ISSN 1342-2804

Research Reports on Mathematical and Computing Sciences

New results on subgradient methods for strongly convex optimization problems with a unified analysis

Masaru Ito

April 2015, B–479

Department of
Mathematical and
Computing Sciences
Tokyo Institute of Technology

SERIES B: Applied Mathematical Science

New results on subgradient methods for strongly convex optimization problems with a unified analysis

Masaru Ito (*ito1@is.titech.ac.jp*)

Department of Mathematical and Computing Sciences, Tokyo Institute of Technology
2-12-1-W8-41 Oh-okayama, Meguro, Tokyo 152-8552 Japan

Research Report B-479
Department of Mathematical and Computing Sciences
Tokyo Institute of Technology
April 2014

Abstract

We develop subgradient- and gradient-based methods for minimizing strongly convex functions under a notion which generalizes the standard Euclidean strong convexity. We propose a unifying framework for subgradient methods which yields two kinds of methods, namely, the Proximal Gradient Method (PGM) and the Conditional Gradient Method (CGM), unifying several existing methods. The unifying framework provides tools to analyze the convergence of PGMs and CGMs for non-smooth, (weakly) smooth, and further for structured problems such as the inexact oracle models. The proposed subgradient methods yield optimal PGMs for several classes of problems and yield (nearly) optimal CGMs for smooth and weakly smooth problems, respectively.

Keywords: non-smooth/smooth convex optimization, structured convex optimization, subgradient/gradient-based proximal method, conditional gradient method, complexity theory, strongly convex functions, weakly smooth functions.

Mathematical Subject Classification (2010): 90C25, 68Q25, 49M37

1 Introduction

Subgradient- and gradient-based methods for convex optimization have been actively investigated in the last decades, providing efficient solutions for large scale optimization problems which arise from image/signal processing, data mining, statistics, *etc.* The efficiency of (sub)gradient-based methods are often analyzed from the viewpoint of oracle complexity [31, 33] to ensure a given absolute accuracy $\varepsilon > 0$ for the optimal value, and so far various "optimal" methods are known for several classes of problems. Achieving the optimal complexity for subgradient methods usually requires a priori problem specific information; sometimes, however, we can attain optimal or nearly optimal complexity with less such requirements (but we may need some restrictions for their implementations).

The following two classes of convex problems have been particularly well studied:

- *Non-smooth problems.* The problems of minimizing Lipschitz continuous convex functions with bounded subgradients;

- *Smooth problems.* The problems of minimizing continuously differentiable convex functions with Lipschitz continuous gradients.

These two classes of convex problems can be simultaneously formulated as *structured convex problems*, which have been receiving much attention in terms of both theory and application aspects. In particular, studies of (sub)gradient-based methods for the class of "smoothable" functions [1, 6, 9, 27, 34, 35], the class of composite functions [1, 5, 8, 17, 18, 19, 37, 40, 41], and the class of weakly smooth functions [11, 12, 38] are notably important.

In general, designing subgradient methods require easy-to-solve subproblems at each iteration. In this paper, we particularly focus on the following two kinds of (sub)gradient methods: the *Proximal (sub)Gradient Method (PGM)* and the *Conditional Gradient Method (CGM)* (all the methods we mention are PGM unless otherwise noted). The PGM is performed using a *prox-function* to define a reasonable proximal operator. Based on the conceptual complexity of Nemirovski and Yudin [31], many important PGMs for the above classes of convex problems can be proposed and their optimality can be discussed. As it will be pointed out in this paper, many of PGMs are modifications, accelerations, and/or combinations of two remarkably important PGMs, namely, the *Mirror-Descent Method (MDM)* [4, 31] and the *Dual-Averaging Method (DAM)* [36], which are optimal for non-smooth problems. The CGMs, on the other hand, are endowed by subproblems which are linear, *i.e.*, problems of minimizing a linear functional over a bounded convex feasible set. Originating from Frank and Wolfe [15], convergence properties of CGMs are well analyzed (see [10, 13, 16, 27, 39] and references therein). Because of its advantages such as easiness of subproblems and sparsity of approximate solutions, CGMs are actively studied with applications to machine learning and statistics [9, 21, 23, 24]; it is important to note that the CGMs have worse convergence rates than the PGMs, but the computational cost of each iteration for the CGMs can be lower, compensating the overall cost. Therefore, it is extremely important to choose between the PGM or the CGM depending on the problem structure of the problem to solve.

In a recent work [22], a unifying framework of PGMs were proposed, giving a unifying treatment to the MDM and the DAM for non-smooth problems as well as their accelerations for smooth (and structured) problems [40, 41]. The unifying framework enables us to generate a family of (optimal) subgradient methods which includes several existing optimal methods, and also to analyze both non-smooth and smooth problems under the same concept whereas existing analysis for them are performed individually. The work [22], however, was developed without assuming the strong convexity of objective functions. Using the knowledge of a strong convexity can help us to obtain much faster rate of convergence. For instance, the MDM [3, 29] for non-smooth problems and Nesterov's methods [33, 37] for smooth (or composite) problems realize the optimal complexity in the strongly convex cases. Moreover, exploiting a multistage procedure is a powerful tool to produce an optimal method [8, 19, 25, 30, 32, 37]. However, the multistage procedures require a priori knowledge of an upper bound of the distance between the initial point and the optimal solution set. Note that the optimality of the DAM for non-smooth problems and of the Tseng's method for smooth problems are not known without the multistage procedure.

This paper proposes a unifying framework of subgradient methods for convex problems with strongly convex objective functions and its convergence analysis for both non-smooth and smooth cases. In order to include the weakly smooth case, we actually consider a more general situation: the *inexact oracle model* studied in [11, 12]. This work can be seen as an extension of the recent work [22] to the strongly convex case with an additional generalization for the construction of auxiliary functions to minimize at each iteration. The proposed methods require a priori knowledge of the convexity parameter of the objective function, while it is not necessary to know an upper bound of the distance between the initial point and the optimal solution set to ensure the optimal rate of convergence with respect to the iteration number.

We emphasize three particular contributions of the current work. At first, the unifying framework yields the MDM and the DAM for non-smooth problems, and Nesterov's and Tseng's methods for smooth problems as special cases. As a consequence, we assert the optimality of these methods including the DAM and Tseng's method for the strongly convex case. Secondly, a family of CGMs can be obtained from the unifying framework including Lan's CGMs [27] and yielding an optimal complexity result for smooth problems in the non-strongly convex case; we further obtain convergence results of the proposed CGMs for the classes of weakly smooth functions. Finally, we provide new optimal convergence results for a weakly smooth extension of the deterministic case of [18] with less a prior requirements for the objective function.

This paper is organized as follows. We firstly discuss some general considerations about strongly convex problems in Section 2. In particular, in Section 2.1, we introduce a kind of "strong convexity" with respect to the prox-function and define the *non-smooth* and the *structured* problems listing some existing methods in the remind part. We consider the unified framework and general guidelines for constructing subproblems in Section 3. We propose general (sub)gradient methods and general convergence results under the framework in Section 4. Finally, in Section 5, we discuss the rate of convergences for the non-smooth and the structured problems providing the optimal complexity for them.

2 Problem settings and existing methods

2.1 Convex optimization problem and assumptions

Throughout this paper, we focus on the following convex optimization problem:

$$\min_{x \in Q} f(x) \tag{1}$$

where Q is a closed convex subset of a finite dimensional real normed space E equipped with a norm $\|\cdot\|$, and $f : E \to \mathbb{R} \cup \{+\infty\}$ is a lower-semicontinuous (lsc) convex function with $Q \subset \operatorname{dom} f$. We denote by E^* the dual space of E equipped with the dual norm $\|s\|_* = \max_{\|x\| \le 1} \langle s, x \rangle$ for $s \in E^*$ where $\langle s, x \rangle$ is the value of $s \in E^*$ at $x \in E$. We always assume that the problem (1) has an optimal solution $x^* \in Q$.

We introduce a *prox-function* $d(x)$ on the feasible set Q, that is, $d : E \to \mathbb{R} \cup \{+\infty\}$ is a nonnegative, continuously differentiable, and strongly convex function on Q (therefore, $Q \subset \operatorname{dom} d$) with a constant $\sigma_d > 0$ such that $d(x_0) = \min_{x \in Q} d(x) = 0$ for the unique minimizer $x_0 \in Q$. We use the notation $l_d(y; x) := d(y) + \langle \nabla d(y), x - y \rangle$ for the linearization of $d(x)$ at $y \in Q$. We also define the *Bregman distance* [7] between x and y for $x, y \in Q$ by

$$\xi(y, x) := d(x) - d(y) - \langle \nabla d(y), x - y \rangle = d(x) - l_d(y; x).$$

Note that the strong convexity of $d(x)$ on Q is equivalent to the property $\xi(y, x) \ge \frac{\sigma_d}{2}\|x - y\|^2$, $\forall x, y \in Q$. The prox-function as well as the Bregman distance will be used for the construction of auxiliary functions in the subproblems solved at each iterations in the methods described in this paper. A simple example for $d(x)$ is the *Euclidean setting*, namely, E is a Euclidean space with $\|x\|_2 = \langle x, x \rangle^{1/2}$, and $d(x) = \frac{1}{2}\|x - x_0\|_2^2$ for some $x_0 \in Q$.

For a lsc convex function $\psi : E \to \mathbb{R} \cup \{+\infty\}$ with $Q \subset \operatorname{dom} \psi$, we introduce the set

$$\sigma(\psi) := \{\tau \ge 0 \ : \ \psi(x) - \tau d(x) \text{ is a lsc convex function on } Q\}.$$

The set $\sigma(\psi)$ corresponds to the set of "convexity parameters" of $\psi(x)$ on Q with respect to the prox-function $d(x)$. In the Euclidean setting $d(x) = \frac{1}{2}\|x - x_0\|_2^2$, the set $\sigma(\psi)$ is the set of convexity parameters of $\psi(x)$ in the usual sense. Furthermore, in general, it can be shown that $\tau \in \sigma(\psi)$ if and only if the following inequality holds:

$$\psi(x) \geq \psi(y) + \psi'(y; x-y) + \tau\xi(y,x), \quad \forall x, y \in Q \ (\subset \mathrm{dom}\psi), \tag{2}$$

where $\psi'(x;d) = \lim_{\alpha\downarrow 0} \frac{\psi(x+\alpha d)-\psi(x)}{\alpha}$ $(x \in \mathrm{dom}\,\psi,\ d \in E)$ [1]. This form is similar to the characterization of the usual strong convexity of $\psi(x)$ on Q with constant $\tau \geq 0$: $\psi(x) \geq \psi(y) + \psi'(y; x-y) + \frac{\tau}{2}\|x-y\|^2$, $\forall x, y \in Q$. Therefore, $\tau \in \sigma(\psi)$ implies the usual strong convexity of $\psi(x)$ on Q with constant $\tau\sigma_d$, because of $\xi(y,x) \geq \frac{\sigma_d}{2}\|x-y\|^2$, $\forall x, y \in Q$. On the other hand, if the Bregman distance $\xi(y,x)$ *grows quadratically* on Q with a constant $A > 0$ (see [18]), *i.e.*, $\xi(y,x) \leq \frac{A}{2}\|x-y\|^2$, $\forall x, y \in Q$, then the usual strong convexity of $\psi(x)$ on Q with a constant $\tau \geq 0$ implies $\tau/A \in \sigma(\psi)$.

We assume a "strong convexity" of the objective function $f(x)$ by supposing that $\sigma(f)\backslash\{0\} \neq \emptyset$. However, in order to deal with several structured optimization problems as we will see in Section 2.3, we need to assume stronger conditions on the objective function as follows. Let us assume that, for each $y \in Q$, there exists a lsc convex function $l_f(y;\cdot) : E \to \mathbb{R}\cup\{+\infty\}$, such that $l_f(y;x) \leq f(x)$ for all $x \in Q$; and we further assume that there exists a convexity parameter $\sigma_f \geq 0$ such that

$$\sigma_f \in \sigma(f) \cap \bigcap_{y \in Q} \sigma(l_f(y;\cdot)). \tag{3}$$

Note that, since $f'(x^*; x-x^*) \geq 0$ holds for all $x \in Q$ by the optimality of x^*, the condition $\sigma_f \in \sigma(f)$ implies that $f(x) - f(x^*) \geq \sigma_f \xi(x^*, x)$ for all $x \in Q$. The function $l_f(y;x)$ can be seen as a strongly convex lower approximation of $f(x)$ at $y \in Q$. The condition (3) is not as restrictive as it is apparent to be specially if the problem (1) is provided by some structure.

The construction of $l_f(y;x)$ also depends on the problem structure. The convex optimization problem (1) which we consider in this paper can be divided into two classes:

- *Non-smooth problems.* We assume that the lower approximation model $l_f(\cdot;\cdot)$ is given by

$$l_f(y;x) := f(y) + \langle g(y), x-y\rangle + \sigma_f \xi(y,x) \tag{4}$$

  where $g(x) \in \partial f(x)$, $x \in Q$ is a subgradient mapping and $\sigma_f \in \sigma(f)$ is a known convexity parameter. Then, the requirement (3) follows because $l_f(y;x) - \sigma_f d(x)$ becomes an affine function. For convenience, we denote $g_k := g(x_k) \in \partial f(x_k)$ for test points x_k. Moreover, we assume that for every $s \in E^*$ and $\beta > 0$, the following optimization problem is solvable:

$$\min_{x \in Q}\{\langle s, x\rangle + \beta d(x)\}. \tag{5}$$

- *Structured problems.* We assume that we can construct a lower approximation model $l_f(\cdot;\cdot)$ of $f(\cdot)$ which admits (3) for some $\sigma_f \geq 0$ (see Section 2.3.1 to see how to define σ_f) and it satisfies

$$f(x) \leq [l_f(y;x) - \bar{\sigma}_f \xi(y,x)] + \frac{L(y)}{2}\|y-x\|^2 + \delta(y,x), \quad \forall x, y \in Q \tag{6}$$

[1] Notice that the function $\varphi(x) := \psi(x) - \tau d(x)$ satisfies $\varphi'(y; x-y) = \psi'(y; x-y) - \tau\,\langle \nabla d(y), x-y\rangle$, $\forall x, y \in Q$. Hence, the convexity of $\varphi(x)$ on Q implies $\varphi(x) \geq \varphi(y) + \varphi'(y; x-y), \forall x, y \in Q$, which is equivalent to (2). Conversely, since $\psi'(y; x-y) \geq -\psi'(y; y-x)$ holds and so is true for $\varphi(\cdot)$ for $x, y \in Q$, (2) implies the two inequalities $\varphi(y) \geq \varphi(z) + \varphi'(z; y-z)$ and $\varphi(x) \geq \varphi(z) - \varphi'(z; z-x)$ for $x, y, z \in Q$. Since $\varphi'(y;\cdot)$ is positively homogeneous, the convexity of $\varphi(\cdot)$ on Q follows by taking a convex combination of the two with $z = \alpha x + (1-\alpha)y, \alpha \in [0,1], x, y \in Q$.

for a nonnegative convex function $\delta(y, \cdot)$ on Q, and constants $\bar{\sigma}_f \geq 0$, $L(y) \geq 0$ with $\sigma_f \geq \bar{\sigma}_f$, $L(y) \geq \bar{\sigma}_f \sigma_d$ [2] . We further assume that for every $\beta \geq 0$, $y \in E$ and $s \in E^*$, the optimization problems of the following form is efficiently solvable:

$$\min_{x \in Q} \{ l_f(y; x) + \langle s, x \rangle + \beta d(x) \}. \tag{7}$$

Note that when $\beta = 0$ and $\sigma_f = 0$, the problem (7) may be a minimization of a convex function which is non-strongly convex, in particular, an affine function on Q. In this case, we additionally assume the boundedness of Q to ensure an existence of its solution. This is the case for the conditional gradient methods.

After developing a general analysis in Section 4.4, the function $\delta(y, x)$ will be finally particularized for the constant case $\delta(y, x) \equiv \delta$ in Sections 5.2, 5.3, and the case $\delta(y, x) := \frac{M}{\rho} \|x - y\|^\rho$, $M \geq 0$, $\rho \in [1, 2)$ in Section 5.4 (see Section 2.3 for several examples and related works).

2.2 Existing methods for non-smooth problems

Consider the non-smooth problems introduced above. We assume for the moment that the subgradient mapping $g(x) \in \partial f(x)$ of $f(x)$ is bounded: $\|g(x)\|_* \leq M$, $\forall x \in Q$. When $\sigma_f = 0$, the MDM and the DAM are known to be optimal methods. They were treated in a unified framework in [22, Method 9(a)] as we describe next. Let

$$x_0 := z_{-1} := \operatorname*{argmin}_{x \in Q} d(x), \quad x_{k+1} := z_k, \quad k \geq 0, \tag{8}$$

where z_k is determined by the solution of the following fixed subproblem either the *extended Mirror-Descent (MD) model*

$$\min_{x \in Q} \{ \lambda_k l_f(x_k; x) + \beta_k d(x) - \beta_{k-1} l_d(z_{k-1}; x) \}, \tag{9}$$

or the *Dual-Averaging (DA) model*

$$\min_{x \in Q} \left\{ \sum_{i=0}^{k} \lambda_i l_f(x_i; x) + \beta_k d(x) \right\}, \tag{10}$$

where $\{\lambda_k\}_{k \geq 0}$ and $\{\beta_k\}_{k \geq -1}$ are positive parameters called *weight* (or *step-size*) and *scaling parameters*, respectively; recall that $l_f(y; x) = f(y) + \langle g(y), x - y \rangle$ by the definition (4) if $\sigma_f = 0$. The MDM, originally proposed by Nemirovski and Yudin [31] and related to proximal subgradient methods by Beck and Teboulle [4], corresponds to the method (8) with the update (9) letting $\beta_k \equiv 1$. On the other hand, the method (8) with the update (10) yields the original DAM proposed by Nesterov [36]. Tuning the scaling parameter $\{\beta_k\}$ enables us to obtain an efficient convergence rate (see [22, 36]); for instance, taking $\lambda_k = 1$ and $\beta_k = O(\sqrt{k})$ yields that $f(\hat{x}_k) - f(x^*) \leq O(1/\sqrt{k})$. Furthermore, the optimal iteration complexity $O(M^2 d(x^*)/(\sigma_d \varepsilon^2))$ to obtain an ε-solution needs the values $d(x^*)$ and M to define λ_k and/or β_k.

