, \pi_2^*) - \mathcal{L}_1(x; \pi_1, \pi_2^*)}_{A} + \underbrace{\pi_1(\mathcal{L}_2(x; \pi_2^*) - \mathcal{L}_2(x; \pi_2))}_{B}. \end{aligned}$$

Then $\pi_1^* \in \operatorname{Arg\,max}_{\pi_1 \in \Pi_1} \{\pi_1 f_2(x) - f_1^*(\pi_1) \equiv \mathcal{L}_1(x; \pi_1, \pi_2^*)\}$ implies that $A \geq 0$. $B \geq 0$ because $\mathcal{L}_2(x; \pi_2^*) - \mathcal{L}_2(x; \pi_2) \geq 0$ and Π_1 is non-negative. Combining these observations, we conclude that $f(x) = \mathcal{L}(x; \pi_1^*, \pi_2^*) \geq \mathcal{L}(x; \pi_1, \pi_2)$. ■

For a given $\bar{z} := (\bar{x}; \bar{\pi}_1, \bar{\pi}_2) \in X \times \Pi_1 \times \Pi_2$, we define the *gap function* with respect to some feasible *reference point* $z := (x; \pi_1, \pi_2)$ as

$$Q(\bar{z}, z) := \mathcal{L}(\bar{x}; \pi_1, \pi_2) - \mathcal{L}(x; \bar{\pi}_1, \bar{\pi}_2). \tag{2.4}$$

Note that $Q(\bar{z}, z) \leq 0, \forall z \in Z$, if and only if $\bar{z}$ is a saddle point of problem (2.3). The above gap function can also provide an upper bound on the functional optimality gap, i.e., the difference between function value $f(\bar{x})$ and the optimal value $f(x^*)$, for problem (2.1).

More specifically, to bound the functional optimality gap at an ergodic solution $\bar{x}^N$, we can focus on only a few reference points. The following notation will be used to construct these reference points:

$$\pi_2^* \in \partial f_2(x^*), \ \pi_1^* \in \partial f_1(f_2(x^*)), \ \hat{\pi}_2 \in \partial f_2(\bar{x}^N), \text{ and } \hat{\pi}_1 \in \partial f_1(f_2(\bar{x}^N)).$$

Proposition 2 below shows how to convert a bound on Q to the functional optimality gap.

Proposition 2 *Let a sequence of feasible solutions* $\{z^t \equiv (x^t; \pi_1^t, \pi_2^t)\}$ *be given and let* $\bar{x}^N$ *denote the ergodic mean of* $\{x^t\}$ *given by* $\sum_{t=0}^{N-1} w^t x^{t+1} / \sum_{t=0}^{N-1} w^t$ *with* $w^t > 0$*. If* $\hat{z} = (x^*; \hat{\pi}_1, \hat{\pi}_2)$ *and* $\sum_{t=0}^{N-1} w^t Q(z^{t+1}, \hat{z}) \le B$*, then*

$$f(\bar{x}^N) - f(x^*) \le \frac{B}{\sum_{t=0}^{N-1} w^t}. \tag{2.5}$$

Proof In view of Proposition 1.b) and the convexity of $\mathcal{L}(\hat{\pi}_2, \hat{\pi}_1, x)$ with respect to x, we have

$$(\textstyle\sum_{t=0}^{N-1} w^t) f(\bar{x}^N) = (\sum_{t=0}^{N-1} w^t)\mathcal{L}(\bar{x}^N; \hat{\pi}_1, \hat{\pi}_2) \le \sum_{t=0}^{N-1} w^t \mathcal{L}(x^{t+1}; \hat{\pi}_1, \hat{\pi}_2).$$

Moreover, Proposition 1.a) implies that $f(x^*) \ge \mathcal{L}(\pi_2^{t+1}, \pi_1^{t+1}, x^*)$ $\forall t$, so

$$\textstyle\sum_{t=0}^{N-1} w^t f(x^*) \ge \sum_{t=0}^{N-1} w^t \mathcal{L}(x^*; \pi_1^{t+1}, \pi_2^{t+1}).$$

Combining the previous two inequalities, we obtain $\sum_{t=0}^{N-1} w^t (f(\bar{x}^N) - f(x^*)) \le \sum_{t=0}^{N-1} w^t Q(z^{t+1}, \hat{z})$. Then the desired inequality follows from dividing both sides by $\sum_{t=0}^{N-1} w^t$. ■

The next functional optimality gap conversion method uses both $\|\bar{x}^N - x^*\|^2$ and the Q function. It will help us to improve the convergence rate for solving strongly convex problems.

Proposition 3 *Let a sequence of feasible solutions* $\{z^t \equiv (x^t; \pi_1^t, \pi_2^t)\}$ *be given and let* $\bar{x}^N$ *denote the ergodic mean of* $\{x^t\}$ *given by* $\sum_{t=0}^{N-1} w^t x^{t+1} / \sum_{t=0}^{N-1} w^t$ *with* $w^t \ge 0$*. If* $\sum_{t=0}^{N-1} w^t Q(z^{t+1}, \tilde{z}) \le B$ *for* $\tilde{z} = (x^*; \pi_1^*, \hat{\pi}_2)$ *and* $\frac{1}{2}\|\bar{x}^N - x^*\|^2 \le C$*, then*

$$f(\bar{x}^N) - f(x^*) \le \frac{B}{\sum_{t=0}^{N-1} w^t} + L_{f_1} M_{\Pi_2}^2 C. \tag{2.6}$$

Proof By an argument similar to Proposition 2, we have $\mathcal{L}(\bar{x}^N; \pi_1^*, \hat{\pi}_2) - f(x^*) \le \frac{B}{\sum_{t=0}^{N-1} w^t}$.

The difference between $f(\bar{x}^N) = \mathcal{L}(\bar{x}^N; \hat{\pi}_1, \hat{\pi}_2)$ and $\mathcal{L}(\bar{x}^N; \pi_1^*, \hat{\pi}_2)$ satisfies

$$\begin{aligned}\mathcal{L}(\bar{x}^N; \hat{\pi}_1, \hat{\pi}_2) - \mathcal{L}(\bar{x}^N; \pi_1^*, \hat{\pi}_2) &= f_1^*(\pi_1^*) - f_1^*(\hat{\pi}_1) - (\pi_1^* - \hat{\pi}_1) f_2(\bar{x}^N) \\ &\overset{(a)}{=} D_{f_1^*}(\hat{\pi}_1, \pi_1^*) \overset{(b)}{=} D_{f_1}(f_2(x^*), f_2(\bar{x}^N)) \overset{(c)}{\le} L_{f_1} M_{\Pi_2}^2 \tfrac{1}{2}\|\bar{x}^N - x^*\|^2,\end{aligned}$$

where (a) follows from the conjugate duality $f_2(\bar{x}^N) \in \partial f_1^*(\hat{\pi}_1)$, (b) follows from the relationship between Bregman distances generated by conjugate functions (1.8), and (c) follows from the L_{f_1}-smoothness of f_1 in Assumption 3. Therefore,

$$\begin{aligned} f(\bar{x}^N) - f(x^*) &= \mathcal{L}(\bar{x}^N; \hat{\pi}_1, \hat{\pi}_2) - \mathcal{L}(\bar{x}^N; \pi_1^*, \hat{\pi}_2) + \mathcal{L}(\bar{x}^N; \pi_1^*, \hat{\pi}_2) - f(x^*) \\ &\le \frac{B}{\sum_{t=0}^{N-1} w^t} + L_{f_1} M_{\Pi_2}^2 C. \end{aligned}$$

■

2.2 Stochastic Layer Functions and Implicit Proximal Updates

With the saddle point reformulation in (2.3), we can solve problem (2.1) possibly by applying the sequential dual (SD) algorithm proposed in [26]. The basic idea of the SD method is to perform proximal updates to the dual variables π_2 and π_1 sequentially before the updating of the primal variable x in each iteration. For example, given $(x^t; \pi_1^t, \pi_2^t)$ the SD method updates π_1^t by

$$\pi_1^{t+1} := \arg\max_{\pi_1 \in \Pi_1} \pi_1 \mathcal{L}_2(x^t; \pi_2^{t+1}) - f_1^*(\pi_1) - \tau_1^t D_{f_1^*}(\pi_1^t, \pi_1), \tag{2.7}$$

after the updating π_2^t to π_2^{t+1}.

However stochastic layer functions pose some new challenges. Specifically, for a deterministic problem with known and simple f_1^* and f_2^*, the dual iterates $\{\pi_i^t\}$ can be computed explicitly through the proximal updates in the SD algorithm [26]. In contrast, (2.1) provide only access to f_1 and f_2 through stochastic first-order oracles, but no direct access to f_1^* or f_2^*. As a consequence, it is impossible to explicitly obtain the arguments, e.g., $\mathcal{L}_2(x^t; \pi_2^t)$ in (2.7), for the proximal updates, and to evaluate these proximal updates. Fortunately, due to the conjugate duality relationship we can obtain stochastic estimators for the arguments and evaluate these proximal updates "implicitly" by calling the stochastic first-order oracles.

More specifically, suppose that the dual point π_i is *associated* with some primal point $\underline{y}_i$, i.e., $\pi_i = f_i'(\underline{y}_i) \in \partial f_i(\underline{y}_i)$. Even though the exact value of π_i is unknown, the relation

$$f_i(\underline{y}_i) = \pi_i \underline{y}_i - f_i^*(\pi_i) \Longleftrightarrow \pi_i \in \partial f_i(\underline{y}_i) \Longleftrightarrow \pi_i \in \arg\max_{\bar{\pi}_i \in \Pi_i} \bar{\pi}_i \underline{y}_i - f_i^*(\bar{\pi}_i) \tag{2.8}$$

allows us to derive stochastic estimators for π_i and $f_i^*(\pi_i)$ with a single query to $\mathcal{SO}_i$,

$$\pi_i(\xi_i) := f_i'(\underline{y}_i, \xi_i) \text{ and } f_i^*(\pi_i, \xi_i) := \pi_i(\xi_i)\underline{y}_i - f_i(\underline{y}_i, \xi_i). \tag{2.9}$$

In other words, we have access to an implicit dual stochastic oracle, denoted by $\mathcal{DSO}_i$, for each stochastic layer function f_i, which upon request at $\pi_i = f_i'(\underline{y}_i)$, returns a pair of unbiased estimators for π_i and $f_i^*(\pi_i)$. Furthermore, we can construct from these dual estimators *stochastic Lagrangian*s to approximate the composition Lagrangians in (2.2),

$$\mathcal{L}_2(\cdot; \pi_2(\xi_2)) := \pi_2(\xi_2) \cdot - f_2^*(\pi_2, \xi_2), \tag{2.10}$$

$$\mathcal{L}_1(\cdot; \pi_1(\xi_1), \pi_2(\xi_2)) := \pi_1(\xi_1)\mathcal{L}_2(\cdot; \pi_2(\xi_2)) - f_1^*(\pi_1, \xi_1), \tag{2.11}$$

and use them as arguments for proximal updates, e.g., $\mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2))$ in place of $\mathcal{L}_2(x^t; \pi_2^{t+1})$ for the proximal update in (2.7).

In addition, we can compute stochastic estimators for the proximal updates by utilizing the stochastic first-order oracles. Let us start with a special case of proximal updates with stepsize 0 (e.g., $\tau_1^t = 0$ in (2.7)). Equivalently, we consider the dual iterate π_i^{t+1} generated simply by maximization, $\pi_i^{t+1} \in \arg\max \pi_i \underline{y}_i - f_i^*(\pi_i)$[④]. It is clear that $\pi_i^{t+1} = f_i'(\underline{y}_i)$. Hence, by calling the $\mathcal{SO}_i$ at the $\underline{y}_i$ (or $\mathcal{DSO}_i$ at the associated $\pi_i = f_i'(\underline{y}_i)$), we can compute an unbiased estimator of π_i^{t+1}. In the sequel, we call such a maximization an *implicit maximization* because the computation only involves the association of π_i^{t+1} with the primal point $\underline{y}_i$ and a call to $\mathcal{SO}_i$, rather than an actual maximization problem.

To compute stochastic estimators for the proximal updates with positive stepsize, we need to generalize the result in [14,13] about the relationshiop between proximal update and gradient computation. In particular, if the Bregman's distance is given by $D_{f_i^*}$, then the dual iterate π_i^{t+1} returned by a proximal update starting from $\pi_i^t = f_i'(\underline{y}_i^t)$ is also associated with some primal point $\underline{y}_i^{t+1}$ as shown in Proposition 4.

Proposition 4 *If π_i^t is associated with some primal point $\underline{y}_i^t$, i.e., $\pi_i^t = f_i'(\underline{y}_i^t) \in \partial f_i(\underline{y}_i^t)$, then the proximal update from π_i^t to π_i^{t+1} with argument y_i^{t+1}, i.e.,*

$$\pi_i^{t+1} \in \arg\min_{\pi_i \in \Pi_i} -\pi_i y_i^{t+1} + f_i^*(\pi_i) + \tau_i^t D_{f_i^*}(\pi_i^t, \pi_i) \tag{2.12}$$

satisfies

$$\pi_i^{t+1} = f_i'(\underline{y}_i^{t+1}) \in \partial f_i(\underline{y}_i^{t+1}) \text{ with } \underline{y}_i^{t+1} := (y_i^{t+1} + \tau_i^t \underline{y}_i^t)/(1 + \tau_i^t). \tag{2.13}$$

Proof This is a direct consequence of the conjugate duality relationship (2.8),

$$\begin{aligned} & \pi_i^{t+1} \in \arg\min_{\pi_i \in \Pi_i} -\pi_i y_i^{t+1} + f_i^*(\pi_i) + \tau_i^t D_{f_i^*}(\pi_i^t, \pi_i) \\ \overset{(a)}{\Longleftrightarrow}\ & \pi_i^{t+1} \in \arg\min_{\pi_i \in \Pi_i} -\pi_i y_i^{t+1} - \tau_i^t \pi_i \underline{y}_i^t + (1+\tau_i^t) f_i^*(\pi_i) \\ \Longleftrightarrow\ & \pi_i^{t+1} \in \arg\min_{\pi_i \in \Pi_i} -\pi_i \underbrace{\left(\tfrac{y_i^{t+1} + \tau_i^t \underline{y}_i^t}{1+\tau_i^t}\right)}_{\underline{y}_i^{t+1}} + f_i^*(\pi_i) \overset{(b)}{\Longleftrightarrow} \pi_i^{t+1} = f_i'(\underline{y}_i^{t+1}) \in \partial f_i(\underline{y}_i^{t+1}), \end{aligned}$$

where (a) follows from $\underline{y}_i^t \in \partial f_i^*(\pi_i^t)$ and (b) follows from (2.8). ∎

In view of Proposition 4, we can compute an unbiased estimator of π_i^{t+1} by calling the stochastic oracle $\mathcal{SO}_i$ at $\underline{y}_i^{t+1}$ (or $\mathcal{DSO}_i$ at the associated point $\pi_i^{t+1} = f_i'(\underline{y}_i^{t+1})$). In the sequel, we call such a proximal update an *implicit proximal update* because the computation involves the calculation of $\underline{y}_i^{t+1}$ in (2.13), the association of (implicit) π_i^{t+1} with it, and the call to the $\mathcal{SO}_i$, but does not require explicit access to f_i^*.

Therefore, the essential ingredients for designing sequential dual type algorithms are still applicable. We just need to replace proximal updates with implicit proximal updates and maximization with implicit maximization, and use appropriate stochastic estimators for the unknown arguments of these operators.

④ We assume that the selection of maximizer is consistent with the subgradient selection, i.e., $\pi_i^{t+1} = f_i'(\underline{y}_i) \in \partial f_i(\underline{y}_i)$.

2.3 Vanilla SSD Algorithm

The design of the vanilla SSD algorithm has been motivated by a decomposition of the gap function Q (c.f., (2.4)) into

$$Q(z^{t+1}, z) = Q_2(z^{t+1}, z) + Q_1(z^{t+1}, z) + Q_0(z^{t+1}, z), \tag{2.14}$$

where Q_2, Q_1, and Q_0 defined below are related to the optimality of π_2^{t+1}, π_1^{t+1} and x^{t+1}, respectively,

$$\begin{aligned} Q_2(z^{t+1}, z) &:= \mathcal{L}(x^{t+1}; \pi_1, \pi_2) - \mathcal{L}(x^{t+1}; \pi_1, \pi_2^{t+1}) \\ &= \pi_1[\pi_2 x^{t+1} - f_2^*(\pi_2)] \boxed{-\pi_1[\pi_2^{t+1} x^{t+1} - f_2^*(\pi_2^{t+1})]}, \end{aligned} \tag{2.15}$$

$$\begin{aligned} Q_1(z^{t+1}, z) &:= \mathcal{L}(x^{t+1}; \pi_1, \pi_2^{t+1}) - \mathcal{L}(x^{t+1}; \pi_1^{t+1}, \pi_2^{t+1}) \\ &= \pi_1 \mathcal{L}_2(x^{t+1}; \pi_2^{t+1}) - f_1^*(\pi_1) \boxed{-[\pi_1^{t+1} \mathcal{L}_2(x^{t+1}; \pi_2^{t+1}) - f_1^*(\pi_1^{t+1})]}, \end{aligned} \tag{2.16}$$

$$\begin{aligned} Q_0(z^{t+1}, z) &:= \mathcal{L}(x^{t+1}; \pi_1^{t+1}, \pi_2^{t+1}) - \mathcal{L}(x; \pi_1^{t+1}, \pi_2^{t+1}) \\ &= \boxed{\pi_1^{t+1} \pi_2^{t+1} x^{t+1} + u(x^{t+1})} - (\pi_1^{t+1} \pi_2^{t+1} x + u(x)). \end{aligned} \tag{2.17}$$

To decrease the gap function Q, we intend to find π_2^{t+1}, π_1^{t+1} and x^{t+1} that can reduce the boxed terms in their respective gap functions Q_2, Q_1, and Q_0. Accordingly, the proposed vanilla SDD method (see Algorithms 1 and 2) for solving problem (2.1) consists of three steps to compute π_2^{t+1}, π_1^{t+1} and x^{t+1} in each iteration. Note that we state this algorithm in its dual form and primal form, respectively, in Algorithms 1 and 2. While the dual form provides us better intuition about the design of the algorithm, the primal form is more convenient for implementation.

More specifically, Algorithm 1 can be interpreted as follows. In Line 2, we aim to reduce $Q_2(z^{t+1}, z)$ by choosing an *associated* π_2^{t+1} with the smallest value of $-\pi_1[\pi_2^{t+1} x^{t+1} - f_i^*(\pi_2^{t+1})]$. Since π_1 is non-negative and Π_2 is row separable, it is equivalent to finding the component-wise smallest vector $-[\pi_2^{t+1} x^{t+1} - f_2^*(\pi_1^{t+1})]$. Then using x^t in place of the to-be-evaluated x^{t+1}, we define π_2^{t+1} through implicit-maximization $\arg\max_{\pi_2 \in \Pi_2} \pi_2 x^t - f_2^*(\pi_2)$, and call the stochastic oracle to compute the unbiased estimators of π_2^{t+1}. Observe that we label the unbiased estimators returned from calls to $\mathcal{DSO}_i$ with a superscript j, i.e., ξ_i^j, to emphasize that they will be used as part of arguments for the proximal update at the j-th layer.

Next, Line 3 of Algorithm 1 tends to reduce $Q_1(z^{t+1}, z)$ by specifying an associated π_1^{t+1} with a small value of $-[\pi_1^{t+1} \mathcal{L}_2(x^{t+1}; \pi_2^{t+1}) - f_1^*(\pi_1^{t+1})]$. Since x^{t+1} is still unavailable and only a stochastic estimator $\mathcal{L}_2(\cdot; \pi_2^{t+1}(\xi_2^1))$ is accessible, it uses the stochastic estimator $\mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1))$ in place of $\mathcal{L}_2(x^{t+1}; \pi_2^{t+1})$ to define π_1^{t+1} as

$$\pi_1^{t+1} \in \arg\max_{\pi_1 \in \Pi_1} \pi_1 \mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1)) - f_1^*(\pi_1) - \tau_1^t D_{f_1^*}(\pi_1^t, \pi_1),$$

and then computes an unbiased estimator of π_1^{t+1} by calling the $\mathcal{DSO}_1$ to perform the implicit proximal update.

Finally, Line 4 needs to choose an x^{t+1} with a small value of $\pi_1^{t+1} \pi_2^{t+1} x^{t+1} + u(x^{t+1})$ in $Q_0(z^{t+1}, z)$. Here we use stochastic estimator $\pi_1^{t+1}(\xi_1^0)\ \pi_2^{t+1}(\xi_2^0)$ in place of $\pi_1^{t+1} \pi_2^{t+1}$ to arrive at

$$x^{t+1} = \arg\min_{x \in X} \pi_1^{t+1}(\xi_1^0) \pi_2^{t+1}(\xi_2^0) x + u(x) + \tfrac{\eta^t}{2} \|x - x^t\|^2. \tag{2.18}$$

Notice that $\pi_2^{t+1}(\xi_2^0)$ is computed from a call to $\mathcal{DSO}_2$ different from the one used to compute $(\pi_2^{t+1}(\xi_2^1), f_2^*(\pi_2^{t+1}, \xi_2^1)))$. This extra call to $\mathcal{DSO}_2$ is needed to construct an unbiased estimator for $\pi_1^{t+1} \pi_2^{t+1}$.

Algorithm 1 Vanilla Stochastic Sequential Dual (SSD) Algorithm for Two-Layer Problems

---

Input: $x_0 \in X$, $\underline{y}_1^0 \in Y_1$, $\pi_1^0 = f_1'(\underline{y}_1^0) \in \partial f_1(\underline{y}_1^0)$.

1: **for** $t = 0, 1, 2, 3 ... T-1$ **do**

2: Call $\mathcal{DSO}_2$ at π_2^{t+1} twice to obtain $\{(\pi_2^{t+1}(\xi_2^j), f_2^*(\pi_2^{t+1}, \xi_2^j))\}_{j \in \{0,1\}}$, where $\pi_2^{t+1} = f_2'(x^t)$ is defined by the implicit maximization $\arg\max_{\pi_2 \in \Pi_2} \pi_2 x^t - f_2^*(\pi_2)$.

3: Call $\mathcal{DSO}_1$ at π_1^{t+1} to obtain $\pi_1^{t+1}(\xi_1^0)$, where $\pi_1^{t+1} = f_1'(\underline{y}_1^{t+1})$ is defined by the implicit-proximal update $\arg\max_{\pi_1 \in \Pi_1} \pi_1 \mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1)) - f_1^*(\pi_1) - \tau_1^t D_{f_1^*}(\pi_1^t, \pi_1)$.

4: Set $x^{t+1} = \arg\min_{x \in X} \pi_1^{t+1}(\xi_1^0) \pi_2^{t+1}(\xi_2^0) x + u(x) + \frac{\eta^t}{2} \|x - x^t\|^2$.

5: **end for**

6: Return $\bar{x}^N = \sum_{t=0}^{N-1} w^t x^{t+1} / (\sum_{t=0}^{N-1} w^t)$.

Algorithm 2 Primal Form of the Vanilla Stochastic Sequential Dual (SSD) Algorithm

Input: $x_0 \in X$, $\underline{y}_1^0 \in Y_1$.
1: **for** $t = 0, 1, 2, 3...T-1$ **do**
2: Call $\mathcal{SO}_2$ twice at x^t to obtain $\{(f_2(x^t, \xi_2^j), f_2'(x^t, \xi_2^j)\}_{j\in\{0,1\}}$.
3: Set $\underline{y}_1^{t+1} := [f_2(x^t, \xi_2^1) + \tau_1^t \underline{y}_i^t]/(1+\tau_1^t)$ and call $\mathcal{SO}_1$ at $\underline{y}_1^{t+1}$ to obtain $f_1'(\underline{y}_1^{t+1}, \xi_1^0)$.
4: Set $x^{t+1} = \arg\min_{x\in X} f_1'(\underline{y}_1^{t+1}, \xi_1^0) f_2'(x^t, \xi_2^0) x + u(x) + \frac{\eta^t}{2}\|x - x^t\|^2$.
5: **end for**
6: Return $\bar{x}^N = \sum_{t=0}^{N-1} w^t x^{t+1} / (\sum_{t=0}^{N-1} w^t)$.

Moreover, by unpacking the implicit proximal update, implicit maximization, and queries to $\mathcal{DSO}_i$ in Algorithm 1, we can obtain an equivalent primal form in Algorithm 2. Clearly, in this primal form, we do not need to access any conjugate dual information of f_1 and f_2.

We now clarify the relations between the vanilla SSD method and a few other existing algorithms in the literature. Firstly, suppose that there is no stochastic noise. Then Algorithm 1 is related to the sequential dual (SD) method in [26]. However, the SD method in [26] uses proximal update (rather than maximization) to compute π_2^{t+1} for the inner layer. Moreover, it incorporates a few extrapolation steps to achieve the accelerated rate of convergence for solving deterministic problems. This explains why we call Algorithm 1 vanilla SSD since it does not have these acceleration steps. We will show later in this section that this simplified algorithm can already achieve nearly optimal rates of convergence for solving convex and strongly convex two-layer NSCO problems. In next section, we will develop a fully-fledged stochastic sequential dual method that can achieve optimal rates of convergence in terms of their dependence not only on target accuracy, but also on other problem parameters.

Secondly, even though the vanilla SSD method was mainly inspired by the SD method in [26] for solving trilinear saddle point problems, its primal form in Algorithm 2 appears to be quite close to the SCGD in [22]. More specifically, Line 4 of Algorithm 2 also performs a descent step along a subgradient-like direction composed of $f_2'(x^t, \xi_2^0)$ and $f_1'(\underline{y}_1^{t+1}, \xi_1^0)$, returned by $\mathcal{SO}_1$ queried at some averaged $\underline{y}_1^{t+1}$. However, SCGD in [22] employed with the two-time-scale stepsizes can only achieve the suboptimal $\mathcal{O}(1/\epsilon^4)$ (resp., $\mathcal{O}(1/\epsilon^{1.25})$[5]) rate of convergence for solving convex (resp., strongly convex) problems. On the other hand, with a simple stepsize policy the vanilla SSD method can achieve the $\mathcal{O}(1/\epsilon^2)$ (resp. $\mathcal{O}(1/\epsilon)$) rate of convergence, which is optimal in terms of its dependence on ϵ for solving convex (resp., strongly convex) stochastic optimization problems.

2.4 Convergence Analysis for General Convex Problems

In this subsection, we establish the convergence of Algorithm 1 for solving general convex problems by assuming that the strong convexity modulus α of $u(x)$ in problem (2.1) is 0. For the sake of simplicity, we fix the weight w^t used in Algorithm 1 to compute the ergodic output solution $\bar{x}^N$ to be 1. Throughout this section, we use $z^t \equiv (x^t; \pi_1^t, \pi_2^t)$ to denote the iterates generated by Algorithm 1 and set the reference point z in the gap function $Q(\cdot, z)$ to be $(x^*; \pi_1, \pi_2)$, where π_1 and π_2 can depend on the stochastic iterates $\{z^t\}$. The analysis consists of the following parts. We first develop upper bounds for the decomposed gap functions Q_2 and Q_1 defined in (2.15) and (2.16), respectively. Then we suggest a stepsize selection for η^t to bound the decomposed gap function Q_0 defined in (2.17), and hence the overall gap function Q. After that, the convergence in terms of function values is derived from Proposition 2.

We start by providing an upper bound on the decomposed gap function Q_2.

Proposition 5 *Let $\{z^t\}$ be generated by Algorithm 1 and let $\tilde{M}_\Pi = M_{\pi_1} M_{\pi_2}$. Then*

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} Q_2(z^{t+1}, z)] \le \sqrt{2N}\tilde{M}_\Pi \mathcal{D}_X + \mathbb{E}[\sum_{t=0}^{N-1} \frac{\sqrt{2N}\tilde{M}_\Pi}{2\mathcal{D}_X}\|x^t - x^{t+1}\|^2].$$

Proof The dual iterate π_2^{t+1} generated by Line 2 in Algorithm 1 satisfies

$$\mathcal{A}_2^t := \pi_1((\pi_2 - \pi_2^{t+1})x^t - (f_2^*(\pi_2) - f_2^*(\pi_2^{t+1}))) \le 0.$$

Comparing $\mathcal{A}_2^t$ with $Q_2(z^{t+1}, z)$, we can see that the only difference exists in that x^{t+1} in $Q_2(z^{t+1}, z)$ is replaced by x^t in $\mathcal{A}_2^t$, and hence that

$$Q_2(z^{t+1}, z) = \mathcal{A}_2^t + \mathcal{B}_2^t \text{ with } \mathcal{B}_2^t := \pi_1(\pi_2 - \pi_2^{t+1})(x^{t+1} - x^t). \tag{2.19}$$

[5] This iteration complexity needs to assume that f_2 is also smooth. The convergence rate for solving non-smooth and strongly convex two-layer problem has not been established in the literature.

Moreover, Young's inequality implies that

$$\mathbb{E}[\sum_{t=0}^{N-1} \mathcal{B}_2^t - \frac{\sqrt{2N}\tilde{M}_\Pi}{2\mathcal{D}_X}\|x^t - x^{t+1}\|^2] \le 4N\tilde{M}_\Pi^2 \frac{1}{2}\frac{\mathcal{D}_X}{\sqrt{2N}\tilde{M}_\Pi} = \sqrt{2N}\tilde{M}_\Pi\mathcal{D}_X.$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}[\sum_{t=0}^{N-1} Q_2(z^{t+1}, z)] \\
&\le \mathbb{E}[\sum_{t=0}^{N-1} \mathcal{A}_2^t] + \mathbb{E}[\sum_{t=0}^{N-1}(\mathcal{B}_2^t - \frac{\sqrt{2N}\tilde{M}_\Pi}{2\mathcal{D}_X}\|x^{t+1} - x^t\|^2)] + \mathbb{E}[\sum_{t=0}^{N-1} \frac{\sqrt{2N}\tilde{M}_\Pi}{2\mathcal{D}_X}\|x^{t+1} - x^t\|^2] \\
&\le \sqrt{2N}\tilde{M}_\Pi\mathcal{D}_X + \mathbb{E}[\sum_{t=0}^{N-1} \frac{\sqrt{2N}\tilde{M}_\Pi}{2\mathcal{D}_X}\|x^t - x^{t+1}\|^2].
\end{aligned}$$

■

Next we move on to provide an upper bound on Q_1. The following three point inequality lemma (Lemma 3.8 [13]) is important for our development.

