Stochastic Optimization for DC Functions and Non-smooth Non-convex Regularizers with Non-asymptotic Convergence

Yi Xu [1] **Qi Qi** [1] **Qihang Lin** [2] **Rong Jin** [3] **Tianbao Yang** [1]

Abstract

Difference of convex (DC) functions cover a broad family of non-convex and possibly non-smooth and non-differentiable functions, and have wide applications in machine learning and statistics. Although deterministic algorithms for DC functions have been extensively studied, stochastic optimization that is more suitable for learning with big data remains under-explored. In this paper, we propose new stochastic optimization algorithms and study their first-order convergence theories for solving a broad family of DC functions. We improve the existing algorithms and theories of stochastic optimization for DC functions from both practical and theoretical perspectives. Moreover, we extend the proposed stochastic algorithms for DC functions to solve problems with a general non-convex non-differentiable regularizer, which does not necessarily have a DC decomposition but enjoys an efficient proximal mapping. To the best of our knowledge, this is the first work that gives the first non-asymptotic convergence for solving non-convex optimization whose objective has a general non-convex non-differentiable regularizer.

1. Introduction

In this paper, we consider a family of non-convex non-smooth optimization problems that can be written in the following form:

$$\min_{\mathbf{x}\in\mathbb{R}^d} g(\mathbf{x}) + r(\mathbf{x}) - h(\mathbf{x}), \tag{1}$$

[1]Department of Computer Science, University of Iowa, Iowa City, IA 52242, USA [2]Department of Management Sciences, University of Iowa, Iowa City, IA 52242, USA [3]Machine Intelligence Technology, Alibaba Group, Bellevue, WA 98004, USA. Correspondence to: Yi Xu <yi-xu@uiowa.edu>, Tianbao Yang <tianbao-yang@uiowa.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

where $g(\cdot)$ and $h(\cdot)$ are real-valued lower-semicontinuous convex functions, $r(\cdot)$ is a proper lower-semicontinuous function. We include the component r in order to capture non-differentiable functions that usually play the role of regularization, e.g., the indicator function of a convex set $\mathcal{X}$ where $r(\mathbf{x}) = \delta_{\mathcal{X}}(\mathbf{x})$ is zero if $\mathbf{x} \in \mathcal{X}$ and infinity otherwise, and a non-differential regularizer such as the convex ℓ_1 norm $\|\mathbf{x}\|_1$ or the non-convex ℓ_0 norm and ℓ_p norm $\|\mathbf{x}\|_p^p$ with $p \in (0, 1)$. We do not necessarily impose smoothness condition on $g(\mathbf{x})$ or $h(\mathbf{x})$ and the convexity condition on $r(\mathbf{x})$.

A special class of the problem (1) is the one with $r(\mathbf{x})$ being a convex function - also known as difference of convex (DC) functions. We would like to mention that even the family of DC functions is broader enough to cover many interesting non-convex problems that are well-studied, including an additive composition of a smooth non-convex function and a non-smooth convex function, weakly convex functions, etc. We postpone this discussion to Section 2 after we formally introduce the definitions of smooth functions and weakly convex functions.

In the literature, deterministic algorithms for DC problems have been studied extensively since its introduction by Pham Dinh Tao in 1985 and are continuously receiving attention from the community (Khamaru & Wainwright, 2018; Wen et al., 2018). Please refer to (Thi & Dinh, 2018) for a survey on this subject. Although stochastic optimization (SO) algorithms for the special cases of DC functions mentioned above (smooth non-convex functions, weakly convex functions) have been studied recently (Davis & Grimmer, 2017; Davis & Drusvyatskiy, 2018b;a; Drusvyatskiy & Paquette, 2018; Chen et al., 2018b; Lan & Yang, 2018; Allen-Zhu, 2017; Chen & Yang, 2018; Allen-Zhu & Hazan, 2016; Reddi et al., 2016b;a; Zhang & He, 2018), a comprehensive study of SO algorithms with a broader applicability to the DC functions and the problem (1) with a non-smooth non-convex regularizer $r(\mathbf{x})$ still remain rare.

The papers by (Mairal, 2013), (Nitanda & Suzuki, 2017) and (Thi et al., 2017) are the most related works dedicated to the stochastic optimization of special DC functions. Mairal (2013) studied a special case of problem (1) with h is smooth and proposed a stochastic majorization-minimization algorithm enjoying an asymptotic convergence result for finding a stationary point. Thi et al. (2017) considered a special class of DC problems and they reformulated the problem into (1) such that h is a sum of n convex functions, and g is a quadratic function and r is the first component of the DC decomposition of the regularizer. Then, they proposed a stochastic variant of the classical DCA (Difference-of-Convex Algorithm) and established an asymptotic convergence result for finding a critical point. To our knowledge, the paper by (Nitanda & Suzuki, 2017) is the probably the first result that gives non-asymptotic convergence for finding an approximate critical point of a special class of DC problems, in which both g and h can be stochastic functions and $r = 0$. Their algorithm consists of multiple stages of solving a convex objective that is constructed by linearizing $h(\mathbf{x})$ and adding a quadratic regularization. However, their algorithm and convergence theory have the following drawbacks. First, at each stage, they need to compute an unbiased stochastic gradient denoted by $\mathbf{v}(\mathbf{x})$ of $\nabla h(\mathbf{x})$ such that $\mathrm{E}[\|\mathbf{v}(\mathbf{x}) - \nabla h(\mathbf{x})\|^2] \leq \epsilon^2$, where ϵ is the accuracy level imposed on the returned solution in terms of the gradient's norm. In reality, one has to resort to mini-batching technique by using a large number of samples to ensure this condition, which is impractical and not user-friendly. An user has to worry about what is the size of the mini-batch in order to find a sufficiently accurate solution while keeping the computational costs minimal. Second, for each constructed convex subproblem, their theory requires running a stochastic algorithm that solves each subproblem to the accuracy level of ϵ, which could waste a lot of computations at earlier stages. Third, their convergence analysis requires that $r(\mathbf{x}) = 0$ and $g(\mathbf{x})$ is a smooth function with a Lipchitz continuous gradient. In addition, they obtained fast convergence result of the problem under Polyak-Łojasiewicz condition, which is not considered in this paper.

Our Contributions - I. In Section 3, we propose new stochastic optimization algorithms and establish their convergence results for solving the DC class of the problem (1) that improves the algorithm and theory in (Nitanda & Suzuki, 2017) from several perspectives. It is our intention to address the aforementioned drawbacks of their algorithm and theory. In particular, (i) our algorithm only requires unbiased stochastic (sub)-gradients of $g(\mathbf{x})$ and $h(\mathbf{x})$ without a requirement on the small variance of the used stochastic (sub)-gradients; (ii) we do not need to solve each constructed subproblem to the accuracy level of ϵ. Instead, we allow the accuracy for solving each constructed subproblem to grow slowly without sacrificing the overall convergence rate; (iii) we improve the convergence theory significantly. First, our convergence analysis does not require $g(\mathbf{x})$ to be smooth with a Lipchitz continuous gradient. Instead, we only require either $g(\mathbf{x}) + r(\mathbf{x})$ or $h(\mathbf{x})$ to be differentiable with a Hölder continuous gradient, under the former condition $h(\mathbf{x})$ can be a non-smooth non-differentiable function and under the later condition $r(\mathbf{x})$ and $g(\mathbf{x})$ can be non-smooth non-differentiable functions. Second, the convergence rate is automatically adaptive to the Hölder continuity of the involved function without requiring the knowledge of the Hölder continuity to run the algorithm. Third, when adaptive stochastic gradient method is employed to solve each subproblem, we establish an adaptive convergence similar to existing theory of AdaGrad for convex problems (Duchi et al., 2011; Chen et al., 2018a) and weakly convex problems (Chen et al., 2018b), which is missing in (Nitanda & Suzuki, 2017).

Our Contributions - II. Moreover, in Section 4 we extend our algorithm and theory to the more general class of non-convex non-smooth problem (1), in which $r(\mathbf{x})$ is a general non-convex non-differentiable regularizer that enjoys an efficient proximal mapping. Although such kind of non-smooth non-convex regularization has been considered in literature (Attouch et al., 2013; Bolte et al., 2014; Bot et al., 2016; Li & Lin, 2015; Yu et al., 2015; Yang, 2018; Liu et al., 2018; An & Nam, 2017; Zhong & Kwok, 2014), existing results are restricted to deterministic optimization and asymptotic or local convergence analysis. In addition, most of them consider a special case of our problem with $g - h$ being a smooth non-convex function. To the best of our knowledge, this is the first work of stochastic optimization with a non-asymptotic first-order convergence result for tackling the non-convex objective (1) with a non-convex non-differentiable regularization and a smooth function g and a possibly non-smooth function h with a Hölder continuous gradient. Our algorithm and theory are based on using the Moreau envelope of $r(\mathbf{x})$ that can be written as a DC function, which then reduces to the problem that is studied in Section 3. By using the algorithms and their convergence results established in Section 3 and carefully controlling the approximation parameter, we establish the first non-asymptotic convergence of stochastic optimization for solving the original non-convex problem with a non-convex non-differentiable regularizer. This non-asymptotic convergence result can be also easily extended to the deterministic optimization, which itself is novel and could be interesting to a broader community. A summary of our results is presented in Table 1.

2. Preliminaries

In this section, we present some preliminaries. Let $\|\cdot\|_p$ denote the standard p-norm with $p \geq 0$. For a non-convex function $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$, let $\hat{\partial} f(\mathbf{x})$ denote the Fréchet subgradient and $\partial f(\mathbf{x})$ denote the limiting subgradient, i.e.,

$$\hat{\partial} f(\bar{\mathbf{x}}) = \left\{ \mathbf{v} \in \mathbb{R}^d : \lim_{\mathbf{x} \to \bar{\mathbf{x}}} \inf \frac{f(\mathbf{x}) - f(\bar{\mathbf{x}}) - \mathbf{v}^\top (\mathbf{x} - \bar{\mathbf{x}})}{\|\mathbf{x} - \bar{\mathbf{x}}\|} \geq 0 \right\},$$

$$\partial f(\bar{\mathbf{x}}) = \{ \mathbf{v} \in \mathbb{R}^d : \exists \mathbf{x}_k \xrightarrow{f} \bar{\mathbf{x}}, v_k \in \hat{\partial} f(\mathbf{x}_k), \mathbf{v}_k \to \mathbf{v} \},$$

Table 1. Summary of our results for finding a (nearly) ϵ-critical point of the problem (1), where g and h are assumed to be convex. HC refers to Hölder continuous gradient condition; SM refers to the smooth condition; CX means convex; NC means non-convex and NS means non-smooth; LP denotes Lipchitz continuous function; LB means lower bounded over $\mathbb{R}^d$; FV means finite-valued over $\mathbb{R}^d$; FVC means finite-valued over a compact set. $\nu \in (0, 1]$ denotes the power constant of the involved function's Hölder continuity. n denotes the total number of components in a finite-sum problem. SPG denotes stochastic proximal gradient algorithm. SVRG denotes stochastic variance reduced gradient algorithm. AdaGrad denotes adaptive stochastic gradient method. Complexity for SPG and AdaGrad means iteration complexity, and for SVRG and AG means gradient complexity. **Better** results were obtained in the arXiv version after ICML.

g	h	r	Algorithms for subproblems	Complexity
-	HC	CX	SPG, AdaGrad	$O(1/\epsilon^{4/\nu})$
SM	HC	CX	SVRG	$O(n/\epsilon^{2/\nu})$
HC	-	CX, HC	SPG, AdaGrad	$O(1/\epsilon^{4/\nu})$
SM	-	CX, HC	SVRG	$O(n/\epsilon^{2/\nu})$
SM	HC	NC, NS, LP	SPG, AdaGrad	$O(1/\epsilon^{4(1+1/\nu)})$
SM	HC	NC, NS, FV, LB	SPG, AdaGrad	$O(1/\epsilon^{4(1+2/\nu)})$
SM	HC	NC, NS, LP	SVRG	$O(n/\epsilon^{2(1+1/\nu)})$
SM	HC	NC, NS, FV, LB	SVRG	$O(n/\epsilon^{2(1+2/\nu)})$
SM	HC	NC, NS, FVC	SVRG	$O(n/\epsilon^{2(1+2/\nu)})$

where the notation $\mathbf{x} \xrightarrow{f} \bar{\mathbf{x}}$ means that $\mathbf{x} \to \bar{\mathbf{x}}$ and $f(\mathbf{x}) \to f(\bar{\mathbf{x}})$. It is known that $\hat{\partial} f(\mathbf{x}) \subseteq \partial f(\mathbf{x})$. If $f(\cdot)$ is differential at $\mathbf{x}$, then $\hat{\partial} f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$. Moreover, if $f(\mathbf{x})$ is continuously differentiable on a neighborhood of $\mathbf{x}$, then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$. When f is convex, the Fréchet and the limiting subgradient reduce to the subgradient in the sense of convex analysis: $\partial f(\mathbf{x}) = \{\mathbf{v} \in \mathbb{R}^d : f(\mathbf{x}) \geq f(\mathbf{y}) + \mathbf{v}^\top(\mathbf{x} - \mathbf{y}), \forall \mathbf{y} \in \mathbb{R}^d\}$. For simplicity, we use $\|\cdot\|$ to denote the Euclidean norm (aka. 2-norm) of a vector. Let $\text{dist}(\mathcal{S}_1, \mathcal{S}_2)$ denote the distance between two sets and $[K] = \{1, \ldots, K\}$.

A function $f(\mathbf{x})$ is smooth with a L-Lipchitz continuous gradient if it is differentiable and the following inequality holds $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y}$. A differentiable function $f(\mathbf{x})$ has (L, ν)-Hölder continuous gradient if there exists $\nu \in (0, 1]$ such that $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|^\nu, \forall \mathbf{x}, \mathbf{y}$. Next, let us characterize the critical points of the considered problem (1) that are standard in the literature (Hiriart-Urruty, 1985; Horst & Thoai, 1999; Thi & Dinh, 2018; An & Nam, 2017), and introduce the convergence measure of an algorithm. First, let us consider the DC problem:

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) := g(\mathbf{x}) - h(\mathbf{x}), \tag{2}$$

where $g : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a proper lower semicontinuous convex function and $h : \mathbb{R}^d \to R$ is convex. Any point $\bar{\mathbf{x}}$ such that $\partial h(\bar{\mathbf{x}}) \cap \partial g(\bar{\mathbf{x}}) \neq \emptyset$ is called a critical point of (2), which is a necessary condition for $\bar{\mathbf{x}}$ to be a local minimizer. For an iterative optimization algorithm, it is hard to find an exactly critical point in a finite-number of iterations. Therefore, we find an ϵ-critical point $\mathbf{x}$ that satisfies

$$\text{dist}(\partial h(\mathbf{x}), \partial g(\mathbf{x})) \leq \epsilon. \tag{3}$$

Similarly, we can extend the above definition of critical points to the general problem (1) with $r(\mathbf{x})$ being a proper and lower semi-continuous (possibly non-convex) function (An & Nam, 2017). In particular, any point $\bar{\mathbf{x}}$ such that $\partial h(\bar{\mathbf{x}}) \cap \hat{\partial}(g + r)(\bar{\mathbf{x}}) \neq \emptyset$ is called a critical point of (1). When g is differentiable, $\hat{\partial}(g + r)(\bar{\mathbf{x}}) = \nabla g(\mathbf{x}) + \hat{\partial} r(\mathbf{x})$ (Rockafellar & Wets, 1998)[Exercise 8.8], and when both g and r are convex and their domains cannot be separated $\hat{\partial}(g + r)(\bar{\mathbf{x}}) = \partial g(\mathbf{x}) + \partial r(\mathbf{x})$ (Rockafellar & Wets, 1998)[Corollary 10.9]. An ϵ-critical point of (1) is a point $\mathbf{x}$ that satisfies $\text{dist}(\partial h(\mathbf{x}), \hat{\partial}(g + r)(\mathbf{x})) \leq \epsilon$. It is notable that when $g + r$ is non-differentiable, finding an ϵ-critical point could become a challenging task for an iterative algorithm even under the condition that r is a convex function. Let us consider the example of $g = |x|, h = r = 0$. As long as $x \neq 0$, we have $\text{dist}(0, \partial|x|) = 1$. To address this challenge when $g + r$ is non-differentiable, we introduce the notion of nearly ϵ-critical points. In particular, a point $\mathbf{x}$ is called a nearly ϵ-critical point of the problem (1) if there exists $\bar{\mathbf{x}}$ such that

$$\|\mathbf{x} - \bar{\mathbf{x}}\| \leq O(\epsilon), \quad \text{dist}(\partial h(\bar{\mathbf{x}}), \hat{\partial}(g + r)(\bar{\mathbf{x}})) \leq \epsilon. \tag{4}$$

A similar notion of nearly critical points for non-smooth and non-convex optimization problems have been utilized in several recent works (Davis & Grimmer, 2017; Davis & Drusvyatskiy, 2018b;a; Chen et al., 2018b).

Examples and Applications of DC functions.

Example 1: Weakly convex functions. Weakly convex functions have been recently studied in numerous papers (Davis & Grimmer, 2017; Davis & Drusvyatskiy, 2018b;a; Chen et al., 2018b; Zhang & He, 2018). A function $f(\mathbf{x})$ is called ρ-weakly convex if $f(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{x}\|^2$ is a convex function. More generally, $f(\mathbf{x})$ is called ρ-relative convex with respect to a strongly convex function $\omega(\mathbf{x})$ if $f(\mathbf{x}) + \rho\omega(\mathbf{x})$ is convex (Zhang & He, 2018). It is obvious that a weakly convex function $f(\mathbf{x})$ is a DC function. Examples of weakly convex functions can be found in deep neural networks with a smooth active function and a smooth/non-smooth loss function (Chen et al., 2018b), robust learning (Xu et al., 2018), robust phase retrieval (Davis & Drusvyatskiy, 2018a).

Example 2: Non-Convex Sparsity-Promoting Regularizers. Many non-convex sparsity-promoting regularizers in statistics can be written as a DC function, including log-sum penalty (LSP) (Candès et al., 2008), minimax concave penalty (MCP) (Zhang, 2010a), smoothly clipped absolute deviation (SCAD) (Fan & Li, 2001), capped ℓ_1 penalty (Zhang, 2010b), transformed ℓ_1 norm (Zhang & Xin, 2018). For detailed DC composition of these regularizers, please refer to (Wen et al., 2018; Gong et al., 2013). We also present the details in the supplement.

Example 3: Least-squares Regression with ℓ_{1-2} Regularization. Recently, a non-convex regularization in the form of $\lambda(\|\mathbf{x}\|_1 - \|\mathbf{x}\|_2)$ was proposed for least-squares regression or compressive sensing (Yin et al., 2015), which is naturally a DC function.

Example 4: Positive-Unlabeled (PU) Learning. In PU learning for binary classification, only positive data $\{(\mathbf{z}_i, +1), i = 1, \ldots, n_+\}$ are observed where $\mathbf{z}_i \in \mathbb{R}^m$ denotes the feature vector of i-th positive example, conventional empirical risk minimization becomes problematic. A remedy to address this challenge is to use unlabeled data for computing an unbiased estimation of $\mathrm{E}_{\mathbf{z},y}[\ell(\mathbf{x}; \mathbf{z}, y)]$, where $\mathbf{y} \in \{1, -1\}$ denotes the label. In particular, the objective in the following problem is an unbiased risk (Kiryo et al., 2017): $\min_{\mathbf{x}\in\mathbb{R}^d} \frac{\pi_p}{n_+}\sum_{i=1}^{n_+}(\ell(\mathbf{x}; \mathbf{z}_i, 1) - \ell(\mathbf{x}; \mathbf{z}_i, -1)) + \frac{\sum_{j=1}^{m}\ell(\mathbf{x};\mathbf{z}_j^u,-1)}{n_u}$, where $\{\mathbf{z}_i^u, i = 1, \ldots, n_u\}$ is a set of unlabeled data, and $\pi_p = \Pr(y = 1)$ is the prior probability of the positive class. It is obvious that if $\ell(\mathbf{x}; \cdot)$ is a convex loss function in terms of $\mathbf{x}$, the above objective function is a DC function. In practice, an estimation of π_p is used.

Examples of Non-Convex Non-Smooth Regularizers. Finally, we present some examples of non-convex non-smooth regularizers $r(\mathbf{x})$ that cannot be written as a DC function or whose DC decomposition is unknown. Thus, the algorithms and theories presented in Section 3 are not directly applicable, but the algorithms discussed in Section 4 are applicable when the proximal mapping of $r(\mathbf{x})$ is efficient to compute. Examples include ℓ_0 norm (i.e., the number of non-zero elements of a vector) and ℓ_p norm regularization for $p \in (0, 1)$ (i.e., $\sum_{i=1}^{d}|x_i|^p$), whose proximal mapping can be efficiently computed (Attouch et al., 2013; Bolte et al., 2014). Let us consider a non-convex optimization problem with domain constraint $\mathbf{x} \in \mathcal{C}$, where $\mathcal{C}$ is a non-convex set. Directly handling a non-convex constrained problem could be difficult. An alternative solution is to convert the constraint into a penalization $r(\mathbf{x}) = \frac{\lambda}{2}\|\mathbf{x} - \mathbb{P}_{\mathcal{C}}(\mathbf{x})\|^2$ with $\lambda > 0$ in the objective, where $\mathbb{P}_{\mathcal{C}}(\cdot)$ denotes the projection of a point to the set $\mathcal{C}$. Note that when $\mathcal{C}$ is a non-convex set, $r(\mathbf{x})$ is a non-convex non-smooth function in general, and its proximal mapping enjoys a closed-form solution (Li & Pong, 2016).