When $\sigma_f > 0$ is known and the objective function $f(x)$ has a specific structure, the extended MDM also admits the optimal complexity $O(M^2/(\sigma_d \sigma_f \varepsilon))$ for the strongly convex case by choosing $\lambda_k := \frac{2}{\sigma_f(k+2)}$, $\beta_k := 1$ ([3, Proposition 3.1]; see also [29, Proposition 2.8] for related results).

[2] In view of $l_f(y; x) \leq f(x)$ and $\xi(y, x) \geq \frac{\sigma_d}{2} \|x - y\|^2$ for $x, y \in Q$, the inequality (6) yields $0 \leq (L(y) - \bar{\sigma}_f \sigma_d) \frac{1}{2} \|y - x\|^2 + \delta(y, x)$ from which the condition $L(y) \geq \bar{\sigma}_f \sigma_d$ is assumed; note that the particular choices $\delta(y, x) = \delta$ or $\frac{M}{\rho} \|y - x\|^\rho$ $(1 \leq \rho < 2)$ in this paper implies $\delta(y, x)/\|y - x\|^2 \to 0$ as $\|y - x\| \to \infty$.

Moreover, it is proved that a multistage procedure for the DAM achieves the same complexity for *uniformly convex* problems [25] with an application to stochastic optimization.

As we mention next, an extended class of problems including non-smooth and smooth ones are considered in [18, 19, 30, 38] which propose optimal methods for these problems and therefore for the non-smooth problems.

2.3 Examples and existing methods for structured problems

2.3.1 Examples of structured problems

The structured problems introduced in Section 2.1 include several special convex problems that are possibly non-smooth. We list some existing examples and results which can be discussed in this setting considering the requirements (3) and (6).

(i) *Smooth problems.* Suppose that $f(x)$ belongs to $C_L^{1,1}(Q)$; that is, $f(x)$ is continuously differentiable on Q and $\nabla f(x)$ satisfies the Lipschitz condition on Q with constant $L > 0$: $\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x-y\|$, $\forall x, y \in Q$. When we know a constant $\sigma_f \in \sigma(f)$, we can define

$$l_f(y;x) := f(y) + \langle \nabla f(y), x-y \rangle + \sigma_f \xi(y,x)$$

to obtain (3) and (6) with $L(\cdot) := L$, $\bar{\sigma}_f := \sigma_f$, and $\delta(\cdot,\cdot) := 0$. The corresponding subproblem (7) reduces to the form (5). The smooth problem with the Euclidean setting $d(x) = \frac{1}{2}\|x - x_0\|_2^2$ is the most basic one among the examples here; in this case, the lower complexity bounds $O(\sqrt{Ld(x^*)/\varepsilon})$ for the case $\sigma_f = 0$ and $O(\sqrt{L/\sigma_f}\log(1/\varepsilon))$ for the case $\sigma_f > 0$ are known for an absolute accuracy $\varepsilon > 0$. The first optimal PGM for the Euclidean case was proposed by Nesterov [32] and its variants were developed in [33], and in [2, 34] for non-strongly convex cases.

CGMs are also considered for the smooth problems, which achieve the complexity $O(LR/\varepsilon)$ where $R := \mathrm{Diam}(Q) = \sup_{x,y \in Q} \|x-y\|$ [10, 13, 15, 16, 27, 39]; excepting Lan's modified CGMs [27], all of these CGMs are based on the classical CGM [15], as we show in the algorithm (13).

(ii) *Composite problems.* Consider an objective function $f(x)$ of the form $f(x) = f_0(x) + \Psi(x)$ where $f_0 \in C_L^{1,1}(Q)$ and $\Psi(x)$ is a lsc convex function on Q with a simple structure. If we know constants $\sigma_{f_0} \in \sigma(f_0)$ and $\sigma_\Psi \in \sigma(\Psi)$, then, we can take

$$l_f(y;x) := f_0(y) + \langle \nabla f_0(y), x-y \rangle + \sigma_{f_0} \xi(y,x) + \Psi(x)$$

from which (3) and (6) hold with $\sigma_f := \sigma_{f_0} + \sigma_\Psi$, $L(\cdot) := L$, $\bar{\sigma}_f := \sigma_{f_0}$, and $\delta(\cdot,\cdot) := 0$. There are many PGMs for this problem [17, 5, 37, 40, 41] and they provide the same iteration complexity as the lowest complexity for the smooth problems in the non-strongly convex case (excepting the work by Fukushima and Mine [17] because they studied this model without assuming the convexity for $f_0(x)$). Nesterov [37] further proposed an optimal method for strongly convex composite problems in the Euclidean setting. The smoothing technique proposed by Nesterov [34] and its extension [6] for a special form of $\Psi(x)$ are also important because of their significant advantage in efficiency, which have further consideration in the strongly convex case [35].

A generalization of CGM to the composite problems was investigated in [1, 3] which also deal with a duality relationship to the MDM.

(iii) *Inexact oracle model.* Suppose that $f(x)$ is equipped with a *first-order* (δ, L, μ)*-oracle* [11], *i.e.*, for each $y \in Q$, we can compute $(f_{\delta,L,\mu}(y), g_{\delta,L,\mu}(y)) \in \mathbb{R} \times E^*$ such that

$$\frac{\mu}{2}\|x-y\|^2 \leq f(x) - (f_{\delta,L,\mu}(y) + \langle g_{\delta,L,\mu}(y), x-y \rangle) \leq \frac{L}{2}\|x-y\|^2 + \delta, \quad \forall x \in Q,$$

where $\delta \geq 0$ and $L \geq \mu \geq 0$. If $\mu = 0$ or the prox-function grows quadratically on Q with constant $A > 0$, then defining

$$l_f(y;x) := f_{\delta,L,\mu}(y) + \langle g_{\delta,L,\mu}(y), x-y\rangle + \frac{\mu}{A}\xi(y,x),$$

admits (3) and (6) with $L(\cdot) := L$, $\sigma_f := \bar{\sigma}_f := \mu/A$, and $\delta(\cdot,\cdot) := \delta$. The inexact oracle model with $\mu = 0$ was firstly studied by Devolder *et al.* [12] and they proposed the classical and the fast (proximal) gradient methods which were extended to the strongly convex case in [11]. A CGM for this model in the case $\mu = 0$ was analyzed by [16].

(iv) *Weakly smooth problems.* Suppose that the objective function $f(x)$ belongs to $C_M^{1,\nu}(Q)$ for some $\nu \in [0,1)$, *i.e.*, $f(x)$ is continuously differentiable on Q and $\nabla f(x)$ satisfies the Hölder condition $\|\nabla f(x) - \nabla f(y)\|_* \leq M\|x-y\|^\nu$, $\forall x,y \in Q$; but in the case $\nu = 0$, we do not assume the smoothness for $f(x)$ and we understand $\nabla f(x)$ as an element in $\partial f(x)$. Since the Hölder condition implies the inequality

$$f(x) - f(y) - \langle \nabla f(y), x-y\rangle \leq \frac{M}{1+\nu}\|x-y\|^{1+\nu}, \quad \forall x,y \in Q,$$

defining $l_f(y;x)$ as (i) for $\sigma_f \in \sigma(f)$, it admits (3) and (6) with $L(\cdot) := 0$, $\bar{\sigma}_f := \sigma_f$, and $\delta(\cdot,\cdot) := \frac{M}{1+\nu}\|x-y\|^{1+\nu}$. The weakly smooth version of the composite and the saddle structures can also be considered in the same way.

For the weakly smooth problems, Nemirovki and Nesterov [30] (see also [14, Section 2.3]) proposed an optimal method with the complexity bounds

$$c_1(\rho)\left(\frac{L}{\varepsilon}\right)^{\frac{2}{3\rho-2}}\left(\frac{d(x^*)}{\sigma_d}\right)^{\frac{\rho}{3\rho-2}} \quad \text{and} \quad c_2(\rho)\left(\frac{M^2}{\sigma^\rho}\frac{1}{\varepsilon^{2-\rho}}\right)^{\frac{1}{3\rho-2}}, \tag{11}$$

for non-strongly and strongly convex cases, respectively, where $\rho := 1+\nu \in [1,2)$, $c_1(\cdot), c_2(\cdot)$ are continuous functions, and $\sigma > 0$ is a convexity parameter of f with respect to the norm $\|\cdot\|$; the proposed method is further applicable for more general classes of problems. Moreover, Nesterov [38] improved a restriction of the method in the non-strongly convex case in the sense that the proposed method ensures the optimal convergence rate without fixing the iteration number. It is important to note that the methods proposed by [30] and [38] can achieve the above complexity of iterations for non-strongly convex case even if we do not know M and ν while the proposed method here needs an additional (but relatively small) "cost" for estimating M. This approach can be also seen in [5, 32, 37] for an estimation of the Lipschitz constant M in the case $\nu = 1$. The studies [11, 12] of the inexact oracle model are also important; they proposed an optimal method for weakly smooth problems in the non-strongly case and a sub-optimal one in the strongly convex case (PGMs for uniformly convex functions are also discussed).

A convergence result for CGMs for this class can be also obtained in the same way as the smooth problems which ensures the complexity $O((MR/\varepsilon)^{1/\nu})$ where $R := \mathrm{Diam}(Q)$ (see [9, Proposition 1.1]).

(v) The objective functions in (i) and (iv) can be simultaneously considered by assuming

$$f(y) - f(x) - \langle g(y), y-x\rangle \leq \frac{L}{2}\|y-x\|^2 + \frac{M}{\rho}\|y-x\|^\rho, \quad \forall x,y \in Q,$$

for a subgradient mapping $g(x) \in \partial f(x)$, $L, M \geq 0$, and $\rho \in [1,2)$. When $\sigma_f \in \sigma(f)$, we can take $l_f(y;x) := f(y) + \langle g(y), x-y\rangle + \sigma_f\xi(y,x)$ to obtain (3) and (6) with $L(\cdot) := L$, $\bar{\sigma}_f := \sigma_f$, and $\delta(y,x) := \frac{M}{\rho}\|y-x\|^\rho$. When $\sigma_f = 0$ or the prox-function grows quadratically on Q, (nearly) optimal methods for this model in the case $\rho = 1$ are studied in [8, 18, 19, 26, 28] with a stochastic setting.

2.3.2 Existing methods for structured problems

We finally describe some particular PGMs and CGMs which will be important for the comparison with the proposed methods in the paper. For that, we introduce two kinds of update formulas of gradient-based methods.

The first is the Classical Gradient Method [22, Method 16], which performs as follows: For given weight $\{\lambda_k\}_{k\geq 0}$ and scaling parameters $\{\beta_k\}_{k\geq -1}$, generate $\{z_k\}_{k\geq -1}$ and $\{x_k\}_{k\geq 0}$ by (8) and set $\{\hat{x}_k\}_{k\geq 0}$ by $\hat{x}_k = \sum_{i=0}^{k} \lambda_i z_i / \sum_{i=0}^{k} \lambda_i$. The *primal* and *dual gradient methods* in [37] for the composite problems (ii) and in [12] for the inexact oracle model (iii) are closely related to this algorithm in the non-strongly convex case. A further relation in the strongly convex case will be presented in this paper.

The second, the Fast Gradient Method (FGM) [22, Method 17], is described as follows: For given weight $\{\lambda_k\}_{k\geq 0}$ and scaling parameters $\{\beta_k\}_{k\geq -1}$, set $x_0 := z_{-1} := \operatorname{argmin}_{x\in Q} d(x)$, $\hat{x}_0 := z_0$ and, for $k \geq 0$, iterate

$$\begin{array}{rcl} x_{k+1} & := & (1-\tau_k)\hat{x}_k + \tau_k z_k, \quad \text{where } \tau_k := \frac{\lambda_{k+1}}{\sum_{i=0}^{k+1}\lambda_i}, \\ \hat{x}_{k+1} & := & (1-\tau_k)\hat{x}_k + \tau_k z_{k+1}, \end{array} \tag{12}$$

where z_k is determined by the fixed subproblem either the extended MD model (9) or the DA model (10). It was indicated in [22] that the FGM with $\lambda_0 := 1, \lambda_{k+1} := \frac{1+\sqrt{1+4\lambda_k^2}}{2}$ $(k \geq 0)$, and $\beta_k \equiv L/\sigma_d$ yields Tseng's accelerated PGMs [41] for the composite problems which achieve the optimal complexity as (i) in the non-strongly convex case. Furthermore, the algorithm (12) is also closely related to the following PGM and CGM, which will be unified in the framework of this paper:

- Replacing the second update in (12) by $\hat{x}_{k+1} := (1-\tau_k)\hat{x}_k + \tau_k w_{k+1}$ and determining w_k and z_k by (9) and (10) with $\beta_k := L/\sigma_d$, respectively, the corresponding method with $\lambda_k := (k+1)/2$ yields the Nesterov's optimal PGM [34, Section 5.3] for the smooth problems in the non-strongly convex case.

- Letting $\lambda_k := (k+1)/2$ and assuming the boundedness of Q, Lan's modified CGMs, Algorithms 4 and 5 in [27] with the stepsize policy $\alpha_k := 2/(k+1)$ and $\theta_k := k$, can be described as the algorithm (12) with the subproblems (9) and (10) with $\beta_k \equiv 0$, respectively.

On the other hand, the classical CGM [10, 15, 39] for smooth problems is basically performed as follows: Choose $x_0 \in Q$ and, for $k \geq 0$, iterate

$$z_k \in \operatorname*{Argmin}_{x\in Q} \langle \nabla f(x_k), x - x_k \rangle, \quad x_{k+1} := (1-\tau_k)x_k + \tau_k z_k, \quad k \geq 0 \tag{13}$$

where $\tau_k \in [0,1)$ (we assume the boundedness of Q). Excepting the Lan's modified CGMs, all the above mentioned CGMs are based on this classical CGM. Notice that the subproblem can be seen as the extended MD model (9) with $\beta_k \equiv 0$.

3 Unified framework for (strongly) convex objective functions

We discuss in this section how we can construct auxiliary functions which need to be minimized at each iteration of the proposed methods. The only assumption required in this section is the (strong) convexity (3), and therefore the characterization of auxiliary functions given in this section can be used for both non-smooth problems (4) and structured problems (6) as we will discuss in Section 4.

Let us set some notations and objects for our development. At first, we introduce the following sequences which take the roles of parameters for tuning our methods: $\{\lambda_k\}_{k\geq 0}$ the sequence of positive real numbers (the *weight parameters*) and $\{\beta_k\}_{k\geq -1}$ the nondecreasing sequence of nonnegative real numbers (the *scaling parameters*). Our (sub)gradient-based methods solve (one or) two subproblems of the forms $\min_{x\in Q}\varphi_k(x)$ and $\min_{x\in Q}\psi_k(x)$ for some auxiliary functions $\varphi_k(x)$ and $\psi_k(x)$, respectively, defined at each iterations. Then, these methods generate the following sequences $\{x_k\}_{k\geq 0}$, $\{z_k\}_{k\geq -1}$, $\{w_k\}_{k\geq -1}$, $\{\hat{x}_k\}_{k\geq 0}$ in Q.

- $\{x_k\}_{k\geq 0}$ is the sequence of test points for which we evaluate $l_f(x_k; x)$.
- $\{z_k\}_{k\geq -1}$ is the sequence of solutions of subproblems $\min_{x\in Q}\varphi_k(x)$ where the auxiliary function $\varphi_k(x)$ can be determined by $\{x_i\}_{i=0}^{k}, \{z_i\}_{i=-1}^{k-1}$, $\{w_i\}_{i=-1}^{k-1}$, $\{\lambda_i\}_{i=0}^{k}$, and $\{\beta_i\}_{i=-1}^{k}$.
- $\{w_k\}_{k\geq -1}$ is the sequence of solutions of subproblems $\min_{x\in Q}\psi_k(x)$ where the auxiliary function $\psi_k(x)$ can be determined by $\{x_i\}_{i=0}^{k}, \{z_i\}_{i=-1}^{k}$, $\{w_i\}_{i=-1}^{k-1}$, $\{\lambda_i\}_{i=0}^{k}$, and $\{\beta_i\}_{i=-1}^{k}$; notice that w_k can possibly depend on z_k.
- $\{\hat{x}_k\}_{k\geq 0}$ is the sequence of approximate solutions for the problem (1).

Therefore, we understand that the auxiliary functions $\{(\varphi_k(x), \psi_k(x))\}_{k\geq -1}$ are constructed associated with weight parameters $\{\lambda_k\}_{k\geq 0}$, scaling parameters $\{\beta_k\}_{k\geq -1}$, and test points $\{x_k\}_{k\geq 0}$; the associated objects are sometimes omitted. We often consider the case of a single sequence $\{\varphi_k(x)\}_{k\geq -1}$ of auxiliary functions which can be regarded as the case $\psi_k(x)\equiv\varphi_k(x)$.