Lemma 1 *Assume that the function $g : Y \to \mathbb{R}$ is μ-strongly convex with respect to some Bregman's distance function V, i.e., $g(y) - g(\bar{y}) - \langle g'(\bar{y}), y - \bar{y}\rangle \ge \mu V(\bar{y}, y)\ \forall y, \bar{y} \in Y$. If $\hat{y} \in \arg\min_{y\in Y}\{\langle \pi, y\rangle + g(y) + \tau V(\bar{y}, y)\}$ for some π and $\tau \ge 0$, then*

$$\langle \hat{y} - y, \pi\rangle + g(\hat{y}) - g(y) \le \tau V(\bar{y}, y) - (\tau + \mu)V(\hat{y}, y) - \tau V(\bar{y}, \hat{y})\ \forall y \in Y. \tag{2.20}$$

We also need a technical lemma (similar to Lemma 4.10 [13]) to bound the error resulted from using the stochastic argument $\mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1))$.

Lemma 2 *Let δ^t be some martingale noise with $\mathbb{E}[\delta^t|t] = 0$ and $\mathbb{E}[\|\delta^t\||t] \le \sigma^2\ \forall t$. If variable $\hat{\pi} \in \Pi$ is correlated with $\{\delta^t\}$ with $M_\pi := \max_{\pi\in\Pi}\|\pi\|$, then*

$$\mathbb{E}[\sum_{t=0}^{N-1}\langle \hat{\pi}, \delta^t\rangle] \le \sqrt{N}M_\pi\sigma. \tag{2.21}$$

Proof We need an auxiliary sequence $\{\hat{\pi}^t\}$ which is conditionally uncorrelated with $\{\delta^t\}$. More specifically, let

$$\hat{\pi}^t = \begin{cases} 0 & \text{if } t = 0, \\ \arg\min_{\pi\in\Pi} -\langle \delta^{t-1}, \pi\rangle + \tau\frac{1}{2}\|\hat{\pi}^{t-1} - \pi\|^2 & \text{if } t \ge 1, \end{cases}$$

where $\tau > 0$ is some positive stepsize. Then we have $\mathbb{E}[\langle \hat{\pi}^t, \delta^t\rangle|t] = 0$ and

$$-\langle \hat{\pi}^{t+1} - \hat{\pi}, \delta^t\rangle \le \tau\frac{1}{2}\|\hat{\pi}^t - \hat{\pi}\|^2 - \tau\frac{1}{2}\|\hat{\pi}^{t+1} - \hat{\pi}\|^2 - \tau\frac{1}{2}\|\hat{\pi}^t - \hat{\pi}^{t+1}\|^2,$$

i.e.,

$$-\langle \hat{\pi}^t - \hat{\pi}, \delta^t\rangle \le \tau[\frac{1}{2}\|\hat{\pi}^t - \hat{\pi}\|^2 - \frac{1}{2}\|\hat{\pi}^{t+1} - \hat{\pi}\|^2] + \frac{1}{2\tau}\|\delta^t\|^2.$$

Next summing up the above relations from 0 to $N - 1$ and taking expectation on both sides, we should get

$$\begin{aligned}
\mathbb{E}[\sum_{t=0}^{N-1}\langle \delta^t, \hat{\pi}\rangle] &\overset{(a)}{\le} \tau\frac{1}{2}\mathbb{E}[\|\hat{\pi}\|^2] + \sum_{t=0}^{N-1}\frac{1}{2\tau}\sigma^2 \\
&\le \frac{\tau}{2}M_\pi^2 + \frac{1}{2\tau}N\sigma^2,
\end{aligned}$$

where (a) follows from $\mathbb{E}[\langle \hat{\pi}^t, \delta^t\rangle|t] = 0$. So the desired bound (2.21) follows from picking $\tau := \frac{\sqrt{N}\sigma}{M_\pi}$. ■

We are now ready to state an upper bound on the decomposed gap function Q_2. For the sake of simplicity, we assume here that a constant stepsize policy is used in Algorithm 1. It should be noted that similar results can also be proved for variable stepsize policies, e.g., $\tau_1^t = \mathcal{O}(\sqrt{t})$.

Proposition 6 *Let $\{z^t\}$ be generated by Algorithm 1 with $\tau_1^t = \tau_1 = \sqrt{N}\sqrt{L_{f_1}}\sigma_{f_2}/(\sqrt{2}\mathcal{D}_{\Pi_1})$, $\mathcal{D}_{\Pi_1}^2 := D_{f_1^*}(\pi_1^0, \pi_i^*)$, $\tilde{M}_\Pi := M_{\Pi_1}M_{\Pi_2}$. Then*

$$\mathbb{E}[\sum_{t=0}^{N-1} Q_1(z^{t+1}, z)] \le \sqrt{N}M_{\Pi_1}\sigma_{f_2} + \sqrt{2N}\sqrt{L_{f_1}}\mathcal{D}_{\Pi_1}\sigma_{f_2} + \sqrt{2N}\tilde{M}_\Pi\mathcal{D}_X + \mathbb{E}[\sum_{t=0}^{N-1}\frac{\sqrt{2N}\tilde{M}_\Pi}{2\mathcal{D}_X}\|x^t - x^{t+1}\|^2].$$

Proof The π_1^{t+1} generated by the implicit proximal update in Line 3 of Algorithm 1 satisfies

$$\begin{aligned}\mathcal{A}_1^t &:= (\pi_1 - \pi_1^{t+1})\mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1)) - (f_1^*(\pi_1) - f_1^*(\pi_1^{t+1})) \\ &\le \tau_1^t D_{f_1^*}(\pi_1^t, \pi_1) - \tau_1^t D_{f_1^*}(\pi_1^t, \pi_1^{t+1}) - (\tau_1^t + 1) D_{f_1^*}(\pi_1^{t+1}, \pi_1).\end{aligned} \tag{2.22}$$

Comparing $\mathcal{A}_1^t$ with Q_1, we see that

$$\begin{aligned}Q_1(z^{t+1}, z) &= \mathcal{A}_1^t + (\pi_1 - \pi_1^{t+1})\left[\mathcal{L}_2(x^{t+1}; \pi_2^{t+1}) - \mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1))\right] \\ &= \mathcal{A}_1^t + (\pi_1 - \pi_1^{t+1})\left[(\mathcal{L}_2(x^{t+1}; \pi_2^{t+1}) - \mathcal{L}_2(x^t; \pi_2^{t+1})) + (\mathcal{L}_2(x^t; \pi_2^{t+1}) - \mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1))\right] \\ &= \mathcal{A}_1^t + \mathcal{B}_1^t + \Delta_1^t,\end{aligned} \tag{2.23}$$

where

$$\begin{aligned}\mathcal{B}_1^t &:= (\pi_1 - \pi_1^{t+1})\pi_2^{t+1}(x^{t+1} - x^t), \\ \Delta_1^t &:= (\pi_1 - \pi_1^{t+1})\left[\mathcal{L}_2(x^t; \pi_2^{t+1}) - \mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1))\right].\end{aligned}$$

So, Similar to Proposition 5, we have

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} \mathcal{B}_1^t] \le \sqrt{2N}\tilde{M}_\Pi \mathcal{D}_X + \mathbb{E}[\sum_{t=0}^{N-1} \frac{\sqrt{2N}\tilde{M}_\Pi}{2\mathcal{D}_X}\|x^t - x^{t+1}\|^2]. \tag{2.24}$$

Moreover, the stochastic error Δ_1^t can be split further into

$$\Delta_1^t \ = \Delta_{\pi_1}^t + \delta_1^t, \tag{2.25}$$

with

$$\begin{aligned}\Delta_{\pi_1}^t &:= (\pi_1^t - \pi_1^{t+1})\left[\mathcal{L}_2(x^t; \pi_2^{t+1}) - \mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1))\right], \\ \delta_1^t &:= (\pi_1 - \pi_1^t)\left[\mathcal{L}_2(x^t; \pi_2^{t+1}) - \mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1))\right].\end{aligned}$$

Because $\mathcal{L}_2(x^t; \pi_2^{t+1}) = f_2(x^t)$ and $\mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1)) = f_2(x^t, \xi_2^1)$ (according to (2.9) and (2.10)) , we have

$$\mathbb{E}[\|\mathcal{L}_2(x^t; \pi_2^{t+1}) - \mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1))\|^2] = \mathbb{E}[\|f_2(x^t) - f_2(x^t, \xi_2^1)\|^2] \le \sigma_{f_2}^2.$$

So the $1/L_{f_1}$-strong convexity of $D_{f_1^*}$ and Young's inequality implies that

$$\mathbb{E}[\Delta_{\pi_1}^t - \tau_1^t D_{f_i^*}(\pi_1^t, \pi_1^{t+1})] \le \tfrac{1}{2\tau_1^t} L_{f_1}\sigma_{f_2}^2, \tag{2.26}$$

hence

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} \Delta_{\pi_1}^t] \le \sqrt{\frac{N}{2}}\sqrt{L_{f_1}}\mathcal{D}_{\Pi_1}\sigma_{f_2} + \mathbb{E}[\sum_{t=0}^{N-1} \tau_1^t D_{f_1^*}(\pi_1^t, \pi_1^{t+1})]. \tag{2.27}$$

Moreover the following bound can be derived from Lemma 2,

$$\begin{aligned}\mathbb{E}[\textstyle\sum_{t=0}^{N-1} \delta_1^t] &= \mathbb{E}\left[\sum_{t=0}^{N-1} -\pi_1^t[\mathcal{L}_2(x^t; \pi_2^{t+1}) - \mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1))]\right] + \mathbb{E}\left[\sum_{t=0}^{N-1} \pi_1[\mathcal{L}_2(x^t; \pi_2^{t+1}) - \mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1))]\right] \\ &\le \sqrt{N}M_{\Pi_1}\sigma_{f_2}.\end{aligned} \tag{2.28}$$

Thus substituting the above bounds on $\mathcal{B}_1^t$ in (2.24), $\Delta_{\pi_1}^t$ in (2.27), $\mathcal{A}_1^t$ in (2.22), and δ_1^t in (2.28) into the decomposed gap function Q_1 in (2.23), we obtain

$$\begin{aligned}&\mathbb{E}[\textstyle\sum_{t=0}^{N-1} Q_1(z^{t+1}, z)] \\ &\le \tau_1^0 D_{f_1^*}(\pi_1^0, \pi_1) + \sqrt{\frac{N}{2}}\sqrt{L_{f_1}}\mathcal{D}_{\Pi_1}\sigma_{f_2} + \sqrt{N}M_{\Pi_1}\sigma_{f_2} + \sqrt{2N}\tilde{M}_\Pi\mathcal{D}_X + \mathbb{E}[\textstyle\sum_{t=0}^{N-1} \frac{\sqrt{2N}\tilde{M}_\Pi}{2\mathcal{D}_X}\|x^t - x^{t+1}\|^2] \\ &\le \sqrt{N}M_{\Pi_1}\sigma_{f_2} + \sqrt{2N}\sqrt{L_{f_1}}\mathcal{D}_{\Pi_1}\sigma_{f_2} + \sqrt{2N}\tilde{M}_\Pi\mathcal{D}_X + \mathbb{E}[\textstyle\sum_{t=0}^{N-1} \frac{\sqrt{2N}\tilde{M}_\Pi}{2\mathcal{D}_X}\|x^t - x^{t+1}\|^2].\end{aligned}$$

■

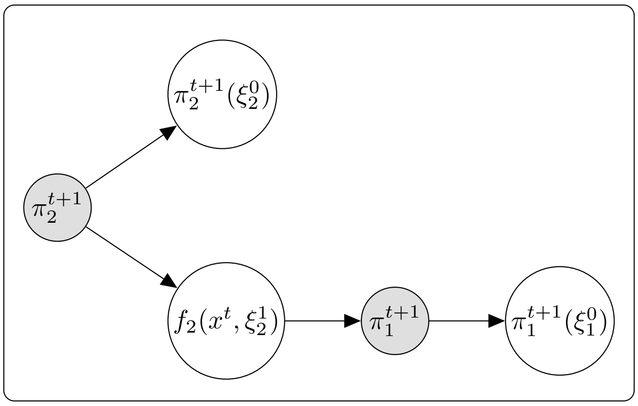

Fig. 1: Illustration of Stochastic Dependency: $\pi_2^{t+1}(\xi_2^0)$ is d-separated from $\pi_1^{t+1}(\xi_1^0)$ conditioned on $(\pi_1^{t+1}, \pi_2^{t+1})$ and $\pi_2^{t+1}(\xi_2^0)$ is d-separated from π_1^{t+1} conditioned on π_2^{t+1}.

Next, We turn our attention to the upper bound on the decomposed gap function Q_0. According to Lemma 1, the x-proximal update in Line 4 of Algorithm 1 satisfies

$$\begin{aligned}\mathcal{A}_0^t &:= \pi_1^{t+1}(\xi_1^0)\pi_2^{t+1}(\xi_2^0)(x^{t+1} - x^*) + u(x^{t+1}) - u(x^*) \\ &\leq \tfrac{\eta^t}{2}\|x^t - x^*\|^2 - \tfrac{\eta^t}{2}\|x^t - x^{t+1}\|^2 - \tfrac{\eta^t}{2}\|x^t - x^*\|^2.\end{aligned} \tag{2.29}$$

Observe that $\mathcal{A}_0^t$ differs from Q_0 only in stochastic error, i.e.,

$$Q_0(z^{t+1}, z) = \mathcal{A}_0^t + \Delta_0^t, \tag{2.30}$$

where $\Delta_0^t := (\pi_1^{t+1}\pi_2^{t+1} - \pi_1^{t+1}(\xi_1^0)\pi_2^{t+1}(\xi_2^0))(x^{t+1} - x^*)$.

Moreover, since $\pi_2^{t+1}(\xi_2^0)$ comes from a separate query to $\mathcal{DSO}_2$ ($\mathcal{SO}_2$), it is independent of $f_2(x^t, \xi_2^1)$ and hence its descendant $\pi_1^{t+1}(\xi_1^0)$ conditioned on π_2^{t+1} (see Figure 1). Therefore, the following expectation and variance of $\pi_1^{t+1}(\xi_1^0)\pi_2^{t+1}(\xi_2^0)$ can be derived:

$$\begin{aligned}\mathbb{E}[\pi_1^{t+1}(\xi_1^0)\pi_2^{t+1}(\xi_2^0)|\pi_1^{t+1}, \pi_2^{t+1}] &= \mathbb{E}[\pi_1^{t+1}(\xi_1^0)|\pi_1^{t+1}, \pi_2^{t+1}]\mathbb{E}[\pi_2^{t+1}(\xi_2^0)|\pi_1^{t+1}, \pi_2^{t+1}] \\ &= \pi_1^{t+1}\mathbb{E}[\pi_2^{t+1}(\xi_2^0)|\pi_1^{t+1}] \\ &= \pi_1^{t+1}\pi_2^{t+1}.\end{aligned}$$

$$\begin{aligned}&\mathbb{E}[\|\pi_1^{t+1}(\xi_1^0)\pi_2^{t+1}(\xi_2^0) - \pi_1^{t+1}\pi_2^{t+1}\|^2|\pi_1^{t+1}, \pi_2^{t+1}] \\ &\quad\leq \mathbb{E}[\|\pi_1^{t+1}(\xi_1^0) - \pi_1^{t+1}\|^2|\pi_1^{t+1}, \pi_2^{t+1}]\mathbb{E}[\|\pi_2^{t+1}(\xi_2^0)\|^2|\pi_1^{t+1}, \pi_2^{t+1}] + \mathbb{E}[\|\pi_1^{t+1}(\pi_2^{t+1}(\xi_2^0) - \pi_2^{t+1})\|^2|\pi_1^{t+1}, \pi_2^{t+1}] \\ &\quad\leq \sigma_{\pi_1}^2 M_{\Pi_2}^2 + M_{\Pi_1}^2\sigma_{\pi_2}^2 := \tilde{\sigma}_{\pi_{1:}}^2.\end{aligned}$$

So we can split the stochastic error term further:

$$\Delta_0^t = \Delta_{x,0}^t + \delta_0^t, \tag{2.31}$$

with

$$\begin{aligned}\Delta_{x,0}^t &:= (\pi_1^{t+1}\pi_2^{t+1} - \pi_1^{t+1}(\xi_1^0)\pi_2^{t+1}(\xi_2^0))(x^{t+1} - x^t), \\ \delta_0^t &:= (\pi_1^{t+1}\pi_2^{t+1} - \pi_1^{t+1}(\xi_1^0)\pi_2^{t+1}(\xi_2^0))(x^t - x^*).\end{aligned}$$

Notice that unbiasedness of $\pi_1^{t+1}(\xi_1^0)\pi_2^{t+1}(\xi_2^0)$ implies $\mathbb{E}[\delta_0^t] = 0, \forall t$.

Now we are ready to bound the decomposed gap function Q_0 and hence the overall gap function Q, and to establish the convergence of Algorithm 1.

Theorem 1 *Let* $\tilde{M}_{\Pi} := M_{\Pi_1} M_{\Pi_2}$ *and* $\tilde{\sigma}^2_{\pi_{1:}} := \sigma^2_{\pi_1} M^2_{\Pi_2} + M^2_{\Pi_1}\sigma^2_{\pi_2}$. *Let* $\{z^t\}$ *be generated by Algorithm 1 with* $\eta^t = \sqrt{2N}\tilde{M}_{\Pi}/\mathcal{D}_X + \sqrt{N}\tilde{\sigma}_{\pi_{1:}}/(\sqrt{2}Dx)$ *and* $\tau_1^t = \sqrt{N}\sqrt{L_{f_1}}\sigma_{f_2}/(\sqrt{2}\mathcal{D}_{\Pi_1})$. *Then*

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} Q(z^{t+1}, z)] \le \sqrt{2N}\tilde{\sigma}_{\pi_{1:}}\mathcal{D}_X + 4\sqrt{2N}\tilde{M}_{\Pi}\mathcal{D}_X + \sqrt{2N}\sqrt{L_{f_1}}\mathcal{D}_{\Pi_1}\sigma_{f_2} + \sqrt{N}M_{\Pi_1}\sigma_{f_2}.$$

Moreover, the ergodic solution $\bar{x}^N = \sum_{t=0}^{N-1} x^{t+1}/N$ *satisfies*

$$\mathbb{E}[f(\bar{x}^N) - f(x^*)] \le \tfrac{1}{\sqrt{N}}(\sqrt{2}\tilde{\sigma}_{\pi_{1:}} + 4\sqrt{2}\tilde{M}_{\Pi}\mathcal{D}_X + \sqrt{2}\sqrt{L_{f_1}}\mathcal{D}_{\Pi_1}\sigma_{f_2} + M_{\Pi_1}\sigma_{f_2}). \tag{2.32}$$

Proof Using (2.29), (2.30) and (2.31), and the fact

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1}(\Delta^t_{x,0} - \frac{\sqrt{N}\tilde{\sigma}_{\pi_{1:}}}{2\sqrt{2}\mathcal{D}_X}\|x^{t+1} - x^t\|^2)] \le \sqrt{\frac{N}{2}}\mathcal{D}_X\tilde{\sigma}_{\pi_{1:}},$$

we can obtain the following bound for Q_0:

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} Q_0(z^{t+1}, z)] \le \eta^0\frac{1}{2}\|x^0 - x^*\|^2 + \sqrt{\frac{N}{2}}\mathcal{D}_X\tilde{\sigma}_{\pi_{1:}} - \sum_{t=0}^{N-1}\frac{\sqrt{N}\tilde{\sigma}_{\pi_{1:}}}{2\sqrt{2}\mathcal{D}_X}\|x^t - x^{t+1}\|^2. \tag{2.33}$$

Then combining (2.33) with the bounds for Q_2 and Q_1 in Proposition 5 and 6, we get

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} Q(z^{t+1}, z)] \le \sqrt{2N}\tilde{\sigma}_{\pi_{1:}}\mathcal{D}_X + 4\sqrt{2N}\tilde{M}_{\Pi}\mathcal{D}_X + \sqrt{2N}\sqrt{L_{f_1}}\mathcal{D}_{\Pi_1}\sigma_{f_2} + \sqrt{N}M_{\Pi_1}\sigma_{f_2}.$$

Therefore, by setting the reference point z in the above inequality to $\hat{z} := (x^*; \hat{\pi}_1, \hat{\pi}_2)$, (2.32) follows immediately from Proposition 2. ∎

2.5 Convergence Analysis for Strongly Convex Problems

In this subsection, we establish the convergence of Algorithm 1 for solving two-layer NSCO problems with strong convexity modulus $\alpha > 0$. We set the weight w^t to be $(t+1)/2$ for defining the ergodic output solution, and focus on reference points of the form $z = (x^*; \pi_1^*, \pi_2)$, where π_2 can depend on the stochastic iterates $\{z^t\}$. Other than increasing η^ts for the x-updates and using Proposition 3 for functional optimality gap conversion, the analysis is similar to the preceding subsection; we first develop upper bounds on gap functions Q_2, Q_1, Q_0, and hence Q, and then show the convergence in function value.

First, we develop a convergence bound for the decomposed gap function Q_2 in the following proposition. Notice that the coefficients for $\|x^t - x^{t+1}\|^2$ is increasing proportionally to $\alpha(t+1)$ in the bound.

Proposition 7 *Let* $\{z^t\}$ *be generated by Algorithm 1 and let* $\tilde{M}_{\Pi} = M_{\pi_1}M_{\pi_2}$. *Then*

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q_2(z^{t+1}, z)] \le \frac{9N}{\alpha}\tilde{M}^2_{\Pi} + \mathbb{E}[\sum_{t=0}^{N-1}\frac{(t+1)\alpha w^t}{18}\|x^t - x^{t+1}\|^2].$$

Proof It is clear that the decomposition of Q_2 in (2.19) still holds:

$$Q_2(z^{t+1}, z) = \mathcal{A}_2^t + \mathcal{B}_2^t, \tag{2.34}$$

where $\mathcal{B}_2^t := \pi_1^*(\pi_2 - \pi_2^{t+1})(x^{t+1} - x^t)$ and $\mathcal{A}_2^t := \pi_1^*((\pi_2 - \pi_2^{t+1})x^t - (f_2^*(\pi_2) - f_2^*(\pi_2^{t+1}))) \le 0$. By Young's inequality, we have

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t(\mathcal{B}_2^t - \frac{(t+1)\alpha w^t}{18}\|x^{t+1} - x^t\|^2)] \le \sum_{t=0}^{N-1} w^t\frac{4\tilde{M}^2_{\Pi}}{4}\frac{18}{(t+1)\alpha} = \frac{9N}{\alpha}\tilde{M}^2_{\Pi}.$$

So we can obtain the desired bound by summing up (2.34):

$$\begin{aligned}
&\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q_2(z^{t+1}, z)] \\
&\le \mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t\mathcal{A}_2^t] + \mathbb{E}[\sum_{t=0}^{N-1} w^t(\mathcal{B}_2^t - \frac{(t+1)\alpha}{18}\|x^{t+1} - x^t\|^2))] + \mathbb{E}[\frac{(t+1)\alpha w^t}{18}\|x^{t+1} - x^t\|^2)] \\
&\le \tfrac{9N}{\alpha}\tilde{M}^2_{\Pi} + \mathbb{E}[\textstyle\sum_{t=0}^{N-1}\frac{(t+1)\alpha w^t}{18}\|x^{t+1} - x^t\|^2].
\end{aligned}$$

∎

Next, we show a convergence bound for the decomposed gap function Q_1. Compared to Proposition 6, the more restricted reference point $z = (x^*; \pi_1^*, \pi_2)$ in this subsection allows for an improved bound.

Proposition 8 *Let $\{z^t\}$ be generated by Algorithm 1 with $\tau_1^t := \frac{t+1}{3}, \forall t$, $\tilde{M}_{\Pi} := M_{\Pi_1} M_{\Pi_2}$ and $\mathcal{D}^2_{\Pi_1} := D_{f_1^*}(\pi_1^0, \pi_i^*)$. Then*

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q_1(z^{t+1}, z)] \leq \frac{3N}{4} L_{f_1} \sigma_{f_2}^2 + \frac{1}{6}\mathcal{D}^2_{\Pi_1} + \frac{9N}{\alpha}\tilde{M}^2_{\Pi} + \mathbb{E}[\textstyle\sum_{t=0}^{N-1} \frac{(t+1)\alpha w^t}{18} \|x^t - x^{t+1}\|^2].$$

Proof The decomposition formulas in (2.23) and (2.25) still hold:

$$Q_1(z^{t+1}, z) = \mathcal{A}_1^t + \mathcal{B}_1^t + \delta_1^t + \Delta_{\pi_1}^t, \tag{2.35}$$

where

$$\begin{aligned}
\mathcal{A}_1^t &:= (\pi_1^* - \pi_1^{t+1})\mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1)) - (f_1^*(\pi_1^*) - f_1^*(\pi_1^{t+1})),\\
\Delta_{\pi_1}^t &:= (\pi_1^t - \pi_1^{t+1})(\mathcal{L}_2(x^t; \pi_2^{t+1}) - \mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1))),\\
\delta_1^t &:= (\pi_1^* - \pi_1^t)(\mathcal{L}_2(x^t; \pi_2^{t+1}) - \mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1))),\\
\mathcal{B}_1^t &:= (\pi_1^* - \pi_1^{t+1})\pi_2^{t+1}(x^{t+1} - x^t).
\end{aligned}$$

Since π_1^* is independent of the stochastic iterates $\{z^t\}$ and $\mathbb{E}[\mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_2^1))|\pi_2^{t+1}] = \mathbb{E}[f_2(x^t, \xi_2^j)|\pi_2^{t+1}] = f_2(x^t) = \mathcal{L}_2(x^t; \pi_2^{t+1})$, we have $\mathbb{E}[\delta_1^t] = 0, \forall t$. Moreover, Line 3 of Algorithm 1 implies that

$$\mathcal{A}_1^t \leq \tau_1^t D_{f_1^*}(\pi_1^t, \pi_1) - \tau_1^t D_{f_1^*}(\pi_1^t, \pi_1^{t+1}) - (\tau_1^t + 1) D_{f_1^*}(\pi_1^{t+1}, \pi_1). \tag{2.36}$$

Thus with $\tau_1^t = \frac{t+1}{3}$, the terms $D_{f_1^*}(\pi_1^t, \pi_1^*)$ and $D_{f_1^*}(\pi_1^{t+1}, \pi_1^*)$ in the above inequality admits telescoping cancellation through a w^t-weighted sum,

$$\begin{aligned}
\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t \mathcal{A}_1^t] &\leq w^0 \tau_1^0 D_{f_1^*}(\pi_1^0, \pi_i^*) - w^{N-1}\tau_1^{N-1} D_{f_1^*}(\pi_1^N, \pi_1^*) - \textstyle\sum_{t=0}^{N-1} w^t \tau_1^t D_{f_1^*}(\pi_1^t, \pi_1^{t+1})\\
&\leq \frac{1}{6}\mathcal{D}^2_{\Pi_1} - \textstyle\sum_{t=0}^{N-1} w^t \tau_1^t D_{f_1^*}(\pi_1^t, \pi_1^{t+1}).
\end{aligned} \tag{2.37}$$

The $1/L_{f_1}$-strong convexity of $D_{f_i^*}$ implies that

$$\begin{aligned}
\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t (\Delta_{\pi_1}^t - \tau_1^t D_{f_1^*}(\pi_1^t, \pi_1^{t+1}))] &\leq \mathbb{E}\left[\textstyle\sum_{t=0}^{N-1} w^t \frac{3}{t+1}\frac{L_{f_1}}{2} \|\mathcal{L}_2(x^t; \pi_2^{t+1}) - \mathcal{L}(x^t; \pi_2^{t+1}(\xi_2^1))\|^2\right]\\
&\leq \textstyle\sum_{t=0}^{N-1} w^t \frac{3}{t+1}\frac{L_{f_1}}{2}\sigma_{f_2}^2 = \frac{3}{4} N L_{f_1} \sigma_{f_2}^2.
\end{aligned} \tag{2.38}$$

Similar to Proposition 7, the following bound of the weighted sum of $\mathcal{B}_1^t$ holds,

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t \mathcal{B}_1^t] \leq \frac{9N}{\alpha}\tilde{M}^2_{\Pi} + \mathbb{E}[\textstyle\sum_{t=0}^{N-1} \frac{(t+1)\alpha w^t}{18} \|x^t - x^{t+1}\|^2]. \tag{2.39}$$

Thus in view of the Q_1 decomposition (2.35), substituting (2.38), (2.37) and (2.39) into $\mathbb{E}[\sum_{t=0}^{N-1} w^t Q_1(z^{t+1}, z)]$ leads to the desired bound.

■

Now we move on to bound Q_0. Obviously, the decomposition formulas (2.30) and (2.31) from the last subsection are still valid:

$$Q_0(z^{t+1}, z) = \mathcal{A}_0^t + \Delta_{x,0}^t + \delta_0^t, \tag{2.40}$$

where

$$\begin{aligned}
\mathcal{A}_0^t &:= \pi_1^{t+1}(\xi_1^0)\pi_2^{t+1}(\xi_2^0)(x^{t+1} - x^*) + u(x^{t+1}) - u(x^*),\\
\Delta_{x,0}^t &:= (\pi_1^{t+1}\pi_2^{t+1} - \pi_1^{t+1}(\xi_1^0)\pi_2^{t+1}(\xi_2^0))(x^{t+1} - x^t),\\
\delta_0^t &:= (\pi_1^{t+1}\pi_2^{t+1} - \pi_1^{t+1}(\xi_1^0)\pi_2^{t+1}(\xi_2^0))(x^t - x^*).
\end{aligned}$$

Moreover, since the stochastic estimators $\pi_1^{t+1}(\xi_1^0)$ and $\pi_2^{t+1}(\xi_2^0)$ are generated in the same way as the last subsection, we conclude

$$\begin{aligned}
&\mathbb{E}[\pi_1^{t+1}(\xi_1^0)\pi_2^{t+1}(\xi_2^0)|\pi_1^{t+1}, \pi_2^{t+1}] = \pi_1^{t+1}\pi_2^{t+1},\\
&\mathbb{E}[\|\pi_1^{t+1}(\xi_1^0)\pi_2^{t+1}(\xi_2^0) - \pi_1^{t+1}\pi_2^{t+1}\|^2|\pi_1^{t+1}, \pi_2^{t+1}] \leq \tilde{\sigma}^2_{\pi_{1:}} := \sigma^2_{\pi_1} M^2_{\Pi_2} + M^2_{\Pi_1}\sigma^2_{\pi_2},
\end{aligned}$$

and

$$\mathbb{E}[\delta_0^t] = 0 \ \forall t.$$

So the following bound on Q_0 convergence follows from setting $\eta^t := (t+1)\alpha/3$.