3. New Stochastic Algorithms of DC functions

In this section, we present new stochastic algorithms for solving the problem (1) when $r(\mathbf{x})$ is a convex function and their convergence results. We assume both $g(\mathbf{x})$ and $h(\mathbf{x})$ have a large number of components such that computing a stochastic gradient is much more efficient than computing a deterministic gradient. Without loss of generality, we assume $g(\mathbf{x}) = \mathrm{E}_\xi[g(\mathbf{x}; \xi)]$ and $h(\mathbf{x}) = \mathrm{E}_\varsigma[h(\mathbf{x}; \varsigma)]$, and consider the following problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) := \mathrm{E}_\xi[g(\mathbf{x}; \xi)] + r(\mathbf{x}) - \mathrm{E}_\varsigma[h(\mathbf{x}; \varsigma)]. \quad (5)$$

where g and h are real-valued lower-semicontinuous convex functions and r is a proper lower-semicontinuous convex function. It is notable that a special case of this problem is the finite-sum form:

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) := \frac{1}{n_1}\sum_{i=1}^{n_1} g_i(\mathbf{x}) + r(\mathbf{x}) - \frac{1}{n_2}\sum_{j=1}^{n_2} h_j(\mathbf{x}), \quad (6)$$

which allows us to develop faster algorithms for smooth functions by using variance reduction techniques.

Since we do not necessarily impose any smoothness assumption on $g(\mathbf{x})$ and $h(\mathbf{x})$, we will postpone the particular assumptions for these functions in the statements of later theorems. For all algorithms presented below, we assume that the **proximal mapping** of $r(\mathbf{x})$ can be efficiently computed, i.e., $\mathrm{prox}_{\eta r}(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2\eta}\|\mathbf{x} - \mathbf{y}\|^2 + r(\mathbf{x})$ can be easily computed for any $\eta > 0$. But it is not necessary for developing subgradient methods when r is convex.

The basic idea of the proposed algorithm is similar to the stochastic algorithm proposed in (Nitanda & Suzuki, 2017). The algorithm consists of multiple stages of solving convex problems. At the k-th stage ($k \geq 1$), given a point $\mathbf{x}_k$, a convex majorant function $F^\gamma_{\mathbf{x}_k}(\mathbf{x})$ is constructed as following such that $F^\gamma_{\mathbf{x}_k}(\mathbf{x}) \geq F(\mathbf{x}), \forall \mathbf{x}$ and $F^\gamma_{\mathbf{x}_k}(\mathbf{x}_k) = F(\mathbf{x}_k)$:

$$\begin{aligned} F^\gamma_{\mathbf{x}_k}(\mathbf{x}) =& g(\mathbf{x}) + r(\mathbf{x}) - (h(\mathbf{x}_k) + \partial h(\mathbf{x}_k)^\top(\mathbf{x} - \mathbf{x}_k)) \\ &+ \frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}_k\|^2, \end{aligned} \quad (7)$$

where $\gamma > 0$ is a constant parameter. Then a stochastic algorithm is employed to optimize $F^\gamma_{\mathbf{x}_k}$. The key difference from the previous work lies at how to solve each convex majorant function. An important change introduced to our design is to make the proposed algorithms more efficient and more practical. Roughly speaking, we only require solving each function $F^\gamma_{\mathbf{x}_k}(\mathbf{x})$ up to an accuracy level of c/k for some constant $c > 0$, i.e., finding $\mathbf{x}_{k+1}$ such that

$$\mathrm{E}[F^\gamma_{\mathbf{x}_k}(\mathbf{x}_{k+1}) - \min_{\mathbf{x}\in\mathbb{R}^d} F^\gamma_{\mathbf{x}_k}(\mathbf{x})] \leq c/k. \quad (8)$$

In contrast, the results presented in (Nitanda & Suzuki,

Algorithm 1 Stagewise Stochastic DC (SSDC) Algorithm

1: **Initialize:** $\mathbf{x}_1 \in \text{dom}(r)$
2: **for** $k = 1, \ldots, K$ **do**
3: Let $F_k(\mathbf{x}) = F^{\gamma}_{\mathbf{x}_k}$ as defined in (7)
4: $\mathbf{x}_{k+1} = \mathcal{A}(F^{\gamma}_{\mathbf{x}_k}, \Theta_k)$ $\diamond$ Θ_k denotes algorithm dependent parameters
5: **end for**

2017) require solving each convex problem up to an accuracy level of ϵ, which is the expected accuracy level on the final solution. This change not only makes our algorithms more efficient by saving unnecessary computations but also more practical without requiring ϵ to run the algorithm. We present a meta algorithm in Algorithm 1, in which $\mathcal{A}$ refers to an appropriate stochastic algorithm for solving each convex majorant function. The Step 4 means that $\mathcal{A}$ is employed for finding $\mathbf{x}_{k+1}$ such that (8) is satisfied (or a more fine-grained condition is satisfied for a particular algorithm as discussed later), where Θ_k denotes the algorithm dependent parameters (e.g., the number of iterations). Our convergence analysis also has its merits compared with the previous work (Nitanda & Suzuki, 2017). We will divide our convergence analysis into three parts. First, in subsection 3.1 we introduce a general convergence measure without requiring any smoothness assumptions of involved functions and conduct a convergence analysis of the proposed algorithm. Second, we analyze different stochastic algorithms and their convergence results in subsection 3.2, including an adaptive convergence result for using AdaGrad. Finally, we discuss the implications of these convergence results for solving the original problem in terms of finding a (nearly) ϵ-stationary point in subsection 3.3.

3.1. A General Convergence Result

For any $\gamma > 0$, define $P_\gamma(\mathbf{z}) = \arg\min_{\mathbf{x}\in\mathbb{R}^d} F^{\gamma}_{\mathbf{z}}(\mathbf{x})$, $G_\gamma(\mathbf{z}) = \gamma(\mathbf{z} - P_\gamma(\mathbf{z}))$. It is notable that $P_\gamma(\mathbf{z})$ is well defined since $F^{\gamma}_{\mathbf{z}}$ is strongly convex. The following proposition shows that when $\mathbf{z} = P_\gamma(\mathbf{z})$, then $\mathbf{z}$ is a critical point of the original problem.

Proposition 1. *If $\mathbf{z} = P_\gamma(\mathbf{z})$, then $\mathbf{z}$ is a critical point of the problem $\min_{\mathbf{x}\in\mathbb{R}^d} g(\mathbf{x}) + r(\mathbf{x}) - h(\mathbf{x})$.*

The above proposition implies that $\|G_\gamma(\mathbf{z})\| = \gamma\|P_\gamma(\mathbf{z}) - \mathbf{z}\|$ can serve as a measure of convergence of an algorithm for solving the considered minimization problem. In subsection 3.3, we will discuss how the convergence in terms of $\gamma\|P_\gamma(\mathbf{z}) - \mathbf{z}\|$ implies the standard convergence in terms of the (sub)gradient norm of the original problem. We use the following basic assumption for our analysis.

Assumption 1. *For an initial solution $\mathbf{x}_1 \in dom(r)$, assume that $F(\mathbf{x}_1) - \inf_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \leq \Delta$ for some $\Delta > 0$.*

The theorems below are the main results of this subsection.

Theorem 1. *Suppose Assumption 1 holds and there exists an stochastic algorithm $\mathcal{A}$ that when applied to $F^{\gamma}_{\mathbf{x}_k}(\mathbf{x})$ can find a solution $\mathbf{x}_{k+1}$ satisfying (8), then we have $\mathrm{E}[\|G_\gamma(\mathbf{x}_\tau)\|^2] \leq (2\gamma\Delta + 2\gamma c(1 + \log(K)))/K$, where $\tau \in [K]$ is uniformly sampled.*

Remark: It is clear that when $K \to \infty$, $\gamma\|\mathbf{x}_\tau - P_\gamma(\mathbf{x}_\tau)\| \to 0$ in expectation, implying the convergence to a critical point. Note that the $\log(K)$ factor will lead to an iteration complexity of $O(\log(1/\epsilon)/\epsilon^4)$ for using stochastic (sub)gradient method. Nevertheless, such a logarithmic factor can be removed by exploiting non-uniform sampling under a slightly stronger condition of the problem.

Theorem 2. *Suppose there exists a stochastic algorithm $\mathcal{A}$ that when applied to $F^{\gamma}_{\mathbf{x}_k}(\mathbf{x})$ can find a solution $\mathbf{x}_{k+1}$ satisfying (8), and there exists $\Delta > 0$ such that $\mathrm{E}[F(\mathbf{x}_k) - \min_{\mathbf{x}} F(\mathbf{x})] \leq \Delta$ for all $k \in [K]$, then we have $\mathrm{E}[\|G_\gamma(\mathbf{x}_\tau)\|^2] \leq \frac{2\gamma(\Delta+c)(\alpha+1)}{K}$, where $\tau \in [K]$ is sampled according to probabilities $p(\tau = k) = k^\alpha / \sum_{k=1}^K k^\alpha$ with $\alpha \geq 1$.*

Remark: Compared to Theorem 1, the condition $\mathrm{E}[F(\mathbf{x}_k) - \min_{\mathbf{x}} F(\mathbf{x})] \leq \Delta$ for all $k \in [K]$ is slightly stronger than Assumption 1. However, it can be easily satisfied if $\mathbf{x}_k \in \text{dom}(r)$ resides in a bounded set (e.g., when $r(\mathbf{x})$ is the indicator function of a bounded set), or if $\mathrm{E}[F(\mathbf{x}_k)]$ is non-increasing (e.g., when using variance-reduction methods for the case that $g(\mathbf{x})$ is smooth). In the following presentation, we assume this condition holds without explicitly mentioned.

3.2. Convergence Results of Different Algorithms

In this section, we will present the convergence results of Algorithm 1 for employing different stochastic algorithms to minimize $F_k(\mathbf{x})$ at each stage. In particular, we consider three representative algorithms, namely stochastic proximal subgradient (SPG) method (Duchi et al., 2010; Zhao & Zhang, 2015), adaptive stochastic gradient (AdaGrad) method (Duchi et al., 2011; Chen et al., 2018a), and proximal stochastic gradient method with variance reduction (SVRG) (Xiao & Zhang, 2014). We refer to Algorithm 1 by using SPG, AdaGrad, SVRG for solving each subproblem as SSDC-SPG, SSDC-AdaGrad, SSDC-SVRG, respectively.

SPG. We make the additional assumptions about the problem for developing SPG, which are typical in the literature (Zhao & Zhang, 2015; Duchi et al., 2010).

Assumption 2. *Assume one of the following conditions: (i) $g(\mathbf{x})$ is L-smooth and $\mathrm{E}[\|(\nabla g(\mathbf{x};\xi) - \partial h(\mathbf{x};\varsigma)) - \mathrm{E}[\nabla g(\mathbf{x};\xi) - \partial h(\mathbf{x};\varsigma)]\|^2] \leq G^2$. (ii) $\mathrm{E}[\|\partial g(\mathbf{x};\xi)\|^2] \leq G^2$, $\mathrm{E}[\|\partial h(\mathbf{x};\varsigma)\|^2\}] \leq G^2$ for $\mathbf{x} \in dom(r)$, and either $r = \delta_{\mathcal{X}}(\mathbf{x})$ for a closed convex set $\mathcal{X}$ or $\|\partial r(\mathbf{x})\| \leq G$ for $\mathbf{x} \in dom(r)$.*

Without loss of generality, we consider minimizing $F^{\gamma}_{\mathbf{x}_1}$ by

SPG. The key update of SPG is the following:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\Omega}\{\mathbf{x}^\top \mathcal{G}(\mathbf{x}_t;\xi_t,\varsigma_t) + r(\mathbf{x}) + \frac{\gamma}{2}\|\mathbf{x}-\mathbf{x}_1\|^2 + \frac{1}{2\eta_t}\|\mathbf{x}-\mathbf{x}_t\|^2\}, t = 1,\dots,T \qquad (9)$$

where $\mathcal{G}(\mathbf{x}_t;\xi_t,\varsigma_t) = \partial g(\mathbf{x}_t;\xi_t) - \partial h(\mathbf{x}_1;\varsigma_t)$. For smooth g, we set $\Omega = \mathbb{R}^d$, and for non-smooth g we set $\Omega = \mathcal{B}_{\mathbf{x}_1} = \{\mathbf{x}\in\mathbb{R}^d, \|\mathbf{x}-\mathbf{x}_1\| \le 3G/\gamma\}$. Restricting the solution to the ball $\mathcal{B}_{\mathbf{x}_1}$ is to accommodate the proximal mapping of $r(\mathbf{x})$ when $g(\mathbf{x})$ is non-smooth. When using the subgradient of $r(\mathbf{x})$ instead of the proximal mapping of $r(\mathbf{x})$ in the update or $r(\mathbf{x})$ is the indicator function of a bounded convex set, the projection onto $\mathcal{B}_{\mathbf{x}_1}$ can be removed. The complete steps of the SPG algorithm are presented in Algorithm 3 in the supplement with the two options to handle smooth and non-smooth g separately. The convergence of Algorithm 1 when using SPG to solve each subproblem is stated below.

Theorem 3. *Suppose Assumption 2 (i) holds and Algorithm 3 with (9) ($\Omega = \mathbb{R}^d$) is employed for solving F_k with $\eta_t = 1/(L(t+1))$, $\gamma \ge 3L$ and $T_k = 4k/c$ iterations where $c \in (0,1]$, then Algorithm 1 guarantees $\mathrm{E}[\|G_\gamma(\mathbf{x}_\tau)\|^2] \le \frac{8\gamma\Delta(\alpha+1)}{K} + \frac{8cG^2\gamma(\alpha+1)}{LK}$. Similarly, Suppose Assumption 2 (ii) holds and Algorithm 3 with (9) ($\Omega = \mathcal{B}_{\mathbf{x}_{k-1}}$) is employed for solving F_k with $\eta_t = 4/(\gamma t)$, and $T_k = k/c$ iterations with $c \in (0,1]$, then $\mathrm{E}[\|G_\gamma(\mathbf{x}_\tau)\|^2] \le \frac{8\gamma\Delta(\alpha+1)}{K} + \frac{448c(\alpha+1)}{K}$, where τ is sampled similarly as in Theorem 2.*

Remark: Let us consider the iteration complexity of using SPG for finding a solution that satisfies $\mathrm{E}[\|G_\gamma(\mathbf{x}_\tau)\|^2] \le \epsilon^2$. For the non-smooth case, by setting $\gamma < 1$ and $c = \gamma$, we need a total number of stages $K = O(\gamma/\epsilon^2)$ and total iteration complexity $\sum_{k=1}^K T_k = \sum_{k=1}^K k/\gamma = O(1/\epsilon^4)$. For the smooth case, by setting $c = 1$ we have $K = O(\max(L,1)/\epsilon^2)$ and total iteration complexity $\sum_{k=1}^K T_k = \sum_{k=1}^K 4k = O(\max(L,1)/\epsilon^4)$.

AdaGrad. AdaGrad (Duchi et al., 2011) is an important algorithm in the literature of stochastic optimization, which uses adaptive step size for each coordinate. It has potential benefit of speeding up the convergence when the cumulative growth of stochastic gradient is slow. Next, we show that AdaGrad can be leveraged to solve each convex majorant function and yield adaptive convergence for the original problem. Similar to (Duchi et al., 2011; Chen et al., 2018a), we make the following assumption.

Assumption 3. *For any $\mathbf{x} \in dom(r)$, there exists $G > 0$ such that $\max(\|\partial g(\mathbf{x};\xi)\|_\infty, \|\partial h(\mathbf{x};\varsigma)\|_\infty) \le G$, either $r = \delta_{\mathcal{X}}(\mathbf{x})$ for a closed convex set $\mathcal{X}$ or $\|\partial r(\mathbf{x})\| \le G_r$.*

The convergence result of Algorithm 1 by using AdaGrad to solve each problem is described by following theorem.

Theorem 4. *Suppose Assumption 3 holds and Algorithm 2 is employed for solving F_k with $\eta_k = c/\sqrt{k}$, T_k being the minimum number that is larger than $M_k \max\{a(2G + \max_i \|g^k_{1:T_k,i}\|), \sum_{i=1}^d \|g^k_{1:T_k,i}\|/a, \frac{G_r}{\eta_k}\|\mathbf{x}^k_1 - \mathbf{x}^k_{T_k+1}\|\}$ where $M_k\eta_k \ge 4/(a\gamma)$, then Algorithm 1 guarantees $\mathrm{E}[\|G_\gamma(\mathbf{x}_\tau)\|^2] \le \frac{8\gamma\Delta(\alpha+1)}{K} + \frac{4\gamma^2c^2a(a+1)(\alpha+1)}{K}$, where $g^k_{1:t,i}$ denotes the cumulative stochastic gradient of the i-th coordinate at the k-th stage, and τ is sampled similarly as in Theorem 2.*

Algorithm 2 ADAGRAD($F^\gamma_{\mathbf{x}_1}, \mathbf{x}_1, \eta$)

Initialize: $t = 1$, $\mathbf{g}_{1:0} = []$, $H_0 \in \mathbb{R}^{d\times d}$, $\Omega = \{\mathbf{x} \in \mathrm{dom}(r) : \|\mathbf{x}-\mathbf{x}_1\| \le \frac{2\sqrt{d}G+G_r}{\gamma}\}$
while T doesn't satisfy the condition in Thm. 4 **do**
  Compute $\mathbf{g}_t = \partial g(\mathbf{x}_t;\xi_t) - \partial h(\mathbf{x}_1;\varsigma_t)$
  Update $g_{1:t} = [g_{1:t-1}, \mathbf{g}_t]$, $s_{t,i} = \|g_{1:t,i}\|_2$
  Set $H_t = H_0 + \mathrm{diag}(\mathbf{s}_t)$
  Let $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\Omega} \mathbf{x}^\top\left(\frac{1}{t}\sum_{\tau=1}^t \mathbf{g}_\tau\right) + r(\mathbf{x}) + \frac{\gamma}{2}\|\mathbf{x}-\mathbf{x}_1\|^2 + \frac{1}{t\eta}\frac{1}{2}(\mathbf{x}-\mathbf{x}_1)^\top H_t(\mathbf{x}-\mathbf{x}_1)$
end while
Output: $\widehat{\mathbf{x}}_T = \sum_{t=1}^T \mathbf{x}_t/T$

Remark: a is a parameter used to balance the two involved terms for minimizing the value of T_k. It is obvious that the total number of iterations $\sum_{k=1}^K T_k$ is adaptive to the data. Next, let us present more discussion on the iteration complexity. Note that $M_k = O(\sqrt{k})$. By the boundness of stochastic gradient $\|g^k_{1:T_k,i}\| \le O(\sqrt{T_k})$, therefore T_k in the order of $O(k)$ will satisfy the condition in Theorem 4. Thus in the worst case, the iteration complexity for finding $\mathrm{E}[\|G_\gamma(\mathbf{x}_\tau)\|^2] \le \epsilon^2$ is in the order of $\sum_{k=1}^K O(k) \le O(1/\epsilon^4)$. We can show the potential advantage of adaptiveness similar to that in (Chen et al., 2018b). In particular, let us consider $r = \delta_{\mathcal{X}}$ and thus $G_r = 0$ in the above result. When the cumulative growth of stochastic gradient is slow, e.g., assuming $\|g^k_{1:T_k,i}\| \le O({T_k}^\beta)$ with $\beta < 1/2$. Then $T_k = O(k^{1/(2(1-\beta))})$ will work, and then the total number of iterations $\sum_{k=1}^K T_k \le K^{1+1/(2(1-\alpha))} \le O(1/\epsilon^{2+1/(1-\alpha)})$, which is better than $O(1/\epsilon^4)$.

SVRG. Next, we discuss SVRG (a variance reduction method) for solving each subproblem when it has a finite-sum form (6) and g is a smooth function. It is notable that the smoothness of h is not necessary for developing the SVRG algorithm since at each stage we linearize $h(\mathbf{x})$. We can use the proximal SVRG proposed in (Xiao & Zhang, 2014) to minimize $F^\gamma_{\mathbf{x}_{k-1}}$, which is presented in Algorithm 4 in the supplement, whose convergence result is stated below.

Theorem 5. *Suppose Assumption 1 and g is smooth, and SVRG (Algorithm 4) is employed for solving F_k with $\eta_k = 0.05/L$, $T_k \ge \max(2, 200L/\gamma)$, $S_k = \lceil\log_2(k)\rceil$, then Algorithm 1 guarantees $\mathrm{E}[\|G_\gamma(\mathbf{x}_\tau)\|^2] \le 12\gamma\Delta(\alpha+1)/K$, where τ is sampled similarly as in Theorem 2.*

Remark: For finding a solution such that $\mathrm{E}[\|G_\gamma(\mathbf{x}_\tau)\|^2] \le \epsilon^2$, the total number of stages $K = O(\gamma/\epsilon^2)$ and the total gradient complexity is $\widetilde{O}((n\gamma + L)/\epsilon^2)$.

3.3. Finding a (nearly) ϵ-critical point

We summarize below the convergence results of the proposed algorithms for finding a (nearly) ϵ-critical point.