We will gradually specify the above general objects by giving explicit update formulas in three steps: The first is for the auxiliary functions $\{(\varphi_k(x), \psi_k(x))\}_{k\geq -1}$ in this section, the second is for the points $\{x_k\}_{k\geq 0}$ and $\{\hat{x}_k\}_{k\geq 0}$ by proposing general methods (Section 4), and the final is for the parameters $\{\lambda_k\}_{k\geq 0}$ and $\{\beta_k\}_{k\geq -1}$ to provide efficient convergences (Section 5).

The key concept for our analysis is based on the inequality

$$(R_k) \quad S_k f(\hat{x}_k) \leq \psi_k(w_k) + C_k$$

where $S_k := \sum_{i=0}^{k}\lambda_i$ and $C_k \geq 0$ is some constant. The purpose of this section is to propose a framework to construct $\varphi_k(x)$ and $\psi_k(x)$ which enables us to have a method satisfying the relation (R_k) and to derive an efficient convergence from this relation.

3.1 General properties for the construction of auxiliary functions

We begin by describing general properties which the auxiliary functions $\{(\varphi_k(x), \psi_k(x))\}_{k\geq -1}$ should satisfy. These properties will guide us in how to iteratively construct the auxiliary functions. The first set of properties is for a sequence of auxiliary functions $\{\varphi_k(x)\}_{k\geq -1}$. We define $\sum_{i=0}^{-1}(\cdot) := 0$ and so $S_{-1} = 0$.

Property A. *Let $\{\varphi_k(x)\}_{k\geq -1}$ be a sequence of auxiliary functions associated with weight parameters $\{\lambda_k\}_{k\geq 0}$, scaling parameters $\{\beta_k\}_{k\geq -1}$, and test points $\{x_k\}_{k\geq 0}$. Let $\sigma_f \geq 0$ be a convexity parameter satisfying (3). Denote $z_k := \operatorname{argmin}_{x\in Q}\varphi_k(x)$[3]. Then, the following conditions hold:*

(A1) $\varphi_{-1}(z_{-1}) = 0$ and $z_{-1} = x_0$.

(A2) $\forall k \geq -1$, $\forall x \in Q$, we have

$$\varphi_{k+1}(x) \geq \varphi_k(z_k) + \lambda_{k+1} l_f(x_{k+1}; x) + \beta_{k+1} d(x) - \beta_k l_d(z_k; x) + S_k \sigma_f \xi(z_k, x).$$

[3] The auxiliary function $\varphi_k(x)$ can possibly be an affine function. In that case, we will assume the boundedness of Q in order to ensure an existence of a minimizer z_k.

(A3) $\forall k \geq -1, \quad \varphi_k(z_k) \leq \min_{x\in Q} \left\{ \sum_{i=0}^{k} \lambda_i l_f(x_i;x) + \beta_k l_d(z_k;x) - S_k \sigma_f \xi(z_k,x) \right\}$.

The above property is an extension of Property 2 in [22] which is particularized by taking $\sigma_f = 0$. As a simple extension of Property A, we further consider a coupled sequence $\{(\varphi_k(x), \psi_k(x))\}_{k\geq -1}$ of auxiliary functions which admits the property below.

Property B. *Let $\{(\varphi_k(x), \psi_k(x))\}_{k\geq -1}$ be a coupled sequence of auxiliary functions associated with weight parameters $\{\lambda_k\}_{k\geq 0}$, scaling parameters $\{\beta_k\}_{k\geq -1}$, and test points $\{x_k\}_{k\geq 0}$. Denote $z_k := \operatorname{argmin}_{x\in Q} \varphi_k(x)$ and $w_k := \operatorname{argmin}_{x\in Q} \psi_k(x)$. Let $\sigma_f \geq 0$ be a convexity parameter satisfying (3). Then, the following conditions hold:*

(B0) $\varphi_k(x) \geq \psi_k(x)$ *for all* $x \in Q$.

(B1) $\psi_{-1}(w_{-1}) = 0$ *and* $z_{-1} = w_{-1} = x_0$.

(B2) $\forall k \geq -1,\ \forall x \in Q$, *we have*

$$\psi_{k+1}(x) \geq \varphi_k(z_k) + \lambda_{k+1} l_f(x_{k+1};x) + \beta_{k+1} d(x) - \beta_k l_d(z_k;x) + S_k \sigma_f \xi(z_k,x).$$

(B3) $\forall k \geq -1, \quad \psi_k(w_k) \leq \min_{x\in Q} \left\{ \sum_{i=0}^{k} \lambda_i l_f(x_i;x) + \beta_k l_d(z_k;x) - S_k \sigma_f \xi(z_k,x) \right\}$.

Note that letting $\psi_k(x) \equiv \varphi_k(x)$, it yields Property A.

3.2 Construction of auxiliary functions

Here we provide some formulas to construct a coupled sequence $\{(\varphi_k(x), \psi_k(x))\}_{k\geq -1}$ of auxiliary functions which admit Property B. For that, we firstly construct a single sequence of auxiliary functions $\{\varphi_k(x)\}_{k\geq -1}$ satisfying Property A.

Theorem 3.1. *Given the weight parameters $\{\lambda_k\}_{k\geq 0}$, the scaling parameters $\{\beta_k\}_{k\geq -1}$, the test points $\{x_k\}_{k\geq 0}$, and a convexity parameter $\sigma_f \geq 0$ satisfying (3), construct the sequence $\{\varphi_k(x)\}_{k\geq -1}$ of auxiliary functions as follows. $\varphi_{-1}(x) := \beta_{-1} d(x)$, $z_{-1} := x_0$ and, for $k \geq -1$, define*

$$\varphi_{k+1}(x) := \varphi_k(z_k) + \lambda_{k+1} l_f(x_{k+1};x) + \beta_{k+1} d(x) - \beta_k l_d(z_k;x) + S_k \sigma_f \xi(z_k,x) \qquad (14)$$

or

$$\varphi_{k+1}(x) := \varphi_k(x) + \lambda_{k+1} l_f(x_{k+1};x) + \beta_{k+1} d(x) - \beta_k d(x). \qquad (15)$$

Then, the sequence $\{\varphi_k(x)\}_{k\geq -1}$ satisfies Property A.

The assumption $z_{-1} := x_0$ is satisfied whenever $\beta_{-1} > 0$ because $\min_{x\in Q} d(x) = d(x_0) = 0$, but it is required when $\beta_{-1} = 0$; in both cases, the condition (A1) holds. To prove Theorem 3.1, it remains to show (A2) and (A3) which will be done in Lemmas 3.4 and 3.5, respectively.

The following theorem is a simple consequence of Theorem 3.1.

Theorem 3.2. *Let $\{\varphi_k(x)\}_{k\geq -1}$ be generated accordingly to the construction in Theorem 3.1. Define $\{\psi_k(x)\}_{k\geq -1}$ by $\psi_{-1}(x) := \varphi_{-1}(x)$ and*

$$\psi_{k+1}(x) := \varphi_k(z_k) + \lambda_{k+1} l_f(x_{k+1};x) + \beta_{k+1} d(x) - \beta_k l_d(z_k;x) + S_k \sigma_f \xi(z_k,x). \qquad (16)$$

Then, the sequence $\{(\varphi_k(x), \psi_k(x))\}$ satisfies Property B.

Proof. Notice that (16) satisfies the condition (B2) as equality. The condition (B1) is immediate from the condition (A1) for $\{\varphi_k(x)\}$ and the definition $\psi_{-1}(x) := \varphi_{-1}(x)$. Since (16) coincides with the right hand side of (A2) for $\{\varphi_k(x)\}$, the condition (B0) is clear. Finally, the condition (B3) is satisfied by (B0) and (A3) for $\{\varphi_k(x)\}$. □

Before proving Theorem 3.1, let us see some particular constructions of auxiliary functions, which will be useful for the comparison with some existing methods.

- *Extended MD model.* Define $\{\varphi_k(x)\}_{k\geq -1}$ by $\varphi_{-1}(x) := \beta_{-1}d(x)$ and

$$\varphi_{k+1}(x) := \varphi_k(z_k) + \lambda_{k+1}l_f(x_{k+1};x) + \beta_{k+1}d(x) - \beta_k l_d(z_k;x) + S_k\sigma_f\xi(z_k,x) \tag{17}$$

  for $k \geq -1$. Then, Property A follows from Theorem 3.1 with the update (14).

- *DA model.* Define $\{\varphi_k(x)\}_{k\geq -1}$ by $\varphi_{-1}(x) := \beta_{-1}d(x)$ and

$$\varphi_k(x) := \sum_{i=0}^{k} \lambda_i l_f(x_i;x) + \beta_k d(x) \tag{18}$$

  for $k \geq -1$. Then, Property A follows from Theorem 3.1 with the update (15).

- *Hybrid model.* Define $\{(\varphi_k(x), \psi_k(x))\}$ by $\psi_{-1}(x) := \beta_{-1}d(x)$ and

$$\begin{array}{rcl} \varphi_k(x) & := & \sum_{i=0}^{k} \lambda_i l_f(x_i;x) + \beta_k d(x), \\ \psi_{k+1}(x) & := & \min_{z\in Q} \varphi_k(z) + \lambda_{k+1}l_f(x_{k+1};x) + \beta_{k+1}d(x) - \beta_k l_d(z_k;x) + S_k\sigma_f\xi(z_k,x) \end{array} \tag{19}$$

  for $k \geq -1$. Then, Property B follows from Theorem 3.2 with the update (16).

Now let us complete the proof of Theorem 3.1.

Lemma 3.3. *Let $\{\varphi_k(x)\}_{k\geq -1}$ be generated accordingly to the construction in Theorem 3.1. Then, for every $k \geq -1$, we have*

$$\varphi_k(x) \geq \varphi_k(z_k) + (\beta_k + S_k\sigma_f)\xi(z_k,x), \quad \forall x \in Q,\ \forall k \geq -1.$$

Proof. Since $\sigma_f \in \sigma(l_f(x_i,\cdot))$ for $i \geq 0$, we can see inductively that $\beta_k + S_k\sigma_f \in \sigma(\varphi_k)$ for all $k \geq -1$. Therefore, using its characterization (2), the optimality condition $\varphi_k'(z_k; x - z_k) \geq 0, \forall x \in Q$ for the minimizer $z_k = \operatorname{argmin}_{x\in Q}\varphi_k(x)$ yields the conclusion. □

Lemma 3.4. *Any sequence $\{\varphi_k(x)\}_{k\geq -1}$ generated accordingly to the construction in Theorem 3.1 satisfies the condition (A2).*

Proof. Notice that the construction (14) satisfies (A2) as equality. In the case of the construction (15), Lemma 3.3 yields for any $x \in Q$ that

$$\begin{array}{rcl} \varphi_{k+1}(x) & = & \varphi_k(x) + \lambda_{k+1}l_f(x_{k+1};x) + \beta_{k+1}d(x) - \beta_k d(x) \\ & \geq & [\varphi_k(z_k) + (\beta_k + S_k\sigma_f)\xi(z_k,x)] + \lambda_{k+1}l_f(x_{k+1};x) + \beta_{k+1}d(x) - \beta_k d(x) \\ & = & \varphi_k(z_k) + \lambda_{k+1}l_f(x_{k+1};x) + \beta_{k+1}d(x) - \beta_k l_d(z_k;x) + S_k\sigma_f\xi(z_k,x) \end{array}$$

which is the condition (A2) for $k \geq -1$. □

Lemma 3.5. *Let $\{\varphi_k(x)\}_{k\geq -1}$ be generated accordingly to the construction in Theorem 3.1. Then, the condition (A3) holds.*

Proof. We prove the assertion by induction. Since $z_{-1} = x_0 = \operatorname{argmin}_{x\in Q} d(x)$, we have $\min_{x\in Q} l_d(z_{-1};x) = \min_{x\in Q} d(x) = 0$ which proves (A3) for $k = -1$. Assume that (A3) holds up to $k \geq -1$. In the case when all $\{\varphi_i(x)\}_{i=0}^{k+1}$ are constructed by (15), it coincides with the formula (18). Therefore, Lemma 3.3 implies that

$$\begin{aligned}\varphi_k(z_k) \leq \varphi_k(x) - (\beta_k + S_k\sigma_f)\xi(z_k, x) &= \sum_{i=0}^{k} \lambda_i l_f(x_i; x) + \beta_k d(x) - (\beta_k + S_k\sigma_f)\xi(z_k, x) \\ &= \sum_{i=0}^{k} \lambda_i l_f(x_i; x) + \beta_k l_d(z_k; x) - S_k\sigma_f\xi(z_k, x)\end{aligned}$$

for every $x \in Q$, from which the condition (A3) follows. If this is not the case, there exists some integer $j \leq k$ such that $\varphi_{k+1}(x)$ is constructed as defining $\varphi_{j+1}(x)$ by (14) and $\varphi_{j+2}(x), \ldots, \varphi_{k+1}(x)$ by (15). Then, we have

$$\varphi_{k+1}(x) = \min_{z\in Q} \varphi_j(z) + \sum_{i=j+1}^{k+1} \lambda_i l_f(x_i; x) + \beta_{k+1} d(x) - \beta_j l_d(z_j; x) + S_j\sigma_f\xi(z_j, x)$$

which yields $\varphi_{k+1}(x) \leq \sum_{i=0}^{k+1} \lambda_i l_f(x_i; x) + \beta_{k+1} d(x)$ by the induction hypothesis (A3) for $\varphi_j(x)$. Therefore, Lemma 3.3 implies for every $x \in Q$ that

$$\begin{aligned}\varphi_{k+1}(z_{k+1}) &\leq \varphi_{k+1}(x) - (\beta_{k+1} + S_{k+1}\sigma_f)\xi(z_{k+1}, x) \\ &\leq \sum_{i=0}^{k+1} \lambda_i l_f(x_i; x) + \beta_{k+1} d(x) - (\beta_{k+1} + S_{k+1}\sigma_f)\xi(z_{k+1}, x) \\ &= \sum_{i=0}^{k+1} \lambda_i l_f(x_i; x) + \beta_{k+1} l_d(z_{k+1}; x) - S_{k+1}\sigma_f\xi(z_{k+1}, x)\end{aligned}$$

which gives the condition (A3) for $\varphi_{k+1}(x)$. □

4 General (sub)gradient-based methods and convergence properties

In this section we propose general (sub)gradient-based methods for smooth and structured problems introduced in Section 2.1. The main strategy is to develop update formulas for test points $\{x_k\}_{k\geq 0}$ and approximate solutions $\{\hat{x}_k\}_{k\geq 0}$ which satisfy the following relation

$$(R_k) \quad S_k f(\hat{x}_k) \leq \psi_k(w_k) + C_k$$

for every $k \geq 0$, where $\{(\varphi_k(x), \psi_k(x))\}_{k\geq -1}$ is a coupled sequence of auxiliary functions satisfying Property B. Furthermore, the relations

$$(P_k) \quad \sum_{i=0}^{k} \lambda_i f(x_i) \leq \psi_k(w_k) + C_k \quad \text{and} \quad (Q_k) \quad \sum_{i=0}^{k} \lambda_i f(w_i) \leq \psi_k(w_k) + C_k$$

for non-smooth and structured problems, respectively, are also useful to provide stronger results. These concepts yield the following convergence rate for subgradient methods.

Lemma 4.1. *Suppose that a sequence $\{\hat{x}_k\}_{k\geq 0} \subset Q$ satisfies the relation (R_k) for a coupled sequence $\{(\varphi_k(x), \psi_k(x))\}_{k\geq -1}$ of auxiliary functions associated with weight parameters $\{\lambda_k\}_{k\geq 0}$, scaling parameters $\{\beta_k\}_{k\geq -1}$, and test points $\{x_k\}_{k\geq 0}$. If the condition (B3) in Property B holds with a convexity parameter $\sigma_f \geq 0$, then we have*

$$f(\hat{x}_k) - f(x) + \sigma_f\xi(z_k, x) \leq \frac{\beta_k l_d(z_k; x) + C_k}{S_k}, \quad \forall x \in Q. \tag{20}$$

Proof. The assertion follows from the condition (B3) and the relation (R_k); for any $x \in Q$, we have

$$S_k f(\hat{x}_k) \le \sum_{i=0}^{k} \lambda_i l_f(x_i; x) + \beta_k l_d(z_k; x) - S_k \sigma_f \xi(z_k, x) + C_k \le S_k f(x) + \beta_k l_d(z_k; x) - S_k \sigma_f \xi(z_k, x) + C_k.$$

□

Remark 4.2. (1) Analogues of Lemma 4.1 easily show that (P_k) and (B3) imply the inequality

$$\min_{0 \le i \le k} f(x_i) - f(x) + \sigma_f \xi(z_k, x) \le \frac{1}{S_k} \sum_{i=0}^{k} \lambda_i f(x_i) - f(x) + \sigma_f \xi(z_k, x) \le \frac{\beta_k l_d(z_k; x) + C_k}{S_k}$$

for $x \in Q$. The conditions (Q_k) and (B3) also conclude the same replacing x_i by w_i.
(2) When $\sigma_f > 0$, (20) provides bounds for the distances to x^* from $\hat{x}_k$ and z_k: According to the facts $f(x) - f(x^*) \ge \sigma_f \xi(x^*, x)$ and $\xi(x, y) \ge \frac{\sigma_d}{2} \|x - y\|^2$ for $x, y \in Q$, the bound (20) implies

$$\min\{\|\hat{x}_k - x^*\|^2, \|z_k - x^*\|^2\} \le \frac{1}{2}\|\hat{x}_k - x^*\|^2 + \frac{1}{2}\|z_k - x^*\|^2 \le \frac{\beta_k l_d(z_k; x^*) + C_k}{\sigma_f \sigma_d S_k}.$$

The general (sub)gradient-based methods for non-smooth and structured problems will be presented in Section 4.4 using the classical and modified updates. We prepare some basic lemmas to gain some insights for the updates before presenting the main results.