Proposition 9 *Let $\{z^t\}$ be generated by Algorithm 1 with $\eta^t = \frac{t+1}{3}\alpha$, then*

$$\begin{aligned}\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q_0(z^{t+1}, z)] + \frac{N(N+3)\alpha}{12}\mathbb{E}[\|x^N - x^*\|^2] \\ \leq \frac{\alpha}{12}\|x^0 - x^*\|^2 + \frac{9N}{4\alpha}\tilde{\sigma}^2_{\pi_{1:}} - \mathbb{E}\left[\textstyle\sum_{t=0}^{N-1} \frac{(t+1)\alpha w^t}{9}\|x^{t+1} - x^t\|^2\right].\end{aligned} \tag{2.41}$$

Proof The strong convexity modulus of $u(x)$, i.e. $\alpha > 0$, and Lemma 1 imply that

$$\mathcal{A}_0^t \leq \eta^t \tfrac{1}{2}\|x^t - x^*\|^2 - \eta^t \tfrac{1}{2}\|x^t - x^{t+1}\|^2 - (\eta^t + \alpha)\tfrac{1}{2}\|x^{t+1} - x^*\|^2.$$

Then the w^t-weighted sum satisfies

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t \mathcal{A}_0^t] \leq \frac{\alpha}{12}\|x^0 - x^*\|^2 - \frac{N(N+3)\alpha}{12}\mathbb{E}[\|x^N - x^*\|^2] - \mathbb{E}[\textstyle\sum_{t=0}^{N-1} \frac{(t+1)\alpha w^t}{6}\|x^{t+1} - x^t\|^2]. \tag{2.42}$$

Moreover a bound for $\Delta_{x,0}^t$ is

$$\begin{aligned}\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t(\Delta_{x,0}^t - \frac{(t+1)\alpha}{18}\|x^t - x^{t+1}\|^2)] &\leq \mathbb{E}[\textstyle\sum_{t=0}^{N-1} \frac{9w^t}{2(t+1)\alpha}\|\pi_1^{t+1}\pi_2^{t+1} - \pi_1^{t+1}(\xi_1^0)\pi_2^{t+1}(\xi_2^0)\|^2] \\ &\leq \frac{9N}{4\alpha}\tilde{\sigma}^2_{\pi_{1:}}.\end{aligned} \tag{2.43}$$

Therefore, in view of the decomposition of Q_0 in (2.40), we can conclude (2.41) by substituting the bound on $\mathcal{A}_0^t$ in (2.42) and the bound on $\Delta_{x,0}^t$ in (2.43) into $\mathbb{E}[\sum_{t=0}^{N-1} w^t Q_0(z^{t+1}, z)]$. ■

Now we put all the pieces together to establish the convergence of Algorithm 1 for the strongly convex NSCO problem.

Theorem 2 *Let $\{z^t\}$ be generated by Algorithm 1 with $\eta^t = \frac{(t+1)\alpha}{3}$ and $\tau_1^t = \frac{t+1}{3}$. Let $\tilde{M}_\Pi := M_{\Pi_1}M_{\Pi_2}$ and $\tilde{\sigma}^2_{\pi_{1:}} := \sigma^2_{\pi_1}M^2_{\Pi_2} + M^2_{\Pi_1}\sigma^2_{\pi_2}$. Then we have*

$$\begin{aligned}&\mathbb{E}[\tfrac{1}{2}\|x^N - x^*\|^2] \leq \frac{3}{N\alpha}C_1 + \frac{3}{N(N+1)\alpha}C_2, \\ &\mathbb{E}[f(\bar{x}^N) - f(x^*)] \leq \frac{1}{N}(2 + \frac{6}{\alpha}L_{f_1}M^2_{\Pi_2})C_1 + \frac{1}{N(N+1)}(2 + \frac{6\log(N+1)}{\alpha}L_{f_1}M^2_{\Pi_2})C_2,\end{aligned}$$

where $C_1 := \frac{36}{\alpha}\tilde{M}^2_\Pi + \frac{3}{2}L_{f_1}\sigma^2_{f_2} + \frac{9}{2\alpha}\tilde{\sigma}^2_{\pi_{1:}}$ and $C_2 := \frac{1}{3}\mathcal{D}^2_{\Pi_1} + \frac{\alpha}{6}\|x^0 - x^*\|^2$.

Proof Adding up the bounds on the decomposed gap functions Q_2, Q_1 and Q_0 bounds from Proposition 7, 8 and 9, we obtain

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q(z^{t+1}, z)] + \frac{N(N+3)\alpha}{6}\mathbb{E}[\tfrac{1}{2}\|x^N - x^*\|^2] \leq \frac{N}{2}C_1 + \frac{1}{2}C_2. \tag{2.44}$$

Then choosing the reference point z to be $z^* := (x^*; \pi_1^*, \pi_2^*)$, noticing that $Q(z^{t+1}, z^*) \geq 0 \; \forall t$ and dividing both sides by $\frac{N(N+3)\alpha}{6}$, we get

$$\mathbb{E}[\tfrac{1}{2}\|x^N - x^*\|^2] \leq \frac{3}{(N+3)\alpha}C_1 + \frac{3}{N(N+3)\alpha}C_2 \leq \frac{3}{N\alpha}C_1 + \frac{3}{N(N+1)\alpha}C_2.$$

Moreover, since $\bar{x}^N$ is the $w^t/\sum_{t=0}^{N-1} w^t$ weighted sum of x^{t+1}, we conclude from the Jensen's inequality that

$$\mathbb{E}[\tfrac{1}{2}\|\bar{x}^N - x^*\|^2] \leq \frac{6}{N\alpha}C_1 + \frac{6\log(N+1)}{N(N+1)\alpha}C_2.$$

Next, to show convergence in function value, we pick the reference point z in (2.44) to be $\tilde{z} := (x^*; \pi_1^*, \hat{\pi}_2)$ such that $\mathbb{E}[\sum_{t=0}^{N-1} w^t Q(z^{t+1}, \tilde{z})] \leq \frac{N}{2}C_1 + \frac{1}{2}C_2$. Then in view of Proposition 3, we have

$$\begin{aligned}\mathbb{E}[f(\bar{x}^N) - f(x^*)] &\leq \mathbb{E}\left[\frac{\sum_{t=0}^{N-1} w^t Q(z^{t+1}, \tilde{z})}{\sum_{t=0}^{N-1} w^t} + \tfrac{1}{2}L_{f_1}M^2_{\Pi_2}\|\bar{x}^N - x^*\|^2\right] \\ &\leq \frac{1}{N}(2 + \frac{6}{\alpha}L_{f_1}M^2_{\Pi_2})C_1 + \frac{1}{N(N+1)}(2 + \frac{6\log(N+1)}{\alpha}L_{f_1}M^2_{\Pi_2})C_2.\end{aligned}$$

■

We conclude this section by raising two issues about the vanilla SSD Algorithm. Firstly, Algorithm 1 is limited in its application. It can only solve the smooth-nonsmooth two-layer NSCO problem. Secondly, Algorithm 1's iteration complexity is optimal only for ϵ, but not other problem parameters like L_{f_1}. Over the next few sections, we will address these issues by developing a nearly optimal modular algorithmic framework for solving general multilayer NSCO problems.

3 SSD Framework

In this section, we propose a general Stochastic Sequential Dual (SSD) algorithmic framework to solve the multilayer NSCO problem (1.3), i.e., $\min_{x\in X}\{f(x) := f_1 \circ f_2 \circ \ldots \circ f_k(x) + u(x)\}$. Focusing on the general algorithmic framework in this section, we impose minimal structural assumptions on the layer functions. More specifically, we assume that every f_i is convex and M_{Π_i}[⑥]-Lipschitz continuous, equipped with a $\mathcal{DSO}_i$ for the dual iterates generated by some dual proximal updates (c.f. Subsection 2.2). The finer details about f_i, such as their smoothness properties (smooth, smoothable or non-smooth), the exact dual proximal updates, and the restrictions on composition, will be presented in later sections.

3.1 Saddle Point Reformulation

As illustrated in Section 1, Assumptions 1 and 2 allow us to reformulate problem (1.3) as

$$\min_{x\in X}\max_{\pi_{1:}\in\Pi_{1:}}\{\mathcal{L}(x;\pi_{1:}) := \mathcal{L}_1(x;\pi_{1:}) + u(x)\}. \tag{3.1}$$

where $\pi_{1:}$ and $\Pi_{1:}$ are the shorthand notations for $\pi_{1:k}$ and $\Pi_{1:k}$[⑦], and

$$\mathcal{L}_i(x;\pi_{i:}) := \begin{cases} x & \text{if } i = k+1, \\ \pi_i \mathcal{L}_{i+1}(x;\pi_{i+1:}) - f_i^*(\pi_i) & \text{if } 0 \le i \le k. \end{cases} \tag{3.2}$$

Then it is straightforward to extend Proposition 1 to describe a basic duality relationship between (1.3) and (3.1) as in the following proposition.

Proposition 10 *Let f and $\mathcal{L}$ be defined in* (1.3) *and* (3.1)*, respectively. Then the following relations hold for all $x \in X$.*

a) Weak Duality: $f(x) \ge \mathcal{L}(x;\pi_{1:})\ \forall \pi_{1:} \in \Pi_{1:}$.
b) Strong Duality: $f(x) = \mathcal{L}(x;\pi^*_{1:})$, for some $\pi^*_{1:} \in \Pi_{1:}$.

Proof The proof is similar to Proposition 1. Let $x \in X$ and $\pi_{1:} \in \Pi_{1:}$ be given. The strong duality result follows from choosing $\pi_i^* \in \partial f_i(f_{i+1:}(x))$ such that

$$\mathcal{L}_i(x;\pi_i^*,\pi_{i+1:}^*) = f_{i:}(x) \text{ for } i = k, k-1, \ldots, 1.$$

For the weak duality, first note that for any feasible $(\pi_1, \pi_2, \ldots, \pi_k)$, we have

$$\begin{aligned} f(x) - \mathcal{L}(x;\pi_{1:}) &= \textstyle\sum_{j=1}^k [\mathcal{L}(x;\pi_{1:j-1},\pi_j^*,\pi_{j+1:}^*) - \mathcal{L}(x;\pi_{1:j-1},\pi_j,\pi_{j+1:}^*)] \\ &= \textstyle\sum_{j=1}^k \underbrace{\pi_{1:j-1}[(\pi_j^* f_{j+1:}(x) - f_j^*(\pi_j^*)) - (\pi_j f_{j+1:}(x) - f_j^*(\pi_j))]}_{A_j}. \end{aligned}$$

Moreover, $A_j = 0$ if f_j is linear since Π_j is a set of singleton and $\pi_j^* = \pi_j$. If f_j is not linear, we have $A_j \ge 0$ since π_j^* is the maximizer of $\pi_j f_{j+1:}(x) - f_j^*(\pi_j)$ and $\pi_{1:j-1}$ is non-negative. Thus we always have $f(x) \ge \mathcal{L}(x;\pi_{1:})$. ■

Accordingly, for a feasible $\bar{z} := (\bar{x};\bar{\pi}_{1:})$, we define the multilayer *gap function* with respect to some reference point $z := (x;\pi_{1:})$ as

$$Q(\bar{z}, z) := \mathcal{L}(\bar{x};\pi_{1:}) - \mathcal{L}(x;\bar{\pi}_{1:}). \tag{3.3}$$

For an ergodic average solution $\bar{x}^N$, the value of the gap function with respect to a few reference points provides an upper bound for the functional optimality gap of problem (1.3). Such reference points are usually constructed from the following dual points:

$$\pi_i^* \in \partial f_i(f_{i+1:}(x^*)) \text{ and } \hat{\pi}_i \in \partial f_i(f_{i+1:}(\bar{x}^N)).$$

For example, the next proposition shows an upper bound derived by using the reference point $\hat{z} := (x^*;\hat{\pi}_{1:})$.

[⑥] M_{Π_i} is defined in Subsection 1.3.
[⑦] We use $\Pi_{i:j}$ to represent $\Pi_i \times \Pi_{i+1} \times \ldots \times \Pi_j$, and $\pi_{i:j}$ to represent either $(\pi_i, \pi_{i+1}, \ldots, \pi_j)$ or $\pi_i\pi_{i+1}\ldots\pi_j$ depending on the context.

Proposition 11 *Let feasible solutions* $\{z^t := (x^t; \pi^t_{1:})\}$ *be given and let* $\bar{x}^N$ *denote the* w^t*-weighted ergodic average of* $\{x^t\}$*, given by* $\sum_{t=0}^{N-1} w^t x^{t+1} / \sum_{t=0}^{N-1} w^t$*. If* $\hat{z} := (x^*; \hat{\pi}_{1:})$ *and* $\sum_{t=0}^{N-1} w^t Q(z^{t+1}, \hat{z}) \le B$*, then*

$$f(\bar{x}^N) - f(x^*) \le \frac{B}{\sum_{t=0}^{N-1} w^t}.$$

Proof The proof is similar to that of Proposition 2. The only difference exists in that we need to use the multilayer duality result, i.e., Proposition 10 instead of Proposition 2. ■

In addition, just like Proposition 3, when the outer layer functions $\{f_1, f_2, \dots, f_{i-1}\}$ are all smooth, another upper bound can be derived using both Q and $\frac{1}{2}\|\bar{x}^N - x^*\|^2$. This alternative bound will help us improve the convergence rate for solving strongly convex NSCO problems.

Proposition 12 *Let* $f_1, \dots, f_{i-1}$ *be Lipschitz-smooth with constants* $L_{f_1}, L_{f_2}, \dots, L_{f_{i-1}}$*. Let feasible solutions* $\{z^t := (x^t; \pi^t_{1:})\}$ *be given and let* $\bar{x}^N$ *denote the* w^t*-weighted ergodic average of* $\{x^t\}$*, given by* $\sum_{t=0}^{N-1} w^t x^{t+1} / \sum_{t=0}^{N-1} w^t$*. If* $\tilde{z} := (x^*; \pi^*_{1:i-1}, \hat{\pi}_{i:})$*,* $\sum_{t=0}^{N-1} w^t Q(z^{t+1}, \tilde{z}) \le B$ *and* $\frac{1}{2}\|\bar{x}^N - x^*\|^2 \le C$*, then*

$$f(\bar{x}^N) - f(x^*) \le \frac{B}{\sum_{t=0}^{N-1} w^t} + \tilde{L}_f C,$$

where $\tilde{L}_f := \sum_{l=1}^{i-1} M_{\Pi_{1:l-1}} L_{f_l} M^2_{\Pi_{l+1:}}$*.*

Proof Clearly, by using the argument in Proposition 2 with the multi-layer duality properties in Proposition 10, we have

$$\mathcal{L}(\bar{x}^N; \pi^*_{1:i-1}, \hat{\pi}_{i:}) - f(x^*) \le \frac{1}{\sum_{t=0}^{N-1} w^t} \sum_{t=0}^{N-1} w^t Q(z^{t+1}, \tilde{z}).$$

But the difference between $f(\bar{x}^N)$ and $\mathcal{L}(\bar{x}^N; \pi^*_{1:i-1}, \hat{\pi}_{i:})$ satisfies

$$\begin{aligned}
f(\bar{x}^N) - \mathcal{L}(\bar{x}^N; \pi^*_{1:i-1}, \hat{\pi}_{i:}) &= \mathcal{L}(\bar{x}^N; \hat{\pi}_{1:i-1}, \hat{\pi}_{i:}) - \mathcal{L}(\bar{x}^N; \pi^*_{1:i-1}, \hat{\pi}_{i:}) \\
&= \textstyle\sum_{l=1}^{i-1} [\mathcal{L}(\bar{x}^N; \pi^*_{1:l-1}, \hat{\pi}_l, \hat{\pi}_{l+1:}) - \mathcal{L}(\bar{x}^N; \pi^*_{1:l-1}, \pi^*_l, \hat{\pi}_{l+1:})] \\
&= \textstyle\sum_{l=1}^{i-1} \pi^*_{1:l-1} [f^*_l(\pi^*_l) - f^*_l(\hat{\pi}_l) - (\pi^*_l - \hat{\pi}_l) f_{l+1:}(\bar{x}^N)] \\
&\overset{(a)}{=} \textstyle\sum_{l=1}^{i-1} \pi^*_{1:l-1} D_{f^*_l}(\hat{\pi}_l, \pi^*_l) \\
&\overset{(b)}{=} \textstyle\sum_{l=1}^{i-1} \pi^*_{1:l-1} D_{f_l}(f_{l+1:}(x^*), f_{l+1:}(\bar{x}^N)) \le \frac{1}{2}\tilde{L}_f \|\bar{x}^N - x^*\|^2 = \tilde{L}_f C,
\end{aligned}$$

where (a) follows from $f_{l+1:}(\bar{x}^N) \in \partial f^*_l(\hat{\pi}_l)$ (the definition of $\hat{\pi}_l$) and (b) follows from the relationship between conjugate Bregman distances (1.8). So we get the desired bound for $f(\bar{x}^N) - f(x^*)$. ■

Moreover, we can utilize stochastic estimators $\{\pi_i(\xi_i), f^*_i(\pi_i, \xi_i)\}$ returned from $\{\mathcal{DSO}_i\}$ to construct the *stochastic composition Lagrangians* for multilayer NSCO as follows:

$$\mathcal{L}_i(x; \pi_{i:}(\xi_{i:})) := \begin{cases} x & \text{if } i = k+1, \\ \pi_i(\xi_i)\mathcal{L}_{i+1}(x; \pi_{i+1:}(\xi_{i+1:})) - f^*_i(\pi_i, \xi_i) & \text{if } 0 \le i \le k, \end{cases} \tag{3.4}$$

where $\pi_{i:}(\xi_{i:})$ is the shorthand for $(\pi_i(\xi_i), \pi_{i+1}(\xi_{i+1}), \dots, \pi_k(\xi_k))$, and use them as arguments for the dual proximal updates. This leads us to the SSD framework.

3.2 SSD Framework

The SSD framework is again inspired by a decomposition of the gap function in (3.3) into

$$Q(z^{t+1}, z) = Q_0(z^{t+1}, z) + \textstyle\sum_{i=1}^k Q_i(z^{t+1}, z), \tag{3.5}$$

where Q_0 measures the optimality of x^{t+1} and Q_i measures the optimality of π^{t+1}_i:

$$\begin{aligned}
Q_0(z^{t+1}, z) &:= \mathcal{L}(x^{t+1}; \pi^{t+1}_{1:}) - \mathcal{L}(x; \pi^{t+1}_{1:}) \\
&= \boxed{\pi^{t+1}_{1:} x^{t+1} + u(x^{t+1})} - \pi^{t+1}_{1:} x - u(x),
\end{aligned} \tag{3.6}$$

$$
\begin{aligned}
Q_i(z^{t+1}, z) &:= \mathcal{L}(x^{t+1}; \pi_{1:i-1}, \pi_i, \pi_{i+1:}^{t+1}) - \mathcal{L}(x^{t+1}; \pi_{1:i-1}, \pi_i^{t+1}, \pi_{i+1:}^{t+1}) \\
&= \pi_{1:i-1}\left(\pi_i \mathcal{L}_{i+1}(x^{t+1}; \pi_{i+1:}^{t+1}) - f_i^*(\pi_i) \boxed{-[\pi_i^{t+1}\mathcal{L}_{i+1}(x^{t+1}; \pi_{i+1:}^{t+1}) - f_i^*(\pi_i^{t+1})]}\right).
\end{aligned}
\tag{3.7}
$$

To decrease the Q function, we can find $\pi_k^{t+1}, \pi_{k-1}^{t+1}, \ldots, \pi_1^{t+1}$ and x^{t+1} to reduce the boxed items in $Q_k, Q_{k-1}, \ldots, Q_0$, respectively. Accordingly, each iteration of the SSD framework (Algorithm 3) performs sequential proximal updates to the dual variables $\pi_k, \ldots, \pi_1$ before updating x. Notice that new π_i proximal update is a natural generalization of the π_2 and π_1 updates in Algorithm 1. Similar to the π_2 update, finding a π_i^{t+1} with a small value of $-\pi_{1:i-1}[\pi_i^{t+1}\mathcal{L}_{i+1}(x^{t+1}; \pi_{i+1:}^{t+1}) - f_i^*(\pi_i^{t+1})]$ is equivalent to finding one with component-wise small values, $-[\pi_i^{t+1}\mathcal{L}_{i+1}(x^{t+1}; \pi_{i+1:}^{t+1}) - f_i^*(\pi_i^{t+1})]$, due to the non-negativity of $\pi_{1:i-1}$ and the row-separability of proximal penalty V_i[8] and domain Π_i. In addition, similar to the π_1 update in Algorithm 1, the argument $\mathcal{L}_{i+1}(x^{t+1}; \pi_{i+1:}^{t+1})$ is not available because x^{t+1} is yet to be computed and only a stochastic estimator $\mathcal{L}_{i+1}(\cdot; \pi_{i+1:}^{t+1}(\xi_{i+1:}^i))$ is accessible. So we employ an momentum-based *guess*, i.e.,

$$
\bar{y}_i^{t+1} := \mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^i)) + \theta^t \pi_{i+1:}^t(\xi_{i+1:}^i)(x^t - x^{t-1}) \tag{3.8}
$$

and define the proximal update to generate π_i^{t+1} by

$$
\pi_i^{t+1} \in \operatorname*{arg\,min}_{\pi_i \in \Pi_i} -\pi_i \bar{y}_i^{t+1} + f_i^*(\pi_i) + \tau_i^t V_i(\pi_i^t, \pi_i).
$$

Next, the new x-proximal update in Line 9 of Algorithm 3 is a straightforward extension of the Line 4 in Algorithm 1, obtained by replacing $\pi_1^{t+1}(\xi_1^0)\pi_2^{t+1}(\xi_2^0)$ with its multilayer counterpart $\pi_{1:}^{t+1}(\xi_{1:}^0)$.

Moreover, if all the layer functions are either smooth or non-smooth, we can rewrite Algorithm 3 in a purely primal form, shown in Algorithm 4.

Algorithm 3 Stochastic Sequential Dual(SSD) Framework

Input: $x_{-1} = x_0 \in X$, feasible $\{\pi_i^0\}_{i=1}^k$.
1: **for** $i = k, k-1, \ldots, 1$ **do**
2: Call $\mathcal{DSO}_i$ i times at π_i^0 to obtain estimates $\{[\pi_i^0(\xi_i^j), f_i^*(\pi_i^0, \xi_i^j)]\}_{j=0}^{i-1}$.
3: **end for**
4: **for** $t = 0, 1, 2, 3 \ldots N-1$ **do**
5: **for** $i = k, k-1, \ldots, 1$ **do**
6: Set $\bar{y}_i^{t+1} := \mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^i)) + \theta^t \pi_{i+1:}^t(\xi_{i+1:}^i)(x^t - x^{t-1})$.
7: Call $\mathcal{DSO}_i$ i times at π_i^{t+1} to obtain estimates $\{[\pi_i^{t+1}(\xi_i^j), f_i^*(\pi_i^{t+1}, \xi_i^j)]\}_{j=0}^{i-1}$, where π_i^{t+1} comes from the (implicit) proximal update, $\pi_i^{t+1} \in \arg\min_{\pi_i \in \Pi_i} -\pi_i \bar{y}_i^{t+1} + f_i^*(\pi_i) + \tau_i^t V_i(\pi_i^t, \pi_i)$.
8: **end for**
9: Set $x^{t+1} := \arg\min_{x \in X} \pi_{1:}^{t+1}(\xi_{1:}^0)x + u(x) + \eta^t \frac{1}{2}\|x - x^t\|^2$.
10: **end for**
11: Return $\bar{x}^N := \sum_{t=0}^{N-1} w^t x^{t+1} / \sum_{t=0}^{N-1} w^t$.

Algorithm 4 Primal Form of the SSD Framework

Input: $\underline{y}_1^0 = x_{-1} = x_0 \in X$, $\{\underline{y}_i^0\}_{i=1}^k$.
for $i = k, k-1, \ldots, 1$ **do**
Call $\mathcal{SO}_i$ i times at $\underline{y}_i^0$ to obtain estimates $\{[f_i'(\underline{y}_i^0, \xi_i^j), f_i^*(f_i'(\underline{y}_i^0), \xi_i^j)]\}_{j=0}^{i-1}$.
end for
for $t = 0, 1, 2, 3 \ldots N-1$ **do**
for $i = k, k-1, \ldots, 1$ **do**
Set $\bar{y}_i^{t+1} := \mathcal{L}_{i+1}(x^t; f_{i:}'(\underline{y}_{i+1:}^{t+1}, \xi_{i+1:}^i)) + \theta^t f_{i+1}'(\underline{y}_{i+1}^{t+1}, \xi_{i+1}^i) \ldots f_k'(\underline{y}_k^{t+1}, \xi_k^i)(x^t - x^{t-1})$
and $\underline{y}_i^{t+1} := (\bar{y}_i^{t+1} + \tau_i^t \underline{y}_i^t)/(1 + \tau_i^t)$.
Call $\mathcal{SO}_i$ i times at $\underline{y}_i^{t+1}$ to obtain estimates $[f_i'(\underline{y}_i^{t+1}, \xi_i^j), f_i^*(f_i'(\underline{y}_i^{t+1}), \xi_i^j)]\}_{j=0}^{i-1}$.
end for
Set $x^{t+1} := \arg\min_{x \in X} f_1'(\underline{y}_1^{t+1}, \xi_1^0) \ldots f_k'(\underline{y}_k^{t+1}, \xi_k^0)x + u(x) + \eta^t \frac{1}{2}\|x - x^t\|^2$.
end for
Return $\bar{x}^N := \sum_{t=0}^{N-1} w^t x^{t+1} / \sum_{t=0}^{N-1} w^t$.

---

[8] A generic row-separable Bregman distance function V_i is used to denote the proximity-penalty term because different types of layer functions require different kinds of Bregman's distances.

3.3 Q_i Decomposition

The effects of Algorithm 3 in reducing the component gap functions Q_i can be illustrated more clearly by comparing each Q_i with the three-point inequality (2.20) of the corresponding proximal update. More specifically, for a fixed $i \geq 1$, Q_i can be written more compactly as

$$Q_i(z^{t+1}, z) := \pi_p \left((\pi_i - \pi_i^{t+1}) \mathcal{L}_q(x^{t+1}; \pi_q^{t+1}) - (f_i^*(\pi_i) - f_i^*(\pi_i^{t+1})) \right), \tag{3.9}$$

where the subscript p denotes the indices of all layers outside layer i, i.e., $p \equiv 1 : i-1$, and the subscript q denotes the indices of all layers inside layer i, i.e., $q \equiv i+1 :$. In this compact form, the left-hand-side of the corresponding three-point inequality (2.20) associated with the proximal update in Line 7 of Algorithm 3 is given by

$$\mathcal{A}_i^t := \pi_p \left((\pi_i - \pi_i^{t+1})[\mathcal{L}_q(x^t; \pi_q^{t+1}(\xi_q^i)) + \theta^t \pi_q^t(\xi_q^i)(x^t - x^{t-1})] - [f_i^*(\pi_i) - f_i^*(\pi_i^{t+1})] \right). \tag{3.10}$$

The differences between them are:

$$\begin{aligned} Q_i(z^{t+1}, z) - \mathcal{A}_i^t = &\\ & \pi_p(\pi_i - \pi_i^{t+1})(\mathcal{L}_q(x^{t+1}; \pi_q^{t+1}) - \mathcal{L}_q(x^{t+1}; \pi_q^{t+1}(\xi_q^i))) \quad (3.11)\\ & + \pi_p(\pi_i - \pi_i^{t+1}) \left(\mathcal{L}_q(x^{t+1}; \pi_q^{t+1}(\xi_q^i)) - [\mathcal{L}_q(x^t; \pi_q^{t+1}(\xi_q^i)) + \theta^t \pi_q^t(\xi_q^i)(x^t - x^{t-1})] \right), \quad (3.12) \end{aligned}$$

where (3.11) and (3.12) capture the errors resulted from using stochastic estimators and from using momentum *guesses* respectively.

Now, let us take a closer look at (3.12):

$$\begin{aligned} (3.12) &= \pi_p(\pi_i - \pi_i^{t+1}) \left([\mathcal{L}_q(x^{t+1}; \pi_q^{t+1}(\xi_q^i)) - \mathcal{L}_q(x^t; \pi_q^{t+1}(\xi_q^i))] - \theta^t \pi_q^t(\xi_q^i)(x^t - x^{t-1}) \right) \\ &= \pi_p \left((\pi_i - \pi_i^{t+1}) \pi_q^{t+1}(\xi_q^i)(x^{t+1} - x^t) - \theta^t [(\pi_i - \pi_i^t) \pi_q^t(\xi_q^i)(x^t - x^{t-1})] \right) \\ &\quad + \theta^t \pi_p(\pi_i^{t+1} - \pi_i^t) \pi_q^t(\xi_q^i)(x^t - x^{t-1}). \end{aligned}$$

In other words, $(3.12) = \mathcal{T}_i^t + \mathcal{C}_i^t$ with

$$\begin{aligned} \mathcal{T}_i^t &:= \pi_p \left((\pi_i - \pi_i^{t+1}) \pi_q^{t+1}(\xi_q^i)(x^{t+1} - x^t) - \theta^t [(\pi_i - \pi_i^t) \pi_q^t(\xi_q^i)(x^t - x^{t-1})] \right), \\ \mathcal{C}_i^t &:= \theta^t \pi_p(\pi_i^{t+1} - \pi_i^t) \pi_q^t(\xi_q^i)(x^t - x^{t-1}). \end{aligned}$$

It is clear that with appropriately chosen θ^t, $\mathcal{T}_i^t$ admits telescopic cancellations and $\mathcal{C}_i^t$ can be canceled with $V_i(\pi_i^t, \pi_i^{t+1})$ and $\frac{1}{2}\|x^t - x^{t-1}\|^2$. We remark here that this decomposition is due to the stochastic momentum *guess* (3.8). Such a *guess* is inspired by the novel momentum *guess* proposed in [26], so similar to how the consequent decomposition in [26] accelerates the SD algorithm, the above decomposition helps Algorithm 3 to achieve improved rates of convergence for NSCO.