Theorem 6. *Assume Algorithm 1 returns a solution $\mathbf{x}_\tau$ such that $\mathrm{E}[\|G_\gamma(\mathbf{x}_\tau)\|^2] \leq O(1/K)$ under appropriate conditions. Then if $g(\mathbf{x}) + r(\mathbf{x})$ is differentiable and has (L, ν)-Hölder continuous gradient, we have $\mathrm{E}[dist(\partial h(\mathbf{x}_\tau), \nabla(g(\mathbf{x}_\tau) + r(\mathbf{x}_\tau))] \leq O\left(\frac{1}{K^{\nu/2}} + \frac{1}{\sqrt{K}}\right)$. If $h(\mathbf{x})$ is differentiable and has (L, ν)-Hölder continuous gradient, we have $\mathrm{E}[\|\mathbf{x}_\tau - \mathbf{z}_\tau\|] \leq O(1/\sqrt{K})$ and $\mathrm{E}[dist(\nabla h(\mathbf{z}_\tau), \partial(g(\mathbf{z}_\tau) + r(\mathbf{z}_\tau))] \leq O\left(\frac{1}{K^{\nu/2}} + \frac{1}{\sqrt{K}}\right)$, where $\mathbf{z}_\tau = P_\gamma(\mathbf{x}_\tau)$.*

Remark: Note that the convergence of the proposed algorithms can be automatically adaptive to the Hölder continuous of the involved functions without requiring the value of ν for running the algorithm. Both SSDC-SPG and SSDC-AdaGrad have an iteration complexity (in the worst-case) of $O(1/\epsilon^{4/\nu})$ for finding a (nearly) ϵ-critical point. When the problem has a finite-sum structure (6) and $g(\mathbf{x})$ is smooth, SSDC-SVRG has a gradient complexity of $O(n/\epsilon^{2/\nu})$ for finding a (nearly) ϵ-critical point.

4. Non-Smooth Non-Convex Regularization

In this section, we consider a more challenging class of problem (1) where $r(\mathbf{x})$ is a proper non-smooth and non-convex lower-semicontinuous function that is not necessarily a DC function (e.g., ℓ_0 norm). Even if $r(\mathbf{x})$ is a DC function such that both components in its DC decomposition $r(\mathbf{x}) = r_1(\mathbf{x}) - r_2(\mathbf{x})$ are non-differentiable functions without Hölder continuous gradients (e.g., ℓ_{1-2} regularization, capped ℓ_1 norm), the theories presented in this section are useful to derive non-asymptotic convergence results in terms of finding an ϵ-critical point. Please note that in this case the results presented in section 3.3 are not applicable. Similarly, we assume $r(\mathbf{x})$ is simple such that its proximal mapping exists and can be efficiently computed.

The problem is challenging due to the presence of non-smooth non-convex function r. To tackle this function, we introduce the Moreau envelope of r:

$$r_\mu(\mathbf{x}) = \min_{\mathbf{y}\in\mathbb{R}^d} \frac{1}{2\mu}\|\mathbf{y} - \mathbf{x}\|^2 + r(\mathbf{y}),$$

where $\mu > 0$. A nice property of the Moreau envelope function is that it can be written as a DC function:

$$r_\mu(\mathbf{x}) = \frac{1}{2\mu}\|\mathbf{x}\|^2 - \underbrace{\max_{\mathbf{y}\in\mathbb{R}^d} \frac{1}{\mu}\mathbf{y}^\top\mathbf{x} - \frac{1}{2\mu}\|\mathbf{y}\|^2 - r(\mathbf{y})}_{R_\mu(\mathbf{x})},$$

where $R_\mu(\mathbf{x})$ is a convex function because it is the max of convex functions of $\mathbf{x}$ (Boyd & Vandenberghe, 2004). The following properties about the Moreau envelope will be useful for our analysis.

Lemma 1. *$\frac{1}{\mu}prox_{\mu r}(\mathbf{x}) \subseteq \partial R_\mu(\mathbf{x})$, and $\frac{1}{\mu}(\mathbf{x} - \mathbf{v}) \subseteq \hat{\partial} r(\mathbf{v}), \forall \mathbf{v} \in prox_{\mu r}(\mathbf{x})$.*

Given the Moreau envelope of r, the key idea is to solve the following DC problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} g(\mathbf{x}) - h(\mathbf{x}) + \frac{1}{2\mu}\|\mathbf{x}\|^2 - R_\mu(\mathbf{x}). \tag{10}$$

By carefully controlling the value of μ and combining the results presented in previous section, we are able to derive non-asymptotic convergence results for the original problem. It is worth mentioning that using the Moreau envelope of r and its DC decomposition for handling non-smooth non-convex function is first proposed in (Liu et al., 2018). However, their algorithms are deterministic and convergence results are only asymptotic. To formally state our non-asymptototic convergence results, we make the following assumptions.

Assumption 4. *Assume g and h are smooth, and one of the following conditions holds: (i) r is Lipchitz continuous; (ii) r is lower bounded and finite-valued over $\mathbb{R}^d$; (iii) $g(\mathbf{x}) - h(\mathbf{x}) + r_\mu(\mathbf{x})$ is level bounded for a small $\mu < 1$, and r is finite-valued on a compact set, and lower bounded over $\mathbb{R}^d$.*

Remark: The above assumptions on r capture many interesting non-convex non-smooth regularizers. For example, ℓ_{1-2} regularization and capped ℓ_1 norm satisfy Assumption 4 (i). The ℓ_0 norm satisfies Assumption 4 (ii). A coersive function r usually satisfies Assumption 4 (iii), e.g., ℓ_p norm $r(\mathbf{x}) = \sum_{i=1}^d |x|^p$ for $p \in (0, 1)$. In Appendix K, we further extend our results to handle a differntiable h that has only a Hölder-continuous gradient.

When employing the presented algorithms in last section to solve the problem (10), we let $r = \frac{1}{2\mu}\|\mathbf{x}\|^2$, $g \leftarrow g$ and $h \leftarrow h + R_\mu$. It is also notable that the new component $R_\mu(\mathbf{x})$ is deterministic, whose subgradient can be computed according to Lemma 1. Thus the condition in Assumption 2 (i) is sufficient for running SPG, and the smoothness condition of g is sufficient for running SVRG. Now we are ready to present our results for solving the problem (1) with a non-smooth and non-convex r.

Theorem 7. *Suppose SSDC is employed for solving (10) and returns $\mathbf{x}_\tau$. Let $\mathbf{w}_\tau = prox_{\mu r}(\mathbf{x}_\tau)$ be the final output, we have the following results to ensure $\mathrm{E}[dist(\nabla h(\mathbf{w}_\tau), \nabla g(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau))] \leq \epsilon$. (a) If Assumption 4 (i) and Assumption 2 (i) hold, then we can set $\mu = \epsilon$, use SPG for solving the subproblems, and have a total gradient complexity of $O(1/\epsilon^8)$. (b) If Assumption 4 (ii) and Assumption 2 (i) hold, then we can set $\mu = \epsilon^2$, use SPG for solving the subproblems, and have a total gradient complexity of $O(1/\epsilon^{12})$. (c) If g and h have a finite-sum form and are smooth, then we can use SVRG for solving the subproblems. Under assumption 4 (i), we can set $\mu = \epsilon$ and have a total gradient complexity of $O(n/\epsilon^4)$. Under assumption 4 (ii) or (iii), we can set $\mu = \epsilon^2$ and have a total gradient complexity of $O(n/\epsilon^6)$.*

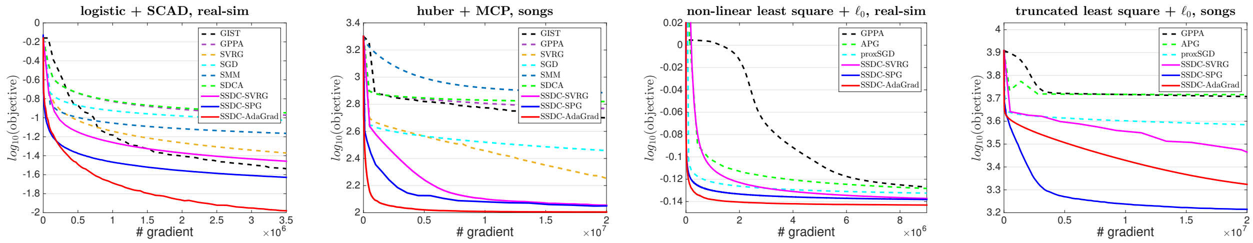

Figure 1. Learning with DC (left two) and non-DC regularizers (right two) on different datasets for classification and regression.

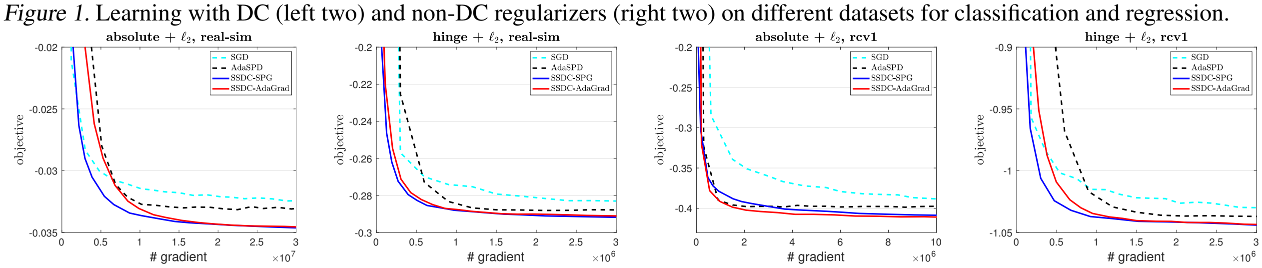

Figure 2. PU learning with different loss functions on different datasets.

5. Numerical Experiments

In this section, we perform some experiments for solving different tasks to demonstrate effectiveness of proposed algorithms by comparing with different baselines. We use very large-scale datasets from libsvm website in experiments, including real-sim (n = 72309) and rcv1 (n=20242) for classification, million songs (n = 463715) for regression. For all algorithms, the initial stepsizes are tuned in the range of $\{10^{-6:1:4}\}$, and the same initial solution with all zero entries is used. The initial iteration number T_0 of SSDC-SPG is tuned in $\{10^{1:1:4}\}$.

First, we compare SSDC algorithms with SDCA (Thi et al., 2017), SMM (Mairal, 2013), SGD (Davis & Drusvyatskiy, 2018b), SVRG (Reddi et al., 2016c), GIST (Gong et al., 2013) and GPPA (An & Nam, 2017) for learning with a DC regularizer: minimizing logistic loss with a SCAD regularizer for classification and huber loss with a MCP regularizer for regression. The parameter in Huber loss is set to be 1. The value of regularization parameter is set to be 10^{-4}. We used the form of weight in SMM following (Mairal, 2013). Since these regularizers are weakly convex, SGD with step size $\eta_0/\sqrt{t}$ is applicable (Davis & Drusvyatskiy, 2018b). We set the inner iteration number of SVRG as n following (Reddi et al., 2016c) and the same value is used as the inner iteration number T of SSDC-SVRG. We set the values of parameters in GIST with their suggested BB rule (Gong et al., 2013). Similar to (Thi et al., 2017), we tune the batch size of SDCA in a wide range and choose the one with the best performance. GIST and GPPA are deterministic algorithms that use all data points in each iteration. For fairness of comparison, we plot the objective in log scale versus the number of gradient computations in Figure 1 (left two).

Second, we consider minimizing ℓ_0 regularized non-linear least square loss function $\frac{1}{n}\sum_{i=1}^{n}(y_i - \sigma(\mathbf{w}^\top\mathbf{x}_i))^2 + \lambda\|\mathbf{w}\|_0$ with a sigmod function $\sigma(s) = \frac{1}{1+e^{-s}}$ for classification and ℓ_0 regularized truncated least square loss function $\frac{1}{2n}\sum_{i=1}^{n}\alpha\log(1 + (y_i - \mathbf{w}^\top\mathbf{x}_i)^2/\alpha) + \lambda\|\mathbf{w}\|_0$ (Xu et al., 2018) for regression. We compare the proposed algorithms with GPPA, APG (Li & Lin, 2015) and proximal version of SGD (proxSGD), where GPPA and APG are deterministic algorithms. We fix the truncation value as $\alpha = \sqrt{10n}$. The loss function in these two tasks are smooth and non-convex. The value of regularization parameter is fixed as 10^{-6}. For APG, we implement both monotone and non-monotone versions following (Li & Lin, 2015), and then the better one is reported. Although the convergence guarantee of proxSGD remains unclear for the considered problems, we still include it for comparison. The results on two data sets are plotted in Figure 1 (right two).

The results of these two experiments indicate that the proposed stochastic algorithms outperform all deterministic baselines (GITS, GPPA, APG) on all tasks, which verify the necessity of using stochastic algorithms on large datasets. In addition, our algorithms especially SSDC-AdaGrad and SSDC-SPG also converge faster than stochastic algorithms SGD, SDCA, and non-convex SVRG verifying that our stochastic algorithms are more practical for the considered problems. We also see that in most cases SSDC-AdaGrad is more effective than SSDC-SPG and SSDC-SVRG.

Finally, we compare SSDC algorithms with two baselines AdaSPD (Nitanda & Suzuki, 2017) and SGD (Davis et al., 2018) for solving two ℓ_2 regularized positive-unlabeled (PU) learning problems (Du Plessis et al., 2015) with non-smooth losses, i.e., hinge loss and absolute loss. The ℓ_2 regularization parameter is set to be 10^{-4}. For SGD, we use the standard stepsize $\eta = \eta_0/\sqrt{t}$ (Ghadimi & Lan, 2013) with η_0 tuned. The mini-batch size and the number of iterations of each stage of AdaSPD are simply set as 10^4. The results on two classification datasets are plotted in Figure 2, which show that SSDC-SPG and SSDC-AdaGrad outperforms SGD and AdaSPD.

Acknowledgements

The authors thank the anonymous reviewers for their helpful comments. Y. Xu and T. Yang are partially supported by National Science Foundation (IIS-1545995).

References

Allen-Zhu, Z. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. In *International Conference on Machine Learning*, pp. 89–97, 2017.

Allen-Zhu, Z. and Hazan, E. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pp. 699–707, 2016.

An, N. T. and Nam, N. M. Convergence analysis of a proximal point algorithm for minimizing differences of functions. *Optimization*, 66(1):129–147, 2017.

Attouch, H., Bolte, J., and Svaiter, B. F. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1):91–129, Feb 2013.

Bolte, J., Sabach, S., and Teboulle, M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2), August 2014.

Bot, R. I., Csetnek, E. R., and László, S. C. An inertial forward–backward algorithm for the minimization of the sum of two nonconvex functions. *EURO Journal on Computational Optimization*, 4(1):3–25, Feb 2016.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

Candès, E. J., Wakin, M. B., and Boyd, S. P. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, Dec 2008.

Chen, Z. and Yang, T. A variance reduction method for non-convex optimization with improved convergence under large condition number. *CoRR*, abs/1809.06754, 2018.

Chen, Z., Xu, Y., Chen, E., and Yang, T. Sadagrad: Strongly adaptive stochastic gradient methods. In *International Conference on Machine Learning*, pp. 912–920, 2018a.

Chen, Z., Yang, T., Yi, J., Zhou, B., and Chen, E. Universal stagewise learning for non-convex problems with convergence on averaged solutions. *CoRR*, /abs/1808.06296, 2018b.

Davis, D. and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *CoRR*, abs/1803.06523, 2018a.

Davis, D. and Drusvyatskiy, D. Stochastic subgradient method converges at the rate $\mathrm{O}(k^{-1/4})$ on weakly convex functions. *CoRR*, /abs/1802.02988, 2018b.

Davis, D. and Grimmer, B. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *arXiv preprint arXiv:1707.03505*, 2017.

Davis, D., Drusvyatskiy, D., Kakade, S., and Lee, J. D. Stochastic subgradient method converges on tame functions. *CoRR*, abs/1804.07795, 2018.

Drusvyatskiy, D. and Paquette, C. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, Jul 2018.

Du Plessis, M., Niu, G., and Sugiyama, M. Convex formulation for learning from positive and unlabeled data. In *International Conference on Machine Learning*, pp. 1386–1394, 2015.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Duchi, J. C., Shalev-Shwartz, S., Singer, Y., and Tewari, A. Composite objective mirror descent. In *Conference on Learning Theory*, pp. 14–26, 2010.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Ghadimi, S. and Lan, G. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Gong, P., Zhang, C., Lu, Z., Huang, J., and Ye, J. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *International Conference on Machine Learning*, pp. 37–45, 2013.

Hiriart-Urruty, J.-B. Generalized differentiability / duality and optimization for problems dealing with differences of convex functions. In Ponstein, J. (ed.), *Convexity and Duality in Optimization*, pp. 37–70, Berlin, Heidelberg, 1985. Springer Berlin Heidelberg.

Horst, R. and Thoai, N. V. Dc programming: Overview. *Journal of Optimization Theory and Applications*, 103(1):1–43, Oct 1999.

Khamaru, K. and Wainwright, M. Convergence guarantees for a class of non-convex and non-smooth optimization problems. In *International Conference on Machine Learning*, pp. 2601–2610, 2018.

Kiryo, R., Niu, G., du Plessis, M. C., and Sugiyama, M. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems 30*, pp. 1675–1685, 2017.

Lan, G. and Yang, Y. Accelerated stochastic algorithms for nonconvex finite-sum and multi-block optimization. *CoRR*, abs/1805.05411, 2018.

Li, G. and Pong, T. K. Douglas-rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. *Mathematical Programming*, 159(1-2): 371–401, 2016.

Li, H. and Lin, Z. Accelerated proximal gradient methods for nonconvex programming. In *Neural Information Processing Systems*, pp. 379–387, 2015.

Liu, T., Pong, T. K., and Takeda, A. A successive difference-of-convex approximation method for a class of nonconvex nonsmooth optimization problems. *Mathematical Programming*, Sep 2018.

Mairal, J. Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems*, pp. 2283–2291, 2013.

Nitanda, A. and Suzuki, T. Stochastic Difference of Convex Algorithm and its Application to Training Deep Boltzmann Machines. In *International Conference on Artificial Intelligence and Statistics*, pp. 470–478, 2017.

Reddi, S. J., Hefny, A., Sra, S., Póczós, B., and Smola, A. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*, pp. 314–323, 2016a.

Reddi, S. J., Sra, S., Póczos, B., and Smola, A. J. Fast incremental method for smooth nonconvex optimization. In *55th IEEE Conference on Decision and Control*, pp. 1971–1977, 2016b.

Reddi, S. J., Sra, S., Póczos, B., and Smola, A. J. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems*, pp. 1145–1153, 2016c.

Rockafellar, R. and Wets, R. J.-B. *Variational Analysis*. Springer Verlag, Heidelberg, Berlin, New York, 1998.

Thi, H. A. L. and Dinh, T. P. Dc programming and dca: thirty years of developments. *Mathematical Programming*, 169 (1):5–68, May 2018.

Thi, H. A. L., Le, H. M., Phan, D. N., and Tran, B. Stochastic dca for the large-sum of non-convex functions problem and its application to group variable selection in classification. In *International Conference on Machine Learning*, pp. 3394–3403, 2017.

Wen, B., Chen, X., and Pong, T. K. A proximal difference-of-convex algorithm with extrapolation. *Computational Optimization and Applications*, 69(2):297–324, Mar 2018.

Xiao, L. and Zhang, T. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Xu, Y., Zhu, S., Yang, S., Zhang, C., Jin, R., and Yang, T. Learning with non-convex truncated losses by SGD. *CoRR*, abs/1805.07880, 2018.

Yang, L. Proximal gradient method with extrapolation and line search for a class of nonconvex and nonsmooth problems. *CoRR*, abs/1711.06831, 2018.

Yin, P., Lou, Y., He, Q., and Xin, J. Minimization of 1-2 for compressed sensing. *SIAM J. Scientific Computing*, 37, 2015.

Yu, Y., Zheng, X., Marchetti-Bowick, M., and Xing, E. P. Minimizing nonconvex non-separable functions. In *International Conference on Artificial Intelligence and Statistics*, 2015.

Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38: 894 – 942, 2010a.

Zhang, S. and He, N. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1806.04781*, 2018.

Zhang, S. and Xin, J. Minimization of transformed l_1 penalty: theory, difference of convex function algorithm, and robust application in compressed sensing. *Mathematical Programming*, 169(1):307–336, 2018.

Zhang, T. Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.*, 11:1081–1107, March 2010b. ISSN 1532-4435.

Zhao, P. and Zhang, T. Stochastic optimization with importance sampling for regularized loss minimization. In *International Conference on Machine Learning*, pp. 1–9, 2015.

Zhong, W. and Kwok, J. T. Gradient descent with proximal average for nonconvex and composite regularization. In *AAAI Conference on Artificial Intelligence*, pp. 2206–2212, 2014.

A. Proof of Proposition 1

According to the first-order optimality condition, we have

$$0 \in \partial(g(P_\lambda(\mathbf{z})) + r(P_\lambda(\mathbf{z}))) - \partial h(\mathbf{z}) + \gamma(P_\gamma(\mathbf{z}) - \mathbf{z}).$$

Since $\mathbf{z} = P_\gamma(\mathbf{z})$, we have

$$0 \in \partial(g + r)(\mathbf{z}) - \partial h(\mathbf{z}),$$

which implies that $\mathbf{z}$ is a critical point of the original minimization problem.