4.1 Update formula for the auxiliary functions when $k = 0$

Here, we develop update formulas of sequences $\{x_k\}_{k \ge -1}$ and $\{\hat{x}_k\}_{k \ge -1}$ which admit (R_k). We consider the case $k = 0$ at first.

Lemma 4.3. *Let $\{(\varphi_k(x), \psi_k(x))\}_{k \ge -1}$ be a coupled sequence of auxiliary functions satisfying Property B.*

(i) In the non-smooth case (4), if $\hat{x}_0 = x_0$ holds, then the relation (R_0) is satisfied with

$$C_0 := \frac{1}{2} \frac{\lambda_0^2}{\sigma_d(\lambda_0 \sigma_f + \beta_{-1})} \|g_0\|_*^2. \tag{21}$$

(ii) In the structured case (6), if $\hat{x}_0 = w_0$ holds, then the relation $(R_0) \equiv (P_0)$ is satisfied with

$$C_0 := \lambda_0 \left(\frac{L(x_0)}{2} - \frac{\sigma_d}{2} \left(\bar{\sigma}_f + \frac{\beta_{-1}}{\lambda_0} \right) \right) \|w_0 - x_0\|^2 + \lambda_0 \delta(x_0, \hat{x}_0). \tag{22}$$

Proof. Note that (B0) implies that $\varphi_k(z_k) = \min_{x \in Q} \varphi_k(x) \ge \min_{x \in Q} \psi_k(x) = \psi_k(w_k)$. Since $\{\beta_k\}$ is non-decreasing, using (B2) with $x = w_{k+1}$ yields that

$$\begin{aligned} \psi_{k+1}(w_{k+1}) &\ge \varphi_k(z_k) + \lambda_{k+1} l_f(x_{k+1}; w_{k+1}) + (\beta_k + S_k \sigma_f) \xi(z_k, w_{k+1}) \\ &\ge \psi_k(w_k) + \lambda_{k+1} l_f(x_{k+1}; w_{k+1}) + (\beta_k + S_k \sigma_f) \xi(z_k, w_{k+1}) \end{aligned} \tag{23}$$

for every $k \ge -1$. In the case $k = -1$, the conditions (B1), $S_{-1} = 0$, and $z_{-1} = x_0$ lead (23) to

$$\begin{aligned} \psi_0(w_0) &\ge \lambda_0 [l_f(x_0; w_0) - \sigma \xi(x_0, w_0) + (\sigma + \beta_{-1}/\lambda_0) \xi(x_0, w_0)] \\ &\ge \lambda_0 \left[l_f(x_0; w_0) - \sigma \xi(x_0, w_0) + \frac{\sigma_d}{2} \left(\sigma + \frac{\beta_{-1}}{\lambda_0} \right) \|w_0 - x_0\|^2 \right]. \end{aligned} \tag{24}$$

for any $\sigma \geq 0$. Let us firstly show (ii). Letting $\sigma := \bar{\sigma}_f$, the settings $\hat{x}_0 = w_0$ and (22) yields

$$\psi_0(w_0) + C_0 \overset{(24)}{\geq} \lambda_0 \left[l_f(x_0; w_0) - \bar{\sigma}_f \xi(x_0, \hat{x}_0) + \frac{L(x_0)}{2} \|\hat{x}_0 - x_0\|^2 + \delta(x_0, \hat{x}_0) \right] \geq \lambda_0 f(\hat{x}_0)$$

which proves the relation (R_0).

It reminds to prove (i). By the definition of $l_f(\cdot; \cdot)$ for the non-smooth case, the inequality (24) with $\sigma := \sigma_f$ implies

$$\begin{aligned} \psi_0(w_0) &\overset{(24)}{\geq} \lambda_0 \left[f(x_0) + \langle g_0, w_0 - x_0 \rangle + \frac{\sigma_d}{2} \left(\sigma_f + \frac{\beta_{-1}}{\lambda_0} \right) \|w_0 - x_0\|^2 \right] \\ &= \lambda_0 f(x_0) + \langle \lambda_0 g_0, w_0 - x_0 \rangle + \frac{\sigma_d}{2} \left(\lambda_0 \sigma_f + \beta_{-1} \right) \|w_0 - x_0\|^2 \\ &\geq \lambda_0 f(x_0) - \frac{1}{2} \frac{\lambda_0^2}{\sigma_d (\lambda_0 \sigma_f + \beta_{-1})} \|g_0\|_*^2, \end{aligned}$$

where the last inequality is due to the basic fact

$$\frac{1}{2}\|x\|^2 + \frac{1}{2}\|s\|_*^2 \geq \langle s, x \rangle \text{ for } x \in E, \ s \in E^*. \tag{25}$$

This means that the relation (R_0) is satisfied with the setting $\hat{x}_0 = x_0$ and (21). □

4.2 Classical update formulas for the auxiliary functions when $k > 0$

Now we develop the update formulas which yield classical updates in (sub)gradient-based methods such as the MDM and the DAM.

Lemma 4.4. *Let $\{(\varphi_k(x), \psi_k(x))\}_{k \geq -1}$ be a coupled sequence of auxiliary functions satisfying Property B. Suppose that the relation (R_k) is satisfied for some $k \geq 0$. Then, the following assertions hold.*

(i) In the non-smooth case (4), if the relations $x_{k+1} = z_k$ and $\hat{x}_{k+1} = \frac{S_k \hat{x}_k + \lambda_{k+1} x_{k+1}}{S_{k+1}}$ hold, then the relation (R_{k+1}) is satisfied with

$$C_{k+1} := C_k + \frac{1}{2\sigma_d} \frac{\lambda_{k+1}^2}{\beta_k + S_{k+1}\sigma_f} \|g_{k+1}\|_*^2. \tag{26}$$

Furthermore, if (P_k) is satisfied, then so is (P_{k+1}) with the same settings of x_{k+1} and C_{k+1}.

(ii) In the structured case (6), if the relations $x_{k+1} = z_k$ and $\hat{x}_{k+1} = \frac{S_k \hat{x}_k + \lambda_{k+1} w_{k+1}}{S_{k+1}}$ hold, then the relation (R_{k+1}) is satisfied with

$$C_{k+1} := C_k + \lambda_{k+1} \left(\frac{L(x_{k+1})}{2} - \frac{\sigma_d}{2} \left(\bar{\sigma}_f + \frac{\beta_k + S_k \sigma_f}{\lambda_{k+1}} \right) \right) \|w_{k+1} - x_{k+1}\|^2 + \lambda_{k+1} \delta(x_{k+1}, w_{k+1}).$$

Furthermore, if (Q_k) is satisfied, then so is (Q_{k+1}) with the same settings of x_{k+1} and C_{k+1}.

Proof. Using (23) and the relation $x_{k+1} = z_k$ imply for any $\sigma \geq 0$ that

$$\begin{aligned} \psi_{k+1}(w_{k+1}) &\geq \psi_k(w_k) + \lambda_{k+1} l_f(x_{k+1}; w_{k+1}) + (\beta_k + S_k \sigma_f) \xi(z_k, w_{k+1}) \\ &= \psi_k(w_k) \\ &\quad + \lambda_{k+1} \left([l_f(x_{k+1}; w_{k+1}) - \sigma \xi(x_{k+1}, w_{k+1})] + \left(\sigma + \frac{\beta_k + S_k \sigma_f}{\lambda_{k+1}} \right) \xi(x_{k+1}, w_{k+1}) \right) \\ &\geq \psi_k(w_k) \\ &\quad + \lambda_{k+1} \left([l_f(x_{k+1}; w_{k+1}) - \sigma \xi(x_{k+1}, w_{k+1})] + \frac{\sigma_d}{2} \left(\sigma + \frac{\beta_k + S_k \sigma_f}{\lambda_{k+1}} \right) \|w_{k+1} - x_{k+1}\|^2 \right). \end{aligned}$$

For the structured problems, letting $\sigma := \bar{\sigma}_f$ and the definition of C_{k+1} in (ii) yield that

$$\psi_{k+1}(w_{k+1}) + C_{k+1} \geq \psi_k(w_k) + C_k + \lambda_{k+1} f(w_{k+1}).$$

Using (R_k) and the convexity of f conclude the relation (R_{k+1}); (Q_{k+1}) follows by using (Q_k) and the inequality above. Hence, the assertion (ii) is proved.

For the non-smooth problems, on the other hand, we can continue by taking $\sigma := \sigma_f$ as follows.

$$\begin{array}{rcl} \psi_{k+1}(w_{k+1}) & \geq & \psi_k(w_k) + \lambda_{k+1} f(x_{k+1}) + \langle \lambda_{k+1} g_{k+1}, w_{k+1} - x_{k+1} \rangle + \frac{\sigma_d}{2}(\beta_k + S_{k+1}\sigma_f)\|w_{k+1} - x_{k+1}\|^2 \\ & \overset{(25)}{\geq} & \psi_k(w_k) + \lambda_{k+1} f(x_{k+1}) - \frac{1}{2} \frac{\lambda_{k+1}^2}{\sigma_d(\beta_k + S_{k+1}\sigma_f)} \|g_{k+1}\|_*^2. \end{array}$$

Hence, the definition (26) of C_{k+1} yields that

$$\psi_{k+1}(w_{k+1}) + C_{k+1} \geq \psi_k(w_k) + C_k + \lambda_{k+1} f(x_{k+1}).$$

Now the assertion (i) follows by the same way as (ii). □

4.3 Modified update formulas for the auxiliary functions when $k > 0$

The modified update formulas described below yields accelerated gradient-based methods for structured problems as Nesterov's and Tseng's methods.

Lemma 4.5. *Let $\{(\varphi_k(x), \psi_k(x))\}_{k \geq -1}$ be a coupled sequence of auxiliary functions satisfying Property B. Suppose that the relation (R_k) is satisfied for some $k \geq 0$. Then, the following assertions hold.*

(i) In the non-smooth case (4), if the relation $\hat{x}_{k+1} = x_{k+1} = \frac{S_k \hat{x}_k + \lambda_{k+1} z_k}{S_{k+1}}$ holds, then the relation (R_{k+1}) is satisfied with

$$C_{k+1} := C_k + \frac{1}{2\sigma_d} \frac{\lambda_{k+1}^2 S_{k+1}}{\lambda_{k+1}^2 \sigma_f + S_{k+1}(\beta_k + S_k \sigma_f)} \|g_{k+1}\|_*^2. \tag{27}$$

(ii) In the structured case (6), if the relations $x_{k+1} = \frac{S_k \hat{x}_k + \lambda_{k+1} z_k}{S_{k+1}}$ and $\hat{x}_{k+1} = \frac{S_k \hat{x}_k + \lambda_{k+1} w_{k+1}}{S_{k+1}}$ hold, then the relation (R_{k+1}) is satisfied with

$$C_{k+1} := C_k + S_{k+1} \left(\frac{L(x_{k+1})}{2} - \frac{\sigma_d}{2} \left(\bar{\sigma}_f + \frac{S_{k+1}(\beta_k + S_k \sigma_f)}{\lambda_{k+1}^2} \right) \right) \|\hat{x}_{k+1} - x_{k+1}\|^2 + S_{k+1} \delta(x_{k+1}, \hat{x}_{k+1}). \tag{28}$$

Proof. Denote $x'_{k+1} := \frac{S_k \hat{x}_k + \lambda_{k+1} w_{k+1}}{S_{k+1}}$. If $x_{k+1} = \frac{S_k \hat{x}_k + \lambda_{k+1} z_k}{S_{k+1}}$ holds, then $x'_{k+1} - x_{k+1} = \frac{\lambda_{k+1}}{S_{k+1}}(w_{k+1} - z_k)$. Using (23) and the relation (R_k), we have

$$\begin{array}{rcl} \psi_{k+1}(w_{k+1}) + C_k & \geq & \psi_k(w_k) + C_k + \lambda_{k+1} l_f(x_{k+1}; w_{k+1}) + (\beta_k + S_k \sigma_f)\xi(z_k, w_{k+1}) \\ & \geq & S_k f(\hat{x}_k) + \lambda_{k+1} l_f(x_{k+1}; w_{k+1}) + (\beta_k + S_k \sigma_f)\xi(z_k, w_{k+1}) \\ & \geq & S_k l_f(x_{k+1}; \hat{x}_k) + \lambda_{k+1} l_f(x_{k+1}; w_{k+1}) + (\beta_k + S_k \sigma_f)\xi(z_k, w_{k+1}) \\ & \geq & S_{k+1} l_f(x_{k+1}; x'_{k+1}) + (\beta_k + S_k \sigma_f)\xi(z_k, w_{k+1}), \end{array} \tag{29}$$

where we used $f(x) \geq l_f(y; x), \forall x, y \in Q$ and the convexity of $l_f(x_{k+1}; \cdot)$ for the last two inequalities. Since $\xi(z_k, w_{k+1}) \geq \frac{\sigma_d}{2}\|w_{k+1} - z_k\|^2 = \frac{\sigma_d}{2} \frac{S_{k+1}^2}{\lambda_{k+1}^2} \|x'_{k+1} - x_{k+1}\|^2$ and

$$\begin{array}{rcl} l_f(x_{k+1}; x'_{k+1}) & = & l_f(x_{k+1}; x'_{k+1}) - \sigma\xi(x_{k+1}, x'_{k+1}) + \sigma\xi(x_{k+1}, x'_{k+1}) \\ & \geq & l_f(x_{k+1}; x'_{k+1}) - \sigma\xi(x_{k+1}, x'_{k+1}) + \frac{\sigma\sigma_d}{2}\|x_{k+1} - x'_{k+1}\|^2 \end{array}$$

hold for any $\sigma \geq 0$, the inequality (29) implies that

$$\begin{aligned}\psi_{k+1}(w_{k+1}) + C_k \quad \geq \quad & S_{k+1}[l_f(x_{k+1}; x'_{k+1}) - \sigma\xi(x_{k+1}, x'_{k+1})] \\ & + \frac{\sigma_d}{2} S_{k+1} \left(\sigma + \frac{S_{k+1}(\beta_k + S_k\sigma_f)}{\lambda_{k+1}^2} \right) \|x'_{k+1} - x_{k+1}\|^2. \end{aligned} \quad (30)$$

Let us prove (ii) at first. Since $\hat{x}_{k+1} = x'_{k+1}$ by the assumption, adding

$$S_{k+1} \left(\frac{L(x_{k+1})}{2} - \frac{\sigma_d}{2} \left(\bar{\sigma}_f + \frac{S_{k+1}(\beta_k + S_k\sigma_f)}{\lambda_{k+1}^2} \right) \right) \|\hat{x}_{k+1} - x_{k+1}\|^2 + S_{k+1}\delta(x_{k+1}, \hat{x}_{k+1})$$

to both sides in (30) with $\sigma := \bar{\sigma}_f$ and using the inequality (6) implies the relation (R_{k+1}) with the setting (28).

To prove (i), on the other hand, letting $\sigma := \sigma_f$ and using $l_f(x_{k+1}; x'_{k+1}) - \sigma\xi(x_{k+1}, x'_{k+1}) = f(x_{k+1}) + \langle g_{k+1}, x'_{k+1} - x_{k+1} \rangle$ leads (30) to

$$\begin{aligned}\psi_{k+1}(w_{k+1}) + C_k \quad & \geq \quad S_{k+1} f(x_{k+1}) + \langle S_{k+1} g_{k+1}, x'_{k+1} - x_{k+1} \rangle \\ & \qquad + \frac{\sigma_d}{2} S_{k+1} \left(\sigma_f + \frac{S_{k+1}(\beta_k + S_k\sigma_f)}{\lambda_{k+1}^2} \right) \|x'_{k+1} - x_{k+1}\|^2 \\ & \overset{(25)}{\geq} \quad S_{k+1} f(x_{k+1}) - \frac{1}{2} \frac{S_{k+1}^2}{\sigma_d S_{k+1} \left(\sigma_f + \frac{S_{k+1}(\beta_k + S_k\sigma_f)}{\lambda_{k+1}^2} \right)} \|g_{k+1}\|_*^2 \\ & = \quad S_{k+1} f(x_{k+1}) - \frac{1}{2\sigma_d} \frac{\lambda_{k+1}^2 S_{k+1}}{\lambda_{k+1}^2 \sigma_f + S_{k+1}(\beta_k + S_k\sigma_f)} \|g_{k+1}\|_*^2. \end{aligned}$$

This means that the relation (R_{k+1}) is obtained with (27). □

4.4 General (sub)gradient-based methods

As a consequence of the previous lemmas, we propose the following unifying framework of (sub)gradient-based methods. We propose two types of updates, the classical and the modified ones, and we will analyze their rate of convergence later.

Method 4.6 (Unifying framework of subgradient-based methods for non-smooth problems)**.** *Consider the non-smooth problem (4). Let $\{\lambda_k\}_{k\geq 0}$ and $\{\beta_k\}_{k\geq -1}$ be sequences of weight and scaling parameters, respectively. Let $\sigma_f \in \sigma(f)$. Generate a sequence $\{(z_{k-1}, x_k, g_k, \hat{x}_k)\}_{k\geq 0}$ by either the classical or the modified method as follows.*

(0) Set $\hat{x}_0 := x_0 := z_{-1} := \operatorname{argmin}_{x\in Q} d(x)$.