Next, we move on to analyzing (3.11). The following decomposition of the stochastic error holds:

$$\begin{aligned} (3.11) =& \pi_p(\pi_i - \pi_i^{t+1})[\mathcal{L}_q(x^{t+1}; \pi_q^{t+1}) - \mathcal{L}_q(x^{t+1}; \pi_q^{t+1}(\xi_q^i))] \\ =& \pi_p(\pi_i - \pi_i^{t+1})[\mathcal{L}_q(x^t; \pi_q^{t+1}) - \mathcal{L}_q(x^t; \pi_q^{t+1}(\xi_q^i))] + \pi_p(\pi_i - \pi_i^{t+1})(\pi_q^{t+1} - \pi_q^{t+1}(\xi_q^i))(x^{t+1} - x^t) \\ =& \pi_p(\pi_i - \pi_i^t)[\mathcal{L}_q(x^t; \pi_q^{t+1}) - \mathcal{L}_q(x^t; \pi_q^{t+1}(\xi_q^i))] + \pi_p(\pi_i^t - \pi_i^{t+1})[\mathcal{L}_q(x^t; \pi_q^{t+1}) - \mathcal{L}_q(x^t; \pi_q^{t+1}(\xi_q^i))] + \varDelta_{x,i}^t \\ =& \delta_i^t + \varDelta_{\pi_i}^t + \varDelta_{x,i}^t, \end{aligned}$$

where

$$\begin{aligned} \delta_i^t &:= \pi_p(\pi_i - \pi_i^t)[\mathcal{L}_q(x^t; \pi_q^{t+1}) - \mathcal{L}_q(x^t; \pi_q^{t+1}(\xi_q^i))], \\ \varDelta_{\pi_i}^t &:= \pi_p(\pi_i^t - \pi_i^{t+1})[\mathcal{L}_q(x^t; \pi_q^{t+1}) - \mathcal{L}_q(x^t; \pi_q^{t+1}(\xi_q^i))], \\ \varDelta_{x,i}^t &:= \pi_p(\pi_i - \pi_i^{t+1})(\pi_q^{t+1} - \pi_q^{t+1}(\xi_q^i))(x^{t+1} - x^t). \end{aligned}$$

Notice that $\varDelta_{\pi_i}^t$ can be canceled using $V_i(\pi_i^t, \pi_i^{t+1})$ and that $\varDelta_{x,i}^t$ can be handled with $\frac{1}{2}\|x^{t+1} - x^t\|^2$. In addition, we will show in the next subsection that $\mathcal{L}_q(x^t; \pi_q^{t+1}(\xi_q^i))$ is an unbiased estimator for $\mathcal{L}_q(x^t; \pi_q^{t+1})$ in δ_i^t. In other words, we have developed a unified decomposition for each Q_i,

$$Q_i(z^{t+1}, z) = \mathcal{A}_i^t + \mathcal{T}_i^t + \mathcal{C}_i^t + \delta_i^t + \varDelta_{\pi_i}^t + \varDelta_{x,i}^t, \tag{3.13}$$

and grouped the terms into components according to how they could possibly be handled. Such a unified decomposition is the key to the forthcoming modular convergence analysis for Algorithm 3.

In addition, we note that for Q_0, it is straightforward to extend (2.30) to obtain

$$Q_0(z^{t+1}, z) = \mathcal{A}_0^t + \Delta_{x,0}^t + \delta_0^t, \tag{3.14}$$

where

$$\begin{aligned} \mathcal{A}_0^t &:= \pi_{1:}^{t+1}(\xi_{1:}^0)(x^{t+1} - x) + u(x^{t+1}) - u(x), \\ \Delta_{x,0}^t &:= (\pi_{1:}^{t+1} - \pi_{1:}^{t+1}(\xi_{1:}^0))(x^{t+1} - x^t), \\ \delta_0^t &:= (\pi_{1:}^{t+1} - \pi_{1:}^{t+1}(\xi_{1:}^0))(x^t - x). \end{aligned}$$

$\mathcal{A}_0^t$ is the left-hand-side of the three-point inequality (2.20) implied by Line 9 of Algorithm 3, $\Delta_{x,0}^t$ can be handled with $\frac{1}{2}\|x^{t+1} - x^t\|^2$, and $\pi_{1:}^{t+1}(\xi_{1:}^0)$ is an unbiased estimator for $\pi_{1:}^{t+1}$ in δ_0^t.

3.4 Repeated Calls to $\mathcal{DSO}_i$

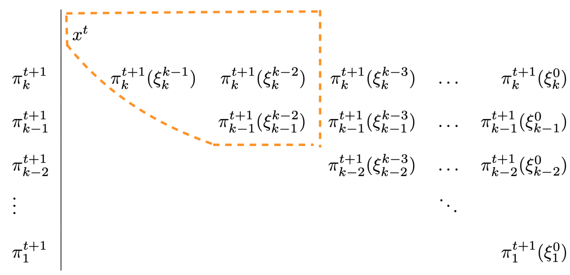

Fig. 2: Illustration of Stochastic Dependency. The stochastic estimators $\pi_i^{t+1}(\xi_i^j)$ in the π_i^{t+1} row are independent estimators of π_i^{t+1}. π_i^{t+1} is generated by a proximal update with an argument $\bar{y}_i^{t+1}$, consisting of all stochastic estimators in the $(k+1-i)$th column, $\{\pi_r^{t+1}(\xi_r^i)\}_{r>i}$. So every $\pi_i^{t+1}(\xi_i^j)$ depends on the entire "triangle" of estimators, $\{\pi_l^{t+1}(\xi_l^r)\}_{l>i, r\geq i}$. Now to obtain an unbiased estimator for $\pi_{i:}^{t+1}$, it is necessary to call all $\{\mathcal{DSO}_j\}_{j>i}$ again to obtain estimators outside the "triangle", i.e., $\{\pi_j^{t+1}(\xi_j^l)\}_{j>i, l<i}$, which are conditionally independent of π_i^{t+1}.

We explain the repeated calls to $\mathcal{DSO}_i$ in Algorithm 3. Algorithm 3 appears to be wasteful in its use of stochastic estimators in that it calls all inner oracles $\{\mathcal{DSO}_j\}_{j>i}$ to draw fresh estimators for every π_i^{t+1} update. But as explained in Figure 2, these fresh estimators are essential for obtaining unbiased estimators for $\{\pi_{i:}^{t+1}\}$. More formally, the following lemma shows that $\pi_{i+1:}^{t+1}(\xi_{i+1:}^i)$ and $\mathcal{L}_{i:}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^i))$ are indeed the desired unbiased estimators for $\pi_{i:}^{t+1}$ and $\mathcal{L}_i(x^t; \pi_{i:}^{t+1})$, respectively.

Proposition 13 *Let the stochastic estimators $\{\pi_i^{t+1}(\xi_i^j)\}_{0\leq j<i\leq k}$ be generated in Algorithm 3. If $j < i$, then*

$$\mathbb{E}[\pi_{i:}^{t+1}(\xi_{i:}^j)|\pi_{i:}^{t+1}] = \pi_{i:}^{t+1} \text{ and } \mathbb{E}[\mathcal{L}_i(x^t; \pi_{i:}^{t+1}(\xi_{i:}^j))|\pi_{i:}^{t+1}] = \mathcal{L}_i(x^t; \pi_{i:}^{t+1}). \tag{3.15}$$

Moreover, if $M_{\mathcal{L}_{i:}}^2 := \max_{x\in X, \pi_{i:}\in\Pi_{i:}} \left\{\mathbb{E}[\|\mathcal{L}_i(x; \pi_{i:}(\xi_{i:}))\|^2]\right\}$ then their variances satisfy

$$\mathbb{E}\left[\|\pi_{i:}^{t+1}(\xi_{i:}^j) - \pi_{i:}^{t+1}\|^2 | \pi_{i:}^{t+1}\right] \leq \sigma_{\pi_{i:}}^2 := \textstyle\sum_{r=i}^k M_{\Pi_{i:r-1}}^2 \sigma_{\pi_r}^2 M_{\Pi_{r+1:}}^2,$$

$$\mathbb{E}\left[\|\mathcal{L}_i(x^t; \pi_{i:}^{t+1}(\xi_{i:}^j)) - \mathcal{L}_i(x^t; \pi_{i:}^{t+1})\|^2 | \pi_{i:}^{t+1}\right] \leq \sigma_{\mathcal{L}_i}^2 := \textstyle\sum_{r=i}^k M_{\Pi_{i:r-1}}^2 (6\sigma_{\pi_r}^2 M_{\mathcal{L}_{r+1:}}^2 + 4\sigma_{f_r}^2).$$

Proof Firstly, we use backward induction on layer indices to show the expectation result. Let $P(i)$ denote the following statement about unbiasedness for layer i:

$$\mathbb{E}[\pi_{i:}^{t+1}(\xi_{i:}^j)|\pi_{i:}^{t+1}] = \pi_{i:}^{t+1}, \text{ and } \mathbb{E}[\mathcal{L}_i(x^t; \pi_{i:}^{t+1}(\xi_{i:}^j))|\pi_{i:}^{t+1}] = \mathcal{L}_i(x^t; \pi_{i:}^{t+1}), \forall j < i.$$

Clearly, the definition of $\mathcal{DSO}_i$ implies that $P(k)$ holds. Now assume that $P(i+1)$ is true for some $1 \le i \le k-1$. Because $\pi_i^{t+1}(\xi_i)$ depends only on π_i^{t+1}, $\pi_i^{t+1}(\xi_i)$ is d-separated from $\{\pi_l^{t+1}(\xi_l^j), f_l^*(\pi_l^{t+1}, \xi_l^j)\}_{l>i}$ conditioned on π_i^{t+1}. So,

$$\mathbb{E}[\pi_{i:}^{t+1}(\xi_{i:}^j)|\pi_{i:}^{t+1}] = \mathbb{E}[\pi_i^{t+1}(\xi_i^j)|\pi_{i:}^{t+1}]\mathbb{E}[\pi_{i+1:}^{t+1}(\xi_{i+1:}^j)|\pi_{i:}^{t+1}] = \pi_i^{t+1}\mathbb{E}[\pi_{i+1:}^{t+1}(\xi_{i+1:}^j)|\pi_{i:}^{t+1}],$$

and

$$\mathbb{E}[\mathcal{L}_i(x^t; \pi_{i:}^{t+1}(\xi_{i:}^j))|\pi_{i:}^{t+1}] = \pi_i^{t+1}\mathbb{E}[\mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^j))|\pi_{i:}^{t+1}] - f_i^*(\pi_i^{t+1}).$$

Moreover, since $j < i$, the estimators $\{\pi_l^{t+1}(\xi_l^j), f_l^*(\pi_l^{t+1}, \xi_l^j)\}_{l>i}$ are used neither directly nor indirectly in generating π_i^{t+1}. So π_i^{t+1} is independent of both $\pi_{i+1:}^{t+1}(\xi_{i+1:}^j)$ and $\mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^j))$ conditioned on $\pi_{i+1:}^{t+1}$. Thus

$$\mathbb{E}[\pi_{i+1:}^{t+1}(\xi_{i+1:}^j)|\pi_{i:}^{t+1}] = \mathbb{E}[\pi_{i+1:}^{t+1}(\xi_{i+1:}^j)|\pi_{i+1:}^{t+1}] = \pi_{i+1:}^{t+1},$$
$$\mathbb{E}[\mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^j))|\pi_{i:}^{t+1}] = \mathbb{E}[\mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^j))|\pi_{i+1:}^{t+1}] = \mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}),$$

where the second equalities in both relations follow from the induction hypothesis, $P(i+1)$. In view of the above two observations, we can conclude that $P(i)$ is true if $P(i+1)$ is true. Thus $P(j)$ holds for all $1 \le j \le k$.

Now we turn our attention to the variance bounds. The next decomposition follows from the definition of nested Lagrangian:

$$\begin{aligned}
&\mathcal{L}_i(x^t; \pi_{i:}^{t+1}(\xi_{i:}^j)) - \mathcal{L}_i(x^t; \pi_{i:}^{t+1}) \\
&= \mathcal{L}_i(x^t; \pi_i^{t+1}(\xi_i), \pi_{i+1:}^{t+1}(\xi_{i+1:})) - [\pi_i^{t+1}\mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^j)) - f_i^*(\pi_i^{t+1})] \\
&\quad + \pi_i^{t+1}[\mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^j)) - \mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1})] \\
&= (\pi_i^{t+1}(\xi_i) - \pi_i^{t+1})\mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^j)) - (f_i^*(\pi_i^{t+1}, \xi_i^j) - f_i^*(\pi_i^{t+1})) \\
&\quad + \pi_i^{t+1}[\mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^j)) - \mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1})].
\end{aligned}$$

Moreover, since both $\pi_i^{t+1}(\xi_i) - \pi_i^{t+1}$ and $f_i^*(\pi_i^{t+1}, \xi_i^j) - f_i^*(\pi_i^{t+1})$ are uncorrelated with $\mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^j))$ conditioned on π_i^{t+1}, we have

$$\mathbb{E}[\|\mathcal{L}_i(x^t; \pi_{i:}^{t+1}(\xi_{i:}^j)) - \mathcal{L}_i(x^t; \pi_{i:}^{t+1})\|^2|\pi_{i:}^{t+1}] \tag{3.16}$$
$$= \mathbb{E}[\|(\pi_i^{t+1}(\xi_i) - \pi_i^{t+1})\mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^j)) - (f_i^*(\pi_i^{t+1}, \xi_i^j) - f_i^*(\pi_i^{t+1}))\|^2|\pi_{i:}^{t+1}] \tag{3.17}$$
$$+ \mathbb{E}[\|\pi_i^{t+1}[\mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^j)) - \mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1})]\|^2|\pi_{i:}^{t+1}]. \tag{3.18}$$

The variance assumption for $\mathcal{SO}_i$ and the conditional independence relationships imply that

$$\mathbb{E}[\|(\pi_i^{t+1}(\xi_i) - \pi_i^{t+1})\mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^j))\|^2||\pi_{i:}^{t+1}] \le \sigma_{\pi_i}^2 M_{\mathcal{L}_{i+1:}}^2.$$

Because $f_i^*(\pi_i^{t+1}, \xi_i^j) := f_i'(\underline{y}_i^t, \xi_i^j)\underline{y}_i^t - f_i(\underline{y}_i^t, \xi_i^j)$ for some $\underline{y}_i^t$ with $\mathbb{E}[\|\underline{y}_i^t\|^2] \le M_{\mathcal{L}_{i+1:}}^2$, we have

$$\begin{aligned}
&\mathbb{E}[\|f_i^*(\pi_i^{t+1}, \xi_i^j) - f_i^*(\pi_i^{t+1})\|^2|\pi_{i:}^{t+1}] \\
&\quad = \mathbb{E}[\|(\pi_i^{t+1}(\xi_i^j) - \pi_i^{t+1})\underline{y}_i^{t+1} - (f_i(\underline{y}_i^{t+1}, \xi_i^j) - f_i(\underline{y}_i^{t+1}))\|^2|\pi_{i:}^{t+1}] \\
&\quad \le 2\sigma_{\pi_i}^2 M_{\mathcal{L}_{i+1:}}^2 + 2\sigma_{f_i}^2.
\end{aligned}$$

So it follows from the algebraic identity $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ that

$$(3.17) \le 4\sigma_{f_i}^2 + 6\sigma_{\pi_i}^2 M_{\mathcal{L}_{i+1:}}^2. \tag{3.19}$$

In addition, since $j < i$, π_i^{t+1} is independent of $\mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^j))$ conditional on $\pi_{i+1:}^{t+1}$, (3.18) can be simplified to

$$\begin{aligned}
(3.18) &\le M_{\Pi_i}^2\mathbb{E}[\|\mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^j)) - \mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1})\|^2|\pi_i^{t+1}, \pi_{i+1:}^{t+1}] \\
&\le M_{\Pi_i}^2\mathbb{E}[\|\mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1}(\xi_{i+1:}^j)) - \mathcal{L}_{i+1}(x^t; \pi_{i+1:}^{t+1})\|^2|\pi_{i+1:}^{t+1}].
\end{aligned} \tag{3.20}$$

So we can substitute (3.19) and (3.20) into (3.16) to obtain

$$\mathbb{E}[\|\mathcal{L}_i(x^t; \pi_{i:}^{t+1}(\xi_{i:}^j)) - \mathcal{L}_i(x^t; \pi_{i:}^{t+1})\|^2|\pi_{i:}^{t+1}] \le$$

$$4\sigma_{f_i}^2 + 6\sigma_{\pi_i}^2 M_{\mathcal{L}_{i+1:}}^2 + M_{\Pi_i}^2 \underbrace{\mathbb{E}[\|\mathcal{L}_{i+1}(x^t;\pi_{i+1:}^{t+1}(\xi_{i+1:}^j)) - \mathcal{L}_{i+1}(x^t;\pi_{i+1:}^{t+1})\|^2|\pi_{i+1:}^{t+1}]}_{A}.$$

Notice that A is the same as right-hand-side of the above inequality except for changing the index from i to $i+1$. So the desired bound on $\mathbb{E}[\|\mathcal{L}_i(x^t;\pi_{i:}^{t+1}(\xi_{i:}^j)) - \mathcal{L}_i(x^t;\pi_{i:}^{t+1})\|^2|\pi_{i:}^{t+1}]$ follows from applying the above inequality recursively.

Finally, the bound for $\mathbb{E}[\|\pi_{i:}^{t+1}(\xi_{i:}^j) - \pi_{i:}^{t+1}\|^2|\pi_{i:}^{t+1}]$ can be derived in a similar fashion. ∎

4 Modular π_i-Updates and Q_i-Bounds

In this section, we provide the π_i-updates for Algorithm 3 in more details and develop the associated Q_i convergence bounds. Obviously, for a given f_i, the update depends on its function class (affine, smooth, smoothable or non-smooth), its oracle type (deterministic or stochastic) and other composing layers. But the dependence on other layer functions is quite weak. In particular, the π_i-update only needs to know whether the argument $\bar{y}_i^{t+1}$ is *exact*, i.e., no stochastic estimators are involved in its construction (e.g., x^t in the π_2-update of Algorithm 1) or *noisy*, i.e., stochastic estimators are involved (e.g., $\mathcal{L}_2(x^t;\pi_2^{t+1}(\xi_2^1))$ in the π_1-update of Algorithm 1). Thus it suffices to propose a modular π_i-update for every type of function, oracle and argument. Moreover, since the Q_i convergence analysis focuses only on the (implicit) dual iterates $\{\pi_i^t\}$ rather than their stochastic estimators, a stochastic f_i should have the same Q_i bound as the deterministic f_i. So we develop a separate Q_i-bound only for every function class and every argument type.

The convergence bounds on Q_i is based on the decomposition in (c.f. (3.13)),

$$Q_i(z^{t+1}, z) = \mathcal{A}_i^t + \mathcal{T}_i^t + \mathcal{C}_i^t + \delta_i^t + \varDelta_{\pi_i}^t + \varDelta_{x,i}^t,$$

and the three point inequality (2.20),

$$\mathcal{A}_i^t \le \tau_i^t \pi_p V_i(\pi_i^t, \pi_i) - \tau_i^t \pi_p V_i(\pi_i^t, \pi_i^{t+1}) - (\tau_i^t + \alpha_{f_i^*})\pi_p V_i(\pi_i^{t+1}, \pi_i), \tag{4.1}$$

where $\alpha_{f_i^*}$ is the strong convexity modulus of f_i^* with respect to V_i. Essentially, we will fix $w^t := (t+1)/2$ and $\theta^t := (t)/(t+1)$ and use the terms from (4.1) to cancel all but $\varDelta_{x,i}^t$ and δ_i^t in Q_i to obtain bounds for $\mathbb{E}[\sum_{t=0}^{N-1} w^t Q_i(z^{t+1}, z)]$ for all $z := (x^*; \pi_1, \pi_2, \dots, \pi_k)$ where $(\pi_1, \pi_2, \dots, \pi_k)$ can depend on $\{z^t\}$.

Now we define a few constants that will become handy. Recall the compact notation (3.9) for Q_i:

$$Q_i(z^{t+1}, z) := \pi_p \left((\pi_i - \pi_i^{t+1}) \mathcal{L}_q(x^{t+1}; \pi_q^{t+1}) - (f_i^*(\pi_i) - f_i^*(\pi_i^{t+1})) \right). \tag{4.2}$$

We let M_p[9] denote an upper bound for $\|\pi_p\|$ for all possible reference points, let $M_q^2 := M_{\Pi_{i+1}}^2 M_{\Pi_{i+2}}^2 \dots M_{\Pi_k}^2 \ge \mathbb{E}[\|\pi_q^{t+1}(\xi_q^i)\|^2]$ and let $\sigma_{\mathcal{L}_q}^2 := \sigma_{\mathcal{L}_{i+1}}^2 \ge \mathbb{E}[\|\mathcal{L}(x^t;\pi_{i+1:}^{t+1}[i] - \mathcal{L}_{i+1}(x^t;\pi_{i+1:}^{t+1})\|^2)$ (derived in Proposition 13). Moreover, we will group the stochastic terms that depend on x-prox update into $\mathrm{STC}_i := \mathbb{E}[\sum_{t=0}^{N-1} w^t \varDelta_{x,i}^t + \delta_i^t]$ and leave their discussion to Section 5.

4.1 Affine Layer Function

In this subsection, we consider the easiest affine layer function, $f_i(y_i) := A_i y_i + b_i$. Clearly, Π_i is a singleton set, $\{A_i\}$, and the subgradient (resp., estimator) returned by the (stochastic) oracle is A_i (resp., $A_i(\xi_i)$). So regardless of the Bregman's distance V_i, the stepsize τ_i^t and the argument $\bar{y}_i^{t+1}$, the proximal update

$$\pi_i^{t+1} \in \operatorname*{arg\,min}_{\pi_i \in \Pi_i} -\pi_i \bar{y}_i^{t+1} + f_i^*(\pi_i) + \tau_i^t V_i(\pi_i^t, \pi_i)$$

always returns A_i. More specifically, given a $\bar{y}_i^{t+1}$, the jth element of $\{[\pi_i^{t+1}(\xi_i^j), f_i^*(\pi_i^{t+1}, \xi_i^j)]\}$, returned by Line 7 in Algorithm 3, is:

1. $[A_i(\xi_i^j), f_i(\bar{y}_i^{t+1}, \xi_i^j) - A_i(\xi_i^j)\bar{y}_i^{t+1}]$ if f_i has a stochastic oracle,
2. $[A_i, f_i(\bar{y}_i^{t+1}) - A_i \bar{y}_i^{t+1}]$ if f_i has a deterministic oracle.

As for the bound on Q_i, notice that $\pi_i^{t+1} - \pi_i = 0$ for any feasible π_i, thus $Q_i(z^{t+1}, z)$ is always zero. This observation leads to the following Q_i convergence bound.

Proposition 14 *If f_i is affine, then the solution sequence $\{z^t := (x^t; \pi_1^t, \pi_2^t, \dots, \pi_k^t)\}$ generated by Algorithm 3 satisfies*

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q_i(z^{t+1}, z)] = 0. \tag{4.3}$$

[9] While $M_{\Pi_1} M_{\Pi_2} \dots M_{\Pi_{i-1}}$ gives such an upper bound, M_p is usually much smaller, especially when we have some rough idea about the location of the reference point.

4.2 Smooth Layer Function

In this subsection, we consider the smooth layer function f_i such that $\|f_i'(y_i) - f_i'(\bar{y}_i)\|$[10] $\leq L_{f_i}\|y_i - \bar{y}_i\|, \forall y_i, \bar{y}_i \in \mathbb{R}^{m_i}$. A direct computation of the corresponding π_i-proximal update is not always possible because f_i^* may not be known or simple. So we use $D_{f_i^*}$ as the proximity penalty function to implement an implicit proximal update from an *associated* π_i^t:

$$\pi_i^{t+1} \in \underset{\pi_i \in \Pi_i}{\arg\min} -\pi_i \bar{y}_i^{t+1} + f_i^*(\pi_i) + \tau_i^t D_{f_i^*}(\pi_i^t, \pi_i).$$

More specifically, given a $\pi_i^t = f_i'(\underline{y}_i^t)$ and a proximal update argument $\bar{y}_i^{t+1}$, π_i^{t+1} is associated with $\underline{y}_i^{t+1} := (\tau_i^t \underline{y}_i^t + \bar{y}_i^{t+1})/(1+\tau_i^t)$ and the j-th element of $\{[\pi_i^{t+1}(\xi_i^j), f_i^*(\pi_i^{t+1}, \xi_i^j)]\}$, returned by Line 7 in Algorithm 3, is

1. $[f_i'(\underline{y}_i^{t+1}, \xi_i^j), f_i(\underline{y}_i^{t+1}, \xi_i^j) - f_i'(\underline{y}_i^{t+1}, \xi_i^j)\bar{y}_i^{t+1}]$ if f_i has a stochastic oracle,
2. $[f_i'(\underline{y}_i^{t+1}), f_i(\bar{y}_i^{t+1}) - f_i'(\underline{y}_i^{t+1})\bar{y}_i^{t+1}]$ if f_i has a deterministic oracle.

In order to develop the Q_i convergence bound, we need to show a technical result for the strong convexity modulus of the multi-dimensional Bregman's distance function $D_{f_i^*}$.

Proposition 15 *Let an m-dimensional vector function g be L_g-smooth , i.e., $\|g'(y) - g'(\bar{y})\|$[1] $\leq L_g\|y - \bar{y}\|$ and let g^* and D_{g^*} denote its (component-wise) conjugate function and (component-wise) conjugate Bregman's distance function. Then given an m-dimensional non-negative weight vector w, we have*

$$\|w\| w^\top D_{g^*}(\bar{\pi}, \pi) \geq \frac{\|w^\top(\bar{\pi} - \pi)\|^2}{2L_g} \tag{4.4}$$

for any associated $\bar{\pi}$ and π. In particular, if f_i is L_{f_i}-smooth, then the dual solutions $\{\pi_i^t\}$ generated by Algorithm 3 satisfy

$$\|\pi_p\| \pi_p V_i(\pi_i^t, \pi_i^{t+1}) \geq \frac{\|\pi_p(\pi_i^{t+1} - \pi_i^t)\|^2}{2L_{f_i}}. \tag{4.5}$$

Proof First, if $w = 0$, (4.4) is clearly true. Now, assume $w \neq 0$. The definition of operator norm implies

$$\|u^\top(g'(y) - g'(\bar{y}))\| \leq L_g\|y - \bar{y}\| \ \forall u \text{ with } \|u\| = 1.$$

So the one-dimensional $g_{u_w}(y) := u_w^\top g(y)$ with $u_w := \frac{w}{\|w\|}$ is L_g-Lipschitz smooth and its Fenchel conjugate $g_{u_w}^*$ is $1/L_g$ strongly convex. More specifically, since $g_{u_w}'(y) = u_w^\top g'(y)$, we have

$$g_{u_w}^*(u_w^\top \pi) - g_{u_w}^*(u_w^\top \bar{\pi}) - u_w^\top(\pi - \bar{\pi})\bar{y} \geq \frac{1}{2L_g}\|u_w^\top(\pi - \bar{\pi})\|^2, \text{ if } \bar{\pi} = g'(\bar{y}) \text{ , i.e., } u_w^\top \bar{\pi} = g_{u_w}'(\bar{y}).$$

Thus the key to showing (4.4) is to relate $g_{u_w}^*(u_w^\top \pi)$ with $u_w^\top g^*(\pi)$ on associated π's.

Those two quantities are quite different in general. In $g_{u_w}^*(u_w^\top \bar{\pi}) := \max_y u_w^\top \bar{\pi} y - g_{u_w}(y)$, we can choose only one overall maximizer, y^*, but in $u_w^\top g^*(\pi) = \sum_j u_{w,j} \max_{y_j} \bar{\pi}_j y_j - g_j(y_j)$, multiple maximizers y_j^* can be selected for every $\bar{\pi}_j$. So we always have $g_{u_w}^*(u_w^\top \pi) \leq u_w^\top g^*(\pi)$. However, for associated π, all those y_j^*s are the same. More specifically, let $\pi = g'(y^*)$, i.e., $\pi_j = g_j'(y^*)$. Then the conjugate duality implies that $g^*(\pi) = \pi y^* - g(y^*)$, so

$$g_{u_w}^*(u_w^\top \pi) := \max_{\bar{y}} u_w^\top \pi \bar{y} - g_{u_w}(\bar{y}) \geq u_w^\top(\pi y^* - g(y^*)) = u_w^\top g^*(\pi).$$

Therefore, $g_{u_w}^*(u_w^\top \pi) = u_w^\top g^*(\pi)$ holds for all associated π's. If $\bar{\pi} = g'(\bar{y})$ and $\pi = g'(y)$, we have

$$\begin{aligned} u_w^\top D_{g^*}(\bar{\pi}, \pi) &= u_w^\top g^*(\pi) - u_w^\top g^*(\bar{\pi}) - u_w^\top(\pi - \bar{\pi})\bar{y} \\ &= g_{u_w}^*(u_w \pi) - g_{u_w}^*(u_w \bar{\pi}) - u_w^\top(\pi - \bar{\pi})\bar{y} \geq \frac{1}{2L_g}\|u_w^\top(\pi - \bar{\pi})\|^2. \end{aligned}$$

Then (4.4) follows from multiplying both sides of the above inequality by $\|w\|^2$. In addition, for smooth functions f_i, the dual solutions $\{\pi_i^t\}$ are always associated with $\underline{y}_i^t$, since they are produced by implicit proximal updates. So (4.5) follows immediately. ■

Now we are ready for the convergence bounds on Q_i. First, we consider the general case with noisy arguments.

[10] l_2 operator norm.

[1] l_2 operator norm.

Proposition 16 *Let f_i be L_{f_i}-smooth with a noisy argument. If solutions $\{z^t := (x^t; \pi_1^t, \ldots, \pi_k^t)\}$ are generated by Algorithm 3 with $\tau_i^t = \frac{t+1}{6} + \frac{t}{4}$, then*

$$\begin{aligned}\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q_i(z^{t+1}, z)] \leq \frac{1}{6} M_p \mathcal{D}_{\Pi_i}^2 + \frac{3N}{2} M_p L_{f_i} \sigma_{\mathcal{L}_q}^2 + \mathbb{E}[\textstyle\sum_{t=0}^{N-1} \frac{2w^t}{t+1} \tilde{L}_{f_i} \|x^{t+1} - x^t\|^2] \\ + STC_i,\end{aligned} \tag{4.6}$$

where $\mathcal{D}_{\Pi_i}^2 := \mathbb{E}[\|D_{f_i^*}(\pi_i^0, \pi_i)\|]$, $\tilde{L}_{f_i} := M_p L_{f_i} M_q^2$ and $STC_i := \mathbb{E}[\sum_{t=0}^{N-1} w^t \delta_i^t + \Delta_{x,i}^t]$.