B. Proof of Theorem 2 and Theorem 1

The proof of Theorem 1 can be obtained by a slight change of the following proof. Define the following notations.

$$\mathbf{z}_k = P_\gamma(\mathbf{x}_k) = \arg\min_{\mathbf{x}\in\mathbb{R}^d} F_k(\mathbf{x}) := \underbrace{g(\mathbf{x}) + r(\mathbf{x}) - \partial h(\mathbf{x}_k)^\top(\mathbf{x} - \mathbf{x}_k)}_{f_k(\mathbf{x})} + \frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}_k\|^2.$$

By the assumption of (8), we have $\mathrm{E}[F_k(\mathbf{x}_{k+1}) - F_k(\mathbf{z}_k)] \leq \epsilon_k = c/k$. By the strong convexity of F_k, we have $F_k(\mathbf{x}_k) \geq F_k(\mathbf{z}_k) + \frac{\gamma}{2}\|\mathbf{x}_k - \mathbf{z}_k\|^2$. Thus we have

$$\begin{aligned}\mathrm{E}[f_k(\mathbf{x}_{k+1}) + \frac{\gamma}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2] &\leq F_k(\mathbf{x}_k) - \frac{\gamma}{2}\|\mathbf{x}_k - \mathbf{z}_k\|^2 + \epsilon_k \\ &= g(\mathbf{x}_k) + r(\mathbf{x}_k) - \frac{\gamma}{2}\|\mathbf{x}_k - \mathbf{z}_k\|^2 + \epsilon_k. \end{aligned} \tag{11}$$

Rearranging the terms, we have

$$\begin{aligned}\mathrm{E}\left[\frac{\gamma}{2}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] &\leq \mathrm{E}[g(\mathbf{x}_k) + r(\mathbf{x}_k) - f_k(\mathbf{x}_{k+1})] + \epsilon_k \\ &\leq \mathrm{E}[g(\mathbf{x}_k) + r(\mathbf{x}_k) - g(\mathbf{x}_{k+1}) - r(\mathbf{x}_{k+1}) + \partial h(\mathbf{x}_k)^\top(\mathbf{x}_{k+1} - \mathbf{x}_k)] + \epsilon_k \\ &\leq \mathrm{E}[g(\mathbf{x}_k) + r(\mathbf{x}_k) - g(\mathbf{x}_{k+1}) - r(\mathbf{x}_{k+1}) + h(\mathbf{x}_{k+1}) - h(\mathbf{x}_k)] + \epsilon_k \\ &= \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + \epsilon_k,\end{aligned}$$

where the last inequality follows the convexity of $h(\cdot)$. Multiplying both sides by $w_k = k^\alpha$ and taking summation over $k = 1, \ldots, K$, we have

$$\mathrm{E}\left[\frac{\gamma}{2}\sum_{k=1}^K w_k\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq \mathrm{E}\left[\sum_{k=1}^K w_k(F(\mathbf{x}_k) - F(\mathbf{x}_{k+1}))\right] + \sum_{k=1}^K w_k\epsilon_k, \tag{12}$$

The second term in the R.H.S of the above inequality can be easily bounded using simple calculus. For the first term, we use similar analysis as that in the proof of Theorem 1 in (Chen et al., 2018b):

$$\begin{aligned}&\sum_{k=1}^K w_k(F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})) = \sum_{k=1}^K (w_{k-1}F(\mathbf{x}_k) - w_kF(\mathbf{x}_{k+1})) + \sum_{k=1}^K (w_k - w_{k-1})F(\mathbf{x}_k) \\ &= w_0F(\mathbf{x}_1) - w_KF(\mathbf{x}_{K+1}) + \sum_{k=1}^K (w_k - w_{k-1})F(\mathbf{x}_k) \\ &= \sum_{k=1}^K (w_s - w_{s-1})(F(\mathbf{x}_k) - F(\mathbf{x}_{K+1})) \leq \sum_{k=1}^K (w_k - w_{k-1})(F(\mathbf{x}_k) - \min_{\mathbf{x}} F(\mathbf{x})),\end{aligned}$$

where we use $\mathbf{w}_0 = 0$. Taking expectation on both sides, we have

$$\mathrm{E}\left[\sum_{k=1}^K w_k(F(\mathbf{x}_k) - F(\mathbf{x}_{k+1}))\right] \leq \sum_{k=1}^K (w_k - w_{k-1})\mathrm{E}[(F(\mathbf{x}_k) - \min_{\mathbf{x}} F(\mathbf{x}))] \leq \Delta w_K$$

Then, we have

$$\mathrm{E}\left[\frac{\gamma}{2}\|\mathbf{z}_\tau - \mathbf{x}_\tau\|^2\right] \leq \frac{\Delta(\alpha+1)}{K} + \frac{c(\alpha+1)}{K},$$

which can complete the proof by multiplying both sides by 2γ. The result in Theorem 1 for the uniform sampling can be easily derived from the equality (16) by using the fact $\sum_{k=1}^K 1/k \leq (1 + \log K)$.

Algorithm 3 SPG($F^{\gamma}_{\mathbf{x}_1}, \mathbf{x}_1, T$)

1: Set step size η_t according to Proposition 2, $\Omega = \{\mathbf{x} \in \text{dom}(r) : \|\mathbf{x} - \mathbf{x}_1\| \leq 3G/\gamma\}$
2: **for** $t = 1, \ldots, T$ **do**
3: Compute $\partial g(\mathbf{x}_t; \xi_t)$ and $\partial h(\mathbf{x}_1; \varsigma_t)$
4: Option 1: $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}} \{\mathbf{x}^\top(\partial g(\mathbf{x}_t; \xi_t) - \partial h(\mathbf{x}_1; \varsigma_t)) + r(\mathbf{x}) + \frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}_1\|^2 + \frac{1}{2\eta_t}\|\mathbf{x} - \mathbf{x}_t\|^2\}$
5: Option 2: $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\Omega} \{\mathbf{x}^\top(\partial g(\mathbf{x}_t; \xi_t) - \partial h(\mathbf{x}_1; \varsigma_t)) + r(\mathbf{x}) + \frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}_1\|^2 + \frac{1}{2\eta_t}\|\mathbf{x} - \mathbf{x}_t\|^2\}$
6: **end for**
7: **Output**: $\widehat{\mathbf{x}}_T = \sum_{t=2}^{T+1} t\mathbf{x}_t / \sum_{t=2}^{T+1} t$ (Option 1) or $\widehat{\mathbf{x}}_T = \sum_{t=1}^{T} t\mathbf{x}_t / \sum_{t=1}^{T} t$ (Option 2)

C. SPG Method

The detailed steps of SPG method are presented in Algorithm 3, with option 1 corresponding to smooth g and option 2 corresponding to non-smooth g.

The following proposition summarizes the convergence of SPG for solving each subproblem, whose proof is presented later.

Proposition 2. *Suppose Assumption 2(i) hold, then by setting $\eta_t = 1/(L(t+1))$ and $\gamma \geq 3L$, Algorithm 3 with Option 1 guarantees that*

$$\mathrm{E}[F^{\gamma}_{\mathbf{x}_1}(\widehat{\mathbf{x}}_T) - F^{\gamma}_{\mathbf{x}_1}(\mathbf{x}_*)] \leq \frac{4L\|\mathbf{x}_* - \mathbf{x}_1\|^2}{T(T+3)} + \frac{2G^2}{(T+3)L}.$$

Suppose Assumption 2(ii) hold, then by setting $\eta_t = 4/(\gamma t)$, Algorithm 3 with Option 2 guarantees that

$$\mathrm{E}\left[F^{\gamma}_{\mathbf{x}_1}(\widehat{\mathbf{x}}_T) - F^{\gamma}_{\mathbf{x}_1}(\mathbf{x}_*)\right] \leq \frac{\gamma\|\mathbf{x}_* - \mathbf{x}_1\|^2}{4T(T+1)} + \frac{28G^2}{\gamma(T+1)},$$

where $\mathbf{x}_* = \arg\min_{\mathbf{x}} F^{\gamma}_{\mathbf{x}_1}(\mathbf{x})$.

C.1. Proof of Proposition 2

Option 1. Let us first prove the case of smooth g. Let $f(\mathbf{x}) = g(\mathbf{x}) - \partial h(\mathbf{x}_1)^\top(\mathbf{x} - \mathbf{x}_1)$ and $\hat{r}(\mathbf{x}) = r(\mathbf{x}) + \frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}_1\|^2$. Then $f(\mathbf{x})$ is L-smooth and $\hat{r}(\mathbf{x})$ is γ-strongly convex. A stochastic gradient of $f(\mathbf{x})$ is given by $\partial f(\mathbf{x}; \xi, \varsigma) = \nabla g(\mathbf{x}; \xi) - \partial h(\mathbf{x}_1, \varsigma)$, which has a variance of G^2 according to the assumption. Let $\eta_t = 1/(L(t+1)) \leq 1/L$. In our proof, we first need the following lemma, which is attributed to (Zhao & Zhang, 2015).

Lemma 2. *Under the same assumptions in Proposition 2, we have*

$$\mathrm{E}[f(\mathbf{x}_{t+1}) + \hat{r}(\mathbf{x}_{t+1}) - f(\mathbf{x}) - \hat{r}(\mathbf{x})] \leq \frac{\|\mathbf{x}_t - \mathbf{x}\|^2}{2\eta_t} - \frac{\|\mathbf{x} - \mathbf{x}_{t+1}\|^2}{2\eta_t} - \frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}_{t+1}\|^2 + \eta_t G^2.$$

The proof of this lemma is similar to the analysis to proof of Lemma 1 in (Zhao & Zhang, 2015). For completeness, we will include its proof later in this section.

Let's continue the proof by following Lemma 2 and letting $w_t = t$, then

$$\begin{aligned}
&\sum_{t=1}^{T} w_{t+1}(f(\mathbf{x}_{t+1}) + \hat{r}(\mathbf{x}_{t+1}) - f(\mathbf{x}) - \hat{r}(\mathbf{x})) \\
&\leq \sum_{t=1}^{T}\left(\frac{w_{t+1}}{2\eta_t}\|\mathbf{x} - \mathbf{x}_t\|^2 - \frac{w_{t+1}}{2\eta_t}\|\mathbf{x} - \mathbf{x}_{t+1}\|^2 - \frac{\gamma w_{t+1}}{2}\|\mathbf{x} - \mathbf{x}_{t+1}\|^2\right) + \sum_{t=1}^{T}\eta_t w_{t+1}G^2 \\
&\leq \sum_{t=1}^{T}\left(\frac{w_{t+1}}{2\eta_t} - \frac{w_t}{2\eta_{t-1}} - \frac{\gamma w_t}{2}\right)\|\mathbf{x} - \mathbf{x}_t\|^2 + \frac{w_1/\eta_0 + \gamma w_1}{2}\|\mathbf{x} - \mathbf{x}_1\|^2 + \sum_{t=1}^{T}\eta_t w_{t+1}G^2 \\
&\leq \frac{L+\gamma}{2}\|\mathbf{x} - \mathbf{x}_1\|^2 + \sum_{t=1}^{T} G^2/L
\end{aligned}$$

where the last inequality uses the fact $\frac{w_{t+1}}{\eta_t} - \frac{w_t}{\eta_{t-1}} - \gamma w_t = L(t+1)^2 - Lt^2 - \gamma t \leq 0$ due to $\gamma \geq 3L$. Then we have

$$f(\widehat{\mathbf{x}}_T) + \hat{r}(\widehat{\mathbf{x}}_T) - f(\mathbf{x}) - \hat{r}(\mathbf{x}) \leq \frac{(L+\gamma)\|\mathbf{x} - \mathbf{x}_1\|^2}{T(T+3)} + \frac{2G^2}{(T+3)L}.$$

Option 2. Next, we prove the case when g is non-smooth. First, we need to show that $\mathbf{x}_* = \arg\min F^{\gamma}_{\mathbf{x}_1}(\mathbf{x})$ in the set $\|\mathbf{x}-\mathbf{x}_1\| \leq 3G/\gamma$. By the optimality condition of $\mathbf{x}_*$ we have

$$(\partial g(\mathbf{x}_*) + \partial r(\mathbf{x}_*) - \partial h(\mathbf{x}_1) + \gamma(\mathbf{x}_* - \mathbf{x}_1))^\top(\mathbf{x} - \mathbf{x}_*) \geq 0, \forall \mathbf{x} \in \mathrm{dom}(r)$$

Plugging $\mathbf{x} = \mathbf{x}_1$ into the above inequality, we have

$$\gamma\|\mathbf{x}_1 - \mathbf{x}_*\|^2 \leq 3G\|\mathbf{x}_1 - \mathbf{x}_*\| \Rightarrow \|\mathbf{x}_1 - \mathbf{x}_*\| \leq 3G/\gamma,$$

where the first inequality uses Assumption 2 (ii). Let us recall the update

$$\mathbf{x}_{t+1} = \arg\min_{\|\mathbf{x}-\mathbf{x}_1\|\leq 3G/\gamma} \mathbf{x}^\top \partial f(\mathbf{x}_t; \xi_t, \varsigma_t) + \hat{r}(\mathbf{x}) + \frac{1}{2\eta_t}\|\mathbf{x} - \mathbf{x}_t\|^2$$

By the optimality condition of $\mathbf{x}_{t+1}$ and the strong convexity of above problem, we have for any $\mathbf{x} \in \Omega = \{\mathbf{x} \in \mathrm{dom}(r); \|\mathbf{x} - \mathbf{x}_1\| \leq 3G/\gamma\}$

$$\begin{aligned}
&\mathbf{x}^\top \partial f(\mathbf{x}_t; \xi_t, \varsigma_t) + \hat{r}(\mathbf{x}) + \frac{1}{2\eta_t}\|\mathbf{x} - \mathbf{x}_t\|^2 \\
&\geq \mathbf{x}_{t+1}^\top \partial f(\mathbf{x}_t; \xi_t, \varsigma_t) + \hat{r}(\mathbf{x}_{t+1}) + \frac{1}{2\eta_t}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \frac{1/\eta_t + \gamma}{2}\|\mathbf{x} - \mathbf{x}_{t+1}\|^2
\end{aligned}$$

Rearranging the terms, we have

$$\begin{aligned}
&(\mathbf{x}_t - \mathbf{x})^\top \partial f(\mathbf{x}_t; \xi_t, \varsigma_t) + \hat{r}(\mathbf{x}_{t+1}) - \hat{r}(\mathbf{x}) \\
&\leq (\mathbf{x}_t - \mathbf{x}_{t+1})^\top \partial f(\mathbf{x}_t; \xi_t, \varsigma_t) - \frac{1}{2\eta_t}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \frac{1}{2\eta_t}\|\mathbf{x} - \mathbf{x}_t\|^2 - \frac{1/\eta_t + \gamma}{2}\|\mathbf{x} - \mathbf{x}_{t+1}\|^2 \\
&\leq \frac{\eta_t\|\partial f(\mathbf{x}_t; \xi_t, \varsigma_t)\|^2}{2} + \frac{1}{2\eta_t}\|\mathbf{x} - \mathbf{x}_t\|^2 - \frac{1/\eta_t + \gamma}{2}\|\mathbf{x} - \mathbf{x}_{t+1}\|^2
\end{aligned}$$

Taking expectation on both sides, we have

$$\begin{aligned}
&\mathrm{E}[f(\mathbf{x}_t) - f(\mathbf{x}) + \hat{r}(\mathbf{x}_{t+1}) - \hat{r}(\mathbf{x})] \\
&\leq \frac{\eta_t \mathrm{E}[\|\partial f(\mathbf{x}_t; \xi_t, \varsigma_t)\|^2]}{2} + \mathrm{E}\left[\frac{1}{2\eta_t}\|\mathbf{x} - \mathbf{x}_t\|^2 - \frac{1/\eta_t + \gamma}{2}\|\mathbf{x} - \mathbf{x}_{t+1}\|^2\right].
\end{aligned}$$

Multiplying both sides w_t and taking summation over $t = 1, \ldots, T$ and expectation, we have

$$\begin{aligned}
&\mathrm{E}\left[\sum_{t=1}^T w_t(f(\mathbf{x}_t) - f(\mathbf{x}) + \hat{r}(\mathbf{x}_{t+1}) - \hat{r}(\mathbf{x}))\right] \\
&\leq \sum_{t=1}^T 2G^2 w_t\eta_t + \mathrm{E}\left[\sum_{t=1}^T \frac{w_t}{2\eta_t}\|\mathbf{x} - \mathbf{x}_t\|^2 - \frac{w_t/\eta_t + w_t\gamma}{2}\|\mathbf{x} - \mathbf{x}_{t+1}\|^2\right].
\end{aligned}$$

Thus,

$$
\begin{aligned}
&\mathrm{E}\Big[\sum_{t=1}^T w_t(f(\mathbf{x}_t) - f(\mathbf{x}) + \hat{r}(\mathbf{x}_t) - \hat{r}(\mathbf{x}))\Big] \\
&\leq \mathrm{E}\Big[\sum_{t=1}^T w_t(r(\mathbf{x}_t) - r(\mathbf{x}_{t+1}))\Big] + \sum_{t=1}^T 2G^2 w_t\eta_t + \mathrm{E}\Big[\sum_{t=1}^T \left(\frac{w_t}{2\eta_t} - \frac{w_{t-1}/\eta_{t-1} + w_{t-1}\gamma}{2}\right)\|\mathbf{x}-\mathbf{x}_t\|^2\Big] \\
&\leq w_0 r(\mathbf{x}_1) - w_T r(\mathbf{x}_{T+1}) + \mathrm{E}\Big[\sum_{t=1}^T (w_t - w_{t-1}) r(\mathbf{x}_t)\Big] + \sum_{t=1}^T 2G^2 w_t\eta_t \\
&+ \mathrm{E}\Big[\sum_{t=1}^T \left(\frac{w_t}{2\eta_t} - \frac{w_{t-1}/\eta_{t-1} + w_{t-1}\gamma}{2}\right)\|\mathbf{x}-\mathbf{x}_t\|^2\Big] \\
&\leq \mathrm{E}\Big[\sum_{t=1}^T (w_t - w_{t-1})(r(\mathbf{x}_t) - r(\mathbf{x}_{T+1}))\Big] + \sum_{t=1}^T 2G^2 w_t\eta_t + \mathrm{E}\Big[\sum_{t=1}^T \left(\frac{w_t}{2\eta_t} - \frac{w_{t-1}/\eta_{t-1} + w_{t-1}\gamma}{2}\right)\|\mathbf{x}-\mathbf{x}_t\|^2\Big] \\
&\leq \mathrm{E}\Big[\sum_{t=1}^T (w_t - w_{t-1}) G\|\mathbf{x}_t - \mathbf{x}_{T+1}\|\Big] + \sum_{t=1}^T 2G^2 w_t\eta_t + \mathrm{E}\Big[\sum_{t=1}^T \left(\frac{w_t}{2\eta_t} - \frac{w_{t-1}/\eta_{t-1} + w_{t-1}\gamma}{2}\right)\|\mathbf{x}-\mathbf{x}_t\|^2\Big] \\
&\leq \mathrm{E}\Big[\sum_{t=1}^T (w_t - w_{t-1}) 6G^2/\gamma\Big] + \sum_{t=1}^T 2G^2 w_t\eta_t + \mathrm{E}\Big[\sum_{t=1}^T \left(\frac{w_t}{2\eta_t} - \frac{w_{t-1}/\eta_{t-1} + w_{t-1}\gamma}{2}\right)\|\mathbf{x}-\mathbf{x}_t\|^2\Big]
\end{aligned}
$$

Plugging the value $w_t = t, \eta_t = 4/(\gamma t)$, and using the fact $\frac{w_t}{2\eta_t} - \frac{w_{t-1}/\eta_{t-1} + w_{t-1}\gamma}{2} \leq 0, \forall t \geq 2$, we have

$$
\mathrm{E}\Big[\sum_{t=1}^T w_t(f(\mathbf{x}_t) - f(\mathbf{x}) + \hat{r}(\mathbf{x}_t) - \hat{r}(\mathbf{x}))\Big] \leq \frac{w_T 6G^2}{\gamma} + \frac{8G^2 T}{\gamma} + \frac{\gamma\|\mathbf{x}-\mathbf{x}_1\|^2}{8}
$$

Thus,

$$
\mathrm{E}\Big[f(\widehat{\mathbf{x}}_T) - f(\mathbf{x}) + \hat{r}(\widehat{\mathbf{x}}_T) - \hat{r}(\mathbf{x}))\Big] \leq \frac{28G^2}{\gamma(T+1)} + \frac{\gamma\|\mathbf{x}-\mathbf{x}_1\|^2}{4T(T+1)}.
$$

C.2. Proof of Lemma 2

First, we need the following lemma, which is attributed to (Zhao & Zhang, 2015).