(1) (k-th iteration, $k \geq 0$) Set $g_k \in \partial f(x_k)$ and compute $z_k, x_{k+1}, \hat{x}_{k+1}$ by

$$\textit{Classical method} \;:\; x_{k+1} := z_k := \operatorname*{argmin}_{x\in Q} \varphi_k(x), \quad \hat{x}_{k+1} := \frac{S_k\hat{x}_k + \lambda_{k+1}z_k}{S_{k+1}},$$

or

$$\textit{Modified method} \;:\; z_k := \operatorname*{argmin}_{x\in Q} \varphi_k(x), \quad \hat{x}_{k+1} := x_{k+1} := \frac{S_k\hat{x}_k + \lambda_{k+1}z_k}{S_{k+1}}.$$

Here, the single sequence $\{\varphi_k(x)\}_{k\geq -1}$ of auxiliary functions is defined by the construction either (17) or (18) with $l_f(x_k; x) := f(x_k) + \langle g_k, x - x_k \rangle + \sigma_f\xi(x_k, x)$, as well as any construction satisfying Property A.

Note that we did not use a coupled sequence $\{(\varphi_k(x), \psi_k(x))\}_{k\geq -1}$ of auxiliary functions because the statements of Lemmas 4.3, 4.4, 4.5 for the non-smooth case are independent of the second object $\{\psi_k(x)\}_{k\geq -1}$ (or w_k). Using the models (17) and (18) enable us to solve the subproblem $z_k := \operatorname{argmin}_{x\in Q} \varphi_k(x)$ since it has the form (5).

Method 4.7 (Unifying framework of gradient-based methods for structured problems)**.** *Consider the structured problem (6). Let $\{\lambda_k\}_{k\geq 0}$ and $\{\beta_k\}_{k\geq -1}$ be sequences of weight and scaling parameters, respectively. Let $\sigma_f \geq 0$ be a convexity parameter satisfying (3). Generate a sequence $\{(z_{k-1}, w_{k-1}, x_k, \hat{x}_k)\}_{k\geq 0}$ by either the classical or the modified method as follows.*

(0) Set $x_0 := z_{-1} := w_{-1} := \operatorname{argmin}_{x\in Q} d(x)$. Compute

$$z_0 := \operatorname*{argmin}_{x\in Q} \varphi_0(x), \quad \hat{x}_0 := w_0 := \operatorname*{argmin}_{x\in Q} \psi_0(x).$$

(1) (k-th iteration, $k \geq 0$) Set

$$\begin{aligned} x_{k+1} &:= \begin{cases} z_k & : \textit{Classical method,} \\ \dfrac{S_k\hat{x}_k + \lambda_{k+1}z_k}{S_{k+1}} & : \textit{Modified method,} \end{cases} \\ z_{k+1} &:= \operatorname*{argmin}_{x\in Q} \varphi_{k+1}(x), \\ w_{k+1} &:= \operatorname*{argmin}_{x\in Q} \psi_{k+1}(x), \\ \hat{x}_{k+1} &:= \frac{S_k\hat{x}_k + \lambda_{k+1}w_{k+1}}{S_{k+1}}. \end{aligned}$$

Here, the coupled sequence $\{(\varphi_k(x), \psi_k(x))\}_{k\geq -1}$ of auxiliary functions is defined by the construction either (17),(18), or (19), as well as any construction satisfying Property B.

Note that in general the classical and the modified methods will provide different efficiency estimates. They yield the same convergence rate for non-smooth problems but the modified method gives much better efficiency than the classical method for smooth problems as discussed in Section 5.

Method 4.6 includes four particularizations; we can choose the classical or the modified updates combined to the choice of the auxiliary functions by the extended MD model (17) or by the DA model (18). Method 4.7 yields six particularizations with the additional choice of the hybrid model (19). Employing the models (17) or (18) in Method 4.7, it reduces the number of subproblem at each iterations since $z_k \equiv w_k$. Note that only the extended MD model (17) turns the subproblem $z_k := \operatorname{argmin}_{x\in Q} \varphi_k(x)$ of the form (7); the others require the solution of the subproblem (10). However, the subproblems with these models have the same difficulty for all the examples cited in Section 2.3.

We conclude this section with general estimates for Methods 4.6 and 4.7 which will be further particularized.

Theorem 4.8. *Let $\{(z_{k-1}, x_k, g_k, \hat{x}_k)\}_{k\geq 0}$ be generated by Method 4.6 for the non-smooth problem (4) associated with weight parameters $\{\lambda_k\}_{k\geq 0}$, scaling parameters $\{\beta_k\}_{k\geq -1}$, and a convexity parameter $\sigma_f \in \sigma(f)$. Then, for every $k \geq 0$, the relation (R_k) holds with*

$$C_k := \begin{cases} \frac{1}{2\sigma_d} \sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_{i-1}+S_i\sigma_f} \|g_i\|_*^2 & : \textit{Classical method,} \\ \frac{1}{2\sigma_d} \sum_{i=0}^{k} \frac{\lambda_i^2 S_i}{\lambda_i^2\sigma_f + S_i(\beta_{i-1}+S_{i-1}\sigma_f)} \|g_i\|_*^2 & : \textit{Modified method.} \end{cases} \tag{31}$$

Therefore, the estimate

$$f(\hat{x}_k) - f(x^*) + \sigma_f \xi(z_k, x^*) \leq \frac{\beta_k l_d(z_k; x^*) + C_k}{S_k}$$

holds for every $k \geq 0$. Furthermore, for every $k \geq 0$, the classical method satisfies the relation (P_k) and therefore the above estimate holds even replacing the left hand side by $\frac{1}{S_k}\sum_{i=0}^k \lambda_i f(x_i) - f(x^*) + \sigma_f \xi(z_k, x^*)$ or by $\min_{0 \leq i \leq k} f(x_i) - f(x^*) + \sigma_f \xi(z_k, x^*)$.

Proof. By the description of Method 4.6, we can apply the part (i) of each Lemmas 4.3,4.4,4.5 to show that the relation (R_k) holds for every $k \geq 0$ with C_k defined by (31); for the classical method, the relation (P_k) can also be verified. The assertion follows from Lemma 4.1 and its analogue for the relation (P_k). □

Using the part (ii) of Lemmas 4.3,4.4,4.5 as analogous to the proof the above theorem, we arrive to a similar result for Method 4.7.

Theorem 4.9. *Let $\{(z_{k-1}, w_{k-1}, x_k, \hat{x}_k)\}_{k \geq 0}$ be generated by Method 4.7 for the structured problem (6) associated with weight parameters $\{\lambda_k\}_{k \geq 0}$, scaling parameters $\{\beta_k\}_{k \geq -1}$, and a convexity parameter $\sigma_f \geq 0$ satisfying (3). Then, for every $k \geq 0$, the relation (R_k) holds with*

$$C_k := \begin{cases} \frac{1}{2}\sum_{i=0}^k \lambda_i \left(L(x_i) - \sigma_d \left(\bar{\sigma}_f + \frac{\beta_{i-1} + S_{i-1}\sigma_f}{\lambda_i}\right)\right) \|w_i - x_i\|^2 + \sum_{i=0}^k \lambda_i \delta(x_i, w_i) \\ \qquad \text{for the classical method; and} \\ \frac{1}{2}\sum_{i=0}^k S_i \left(L(x_i) - \sigma_d \left(\bar{\sigma}_f + \frac{S_i(\beta_{i-1} + S_{i-1}\sigma_f)}{\lambda_i^2}\right)\right) \|\hat{x}_i - x_i\|^2 + \sum_{i=0}^k S_i \delta(x_i, \hat{x}_i) \\ \qquad \text{for the modified method.} \end{cases} \tag{32}$$

Therefore, the estimate

$$f(\hat{x}_k) - f(x^*) + \sigma_f \xi(z_k, x^*) \leq \frac{\beta_k l_d(z_k; x^*) + C_k}{S_k} \tag{33}$$

holds for every $k \geq 0$. Furthermore, for every $k \geq 0$, the classical method satisfies the relation (Q_k) and therefore the above estimate holds even replacing the left hand side by $\frac{1}{S_k}\sum_{i=0}^k \lambda_i f(w_i) - f(x^*) + \sigma_f \xi(z_k, x^*)$ or by $\min_{0 \leq i \leq k} f(w_i) - f(x^*) + \sigma_f \xi(z_k, x^*)$.

Remark 4.10. Method 4.7 with $\sigma_f = \bar{\sigma}_f = 0$ and $\beta_k \equiv 0$ yields several versions of CGM because the constructed auxiliary functions are non-negative linear combinations of constants and $\{l_f(x_i; x)\}_{i=0}^k$. In this case, Theorem 4.9 implies that the modified method ensures

$$f(\hat{x}_k) - f(x^*) \leq \frac{C_k}{S_k} \leq \frac{\frac{1}{2}\mathrm{Diam}(Q)^2 \sum_{i=0}^k L(x_i)\frac{\lambda_i^2}{S_i}}{S_k} + \frac{\sum_{i=0}^k S_i \delta(x_i, \hat{x}_i)}{S_k} \tag{34}$$

for all $k \geq 0$, because $\|\hat{x}_i - x_i\|^2 = \frac{\lambda_i^2}{S_i^2}\|w_i - z_{i-1}\|^2 \leq \frac{\lambda_i^2}{S_i^2}\mathrm{Diam(Q)}^2$. Note that, if $l_f(y; \cdot)$ is affine for each $y \in Q$, then the classical CGM (13) with $\tau_k := \lambda_{k+1}/S_{k+1}$ and $\hat{x}_k := x_k$ also admits a similar estimate[4]

$$f(x_k) - f(x^*) \leq \frac{\lambda_0[f(x_0) - l_f(x_0; z_0)]}{S_k} + \frac{\frac{1}{2}\mathrm{Diam}(Q)^2 \sum_{i=1}^k L(x_{i-1})\frac{\lambda_i^2}{S_i}}{S_k} + \frac{\sum_{i=1}^k S_i \delta(x_{i-1}, x_i)}{S_k}. \tag{35}$$

[4] The proof of [16, Theorem 5.3] replacing the notation $(h(\cdot), \lambda_{k+1}, \tilde{\lambda}_{k+1}, L_{k+1}, \delta_{k+1}, \tilde{\alpha}_{k+1}, \beta_{k+1}, \alpha_k)$ of [16] by $(-f(\cdot), x_k, z_k, L(x_k), \delta(x_k, x_{k+1}), \tau_k, S_k/\lambda_0, \lambda_k/\lambda_0)$ for $k \geq 0$ shows the desired estimate because showing the result uses the assumption [16, eq.(52)] with $(L, \delta) = (L_{k+1}, \delta_{k+1})$ only at $(\lambda, \bar{\lambda}) = (\lambda_{k+2}, \lambda_{k+1})$, which corresponds to our assumption (6) at $(x, y) = (x_k, x_{k+1})$.

5 Convergence analysis of subgradient-based methods

In this section, we finally obtain the actual convergence rates for Methods 4.6 and 4.7 for particular classes of convex problems based on the general estimates presented in Section 4.4, and compare these results with existing ones. Our choices for weight $\{\lambda_k\}_{k\geq 0}$ and scaling parameters $\{\beta_k\}_{k\geq -1}$ resemble and extend the existing ones to produce approximate solutions $\{\hat{x}_k\}_{k\geq 0}$ which yields a nice convergence for $f(\hat{x}_k) - f(x^*)$. We show convergence properties of PGMs for the non-smooth problems in the next subsection, for the structured problems with inexact oracle in Sections 5.2, 5.3, and for the weakly smooth problems in the last subsection, for the strongly convex case. Optimal and nearly optimal convergences of CGMs are developed in Sections 5.3 and 5.4.4. All of convergence rates matches the known optimal rates of convergence or give a slight improvement of them with an advantage of obtaining them with a unified analysis.

5.1 Efficiency for non-smooth problems

Method 4.6 generates a sequence $\{\hat{x}_k\}$ which satisfies the relation (R_k) with C_k defined by (31).

When $\sigma_f = 0$, the definitions of C_k for the classical method and modified methods become the same: $C_k = \frac{1}{2\sigma_d}\sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2$; this case is analyzed in [22, Corollary 11] which ensures the optimal complexity $O(M^2 d(x^*)/(\sigma_d \varepsilon^2))$ with an advantage for the choice of the parameters $\{\lambda_k\}$ and $\{\beta_k\}$ to ensure the optimal convergence rate.

When $\sigma_f > 0$, note that

$$\frac{\lambda_i^2 S_i}{\lambda_i^2 \sigma_f + S_i(\beta_{i-1} + S_{i-1}\sigma_f)} = \frac{\lambda_i^2}{\beta_{i-1} + S_{i-1}\sigma_f + \frac{\lambda_i^2}{S_i}\sigma_f} \geq \frac{\lambda_i^2}{\beta_{i-1} + S_i \sigma_f}$$

holds since $\lambda_i / S_i \leq 1$. In this case, theoretically, the modified method ensures not a worse convergence rate than the classical counterpart. We give an optimal convergence result with a simple choice for the parameters $\lambda_k = (k+1)/2$ and $\beta_k \equiv 0$ below. Note that every subproblem $\min_{x\in Q} \varphi_k(x)$ has a unique solution even if $\beta_k \equiv 0$ because $\sigma(\varphi_k) \ni \beta_k + S_k\sigma_f = S_k \sigma_f > 0$ (see the proof of Lemma 3.3).

Theorem 5.1. *Let $\{(z_{k-1}, x_k, g_k, \hat{x}_k)\}_{k\geq 0}$ be generated by Method 4.6 for the non-smooth problem (4) associated with $\lambda_k = (k+1)/2$ and $\beta_k \equiv 0$. Assume that $\sigma_f > 0$ and $\sup_{k\geq 0}\|g_k\|_* \leq M_f < +\infty$. Then, we have*

$$\max\{f(\hat{x}_k) - f(x^*),\ \min_{0\leq i\leq k} f(x_i) - f(x^*)\} + \sigma_f \xi(x_{k+1}, x^*) \leq \frac{2M_f^2}{\sigma_d \sigma_f (k+4)}, \quad \forall k \geq 0$$

with the classical method, and

$$f(\hat{x}_k) - f(x^*) + \sigma_f \xi(z_k, x^*) \leq \frac{2M_f^2}{\sigma_d\sigma_f}\frac{k + \log k + 3/2}{(k+1)(k+2)} = O\left(\frac{M_f^2}{\sigma_d\sigma_f k}\right), \quad \forall k \geq 1$$

with the modified method.

Proof. Since $\beta_k \equiv 0$ and $S_k = \frac{(k+1)(k+2)}{4}$, Theorem 4.8 implies the estimate

$$f(\hat{x}_k) - f(x^*) + \sigma_f \xi(z_k, x^*) \leq \frac{C_k}{S_k} = \frac{4C_k}{(k+1)(k+2)} \tag{36}$$

with C_k defined by (31). The classical method also admits the same estimate replacing $f(\hat{x}_k)-f(x^*)$ by $\min_{0\leq i\leq k} f(x_i) - f(x^*)$ and we have

$$C_k = \frac{1}{2\sigma_d}\sum_{i=0}^{k}\frac{\lambda_i^2}{\beta_{i-1}+S_i\sigma_f}\|g_i\|_*^2 \leq \frac{M_f^2}{2\sigma_d\sigma_f}\sum_{i=0}^{k}\frac{\lambda_i^2}{S_i}.$$

Using the inequality

$$\sum_{i=0}^{k}\frac{\lambda_i^2}{S_i} = \sum_{i=0}^{k}\frac{i+1}{i+2} \leq \frac{(k+1)(k+2)}{k+4} \tag{37}$$

(see [16, Proposition A.3]), we obtain the first assertion for the classical method.

In the modified method, on the other hand, we have

$$C_k = \frac{1}{2\sigma_d}\sum_{i=0}^{k}\frac{\lambda_i^2 S_i}{\lambda_i^2\sigma_f + S_i(\beta_{i-1}+S_{i-1}\sigma_f)}\|g_i\|_*^2 \leq \frac{M_f^2}{2\sigma_d\sigma_f}\sum_{i=0}^{k}\frac{(i+1)(i+2)}{i(i+2)+4}$$

and

$$\sum_{i=0}^{k}\frac{(i+1)(i+2)}{i(i+2)+4} \leq \frac{1}{2}+\sum_{i=1}^{k}\frac{(i+1)(i+2)}{i(i+2)} = \frac{1}{2}+\sum_{i=1}^{k}\left(1+\frac{1}{i}\right) \leq \frac{1}{2}+k+(1+\log k)$$

for all $k \geq 1$, which leads (36) to the second assertion. □

Note that the choices of parameters $\lambda_k = (k+1)/2$ and $\beta_k \equiv 0$ do not depend on M_f and σ_f. However, we require σ_f when we solve the subproblems. For instance, the classical method with the extended MD model (17) associated with the above parameters $\lambda_k = (k+1)/2$, $\beta_k \equiv 0$ becomes

$$\begin{aligned}
x_{k+1} := z_k \quad &:= \quad \operatorname*{argmin}_{x\in Q}\{\lambda_k[f(x_k)+\langle g_k, x-x_k\rangle + \sigma_f\xi(x_k,x)] + S_{k-1}\sigma_f\xi(x_k,x)\} \\
&= \quad \operatorname*{argmin}_{x\in Q}\{\lambda_k[f(x_k)+\langle g_k, x-x_k\rangle] + S_k\sigma_f\xi(x_k,x)\} \\
&= \quad \operatorname*{argmin}_{x\in Q}\left\{\frac{\lambda_k}{S_k\sigma_f}[f(x_k)+\langle g_k, x-x_k\rangle] + \xi(x_k,x)\right\} \\
&= \quad \operatorname*{argmin}_{x\in Q}\left\{\frac{2}{\sigma_f(k+2)}[f(x_k)+\langle g_k, x-x_k\rangle] + \xi(x_k,x)\right\}, \\
\hat{x}_k \quad &:= \quad \frac{1}{S_k}\sum_{i=0}^{k}\lambda_i x_i = \frac{2}{(k+1)(k+2)}\sum_{i=0}^{k}(i+1)x_i,
\end{aligned}$$

which gives the estimates

$$\begin{aligned}
\max\{f(\hat{x}_k)-f(x^*),\ \min_{0\leq i\leq k} f(x_i)-f(x^*)\} + \sigma_f\xi(x_{k+1},x^*) \quad &\leq \quad \frac{2M_f^2}{\sigma_d\sigma_f(k+4)}, \\
\min\{\|\hat{x}_k-x^*\|^2,\ \|x_{i(k)}-x^*\|^2,\ \|x_{k+1}-x^*\|^2\} \quad &\leq \quad \frac{2M_f^2}{\sigma_d^2\sigma_f^2(k+4)},
\end{aligned} \tag{38}$$

for all $k \geq 0$, where $i(k) \in \operatorname{Argmin}_{0\leq i\leq k} f(x_i)$ (see Lemma 4.1 and Remark 4.2). Notice that the computation of z_k is equivalent to the subproblem (9) (the extended MD model for non-strongly convex case) with $\lambda_k := \frac{2}{\sigma_f(k+2)}$ and $\beta_k \equiv 1$. This result is closely related to [3, Proposition 1] and [29, Proposition 2.8]. The convergence result (38) is also valid for the DA model (18), and then we conclude that a strongly convex version of the DAM achieves the optimal complexity for non-smooth problems (see Section 2.2). Note that we do not exploit the multistage procedure and do not require an upper bound of $d(x^*)$ to obtain the optimality which are different features from [25].