Proof Since f_i^* is 1-strongly convex with respect to $D_{f_i^*}$, π_i^{t+1} satisfies the following three point inequality (2.20):

$$\mathcal{A}_i^t \leq \tau_i^t \pi_p D_{f_i^*}(\pi_i^t, \pi_i) - \tau_i^t \pi_p D_{f_i^*}(\pi_i^t, \pi_i^{t+1}) - (\tau_i^t + 1)\pi_p D_{f_i^*}(\pi_i^{t+1}, \pi_i). \tag{4.7}$$

So the w^t-weighted sum of $\mathcal{A}_i^t$ admits a telescopic cancellation:

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t \mathcal{A}_i^t] \leq w_0 \tau_i^0 \underbrace{(\pi_p D_{f_i^*}(\pi_i^0, \pi_i)}_{\leq M_p \mathcal{D}_{\Pi_i}^2} - w^{N-1}(\tau_i^{N-1} + 1)\mathbb{E}[\pi_p D_{f_i^*}(\pi_i^N, \pi_i)] - \mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t \tau_i^t \pi_p D_{f_i^*}(\pi_i^t, \pi_i^{t+1})]. \tag{4.8}$$

We are going to use the negative terms in the above inequality to cancel $\mathcal{T}_i^t$, $\mathcal{C}_i^t$ and $\Delta_{\pi_i}^t$ in Q_i (see (3.13)).

Firstly, observe that $\sum_{t=0}^{N-1} w^t \mathcal{T}_i^t$ admits a telescopic cancellation and the remaining terms can be handled by $D_{f_i^*}(\pi_i^N, \pi_i)$:

$$\begin{aligned}&\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t \mathcal{T}_i^t - w^{N-1}(\tau_i^{N-1} + 1) D_{f_i^*}(\pi_i^N, \pi_i)] \\ &\leq \mathbb{E}[-w^{N-1}\pi_p(\pi_i^N - \pi_i)\pi_q^N(\xi_q^i)(x^N - x^{N-1}) - \frac{N w^{N-1}}{4} \pi_p D_{f_i^*}(\pi_i^N, \pi_i)] \\ &\overset{(a)}{\leq} \frac{2w^{N-1}}{N} M_p L_{f_i} \mathbb{E}[\|\pi_q^N(\xi_q^i)(x^N - x^{N-1})\|^2] \\ &\overset{(b)}{=} \frac{2w^{N-1}}{N} M_p L_{f_i} \mathbb{E}\left[\mathbb{E}[\|\pi_q^N(\xi_q^i)\|^2 | \pi_q^N] \mathbb{E}[\|x^N - x^{N-1}\|^2 | \pi_q^N]\right] \\ &\overset{(c)}{\leq} \frac{2w^{N-1}}{N} \mathbb{E}[M_p L_{f_i} M_q^2 \|x^N - x^{N-1}\|^2] = \mathbb{E}[\frac{2w^{N-1}}{N} \tilde{L}_{f_i} \|x^N - x^{N-1}\|^2],\end{aligned} \tag{4.9}$$

where (a) follows from $M_p L_{f_i} \pi_p D_{f_i^*}(\pi_i^N, \pi_i) \geq \frac{1}{2}\|\pi_p(\pi_i - \pi_i^N)\|^2$ by Proposition 15 and the Young's inequality, (b) follows from the conditional independence between $\pi_q^N(\xi_q^i)$ and x^N, and (c) follows from the definition of M_q.

Secondly, $\mathcal{C}_i^t$ can be canceled by $\sum_{t=0}^{N-1} \frac{tw^t}{4} D_{f_i^*}(\pi_i^t, \pi_i^{t+1})$ in similar fashion:

$$\begin{aligned}&\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t \left(\mathcal{C}_i^t - \frac{t}{4}\pi_p D_{f_i^*}(\pi_i^t, \pi_i^{t+1})\right)] \\ &= \mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t (\theta^t \pi_p(\pi_i^{t+1} - \pi_i^t)\pi_q^t(\xi_q^i)(x^t - x^{t-1}) - \frac{tw^t}{4}\pi_p D_{f_i^*}(\pi_i^t, \pi_i^{t+1}))] \\ &\overset{(b)}{\leq} \mathbb{E}[\textstyle\sum_{t=1}^{N-1} \frac{tw^t(\theta^t)^2}{2} M_p L_{f_i} M_q^2 \|x^t - x^{t-1}\|^2) \leq \mathbb{E}[\textstyle\sum_{t=1}^{N-1} \frac{2w^{t-1}}{t} \tilde{L}_{f_i} \|x^t - x^{t-1}\|^2],\end{aligned} \tag{4.10}$$

where (b) follows from $x_0 = x_{-1}$.

Furthermore the stochastic error $\Delta_{\pi_i}^t$ due to noisy arguments can be canceled by the remaining term $\sum_{t=0}^{N-1} w^t \frac{t+1}{6} \pi_p D_{f_i^*}(\pi_i^t, \pi_i^{t+1})$:

$$\begin{aligned}&\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t (\Delta_{\pi_i}^t - \frac{t+1}{6} D_{f_i^*}(\pi_i^t, \pi_i^{t+1}))] \\ &= \mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t ((\pi_p(\pi_i^{t+1} - \pi_i^t)(\mathcal{L}(\pi_q^{t+1}(\xi_q^i), x^t) - \mathcal{L}(\pi_q^{t+1}, x^t)) - \frac{t+1}{6}\pi_p D_{f_i^*}(\pi_i^t, \pi_i^{t+1}))] \\ &\leq \textstyle\sum_{t=0}^{N-1} w^t \frac{6}{t+1} \frac{1}{2} M_p L_{f_i} \sigma_{\mathcal{L}_q}^2 = \frac{3N}{2} M_p L_{f_i} \sigma_{\mathcal{L}_q}^2.\end{aligned} \tag{4.11}$$

The desired convergence bound (4.6) then follows from substituting (4.8), (4.9), (4.10) and (4.11) into $\mathbb{E}[\sum_{t=0}^{N-1} w^t Q_i(z^{t+1}, z)]$. ∎

Clearly, in the easier case with exact arguments, only a slight modification to the above analysis is needed to develop the Q_i bound. In fact, since $\Delta_{\pi_i}^t = 0$, we can set $\tau_i^t = t/2$ to obtain a simpler bound, as illustrated in the next corollary.

Corollary 1 *Let f_i be L_{f_i}-smooth with an exact argument. If solutions $\{z^t := (x^t; \pi_1^t, \ldots, \pi_k^t)\}$ are generated by Algorithm 3 with $\tau_i^t = \frac{t}{2}$, then*

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q_i(z^{t+1}, z)] \leq \mathbb{E}[\textstyle\sum_{t=0}^{N-1} \frac{w^t}{t+1} \tilde{L}_{f_i} \|x^{t+1} - x^t\|^2], \tag{4.12}$$

where $\mathcal{D}_{\Pi_i}^2 := \mathbb{E}[\|D_{f_i^*}(\pi_i^0, \pi_i)\|]$ and $\tilde{L}_{f_i} := M_p L_{f_i} M_q^2$.

4.3 Smoothable Layer Function

Now we move on to smoothable layer functions. Nesterov [17] shows that a one-dimensional function g is *smoothable* if it can be expressed as $g(y) \equiv \max_{\pi \in \Pi} \pi y - g^*(y)$ for some known and *simple* g^*, i.e., the proximal update $\pi^+ \in \arg\max_{\pi \in \Pi} \pi y - g^*(y) - \frac{\tau}{2}\|\pi - \pi^-\|^2$ can be computed efficiently for any y, $\pi^- \in \Pi$ and $\tau \geq 0$. We call the multi-dimensional layer function f_i *smoothable* if all its components, $f_{i,j}$, are smoothable. Note that such a structural assumption is stronger than the first-order oracle assumption because both the value and the subgradient of f_i at any given $\bar{y}_i$ can be computed by solving

$$\arg\max_{\pi_{i,j} \in \Pi_{i,j}} \pi_{i,j} y_i - f^*_{i,j}(\pi_{i,j}) - (0)\|\pi_{i,j} - \bar{\pi}_{i,j}\|^2.$$

More specifically, to implement the proximal update in Line 7 of Algorithm 3,

$$\pi_i^{t+1} \in \arg\min_{\pi_i \in \Pi_i} -\pi_i \bar{y}_i^{t+1} + f_i^*(\pi_i) + \tau_i^t V_i(\pi_i^t, \pi_i),$$

we set $V_i(\bar{\pi}_i, \pi_i) := \sqrt{m_{i-1}}/2\left[\|\pi_{i,1} - \bar{\pi}_{i,1}\|^2, \|\pi_{i,2} - \bar{\pi}_{i,2}\|^2 \ldots, \|\pi_{i,m_{i-1}} - \bar{\pi}_{i,m_{i-1}}\|^2\right]$ such that each row of π_i^{t+1} is computed by

$$\pi_{i,j}^{t+1} \in \arg\min_{\pi_{i,j} \in \Pi_{i,j}} -\pi_{i,j} \bar{y}_i^{t+1} + f^*_{i,j}(\pi_{i,j}) + \tfrac{\tau_i^t}{2}\|\pi_{i,j}^t - \pi_{i,j}\|^2.$$

Then the l-th element of the estimate $\{[\pi_i^{t+1}(\xi_i^l), f_i^*(\pi_i^{t+1}, \xi_i^l)]\}_{l=0}^{i-1}$ is given by the exact value $[\pi_i^{t+1}, f_i^*(\pi_i^{t+1})]$.

Similar to the preceding subsection, we need to determine the strong convexity modulus of $\pi_p V_i(\cdot,\cdot)$ to derive the convergence bound on Q_i. The next proposition finds such a constant.

Proposition 17 *If $f_i : \mathbb{R}^{m_i} \to \mathbb{R}^{m_{i-1}}$ is smoothable and $V_{i,j}(\pi_{i,j}, \bar{\pi}_{i,j}) = \frac{1}{2}\|\pi_{i,j} - \bar{\pi}_{i,j}\|^2$, then*

$$\|\pi_p\| \pi_p V_i(\pi_i^t, \pi_i^{t+1}) \geq \frac{\|\pi_p(\pi_i^{t+1} - \pi_i^t)\|^2}{2}.$$

Proof If π_p is zero, then the desired inequality clearly holds. Now assume that $\pi_p \neq 0$ and let $u_w := \frac{\pi_p}{\|\pi_p\|_1}$. Assumption 2 implies that u_w is a non-negative weight vector that sums up to one. Therefore we conclude from Jensen's inequality that

$$u_w V_i(\pi_i, \bar{\pi}_i)/\sqrt{m_{i-1}} = \sum_{j=1}^{n_i} u_{w,i} \tfrac{1}{2}\|\pi_{i,j} - \bar{\pi}_{i,j}\|^2 \geq \tfrac{1}{2}\|\sum_{j=1}^{n_i} u_{w,j}(\pi_{i,j} - \bar{\pi}_{i,j})\|^2 = \frac{\|u_w(\pi_i^{t+1} - \pi_i^t)\|^2}{2}.$$

The desired inequality then follows from multiplying both sides with $\|\pi_p\|_1^2$ and noticing $\|\pi_p\|_1 \leq \sqrt{m_{i-1}}\|\pi_p\|_2$.
■

Now we are ready to show the convergence bounds on Q_i. One difference from the smooth layer exists in that f_i^* is no longer strongly convex with respect to V_i, so these bounds have worse dependencies on N. In the following proposition, two separate Q_i bounds are proposed for the general convex NSCO problem and the strongly convex NSCO problem respectively.

Proposition 18 *Let f_i be smoothable and let its argument be noisy. If $STC_i := \mathbb{E}[\sum_{t=0}^{N-1} w^t(\delta_i^t + \Delta_{x,i}^t)]$, $\mathcal{D}^2_{\Pi_i} := \max_{\pi_i, \bar{\pi}_i \in \Pi_i}\|V_i(\pi_i, \bar{\pi}_i)\|$ and $\tilde{\mathcal{D}}_{\Pi_i} := M_p \mathcal{D}_{\Pi_i} M_q$, then the solution sequence $\{z^t := (x^t; \pi_1^t, \pi_2^t, \ldots, \pi_k^t)\}$ generated by Algorithm 3 satisfies the following bounds.*

a) If $\tau_i^t = \frac{\sigma_{\mathcal{L}_q}\sqrt{N+1}}{2\mathcal{D}_{\Pi_i}} + \frac{\mathcal{D}_X M_q}{\mathcal{D}_{\Pi_i}}$, then

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q(z^{t+1}, z)] \leq \frac{N\sqrt{N+1}}{2} M_p \mathcal{D}_{\Pi_i}\sigma_{\mathcal{L}_q} + \frac{N}{2}\tilde{\mathcal{D}}_{\Pi_i}\mathcal{D}_X + \sum_{t=0}^{N-1}\frac{w^t}{2}\frac{\tilde{\mathcal{D}}_{\Pi_i}}{\mathcal{D}_X}\|x^{t+1} - x^t\|^2 + STC_i. \quad (4.13)$$

b) If $\tau_i^t = \frac{\sigma_{\mathcal{L}_q}\sqrt{N+1}}{2\mathcal{D}_{\Pi_i}} + \frac{2}{(t+1)c} M_p M_q^2$ for some $c > 0$, then

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q(z^{t+1}, z)] \leq \frac{N\sqrt{N+1}}{2} M_p \mathcal{D}_{\Pi_i}\sigma_{\mathcal{L}_q} + \frac{2}{c}\tilde{\mathcal{D}}^2_{\Pi_i} + \sum_{t=0}^{N-1}\frac{c(t+1)w^t}{4}\|x^{t+1} - x^t\|^2 + STC_i. \quad (4.14)$$

Proof The three point inequality (2.20) becomes

$$w^t\mathcal{A}_i^t \le w^t\tau_i^t\pi_p V_i(\pi_i^t,\pi_i) - w^t\tau_i^t\pi_p V_i(\pi_i^t,\pi_i^{t+1}) - w^t\tau_i^t\pi_p V_i(\pi_i^{t+1},\pi_i). \tag{4.15}$$

Compared with the increasing stepsizes τ_i^t used in Proposition 16, we have to keep τ_i^t almost constant, leading to a faster accumulation of errors. Moreover, we use different weighted temporal sums of $\{\|x^t - x^{t+1}\|^2\}$ to develop bounds; in a) each $\|x^t - x^{t+1}\|^2$ is multiplied by w^t, while in b) each $\|x^t - x^{t+1}\|^2$ is multiplied by a rapidly increasing $w^t(t+1)/2$. But the resulting $\sum_{t=0}^{N-1} w^t(t+1)/2\|x^t - x^{t+1}\|^2$ can be compensated by an increasing η^t, so b) is useful for improving convergence rates in strongly convex problems.

Now we show a). Since $w^t\tau_i^t$ is monotonically non-decreasing, the telescoping sum of (4.15) satisfies

$$\begin{aligned}
&\textstyle\sum_{t=0}^{N-1} w^t\mathcal{A}_i^t \\
&\quad\le \Big[w_0\tau_i^0 M_p V_i(\pi_i^0,\pi_i) + \textstyle\sum_{t=1}^{N-1}(w^t\tau_i^t - w^{t-1}\tau_i^{t-1})\pi_p V_i(\pi_i^t,\pi_i)\Big] - w^{N-1}\tau_i^{N-1}\pi_p V_i(\pi_i^N,\pi_i) \\
&\qquad - \textstyle\sum_{t=0}^{N-1} w^t\tau_i^t\pi_p V_i(\pi_i^t,\pi_i^{t+1}) \\
&\quad\le w^{N-1}\tau_i^{N-1}M_p\mathcal{D}_{\Pi_i}^2 - w^{N-1}\tau_i^{N-1}\pi_p V_i(\pi_i^N,\pi_i) - \textstyle\sum_{t=0}^{N-1} w^t\tau_i^t\pi_p V_i(\pi_i^t,\pi_i^{t+1}) \\
&\quad= \tfrac{N\sqrt{N+1}}{4}M_p\mathcal{D}_{\Pi_i}\sigma_{\mathcal{L}_q} + \tfrac{N}{2}\tilde{\mathcal{D}}_{\Pi_i}\mathcal{D}_X - \textstyle\sum_{t=0}^{N-1} w^t\tau_i^t\pi_p V_i(\pi_i^t,\pi_i^{t+1}) - w^{N-1}\tau_i^{N-1}\pi_p V_i(\pi_i^N,\pi_i).
\end{aligned} \tag{4.16}$$

The negative terms in concluding inequality above will be used to cancel out $\mathcal{C}_i^t$, $\Delta_{\pi_i}^t$ and $\mathcal{T}_i^t$. In particular, we split τ_i^t into components according to their roles in handling $\mathcal{C}_i^t$ and $\Delta_{\pi_i}^t$:

$$\tau_{i,\sigma}^t = \tfrac{\sigma_{\mathcal{L}_q}\sqrt{N+1}}{2\mathcal{D}_{\Pi_i}}, \quad \tau_{i,\pi}^t = \tfrac{\mathcal{D}_X M_q}{\mathcal{D}_{\Pi_i}}.$$

First, $\mathcal{T}_i^t$ admits a telescopic cancellation and the remainder can be canceled by $w^{N-1}\tau_{i,\pi}^{N-1}\pi_p V_i(\pi_i^N,\pi_i)$:

$$\begin{aligned}
&\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t\mathcal{T}_i^t - w^{N-1}\tau_{i,\pi}^{N-1}\pi_p V_i(\pi_i^N,\pi_i)] \\
&\quad= \mathbb{E}[-w^{N-1}\pi_p(\pi_i^N - \pi_i)\pi_q^N(\xi_q^i)(x^N - x^{N-1}) - w^{N-1}\tau_{i,\pi}^{N-1}\pi_p V_i(\pi_i^N,\pi_i)] \\
&\quad\overset{(a)}{\le} \tfrac{w^{N-1}}{2}\tfrac{M_p\mathcal{D}_{\Pi_i}M_q}{\mathcal{D}_X}\mathbb{E}[\|x^N - x^{N-1}\|^2] = \tfrac{w^{N-1}}{2}\tfrac{\tilde{\mathcal{D}}_{\Pi_i}}{\mathcal{D}_X}\mathbb{E}[\|x^N - x^{N-1}\|^2],
\end{aligned} \tag{4.17}$$

where (a) follows from $-\pi_p(\pi_i^N - \pi_i)\pi_q^N(\xi_q^i)(x^N - x^{N-1}) \le \|\pi_p(\pi_i^N - \pi_i)\|\|\pi_q^N(\xi_q^i)\|\|x^N - x^{N-1}\|$, $M_p\pi_p V_i(\pi_i^N,\pi_i) \ge \frac{1}{2}\|\pi_p(\pi_i - \pi_i^N)\|^2$ in Proposition 17, the Young's inequality and the conditional independence between x^N and $\pi_q^N(\xi_q^i)$. As for $\mathcal{C}_i^t$, we have

$$\begin{aligned}
\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t(\mathcal{C}_i^t - \tau_{i,\pi}^t V_i(\pi_i^t,\pi_i^{t+1}))] &\le \textstyle\sum_{t=0}^{N-1} w^t(\theta^t)^2\tfrac{\tilde{\mathcal{D}}_{\Pi_i}}{\mathcal{D}_X}\mathbb{E}[\tfrac{1}{2}\|x^t - x^{t-1}\|^2] \\
&\overset{(b)}{\le} \textstyle\sum_{t=0}^{N-2}\tfrac{w^t}{2}\tfrac{\tilde{\mathcal{D}}_{\Pi_i}}{\mathcal{D}_X}\mathbb{E}[\|x^t - x^{t+1}\|^2],
\end{aligned} \tag{4.18}$$

where (b) follows from the relation $w^t\theta^t = w^{t-1}$ and $\theta^{t-1} \le 1$.

Furthermore, since $\mathbb{E}[\|\mathcal{L}(\pi_q^{t+1}(\xi_i),x^t) - \mathcal{L}(\pi_q^{t+1},x^t)\|^2] \le \sigma_{\mathcal{L}_q}^2$, we obtain

$$\begin{aligned}
&\mathbb{E}[\textstyle\sum_{t=0}^{N-1}\Delta_{\pi_i}^t - w^t\tau_{i,\sigma}^t\pi_p V_i(\pi_i^t,\pi_i^{t+1})] \\
&\quad\le \textstyle\sum_{t=0}^{N-1} w^t\mathbb{E}[\|\pi_p(\pi_i^{t+1} - \pi_i^t)\|\|\mathcal{L}(\pi_q^{t+1}(\xi_i),x^t) - \mathcal{L}(\pi_q^{t+1},x^t)\| - w^t\tau_{i,\sigma}^t\tfrac{1}{M_p}\tfrac{1}{2}\|\pi_p(\pi_i^{t+1} - \pi_i^t)\|^2] \\
&\quad\le \textstyle\sum_{t=0}^{N-1} w^t\tfrac{M_p\mathcal{D}_{\Pi_i}\sigma_{\mathcal{L}_q}}{\sqrt{N+1}} = \tfrac{N\sqrt{N+1}}{4}M_p\mathcal{D}_{\Pi_i}\sigma_{\mathcal{L}_q}.
\end{aligned} \tag{4.19}$$

Thus the desired inequality (4.13) follows from substituting (4.15), (4.16), (4.17), (4.18) and (4.19) into

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q_i(z^{t+1},z)) \le \mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t(\mathcal{A}_i^t + \mathcal{T}_i^t + \mathcal{C}_i^t + \Delta_{\pi_i}^t + \Delta_{x,i}^t + \delta_i^t)].$$

In addition, the bound in (4.14) can be derived similarly, but with $\tau_{i,\pi}^t = \frac{2}{(t+1)c}M_p M_q^2$. ■

For the simpler exact arguments case, the next result is an immediate corollary.

Corollary 2 *Let f_i be smoothable with exact arguments. If $\mathcal{D}_{\Pi_i}^2 := \max_{\pi_i,\bar{\pi}_i\in\Pi_i}\|V_i(\pi_i,\bar{\pi}_i)\|$ and $\tilde{\mathcal{D}}_{\Pi_i} := M_p\mathcal{D}_{\Pi_i}M_q$, then the solution sequence $\{z^t := (x^t;\pi_1^t,\pi_2^t,\ldots,\pi_k^t)\}$ generated by Algorithm 3 satisfies the following bounds.*

a) If $\tau_i^t = \frac{\mathcal{D}_X M_q}{\mathcal{D}_{\Pi_i}}$, then

$$\mathbb{E}[\sum_{t=0}^{N-1} w^t Q_i(z^{t+1}, z)] \leq \frac{N}{2}\tilde{\mathcal{D}}_{\Pi_i}\mathcal{D}_X + \sum_{t=0}^{N-1} \frac{w^t}{4}\frac{\tilde{\mathcal{D}}_{\Pi_i}}{\mathcal{D}_X}\mathbb{E}[\|x^{t+1} - x^t\|^2]. \tag{4.20}$$

b) If $\tau_i^t = \frac{2}{(t+1)c} M_p M_q^2$ for some $c > 0$, then

$$\mathbb{E}[\sum_{t=0}^{N-1} w^t Q_i(z^{t+1}, z)] \leq \frac{2}{c}\tilde{\mathcal{D}}_{\Pi_i}^2 + \sum_{t=0}^{N-1} \frac{(t+1)c}{4} w^t \mathbb{E}[\|x^{t+1} - x^t\|^2]. \tag{4.21}$$

4.4 Non-smooth Layer Function

Now we study the non-smooth layer function f_i. Since f_i^* is not accessible explicitly, we choose $D_{f_i^*}$ as the proximity penalty function to implement the implicit proximal update from an associated π_i^t. So other than different stepsizes τ_i^t, the π_i^{t+1} update rule is exactly the same as that for smooth layer functions.

The non-smooth layer function is nearly the most general convex function in that it only needs to be Lipschitz continuous. So it is tempting to treat every layer function as non-smooth for simplicity. However, this does not work because we need a rather strong assumption for the non-smooth layer function to ensure the convergence of Algorithm 3.

Assumption 4 *If f_i is non-smooth, then its argument is exact, i.e., the inner layer functions $f_{i+1}, f_{i+2}, \ldots, f_k$ are all deterministic functions.*

This assumption can be explained in view of the decomposition of Q_i in (3.13). Just like the smooth layer function, the proximal update π_i^{t+1} satisfies

$$\mathcal{A}_i^t \leq \tau_i^t \pi_p D_{f_i^*}(\pi_i^t, \pi_i) - \tau_i^t \pi_p D_{f_i^*}(\pi_i^t, \pi_i^{t+1}) - (\tau_i^t + 1)\pi_p D_{f_i^*}(\pi_i^{t+1}, \pi_i). \tag{4.22}$$

But the Bregman's distance $D_{f_i^*}$ associated with the non-smooth f_i is not necessarily strongly convex, so the stochastic error $\Delta_{\pi_i}^t$ from a noisy argument can not be canceled in our framework. Indeed, it appears that such an assumption is often implicitly used in the literature. For example, in [22] and [24], the authors only consider problems where the innermost layer function f_k can be non-smooth. Since the argument x to f_k is always exact, those problems satisfy Assumption 4 automatically.

Since $D_{f_i^*}$ cannot help us to cancel any terms, we might as well set the penalty parameter τ_i^t to zero. This way of specifying stepsizes leads us to the following convergence result.

Proposition 19 *Let f_i be M_{Π_i}-Lipschitz continuous with exact argument. If $\tilde{M}_{\Pi_i} := M_p M_{\Pi_i} M_q$, then the solution sequence $\{z^t := (x^t; \pi_1^t, \pi_2^t, \ldots, \pi_k^t)\}$ generated by Algorithm 3 with $\tau_i^t = 0$ satisfies the following bounds.*

$$\sum_{t=0}^{N-1} w^t Q_i(z^{t+1}, z) \leq \frac{N\sqrt{N+1}}{2}\tilde{M}_{\Pi_i}\mathcal{D}_X + \sum_{t=0}^{N-1} \frac{w^t\sqrt{N+1}}{2}\frac{\tilde{M}_{\Pi_i}}{\mathcal{D}_X}\|x^{t+1} - x^t\|^2, \tag{4.23}$$

$$\sum_{t=0}^{N-1} w^t Q_i(z^{t+1}, z) \leq \frac{2N}{c}\tilde{M}_{\Pi_i}^2 + \sum_{t=0}^{N-1} \frac{c(t+1)w^t}{4}\|x^{t+1} - x^t\|^2, \quad \forall c > 0. \tag{4.24}$$

Proof Notice that $\tau_i^t := 0$ and (4.22) imply

$$\mathcal{A}_i^t \leq 0 \ \forall t, \text{ and } \mathbb{E}[\sum_{t=0}^{N-1} w^t \mathcal{A}_i^t] \leq 0. \tag{4.25}$$

We now derive (4.23) by using $\sum_{t=0}^{N-1} \frac{w^t\sqrt{N+1}}{2}\frac{\tilde{M}_{\Pi_i}}{\mathcal{D}_X}\|x^{t+1} - x^t\|^2$ to cancel out $\mathcal{C}_i^t$ and $\mathcal{T}_i^t$ in (3.13). For $\mathcal{T}_i^t$, after the telescopic cancellation, the remaining term can be canceled by $\frac{w^{N-1}\sqrt{N+1}}{2}\frac{\tilde{M}_{\Pi_i}}{\mathcal{D}_X}\|x^N - x^{N-1}\|^2$:

$$\begin{aligned}
&\sum_{t=0}^{N-1} w^t \mathcal{T}_i^t - \frac{w^{N-1}\sqrt{N+1}}{2}\frac{\tilde{M}_{\Pi_i}}{\mathcal{D}_X}\|x^N - x^{N-1}\|^2 \\
&\quad = -w^{N-1}\pi_p(\pi_i^N - \pi_i)\pi_q^N(x^N - x^{N-1}) - \frac{w^{N-1}\sqrt{N+1}}{2}\frac{\tilde{M}_{\Pi_i}}{\mathcal{D}_X}\|x^N - x^{N-1}\|^2 \\
&\quad \leq w^{N-1}\frac{\mathcal{D}_X}{\sqrt{N+1}\tilde{M}_{\Pi_i}}\frac{4\tilde{M}_{\Pi_i}^2}{2} = \frac{2w^{N-1}}{\sqrt{N+1}}\mathcal{D}_X\tilde{M}_{\Pi_i}.
\end{aligned} \tag{4.26}$$

Then $\mathcal{C}_i^t$ can be handled by the unused $\sum_{t=0}^{N-2} \frac{w^t\sqrt{N+1}}{2}\frac{\tilde{M}_{\Pi_i}}{\mathcal{D}_X}\|x^{t+1} - x^t\|^2$:

$$\begin{aligned}
&\sum_{t=0}^{N-1} w^t \mathcal{C}_i^t - \sum_{t=0}^{N-2} \frac{w^t\sqrt{N+1}}{2}\frac{\tilde{M}_{\Pi_i}}{\mathcal{D}_X}\|x^{t+1} - x^t\|^2 \\
&\quad = \sum_{t=1}^{N-1} w^{t-1}(\pi_p(\pi_i^{t+1} - \pi_i^t)\pi_q^{t+1}(x^t - x^{t-1}) - \frac{w^t\sqrt{N+1}}{2}\frac{\tilde{M}_{\Pi_i}}{\mathcal{D}_X}\|x^t - x^{t-1}\|^2) \\
&\quad \leq \sum_{t=0}^{N-2} \frac{2w^t}{\sqrt{N+1}}\mathcal{D}_X\tilde{M}_{\Pi_i}.
\end{aligned} \tag{4.27}$$

Substituting (4.25), (4.26) and (4.27) into $\sum_{t=0}^{N-1} w^t Q_i(z^{t+1}, z)$ and noting that $\sum_{t=0}^{N-1} \frac{2w^t}{\sqrt{N+1}} \mathcal{D}_X \tilde{M}_{\Pi_i} = \frac{1}{2} N\sqrt{N+1} \mathcal{D}_X \tilde{M}_{\Pi_i}$, we have (4.23). The derivation of bound (4.24) follows similarly except that

$$\sum_{t=0}^{N-1} w^t \mathcal{T}_i^t - \tfrac{cN}{2} w^{N-1} \tfrac{1}{2} \|x^N - x^{N-1}\|^2 \le \tfrac{2w^{N-1}}{cN} \tfrac{4\tilde{M}_{\Pi_i}^2}{2} = \tfrac{2}{c} \tilde{M}_{\Pi_i}^2,$$
$$\sum_{t=0}^{N-1} w^t \mathcal{C}_i^t - \sum_{t=0}^{N-2} \tfrac{c(t+1)}{2} \tfrac{w^t}{2} \|x^{t+1} - x^t\|^2 \le \tfrac{2(N-1)}{c} \tilde{M}_{\Pi_i}^2.$$

■

4.5 Separable Mixture Layer Function

Starting from this subsection, we move on to layer functions constructed from other functions of different types, referred to as sub-functions. Of course, one could ignore the mixture structure, assign the weakest type to the entire function, and design an algorithm accordingly. But our SSD framework can exploit certain mixture structures to obtain better convergence results. Two such structures will be analyzed over this and the next subsection.