Lemma 3. *If $\hat{r}(\mathbf{x})$ is convex and*

$$
\widehat{\mathbf{u}} = \arg\min_{\mathbf{x}} \mathbf{x}^\top \mathbf{g}_u + \hat{r}(\mathbf{x}) + \frac{1}{2\eta}\|\mathbf{x}-\mathbf{z}\|^2,\ \widehat{\mathbf{v}} = \arg\min_{\mathbf{x}} \mathbf{x}^\top \mathbf{g}_v + \hat{r}(\mathbf{x}) + \frac{1}{2\eta}\|\mathbf{x}-\mathbf{z}\|^2,
$$

then we have

$$
\|\widehat{\mathbf{u}} - \widehat{\mathbf{v}}\| \leq \eta\|\mathbf{g}_u - \mathbf{g}_v\|.
$$

Proof. **of Lemma 3.** The proof can be found in the analysis of Lemma 1 in (Zhao & Zhang, 2015), but we still include it here for completeness. By the optimilaty of $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{v}}$ we have

$$
\begin{aligned}
\mathbf{a} :=& \frac{\mathbf{z} - \widehat{\mathbf{u}}}{\eta} - \mathbf{g}_u \in \partial\hat{r}(\widehat{\mathbf{u}}) \\
\mathbf{b} :=& \frac{\mathbf{z} - \widehat{\mathbf{v}}}{\eta} - \mathbf{g}_v \in \partial\hat{r}(\widehat{\mathbf{v}}).
\end{aligned}
$$

Since $\hat{r}(\mathbf{x})$ is convex, then

$$
0 \leq \langle \mathbf{a} - \mathbf{b}, \widehat{\mathbf{u}} - \widehat{\mathbf{v}}\rangle = \frac{1}{\eta}\langle \eta\mathbf{g}_v - \eta\mathbf{g}_u + \widehat{\mathbf{v}} - \widehat{\mathbf{u}}, \widehat{\mathbf{u}} - \widehat{\mathbf{v}}\rangle,
$$

which implies

$$
1/\eta\|\widehat{\mathbf{u}} - \widehat{\mathbf{v}}\|^2 \leq \langle \mathbf{g}_v - \mathbf{g}_u, \widehat{\mathbf{u}} - \widehat{\mathbf{v}}\rangle \leq \|\mathbf{g}_v - \mathbf{g}_u\|\|\widehat{\mathbf{u}} - \widehat{\mathbf{v}}\|.
$$

Thus

$$\|\widehat{\mathbf{u}} - \widehat{\mathbf{v}}\| \leq \eta \|\mathbf{g}_v - \mathbf{g}_u\|.$$

□

Then, let's start the proof of Lemma 2. Since $f(\mathbf{x}) = g(\mathbf{x}) - \partial h(\mathbf{x}_1)^\top(\mathbf{x} - \mathbf{x}_1)$ and $\hat{r}(\mathbf{x}) = r(\mathbf{x}) + \frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}_1\|^2$, then $f(\mathbf{x})$ is L-smooth and $\hat{r}(\mathbf{x})$ is γ-strongly convex. Recall that a stochastic gradient of $f(\mathbf{x})$ is given by $\partial f(\mathbf{x}; \xi, \varsigma) = \nabla g(\mathbf{x}; \xi) - \partial h(\mathbf{x}_1, \varsigma)$, which has a variance of G^2 according to the assumption, and $\eta_t = 1/(L(t+1)) \leq 1/L$. By the convexity of $f(\mathbf{x})$ and strong convexity of $\hat{r}(\mathbf{x})$ we have

$$f(\mathbf{x}) + \hat{r}(\mathbf{x}) \geq f(\mathbf{x}_t) + \langle \partial f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle + \hat{r}(\mathbf{x}_{t+1}) + \langle \partial \hat{r}(\mathbf{x}_{t+1}), \mathbf{x} - \mathbf{x}_{t+1} \rangle + \frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}_{t+1}\|^2.$$

By the smoothness of $f(\mathbf{x})$, we also have

$$f(\mathbf{x}_t) \geq f(\mathbf{x}_{t+1}) - \langle \partial f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle - \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2.$$

Combing the above two inequalities, we have

$$\begin{aligned}
& f(\mathbf{x}_{t+1}) + \hat{r}(\mathbf{x}_{t+1}) - f(\mathbf{x}) - \hat{r}(\mathbf{x}) \\
\leq & \langle \partial f(\mathbf{x}_t) + \partial \hat{r}(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{x} \rangle - \frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}_{t+1}\|^2 + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\
= & \langle \partial f(\mathbf{x}_t) - \partial f(\mathbf{x}_t; \xi_t, \varsigma_t), \mathbf{x}_{t+1} - \mathbf{x} \rangle + \frac{1}{\eta_t}\langle \mathbf{x}_t - \mathbf{x}_{t+1}, \mathbf{x}_{t+1} - \mathbf{x} \rangle \\
& - \frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}_{t+1}\|^2 + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\
= & \langle \partial f(\mathbf{x}_t) - \partial f(\mathbf{x}_t; \xi_t, \varsigma_t), \mathbf{x}_{t+1} - \mathbf{x} \rangle + \frac{\|\mathbf{x}_t - \mathbf{x}\|^2}{2\eta_t} - \frac{\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2}{2\eta_t} - \frac{\|\mathbf{x}_{t+1} - \mathbf{x}\|^2}{2\eta_t} \\
& - \frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}_{t+1}\|^2 + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2, \qquad (13)
\end{aligned}$$

where the last second equality is due to the update and optimilaty of $\mathbf{x}_{t+1}$ (Option 1 in Algorithm 3); the last equality uses the fact that $2\langle \mathbf{x} - \mathbf{y}, \mathbf{y} - \mathbf{z} \rangle = \|\mathbf{x} - \mathbf{z}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2 - \|\mathbf{y} - \mathbf{z}\|^2$. To deal with the term $\langle \partial f(\mathbf{x}_t) - \partial f(\mathbf{x}_t; \xi_t, \varsigma_t), \mathbf{x}_{t+1} - \mathbf{x} \rangle$, we define $\widehat{\mathbf{x}}_{t+1} = \arg\min_{\mathbf{x}} \mathbf{x}^\top \partial f(\mathbf{x}_t) + \hat{r}(\mathbf{x}) + \frac{1}{2\eta_t}\|\mathbf{x} - \mathbf{x}_t\|^2$, which is independent of $\partial f(\mathbf{x}_t; \xi_t, \varsigma_t)$. Taking expectation over ξ_t and ς_t over this term we get

$$\begin{aligned}
& \mathrm{E}[\langle \partial f(\mathbf{x}_t) - \partial f(\mathbf{x}_t; \xi_t, \varsigma_t), \mathbf{x}_{t+1} - \mathbf{x} \rangle] \\
= & \mathrm{E}[\langle \partial f(\mathbf{x}_t) - \partial f(\mathbf{x}_t; \xi_t, \varsigma_t), \mathbf{x}_{t+1} - \widehat{\mathbf{x}}_{t+1} + \widehat{\mathbf{x}}_{t+1} - \mathbf{x} \rangle] \\
= & \mathrm{E}[\langle \partial f(\mathbf{x}_t) - \partial f(\mathbf{x}_t; \xi_t, \varsigma_t), \mathbf{x}_{t+1} - \widehat{\mathbf{x}}_{t+1} \rangle] + \mathrm{E}[\langle \partial f(\mathbf{x}_t) - \partial f(\mathbf{x}_t; \xi_t, \varsigma_t), \widehat{\mathbf{x}}_{t+1} - \mathbf{x} \rangle] \\
= & \mathrm{E}[\langle \partial f(\mathbf{x}_t) - \partial f(\mathbf{x}_t; \xi_t, \varsigma_t), \mathbf{x}_{t+1} - \widehat{\mathbf{x}}_{t+1} \rangle] \\
\leq & \mathrm{E}[\|\partial f(\mathbf{x}_t) - \partial f(\mathbf{x}_t; \xi_t, \varsigma_t)\| \|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_{t+1}\|] \\
\leq & \eta_t \mathrm{E}[\|\partial f(\mathbf{x}_t) - \partial f(\mathbf{x}_t; \xi_t, \varsigma_t)\|^2] \leq \eta_t G^2,
\end{aligned}$$

where the third equality is due to $\mathrm{E}[\langle \partial f(\mathbf{x}_t) - \partial f(\mathbf{x}_t; \xi_t, \varsigma_t), \widehat{\mathbf{x}}_{t+1} - \mathbf{x} \rangle | \mathbf{x}_t] = 0$; the last third inequality uses Cauchy-Schwartz inequality; the last second inequality is due to Lemma 3; the last inequality uses Assumption 2 (i). With above inequality, taking the expectation on both sides of (13) and using the fact that $\eta_t \leq 1/L$, we get

$$\mathrm{E}[f(\mathbf{x}_{t+1}) + \hat{r}(\mathbf{x}_{t+1}) - f(\mathbf{x}) - \hat{r}(\mathbf{x})] \leq \frac{\|\mathbf{x}_t - \mathbf{x}\|^2}{2\eta_t} - \frac{\|\mathbf{x} - \mathbf{x}_{t+1}\|^2}{2\eta_t} - \frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}_{t+1}\|^2 + \eta_t G^2.$$

D. Proof of Theorem 3

One might directly use the result in Proposition 2 to argue that the condition (8) holds by assuming that $\|\mathbf{x}_k - \mathbf{z}_k\|$ is bounded, which is true in the non-smooth case due to the domain constraint $\mathbf{x} \in \Omega$ in the update. In the smooth case, the upper bound is not directly available for setting T_k such that the condition (8) holds. Fortunately, when we apply the above result in the convergence analysis of Algorithm 1, we can utilize the strong convexity of F_k to cancel the term $O(\frac{\gamma\|\mathbf{z}_k - \mathbf{x}_k\|^2}{T_k(T_k+1)})$ by setting T_k to be larger than a constant.

Let us use the same notations as in the proof of Theorem 2 and prove the case of smooth g. By Proposition 2, we have

$$\mathrm{E}[F_k(\mathbf{x}_{k+1}) - F_k(\mathbf{z}_k)] \leq \frac{4L\|\mathbf{x}_k - \mathbf{z}_k\|^2}{T_k(T_k+3)} + \frac{2G^2}{(T_k+3)L}.$$

To continue the analysis, we have

$$\begin{aligned}\mathrm{E}[f_k(\mathbf{x}_{k+1}) + \frac{\gamma}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2] &\leq F_k(\mathbf{z}_k) + \frac{4L\|\mathbf{x}_k - \mathbf{z}_k\|^2}{T_k(T_k+3)} + \frac{2G^2}{(T_k+3)L} \\ &\leq F_k(\mathbf{x}_k) - \frac{\gamma}{2}\|\mathbf{x}_k - \mathbf{z}_k\|^2 + \frac{4L\|\mathbf{x}_k - \mathbf{z}_k\|^2}{T_k(T_k+3)} + \frac{2G^2}{(T_k+3)L} \\ &\leq g(\mathbf{x}_k) + r(\mathbf{x}_k) + \frac{2G^2}{(T_k+3)L}\end{aligned}$$

where we use $F_k(\mathbf{x}_k) \geq F_k(\mathbf{z}_k) + \frac{\gamma}{2}\|\mathbf{x}_k - \mathbf{z}_k\|^2$ due to the strong convexity of $F(\mathbf{x})$, and $\gamma \geq 3L, T_k \geq 4$. On the other hand, we have that

$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 =& \|\mathbf{x}_{k+1} - \mathbf{z}_k + \mathbf{z}_k - \mathbf{x}_s\|^2 \\ =& \|\mathbf{x}_{k+1} - \mathbf{z}_k\|^2 + \|\mathbf{z}_k - \mathbf{x}_k\|^2 + 2\langle \mathbf{x}_{k+1} - \mathbf{z}_k, \mathbf{z}_k - \mathbf{x}_k\rangle \\ \geq& (1-\hat{\alpha}^{-1})\|\mathbf{x}_{k+1} - \mathbf{z}_k\|^2 + (1-\hat{\alpha})\|\mathbf{x}_k - \mathbf{z}_k\|^2\end{aligned}$$

where the inequality follows from the Young's inequality with $0 < \hat{\alpha} < 1$. Thus we have that

$$\begin{aligned}\mathrm{E}\left[\frac{\gamma(1-\hat{\alpha})}{2}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq& \mathrm{E}[g(\mathbf{x}_k) + r(\mathbf{x}_k) - f_k(\mathbf{x}_{k+1})] + \frac{\gamma(\hat{\alpha}^{-1}-1)}{2}\mathrm{E}[\|\mathbf{x}_{k+1} - \mathbf{z}_k\|^2] \\ &+ \frac{2G^2}{(T_k+3)L}.\end{aligned}$$

On the other hand, by the convexity of $h(\cdot)$ we have

$$\begin{aligned}&\mathrm{E}[g(\mathbf{x}_k) + r(\mathbf{x}_k) - f_k(\mathbf{x}_{k+1})] \\ &\leq \mathrm{E}[g(\mathbf{x}_k) + r(\mathbf{x}_k) - g(\mathbf{x}_{k+1}) - r(\mathbf{x}_{k+1}) + \partial h(\mathbf{x}_k)^\top(\mathbf{x}_{k+1} - \mathbf{x}_k)] \\ &\leq \mathrm{E}[g(\mathbf{x}_k) + r(\mathbf{x}_k) - g(\mathbf{x}_{k+1}) - r(\mathbf{x}_{k+1}) + h(\mathbf{x}_{k+1}) - h(\mathbf{x}_k)] \\ &= \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})],\end{aligned}$$

By the strong convexity of $F_k(\mathbf{x})$, we also have

$$\frac{\gamma}{2}\mathrm{E}[\|\mathbf{x}_{k+1} - \mathbf{z}_k\|^2] \leq \mathrm{E}[F_k(\mathbf{x}_{k+1}) - F_k(\mathbf{z}_k)] \leq \frac{4L\|\mathbf{x}_k - \mathbf{z}_k\|^2}{T_k(T_k+3)} + \frac{2G^2}{(T_k+3)L}$$

Then we have

$$\mathrm{E}\left[\frac{\gamma(1-\hat{\alpha})}{2}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + (\hat{\alpha}^{-1}-1)\left(\frac{4L\|\mathbf{x}_k - \mathbf{z}_k\|^2}{T_k(T_k+3)} + \frac{2G^2}{(T_k+3)L}\right) + \frac{2G^2}{(T_k+3)L}$$

Let $\hat{\alpha} = 1/2$, we have

$$\begin{aligned}\mathrm{E}\left[\frac{\gamma}{4}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] &\leq \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + \frac{4L\|\mathbf{x}_k - \mathbf{z}_k\|^2}{T_k(T_k+3)} + \frac{4G^2}{(T_k+3)L} \\ &\leq \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + \frac{\gamma\|\mathbf{x}_k - \mathbf{z}_k\|^2}{8} + \frac{4G^2}{(T_k+3)L}\end{aligned}$$

where we use the fact $\gamma \geq 3L, T_k \geq 4$ and hence $4L/T_k^2 \leq \gamma/8$. It then gives us

$$\mathrm{E}\left[\frac{\gamma}{8}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + \frac{4G^2}{(T_k+3)L}$$

By setting $T_k = 4k/c$ with $0 < c \leq 1$, we have

$$\mathrm{E}\left[\frac{\gamma}{8}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + \frac{cG^2}{kL}$$

Following similar analysis to the proof of Theorem 2, we can finish the proof. For completeness, we include the remaining analysis here. Multiplying both sides by $w_k = k^\alpha$ and taking summation over $k = 1, \ldots, K$, we have

$$\mathrm{E}\left[\frac{\gamma}{8}\sum_{k=1}^{K} w_k\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq \mathrm{E}\left[\sum_{k=1}^{K} w_k(F(\mathbf{x}_k) - F(\mathbf{x}_{k+1}))\right] + \sum_{k=1}^{K} w_k\frac{cG^2}{kL}, \tag{14}$$

Similar to proof of of Theorem 2, we have

$$\mathrm{E}\left[\sum_{k=1}^{K} w_k(F(\mathbf{x}_k) - F(\mathbf{x}_{k+1}))\right] \leq \sum_{k=1}^{K}(w_k - w_{k-1})\mathrm{E}[(F(\mathbf{x}_k) - \min_{\mathbf{x}} F(\mathbf{x}))] \leq \Delta w_K$$

Then, by $\sum_{k=1}^{K} k^\alpha \geq \int_0^K x^\alpha dx = \frac{K^{\alpha+1}}{\alpha+1}$ and $\sum_{k=1}^{K} k^{\alpha-1} \leq K^\alpha,\ (\alpha \geq 1)$, we have

$$\mathrm{E}\left[\frac{\gamma}{8}\|\mathbf{z}_\tau - \mathbf{x}_\tau\|^2\right] \leq \frac{\Delta K^\alpha}{\sum_{k=1}^{K} k^\alpha} + \frac{cG^2\sum_{k=1}^{K} k^{\alpha-1}}{L\sum_{k=1}^{K} k^\alpha} \leq \frac{\Delta(\alpha+1)}{K} + \frac{cG^2(\alpha+1)}{LK},$$

which can complete the proof by multiplying both sides by 8γ.

Similarly, we can prove the case of non-smooth $g(\mathbf{x})$. For completeness, we include the details here. By Proposition 2 we have

$$\begin{aligned}
\mathrm{E}[f_k(\mathbf{x}_{k+1}) + \frac{\gamma}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2] &\leq F_k(\mathbf{z}_k) + \frac{\gamma\|\mathbf{x}_k - \mathbf{z}_k\|^2}{4T_k(T_k+1)} + \frac{28G^2}{\gamma(T_k+1)} \\
&\leq F_k(\mathbf{x}_k) - \frac{\gamma}{2}\|\mathbf{x}_k - \mathbf{z}_k\|^2 + \frac{\gamma\|\mathbf{x}_k - \mathbf{z}_k\|^2}{4T_k(T_k+1)} + \frac{28G^2}{\gamma(T_k+1)} \\
&= g(\mathbf{x}_k) + r(\mathbf{x}_k) + \frac{28G^2}{\gamma(T_k+1)}
\end{aligned}$$

where we use $F_k(\mathbf{x}_k) \geq F_k(\mathbf{z}_k) + \frac{\gamma}{2}\|\mathbf{x}_k - \mathbf{z}_k\|^2$ due to the strong convexity of $F(\mathbf{x})$, and $T_k \geq 1$. On the other hand, we have that

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 =& \|\mathbf{x}_{k+1} - \mathbf{z}_k + \mathbf{z}_k - \mathbf{x}_s\|^2 \\
=& \|\mathbf{x}_{k+1} - \mathbf{z}_k\|^2 + \|\mathbf{z}_k - \mathbf{x}_k\|^2 + 2\langle\mathbf{x}_{k+1} - \mathbf{z}_k, \mathbf{z}_k - \mathbf{x}_k\rangle \\
\geq& (1 - \hat{\alpha}^{-1})\|\mathbf{x}_{k+1} - \mathbf{z}_k\|^2 + (1 - \hat{\alpha})\|\mathbf{x}_k - \mathbf{z}_k\|^2
\end{aligned}$$

where the inequality follows from the Young's inequality with $0 < \hat{\alpha} < 1$. Thus we have that

$$\mathrm{E}\left[\frac{\gamma(1-\hat{\alpha})}{2}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq \mathrm{E}[g(\mathbf{x}_k) + r(\mathbf{x}_k) - f_k(\mathbf{x}_{k+1})] + \frac{\gamma(\hat{\alpha}^{-1} - 1)}{2}\mathrm{E}[\|\mathbf{x}_{k+1} - \mathbf{z}_k\|^2] + \frac{28G^2}{\gamma(T_k+1)}.$$

On the other hand, by the convexity of $h(\cdot)$ we have

$$\begin{aligned}
&\mathrm{E}[g(\mathbf{x}_k) + r(\mathbf{x}_k) - f_k(\mathbf{x}_{k+1})] \\
&\leq \mathrm{E}[g(\mathbf{x}_k) + r(\mathbf{x}_k) - g(\mathbf{x}_{k+1}) - r(\mathbf{x}_{k+1}) + \partial h(\mathbf{x}_k)^\top(\mathbf{x}_{k+1} - \mathbf{x}_k)] \\
&\leq \mathrm{E}[g(\mathbf{x}_k) + r(\mathbf{x}_k) - g(\mathbf{x}_{k+1}) - r(\mathbf{x}_{k+1}) + h(\mathbf{x}_{k+1}) - h(\mathbf{x}_k)] \\
&= \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})],
\end{aligned}$$

By the strong convexity of $F_k(\mathbf{x})$, we also have

$$\frac{\gamma}{2}\mathrm{E}[\|\mathbf{x}_{k+1} - \mathbf{z}_k\|^2] \leq \mathrm{E}[F_k(\mathbf{x}_{k+1}) - F_k(\mathbf{z}_k)] \leq \frac{\gamma\|\mathbf{x}_k - \mathbf{z}_k\|^2}{4T_k(T_k+1)} + \frac{28G^2}{\gamma(T_k+1)}$$

Then we have

$$\mathrm{E}\left[\frac{\gamma(1-\hat{\alpha})}{2}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + (\hat{\alpha}^{-1} - 1)\left(\frac{\gamma\|\mathbf{x}_k - \mathbf{z}_k\|^2}{4T_k(T_k+1)} + \frac{28G^2}{\gamma(T_k+1)}\right) + \frac{28G^2}{\gamma(T_k+1)}$$

Let $\hat{\alpha} = 1/2$, we have

$$\begin{aligned}
\mathrm{E}\left[\frac{\gamma}{4}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] &\leq \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + \frac{\gamma\|\mathbf{x}_k - \mathbf{z}_k\|^2}{4T_k(T_k+1)} + \frac{56G^2}{\gamma(T_k+1)} \\
&\leq \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + \frac{\gamma\|\mathbf{x}_k - \mathbf{z}_k\|^2}{8} + \frac{56G^2}{\gamma(T_k+1)}
\end{aligned}$$

where we use the fact $T_k \geq 1$. It then gives us

$$\mathrm{E}\left[\frac{\gamma}{8}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + \frac{56G^2}{\gamma(T_k+1)}$$

By setting $T_k = k/c$ with $0 < c \leq 1$, we have

$$\mathrm{E}\left[\frac{\gamma}{8}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + \frac{56cG^2}{k\gamma}$$

Multiplying both sides by $w_k = k^\alpha$ and taking summation over $k = 1, \ldots, K$, we have

$$\mathrm{E}\left[\frac{\gamma}{8}\sum_{k=1}^{K} w_k\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq \mathrm{E}\left[\sum_{k=1}^{K} w_k(F(\mathbf{x}_k) - F(\mathbf{x}_{k+1}))\right] + \sum_{k=1}^{K} w_k\frac{56cG^2}{k\gamma}, \tag{15}$$

Similar to proof of of Theorem 2, we have

$$\mathrm{E}\left[\sum_{k=1}^{K} w_k(F(\mathbf{x}_k) - F(\mathbf{x}_{k+1}))\right] \leq \sum_{k=1}^{K}(w_k - w_{k-1})\mathrm{E}[(F(\mathbf{x}_k) - \min_{\mathbf{x}} F(\mathbf{x}))] \leq \Delta w_K$$

Then, by $\sum_{k=1}^{K} k^\alpha \geq \int_0^K x^\alpha dx = \frac{K^{\alpha+1}}{\alpha+1}$ and $\sum_{k=1}^{K} k^{\alpha-1} \leq K^\alpha,\ (\alpha \geq 1)$, we have

$$\mathrm{E}\left[\frac{\gamma}{8}\|\mathbf{z}_\tau - \mathbf{x}_\tau\|^2\right] \leq \frac{\Delta K^\alpha}{\sum_{k=1}^{K} k^\alpha} + \frac{56cG^2\sum_{k=1}^{K} k^{\alpha-1}}{\gamma\sum_{k=1}^{K} k^\alpha} \leq \frac{\Delta(\alpha+1)}{K} + \frac{56cG^2(\alpha+1)}{L\gamma},$$

which can complete the proof by multiplying both sides by 8γ.