5.2 Efficiency of the classical method for structured problems with constants L and δ

Now, let us consider the structured problems (6) for the particular case $L(\cdot) = L \geq 0$ and $\delta(\cdot,\cdot) = \delta \geq 0$. We firstly show the convergence result of the classical method of Method 4.7 which does not ensure the optimal convergence rate for the class $C_L^{1,1}(Q)$. This rate is as better as the existing methods compared in this subsection.

Theorem 5.2. *Consider the structured problem (6) for the special case $L(\cdot) = L \geq 0$ and $\delta(\cdot,\cdot) = \delta \geq 0$. Let $\{(z_{k-1}, w_{k-1}, x_k, \hat{x}_k)\}_{k\geq 0}$ be generated by the classical method of Method 4.7 with*

$$\beta_k \equiv \frac{L - \bar{\sigma}_f\sigma_d}{\sigma_d}, \quad \lambda_0 = 1, \quad \lambda_{k+1} = \frac{\beta_k + S_k\sigma_f}{\beta_k}. \tag{39}$$

Then, for every $k \geq 0$, we have

$$f(\hat{x}_k) - f(x^*) + \sigma_f\xi(z_k, x^*) \leq \frac{L - \bar{\sigma}_f\sigma_d}{\sigma_d} l_d(z_k; x^*) \min\left\{\left(1 - \frac{\sigma_f\sigma_d}{L - \bar{\sigma}_f\sigma_d + \sigma_f\sigma_d}\right)^k, \frac{1}{k+1}\right\} + \delta. \tag{40}$$

Furthermore, the left hand side of (40) can be replaced by $\frac{1}{S_k}\sum_{i=0}^k \lambda_k f(w_k) - f(x^*) + \sigma_f\xi(z_k, x^*)$ or by $\min_{0\leq i\leq k} f(w_i) - f(x^*) + \sigma_f\xi(z_k, x^*)$.

Proof. The classical method admits the relation (R_k) and (Q_k) with

$$C_k = \frac{1}{2}\sum_{i=0}^k \lambda_i \left(L - \sigma_d\left(\bar{\sigma}_f + \frac{\beta_{i-1} + S_{i-1}\sigma_f}{\lambda_i}\right)\right)\|w_i - x_i\|^2 + \sum_{i=0}^k \lambda_i\delta.$$

The definitions of λ_k and β_k implies that $C_k = \sum_{i=0}^k \lambda_i\delta = S_k\delta$ (since $\frac{\beta_{i-1}+S_{i-1}\sigma_f}{\lambda_i} = \beta_{i-1} = \frac{L-\bar{\sigma}_f\sigma_d}{\sigma_d}$) and $S_k = 1 + \left(1 + \frac{\sigma_f}{\beta_{-1}}\right)S_{k-1}$ for all $k \geq 0$. Therefore, we have $S_k \geq k+1$ and $S_k \geq (1 + \frac{\sigma_f}{\beta_{-1}})^k S_0 = (1 - \frac{\sigma_f}{\beta_{-1}+\sigma_f})^{-k}$, and the result follows from Theorem 4.9. □

It is interesting to notice that the particular choice of parameters (39) does not necessarily require the knowledge of σ_f and $\bar{\sigma}_f$ for the implementation of the classical gradient method with the extended MD model (17); for smooth problems (*i.e.*, $f \in C_L^{1,1}(Q)$), for instance, the corresponding subproblem can be rewritten as follows:

$$\begin{aligned}
z_k &:= \operatorname*{argmin}_{x\in Q}\{\lambda_k[f(x_k) + \langle\nabla f(x_k), x - x_k\rangle + \bar{\sigma}_f\xi(x_k, x)] + \beta_k\xi(x_k, x) + S_{k-1}\sigma_f\xi(x_k, x)\} \\
&= \operatorname*{argmin}_{x\in Q}\left\{f(x_k) + \langle\nabla f(x_k), x - x_k\rangle + \left(\bar{\sigma}_f + \frac{\beta_k + S_{k-1}\sigma_f}{\lambda_k}\right)\xi(x_k, x)\right\} \\
&\overset{(39)}{=} \operatorname*{argmin}_{x\in Q}\left\{f(x_k) + \langle\nabla f(x_k), x - x_k\rangle + \frac{L}{\sigma_d}\xi(x_k, x)\right\},
\end{aligned} \tag{41}$$

which requires only L; in the Euclidean setting (*i.e.*, $\frac{1}{\sigma_d}\xi(x_k, x) = \frac{1}{2}\|x_k - x\|_2^2$), furthermore, the Lipschitz condition (6) ensures that $f(x_{k+1}) \leq f(x_k)$ because $x_{k+1} = z_k$ is given by (41). The classical gradient method with the DA model (18) and the hybrid model (19), on the other hand, do not possess this advantage.

Let us see the corresponding methods for other particular structures.

- Consider the composite problem $\min_{x\in Q}[f(x)\equiv f_0(x)+\Psi(x)]$ as the example (ii) in Section 2.3 with the structure $\bar{\sigma}_f=\sigma_{f_0}=0$ (and thus $\sigma_f=\sigma_\Psi$) in the Euclidean setting (then, $\sigma_d=1$). Choosing parameters by (39), the classical gradient methods with the extended MD model and the hybrid model yield the Gradient Method $\mathcal{GM}(x_0,L)$ and the Dual Gradient Method $\mathcal{DG}(x_0,L)$ in [37], respectively (in this case, we do not exploit the procedure to estimate the Lipschitz constant L). Then, Theorem 5.2 improves the assertions [37, Theorems 4,5,6] in the following sense: The linear convergence factor $1-\frac{\sigma_f}{L+\sigma_f}=\frac{L}{L+\sigma_f}$ provided by (40) is less than the one in [37, Theorem 5] (because $\frac{L}{L+\sigma_f}\le\min\{\frac{\gamma L}{\sigma_f},1-\frac{\sigma_f}{4\gamma L}\}$ for any $\gamma>1$) and the same linear convergence is also valid for the method $\mathcal{DG}(x_0,L)$ which is not presented in the paper (the linear convergence for the dual gradient method was firstly demonstrated in [11]).
- For the convex problems with inexact oracle model as the example (iii) in Section 2.3 in the Euclidean setting (then, $\sigma_f=\bar{\sigma}_f$, $\sigma_d=1$), the classical gradient method with the extended MD model and the hybrid model yields the primal and the dual gradient methods in [11], respectively (but the definition (39) of $\{\lambda_k\}$ is slightly different from (4.1) and (4.2) in [11]). Because of $\sigma_d=1$ and $(L-\bar{\sigma}_f)l_d(z_k;x^*)\le Ld(x^*)=\frac{L}{2}\|x_0-x^*\|_2^2$, the estimate (40) slightly improves Theorems 4 and 5 in [11] (Since $\sigma_f=\bar{\sigma}_f$, the factor of linear convergence is the same).

Note that the classical gradient method with the DA model (18) can reduce the subproblems of the dual gradient method to one in [11, 37] preserving the same convergence property.

5.3 Efficiency of the modified method for structured problems with constants L and δ

The modified method of Method 4.7 for the structured problem (6) for the particular case $L(\cdot)=L\ge 0$, $\delta(\cdot,\cdot)=\delta\ge 0$ can be analyzed as follows. Differently from the classical method, it achieves the optimal convergence rate for the class $C_L^{1,1}(Q)$. The result below further implies to efficient rates for the conditional gradient methods, too.

Theorem 5.3. *Consider the structured problem (6) for the particular case $L(\cdot)=L\ge 0$ and $\delta(\cdot,\cdot)=\delta\ge 0$.*
(1) Let $\{(z_{k-1},w_{k-1},x_k,\hat{x}_k)\}_{k\ge 0}$ be generated by the modified method of Method 4.7 with

$$\beta_k\equiv\frac{L-\bar{\sigma}_f\sigma_d}{\sigma_d},\quad \lambda_0=1,\quad (L-\bar{\sigma}_f\sigma_d)\lambda_{k+1}^2=\sigma_d(S_k\sigma_f+\beta_{k-1})(\lambda_{k+1}+S_k)\ (k\ge 0) \tag{42}$$

(i.e., λ_{k+1} is determined as the largest root of the above quadratic equation). Then, for every $k\ge 0$, we have

$$\begin{aligned}f(\hat{x}_k)-f(x^*)+\sigma_f\xi(z_k,x^*) \;&\le\; \frac{L-\bar{\sigma}_f\sigma_d}{\sigma_d}l_d(z_k;x^*)\min\left\{\frac{4}{(k+2)^2},\left(1+\frac{1}{2}\sqrt{\frac{\sigma_f\sigma_d}{L-\bar{\sigma}_f\sigma_d}}\right)^{-2k}\right\}\\ &\quad+\min\left\{\frac{1}{3}k+\frac{1}{6}\log(k+2)+1,\ 1+\sqrt{\frac{L-\bar{\sigma}_f\sigma_d}{\sigma_f\sigma_d}}\right\}\delta.\end{aligned}$$

(2) Suppose further that $\sigma_f=0$ and Q is bounded. Let $\{(z_{k-1},w_{k-1},x_k,\hat{x}_k)\}_{k\ge 0}$ be generated by the modified method of Method 4.7 with $\beta_k\equiv 0$, $\lambda_k:=(k+1)/2$ as a conditional gradient method (refer Remark 4.10). Then, for every $k\ge 0$, we have

$$f(\hat{x}_k)-f(x^*)\le\frac{2L\max_{0\le i\le k}\|w_i-z_{i-1}\|^2}{k+4}+\frac{k+3}{3}\delta.$$

Proof. By Theorem 4.9, we have the estimate (33) with

$$\begin{aligned} C_k &= \frac{1}{2}\sum_{i=0}^{k} S_i \left(L(x_i) - \sigma_d \left(\bar{\sigma}_f + \frac{S_i(\beta_{i-1} + S_{i-1}\sigma_f)}{\lambda_i^2} \right) \right) \|\hat{x}_i - x_i\|^2 + \sum_{i=0}^{k} S_i \delta(x_i, \hat{x}_i) \\ &= \frac{1}{2}\sum_{i=0}^{k} \frac{\lambda_i^2}{S_i} \left(L - \sigma_d \left(\bar{\sigma}_f + \frac{S_i(\beta_{i-1} + S_{i-1}\sigma_f)}{\lambda_i^2} \right) \right) \|w_i - z_{i-1}\|^2 + \sum_{i=0}^{k} S_i \delta. \end{aligned}$$

(1) Notice that, since $\lambda_{k+1} + S_k = S_{k+1}$, (42) eliminates the above first summation so that we have $C_k = \sum_{i=0}^{k} S_i \delta$. Therefore, using Lemmas A.1 to A.4, given at Appendix, for the analysis of (42), (33) leads to the assertion.

(2) Letting $\lambda_k = (k+1)/2$, $\beta_k = 0$, and $\sigma_f = 0$ in Theorem 4.9 with C_k described above and using the inequality (37) establish that

$$f(\hat{x}_k) - f(x^*) \le \frac{C_k}{S_k} = \frac{L\sum_{i=0}^{k} \frac{\lambda_i^2}{S_i} \|w_i - z_{i-1}\|^2}{2S_k} + \frac{\sum_{i=0}^{k} S_i \delta}{S_k} \le \frac{2L \max_{0 \le i \le k} \|w_i - z_{i-1}\|^2}{k+4} + \frac{k+3}{3}\delta.$$

□

In the non-strongly convex case $\sigma_f = \bar{\sigma}_f = 0$, Tseng's methods [41] are derived from the modified method with the model (17) or (18) and Nesterov's method [34] is derived with the hybrid model (19) described in Section 2.3.2. From these facts one can conclude that the first result of Theorem 5.3 yields the strongly convex versions of Tseng's and Nesterov's methods with optimal complexity (see [11] for the verification of the optimality). The fast/accelerated gradient method in [11, 12, 37] for strongly convex problems are different from these three particularizations of the models (17) to (19).

Let us consider the Euclidean setting $d(x) = \frac{1}{2}\|x - x_0\|_2^2$, $\sigma_d = 1$. The first assertion of Theorem 5.3, applied to the convex problems with inexact oracle model (the example (iii) in Section 2.3), is slightly better than the estimate [11, Theorem 7] in view of $(L - \sigma_f) l_d(z_k; x^*) \le L d(x^*)$ and $\frac{L - \sigma_f}{\sigma_f} \le \frac{L}{\sigma_f}$. Furthermore, the first assertion applied to the composite problems $\min_{x \in Q}[f(x) \equiv f_0(x) + \Psi(x)]$ (the example (ii) in Section 2.3) is the same as Nesterov's one [37, Theorem 6] with $\gamma_u = 2$ (Recall that $\bar{\sigma}_f = \sigma_{f_0} = 0, \sigma_f = \sigma_\Psi$). Therefore, Method 4.7 achieves the optimal complexity for smooth and strongly convex problems (see Section 2.3).

The second result of Theorem 5.3 matches the conclusion for the classical CGM observed in [16, Section 5.2.1]. If we further assume $f \in C_L^{1,1}(Q)$, then the corresponding implementation of the second assertion with the extended MD model (17) and the DA model (18) yield particular instances of the CGMs proposed by Lan [27] (see Section 2.3.2).

5.4 Efficiency of the modified method for weakly smooth problems

Considering structured problems in the case when $\delta(y, x) = \frac{M(y)}{\rho}\|y - x\|^\rho$, $\rho \in [1, 2)$, we can provide convergence analysis for problems involving weakly smooth functions of the class $C_M^{1,\rho-1}(Q)$ (see examples (iv) and (v) in Section 2.3). Note that the smooth case $\rho = 2$ reduces to the situation $\delta(y, x) = 0$ which has been already discussed. In this section, we show convergence results of modified proximal/conditional gradient methods for this setting. In the case $\rho = 1$, the results of Sections 5.4.1 to 5.4.3 can be seen as variants of stochastic gradient methods developed in [8, 18] for the deterministic setting.

5.4.1 General bound

Our analysis for proximal gradient methods is based on the following lemma.

Lemma 5.4. *Let $\{(z_{k-1}, w_{k-1}, x_k, \hat{x}_k)\}_{k\geq 0}$ be generated by the modified method of Method 4.7 with parameters $\{(\lambda_k, \beta_{k-1})\}_{k\geq 0}$ and $\sigma_f \geq 0$ for the structured problem (6) for the special case $\delta(y,x) = \frac{M(y)}{\rho}\|y-x\|^\rho$, $\rho \in [1,2)$. Put $\alpha_k := L(x_k) - \sigma_d\left(\bar{\sigma}_f + \frac{S_k(\beta_{k-1}+S_{k-1}\sigma_f)}{\lambda_k^2}\right)$. If $\alpha_i < 0$ for each $0 \leq i \leq k$, then we have*

$$f(\hat{x}_k) - f(x^*) + \sigma_f \xi(z_k, x^*) \leq \frac{\beta_k l_d(z_k; x^*)}{S_k} + \frac{(2-\rho)\max_{0\leq i\leq k} M(x_i)^{\frac{2}{2-\rho}}}{2\rho S_k}\sum_{i=0}^{k}\frac{S_i}{(-\alpha_i)^{\frac{\rho}{2-\rho}}}.$$

Proof. Note that the function $g(r) = ar^2 + br^\rho$ for $r \geq 0, a < 0, b \in \mathbb{R}$ satisfies $\max_{r\geq 0} g(r) = \frac{2-\rho}{2\rho}(-2a)^{\frac{-\rho}{2-\rho}}(\rho b)^{\frac{2}{2-\rho}}$. Hence, Theorem 4.9 concludes that

$$\begin{aligned} f(\hat{x}_k) - f(x^*) + \sigma_f\xi(z_k, x^*) &\leq \frac{\beta_k l_d(z_k; x^*)}{S_k} + \frac{1}{S_k}\sum_{i=0}^{k} S_i\left(\frac{1}{2}\alpha_i\|\hat{x}_i - x_i\|^2 + \frac{M(x_i)}{\rho}\|\hat{x}_i - x_i\|^\rho\right) \\ &\leq \frac{\beta_k l_d(z_k; x^*)}{S_k} + \frac{1}{S_k}\sum_{i=0}^{k} S_i \times \frac{2-\rho}{2\rho}(-\alpha_i)^{\frac{-\rho}{2-\rho}} M(x_i)^{\frac{2}{2-\rho}}, \end{aligned}$$

which proves the assertion. □

5.4.2 Convergence analysis for the non strongly convex case

Let us deduce a convergence result of modified proximal gradient methods for the non strongly convex case $\sigma_f = \bar{\sigma}_f = 0$. The result with $\rho = 1$ is closely related to the deterministic versions of [18, Proposition 8] and [8, Corollary 1].