In this subsection, we focus on separable mixture function f_i where parts of the dual variable π_i can be updated independently according to the type of associated sub-functions. First, we consider the most straightforward structure, *output mixing*, where the vector function f_i can decomposed into sub-vector functions of different types. In other words, the function f_i, the dual variable π_i and the conjugate function f_i^* admit a row partitioning:

$$f_i(y_i) := \begin{bmatrix} f_{i,p_1}(y_i) \\ \hline f_{i,p_2}(y_i) \\ \hline \vdots \\ \hline f_{i,p_L}(y_i) \end{bmatrix}^{❷}, \quad \pi_i := \begin{bmatrix} \pi_{i,p_1} \\ \hline \pi_{i,p_2} \\ \hline \vdots \\ \hline \pi_{i,p_L} \end{bmatrix}, \quad f_i^*(\pi_i) := \begin{bmatrix} f_{i,p_1}^*(\pi_{i,p_1}) \\ \hline f_{i,p_2}^*(\pi_{i,p_2}) \\ \hline \vdots \\ \hline f_{i,p_L}^*(\pi_{i,p_L}) \end{bmatrix}.$$

So we can update each π_{i,p_l} using one of the update rules developed in the preceding subsections according to the type of f_{i,p_l}. Now we develop the Q_i convergence bound. If we separate the outer layers' reference vector π_p accordingly, i.e., $\pi_p := [\pi_{p_1} | \pi_{p_2} | \dots | \pi_{p_L}]$, then the gap function Q_i becomes separable,

$$Q_i(z^{t+1}, z) = \sum_{l=1}^{L} Q_{i,l}(z^{t+1}, z),$$

where

$$Q_{i,l}(z^{t+1}, z) := \pi_{p_l} \left((\pi_{i,p_l} - \pi_{i,p_l}^{t+1}) \mathcal{L}_q(x^{t+1}; \pi_q^{t+1}) - [f_{p_l}^*(\pi_{i,p_l}) - f_{p_l}^*(\pi_{i,p_l}^{t+1})] \right).$$

As a consequence, the bounds for $\mathbb{E}[\sum_{t=0}^{N-1} w^t Q_{i,l}(z^{t+1}, z)]$ are the same as those developed before; we just have to replace the layer function constants, M_p, M_{Π_i}, L_{f_i}, and $\mathcal{D}_{\Pi_i}$, with the corresponding sub-function constants, M_{p_l}, $M_{\Pi_{i,l}}$, $L_{f_{i,l}}$, and $\mathcal{D}_{\Pi_{i,l}}$. Moreover, if there are affine sub-functions, we can relax the compositional monotonicity in Assumption 2, $\pi_p \ge 0$, to requiring $\pi_{i,p_l} \ge 0$ only for non-affine f_{i,p_l}.

Next, we consider the so-called *separable input mixing*, which means there exists some partitioning $y_i := [y_{i,q_1} | y_{i,q_2} | \dots | y_{i,q_L}]$ such that

$$f_i(y_i) = \sum_{l=1}^{L} f_{i,q_l}(y_{i,q_l}).$$

It is clear that both π_i and f_i^* allow decomposition as follows:

$$\pi_i := [\pi_{i,q_1} | \pi_{i,q_2} | \dots | \pi_{i,q_L}], \quad f_i^*(\pi_i) := \sum_{l=1}^{L} f_{i,q_l}^*(\pi_{i,q_l}).$$

Thus the gap function Q_i is also separable, $Q_i(z^{t+1}, z) = \sum_{l=1}^{L} Q_{i,j}(z^{t+1}, z)$, where

$$Q_{i,l}(z^{t+1}, z) := \pi_i \left[(\pi_{i,q_l} - \pi_{i,q_l}^{t+1}) \mathcal{L}_{q_l}(x^{t+1}; \pi_q^{t+1}) - [f_{q_l}^*(\pi_{i,p_l}) - f_{q_l}^*(\pi_{i,p_l}^{t+1})] \right].$$

Therefore, we can also update each π_{i,q_l} separately using the previous update rules according to the type of f_{i,q_l}. Moreover, the bound on Q_i is just the sum of individual convergence bounds on $Q_{i,j}$.

❷ The Ξ (or [|]) notation merely represents that the non-overlapping sub-vector functions $f_{i,p_1}, f_{i,p_2}, \dots, f_{i,p_L}(y_i)$ cover all components of f_i, i.e. $p_1 \cup p_2 \cup \dots \cup p_L = 1, 2, 3, \dots, m_{i-1}$. But it should not be interpreted as f_{i,p_l} being a consecutive block of f_i. This general subset notation will be used in the sequel for simplicity.

We consider the semi-smooth function with a semi-noisy argument, i.e., there exists some partitioning of the argument, $y_i := [y_{i,S}|y_{i,N}]$, such that $y_{i,S}$ is noisy and $\nabla_{y_{i,S}} f_i([y_{i,S}|y_{i,N}])$ is $L_{f_{i,S}}$-Lipschitz continuous with respect to $y_{i,S}$ for any fixed $y_{i,N}$, while $y_{i,N}$ is exact and $\|\pi_{i,N}\| \leq M_{\Pi_i,N}$ for all $\{\pi_{i,N} : [\pi_{i,S}|\pi_{i,N}] \in \partial f_i(y_i)\}$. In other words, the dual variable π_i and the input $\mathcal{L}_q(x;\pi_q)$ can be separated into

$$\pi_i = [\pi_{i,S}|\pi_{i,N}], \quad \mathcal{L}_q(x;\pi_q) = \begin{bmatrix} \mathcal{L}_{q,S}(x;\pi_q) \\ \mathcal{L}_{q,N}(x;\pi_q) \end{bmatrix}.$$

However, since we do not assume $f_i(y_i) \equiv f_{i,S}(y_{i,S}) + f_{i,N}(y_{i,N})$, the conjugate function is not separable and separate proximal updates for $\pi_{i,S}$ and $\pi_{i,N}$ is not possible. Instead we update them jointly with the following implicit proximal update,

$$\pi_i^{t+1} \in \underset{\pi_i \in \Pi_i}{\arg\min} -\pi_i \bar{y}_i^{t+1} + f_i^*(\pi_i) + \tau_i^t D_{f_i^*}(\pi_i^t, \pi_i),$$

which has the same implementation as that of the smooth layer function in Subsection 4.2

Now we derive the convergence bounds of Q_i. First, we need to determine the strong convexity modulus of $\pi_p D_{f_i^*}(\pi_i^t, \pi_i^{t+1})$ with respect to $\|\pi_p(\pi_{i,S}^t - \pi_{i,S}^{t+1})\|$, which is found in the next proposition.

Proposition 20 *Let the convex mapping $g(y_S, y_N)$ be semi-smooth, i.e., $\|\nabla_{y_S} g(y_S, y_N) - \nabla_{y_S} g(\bar{y}_S, y_N)\| \leq L_{g_S}\|y_S - \bar{y}_S\|\ \forall y_S, y_N, \bar{y}_S$. Then given a non-negative weight vector w, the following relation holds for any $\pi := (\pi_N, \pi_S) \in \partial g(y)$ and $\bar{\pi} := (\bar{\pi}_N, \bar{\pi}_S) \in \partial g(\bar{y})$:*

$$\|w\| w^\top D_{g^*}(\bar{\pi}, \pi) \geq \tfrac{1}{2L_{g_S}} \|w^\top(\pi_S - \bar{\pi}_S)\|^2. \tag{4.28}$$

Proof We first show (4.28) for a one-dimensional g.

Claim If the convex $h : \mathbb{R}^n \to \mathbb{R}$ is semi-smooth with constant L_{h_S}, then $D_{h^*}(\bar{\pi}, \pi) \geq \frac{1}{2L_{h_S}}\|w(\pi_S - \bar{\pi}_S)\|^2$ for sub-gradients $\bar{\pi} \in \partial h(\bar{y}), \pi \in h(y)$.

The following analysis is an extension of Lemma 5.8 in [13]. Let $\phi(y) := h(y) - h(\bar{y}) - \langle \bar{\pi}, y - \bar{y}\rangle$ and let $\phi_S(y)$ denote $\nabla_{y_S}\phi(y)$. Clearly, $\|\phi_S(y_N, y_S) - \phi_S(y_N, \hat{y}_S)\| \leq L_{h_S}\|y_S - \hat{y}_S\|$. Moreover, since $0 \in \partial\phi(\bar{y})$, the convexity of h implies that $\phi(\bar{y}) \leq \phi(\hat{y}) \forall \hat{y}$. Therefore,

$$\begin{aligned}
\phi(\bar{y}) &\leq \phi(y - \tfrac{1}{L_{h_S}}[0, \phi_S(y)]) \\
&= \phi(y) + \int_0^1 \langle \phi_S(y - \tfrac{\tau}{L_{h_S}}[0, \phi_S(y)]), -\tfrac{1}{L_{h_S}}\phi_S(y)\rangle d(\tau) \\
&= \phi(y) - \tfrac{1}{L_{h_S}}\|\phi_S(y)\|^2 + \int_0^1 \langle \phi_S(y - \tfrac{\tau}{L_{h_S}}[0, \phi_S(y)]) - \phi_S(y), -\tfrac{1}{L_{h_S}}\phi_S(y)\rangle d(\tau) \\
&\leq \phi(y) - \tfrac{1}{L_{h_S}}\|\phi_S(y)\|^2 + \int_0^1 \tau\|\phi_S(y)\|\|\tfrac{1}{L_{h_S}}\phi_S(y)\| d(\tau) \\
&= \phi(y) - \tfrac{1}{2L_{h_S}}\|\phi_S(y)\|^2.
\end{aligned}$$

Noting $\phi(\bar{y}) = 0$, we have $\frac{1}{2L_{h_S}}\|\bar{\pi}_S - \pi_S\|^2 \leq \phi(y) = D_h(\bar{y}, y) = D_{h^*}(\pi, \bar{\pi})$.

Next, the extension to multi-dimensional g is similar to that of Proposition 15. ∎

Now we are ready to prove the Q_i bounds.

Proposition 21 *Let f_i be a semi-smooth function with a partial smoothness constant of $L_{f_{i,S}}$ and a partial Lipschitz constant of $M_{\Pi_i,N}$ (for the non-smooth part). Moreover, let f_i's argument be semi-noisy with constant $M_{q,S}^2 := \max_{\pi_q \in \Pi_q} \mathbb{E}[\|\pi_{q,S}(\xi_q^i)\|^2]$ and $M_{q,N} := \max_{\pi_q \in \Pi_q}\|\pi_{q,N}\|$. If $\tilde{L}_{f_{i,S}} := M_p L_{f_{i,s}} M_{q,S}^2$, $\tilde{M}_{\Pi_i,N} := M_p \tilde{M}_{\Pi_i,N} M_{q,N}$ and $STC_i := \mathbb{E}(\sum_{t=0}^{N-1} w^t(\Delta_{x,i}^t + \delta_i^t))$, then the solution sequence $\{z^t := (x^t; \pi_1^t, \ldots, \pi_k^t)\}$ generated by Algorithm 3 with $\tau_i^t := \frac{t+1}{6} + \frac{t}{4}$ satisfies*

$$\begin{aligned}
\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q_i(z^{t+1}, z)] \leq& \tfrac{1}{6} M_p \mathcal{D}_{\Pi_i}^2 + \tfrac{3N}{2} M_p L_{f_{i,s}} \sigma_{\mathcal{L}_q}^2 + \textstyle\sum_{t=0}^{N-1}(\frac{2\tilde{L}_{f_{i,S}}}{t+1} + \frac{\sqrt{N+1}\tilde{M}_{\Pi_i,N}}{2\mathcal{D}_X})\|x^t - x^{t+1}\|^2 \\
&+ \tfrac{N\sqrt{N+1}}{2}\tilde{M}_{\Pi_i,N}\mathcal{D}_X + STC_i.
\end{aligned} \tag{4.29}$$

$$\begin{aligned}
\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q_i(z^{t+1}, z)] \leq& \tfrac{1}{6} M_p \mathcal{D}_{\Pi_i}^2 + \tfrac{3N}{2} M_p L_{f_{i,s}} \sigma_{\mathcal{L}_q}^2 + \textstyle\sum_{t=0}^{N-1}(\frac{\tilde{L}_{f_{i,S}}}{t+1} + \frac{c(t+1)}{4})\|x^t - x^{t+1}\|^2 \\
&+ \tfrac{2N}{c}\tilde{M}_{\Pi_i,N}^2 + STC_i,\ \forall c > 0.
\end{aligned} \tag{4.30}$$

Proof Since

$$
\begin{aligned}
Q_i(z^{t+1}, z) :=& \pi_p \left[(\pi_{i,S} - \pi_{i,S}^{t+1}) \mathcal{L}_{q,S}(x^{t+1}; \pi_q^{t+1}) + (\pi_{i,N} - \pi_{i,N}^{t+1}) \mathcal{L}_{q,N}(x^{t+1}; \pi_q^{t+1}) \right] \\
& - \pi_p [f_i^*(\pi_i) - f_i^*(\pi_i^{t+1})], \\
\mathcal{A}_i^t \ \ :=& \pi_p \left[(\pi_{i,N} - \pi_{i,N}^{t+1}) \bar{y}_{i,N}^{t+1} + (\pi_{i,S} - \pi_{i,S}^{t+1}) \bar{y}_{i,S}^{t+1} - (f_i^*(\pi_i) - f_i^*(\pi_i^{t+1})) \right] \\
\leq & \tau_i^t \pi_p D_{f_i^*}(\pi_i^t, \pi_i) - \tau_i^t \pi_p D_{f_i^*}(\pi_i^t, \pi_i^{t+1}) - (\tau_i^t + 1) \pi_p D_{f_i^*}(\pi_i^{t+1}, \pi_i),
\end{aligned}
$$

we can group the terms into those related to $\pi_{i,S}$ and $\pi_{i,N}$ respectively, i.e., $Q_i(z^{t+1}, z) = \mathcal{A}_i^t + Q_{i,S}^t + Q_{i,N}^t$, where

$$
\begin{aligned}
Q_{i,S}^t &:= \pi_p (\pi_{i,S} - \pi_{i,S}^{t+1}) \left\{ \mathcal{L}_{q,S}(x^{t+1}; \pi_q^{t+1}) - [\mathcal{L}_{q,S}(x^t; \pi_q^{t+1}(\xi_q^i)) + \theta^t \pi_{q,S}^t(\xi_q^i)(x^t - x^{t-1})] \right\}, \\
Q_{i,N}^t &:= \pi_p (\pi_{i,N} - \pi_{i,N}^{t+1}) \left\{ \mathcal{L}_{q,N}(x^{t+1}; \pi_q^{t+1}) - [\mathcal{L}_{q,N}(x^t; \pi_q^{t+1}(\xi_q^i)) + \theta^t \pi_{q,N}^t(\xi_q^i)(x^t - x^{t-1})] \right\}.
\end{aligned}
$$

Clearly, they can be decomposed further into

$$
\begin{aligned}
Q_{i,S}^t &= \mathcal{C}_{i,S}^t + \mathcal{T}_{i,S}^t + \delta_{i,S}^t + \Delta_{\pi_{i,S}}^t + \Delta_{x,i,S}^t, \\
Q_{i,N}^t &= \mathcal{C}_{i,N}^t + \mathcal{T}_{i,N}^t,
\end{aligned}
$$

where the individual error terms are defined in a similar fashion as those in (3.13). So we can just apply Proposition 19 to bound $\mathbb{E}[\sum_{t=0}^{N-1} w^t Q_{i,N}^t]$ and apply Proposition 16 and (4.28) to bound

$$
\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t \left(Q_{i,S}^t - \tau_i^t \pi_p D_{f_i^*}(\pi_i^t, \pi_i) - \tau_i^t \pi_p D_{f_i^*}(\pi_i^t, \pi_i^{t+1}) - (\tau_i^t + 1) \pi_p D_{f_i^*}(\pi_i^{t+1}, \pi_i) \right)].
$$

Then the desired convergence bound for Q_i is simply the sum of the smooth bound (4.6) (with L_{f_i} and M_q replaced by $L_{f_{i,s}}$ and $M_{q,S}$) and the non-smooth bounds in (4.23) or (4.24) (with M_π and M_q replaced by $M_{\Pi_i,N}$ and $M_{q,N}$).

■

We remark here that the above stepsizes τ_i^t are the same as those in Proposition 16, so the non-smooth part, $\pi_{i,N}$, is also updated using the smooth rule. Compared with the separable input mixing layer discussed in the preceding subsection, the difference is that $D_{f_i^*}$, rather than $\mathcal{D}_{f_{i,S}^*}$, appears in the convergence bounds (4.29) and (4.30).

4.7 Summary of Modular Updates

For easy reference, we summarize all the convergence bounds on Q_i into an abstract form,

$$
\mathbb{E}(\textstyle\sum_{t=0}^{N-1} w^t Q_i(z^{t+1}, z)) \leq \mathrm{CST}_i + \sum_{t=0}^{N-1} (H_{i,-}^t + H_{i,+}^t + H_{i,\alpha}^t) \left(\frac{w^t}{2} \|x^{t+1} - x^t\|^2 \right) + \mathrm{STC}_i, \tag{4.31}
$$

where the coefficient for $\frac{w^t}{2}\|x^{t+1} - x^t\|^2$ are split further into three categories: H_α^t if it is some multiple of $\frac{t+1}{2}$ (useful for strongly convex problems), H_-^t if $w^t H_-^t$ is monotonically non-increasing and H_+^t if $w^t H_+^t$ is monotonically non-decreasing. The detailed coefficients are shown in Table 3. Such an abstract representation can simplify the derivation of the overall convergence rate in the next section.

5 Convergence Analysis

In this section, we study the convergence properties of Algorithm 3 by combining the preceding Q_i bounds with Q_0 bounds from some x-updates. So we will reuse a few constants defined previously, including $w^t := (t+1)/2$, $\theta^t := (t)/(t+1)$, the compact Q_i, M_p, M_q and $\sigma_{\mathcal{L}_q}$ notation associated with (4.2), and the noise bounds $\sigma_{\mathcal{L}_i}^2$ and $\sigma_{\pi_{i:}}^2$ defined in Proposition 13. Moreover, it is convenient to abstract away from the details about the layer functions by summing up Q_i in (4.31). More specifically, if we set $Q_{1:} := \sum_{i=1}^k Q_i$, then

$$
\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q_{1:}(z^{t+1}, z)] \leq \mathrm{CST} + \sum_{t=0}^{N-1} \frac{w^t}{2} (H_+^t + H_\alpha^t + H_-^t) \|x^{t+1} - x^t\|^2 + \mathrm{STC}, \tag{5.1}
$$

where $\mathrm{CST} := \sum_{i=1}^k \mathrm{CST}_i$, $H_+^t := \sum_{i=1}^k H_{i,+}^t$, $H_-^t := \sum_{i=1}^k H_{i,-}^t$, $H_\alpha^t := \sum_{i=1}^k H_{i,\alpha}^t$, and $\mathrm{STC} := \sum_{i=1}^k \mathrm{STC}_i$.

Table 3: Summary of the Selection of Stepsizes and Bounds on $\mathbb{E}(\sum_{t=0}^{N-1} w^t Q_i(z^{t+1}, z))$

Arg Type	Stepsize	CST_i	$H^t_{i,-}$	$H^t_{i,+}$	$H^t_{i,\alpha}$	STC_i?
			Stochastic/Deterministic Affine			
Exct ($\alpha > 0$)	0					
Nsy ($\alpha > 0$)	0					
			Deterministic/Stochastic Smooth f_i with $\tilde{L}_{f_i} = M_p L_{f_i} M_q^2$			
Exct ($\alpha > 0$)	$\frac{t}{2}$		$\frac{2}{t+1}\tilde{L}_{f_i}$			
Nsy ($\alpha > 0$)	$\frac{t}{4} + \frac{t+1}{6}$	$\frac{1}{6}M_p\mathcal{D}^2_{f_i} + \frac{3}{2}NM_pL_{f_i}\sigma^2_{\mathcal{L}_q}$	$\frac{4}{t+1}\tilde{L}_{f_i}$			✓
			(Deterministic) Smoothable f_i with $\tilde{\mathcal{D}}_{\Pi_i} = M_p \mathcal{D}_{\Pi_i} M_q$			
Exct	$\frac{\mathcal{D}_X M_q}{\mathcal{D}_{\Pi_i}}$	$\frac{N}{2}\tilde{\mathcal{D}}_{\Pi_i}\mathcal{D}_X$		$\frac{\tilde{\mathcal{D}}_{\Pi_i}}{\mathcal{D}_X}$		
Exct $\alpha > 0$	$\frac{2}{(t+1)c}M_pM_q^2$	$\frac{2}{c}\tilde{\mathcal{D}}^2_{\Pi_i}$			$\frac{c(t+1)}{2}$	
Nsy	$\frac{\mathcal{D}_X M_q}{\mathcal{D}_{\Pi_i}} + \frac{\sqrt{N+1}\sigma_{\mathcal{L}_q}}{2\mathcal{D}_{\Pi_i}}$	$\frac{N}{2}\tilde{\mathcal{D}}_{\Pi_i}\mathcal{D}_X + \frac{N\sqrt{N+1}M_p\mathcal{D}_{\Pi_i}\sigma_{\mathcal{L}_q}}{2}$		$\frac{\tilde{\mathcal{D}}_{\Pi_i}}{\mathcal{D}_X}$		✓
Nsy $\alpha > 0$	$\frac{2M_pM_q^2}{(t+1)c} + \frac{\sqrt{N+1}\sigma_{\mathcal{L}_q}}{2\mathcal{D}_{\Pi_i}}$	$\frac{2}{c}\tilde{\mathcal{D}}^2_{\Pi_i} + \frac{N\sqrt{N+1}M_p\mathcal{D}_{\Pi_i}\sigma_{\mathcal{L}_q}}{2}$			$\frac{c(t+1)}{2}$	✓
			Deterministic/Stochastic Non-Smooth f_i with $\tilde{M}_{\Pi_i} = M_p M_{\Pi_i} M_q$			
Exct	0	$\frac{N\sqrt{N+1}}{2}\tilde{M}_{\Pi_i}\mathcal{D}_X$		$\frac{\sqrt{N+1}\tilde{M}_{\Pi_i}}{\mathcal{D}_X}$		
Exct $\alpha > 0$	0	$\frac{2N}{c}\tilde{M}^2_{\Pi_i}$			$\frac{c(t+1)}{2}$	
			Semi-Smooth/Semi-Nsy f_i with $\tilde{L}_{f_i,S} = M_p L_{f_{i,s}} M^2_{q,S}$, $\tilde{M}_{\Pi_i,N} = M_p M_{\Pi_i,N} M_{q,N}$.			
Semi-Nsy	$\frac{t}{4} + \frac{t+1}{6}$	$\frac{1}{6}M_p\mathcal{D}^2_{\Pi_i} + \frac{3}{2}NM_pL_{f_{i,s}}\sigma^2_{\mathcal{L}_q}$	$\frac{4}{t+1}\tilde{L}_{f_i,S}$			✓
		$+\frac{N\sqrt{N+1}}{2}\tilde{M}_{\Pi_i,N}\mathcal{D}_X$		$+\frac{\sqrt{N+1}\tilde{M}_{\Pi_i,N}}{\mathcal{D}_X}$		
Semi-Nsy	$\frac{t}{4} + \frac{t+1}{6}$	$\frac{1}{6}M_p\mathcal{D}^2_{\Pi_i} + \frac{3}{2}NM_pL_{f_{i,s}}\sigma^2_{\mathcal{L}_q}$	$\frac{4}{t+1}\tilde{L}_{f_i,S}$			✓
$\alpha > 0$		$+\frac{2N}{c}\tilde{M}^2_{\Pi_i,N}$			$+\frac{c(t+1)}{2}$	

$\mathbb{E}(\sum_{t=0}^{N-1} w^t Q_i(z^{t+1}, z)) \leq \text{CST}_i + \sum_{t=0}^{N-1} \frac{w^t}{2}(H^t_{i,-} + H^t_{i,+} + H^t_{i,\alpha})\|x^{t+1} - x^t\|^2 + \text{STC}_i$, with $\text{STC}_i := \mathbb{E}(\sum_{t=0}^{N-1} w^t(\Delta^t_{x,i} + \delta^t_i))$.

In our analysis, we will make the distinction between *deterministic problems*, which only have deterministic layer functions, and *stochastic problems*, which have at least one stochastic layer function, because the simpler proofs for deterministic problems can serve as stepping stones for analyzing the more complicated stochastic problems. Moreover, general convex problems ($\alpha = 0$) and strongly convex problems ($\alpha > 0$) will be considered separately. Thus this section will have four subsections for deterministic convex problems, stochastic convex problems, deterministic strongly convex problems and stochastic strongly convex problems, respectively. In each subsection, we will develop a general convergence result using the abstract bound (5.1), and then illustrate it with a concrete example.

5.1 Deterministic Convex Problems

We first consider the simplest case, the deterministic convex problem. The next convergence result follows from choosing $\eta^t := H^t_+ + H^t_-$.

Theorem 3 *If the solution sequence $\{z^t := (x^t; \pi^t_1, \pi^t_2, \ldots, \pi^t_k)\}$ is generated by Algorithm 3 with $\eta^t := H^t_- + H^t_+$ and τ^t_is specified according to Table 3, then for any feasible $z := (x^*; \pi_1, \ldots, \pi_k)$,*

$$\textstyle\sum_{t=0}^{N-1} w^t Q(z^{t+1}, z) \leq \frac{w^0}{2} H^0_- \|x^0 - x^*\|^2 + w^{N-1} H^{N-1}_+ \mathcal{D}^2_X + CST. \tag{5.2}$$

Moreover, if $\bar{x}^N = \sum_{t=0}^{N-1} w^t x^{t+1} / \sum_{t=0}^{N-1} w^t$ with $w^t = (t+1)/2$, then

$$f(\bar{x}^N) - f(x^*) \leq \tfrac{4}{N(N+1)} CST + \tfrac{1}{N(N+1)} H^0_- \|x^0 - x^*\|^2 + \tfrac{2}{N+1} H^{N-1}_+ \mathcal{D}^2_X. \tag{5.3}$$

Proof Since no stochastic estimators are used and $\alpha = 0$, (5.1) can be simplified to

$$\textstyle\sum_{t=0}^{N-1} w^t Q_{1:}(z^{t+1}, z) \leq \text{CST} + \sum \frac{w^t}{2}(H^t_+ + H^t_-)\|x^{t+1} - x^t\|^2, \tag{5.4}$$

Layer Type	Smooth	Smoothable	Non-smooth
Optimal Complexity	$\mathcal{O}\{\frac{\sqrt{\tilde{L}_{f_i}}\|x_0-x^*\|}{\sqrt{\epsilon}}\}$	$\mathcal{O}\{\frac{\tilde{\mathcal{D}}_{\Pi_i}\mathcal{D}_X}{\epsilon}\}$	$\mathcal{O}\{\frac{\tilde{M}^2_{\Pi_k}\mathcal{D}^2_X}{\epsilon^2}\}$

Table 4: Optimal Iteration Complexity for One-layer Problem (5.7)

and Q_0 in (3.14) satisfies

$$w^t Q_0(z^{t+1}, z) = w^t \mathcal{A}_0^t \le \tfrac{1}{2}\left(w^t\eta^t\|x^t - x\|^2 - w^t\eta^t\|x^{t+1} - x\|^2 - w^t\eta^t\|x^{t+1} - x^t\|^2\right).$$

Since $w^t H_+^t \ge w^{t-1}H_+^{t-1}$ and $w^t H_-^t \le w^{t-1}H_-^{t-1}$, the w^t-weighted sum of $\{Q_0(z^{t+1}, z)\}$ satisfies

$$\textstyle\sum_{t=0}^{N-1} w^t Q_0(z^{t+1}, z) \le \frac{w^0}{2} H_-^0\|x^0 - x^*\|^2 + w^{N-1}H_+^{N-1}\mathcal{D}_X^2 - \sum \frac{w^t}{2}(H_-^t + H_+^t)\|x^{t+1} - x^t\|^2.$$

Then adding the above inequality to (5.4), (5.2) follows immediately.

Next we can set z in (5.2) to $\hat{z} := (x^*; \hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_k)$ such that (5.3) is a direct consequence of Proposition 11.

■

We demonstrate the implication of the above result by applying Algorithm 3 to solve

$$\min_{x\in X}\{f(x) := f_1 \circ f_2 \circ \ldots \circ f_k(x) + u(x)\}, \tag{5.5}$$

where f_k is non-smooth, a subset of the layer functions $\{f_i\}_{i\in P}$ are smoothable and the remaining layer functions $\{f_i\}_{i\in S}$ are smooth. Clearly, Theorem 3 implies that Algorithm 3 has an iteration complexity of

$$\mathcal{O}\left\{\frac{\sqrt{\sum_{i\in S}\tilde{L}_{f_i}}\|x_0-x^*\|}{\sqrt{\epsilon}} + \frac{\sum_{i\in P}\tilde{\mathcal{D}}_{\Pi_i}\mathcal{D}_X}{\epsilon} + \frac{\tilde{M}^2_{\Pi_k}\mathcal{D}^2_X}{\epsilon^2}\right\}. \tag{5.6}$$

In addition, (5.6) can be interpreted as the sum of optimal iteration complexities for solving the simplified one-layer problems

$$\min_{x\in X}\; b_i^\top f_i(A_i x + c_i) + \tfrac{1}{k}u(x)\;\forall i \in [k], \tag{5.7}$$

where b_i and (A_i, c_i) are some linearization of $f_{1:i-1}$ and $f_{i+1:}$. More specifically, if the layer function f_i is smooth, smoothable or non-smooth, then $\tilde{L}_{f_i}$, $\tilde{\mathcal{D}}_{\Pi_i}$ and $\tilde{M}_{\Pi_k}$ (for f_k) are the corresponding Lipschitz-smoothness constant, dual radius and Lipschitz continuity constant for $b_i^\top f_i(A_i x + c_i)$, so the optimal iteration complexity for solving it is shown in Table 4.