E. Proof of Theorem 4

The convergence analysis of using AdaGrad is build on the following proposition about the convergence AdaGrad for minimizing $F_{\mathbf{x}}^{\gamma}$, which is attributed to (Chen et al., 2018a), whose proof is presented later for completeness.

Proposition 3. *Let $H_0 = 2GI$ with $2G \geq \max_t \|\mathbf{g}_t\|_\infty$, and iteration number T be the smallest integer satisfying $T \geq M\max\{a(2G + \max_i \|g_{1:T,i}\|), \sum_{i=1}^{d}\|g_{1:T,i}\|/a, G_r\|\mathbf{x}_1 - \mathbf{x}_{T+1}\|/\eta\}$ for any $a > 0$. Algorithm 2 guarantees that*

$$\mathrm{E}[F_{\mathbf{x}_1}^{\gamma}(\widehat{\mathbf{x}}_T) - F_{\mathbf{x}_1}^{\gamma}(\mathbf{x}_*)] \leq \frac{1}{2aM\eta}\|\mathbf{x}_1 - \mathbf{x}_*\|^2 + \frac{(a+1)\eta}{M},$$

where $\mathbf{x}_* = \arg\min_{\mathbf{x}} F_{\mathbf{x}_1}^{\gamma}(\mathbf{x})$, and $g_{1:t,i}$ denotes the i-th row of $g_{1:t}$.

Let us use the same notations as in the proof of Theorem 2. By Proposition 3, we have

$$\mathrm{E}[F_k(\mathbf{x}_{k+1}) - F_k(\mathbf{z}_k)] \leq \frac{\|\mathbf{x}_k - \mathbf{z}_k\|^2}{2aM_k\eta_k} + \frac{(a+1)\eta_k}{M_k}$$

To continue the analysis, we have

$$\begin{aligned}
\mathrm{E}[f_k(\mathbf{x}_{k+1}) + \frac{\gamma}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2] &\leq F_k(\mathbf{z}_k) + \frac{\|\mathbf{x}_k - \mathbf{z}_k\|^2}{2aM_k\eta_k} + \frac{(a+1)\eta_k}{M_k} \\
&\leq F_k(\mathbf{x}_k) - \frac{\gamma}{2}\|\mathbf{x}_k - \mathbf{z}_k\|^2 + \frac{\|\mathbf{x}_k - \mathbf{z}_k\|^2}{2aM_k\eta_k} + \frac{(a+1)\eta_k}{M_k} \\
&\leq g(\mathbf{x}_k) + r(\mathbf{x}_k) + \frac{(a+1)\eta_k}{M_k}
\end{aligned}$$

where we use $F_k(\mathbf{x}_k) \geq F_k(\mathbf{z}_k) + \frac{\gamma}{2}\|\mathbf{x}_k - \mathbf{z}_k\|^2$ due to the strong convexity of $F_k(\mathbf{x})$, and $M_k\eta_k \geq \frac{4}{a\gamma}$. On the other hand, we have that

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 =& \|\mathbf{x}_{k+1} - \mathbf{z}_k + \mathbf{z}_k - \mathbf{x}_s\|^2 \\
=& \|\mathbf{x}_{k+1} - \mathbf{z}_k\|^2 + \|\mathbf{z}_k - \mathbf{x}_k\|^2 + 2\langle \mathbf{x}_{k+1} - \mathbf{z}_k, \mathbf{z}_k - \mathbf{x}_k\rangle \\
\geq& (1 - \hat{\alpha}^{-1})\|\mathbf{x}_{k+1} - \mathbf{z}_k\|^2 + (1 - \hat{\alpha})\|\mathbf{x}_k - \mathbf{z}_k\|^2
\end{aligned}$$

where the inequality follows from the Young's inequality with $0 < \hat{\alpha} < 1$. Thus we have that

$$\mathrm{E}\left[\frac{\gamma(1-\hat{\alpha})}{2}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq \mathrm{E}[g(\mathbf{x}_k) + r(\mathbf{x}_k) - f_k(\mathbf{x}_{k+1})] + \frac{\gamma(\hat{\alpha}^{-1} - 1)}{2}\mathrm{E}[\|\mathbf{x}_{k+1} - \mathbf{z}_k\|^2] + \frac{(a+1)\eta_k}{M_k}.$$

On the other hand, by the convexity of $h(\cdot)$ we have

$$\begin{aligned}
&\mathrm{E}[g(\mathbf{x}_k) + r(\mathbf{x}_k) - f_k(\mathbf{x}_{k+1})] \\
&\leq \mathrm{E}[g(\mathbf{x}_k) + r(\mathbf{x}_k) - g(\mathbf{x}_{k+1}) - r(\mathbf{x}_{k+1}) + \partial h(\mathbf{x}_k)^\top(\mathbf{x}_{k+1} - \mathbf{x}_k)] \\
&\leq \mathrm{E}[g(\mathbf{x}_k) + r(\mathbf{x}_k) - g(\mathbf{x}_{k+1}) - r(\mathbf{x}_{k+1}) + h(\mathbf{x}_{k+1}) - h(\mathbf{x}_k)] \\
&= \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})],
\end{aligned}$$

By the strong convexity of $F_k(\mathbf{x})$, we also have

$$\frac{\gamma}{2}\mathrm{E}[\|\mathbf{x}_{k+1} - \mathbf{z}_k\|^2] \leq \mathrm{E}[F_k(\mathbf{x}_{k+1}) - F_k(\mathbf{z}_k)] \leq \frac{\|\mathbf{x}_k - \mathbf{z}_k\|^2}{2aM_k\eta_k} + \frac{(a+1)\eta_k}{M_k}$$

Then we have

$$\mathrm{E}\left[\frac{\gamma(1-\hat{\alpha})}{2}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + (\hat{\alpha}^{-1} - 1)\left(\frac{\|\mathbf{x}_k - \mathbf{z}_k\|^2}{2aM_k\eta_k} + \frac{(a+1)\eta_k}{M_k}\right) + \frac{(a+1)\eta_k}{M_k}$$

Let $\hat{\alpha} = 1/2$, we have

$$\begin{aligned}
\mathrm{E}\left[\frac{\gamma}{4}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] &\leq \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + \frac{\|\mathbf{x}_k - \mathbf{z}_k\|^2}{2aM_k\eta_k} + \frac{2(a+1)\eta_k}{M_k} \\
&\leq \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + \frac{\gamma\|\mathbf{x}_k - \mathbf{z}_k\|^2}{8} + \frac{2(a+1)\eta_k}{M_k}
\end{aligned}$$

where we use the fact $M_k\eta_k \geq \frac{4}{a\gamma}$. It then gives us

$$\mathrm{E}\left[\frac{\gamma}{8}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + \frac{2(a+1)\eta_k}{M_k}$$

By setting $\eta_k = c/\sqrt{k}$, we have

$$\mathrm{E}\left[\frac{\gamma}{8}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + \frac{a(a+1)\gamma c^2}{2k}$$

Multiplying both sides by $w_k = k^\alpha$ and taking summation over $k = 1, \ldots, K$, we have

$$\mathrm{E}\left[\frac{\gamma}{8}\sum_{k=1}^{K} w_k\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq \mathrm{E}\left[\sum_{k=1}^{K} w_k(F(\mathbf{x}_k) - F(\mathbf{x}_{k+1}))\right] + \sum_{k=1}^{K} w_k\frac{a(a+1)\gamma c^2}{2k}, \tag{16}$$

Similar to proof of of Theorem 2, we have

$$\mathrm{E}\left[\sum_{k=1}^{K} w_k(F(\mathbf{x}_k) - F(\mathbf{x}_{k+1}))\right] \leq \sum_{k=1}^{K}(w_k - w_{k-1})\mathrm{E}[(F(\mathbf{x}_k) - \min_{\mathbf{x}} F(\mathbf{x}))] \leq \Delta w_K$$

Then, by $\sum_{k=1}^{K} k^\alpha \geq \int_0^K x^\alpha dx = \frac{K^{\alpha+1}}{\alpha+1}$ and $\sum_{k=1}^{K} k^{\alpha-1} \leq K^\alpha,\ (\alpha \geq 1)$, we have

$$\mathrm{E}\left[\frac{\gamma}{8}\|\mathbf{z}_\tau - \mathbf{x}_\tau\|^2\right] \leq \frac{\Delta K^\alpha}{\sum_{k=1}^{K} k^\alpha} + \frac{\gamma c^2 \sum_{k=1}^{K} k^{\alpha-1}}{2a^2\sum_{k=1}^{K} k^\alpha} \leq \frac{\Delta(\alpha+1)}{K} + \frac{a(a+1)\gamma c^2(\alpha+1)}{2K},$$

which can complete the proof by multiplying both sides by 8γ.

F. Proof of Proposition 3

First, we need to show that $\mathbf{x}_* = \arg\min F^\gamma_{\mathbf{x}_1}(\mathbf{x})$ in the set $\|\mathbf{x} - \mathbf{x}_1\| \leq \frac{2G+G_r}{\gamma}$. By the optimality condition of $\mathbf{x}_*$ we have

$$(\partial g(\mathbf{x}_*) + \partial r(\mathbf{x}_*) - \partial h(\mathbf{x}_1) + \gamma(\mathbf{x}_* - \mathbf{x}_1))^\top(\mathbf{x} - \mathbf{x}_*) \geq 0, \forall \mathbf{x} \in \mathrm{dom}(r)$$

Plugging $\mathbf{x} = \mathbf{x}_1$ into the above inequality, we have

$$\gamma\|\mathbf{x}_1 - \mathbf{x}_*\|^2 \leq (2G + G_r)\|\mathbf{x}_1 - \mathbf{x}_*\| \Rightarrow \|\mathbf{x}_1 - \mathbf{x}_*\| \leq \frac{2G + G_r}{\gamma},$$

where the first inequality uses Assumption 3.

Denote by $\psi_t(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_1)^\top H_t(\mathbf{x} - \mathbf{x}_1)$, $\psi_0(\mathbf{x}) = 0$ and $\|\mathbf{x}\|_H = \sqrt{\mathbf{x}^\top H\mathbf{x}}$, then we can see that $\psi_{t+1}(\mathbf{x}) \geq \psi_t(\mathbf{x})$ for any $t \geq 0$. Let $f(\mathbf{x}) = g(\mathbf{x}) - \partial h(\mathbf{x}_1)^\top(\mathbf{x} - \mathbf{x}_1)$ and $\hat{r}(\mathbf{x}) = r(\mathbf{x}) + \frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}_1\|^2$. Then $f(\mathbf{x})$ is convex and $\hat{r}(\mathbf{x})$ is

γ-strongly convex. Let $\mathbf{z}_t = \sum_{\tau=1}^{t} \mathbf{g}_t$, $\Delta_t = (\partial f(\mathbf{x}_t) - \mathbf{g}_t)^\top (\mathbf{x}_t - \mathbf{x})$ and

$$\psi_t^*(g) = \sup_{\mathbf{x}\in\Omega} g^\top \mathbf{x} - \frac{1}{\eta}\psi_t(\mathbf{x}) - t\hat{r}(\mathbf{x})$$

By the convexity of $f(\mathbf{x})$, and then taking the summation over all iterations, we get

$$\begin{aligned}
&\sum_{t=1}^{T}(f(\mathbf{x}_t) - f(\mathbf{x}) + \hat{r}(\mathbf{x}_t) - \hat{r}(\mathbf{x}))\\
\leq &\sum_{t=1}^{T}(\partial f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}) + \hat{r}(\mathbf{x}_t) - \hat{r}(\mathbf{x})) = \sum_{t=1}^{T}\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}) + \sum_{t=1}^{T}\Delta_t + \sum_{t=1}^{T}(\hat{r}(\mathbf{x}_t) - \hat{r}(\mathbf{x}))\\
= &\sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{x}_t - \sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{x} - \frac{1}{\eta}\psi_T(\mathbf{x}) - T\hat{r}(\mathbf{x}) + \frac{1}{\eta}\psi_T(\mathbf{x}) + \sum_{t=1}^{T}\Delta_t + \sum_{t=1}^{T}\hat{r}(\mathbf{x}_t)\\
\leq &\sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{x}_t + \sup_{\mathbf{x}\in\Omega}\left\{-\sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{x} - \frac{1}{\eta}\psi_T(\mathbf{x}) - T\hat{r}(\mathbf{x})\right\} + \frac{1}{\eta}\psi_T(\mathbf{x}) + \sum_{t=1}^{T}\Delta_t + \sum_{t=1}^{T}\hat{r}(\mathbf{x}_t)\\
= &\sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{x}_t + \psi_T^*(-\mathbf{z}_T) + \frac{1}{\eta}\psi_T(\mathbf{x}) + \sum_{t=1}^{T}\Delta_t + \sum_{t=1}^{T}\hat{r}(\mathbf{x}_t) \qquad (17)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\psi_T^*(-\mathbf{z}_T) &= -\sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{x}_{T+1} - \frac{1}{\eta}\psi_T(\mathbf{x}_{T+1}) - T\hat{r}(\mathbf{x}_{T+1})\\
&\leq -\sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{x}_{T+1} - \frac{1}{\eta}\psi_{T-1}(\mathbf{x}_{T+1}) - (T-1)\hat{r}(\mathbf{x}_{T+1}) - \hat{r}(\mathbf{x}_{T+1})\\
&\leq \sup_{\mathbf{x}\in\Omega}\left\{-\mathbf{z}_T^\top\mathbf{x} - \frac{1}{\eta}\psi_{T-1}(\mathbf{x}) - (T-1)\hat{r}(\mathbf{x})\right\} - \hat{r}(\mathbf{x}_{T+1})\\
&= \psi_{T-1}^*(-\mathbf{z}_T) - \hat{r}(\mathbf{x}_{T+1})\\
&\leq \psi_{T-1}^*(-\mathbf{z}_{T-1}) - \mathbf{g}_T^\top\nabla\psi_{T-1}^*(-\mathbf{z}_{T-1}) + \frac{\eta}{2}\|\mathbf{g}_T\|_{\psi_{T-1}^*}^2 - \hat{r}(\mathbf{x}_{T+1})
\end{aligned}$$

where the last inequality is due to $\psi_t(\mathbf{x})$ is 1-strongly convex w.r.t $\|\cdot\|_{\psi_t} = \|\cdot\|_{H_t}$ and consequentially $\psi_t^*(\mathbf{x})$ is η-smooth w.r.t. $\|\cdot\|_{\psi_t^*} = \|\cdot\|_{H_t^{-1}}$. Then

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{x}_t + \psi_T^*(-\mathbf{z}_T)\\
\leq &\sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{x}_t + \psi_{T-1}^*(-\mathbf{z}_{T-1}) - \mathbf{g}_T^\top\nabla\psi_{T-1}^*(-\mathbf{z}_{T-1}) + \frac{\eta}{2}\|\mathbf{g}_T\|_{\psi_{T-1}^*}^2 - \hat{r}(\mathbf{x}_{T+1})\\
= &\sum_{t=1}^{T-1}\mathbf{g}_t^\top\mathbf{x}_t + \psi_{T-1}^*(-\mathbf{z}_{T-1}) + \frac{\eta}{2}\|\mathbf{g}_T\|_{\psi_{T-1}^*}^2 - \hat{r}(\mathbf{x}_{T+1})
\end{aligned}$$

By repeating this process, we get

$$\sum_{t=1}^{T}\mathbf{g}_t^\top\mathbf{x}_t + \psi_T^*(-\mathbf{z}_T) \leq \psi_0^*(-\mathbf{z}_0) + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|_{\psi_{t-1}^*}^2 - \sum_{t=1}^{T}\hat{r}(\mathbf{x}_{t+1}) = \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|_{\psi_{t-1}^*}^2 - \sum_{t=1}^{T}\hat{r}(\mathbf{x}_{t+1}) \qquad (18)$$

Plugging the inequality (18) in the inequality (17), we have

$$\sum_{t=1}^{T}(f(\mathbf{x}_t) - f(\mathbf{x}) + \hat{r}(\mathbf{x}_t) - \hat{r}(\mathbf{x})) \leq \frac{1}{\eta}\psi_T(\mathbf{x}) + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|_{\psi_{t-1}^*}^2 + \sum_{t=1}^{T}\Delta_t + \hat{r}(\mathbf{x}_1) - \hat{r}(\mathbf{x}_{T+1}).$$

Algorithm 4 SVRG($F^{\gamma}_{\mathbf{x}_1}, \mathbf{x}_1, T, S$)

1: **Input**: $\mathbf{x}_1 \in \text{dom}(r)$, the number of inner initial iterations T_1, and the number of outer loops S.
2: $\bar{\mathbf{x}}^{(0)} = \mathbf{x}_1$
3: **for** $s = 1, 2, \ldots, S$ **do**
4: $\quad \bar{\mathbf{g}}_s = \nabla g(\bar{\mathbf{x}}^{(s-1)}) - \partial h(\mathbf{x}_1)$, $\mathbf{x}_0^{(s)} = \bar{\mathbf{x}}^{(s-1)}$
5: $\quad$ **for** $t = 1, 2, \ldots, T$ **do**
6: $\quad\quad$ Choose $i_t \in \{1, \ldots, n_1\}$ uniformly at random.
7: $\quad\quad \nabla_t^{(s)} = \nabla g_{i_t}(\mathbf{x}_{t-1}^{(s)}) - \nabla g_{i_t}(\bar{\mathbf{x}}^{(s-1)}) + \bar{\mathbf{g}}_s$
8: $\quad\quad \mathbf{x}_t^{(s)} = \arg\min_{\mathbf{x}} \{\langle \nabla_t^{(s)}, \mathbf{x} - \mathbf{x}_{t-1}^{(s)} \rangle + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{x}_{t-1}^{(r)}\|_2^2 + r(\mathbf{x}) + \frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}_1\|^2\}$
9: $\quad$ **end for**
10: $\quad \bar{\mathbf{x}}^{(s)} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t^{(s)}$
11: **end for**
12: **Output:** $\bar{\mathbf{x}}^{(S)}$

It is known from the analysis in (Duchi et al., 2011) that

$$\sum_{t=1}^{T} \|\mathbf{g}_t\|^2_{\psi^*_{t-1}} \leq 2\sum_{i=1}^{d} \|\mathbf{g}_{1:T,i}\|_2$$

Thus

$$\begin{aligned}
&\sum_{t=1}^{T} (f(\mathbf{x}_t) - f(\mathbf{x}) + \hat{r}(\mathbf{x}_t) - \hat{r}(\mathbf{x})) \\
&\leq \frac{2G\|\mathbf{x} - \mathbf{x}_1\|_2^2}{2\eta} + \frac{(\mathbf{x} - \mathbf{x}_1)^\top \text{diag}(\mathbf{s}_T)(\mathbf{x} - \mathbf{x}_1)}{2\eta} + \eta \sum_{i=1}^{d} \|g_{1:T,i}\|_2 + \sum_{t=1}^{T} \Delta_t + \hat{r}(\mathbf{x}_1) - \hat{r}(\mathbf{x}_{T+1}) \\
&\leq \frac{2G + \max_i \|g_{1:T,i}\|_2}{2\eta} \|\mathbf{x} - \mathbf{x}_1\|_2^2 + \eta \sum_{i=1}^{d} \|g_{1:T,i}\|_2 + \sum_{t=1}^{T} \Delta_t + (\partial \hat{r}(\mathbf{x}_1))^\top (\mathbf{x}_1 - \mathbf{x}_{T+1}) \\
&\leq \frac{2G + \max_i \|g_{1:T,i}\|_2}{2\eta} \|\mathbf{x} - \mathbf{x}_1\|_2^2 + \eta \sum_{i=1}^{d} \|g_{1:T,i}\|_2 + \sum_{t=1}^{T} \Delta_t + G_r \|\mathbf{x}_1 - \mathbf{x}_{T+1}\|_2
\end{aligned}$$

where the last inequality holds by using the fact that $\|\partial \hat{r}(\mathbf{x}_1)\| = \|\partial r(\mathbf{x}_1)\| \leq G_r$. Dividing by T and taking the expecation on both sides, then by using the convexity of $f(\mathbf{x}) + \hat{r}(\mathbf{x})$ and $\mathrm{E}[\sum_{t=1}^{T} \Delta_t / T] = 0$ according to the stopping time argument (Chen et al., 2018a)[Lemma 1, Supplement] we get

$$\begin{aligned}
&\mathrm{E}[F^{\gamma}_{\mathbf{x}_1}(\widehat{\mathbf{x}}_T) - F^{\gamma}_{\mathbf{x}_1}(\mathbf{x}_*)] \\
\leq &\mathrm{E}\left[\frac{2G + \max_i \|g_{1:T,i}\|_2}{2\eta T} \|\mathbf{x}_* - \mathbf{x}_1\|_2^2 + \frac{\eta}{T} \sum_{i=1}^{d} \|g_{1:T,i}\|_2 + \frac{G_r \|\mathbf{x}_1 - \mathbf{x}_{T+1}\|_2}{T}\right] \\
\leq &\frac{1}{2aM\eta} \|\mathbf{x}_1 - \mathbf{x}_*\|^2 + \frac{(a+1)\eta}{M},
\end{aligned}$$

where the last inequality is due to $T \geq M \max\{a(2G + \max_i \|g_{1:T,i}\|), \sum_{i=1}^{d} \|g_{1:T,i}\|/a, G_r\|\mathbf{x}_1 - \mathbf{x}_{T+1}\|/\eta\}$.