Theorem 5.5. *Consider the structured problem (6) for the special case $L(\cdot) = L \geq 0$, $\sigma_f = \bar{\sigma}_f = 0$, and $\delta(y,x) = \frac{M(y)}{\rho}\|y-x\|^\rho$ for $M(\cdot) \geq 0$, $\rho \in [1,2)$. Let $\{(z_{k-1}, w_{k-1}, x_k, \hat{x}_k)\}_{k\geq 0}$ be generated by the modified method of Method 4.7 with*

$$\lambda_k := \frac{k+1}{2}, \quad \beta_k := \frac{L}{\sigma_d} + \frac{\gamma}{\sigma_d}(k+3)^{\frac{3}{2}(2-\rho)}, \quad \gamma > 0.$$

Then, for every $k \geq 0$, we have

$$f(\hat{x}_k) - f(x^*) \leq \frac{4Ll_d(z_k; x^*)}{\sigma_d(k+1)(k+2)} + \left[\frac{4\gamma l_d(z_k; x^*)}{\sigma_d} + \frac{\max_{0\leq i\leq k} M(x_i)^{\frac{2}{2-\rho}}}{3\rho\gamma^{\frac{\rho}{2-\rho}}}\right]\frac{(k+3)^{\frac{3}{2}(2-\rho)}}{(k+1)(k+2)}.$$

Proof. We apply Lemma 5.4 to prove the assertion. Note that

$$\frac{\beta_k}{S_k} = \frac{4L}{\sigma_d(k+1)(k+2)} + \frac{4\gamma(k+3)^{\frac{3}{2}(2-\rho)}}{\sigma_d(k+1)(k+2)} \tag{43}$$

and α_k in Lemma 5.4 becomes now $\alpha_k = -\frac{L}{k+1} - \gamma\frac{(k+2)^{\frac{3}{2}(2-\rho)+1}}{k+1} \leq -\gamma\frac{(k+2)^{\frac{3}{2}(2-\rho)+1}}{k+1} < 0$. Furthermore, we have

$$\begin{aligned} \frac{1}{S_k}\sum_{i=0}^{k}\frac{S_i}{(-\alpha_i)^{\frac{\rho}{2-\rho}}} &\leq \frac{1}{S_k}\sum_{i=0}^{k}\frac{(i+1)^{\frac{\rho}{2-\rho}+1}}{4\gamma^{\frac{\rho}{2-\rho}}(i+2)^{\frac{3}{2}\rho+\frac{\rho}{2-\rho}-1}} \leq \frac{1}{4\gamma^{\frac{\rho}{2-\rho}}S_k}\sum_{i=0}^{k}(i+2)^{2-\frac{3}{2}\rho} \\ &\leq \frac{1}{4\gamma^{\frac{\rho}{2-\rho}}S_k}\frac{2}{3(2-\rho)}(k+3)^{3-\frac{3}{2}\rho} = \frac{2(k+3)^{\frac{3}{2}(2-\rho)}}{3(2-\rho)\gamma^{\frac{\rho}{2-\rho}}(k+1)(k+2)}, \end{aligned} \tag{44}$$

where the second and the third inequalities are due to $i+1 \le i+2$ and the fact $\sum_{i=0}^{k}(i+2)^q \le \frac{1}{1+q}(k+3)^{1+q}$, $\forall q > -1$, respectively. Consequently, the theorem follows by applying Lemma 5.4 with the inequalities (43) and (44). □

Notice that we need the parameter ρ to define β_k but not the $M(\cdot)$. Now let us calculate an efficient choice for γ. Suppose that $M(\cdot) \le \hat{M} < +\infty$. Using $l_d(z_k; x^*) \le d(x^*)$ and the fact that the function $g(\gamma) = a\gamma + \frac{b}{\gamma^p}$ $(a, b, p > 0)$ attains its minimum at $\gamma^* = (pb/a)^{\frac{1}{p+1}}$ on $(0, \infty)$ with $g(\gamma^*) = (p+1)p^{\frac{-p}{p+1}} a^{\frac{p}{p+1}} b^{\frac{1}{p+1}}$, the choice

$$\gamma = \gamma^* := \left(\frac{\rho}{2-\rho} \frac{\hat{M}^{\frac{2}{2-\rho}}}{3\rho} \frac{\sigma_d}{4d(x^*)} \right)^{\frac{2-\rho}{2}} = \hat{M} \left(\frac{\sigma_d}{12(2-\rho)d(x^*)} \right)^{\frac{2-\rho}{2}}$$

makes the estimate of Theorem 5.5 as follows:

$$\begin{aligned} f(\hat{x}_k) - f(x^*) &\le \frac{4Ld(x^*)}{\sigma_d(k+1)(k+2)} + \frac{2}{2-\rho} \left(\frac{\rho}{2-\rho} \right)^{-\frac{\rho}{2}} \left(\frac{4d(x^*)}{\sigma_d} \right)^{\frac{\rho}{2}} \left(\frac{\hat{M}^{\frac{2}{2-\rho}}}{3\rho} \right)^{\frac{2-\rho}{2}} \frac{(k+3)^{\frac{3}{2}(2-\rho)}}{(k+1)(k+2)} \\ &= \frac{4Ld(x^*)}{\sigma_d(k+1)(k+2)} + \frac{2(2\sqrt{3})^\rho}{3\rho(2-\rho)^{\frac{2-\rho}{2}}} \hat{M} \left(\frac{d(x^*)}{\sigma_d} \right)^{\frac{\rho}{2}} \frac{(k+3)^{\frac{3}{2}(2-\rho)}}{(k+1)(k+2)}. \end{aligned}$$

Note that $\min_{x>0} x^x = (1/e)^{1/e}$ and $\max_{\rho \in [1,2]} \frac{2}{3\rho}(2\sqrt{3})^\rho = \frac{2}{3 \cdot 2}(2\sqrt{3})^2 = 4$ because $\log(2\sqrt{3}) > 1$ implies the positivity of the derivative of $\frac{2}{3\rho}(2\sqrt{3})^\rho$. Therefore, we have $\frac{2(2\sqrt{3})^\rho}{3\rho(2-\rho)^{\frac{2-\rho}{2}}} \le 4e^{1/(2e)}$ which shows $f(\hat{x}_k) - f(x^*) \le O\left(\frac{Ld(x^*)}{\sigma_d} k^{-2} + \hat{M} \left(\frac{d(x^*)}{\sigma_d} \right)^{\frac{\rho}{2}} k^{-\frac{3\rho-2}{2}} \right)$. Consequently, we obtain an upper bound of the iteration complexity to obtain $f(\hat{x}_k) - f(x^*) \le \varepsilon$ which is proportional to

$$\left(\frac{Ld(x^*)}{\sigma_d \varepsilon} \right)^{\frac{1}{2}} + \left(\frac{d(x^*)}{\sigma_d} \right)^{\frac{\rho}{3\rho-2}} \left(\frac{\hat{M}}{\varepsilon} \right)^{\frac{2}{3\rho-2}}.$$

In view of the lower complexity (11) (with L replaced by $\hat{M}$ there), it turns out that the order of the second term is optimal for the class $C_{\hat{M}}^{1,\rho-1}(E)$.

5.4.3 Convergence analysis for the strongly convex case

Now we show a result for the strongly convex case $\sigma_f > 0$.

Theorem 5.6. *Consider the structured problem (6) for the special case $L(\cdot) = L \ge 0$ and $\delta(y,x) = \frac{M(y)}{\rho}\|y - x\|^\rho$ for $M(\cdot) \ge 0$, $\rho \in [1,2)$. Assume that $\sigma_f > 0$. Let $\{(z_{k-1}, w_{k-1}, x_k, \hat{x}_k)\}_{k \ge 0}$ be generated by the modified method of Method 4.7 with*

$$\lambda_k := \frac{1}{p+1}(k+1)^p, \quad \beta_k := \left(\frac{L}{\sigma_d} + \beta \right)(k+2)^{p-1}$$

where $p \ge 1$ and $\beta \ge 0$ with $\sigma_d \bar{\sigma}_f + pL + (p+1)\sigma_d \beta > 0$. Then, for every $k \ge 0$, we have

$$\begin{aligned} f(\hat{x}_k) - f(x^*) + \sigma_f \xi(z_k, x^*) \le & \left(\frac{L}{\sigma_d} + \beta \right)(p+1)^2 l_d(z_k; x^*) \frac{(k+2)^{p-1}}{(k+1)^{p+1}} \\ & + \frac{(p+1)(2-\rho)\max_{0 \le i \le k} M(x_i)^{\frac{2}{2-\rho}}}{2\rho(\sigma_d \bar{\sigma}_f + pL + (p+1)\sigma_d \beta)^{\frac{\rho}{2-\rho}}} \frac{1}{(k+1)^{p+1}} \\ & + \frac{3^{p+1}(2-\rho)\max_{0 \le i \le k} M(x_i)^{\frac{2}{2-\rho}}}{2\rho} \left(\frac{2^{p-1}(p+1)^2}{\sigma_d \sigma_f} \right)^{\frac{\rho}{2-\rho}} P(k), \end{aligned}$$

where

$$P(k) = \begin{cases} \left(p+2-\frac{2\rho}{2-\rho}\right)^{-1}(k+1)^{-\frac{3\rho-2}{2-\rho}} & : p+1 > \frac{3\rho-2}{2-\rho}, \\ \dfrac{1+\log k}{(k+1)^{p+1}} & : p+1 = \frac{3\rho-2}{2-\rho}, \\ \dfrac{1-\left(p+2-\frac{2\rho}{2-\rho}\right)^{-1}}{(k+1)^{p+1}} & : p+1 < \frac{3\rho-2}{2-\rho}. \end{cases}$$

Proof. Note that β_k is non-decreasing and $\frac{1}{(p+1)^2}(k+1)^{p+1} \leq S_k \leq \frac{1}{(p+1)^2}(k+2)^{p+1}$. Then, we have

$$\frac{\beta_k}{S_k} \leq \left(\frac{L}{\sigma_d}+\beta\right)(p+1)^2\frac{(k+2)^{p-1}}{(k+1)^{p+1}} = O(k^{-2}). \tag{45}$$

Since the inequalities $\frac{S_k}{\lambda_k^2} \geq \frac{1}{(k+1)^{p-1}}$ and $\frac{S_kS_{k-1}}{\lambda_k^2} \geq \frac{1}{(p+1)^2}\frac{k^{p+1}}{(k+1)^{p-1}} \geq \frac{k^2}{2^{p-1}(p+1)^2}$ for $k \geq 1$ imply

$$-\alpha_k := \sigma_d\left(\bar{\sigma}_f + \frac{S_k(\beta_{k-1}+S_{k-1}\sigma_f)}{\lambda_k^2}\right) - L \geq \sigma_d\bar{\sigma}_f + \beta\sigma_d + \frac{\sigma_d\sigma_f}{2^{p-1}(p+1)^2}k^2 > 0, \quad k \geq 1,$$

we obtain

$$\frac{S_k}{(-\alpha_k)^{\frac{\rho}{2-\rho}}} < \frac{1}{(p+1)^2}\left(\frac{2^{p-1}(p+1)^2}{\sigma_d\sigma_f}\right)^{\frac{\rho}{2-\rho}}\frac{(k+2)^{p+1}}{k^{\frac{2\rho}{2-\rho}}} \leq \frac{3^{p+1}}{(p+1)^2}\left(\frac{2^{p-1}(p+1)^2}{\sigma_d\sigma_f}\right)^{\frac{\rho}{2-\rho}}k^{p+1-\frac{2\rho}{2-\rho}}$$

for all $k \geq 1$. Combining with $\frac{S_0}{(-\alpha_0)^{\frac{\rho}{2-\rho}}} = \frac{1}{(p+1)(\sigma_d\bar{\sigma}_f+pL+(p+1)\sigma_d\beta)^{\frac{\rho}{2-\rho}}}$ yields that

$$\frac{1}{S_k}\sum_{i=0}^{k}\frac{S_i}{(-\alpha_i)^{\frac{\rho}{2-\rho}}} \leq \frac{p+1}{(\sigma_d\bar{\sigma}_f+pL+(p+1)\sigma_d\beta)^{\frac{\rho}{2-\rho}}}\frac{1}{(k+1)^{p+1}} + 3^{p+1}\left(\frac{2^{p-1}(p+1)^2}{\sigma_d\sigma_f}\right)^{\frac{\rho}{2-\rho}}P(k), \tag{46}$$

where the factor $P(k)$ is due to the following inequality:

$$\sum_{i=1}^{k} i^q \leq \begin{cases} \frac{1}{1+q}(k+1)^{q+1} & : q > -1, \\ 1+\log k & : q = -1, \\ 1-\frac{1}{1+q} & : q < -1. \end{cases}$$

Consequently, the assertion follows from Lemma 5.4 with the inequalities (45) and (46). □

Notice that we do not need ρ and $M(\cdot)$ in the definition the parameters λ_k, β_k; the result holds for all acceptable $\rho \in [1,2)$. If we further have $p+1 > \frac{3\rho-2}{2-\rho}$, then $P(k)$ has the best rate of convergence for a fixed ρ. Now let us see the above upper bound in the case $L = \bar{\sigma}_f = 0,\ M(\cdot) = M,\ \beta > 0,\ \sigma_f > 0,\ p+1 > \frac{3\rho-2}{2-\rho}$:

$$\begin{aligned} f(\hat{x}_k) - f(x^*) + \sigma_f\xi(z_k,x^*) \quad \leq \quad & \beta(p+1)^2 l_d(z_k;x^*)\tfrac{(k+2)^{p-1}}{(k+1)^{p+1}} + \tfrac{(p+1)(2-\rho)M^{\frac{\rho}{2-\rho}}}{2\rho((p+1)\sigma_d\beta)^{\frac{\rho}{2-\rho}}}\tfrac{1}{(k+1)^{p+1}} \\ & + \tfrac{3^{p+1}(2-\rho)}{2\rho}M^{\frac{2}{2-\rho}}\left(\tfrac{2^{p-1}(p+1)^2}{\sigma_d\sigma_f}\right)^{\frac{\rho}{2-\rho}}\left(p+2-\tfrac{2\rho}{2-\rho}\right)^{-1}(k+1)^{-\frac{3\rho-2}{2-\rho}}. \end{aligned}$$

Since this bound is of $O\left(c(p,\rho)\frac{M^{2/(2-\rho)}}{(\sigma_d\sigma_f)^{\rho/(2-\rho)}}k^{-\frac{3\rho-2}{2-\rho}}\right)$ for a continuous function $c(p,\rho)$, it achieves the optimal rate of convergence (11) for the strongly convex case (recall that $\sigma_d\sigma_f$ becomes a convexity parameter of f with respect to the norm $\|\cdot\|$; see Section 2.1).

Let us consider the non-smooth case $\rho = 1,\ \bar{\sigma}_f = \sigma_f > 0$. Then, taking $p = 1$ and $\beta = 0$ yields $\lambda_k = (k+1)/2,\ \beta_{k-1} = L/\sigma_d$, and

$$f(\hat{x}_k) - f(x^*) + \sigma_f\xi(z_k,x^*) \leq \frac{4Ll_d(z_k;x^*)}{\sigma_d(k+1)^2} + \frac{\max_{0\leq i\leq k} M(x_i)^2}{(\sigma_d\sigma_f+L)(k+1)^2} + \frac{18\max_{0\leq i\leq k} M(x_i)^2}{\sigma_d\sigma_f(k+1)}.$$

This result is similar to the ones [18, Proposition 9] and [8, Corollary 2] in the deterministic case.

5.4.4 Convergence analysis of conditional gradient methods

We finally consider the case of conditional gradient methods: $\beta_k \equiv 0,\ \sigma_f = \bar{\sigma}_f = 0$. This case can be analyzed without Lemma 5.4.

Theorem 5.7. *Suppose that the structured problem (6) is equipped with $L(\cdot) = L \geq 0$, $\sigma_f = \bar{\sigma}_f = 0$, and $\delta(y,x) = \frac{M}{\rho}\|y-x\|^\rho$ for $M \geq 0,\ \rho \in [1,2)$. Then, the modified method of Method 4.7 for the problem with $\lambda_k = (k+1)/2$ and $\beta_k \equiv 0$ generates a sequence $\{\hat{x}_k\}_{k\geq 0} \subset Q$ satisfying*

$$f(\hat{x}_k) - f(x^*) \leq \frac{2L\mathrm{Diam}(Q)^2}{k+4} + \frac{2^{\rho+1}M\mathrm{Diam}(Q)^\rho}{\rho(3-\rho)(k+2)^{\rho-1}} \tag{47}$$

for every $k \geq 0$.