Moreover, if we ignore the layered composition structure in (5.5) and treat the entire function as non-smooth, then the optimal iteration complexity is $\mathcal{O}\{(M_{\Pi_1}M_{\Pi_2}\ldots M_{\Pi_k})^2\mathcal{D}_X^2/\epsilon^2\}$. But since $\|\hat{\pi}_1\ldots\hat{\pi}_{k-1}\|M_{\Pi_k} \le M_P M_{\Pi_k} := \tilde{M}_{\Pi_k} \le M_{\Pi_1}M_{\Pi_2}\ldots M_{\Pi_k}$ for any feasible $\hat{z} := (x^*; \hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_k)$, the iteration complexity of Algorithm 3 in (5.6) is never worse than that of the naive approach. Indeed, $\tilde{M}_{\Pi_k} := M_P M_{\Pi_k}$ is usually much smaller than $M_{\Pi_1}M_{\Pi_2}\ldots M_{\Pi_k}$. In addition, the complexiy bound in (5.6) also much weaker dependence on $\tilde{L}_{f_i}$, $\tilde{\mathcal{D}}_{\Pi_i}$ (see [9] for relevant discussions). So the iteration complexity in (5.6) often improves over the naive approach.

5.2 Stochastic Convex Problems

Next, we move on to stochastic convex problems with $\alpha = 0$. Clearly, both the primal gap function Q_0 and the aggregate gap function $Q_{1:}$ have additional stochastic error terms resulting from noisy arguments:

$$Q_0(z^{t+1}, z) \le \mathcal{A}_0^t + \varDelta_{x,0}^t + \delta_0^t, \tag{5.8}$$

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q_{1:}(z^{t+1}, z)] \le \text{CST} + \sum_{t=0}^{N-1}(H_+^t + H_-^t)\mathbb{E}\left[\frac{w^t}{2}\|x^{t+1} - x^t\|^2\right] + \mathbb{E}[\sum_{t=0}^{N-1} w^t \sum_{i=1}^{k-1}\left(\delta_i^t + \varDelta_{x,i}^t\right)]. \tag{5.9}$$

Note that the last summation in (5.9) ends at $k-1$ because the innermost layer f_k always has exact argument. Moreover, since the solution sequence $\{z^t := (x^t; \pi_1^t, \ldots, \pi_k^t)\}$ generated by Algorithm 3 is stochastic, the reference point z required by Proposition 11 for bounding functional optimality gap is a random vector,

$\hat{z} := (x^*; \hat{\pi}_1, \ldots, \hat{\pi}_k)$, where $(\hat{\pi}_1, \ldots, \hat{\pi}_k)$ is dependent on $\{z^t\}$. Taken together, those extra stochastic terms become:

$$\delta^t := \textstyle\sum_{i=0}^{k-1} \delta_i^t = \sum_{i=1}^{k-1} \hat{\pi}_p(\hat{\pi}_i - \pi_i^t)[\mathcal{L}_q(x^t; \pi_q^{t+1}) - \mathcal{L}_q(x^t; \pi_q^{t+1}(\xi_q^i))] + (\pi_{1:}^{t+1} - \pi_{1:}^{t+1}(\xi_{1:}^0))(x^t - x), \tag{5.10}$$

$$\varDelta_x^t := \textstyle\sum_{i=0}^{k-1} \varDelta_{x,i}^t = \left\{[\sum_{i=1}^{k-1} \hat{\pi}_p(\hat{\pi}_i - \pi_i^{t+1})(\pi_q^{t+1} - \pi_q^{t+1}(\xi_q^i))] + [\pi_{1:}^{t+1} - \pi_{1:}^{t+1}(\xi_{1:}^0)]\right\}(x^{t+1} - x^t). \tag{5.11}$$

Hence we need the following bounds on δ^t and $\varDelta_x^t$.

Lemma 3 *Let the solution sequence $\{z^t\}$ be generated by Algorithm 3 and let $\hat{z}$ be the random reference point. If*

$$\tilde{\sigma}_\varDelta := \textstyle\sum_{i=1}^{k-1} M_p M_{\Pi_i} \sigma_{\mathcal{L}_q},\ \tilde{\sigma}_\Pi^2 := 4\sum_{i=1}^{k-1} M_p^2 M_{\Pi_i}^2 \sigma_{\pi_{i+1:}}^2 \le \sum_{i=1}^{k} (4i) M_{\Pi_{1:i-1}}^2 \sigma_{\pi_i}^2 M_q^2,\ \text{and } w^t := (t+1)/2,$$

then we have

$$\mathbb{E}[\|[\textstyle\sum_{i=1}^{k-1} \hat{\pi}_p(\hat{\pi}_i - \pi_i^{t+1})(\pi_q^{t+1} - \pi_q^{t+1}(\xi_q^i))] + [\pi_{1:}^{t+1} - \pi_{1:}^{t+1}(\xi_{1:}^0)]\|^2] \le k\tilde{\sigma}_\Pi^2, \tag{5.12}$$

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t \delta^t] \le (N+1)\sqrt{N}\tilde{\sigma}_\varDelta. \tag{5.13}$$

Proof First, we show (5.12). Proposition 13 implies that $\forall 1 \le i \le k-1$

$$\begin{aligned}\mathbb{E}[\|[\hat{\pi}_p(\hat{\pi}_i - \pi_i^{t+1})(\pi_q^{t+1} - \pi_q^{t+1}(\xi_q^i))]\|^2] &\le 4M_p^2 M_{\Pi_i}^2 \mathbb{E}[\|\pi_q^{t+1} - \pi_q^{t+1}(\xi_q^i)\|^2] \\ &\le 4M_p^2 M_{\Pi_i}^2 \sigma_{\pi_{i+1:}}^2\ ,\end{aligned}$$

and that

$$\mathbb{E}[\|\pi_{1:}^{t+1} - \pi_{1:}^{t+1}(\xi_{1:}^0)\|^2] \le \sigma_{\pi_{1:}}^2.$$

So in view of the algebraic fact $\|\sum_{i=1}^k a_i\|^2 \le k\|a_i\|^2$, we conclude (5.12) by summing up the above inequalities.

Next, we show (5.13). Clearly, $\mathbb{E}[\delta_0^t] := \mathbb{E}[(\pi_{1:}^{t+1} - \pi_{1:}^{t+1}(\xi_{1:}^0))(x^t - x^*)] = 0$, so $\mathbb{E}[\sum_{t=0}^{N-1} w^t \delta_0^t] = 0$. As for $i \ge 1$, because

$$\begin{aligned}&\mathbb{E}[\|\pi_i^t\left(\mathcal{L}_q(x^t; \pi_q^{t+1}) - \mathcal{L}_q(x^t; \pi_q^{t+1}(\xi_q^i))\right)\|^2] \le M_{\Pi_i}^2 \sigma_{\mathcal{L}_q}^2, \|w^t \hat{\pi}_p\| \le \tfrac{N}{2} M_p, \\ &\mathbb{E}[\|\left(\mathcal{L}_q(x^t; \pi_q^{t+1}) - \mathcal{L}_q(x^t; \pi_q^{t+1}(\xi_q^i))\right)\|^2] \le \sigma_{\mathcal{L}_q}^2, \|w^t \hat{\pi}_p \hat{\pi}_i\| \le \tfrac{N}{2} M_p M_{\Pi_i},\ \forall t \le N-1,\end{aligned}$$

Lemma 2 implies that

$$\begin{aligned}&\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t \delta_i^t] \\ &= \mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t \hat{\pi}_p \hat{\pi}_i (\mathcal{L}(\pi_q^{t+1}, x^t) - \mathcal{L}(\pi_q^{t+1}(\xi_q^i), x^t))] + \mathbb{E}[\sum_{t=0}^{N-1} w^t \hat{\pi}_p(-\pi_i^t)(\mathcal{L}(\pi_q^{t+1}, x^t) - \mathcal{L}(\pi_q^{t+1}(\xi_q^i), x^t))] \\ &\le \tfrac{N\sqrt{N}}{2}(M_p M_{\Pi_i} \sigma_{\mathcal{L}_q} + M_p M_{\Pi_i} \sigma_{\mathcal{L}_q}) \\ &= N\sqrt{N} M_p M_{\Pi_i} \sigma_{\mathcal{L}_q}.\end{aligned}$$

In view of $\delta^t := \sum_{i=0}^k \delta_i^t$, the desired bound in (5.13) follows immediately. ∎

Now we are ready for the functional value convergence result. Because of the need to cancel additional stochastic errors, we set $\eta^t := H_-^t + H_+^t + (\sqrt{N+1}\sqrt{k}\tilde{\sigma}_\Pi)/(2\mathcal{D}_X)$ in the following convergence theorem.

Theorem 4 *Let the solution sequence $\{z^t\}$ be generated by Algorithm 3 with τ_i^t chosen according to Table 3 and let $\eta^t := \tilde{H}_-^t + \tilde{H}_+^t + (\sqrt{N+1}\sqrt{k}\tilde{\sigma}_\Pi)/2\mathcal{D}_X$. Then we have*

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q(z^{t+1}, \hat{z})] \le CST + \frac{w_0}{2} H_-^0 \|x^0 - x^*\|^2 + w^{N-1} H_+^{N-1} \mathcal{D}_X^2 + \frac{N\sqrt{N+1}\sqrt{k}}{2} \mathcal{D}_X \tilde{\sigma}_\Pi + N\sqrt{N}\tilde{\sigma}_\varDelta.$$

Moreover, let $\bar{x}^N = \sum_{t=0}^{N-1} w^t x^{t+1} / \sum_{t=0}^{N-1} w^t$ with $w^t = (t+1)/2$. Then

$$\begin{aligned}\mathbb{E}(f(\bar{x}^N) - f(x^*)) \le\ & \tfrac{4}{N(N+1)} CST + \tfrac{1}{N(N+1)} H_-^0 \|x^0 - x^*\|^2 + \tfrac{2}{N+1} H_+^{N-1} \mathcal{D}_X^2 \\ &+ \tfrac{4}{\sqrt{N}} \tilde{\sigma}_\varDelta + \tfrac{2\sqrt{k}}{\sqrt{N+1}} \tilde{\sigma}_\Pi \mathcal{D}_X.\end{aligned} \tag{5.14}$$

Layer Type	Complexity
Smooth	$\frac{\sqrt{\tilde{L}_{f_i}}\|x_0-x^*\|+\sqrt{M_p\mathcal{D}_{\Pi_i}^2}}{\sqrt{\epsilon}}+\frac{M_pL_{f_i}\sigma_{\mathcal{L}_q}^2}{\epsilon}+\frac{M_p^2\sigma_{\pi_{i:}}^2\mathcal{D}_X^2}{\epsilon^2}+\frac{M_p^2M_{\Pi_i}^2\sigma_{\mathcal{L}_q}^2}{\epsilon^2}$
Smoothable[3]	$\frac{\tilde{\mathcal{D}}_{\Pi_i}\mathcal{D}_X}{\epsilon}+\frac{(M_p\mathcal{D}_{\Pi_i}\sigma_{\mathcal{L}_q})^2}{\epsilon^2}+\frac{M_p^2M_{\Pi_i}^2\sigma_{\pi_{i+1:}}^2\mathcal{D}_X^2}{\epsilon^2}$
Non-Smooth(Layer k)	$\frac{\tilde{M}_{\Pi_k}^2\mathcal{D}_X^2}{\epsilon^2}+\frac{M_p^2\sigma_{\pi_k}^2\mathcal{D}_X^2}{\epsilon^2}$

Table 5: Iteration Complexity for Solving Stochastic Linearized One-layer Problem

Proof First, since $\{w^t\left(H_+^t+(\sqrt{N+1}\sqrt{k}\tilde{\sigma}_\Pi)/2\mathcal{D}_X\right)\}$ is monotonically non-decreasing and $\{w^tH_-^t\}$ is monotonically non-increasing, the three point inequality (2.20) for $\mathcal{A}_0^t$ implies that the w^t weighted sum of Q_0 satisfies

$$\begin{aligned}\mathbb{E}[\textstyle\sum_{t=0}^{N-1}w^tQ_0(z^{t+1},\hat{z})]\le{}&\tfrac{w^0}{2}H_-^0\|x^0-x^*\|^2-\textstyle\sum_{t=0}^{N-1}\tfrac{w^t}{2}(H_-^t+H_+^t+\tfrac{\sqrt{N+1}\sqrt{k}\tilde{\sigma}_\Pi}{2\mathcal{D}_X})\mathbb{E}\left[\|x^t-x^{t+1}\|^2\right]\\&+w^{N-1}(H_+^t+\tfrac{\sqrt{N+1}\sqrt{k}\tilde{\sigma}_\Pi}{2\mathcal{D}_X})\mathcal{D}_X^2+\mathbb{E}[\textstyle\sum_{t=0}^{N-1}w^t(\Delta_{x,0}^t+\delta_0^t)].\end{aligned}\tag{5.15}$$

Then adding (5.15) to (5.9), we conclude from Lemma 3 and the Young's inequality that

$$\begin{aligned}\mathbb{E}[\textstyle\sum_{t=0}^{N-1}w^tQ(z^{t+1},\hat{z})]\le{}&\text{CST}+\tfrac{w_0}{2}H_-^0\|x^0-x^*\|^2+w^{N-1}H_+^{N-1}\mathcal{D}_X^2+\tfrac{N\sqrt{N+1}\sqrt{k}}{4}\mathcal{D}_X\tilde{\sigma}_\Pi\\&+\mathbb{E}[\textstyle\sum_{t=0}^{N-1}w^t(-\tfrac{\sqrt{N+1}\sqrt{k}\tilde{\sigma}_\Pi}{2\mathcal{D}_X}\tfrac{1}{2}\|x^{t+1}-x^t\|^2+\Delta_x^t)]+\mathbb{E}[\textstyle\sum_{t=0}^{N-1}w^t\delta^t]\\\le{}&\text{CST}+\tfrac{w_0}{2}H_-^0\|x^0-x^*\|^2+w^{N-1}H_+^{N-1}\mathcal{D}_X^2+\tfrac{N\sqrt{N+1}\sqrt{k}}{2}\mathcal{D}_X\tilde{\sigma}_\Pi+(N)\sqrt{N}\tilde{\sigma}_\Delta.\end{aligned}$$

Next, since Proposition 11 implies $f(\bar{x}^N)-f(x^*)\le\sum_{t=0}^{N-1}w^tQ(z^{t+1},\hat{z})/(\sum_{t=0}^{N-1}w^t)$, (5.14) follows from dividing both sides of the above inequality by $(\sum_{t=0}^{N-1}w^t)$. ■

We illustrate Theorem 4 by applying it to (5.5) again, but with access to only stochastic oracles for layer functions. A straightforward application of Theorem 4 implies that Algorithm 3 has an iteration complexity of

$$\begin{aligned}\mathcal{O}\Bigg\{&\frac{\sqrt{\sum_{i\in S}\tilde{L}_{f_i}}\|x_0-x^*\|+\sqrt{\sum_{i\in S}M_p\mathcal{D}_{\Pi_i}^2}}{\sqrt{\epsilon}}+\frac{\sum_{i\in S}M_pL_{f_i}\sigma_{\mathcal{L}_q}^2}{\epsilon}\\&+\frac{\sum_{i\in P}\tilde{\mathcal{D}}_{\Pi_i}\mathcal{D}_X}{\epsilon}+\frac{\sum_{i\in P}(M_p\mathcal{D}_{\Pi_i}\sigma_{\mathcal{L}_q})^2}{\epsilon^2}\\&+\frac{\tilde{M}_{\Pi_k}^2\mathcal{D}_X^2}{\epsilon^2}+\frac{(k)\tilde{\sigma}_{\pi_i}^2\mathcal{D}_X^2}{\epsilon^2}+\frac{\tilde{\sigma}_\Delta^2}{\epsilon^2}\Bigg\}.\end{aligned}\tag{5.16}$$

Notice that the above iteration complexity bound is influenced more heavily by the noise of inner layer functions through $\sigma_{\mathcal{L}_q}$, $\tilde{\sigma}_\Pi$ and $\tilde{\sigma}_\Delta$ because every proximal update uses stochastic estimators from all its inner layers.

Moreover, the complexity bound in (5.16) also admits a linearized one-layer interpretation. By substituting in $\tilde{\sigma}_\Delta:=\sum_{i=1}^{k-1}M_pM_{\Pi_i}\sigma_{\mathcal{L}_q}$ and $\tilde{\sigma}_\Pi^2:=4\sum_{i=1}^{k-1}M_p^2M_{\Pi_i}^2\sigma_{\pi_{i+1:}}^2$, (5.16) reduces to

$$\begin{aligned}\mathcal{O}\Bigg\{&\sum_{i\in S}\frac{\sqrt{\tilde{L}_{f_i}}\|x_0-x^*\|+\sqrt{M_p\mathcal{D}_{\Pi_i}^2}}{\sqrt{\epsilon}}+\frac{M_pL_{f_i}\sigma_{\mathcal{L}_q}^2}{\epsilon}+\frac{kM_p^2\sigma_{\pi_{i:}}^2\mathcal{D}_X^2}{\epsilon^2}+\frac{(k-1)M_p^2M_{\Pi_i}^2\sigma_{\mathcal{L}_q}^2}{\epsilon^2}\\&+\sum_{i\in P}\frac{\tilde{\mathcal{D}}_{\Pi_i}\mathcal{D}_X}{\epsilon}+\frac{k(M_p\mathcal{D}_{\Pi_i}\sigma_{\mathcal{L}_q})^2}{\epsilon^2}+\frac{kM_p^2M_{\Pi_i}^2\sigma_{\pi_{i+1:}}^2\mathcal{D}_X^2}{\epsilon^2}\\&+\frac{\tilde{M}_{\Pi_k}^2\mathcal{D}_X^2}{\epsilon^2}+\frac{kM_p^2\sigma_{\pi_k}^2\mathcal{D}_X^2}{\epsilon^2}\Bigg\}.\end{aligned}\tag{5.17}$$

If we linearize every f_i in a similar fashion as (5.7), but assuming only stochastic oracle access to (A_i,c_i) with $\mathbb{E}[\|A_i(\xi_i)x+c_i(\xi_i)-A_ix-c_i\|^2]\le\sigma_{\mathcal{L}_q}^2$, $\mathbb{E}[\|A_i(\xi_i)-A_i\|^2]\le\sigma_{\pi_{i+1:}}$, then the complexities for solving them are shown in Table 5 and their sum matches (5.17) up to a factor of k.

5.3 Deterministic Strongly Convex Problems

We start to consider deterministic strongly convex problems. Since $u(x)$ have a strong convexity modulus of $\alpha>0$, we can choose increasing η^ts to obtain an improved convergence result, which is shown in the next proposition.

Theorem 5 *Let f be a deterministic NSCO problem (1.3) with $\alpha > 0$ and let k_n denote the number of layer functions which are neither smooth nor linear. Let $\{z^t := (x^t; \pi_1^t, \ldots, \pi_k^t)\}$ be generated by Algorithm 3 with $\eta^t := H_-^t + \frac{t+1}{3}\alpha$ and τ_i^t specified according to Table 3 and $c = \frac{2}{3k_n}\alpha$. If $w^t := (t+1)/2$, the following bounds hold for both the last iterate x^N and the ergodic average solution $\bar{x}^N := \sum_{t=0}^{N-1} w^t x^t / (\sum_{t=0}^{N-1} w^t)$:*

$$\tfrac{1}{2}\|x^N - x^*\|^2 \le \tfrac{6}{N(N+3)} \tfrac{CST}{\alpha} + \tfrac{3}{2N(N+3)} (\tfrac{H_-^0}{\alpha} + \tfrac{1}{3})\|x_0 - x^*\|^2, \tag{5.18a}$$

$$f(\bar{x}^N) - f(x^*) \le \tfrac{4}{N(N+1)} CST + \tfrac{1}{N(N+1)} (H_-^0 + \tfrac{\alpha}{3})\|x_0 - x^*\|^2. \tag{5.18b}$$

Proof The x-update in Line 7 of Algorithm 3 satisfies the following three-point inequality (see (2.20)):

$$\mathcal{A}_0^t \le \tfrac{\eta^t}{2}\tfrac{1}{2}\|x^t - x\|^2 - \tfrac{\eta^t + \alpha}{2}\tfrac{1}{2}\|x^{t+1} - x\|^2 - \tfrac{\eta^t}{2}\|x^t - x^{t+1}\|^2. \tag{5.19}$$

In view of this relation, we can use increasing $\eta^t := (t+1)\alpha/3$, while keeping $\sum_{t=0}^{N-1} w^t \mathcal{A}_0^t$ small. Such an increasing stepsize policy for η^t is the key to improve the convergence rate because it allows a more effective cancellation of error terms from the aggregate dual gap function $Q_{1:}$. More specifically, the bound on $Q_{1:}$ in (5.1) can be simplified to

$$\textstyle\sum_{t=0}^{N-1} w^t Q_{1:}(z^{t+1}, z) \le \mathrm{CST} + \sum_{t=0}^{N-1} \frac{w^t}{2}(H_\alpha^t + H_-^t)\|x^{t+1} - x^t\|^2. \tag{5.20}$$

Since $H_{i,\alpha}^t$ in Table 3 are given by

$$H_{i,\alpha}^t = \begin{cases} 0 & \text{if } f_i \text{ is either smooth or affine,} \\ \frac{(t+1)c}{2} & \text{otherwise,} \end{cases}$$

$c = 2\alpha/(3k_n)$ implies $H_\alpha^t = (t+1)\alpha/3$.

Now let a feasible reference point $z := (x^*, \pi_1, ..., \pi_k)$ be given. Since η^t satisfy $w^t\eta^t \le w^{t-1}(\eta^{t-1} + \alpha)$ and $Q_0(z^{t+1}, z) = \mathcal{A}_0^t$, we have

$$\textstyle\sum_{t=0}^{N-1} w^t Q_0(z^{t+1}, z) \le w^0\eta^0\frac{1}{2}\|x^0 - x\|^2 - w^{N-1}(\eta^{N-1} + \alpha)\frac{1}{2}\|x^N - x\|^2 - \sum_{t=0}^{N-1} w^t\eta^t\frac{1}{2}\|x^{t+1} - x^t\|^2. \tag{5.21}$$

Then by adding the above inequality to (5.20), we get

$$\textstyle\sum_{t=0}^{N-1} w^t Q(z^{t+1}, z) + w^{N-1}(\eta^{N-1} + \alpha)\frac{1}{2}\|x^N - x^*\|^2 \le w^0(\eta^0 + H_-^0)\frac{1}{2}\|x_0 - x^*\|^2 + \mathrm{CST}.$$

To obtain (5.18a), we can set the reference point to $z^* := (x^*; \pi_1^*, \ldots, \pi_k^*)$. The desired bound for $\frac{1}{2}\|x^N - x^*\|^2$ then follows immediately from the previous conclusion and the fact that $Q(z^t, z^*) \ge 0\ \forall t,$. Moreover, if we set the reference point to $\hat{z} := (x^*; \hat{\pi}_1, \ldots, \hat{\pi}_k)$, (5.18b) can be derived by applying Proposition 11. ■

To illustrate Theorem 5, we return to (5.5) again, but with $\alpha > 0$. A straightforward application of (5.18b) implies that Algorithm 3 exhibits an iteration complexity of

$$\mathcal{O}\left\{ \frac{\sqrt{\sum_{i \in S} \tilde{L}_{f_i} + \alpha}\,\|x_0 - x^*\|}{\sqrt{\epsilon}} + \frac{\sqrt{k_n \sum_{i \in P} \tilde{\mathcal{D}}_{\Pi_i}^2}}{\sqrt{\alpha\epsilon}} + \frac{k_n \tilde{M}_{\Pi_k}^2}{\alpha\epsilon} \right\}.$$

It should be noted, however, that the above iteration complexity is not optimal for smooth layers. Instead, we can use both (5.18a) and (5.18b) to design a multi-epoch restarting scheme (see Section 4.2.3 [13]) to improve the complexity to

$$\mathcal{O}\left\{ \sqrt{\frac{\sum_{i \in S} \tilde{L}_{f_i}}{\alpha} + 1}\left(\log(\tfrac{\|x_0 - x^*\|^2\alpha}{\epsilon}) + 1\right) + \frac{\sqrt{k_n \sum_{i \in P} \tilde{\mathcal{D}}_{\Pi_i}^2}}{\sqrt{\alpha\epsilon}} + \frac{k_n \tilde{M}_{\Pi_k}^2}{\alpha\epsilon} \right\}. \tag{5.22}$$

which matches the sum of optimal iteration complexities for solving linearized one-layer problems (5.7), shown in Table 6.

Layer Type, $\alpha > 0$	Smooth	Smoothable	Non-smooth
Optimal Complexity	$\mathcal{O}\{(\sqrt{\frac{k\tilde{L}_{f_i}}{\alpha}})\log\frac{\|x_0-x^*\|^2}{\epsilon}\}$	$\mathcal{O}\{\frac{k\tilde{\mathcal{D}}_{\Pi_i}\mathcal{D}_X}{\sqrt{\epsilon}}\}$	$\mathcal{O}\{\frac{k\tilde{M}^2_{\Pi_k}\mathcal{D}^2_X}{\epsilon\alpha}\}$

Table 6: Optimal Complexity for Strongly Convex One-layer Problem (5.7).

5.4 Stochastic Strongly Convex Problems

Now we consider stochastic strongly convex problems. In general, we can only prove a $\mathcal{O}(1/\sqrt{N})$ functional value convergence rate for Algorithm 3 because the α-strong convexity cannot improve the convergence bound for $\mathbb{E}[\sum_{t=0}^{N-1} w^t\delta_i^t]$ from (5.10). However, when all layer functions outside the innermost stochastic layer function are (stochastic) smooth, i.e.,

$$f(x) := \underbrace{f_1 \circ \ldots \circ f_{l-1}}_{(stochastic) smooth} \circ \underbrace{f_l}_{stochastic} \circ \underbrace{f_{l+1} \circ \ldots \circ f_k(x)}_{deterministic} + u(x), \tag{5.23}$$

the alternative gap to functional optimality conversion rule in Proposition 12 allows us to select reference points, for which $\mathbb{E}[\sum_{t=0}^{N-1} w^t\delta_i^t] = 0$, to accelerate it to $\mathcal{O}(1/N)$. We will study this special case for the rest of the subsection.

More specifically, since Proposition 12 only needs a bound for $Q(\cdot,\tilde{z})$ with $\tilde{z} := (x^*; \pi^*_{1:l-1}, \hat{\pi}_{l:})$ and a bound for $\frac{1}{2}\|\bar{x}^N - x^*\|^2$, which can be derived from $Q(\cdot, z^*)$ with $z^* := (x^*; \pi_1^*, \ldots, \pi_k^*)$, we can focus on (random) reference points of the form $z := (x^*; \pi^*_{1:l-1}, \pi_{l:})$, where only $\pi_{l:}$ could depend on the stochastic solution sequence $\{z^t\}$. Lemma 4 below states some simplified bounds on δ^t (c.f. (5.10)) and Δ_x^t (c.f. (5.11)) for these reference points,

$$\delta^t := \textstyle\sum_{i=0}^k \delta_i^t = \sum_{i=0}^{l-1}\delta_i^t = \sum_{i=1}^{l-1}\pi_p^*(\pi_i^* - \pi_i^t)[\mathcal{L}_q(x^t;\pi_q^{t+1}) - \mathcal{L}_q(x^t;\pi_q^{t+1}(\xi_q^i))] + (\pi_{1:}^{t+1} - \pi_{1:}^{t+1}(\xi_{1:}^0))(x^t - x),$$
$$\Delta_x^t := \textstyle\sum_{i=0}^k \Delta_{x,i}^t = \sum_{i=0}^{l-1}\Delta_{x,i}^t = \left\{[\sum_{i=1}^{l-1}\pi_p^*(\pi_i^* - \pi_i^{t+1})(\pi_q^{t+1} - \pi_q^{t+1}(\xi_q^i))] + [\pi_{1:}^{t+1} - \pi_{1:}^{t+1}(\xi_{1:}^0)]\right\}(x^{t+1} - x^t).$$

Lemma 4 *Let the solution sequence $\{z^t\}$ be generated by Algorithm 3. If*

$$\tilde{\sigma}_\Pi^2 := 4\textstyle\sum_{i=1}^{l-1} M_p^2 M_{\Pi_i}^2 \sigma^2_{\pi_{i+1:}} \leq \sum_{i=1}^{l}(4i) M^2_{\Pi_{1:i-1}}\sigma^2_{\pi_i} M_q^2 \text{ and } w^t := (t+1)/2,$$

then

$$\mathbb{E}[\|\textstyle\sum_{i=1}^{l-1}\pi_p^*(\pi_i^* - \pi_i^{t+1})[\pi_q^{t+1} - \pi_q^{t+1}(\xi_q^i)] + [\pi_{1:}^{t+1} - \pi_{1:}^{t+1}(\xi_{1:}^0)]\|^2] \leq \tilde{\sigma}_\Pi^2, \tag{5.24}$$

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t\delta^t] = 0. \tag{5.25}$$

Proof The derivation for (5.24) is similar to that in Lemma 3 except the terms in $\{\pi_p^*(\pi_i^* - \pi_i^{t+1})[\pi_q^{t+1} - \pi_q^{t+1}(\xi_q^i)]\}_{i=1}^{l-1} \cup \{\pi_{1:}^{t+1} - \pi_{1:}^{t+1}(\xi_{1:}^0)\}$ are all conditionally uncorrelated. Next, to show (5.25), first notice $\mathbb{E}[\delta_0^t] := \mathbb{E}[\pi_{1:}^{t+1} - \pi_{1:}^{t+1}(\xi_{1:}^0))(x^t - x^*)] = 0$. Moreover, for any $i \leq l-1$, $\pi_p^*(\pi_i^* - \pi_i^t)$ is conditionally independent of $[\mathcal{L}_q(x^t;\pi_q^{t+1}) - \mathcal{L}_q(x^t;\pi_q^{t+1}(\xi_q^i))]$, so $\mathbb{E}[\delta_i^t] = 0$. Thus we have $\mathbb{E}[\sum_{t=0}^{N-1} w^t\delta^t] = 0$. ∎

We establish a bound on the gap function Q in the following proposition. Just like the previous subsection, the α-strong convexity allows us to choose an increasing η^t.