G. Analysis of using SVRG

The algorithm of SVRG for solving $F^{\gamma}_{\mathbf{x}_1}$ is presented in Algorithm 4 and its convergence result is given below.

Proposition 4. *By setting $\eta < 1/(4L)$ and T is large enough such that $\rho = \frac{1}{\gamma\eta(1-4L\eta)T} + \frac{4L\eta(T+1)}{(1-4L\eta)T} < 1$, then*

$$\mathrm{E}[F^{\gamma}_{\mathbf{x}_1}(\bar{\mathbf{x}}^{(S)}) - F^{\gamma}_{\mathbf{x}_1}(\mathbf{x}_*)] \leq \rho^S [F^{\gamma}_{\mathbf{x}_1}(\mathbf{x}_1) - F^{\gamma}_{\mathbf{x}_1}(\mathbf{x}_*)].$$

In particular, if we set $\eta = 0.05/L$, $T \geq \max(2, 200L/\gamma)$, we have

$$\mathrm{E}[F^{\gamma}_{\mathbf{x}_1}(\bar{\mathbf{x}}^{(S)}) - F^{\gamma}_{\mathbf{x}_1}(\mathbf{x}_*)] \leq 0.5^S [F^{\gamma}_{\mathbf{x}_1}(\mathbf{x}_1) - F^{\gamma}_{\mathbf{x}_1}(\mathbf{x}_*)].$$

Remark: The gradient complexity of SVRG is $(n + T)S$, where $n = n_1 + n_2$.

G.1. Proof of Theorem 5

Recall that $F(\mathbf{x}) = \frac{1}{n_1}\sum_{i=1}^{n_1} g_i(\mathbf{x}) + r(\mathbf{x}) - \frac{1}{n_2}\sum_{j=1}^{n_2} h_j(\mathbf{x})$ and $F_k(\mathbf{x}) = \frac{1}{n_1}\sum_{i=1}^{n_1} g_i(\mathbf{x}) + r(\mathbf{x}) - h(\mathbf{x}_k) - \frac{1}{n_2}\sum_{j=1}^{n_2} \partial h_j(\mathbf{x}_k)^\top(\mathbf{x}-\mathbf{x}_k) + \frac{\gamma}{2}\|\mathbf{x}-\mathbf{x}_k\|^2$. For any $\mathbf{x}$, by the convexity of $h(\mathbf{x})$ we know $F_k(\mathbf{x}) \geq F(\mathbf{x}) + \frac{\gamma}{2}\|\mathbf{x}-\mathbf{x}_k\|^2$. By applying the result in Proposition 4 to the k-th stage, we have

$$\mathrm{E}_k[F_k(\mathbf{x}_{k+1}) - F_k(\mathbf{z}_k)] \leq 0.5^{S_k}\mathrm{E}[F_k(\mathbf{x}_k) - F_k(\mathbf{z}_k)],$$

where $\mathbf{z}_k = \arg\min_{\mathbf{x}} F_k(\mathbf{x})$. Since $F_k(\mathbf{x}_{k+1}) - F_k(\mathbf{z}_k) \geq 0$ and $F_k(\mathbf{x}_k) - F_k(\mathbf{z}_k) \geq 0$, then $\mathrm{E}_k[F_k(\mathbf{x}_{k+1}) - F_k(\mathbf{z}_k)] \leq \mathrm{E}[F_k(\mathbf{x}_k) - F_k(\mathbf{z}_k)]$, which implies $\mathrm{E}_k[F_k(\mathbf{x}_{k+1})] \leq \mathrm{E}[F_k(\mathbf{x}_k)]$. Due to $F(\mathbf{x}_{k+1}) \leq F_k(\mathbf{x}_{k+1})$ and $F_k(\mathbf{x}_k) = F(\mathbf{x}_k)$, we have $\mathrm{E}_k[F(\mathbf{x}_{k+1})] \leq F(\mathbf{x}_k)$. Hence $\mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_*)] \leq F(\mathbf{x}_0) - F(\mathbf{x}_*) \leq \Delta$ for all k. Since

$$F_k(\mathbf{x}_k) - F_k(\mathbf{z}_k) \leq F(\mathbf{x}_k) - F(\mathbf{z}_k) \leq F(\mathbf{x}_k) - F(\mathbf{x}_*),$$

as a result, we have

$$\mathrm{E}_k[F_k(\mathbf{x}_{k+1}) - F_k(\mathbf{z}_k)] \leq 0.5^{S_k}[F(\mathbf{x}_k) - F(\mathbf{x}_*)].$$

To continue the analysis, we have

$$\begin{aligned}\mathrm{E}[F(\mathbf{x}_{k+1}) + \frac{\gamma}{2}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2] &\leq F_k(\mathbf{z}_k) + 0.5^{S_k}[F(\mathbf{x}_k) - F(\mathbf{x}_*)]\\ &\leq F_k(\mathbf{x}_k) - \frac{\gamma}{2}\|\mathbf{x}_k - \mathbf{z}_k\|^2 + 0.5^{S_k}[F(\mathbf{x}_k) - F(\mathbf{x}_*)]\\ &\leq F(\mathbf{x}_k) + 0.5^{S_k}[F(\mathbf{x}_k) - F(\mathbf{x}_*)],\end{aligned}$$

where we use $F_k(\mathbf{x}_k) \geq F_k(\mathbf{z}_k) + \frac{\gamma}{2}\|\mathbf{x}_k - \mathbf{z}_k\|^2$ due to the strong convexity of $F(\mathbf{x})$, and $F_k(\mathbf{x}_k) = F(\mathbf{x}_k)$. On the other hand, we have that

$$\begin{aligned}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 =& \|\mathbf{x}_{k+1} - \mathbf{z}_k + \mathbf{z}_k - \mathbf{x}_s\|^2\\ =& \|\mathbf{x}_{k+1} - \mathbf{z}_k\|^2 + \|\mathbf{z}_k - \mathbf{x}_k\|^2 + 2\langle \mathbf{x}_{k+1} - \mathbf{z}_k, \mathbf{z}_k - \mathbf{x}_k\rangle\\ \geq& (1-\hat{\alpha}^{-1})\|\mathbf{x}_{k+1} - \mathbf{z}_k\|^2 + (1-\hat{\alpha})\|\mathbf{x}_k - \mathbf{z}_k\|^2\end{aligned}$$

where the inequality follows from the Young's inequality with $0 < \hat{\alpha} < 1$. Thus we have that

$$\begin{aligned}\mathrm{E}\left[\frac{\gamma(1-\hat{\alpha})}{2}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq& \mathrm{E}[F(\mathbf{x}_k) - F_k(\mathbf{x}_{k+1})] + \frac{\gamma(\hat{\alpha}^{-1}-1)}{2}\mathrm{E}[\|\mathbf{x}_{k+1} - \mathbf{z}_k\|^2]\\ &+ 0.5^{S_k}[F(\mathbf{x}_k) - F(\mathbf{x}_*)].\end{aligned}$$

On the other hand, by the strong convexity of $F_k(\mathbf{x})$, we also have

$$\frac{\gamma}{2}\mathrm{E}[\|\mathbf{x}_{k+1} - \mathbf{z}_k\|^2] \leq \mathrm{E}[F_k(\mathbf{x}_{k+1}) - F_k(\mathbf{z}_k)] \leq 0.5^{S_k}[F(\mathbf{x}_k) - F(\mathbf{x}_*)]$$

Then we have

$$\begin{aligned}\mathrm{E}\left[\frac{\gamma(1-\hat{\alpha})}{2}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq& \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + (\hat{\alpha}^{-1}-1)\left(0.5^{S_k}[F(\mathbf{x}_k) - F(\mathbf{x}_*)]\right)\\ &+ 0.5^{S_k}[F(\mathbf{x}_k) - F(\mathbf{x}_*)]\end{aligned}$$

Let $\hat{\alpha} = 1/2$, we have

$$\mathrm{E}\left[\frac{\gamma}{4}\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq \mathrm{E}[F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})] + 2\times 0.5^{S_k}[F(\mathbf{x}_k) - F(\mathbf{x}_*)].$$

Multiplying both sides by $w_k = k^\alpha$ and taking summation over $k = 1, \ldots, K$, we have

$$\mathrm{E}\left[\frac{\gamma}{4}\sum_{k=1}^K w_k\|\mathbf{z}_k - \mathbf{x}_k\|^2\right] \leq \mathrm{E}\left[\sum_{k=1}^K w_k(F(\mathbf{x}_k) - F(\mathbf{x}_{k+1}))\right] + 2\mathrm{E}\left[\sum_{k=1}^K w_k 0.5^{S_k}[F(\mathbf{x}_k) - F(\mathbf{x}_*)]\right],$$

Then following the similar analysis as the proof of Theorem 2, we have

$$\mathrm{E}\left[\frac{\gamma}{4}\|\mathbf{z}_\tau - \mathbf{x}_\tau\|^2\right] \leq \frac{\Delta w_K}{\sum_{k=1}^K w_k} + \frac{2\Delta\sum_{k=1}^K w_k 0.5^{S_k}}{\sum_{k=1}^K w_k},$$

By noting that $0.5^{S_k} \leq 1/k$,

$$\mathrm{E}\left[\frac{\gamma}{4}\|\mathbf{z}_\tau - \mathbf{x}_\tau\|^2\right] \leq \frac{3\Delta(\alpha+1)}{K},$$

which can complete the proof by multiplying both sides by 4γ.

H. Proof of Theorem 6

The key is to connect the convergence in terms of $\|G_\gamma(\mathbf{x})\|$ to the convergence in terms of (sub)gradient. To this end, we present the following result.

Proposition 5. *If $g(\mathbf{x}) + r(\mathbf{x})$ is differentiable and has L_{g+r}-Hölder continuous gradient, we have*

$$dist(\partial h(\mathbf{x}), \nabla g(\mathbf{x}) + \nabla r(\mathbf{x})) \leq \frac{L_{g+r}}{\gamma^\nu}\|G_\gamma(\mathbf{x})\|^\nu + \|G_\gamma(\mathbf{x})\|.$$

If $h(\mathbf{x})$ is differentiable and has L_h-Hölder continuous gradient, we have

$$dist(\nabla h(\mathbf{x}_+), \partial g(\mathbf{x}_+) + \partial r(\mathbf{x}_+)) \leq \frac{L_h}{\gamma^\nu}\|G_\gamma(\mathbf{x})\|^\nu + \|G_\gamma(\mathbf{x})\|.$$

where $\mathbf{x}_+ = P_\gamma(\mathbf{x})$ and $G_\gamma(\mathbf{x}) = \gamma(\mathbf{x} - \mathbf{x}_+)$.

Proof of Proposition 5. From the proof of Proposition 1, we have

$$0 \in \partial g(\mathbf{x}_+) + \partial r(\mathbf{x}_+) - \partial h(\mathbf{x}) + \gamma(\mathbf{x}_+ - \mathbf{x}),$$

When $g(\mathbf{x}) + r(\mathbf{x})$ is differentiable and has L-Hölder continuous gradient, there exists $\mathbf{v} \in \partial h(\mathbf{x})$ such that

$$\begin{aligned}\|\nabla g(\mathbf{x}) + \nabla r(\mathbf{x}) - \mathbf{v}\| &= \|\nabla g(\mathbf{x}_+) + \nabla r(\mathbf{x}_+) - \nabla g(\mathbf{x}) - \nabla r(\mathbf{x})\| + \gamma\|\mathbf{x} - \mathbf{x}^+\| \\ &\leq L_{g+r}\|\mathbf{x} - \mathbf{x}_+\|^\nu + \|G_\gamma(\mathbf{x})\| = \frac{L_{g+r}}{\gamma^\nu}\|G_\gamma(\mathbf{x})\|^\nu + \|G_\gamma(\mathbf{x})\|.\end{aligned}$$

Similarly, we can prove the case when h is differentiable and had Hölder continuous gradient. □

Next, let's start with the proof of this theorem.

Proof of Theorem 6. By Jensen's inequality we know for any random variable X and a convex function $g(x)$,

$$\mathrm{E}[g(X)] \geq g(\mathrm{E}[X]).$$

Let $X = \|G_\gamma(\mathbf{x}_\tau)\|^v$ and $g(x) = x^{2/v}$, then $g(x)$ is a convex function since $0 < v \leq 1$. Therefore, we have

$$(\mathrm{E}[\|G_\gamma(\mathbf{x}_\tau)\|^v])^{2/v} \leq \mathrm{E}[\|G_\gamma(\mathbf{x}_\tau)\|^2] \leq O(1/K),$$

which implies

$$\mathrm{E}[\|G_\gamma(\mathbf{x}_\tau)\|^v] \leq O(1/K^{v/2}),\ (0 < v \leq 1). \tag{19}$$

We finish the proof by combining inequality (19) and the results in Proposition 5.

□

I. Proof of Lemma 1

The proof of the first fact can be found in (Liu et al., 2018)[Eqn. 7], and the second fact follows (Rockafellar & Wets, 1998)[Theorem 10.1 and Exercise 8.8].

Since $r(\cdot)$ is nonnegative proper closed, then $R_\mu(\mathbf{x})$ is convex continuous. By the definition of $r_\mu(x)$ and $\mathrm{prox}_{\mu r}(\mathbf{x})$ we know the supremum in $R_\mu(\mathbf{x})$ is attained at any point in $\mathrm{prox}_{\mu r}(\mathbf{x})$. Let $\mathbf{v} \in \mathrm{prox}_{\mu r}(\mathbf{x})$, then for any $\mathbf{w}$ we get

$$\begin{aligned}R_\mu(\mathbf{w}) - R_\mu(\mathbf{x}) =& \max_{y\in\mathbb{R}^d}\left\{\frac{1}{\mu}\mathbf{y}^\top\mathbf{w} - \frac{1}{2\mu}\|\mathbf{y}\|^2 - r(\mathbf{y})\right\} - \max_{y\in\mathbb{R}^d}\left\{\frac{1}{\mu}\mathbf{y}^\top\mathbf{x} - \frac{1}{2\mu}\|\mathbf{y}\|^2 - r(\mathbf{y})\right\} \\ \geq& \left\{\frac{1}{\mu}\mathbf{v}^\top\mathbf{w} - \frac{1}{2\mu}\|\mathbf{v}\|^2 - r(\mathbf{v})\right\} - \left\{\frac{1}{\mu}\mathbf{v}^\top\mathbf{x} - \frac{1}{2\mu}\|\mathbf{v}\|^2 - r(\mathbf{v})\right\} \\ =& \frac{1}{\mu}\mathbf{v}^\top(\mathbf{w} - \mathbf{x}),\end{aligned}$$

which implies $\frac{1}{\mu}\mathrm{prox}_{\mu r}(\mathbf{x}) \subseteq \partial R_\mu(\mathbf{x})$.

By Theorem 1.25 of (Rockafellar & Wets, 1998), the set $\mathrm{prox}_{\mu r}(\mathbf{x}) := \mathrm{Arg}\min_{\mathbf{y}\in\mathbb{R}^d}\left\{\frac{1}{2\mu}\|\mathbf{y} - \mathbf{x}\|^2 + r(\mathbf{y})\right\}$ is always nonempty since r is proper lower-semicontinuous and bounded below. Let $\mathbf{v} \in \mathrm{prox}_{\mu r}(\mathbf{x})$, then by Exercise 8.8 (c) and

Theorem 10.1 of (Rockafellar & Wets, 1998) we have

$$\frac{1}{\mu}(\mathbf{x}-\mathbf{v}) \in \hat{\partial} r(\mathbf{v})$$

J. Proof of Theorem 7

Theorem 7 is a corollary of the following theorem.

Theorem 8. *We have the following results:*

a. *If Assumption 4 (i) and Assumption 2 (i) hold, then we can use Algorithm 1 with Algorithm 3 (option 1) to solve (10) with $\mu=\epsilon$, which returns a solution $\mathbf{x}_\tau$ after $K=O(1/\epsilon^4)$ stages satisfying*

$$\mathrm{E}[\|\mathbf{x}_\tau - \mathbf{w}_\tau\|] \leq O(\epsilon),$$
$$\mathrm{E}[\mathit{dist}(\nabla h(\mathbf{w}_\tau), \nabla g(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau))] \leq O(\epsilon),$$

where $\mathbf{w}_\tau = \mathit{prox}_{\mu r}(\mathbf{x}_\tau)$.

b. *If Assumption 4 (ii) and Assumption 2 (i) hold, then we can use Algorithm 1 with Algorithm 3 (option 1) to solve (10) with $\mu=\epsilon^2$, which returns a solution $\mathbf{x}_\tau$ after $K=O(1/\epsilon^6)$ stages satisfying*

$$\mathrm{E}[\|\mathbf{x}_\tau - \mathbf{w}_\tau\|] \leq O(\epsilon),$$
$$\mathrm{E}[\mathit{dist}(\nabla h(\mathbf{w}_\tau), \nabla g(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau))] \leq O(\epsilon),$$

where $\mathbf{w}_\tau = \mathit{prox}_{\mu r}(\mathbf{x}_\tau)$.

c. *If g and h have a finite-sum form and g is smooth, then we can use Algorithm 1 with Algorithm 4 to solve (10). We can set $\mu=\epsilon$ if Assumption 4 (i) holds or $\mu=\epsilon^2$ if Assumption 4 (ii) or (iii) holds. The algorithm will return a solution $\mathbf{x}_\tau$ after $K=O(1/\epsilon^4)$ (corresponding to Assumption 4 (i)) or $K=O(1/\epsilon^6)$ (corresponding to Assumption 4 (ii) or (iii)) stages satisfying*

$$\mathrm{E}[\|\mathbf{x}_\tau - \mathbf{w}_\tau\|] \leq O(\epsilon),$$
$$\mathrm{E}[\mathit{dist}(\nabla h(\mathbf{w}_\tau), \nabla g(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau))] \leq O(\epsilon),$$

where $\mathbf{w}_\tau = \mathit{prox}_{\mu r}(\mathbf{x}_\tau)$.

Proof. In the following proof, $\hat{\partial} r(\mathbf{x})$ means there exists $\mathbf{v} \in \hat{\partial} r(\mathbf{x})$. By applying the stochastic algorithms for DC functions in last section, at each stage the following problem is solved approximately

$$\mathbf{z}_k = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \hat{g}(\mathbf{x}) + \frac{\gamma}{2}\|\mathbf{x}-\mathbf{x}_k\|^2 - (\nabla h(\mathbf{x}_k) + \frac{1}{\mu}\mathrm{prox}_{\mu r}(\mathbf{x}_k))^\top(\mathbf{x}-\mathbf{x}_k).$$

Then we have

$$\begin{aligned}\mathrm{E}[\|\nabla \hat{g}(\mathbf{x}_\tau) - \nabla h(\mathbf{x}_\tau) - \frac{1}{\mu}\mathrm{prox}_{\mu r}(\mathbf{x}_\tau)\|] &\leq \mathrm{E}[\|\nabla \hat{g}(\mathbf{z}_\tau) - \nabla \hat{g}(\mathbf{x}_\tau)\| + \gamma\|\mathbf{z}_\tau - \mathbf{x}_\tau\|]\\ &\leq \mathrm{E}[L\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + \frac{1}{\mu}\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + \gamma\|\mathbf{z}_\tau - \mathbf{x}_\tau\|],\end{aligned}$$

for any $\tau \in \{1,\ldots,K\}$. Denote by $\mathbf{w}_\tau = \mathrm{prox}_{\mu r}(\mathbf{x}_\tau)$. It is notable that $\frac{1}{\mu}\left(\mathbf{x}_\tau - \mathrm{prox}_{\mu r}(\mathbf{x}_\tau)\right) \in \hat{\partial} r(\mathbf{w}_\tau)$. Then we have $\nabla \hat{g}(\mathbf{x}_\tau) - \nabla h(\mathbf{x}_\tau) - \frac{1}{\mu}\mathrm{prox}_{\mu r}(\mathbf{x}_\tau) = \nabla g(\mathbf{x}_\tau) - \nabla h(\mathbf{x}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)$ and

$$\mathrm{E}[\|\nabla g(\mathbf{x}_\tau) - \nabla h(\mathbf{x}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)\|] \leq \mathrm{E}[L\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + \frac{1}{\mu}\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + \gamma\|\mathbf{z}_\tau - \mathbf{x}_\tau\|],$$

which implies that

$$\mathrm{E}[\|\nabla g(\mathbf{w}_\tau) - \nabla h(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)\|] \leq \mathrm{E}[(L+\gamma)\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + \frac{1}{\mu}\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + 2L\|\mathbf{x}_\tau - \mathbf{w}_\tau\|],$$

where uses the facts that g and h are smooth.