Proposition 22 *Let k_n denotes the number of layer functions which are neither smooth nor linear. Let $\{z^t := (x^t; \pi_1^t, \ldots, \pi_k^t)\}$ be generated by Algorithm 3 with $\eta^t = H_-^t + \frac{t+1}{3}\alpha$ and τ_i^t specified according to Table 3 and $c := \frac{1}{3k_n}\alpha$. Then*

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q(z^{t+1}, z)] + \frac{N(N+1)\alpha}{12}\mathbb{E}[\|x^N - x\|^2] \leq (\frac{\alpha}{6} + \frac{H_-^0}{2})\|x_0 - x\|^2 + CST + \frac{3N}{2\alpha}\tilde{\sigma}_\Pi^2. \tag{5.26}$$

Proof Clearly, the aggregate dual gap function $Q_{1:}$ can be bounded by

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q_{1:}(z^{t+1}, z)] \leq \text{CST} + \sum_{t=0}^{N-1}\frac{w^t}{2}(H_\alpha^t + H_-^t)\mathbb{E}[\|x^{t+1} - x^t\|^2] + \mathbb{E}[\sum_{t=0}^{N-1} w^t \sum_{i=1}^{k-1}(\delta_i^t + \Delta_{x,i}^t)]. \tag{5.27}$$

The primal gap function Q_0 can be decomposed to $Q_0(z^{t+1}, z) = \mathcal{A}_0^t + \delta_0^t + \Delta_{x,0}^t$ with

$$\mathcal{A}_0^t \leq \frac{1}{2}\left(\eta^t\|x^t - x\|^2 - (\eta^t + \alpha)\|x^{t+1} - x\|^2 - \eta^t\|x^t - x^{t+1}\|^2\right). \tag{5.28}$$

So the w^t-weighted sum of $Q_0(z^{t+1}, z)$ satisfies

$$\begin{aligned}\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q_0(z^{t+1}, z)] \leq &\tfrac{w^0\eta^0}{2}\|x^0 - x\|^2 - \tfrac{w^{N-1}(\eta^{N-1}+\alpha)}{2}\mathbb{E}[\|x^N - x\|^2] + \mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t\delta_0^t + \varDelta_{x,0}^t] \\ &- \textstyle\sum_{t=0}^{N-1} \tfrac{w^t}{2}(\tfrac{t+1}{3}\alpha + H_-^t)\mathbb{E}[\|x^{t+1} - x^t\|^2].\end{aligned} \tag{5.29}$$

Next we use the above $(t+1)\alpha/3\|x^t - x^{t+1}\|^2$ to cancel terms from (5.27). Since $c := 1/(3k_n)$ in Table 3, $H_\alpha^t = (t+1)\alpha/6$. The remaining $(t+1)\alpha/12\|x^{t+1} - x^t\|^2$ can bound $\varDelta_x^t$ using the Young's inequality:

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t(\varDelta_x^t - \tfrac{(t+1)\alpha}{12}\|x^t - x^{t+1}\|^2)] \leq \tfrac{3N}{2\alpha}\tilde{\sigma}_\varPi^2. \tag{5.30}$$

Finally in view of $\mathbb{E}[\sum_{t=0}^{N-1} w^t\delta_i^t] = 0$, the desired bound in (5.26) can be obtained by adding (5.29) to (5.27). ■

Now we use Proposition 12 to show the convergence of Algorithm 3 in function values. Towards that end, it is convenient to split CST into CST $:= CST^{(0)} + NCST^{(1)}$ such that constants $CST^{(0)}$ and $CST^{(1)}$ are independent of N.

Theorem 6 *Let $\{z^t := (x^t; \pi_1^t, \ldots, \pi_k^t)\}$ be generated by Algorithm 3 using $\eta^t = H_-^t + \frac{t+1}{3}\alpha$ and τ_i^ts specified according to Table 3 with $c = \frac{1}{3k_n}\alpha$. If $w^t := (t+1)/2$ and $\bar{x}^N = \sum_{t=0}^{N-1} w^t x^t / \sum_{t=0}^{N-1} w^t$, then*

$$\begin{aligned}\mathbb{E}[f(\bar{x}^N) - f(x^*)] \leq &\tfrac{4}{N(N+1)}(1 + \tfrac{3\tilde{L}_R\log(N+1)}{\alpha})CST^{(0)} + \tfrac{4}{N+1}(1 + \tfrac{3\tilde{L}_R}{\alpha})CST^{(1)} + \tfrac{6}{N+1}(1 + \tfrac{3\tilde{L}_R}{\alpha})\tfrac{\tilde{\sigma}_\varPi^2}{\alpha} \\ &+ \tfrac{1}{N(N+1)}(1 + \tfrac{3\tilde{L}_R\log(N+1)}{\alpha})(H_-^0 + \tfrac{1}{3}\alpha)\|x^* - x_0\|^2,\end{aligned} \tag{5.31}$$

where $\tilde{L}_R := \sum_{i=1}^{l-1} M_p L_{f_i} M_q^2$ is smoothness constant associated with outer layer functions $f_1, \ldots, f_{l-1}$.

Proof Just like Theorem 5, we can set the reference point z in Proposition 22 to be $\tilde{z}$ and then z^* to derive:

$$\mathbb{E}[\textstyle\sum_{t=0}^{N-1} w^t Q(z^{t+1}, \tilde{z})] \leq (\tfrac{\alpha}{6} + \tfrac{H_-^0}{2})\|x_0 - x\|^2 + \text{CST} + \tfrac{3N}{2\alpha}\tilde{\sigma}_\varPi^2, \tag{5.32}$$

$$\mathbb{E}[\tfrac{1}{2}\|x^t - x^*\|^2] \leq \tfrac{1}{t(t+1)}(\tfrac{1}{2} + \tfrac{3H_-^0}{2\alpha})\|x_0 - x\|^2 + \tfrac{6}{t(t+1)}\tfrac{\text{CST}^t}{\alpha} + \tfrac{9}{(t+1)}\tfrac{\tilde{\sigma}_\varPi^2}{\alpha^2},\ \forall t. \tag{5.33}$$

Then (5.33) and the Jensen's inequality imply

$$\begin{aligned}\mathbb{E}[\tfrac{1}{2}\|\bar{x}^N - x^*\|^2] \leq &\frac{\mathbb{E}[\sum_{t=0}^{N-1} w^t\frac{1}{2}\|x^t - x^*\|^2]}{\sum_{t=0}^{N-1} w^t} \\ \leq &\tfrac{4}{N(N+1)}\tfrac{3\log(N+1)}{\alpha}CST^{(0)} + \tfrac{4}{N+1}\tfrac{3}{\alpha}CST^{(1)} + \tfrac{6}{N+1}\tfrac{3}{\alpha}\tfrac{\tilde{\sigma}_\varPi^2}{\alpha} \\ &+ \tfrac{2}{N(N+1)}\tfrac{3\log(N+1)}{\alpha}(H_-^0 + \tfrac{1}{3}\alpha)\tfrac{1}{2}\|x^* - x_0\|^2,\end{aligned}$$

and (5.32) imply

$$\mathbb{E}(\frac{\sum_{t=0}^{N-1} w^t Q(z^{t+1}, \tilde{z})}{\sum_{t=0}^{N-1} w^t}) \leq \tfrac{4}{N(N+1)}CST^{(0)} + \tfrac{4}{N+1}CST^{(1)} + \tfrac{2}{N(N+1)}(H_-^0 + \tfrac{1}{3}\alpha)\tfrac{1}{2}\|x^* - x_0\|^2 + \tfrac{6}{N+1}\tfrac{\tilde{\sigma}_\varPi^2}{\alpha}.$$

So the desired bound (5.31) follows directly from Proposition 12. ■

Now, we illustrate the use of Theorem 6 by considering the problem, $\min_{x\in X} f(x) := f_1 \circ \ldots \circ f_k(x) + u(x)$, where f_k is stochastic non-smooth and $\{f_i\}_{i\geq 2}$ are stochastic smooth. Because there is one nonsmooth layer, i.e., $k_n = 1$, Theorem 6 implies that Algorithm 3 has an iteration complexity of

$$\begin{aligned}\mathcal{O}\Bigg\{&(\tfrac{\sqrt{\sum_{i=1}^{k-1}\tilde{L}_{f_i}+\alpha}\|x_0-x^*\|}{\sqrt{\epsilon}} + \tfrac{\sqrt{\sum_{i=1}^{k-1} M_p\mathcal{D}_{\varPi_i}^2}}{\sqrt{\epsilon}})\sqrt{1 + \tfrac{\tilde{L}_R}{\alpha}\log(\tfrac{1}{\epsilon})} + \tfrac{\sum_{i=1}^{k-1} M_p L_{f_i}\sigma_{\mathcal{L}_q}^2}{\epsilon}(1 + \tfrac{\tilde{L}_R}{\alpha}) \\ &+ \tfrac{\tilde{M}_{\varPi_k}^2}{\alpha\epsilon}(1 + \tfrac{\tilde{L}_R}{\alpha}) + \tfrac{\tilde{\sigma}_\varPi^2}{\alpha\epsilon}(1 + \tfrac{\tilde{L}_R}{\alpha})\Bigg\}.\end{aligned}$$

Moreover, the complexity for finding a solution x^N s.t. $\mathbb{E}[\|x^N - x^*\|^2] \leq \epsilon$ can also be interpreted as the sum of iteration complexities (listed in Table 7) for finding ϵ-close solutions for linearized stochastic one-layer problems (c.f. (5.7)) $\min_{x\in X}\ b_i^\top f_i(A_i x + c_i) + u(x)$ where (A_i, c_i) are accessible only through stochastic oracles.

6 Applications

In this section, we demonstrate the practical use of the SSD Algorithm 3 by applying it to two concrete problems.

Layer Type, $\alpha > 0$	Complexity for Finding an ϵ-close Solution
Smooth	$\frac{\sqrt{\tilde{L}_{f_i}+\alpha}\|x_0-x^*\|+\sqrt{M_p\mathcal{D}^2_{\Pi_i}}}{\sqrt{\alpha\epsilon}} + \frac{M_{\Pi_{1:i-1}}L_{f_i}\sigma^2_{\mathcal{L}_q}}{\alpha\epsilon} + \frac{M^2_{\Pi_{1:i-1}}\sigma^2_{\pi_{i:}}}{\alpha^2\epsilon}$
Non-Smooth(Layer k)	$\frac{\|x_0-x^*\|}{\sqrt{\epsilon}} + \frac{\tilde{M}^2_{\Pi_k}}{\alpha^2\epsilon} + \frac{M_p\sigma^2_{\pi_k}}{\alpha^2\epsilon}$

Table 7: Iteration Complexities for Strongly Convex Stochastic Linearized One-layer Problems

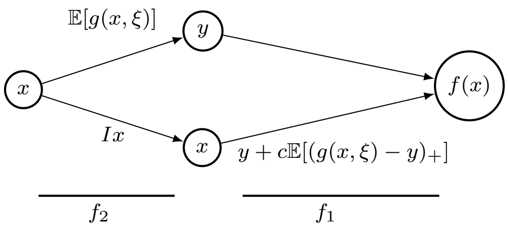

Fig. 3: Two Layer Formulation for (6.1).

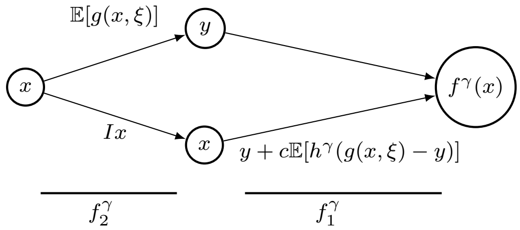

Fig. 4: Two Layer Formulation for (6.2).

Algorithm 5 SSD Algorithm for Mean-upper-semideviation Risk

Input: $x_{-1} = x_0 \in X$, $\underline{y}_1^0 \in \mathrm{dom}(f_1^\gamma)$.

Draw scenario $\xi_{2,1}^0$ and compute $\pi_2^0(\xi_{2,1}^0) := (f_2^\gamma)'(x_0, \xi_{2,1}^0)$.

for $t = 0, 1, 2, 3 ... N-1$ **do**

Draw independent scenarios $\xi_{2,1}^{t+1}, \xi_{2,0}^{t+1}, \xi_{1,0}^{t+1}$ from Ξ according to the probability distribution of ξ.

Let $\bar{x}^{t+1} := x^t + \theta^t(x^t - x^{t-1})$.
Set $\pi_2^{t+1}(\xi_{2,1}^{t+1}) := [g'(\bar{x}^{t+1}, \xi_{2,1}^{t+1}); I]$ and $f_2^*(\pi_2^{t+1}, \xi_{2,1}^{t+1}) := [g(\bar{x}^{t+1}, \xi_{2,1}^{t+1}) - g'(\bar{x}^{t+1}, \xi_{2,1}^{t+1})\bar{x}^{t+1}; 0]$.
Set $\pi_2^{t+1}(\xi_{2,0}^{t+1}) := [g'(\bar{x}^{t+1}, \xi_{2,0}^{t+1}); I]$.

Set $\bar{y}_1^{t+1} := \mathcal{L}_2(x^t; \pi_2^{t+1}(\xi_{2,1}^{t+1})) + \theta^t \pi_2^t(\xi_{2,1}^t)(x^t - x^{t-1})$.
Compute $\underline{y}_1^{t+1} := (\tau_1^t \underline{y}_1^t + \bar{y}_1^{t+1})/(1 + \tau_1^t)$ and set $\pi_1^{t+1}(\xi_{1,0}^{t+1}) := \nabla f_1^\gamma(\underline{y}_i^{t+1}, \xi_{1,0}^{t+1})$.

Set $x^{t+1} := \arg\min_{x \in X} \pi_1^{t+1}(\xi_{1,0}^{t+1})\pi_2^{t+1}(\xi_{2,0}^{t+1})x + \eta^t \frac{1}{2}\|x - x^t\|^2$.

end for

Return $\bar{x}^N := \sum_{t=0}^{N-1} w^t x^{t+1} / \sum_{t=0}^{N-1} w^t$.

6.1 Risk Averse Optimization

In risk-averse optimization, one important risk measure is the mean-upper-semideviation of order one [20]. More specifically, the risk associated with a random cost variable Z is:

$$\rho(Z) := \mathbb{E}[Z] + c\mathbb{E}[(Z - \mathbb{E}[Z])_+]^{❸},$$

where the risk-expectation trade-off parameter $c \leq 1$ reflects the modeler's risk aversiveness. Now if random variable $Z(x)$ represents the cost incurred by a decision x under a random scenario ξ (distributed according to some probability measure over Ξ), i.e., $Z(x) \sim g(x, \xi)$, then the risk minimization problem is

$$\min_{x \in X} \{\rho(Z(x)) := \mathbb{E}[g(x,\xi)] + c\mathbb{E}[(g(x,\xi) - \mathbb{E}[g(x,\xi)])_+]\}. \tag{6.1}$$

We assume $g(x, \xi)$ to be non-smooth for generality. For example, in the two-stage linear program (with relative complete recourse), $g(x, \xi)$ is the minimum objective value of the second stage LP, which is clearly non-smooth.

To apply our SSD framework, we can formulate (6.1) as a two layer problem. However, this formulation violates Assumption 4 because the non-smooth layer function $f_1(x, y) := y + c\mathbb{E}[(g(x, \xi) - y)_+]$ has a noisy argument y. One way around the issue is to replace the non-smooth $(z)_+$ by a Nesterov's smooth approximation function with a parameter $\gamma > 0$ [17] (the one-sided Huber Loss function):

$$h^\gamma(z) := \max_{\pi \in [0,1]} \pi z - \frac{\gamma}{2}\|\pi\|^2 \equiv \begin{cases} 0 & \text{if } z < 0 \\ \frac{1}{2}z^2 & \text{if } 0 \leq z \leq \gamma \\ z - \frac{1}{2}\gamma & \text{if } \gamma < z \end{cases}.$$

❸ $(y)_+ := \max\{0, y\}$.

Since h^γ approximates $(z)_+$ uniformly, i.e., $h^\gamma(z) \le (z)_+ \le h^\gamma(z) + \frac{\gamma}{2}$, we can set $\gamma = \epsilon$ such that an $\epsilon/2$-optimal solution for the smoothed problem f^γ is an ϵ-optimal solution for (6.1). More specifically, the smoothed problem is defined as

$$\min_{x\in X}\{f^\gamma(x) := \mathbb{E}[g(x,\xi)] + c\mathbb{E}[h^\gamma\,(g(x,\xi) - \mathbb{E}[g(x,\xi)])]\}, \tag{6.2}$$

and its two layer formulation is shown in Figure 4. Observe that $\partial f_1^\gamma(x,y)/\partial y$ is a c/γ continuous function of y for a fixed x, so f_1^γ is a semi-smooth-noisy layer function. Then by filling in the abstract dual updates in the SSD Framework Algorithm 3, a concrete implementation for solving (6.2) can be easily deduced, which is shown in Algorithm 5. Moreover, if the problem satisfies

$$\mathbb{E}[|g(x,\xi)|^2] \le M_g^2 \text{ and } \mathbb{E}[|g(x,\xi) - \mathbb{E}[g(x,\xi)]|^2] \le \sigma_g^2, \forall x,$$

$$\mathbb{E}[\|g'(x,\xi)\|^2] \le M_{\pi_g}^2 \text{ and } \mathbb{E}[\|g'(x,\xi) - \mathbb{E}(g'(x,\xi))\|^2] \le \sigma_{\pi_g}^2,\ \forall x,$$

we can pick the stepsizes in Algorithm 5 to be

$$\tau_1^t = \tfrac{t}{4} + \tfrac{t+1}{6}, \quad \eta^t = \tfrac{4c}{(t+1)\gamma} M_{\pi_g}^2 + \tfrac{\sqrt{N+1}cM_{\pi_g}}{\mathcal{D}_X} + \tfrac{\sqrt{2(N+1)c}}{2\mathcal{D}_X}\sqrt{6\sigma_{\pi_g}^2 + M_{\pi_g}^2}.$$

Then according to Theorem 4, Algorithm 5 has a convergence rate of

$$\mathcal{O}\{\tfrac{1}{N(N+1)}(M_g^2 + \tfrac{M_{\pi_g}^2\mathcal{D}_X^2}{\gamma}) + \tfrac{1}{N\gamma}(\sigma_{\pi_g}^2 + n\sigma_g^2) + \tfrac{1}{\sqrt{N}}(M_{\pi_g}\mathcal{D}_X + \sigma_{\pi_g}\mathcal{D}_X + \sigma_g)\}.$$

In addition, by substituting in $\gamma = \epsilon$, we can show that Algorithm 5 has an iteration complexity of

$$\mathcal{O}\{\tfrac{M_g^2}{\sqrt{\epsilon}} + \tfrac{M_{\pi_g}\mathcal{D}_X}{\epsilon} + \tfrac{1}{\epsilon^2}(\sigma_{\pi_g}^2\mathcal{D}_X^2 + \sigma_g^2 + M_{\pi_g}^2\mathcal{D}_X^2)\},$$

which matches the $\mathcal{O}\{1/\epsilon^2\}$ complexity for solving risk-neutral convex stochastic program.

6.2 Stochastic Composite Optimization

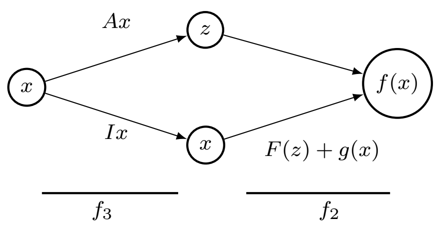

Fig. 5: Two Layer Formulation for (6.3).

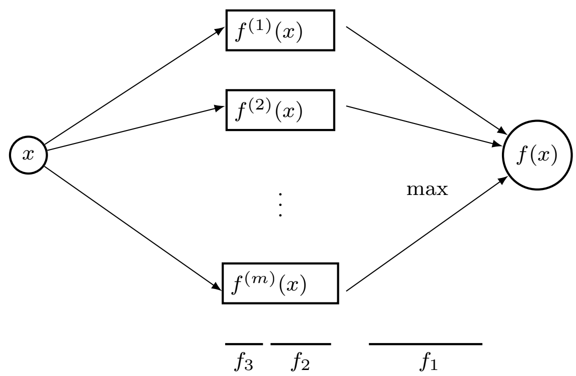

Fig. 6: Three Layer Formulation for (6.4).

Next we consider a stochastic composite optimization problem that arises frequently in machine learning and data analysis [3]:

$$\min_{x\in X}\{f(x) := F(Ax) + g(x) \equiv \max_{\pi_F\in\Pi_F} \langle \pi_F, Ax\rangle - F^*(\pi_F) + g(x)\}, \tag{6.3}$$

where F is a smoothable function, for example, the total variation loss function, and g is a stochastic smooth function, for example, the data fidelity loss function. In addition, the dimension of A is usually large, so we assume that there exist a stochastic oracle $\mathcal{SO}$ to return unbiased estimates $A(\xi)$, $A^\top(\xi)$, and $g'(x,\xi)$ for A, $A^\top$ and $g'(x)$, and that their variances are uniformly bounded by σ_A^2, $\sigma_{A^\top}^2$ and $\sigma_{\pi_g}^2$ respectively.

Clearly, (6.3) can be formulated as a two layer problem shown in Figure 5. By noticing that f_3 is a stochastic linear layer and f_2 is a separable mixture layer with input mixing, it is straightforward to derive from the abstract

Algorithm 6 SSD Algorithm for Composite Optimization

Input: $x_{-1} = x_0 \in X$ and $\pi_F^0 \in \Pi_F$.

1: Set $\underline{y}_g^0 := x_0$ and call $\mathcal{SO}$ to obtain estimate $A(\xi_{3,2}^0)$.
2: **for** $t = 0, 1, 2, 3 \ldots N-1$ **do**
3: Call $\mathcal{SO}$ to obtain estimates $A(\xi_{3,2}^{t+1})$ and $A^\top(\xi_{3,0}^{t+1})$.
4: Let $\bar{x}^{t+1} := x^t + \theta^t(x^t - x^{t-1})$.
Let $\underline{y}_g^{t+1} := (\tau_g^t \underline{y}_g^t + \bar{x}^{t+1})/(1+\tau_g^t)$ and call $\mathcal{SO}$ to obtain $\pi_g^{t+1}(\xi_{2,0}^{t+1}) := g'(\underline{y}_g^{t+1}, \xi_{2,0}^{t+1})$.
5: Let $\underline{y}_F^{t+1} := A(\xi_{3,2}^{t+1})x^t + A(\xi_{3,2}^t)(x^t - x^{t-1})$.
Compute $\pi_F^{t+1} := \arg\min_{\pi_F \in \Pi_F} -\langle \pi_F, \underline{y}_F^{t+1}\rangle + F^*(\pi_F) + \tau_F^t \frac{1}{2}\|\pi_F - \pi_F^t\|^2$
6: Set $x^{t+1} := \arg\min_{x \in X} \langle \pi_g^{t+1}(\xi_{2,0}^{t+1}) + A^\top(\xi_{3,0}^{t+1})\pi_F^{t+1}, x\rangle + \eta^t \frac{1}{2}\|x - x^t\|^2$.
7: **end for**
8: Return $\bar{x}^N := \sum_{t=0}^{N-1} w^t x^{t+1} / \sum_{t=0}^{N-1} w^t$.

SSD Framework in Algorithm 3 a concrete implementation for solving (6.3), shown in Algorithm 6. Moreover, with appropriately chosen stepsizes, Theorem 4 implies that Algorithm 6 has an iteration complexity of

$$\mathcal{O}\{\frac{\sqrt{L_g}\|x-x^*\|}{\sqrt{\epsilon}} + \frac{\|A\|\mathcal{D}_X D_{\Pi_F}}{\epsilon} + \frac{(\sigma_A^2 + \sigma_{A^\top}^2)\mathcal{D}_X^2 \mathcal{D}_{\Pi_F}^2}{\epsilon^2} + \frac{\sigma_{\pi_g}^2 \mathcal{D}_X^2}{\epsilon^2}\},$$

which is not improvable under our setting.

Next, comparing Algorithm 6 with the accelerated primal dual (APD) algorithm [3] designed specifically for problem (6.3), we see that APD can achieve the same convergence rate. However, our general approach allows an easy extension to handle more complicated problems in which f in (6.3) is but one sub-component. In particular, if $f^{(i)} := g^{(i)}(x) + F^{(i)}(A^{(i)}x)$, then a minimax problem which arises frequently from constrained optimization or multi-objective optimization is

$$\min_{x \in X} \{f(x) := \max\{f^{(1)}(x), f^{(2)}(x), \ldots, f^{(m)}(x)\} \equiv \max_{\pi_1 \in \Delta_m^+} \sum \pi_1^{(i)} f^{(i)}(x)\}^{[4]}. \quad (6.4)$$

Clearly, (6.4) admits a three layer formulation shown in Figure 6, so we only need to add an additional proximal step to update π_1^{t+1} after parallel (implicit) proximal updates for π_3^{t+1}, π_2^{t+1} in Algorithm 6. Moreover, if the variances for $g^{(i)}(\cdot, \xi_i)$, $\nabla g^{(i)}(\cdot, \xi_i)$, $A^{(i)}(\xi_i)$, ${A^{(i)}}^\top(\xi_i)$ are uniformly bounded by σ_g, $\sigma_{\pi_g}^2$, σ_A^2 and $\sigma_{A^\top}^2$, and if $\|A^{(i)}\|$, $\mathcal{D}_{\Pi_{F^{(i)}}}$ and $\|\nabla g^{(i)}(\cdot)\|$ are uniformly bounded by $\|A\|$, $\mathcal{D}_{\Pi_F}$ and M_g, then a straightforward application of Theorem 4 implies that the extended algorithm has an iteration complexity of

$$\mathcal{O}\{\frac{\sqrt{L_g}\|x-x^*\|}{\sqrt{\epsilon}} + \frac{\sqrt{m}\|A\|\mathcal{D}_X D_{\Pi_F}}{\epsilon} + \frac{m(\sigma_A^2 + \sigma_{A^\top}^2)\mathcal{D}_X^2 \mathcal{D}_{\Pi_F}^2}{\epsilon^2} + \frac{m(\sigma_{\pi_g}^2 \mathcal{D}_X^2 + \sigma_g^2)}{\epsilon^2} + \frac{\sqrt{m}M_g}{\epsilon}\}.$$

Moreover, if the entropy Bregman's distance function is used for the π_1 proximal update like [26], the above complexity can be improved to be nearly independent of the number of sub-components,

$$\mathcal{O}\{\frac{\sqrt{L_g}\|x-x^*\|}{\sqrt{\epsilon}} + \frac{\sqrt{\log(m)}\|A\|\mathcal{D}_X D_{\Pi_F}}{\epsilon} + \frac{\log(m)(\sigma_A^2 + \sigma_{A^\top}^2)\mathcal{D}_X^2 \mathcal{D}_{\Pi_F}^2}{\epsilon^2} + \frac{\log(m)(\sigma_{\pi_g}^2 \mathcal{D}_X^2 + \sigma_g^2)}{\epsilon^2} + \frac{\sqrt{\log(m)}M_g}{\epsilon}\}.$$

.

7 Conclusion

In this paper we showed that by imposing a layer-wise convexity assumption and a compositional monotonicity assumption, convex NSCO problems can be solved with tight iteration complexities. For the two-layer problem, we introduced a simple vanilla-SSD algorithm which can be implemented purely in the primal form. For the multi-layer problem, we proposed a general Stochastic Sequential Dual (SSD) framework. The framework consists of modular dual updates for different types of functions (smooth, smoothable, non-smooth, etc.), and so is capable of handling the general composition of different layer functions. Moreover, we presented modular convergence proofs to show that the complexity of SSD is optimal for nearly all problem parameters.

[4] $\Delta_+^m := \{\pi_1 \in \mathbb{R}_+^m | \sum_{i=1}^m \pi_1^{(i)} = 1\}$ is the probability simplex.

Reference

1. Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.
2. Tianyi Chen, Yuejiao Sun, and Wotao Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *arXiv preprint arXiv:2008.10847*, 2020.
3. Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.
4. S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23:2061–2089, 2013.
5. Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.
6. Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
7. 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.
8. G. Lan. Efficient methods for stochastic composite optimization. Manuscript, Georgia Institute of Technology, 2008.
9. G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, pages 365–397, 2012.
10. G. Lan. Gradient sliding for composite optimization. *Mathematical Programming*, 159(1-2):201–235, 2016.
11. G. Lan and Y. Ouyang. Accelerated gradient sliding for structured convex optimization. *Computational Optimization and Applications*, 2020. under revision.
12. Guanghui Lan. Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization. *Mathematical Programming*, 149(1-2):1–45, 2015.
13. Guanghui Lan. *Lectures on Optimization Methods for Machine Learning*. Springer-Nature, 2020.
14. Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical Programming*, 171:167–215, 2018.
15. Adrian S Lewis and Stephen J Wright. A proximal method for composite minimization. *Mathematical Programming*, 158(1-2):501–546, 2016.
16. Arkadi Semenovich Nemirovsky and David Borisovich Yudin. *Problem complexity and method efficiency in optimization.* John Wiley UK/USA, 1983.
17. Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.
18. Yurii Nesterov. Gradient methods for minimizing composite objective function. core discussion papers 2007076, université catholique de louvain. *Center for Operations Research and Econometrics (CORE)*, 5(5.3), 2007.
19. Andrzej Ruszczynski. A stochastic subgradient method for nonsmooth nonconvex multi-level composition optimization. *arXiv preprint arXiv:2001.10669*, 2020.
20. Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory.* SIAM, 2014.
21. P Tseng. On accelerated proximal gradient methods for convex-concave optimization, manuscript. *University of Washington, USA*, 2008.
22. Mengdi Wang, Ethan X Fang, and Han Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449, 2017.
23. Mengdi Wang, Ji Liu, and Ethan X Fang. Accelerating stochastic composition optimization. *The Journal of Machine Learning Research*, 18(1):3721–3743, 2017.
24. Shuoguang Yang, Mengdi Wang, and Ethan X Fang. Multilevel stochastic gradient methods for nested composition optimization. *SIAM Journal on Optimization*, 29(1):616–659, 2019.
25. Junyu Zhang and Lin Xiao. Multi-level composite stochastic optimization via nested variance reduction. *arXiv preprint arXiv:1908.11468*, 2019.
26. Zhe Zhang, Shabbir Ahmed, and Guanghui Lan. Efficient algorithms for distributionally robust stochastic optimization with discrete scenario support. *arXiv preprint arXiv:1909.11216*, 2019.