Next, we need to show that $\mathrm{E}[\|\mathbf{x}_\tau - \mathbf{w}_\tau\|]$ is small. The argument will be different for part (a), part (b) and part (c). For part (a), using

$$\frac{1}{2\mu}\|\mathbf{x}_\tau - \mathbf{w}_\tau\|^2 + r(\mathbf{w}_\tau) \leq r(\mathbf{x}_\tau)$$

we have

$$\frac{1}{2\mu}\|\mathbf{x}_\tau - \mathbf{w}_\tau\|^2 \leq r(\mathbf{x}_\tau) - r(\mathbf{w}_\tau) \leq G\|\mathbf{x}_\tau - \mathbf{w}_\tau\| \Rightarrow \|\mathbf{x}_\tau - \mathbf{w}_\tau\| \leq 2G\mu,$$

where the second inequality with an appropriate $G > 0$ follows the Lipchitz continuity of r. Then

$$\mathrm{E}[\|\nabla g(\mathbf{w}_\tau) - \nabla h(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)\|] \leq \mathrm{E}[(\gamma + L)\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + \frac{1}{\mu}\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + 4GL\mu].$$

By setting $\mu = \epsilon$ and $K = O(1/\epsilon^4)$ and τ randomly sampled, we have $\mathrm{E}[\|\mathbf{x}_\tau - \mathbf{z}_\tau\|] \leq \epsilon^2$ and hence

$$\mathrm{E}[\|\nabla g(\mathbf{w}_\tau) - \nabla h(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)\|] \leq O(\epsilon).$$

For part (b), using

$$\frac{1}{2\mu}\|\mathbf{x}_\tau - \mathbf{w}_\tau\|^2 + r(\mathbf{w}_\tau) \leq r(\mathbf{x}_\tau)$$

we have

$$\|\mathbf{x}_\tau - \mathbf{w}_\tau\| \leq \sqrt{2\mu\left(r(\mathbf{x}_\tau) - \min_{\mathbf{x}\in\mathbb{R}^d} r(\mathbf{x})\right)} \leq \sqrt{2\mu M},$$

where $M > 0$ exists due to Assumption 4(ii). Then

$$\mathrm{E}[\|\nabla g(\mathbf{w}_\tau) - \nabla h(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)\|] \leq \mathrm{E}[(\gamma + L)\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + \frac{1}{\mu}\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + 2L\sqrt{2\mu M}].$$

By setting $\mu = \epsilon^2$ and $K = O(1/\epsilon^6)$ and τ randomly sampled, we have $\mathrm{E}[\|\mathbf{x}_\tau - \mathbf{z}_\tau\|] \leq \epsilon^3$ and hence

$$\mathrm{E}[\|\nabla g(\mathbf{w}_\tau) - \nabla h(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)\|] \leq O(\epsilon).$$

For part (c), we take expectation over the above inequality giving

$$\mathrm{E}[\|\mathbf{x}_\tau - \mathbf{w}_\tau\|] \leq \sqrt{2\mu\left(\mathrm{E}[r(\mathbf{x}_\tau) - \min_{\mathbf{x}\in\mathbb{R}^d} r(\mathbf{x})]\right)}.$$

Since using the SVRG, we can show $\mathrm{E}[f(\mathbf{x}_\tau) + r_\mu(\mathbf{x}_\tau)]$ is bounded above, i.e., $\mathbf{x}_\tau$ is in a bounded set (in expectation), which together with the assumption r is lower bounded implies that there exists a constant $M > 0$ such that $\mathrm{E}[r(\mathbf{x}_\tau) - \min_{\mathbf{x}\in\mathbb{R}^d} r(\mathbf{x})] \leq M$ for $\tau = 1, \ldots, K$. Then

$$\mathrm{E}[\|\nabla g(\mathbf{w}_\tau) - \nabla h(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)\|] \leq \mathrm{E}[(\gamma + 3L)\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + \frac{1}{\mu}\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + 2L\sqrt{2\mu M}].$$

By setting $\mu = \epsilon^2$ and $K = O(1/\epsilon^6)$ and τ randomly sampled, we have $\mathrm{E}\|\mathbf{x}_\tau - \mathbf{z}_\tau\| \leq \epsilon^3$ and hence

$$\mathrm{E}[\|\nabla g(\mathbf{w}_\tau) - \nabla h(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)\|] \leq O(\epsilon).$$

□

K. Additional Analysis of Section 4

We mentioned that our results in Section 4 can be extended without much efforts to handle a differentiable h that has only a Hölder-continuous gradient. To formally state our new non-asymptototic convergence results, we rewrite Assumption 4 as follows.

Assumption 5. *Assume g is L-smooth and h has a Hölder-continuous gradient, and one of the following conditions holds:*

(i) r is Lipchitz continuous.

(ii) r is lower bounded and finite-valued over $\mathbb{R}^d$.

(ii) $f(\mathbf{x}) + r_\mu(\mathbf{x})$ is level bounded for a small $\mu < 1$, and r is finite-valued on a compact set, and lower bounded over $\mathbb{R}^d$.

Then our new convergence results for solving the problem (1) are presented in the following theorem.

Theorem 9. *We have the following results:*

a. *If Assumption 5 (i) and Assumption 2 (i) hold, then we can use Algorithm 1 with Algorithm 3 (option 1) to solve (10) with $\mu = \epsilon$, which returns a solution $\mathbf{x}_\tau$ after $K = O(1/\epsilon^4)$ stages satisfying*

$$\mathrm{E}[\|\mathbf{x}_\tau - \mathbf{w}_\tau\|] \leq O(\epsilon), \quad \mathrm{E}[\mathit{dist}(\nabla h(\mathbf{w}_\tau), \nabla g(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau))] \leq O(\epsilon^\nu),$$

where $\mathbf{w}_\tau = \mathit{prox}_{\mu r}(\mathbf{x}_\tau)$.

b. *If Assumption 5 (ii) and Assumption 2 (ii) hold, then we can use Algorithm 1 with Algorithm 3 (option 1) to solve (10) with $\mu = \epsilon^2$, which returns a solution $\mathbf{x}_\tau$ after $K = O(1/\epsilon^6)$ stages satisfying*

$$\mathrm{E}[\|\mathbf{x}_\tau - \mathbf{w}_\tau\|] \leq O(\epsilon), \quad \mathrm{E}[\mathit{dist}(\nabla h(\mathbf{w}_\tau), \nabla g(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau))] \leq O(\epsilon^\nu),$$

where $\mathbf{w}_\tau = \mathit{prox}_{\mu r}(\mathbf{x}_\tau)$.

c. *If g and h have a finite-sum form and g is smooth, then we can use Algorithm 1 with Algorithm 4 to solve (10). We can set $\mu = \epsilon$ if Assumption 5 (i) holds or $\mu = \epsilon^2$ if Assumption 5 (ii) or (iii) holds. The algorithm will return a solution $\mathbf{x}_\tau$ after $K = O(1/\epsilon^4)$ (corresponding to Assumption 5 (i)) or $K = O(1/\epsilon^6)$ (corresponding to Assumption 5 (ii) or (iii)) stages satisfying*

$$\mathrm{E}[\|\mathbf{x}_\tau - \mathbf{w}_\tau\|] \leq O(\epsilon), \quad \mathrm{E}[\mathit{dist}(\nabla h(\mathbf{w}_\tau), \nabla g(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau))] \leq O(\epsilon^\nu),$$

where $\mathbf{w}_\tau = \mathit{prox}_{\mu r}(\mathbf{x}_\tau)$.

Proof. In the following proof, $\hat{\partial} r(\mathbf{x})$ means there exists $\mathbf{v} \in \hat{\partial} r(\mathbf{x})$. By applying the stochastic algorithms for DC functions in last section, at each stage the following problem is solved approximately

$$\mathbf{z}_k = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \hat{g}(\mathbf{x}) + \frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}_k\|^2 - (\partial h(\mathbf{x}_k) + \frac{1}{\mu}\mathrm{prox}_{\mu r}(\mathbf{x}_k))^\top(\mathbf{x} - \mathbf{x}_k).$$

Then we have

$$\begin{aligned}\mathrm{E}[\|\nabla \hat{g}(\mathbf{x}_\tau) - \partial h(\mathbf{x}_\tau) - \frac{1}{\mu}\mathrm{prox}_{\mu r}(\mathbf{x}_\tau)\|] &\leq \mathrm{E}[\|\nabla \hat{g}(\mathbf{z}_\tau) - \nabla \hat{g}(\mathbf{x}_\tau)\| + \gamma\|\mathbf{z}_\tau - \mathbf{x}_\tau\|] \\ &\leq \mathrm{E}[L\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + \frac{1}{\mu}\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + \gamma\|\mathbf{z}_\tau - \mathbf{x}_\tau\|],\end{aligned}$$

for any $\tau \in \{1, \ldots, K\}$. Denote by $\mathbf{w}_\tau = \mathrm{prox}_{\mu r}(\mathbf{x}_\tau)$. It is notable that $\frac{1}{\mu}\left(\mathbf{x}_\tau - \mathrm{prox}_{\mu r}(\mathbf{x}_\tau)\right) \in \hat{\partial} r(\mathbf{w}_\tau)$. Then we have $\nabla \hat{g}(\mathbf{x}_\tau) - \partial h(\mathbf{x}_\tau) - \frac{1}{\mu}\mathrm{prox}_{\mu r}(\mathbf{x}_\tau) = \nabla g(\mathbf{x}_\tau) - \partial h(\mathbf{x}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)$ and

$$\mathrm{E}[\|\nabla g(\mathbf{x}_\tau) - \partial h(\mathbf{x}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)\|] \leq \mathrm{E}[L\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + \frac{1}{\mu}\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + \gamma\|\mathbf{z}_\tau - \mathbf{x}_\tau\|],$$

which implies that

$$\begin{aligned}&\mathrm{E}[\|\nabla g(\mathbf{w}_\tau) - \partial h(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)\|] \\ \leq &\mathrm{E}[(L+\gamma)\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + \frac{1}{\mu}\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + L\|\mathbf{x}_\tau - \mathbf{w}_\tau\| + L\|\mathbf{x}_\tau - \mathbf{w}_\tau\|^\nu],\end{aligned}$$

where uses the facts that g is L-smooth and h has L-Hölder-continuous gradient with parameter $\nu \in (0, 1]$.

Next, we need to show that $\mathrm{E}[\|\mathbf{x}_\tau - \mathbf{w}_\tau\|]$ is small. The argument will be different for part (a), part (b) and part (c). For part (a), using

$$\frac{1}{2\mu}\|\mathbf{x}_\tau - \mathbf{w}_\tau\|^2 + r(\mathbf{w}_\tau) \leq r(\mathbf{x}_\tau)$$

we have

$$\frac{1}{2\mu}\|\mathbf{x}_\tau - \mathbf{w}_\tau\|^2 \leq r(\mathbf{x}_\tau) - r(\mathbf{w}_\tau) \leq G\|\mathbf{x}_\tau - \mathbf{w}_\tau\| \Rightarrow \|\mathbf{x}_\tau - \mathbf{w}_\tau\| \leq 2G\mu,$$

where the second inequality with an appropriate $G > 0$ follows the Lipchitz continuity of r. Then

$$\mathrm{E}[\|\nabla g(\mathbf{w}_\tau) - \nabla h(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)\|] \leq \mathrm{E}[(\gamma + L)\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + \frac{1}{\mu}\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + 2GL\mu + L(2G\mu)^\nu].$$

By setting $\mu = \epsilon$ and $K = O(1/\epsilon^4)$ and τ randomly sampled, we have $\mathrm{E}[\|\mathbf{x}_\tau - \mathbf{z}_\tau\|] \leq \epsilon^2$ and hence

$$\mathrm{E}[\|\nabla g(\mathbf{w}_\tau) - \nabla h(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)\|] \leq O(\epsilon^\nu).$$

For part (b), using

$$\frac{1}{2\mu}\|\mathbf{x}_\tau - \mathbf{w}_\tau\|^2 + r(\mathbf{w}_\tau) \leq r(\mathbf{x}_\tau)$$

we have

$$\|\mathbf{x}_\tau - \mathbf{w}_\tau\| \leq \sqrt{2\mu\left(r(\mathbf{x}_\tau) - \min_{\mathbf{x}\in\mathbb{R}^d} r(\mathbf{x})\right)} \leq \sqrt{2\mu M},$$

where $M > 0$ exists due to Assumption 4(ii). Then

$$\mathrm{E}[\|\nabla g(\mathbf{w}_\tau) - \nabla h(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)\|] \leq \mathrm{E}[(\gamma + L)\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + \frac{1}{\mu}\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + L\sqrt{2\mu M} + L(2\mu M)^{\nu/2}].$$

By setting $\mu = \epsilon^2$ and $K = O(1/\epsilon^6)$ and τ randomly sampled, we have $\mathrm{E}[\|\mathbf{x}_\tau - \mathbf{z}_\tau\|] \leq \epsilon^3$ and hence

$$\mathrm{E}[\|\nabla g(\mathbf{w}_\tau) - \nabla h(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)\|] \leq O(\epsilon^\nu).$$

For part (c), we take expectation over the above inequality giving

$$\mathrm{E}[\|\mathbf{x}_\tau - \mathbf{w}_\tau\|] \leq \sqrt{2\mu\left(\mathrm{E}[r(\mathbf{x}_\tau) - \min_{\mathbf{x}\in\mathbb{R}^d} r(\mathbf{x})]\right)}.$$

Since using the SVRG, we can show $\mathrm{E}[f(\mathbf{x}_\tau) + r_\mu(\mathbf{x}_\tau)]$ is bounded above, i.e., $\mathbf{x}_\tau$ is in a bounded set (in expectation), which together with the assumption r is lower bounded implies that there exists a constant $M > 0$ such that $\mathrm{E}[r(\mathbf{x}_\tau) - \min_{\mathbf{x}\in\mathbb{R}^d} r(\mathbf{x})] \leq M$ for $\tau = 1, \ldots, K$. Then

$$\mathrm{E}[\|\nabla g(\mathbf{w}_\tau) - \nabla h(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)\|] \leq \mathrm{E}[(\gamma + L)\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + \frac{1}{\mu}\|\mathbf{x}_\tau - \mathbf{z}_\tau\| + L\sqrt{2\mu M} + L(2\mu M)^{\nu/2}].$$

By setting $\mu = \epsilon^2$ and $K = O(1/\epsilon^6)$ and τ randomly sampled, we have $\mathrm{E}\|\mathbf{x}_\tau - \mathbf{z}_\tau\| \leq \epsilon^3$ and hence

$$\mathrm{E}[\|\nabla g(\mathbf{w}_\tau) - \nabla h(\mathbf{w}_\tau) + \hat{\partial} r(\mathbf{w}_\tau)\|] \leq O(\epsilon^\nu).$$

□

L. DC decomposition of non-convex sparsity-promoting regularizers

We present the details of DC decomposition for several regularizers. The following examples are from (Wen et al., 2018; Gong et al., 2013).

Example 1. The DC decomposition of log-sum penalty (LSP) (Candès et al., 2008) is given by

$$r(\mathbf{x}) := \lambda\sum_{i=1}^d \log(|\mathbf{x}_i| + \theta) = \lambda\frac{\|\mathbf{x}\|_1}{\theta} - \underbrace{\lambda\sum_{i=1}^d\left(\frac{|\mathbf{x}_i|}{\theta} - \log(|\mathbf{x}_i| + \theta)\right)}_{r_2(\mathbf{x})},$$

where $\lambda > 0$ and $\theta > 0$. It was shown that $r_2(\mathbf{x})$ is convex and smooth with smoothness parameter $\frac{\lambda}{\theta^2}$.

Example 2. The DC decomposition of minimax concave penalty (MCP) (Zhang, 2010a) is given by

$$r(\mathbf{x}) := \lambda\sum_{i=1}^d\int_0^{|\mathbf{x}_i|}\left[1 - \frac{z}{\theta\lambda}\right]_+ dz = \lambda\|\mathbf{x}\|_1 - \underbrace{\lambda\sum_{i=1}^d\int_0^{|\mathbf{x}_i|}\min\left\{1, \frac{z}{\theta\lambda}\right\}dz}_{r_2(\mathbf{x})},$$

where $\lambda > 0$, $\theta > 0$, $[z]_+ = \max\{0, z\}$,

$$\lambda\int_0^{|\mathbf{x}_i|}\left[1 - \frac{z}{\theta\lambda}\right]_+ dz = \begin{cases} \lambda|\mathbf{x}_i| - \frac{\mathbf{x}_i^2}{2\theta} & \text{if } |\mathbf{x}_i| \leq \theta\lambda \\ \frac{\theta\lambda^2}{2} & \text{if } |\mathbf{x}_i| > \theta\lambda \end{cases}$$

and

$$\lambda\int_0^{|\mathbf{x}_i|}\min\left\{1, \frac{z}{\theta\lambda}\right\}dz = \begin{cases} \frac{\mathbf{x}_i^2}{2\theta} & \text{if } |\mathbf{x}_i| \leq \theta\lambda \\ \lambda|\mathbf{x}_i| - \frac{\theta\lambda^2}{2} & \text{if } |\mathbf{x}_i| > \theta\lambda \end{cases}$$

Then $r_2(\mathbf{x})$ is a convex and smooth function, and the smoothness parameter $\frac{1}{\theta}$.

Example 3. The DC decomposition of smoothly clipped absolute deviation (SCAD) (Fan & Li, 2001) is given by

$$r(\mathbf{x}) = \lambda \sum_{i=1}^{d} \int_0^{|\mathbf{x}_i|} \min\left\{1, \frac{[\theta\lambda - z]_+}{(\theta-1)\lambda}\right\} dz = \lambda\|\mathbf{x}\|_1 - \underbrace{\lambda \sum_{i=1}^{d} \int_0^{|\mathbf{x}_i|} \frac{[\min\{\theta\lambda, z\} - \lambda]_+}{(\theta-1)\lambda} dz}_{r_2(\mathbf{x})},$$

where $\lambda > 0$, $\theta > 2$,

$$\lambda \int_0^{|\mathbf{x}_i|} \min\left\{1, \frac{[\theta\lambda - z]_+}{(\theta-1)\lambda}\right\} dz = \begin{cases} \lambda|\mathbf{x}_i| & \text{if} \quad |\mathbf{x}_i| \le \lambda \\ \frac{-\mathbf{x}_i^2 + 2\theta\lambda|\mathbf{x}_i| - \lambda^2}{2(\theta-1)} & \text{if} \quad \lambda < |\mathbf{x}_i| \le \theta\lambda \\ \frac{(\theta+1)\lambda^2}{2} & \text{if} \quad |\mathbf{x}_i| > \theta\lambda \end{cases}$$

and

$$\lambda \int_0^{|\mathbf{x}_i|} \frac{[\min\{\theta\lambda, z\} - \lambda]_+}{(\theta-1)\lambda} dz = \begin{cases} 0 & \text{if} \quad |\mathbf{x}_i| \le \lambda \\ \frac{\mathbf{x}_i^2 - 2\lambda|\mathbf{x}_i| + \lambda^2}{2(\theta-1)} & \text{if} \quad \lambda < |\mathbf{x}_i| \le \theta\lambda \\ \lambda|\mathbf{x}_i| - \frac{(\theta+1)\lambda^2}{2} & \text{if} \quad |\mathbf{x}_i| > \theta\lambda \end{cases}$$

Then $r_2(\mathbf{x})$ was shown to be convex and smooth with modulus $\frac{1}{\theta-1}$.

Example 4. The DC decomposition of transformed ℓ_1 norm (Zhang & Xin, 2018) is given by

$$r(\mathbf{x}) := \sum_{i=1}^{d} \frac{(\theta+1)|\mathbf{x}_i|}{\theta + |\mathbf{x}_i|} = \frac{(1+\theta)\|\mathbf{x}\|_1}{\theta} - \underbrace{\sum_{i=1}^{d} \left[\frac{(\theta+1)|\mathbf{x}_i|}{\theta} - \frac{(\theta+1)|\mathbf{x}_i|}{\theta + |\mathbf{x}_i|}\right]}_{r_2(\mathbf{x})},$$

where $\theta > 0$. The function $r_2(\mathbf{x})$ is smooth with parameter $\frac{2(1+\theta)}{\theta^2}$.

Example 5. The DC decomposition of capped ℓ_1 penalty (Zhang, 2010b) is given by

$$r(\mathbf{x}) := \lambda \sum_{i=1}^{d} \min\{|\mathbf{x}_i|, \theta\} = \lambda\|\mathbf{x}\|_1 - \underbrace{\lambda \sum_{i=1}^{d} [|\mathbf{x}_i| - \theta]_+}_{r_2(\mathbf{x})},$$

where $\lambda > 0$, $\theta > 0$.