Proof. Theorem 4.9 yields that $f(\hat{x}_k) - f(x^*) \leq C_k/S_k$ with $S_k = (k+1)(k+2)/4$ and

$$C_k = \sum_{i=0}^k S_i\left(\frac{L}{2}\|\hat{x}_i - x_i\|^2 + \frac{M}{\rho}\|\hat{x}_i - x_i\|^\rho\right) = \sum_{i=0}^k \left(\frac{L}{2}\frac{\lambda_i^2}{S_i}\|w_i - z_{i-1}\|^2 + \frac{M}{\rho}\frac{\lambda_i^\rho}{S_i^{\rho-1}}\|w_i - z_{i-1}\|^\rho\right)$$

(see Remark 4.10). Using the inequality (37) and

$$\sum_{i=0}^k \frac{\lambda_i^\rho}{S_i^{\rho-1}} = \frac{1}{2^{2-\rho}}\sum_{i=0}^k \frac{i+1}{(i+2)^{\rho-1}} \leq \frac{1}{2^{2-\rho}}\sum_{i=0}^k (i+1)^{2-\rho} \leq \frac{1}{2^{2-\rho}(3-\rho)}(k+2)^{3-\rho}$$

(the first and the second inequalities are due to $i+1 \leq i+2$ and the fact $\sum_{i=0}^k (i+1)^q \leq \frac{1}{1+q}(k+2)^{1+q}$ for $q \geq 0$, respectively), we conclude that

$$f(\hat{x}_k) - f(x^*) \leq \frac{C_k}{S_k} \leq \frac{2L\mathrm{Diam(Q)}^2}{k+4} + \frac{2^\rho M\mathrm{Diam(Q)}^\rho}{\rho(3-\rho)}\frac{(k+2)^{2-\rho}}{k+1}.$$

The estimate (47) now follows from $\frac{k+2}{k+1} \leq 2$ for $k \geq 0$. □

The bound (47) is also valid for the classical CGM (13) with $\tau_k := \lambda_{k+1}/S_{k+1} = \frac{2}{k+3}$, $\hat{x}_k := x_k$; it can be derived in the same way as Theorem 5.7 based on the estimate (35) since $f(x_0) - l_f(x_0; z_0) \overset{(6)}{\leq} \frac{L}{2}\mathrm{Diam}(Q)^2 + \frac{M}{\rho}\mathrm{Diam}(Q)^\rho$ and $\delta(x_{k-1}, x_k) = \frac{M}{\rho}\|x_k - x_{k-1}\|^\rho \overset{(13)}{=} \frac{M}{\rho}\frac{\lambda_k^\rho}{S_k^\rho}\|x_{k-1} - z_{k-1}\|^\rho \leq \frac{M}{\rho}\frac{\lambda_k^\rho}{S_k^\rho}\mathrm{Diam}(Q)^\rho$ for $k \geq 1$. This result in the case $L = 0$ is very similar to a known result for the classical CGM (see [9, Proposition 1.1]).

Since the choice $\lambda_k = (k+1)/2$ and $\beta_k \equiv 0$ are independent of L, M, and ρ, the conditional gradient methods can be applied to the classes $C_L^{1,\nu}(E)$, $\nu \in (0,1]$ ensuring the convergence $f(\hat{x}_k) - f(x^*) \leq O\left(\frac{LR^{1+\nu}}{k^\nu}\right)$ where $R = \mathrm{Diam}(Q)$. This rate of convergence is optimal when $\nu = 1$ in the sense of linear optimization oracle [27] and nearly optimal otherwise [20].

6 Conclusion

This paper proposes a new framework for applying subgradient methods to minimize strongly convex functions. It unifies the analysis of PGMs and CGMs for several classes of problems including non-smooth, smooth, and weakly smooth problems. We have introduced the notion of strong convexity with respect to the prox-function, which generalizes the one in the Euclidean setting. The proposed PGMs establish optimal convergence rates for these problems with slight improvements than some existing methods. Furthermore, particular cases of the framework yield a family of variations of the classical CGM with optimal and nearly optimal guarantee of convergence in the non-strongly convex case.

A remarkable novel result in this paper, in view of method efficiency, is the achievement of the optimal complexity for the weakly smooth problems (the class $C_M^{1,\nu}(Q)$, $\nu \in [0,1)$) in the strongly convex case without knowing the constant M and an upper bound of $d(x^*)$ (Section 5.4.3; see also Section 2.3.1 (iv) for remarks on the literature). The theoretical approach for that is similar to the ones in [11, 12, 38] because the structure (6) assumes an oracle inexactness of the original problem. However, the essential of the analysis in Sections 5.4.2, 5.4.3 is not the same; it can be seen as a generalization of the techniques of [18, 19] in the deterministic case.

We finally describe several topics for further considerations. At first, we can consider a generalization/combination of the (sub)gradient-besed methods here with smoothing technique, stochastic situation, or uniformly convex setting. Related studies can be seen in [18, 19, 25, 27]. Secondly, one can further consider to tune the parameters, the weight and the scaling ones, to obtain an efficient convergence. The proposed choices in Section 5 are not the only way to ensure the optimal convergence; see, e.g., [16, 29] for some discussions on other choices. Thirdly, it is important to note that the convergence results for the class $C_M^{1,\nu}(Q)$ in Sections 5.4.2, 5.4.3 are not *adaptive* in contrast to the known method [38] proposed by Nesterov; namely , it does not ensure the optimal convergence without knowing the parameter ν. From the practical viewpoint, it will be important to develop techniques to ensure efficient convergence rates without such problem specific information.

Acknowledgements

This work was partially supported by JSPS Grant-in-Aid for Scientific Research (C) number 26330024. The author is thankful to Prof. Mituhiro Fukuda for comments and suggestions and also to Prof. Guanghui Lan for pointing out some related results.

References

[1] A. Argyriou, M. Signoretto, and J. Suykens, Hybrid conditional gradient - smoothing algorithms with applications to sparse and low rank regularization, *in Regularization, Optimization, Kernels, and Support Vector Machines* (J. Suykens, A. Argyriou, and M. Signoretto, eds.), pp. 53–82, Chapman & Hall/CRC, Boca Raton, USA, 2014.

[2] A. Auslender and M. Teboulle, Interior gradient and proximal method for convex and conic optimization, *SIAM Journal on Optimization*, **16**, pp. 697–725, 2006.

[3] F. Bach, Duality between subgradient and conditional gradient methods, *SIAM Journal on Optimization*, **25**, pp. 115–129, 2015.

[4] A. Beck and M. Teboulle, Mirror descent and nonlinear projected subgradient methods for convex optimization, *Operations Research Letters*, **31**, pp. 167–175, 2003.

[5] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM Journal on Imaging Sciences*, **2**, pp. 183–202, 2009.

[6] A. Beck and M. Teboulle, Smoothing and first order methods: A unified framework, *SIAM Journal on Optimization*, **22**, pp. 557–580, 2012.

[7] L. Bregman, The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming, *USSR Computational Mathematics and Mathematical Physics*, **7**, pp. 200–217, 1967.

[8] X. Chen, Q. Lin, and J. Peña, Optimal regularized dual averaging methods for stochastic optimization, *Advances in Neural Information Processing Systems*, **25**, pp. 395–403, 2012.

[9] B. Cox, B. Juditsky, and A. Nemirovski, Dual subgradient algorithms for large-scale nonsmooth learning problems, *Mathematical Programming*, **148**, pp. 143–180, 2013.

[10] V. F. Demyanov and A. M. Rubinov, *Approximate methods in optimization problems*, American Elsevier Publishing Company, New York, 1970.

[11] O. Devolder, F. Glineur, and Y. Nesterov, First-order methods with inexact oracle: The strongly convex case, *CORE Discussion Paper*, **2013/16**, 2013.

[12] O. Devolder, F. Glineur, and Y. Nesterov, First-order methods of smooth convex optimization with inexact oracle, *Mathematical Programming*, **146**, pp. 37–75, 2014.

[13] J. Dunn and S. Harshbarger, Conditional gradient algorithms with open loop step size rules, *Journal of Mathematical Analysis and Applications*, **62**, pp. 432–444, 1978.

[14] K.-H. Elster (ed.), *Modern mathematical methods in optimization*, Academie Verlag, Berlin, 1993.

[15] M. Frank and P. Wolfe, An algorithm for quadratic programming, *Naval Research Logistics Quarterly*, **3**, pp. 95–110, 1956.

[16] R. M. Freund and P. Grigas, New analysis and results for the conditional gradient method, *Mathematical Programming*, Online First, DOI 10.1007/s10107-014-0841-6, 2014.

[17] M. Fukushima and H. Mine, A generalized proximal point algorithm for certain non-convex minimization problems, *International Journal of Systems Science*, **12**, pp. 989–1000, 1981.

[18] S. Ghadimi and G. Lan, Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, I: A generic algorithmic framework, *SIAM Journal on Optimization*, **22**, pp. 1469–1492, 2012.

[19] S. Ghadimi and G. Lan, Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: Shrinking procedures and optimal algorithms, *SIAM Journal on Optimization*, **23**, pp. 2061–2089, 2013.

[20] C. Guzmán and A. Nemirovski, On lower complexity bounds for large-scale convex optimization, *Journal of Complexity*, **31**, pp. 1–14, 2015.

[21] Z. Harchaoui, A. Juditsky, and A. Nemirovski, Conditional gradient algorithms for norm-regularized smooth convex optimization, *Mathematical Programming*, Online first, DOI 10.1007/s10107-014-0778-9, 2014.

[22] M. Ito and M. Fukuda, A family of subgradient-based methods for convex optimization problems in a unifying framework, *Technical Report B-477*, Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, 2014.

[23] M. Jaggi, *Sparse convex optimization methods for machine learning*, Ph.D. thesis, ETH Zurich, 2011.

[24] M. Jaggi, Revisiting Frank-Wolfe: Projection-free sparse convex optimization, *Proceedings of the 30th International Conference on Machine Learning*, pp. 427–435, 2013.

[25] A. Juditsky and Y. Nesterov, Primal-dual subgradient methods for minimizing uniformly convex functions, *arXiv:1401.1792v1*, 2014.

[26] G. Lan, An optimal method for stochastic composite optimization, *Mathematical Programming*, **133**, pp.365–397, 2012.

[27] G. Lan, The complexity of large-scale convex programming under a linear optimization oracle, *arXiv:1309.5550v2*, 2014.

[28] G. Lan, Gradient sliding for composite optimization, *arXiv:1406.0919v2*, 2014.

[29] A. Nedić and D. Bertsekas, Convergence rate of incremental subgradient algorithms, *in Stochastic Optimization: Algorithms and Applications* (S. Uryasev and P. Pardalos, eds.), pp. 223–264, Kluwer Academic Publishers, Dordrecht, Netherlands, 2001.

[30] A. Nemirovski and Y. Nesterov, Optimal methods for smooth convex minimization, *Zh. Vychishl. Mat. i Mat. Fiz.*, **25**, pp. 356–369, 1985 (in Russian).

[31] A. Nemirovski and D. Yudin, *Problem complexity and method efficiency in optimization*, Nauka Publishers, Moscow, Russia, 1979 (in Russian); English translation: John Wiley & Sons, New York, USA, 1983.

[32] Y. Nesterov, A method of solving a convex programming problem with convergence rate $O(1/k^2)$, *Soviet Mathematics Doklady*, **27**, pp. 372–376, 1983.

[33] Y. Nesterov, *Introductory lectures on convex optimization : A basic course*, Kluwer Academic Publishers, Boston, 2004.

[34] Y. Nesterov, Smooth minimization of non-smooth functions, *Mathematical Programming*, **103**, pp. 127–152, 2005.

[35] Y. Nesterov, Excessive gap technique in nonsmooth convex minimization, *SIAM Journal on Optimization*, **16**, pp. 235–249, 2005.

[36] Y. Nesterov, Primal-dual subgradient methods for convex problems, *Mathematical Programming*, Ser. B, **120**, pp. 221–259, 2009.

[37] Y. Nesterov, Gradient methods for minimizing composite functions, *Mathematical Programming*, Ser. B, **140**, pp. 125–161, 2013.

[38] Y. Nesterov, Universal gradient methods for convex optimization problems, *Mathematical Programming*, Online First, DOI 10.1007/s10107-014-0790-0, 2014.

[39] B. N. Pshenichny and Y. M. Danilin, *Numerical methods in extremal problems*, MIR Publishers, Moscow, 1978.

[40] P. Tseng, On accelerated proximal gradient methods for convex-concave optimization, Technical Report, University of Washington, 2008.

[41] P. Tseng, Approximation accuracy, gradient methods, and error bound for structured convex optimization, *Mathematical Programming*, Ser. B, **125**, pp. 263–295, 2010.

A Appendix

In order to complete the proof of Theorem 5.3, we need to obtain upper bounds for $1/S_k$ and $\sum_{i=0}^{k} S_i/S_k$ for the sequence $\{S_k\}_{k\geq 0}$ defined by (42). Since $\lambda_{k+1} = S_{k+1} - S_k$, writing $r := \frac{\sigma_f \sigma_d}{L - \bar{\sigma}_f \sigma_d} \geq 0$, the sequence $\{S_k\}_{k\geq 0}$ in (42) is determined by the recurrence

$$S_0 = 1, \quad (S_{k+1} - S_k)^2 = S_{k+1}(1 + rS_k), \quad k \geq 0 \tag{48}$$

where the root of the equation in S_{k+1} takes the largest one, namely,

$$S_{k+1} = \frac{1 + (2+r)S_k + \sqrt{(1 + (2+r)S_k)^2 - 4S_k^2}}{2}. \tag{49}$$

The essentials of lemmas below are the same as [11, Lemma 4-7] excepting the replacement of μ/L in the article by an arbitrary $r \geq 0$.

Lemma A.1. *For any sequence $\{S_k\}_{k\geq 0}$ defined by (48) for $r \geq 0$, we have*

$$\frac{1}{S_k} \leq \min\left\{\frac{4}{(k+1)(k+4)}, \left(\frac{2}{2 + r + \sqrt{r^2 + 4r}}\right)^k\right\}, \quad \forall k \geq 0.$$

Proof. Since $S_{k+1} \geq S_k$, we have

$$\sqrt{S_{k+1}} - \sqrt{S_k} = \frac{S_{k+1} - S_k}{\sqrt{S_{k+1}} + \sqrt{S_k}} \geq \frac{S_{k+1} - S_k}{2\sqrt{S_{k+1}}} \overset{(48)}{=} \frac{1}{2}\sqrt{1 + rS_k} \geq \frac{1}{2} \tag{50}$$

which shows $\sqrt{S_k} \geq \frac{k}{2} + \sqrt{S_0} = \frac{k+2}{2}$ for all $k \geq 0$. Then, we have

$$S_k - S_0 = \sum_{i=0}^{k-1}(S_{i+1} - S_i) \overset{(48)}{=} \sum_{i=0}^{k-1}\sqrt{S_{i+1}(1 + rS_i)} \geq \sum_{i=0}^{k-1}\sqrt{S_{i+1}} \geq \sum_{i=0}^{k-1}\frac{i+3}{2} = \frac{k(k+5)}{4}$$

which gives $S_k \geq S_0 + \frac{k(k+5)}{4} = \frac{(k+1)(k+4)}{4}$. On the other hand, using (49) yields that

$$\frac{S_{k+1}}{S_k} = \frac{\frac{1}{S_k} + 2 + r + \sqrt{\left(\frac{1}{S_k} + (2+r)\right)^2 - 4}}{2} \geq \frac{2 + r + \sqrt{(2+r)^2 - 4}}{2} = \frac{2 + r + \sqrt{r^2 + 4r}}{2} \tag{51}$$

for all $k \geq 0$. Hence, we have $S_k \geq S_0\left(\frac{2+r+\sqrt{r^2+4r}}{2}\right)^k = \left(\frac{2+r+\sqrt{r^2+4r}}{2}\right)^k$. □

Remark. The linear convergence factor $\frac{2}{2+r+\sqrt{r^2+4r}}$ in the above lemma satisfies

$$1 - \sqrt{\frac{r}{r+1}} \leq \frac{2}{2 + r + \sqrt{r^2 + 4r}} \leq \left(1 + \frac{1}{2}\sqrt{r}\right)^{-2}.$$

In fact, since

$$\left(1 - \sqrt{\frac{r}{r+1}}\right)^{-1} = \frac{\sqrt{r+1}}{\sqrt{r+1} - \sqrt{r}} = \sqrt{r+1}(\sqrt{r+1} + \sqrt{r}) = \frac{2 + 2r + \sqrt{4r^2 + 4r}}{2},$$

we obtain

$$\left(1 + \frac{1}{2}\sqrt{r}\right)^2 = \frac{2 + r/2 + \sqrt{4r}}{2} \leq \frac{2 + r + \sqrt{r^2 + 4r}}{2} \leq \frac{2 + 2r + \sqrt{4r^2 + 4r}}{2} = \left(1 - \sqrt{\frac{r}{r+1}}\right)^{-1}.$$

Note that if $\bar{\sigma}_f = \sigma_f$ and $r = \frac{\sigma_f \sigma_d}{L - \bar{\sigma}_f \sigma_d}$, then $\sqrt{\frac{r}{r+1}} = \sqrt{\frac{\sigma_f \sigma_d}{L}}$.

Lemma A.2. *The sequence $\{S_k\}_{k\geq 0}$ defined by (48) for $r>0$ satisfies*

$$\frac{\sum_{i=0}^k S_i}{S_k} \leq \frac{1+\sqrt{1+4r^{-1}}}{2} \leq 1+\sqrt{\frac{1}{r}}, \quad \forall k \geq 0.$$

Proof. Notice that $\gamma := \frac{1+\sqrt{1+4r^{-1}}}{2}$ satisfies

$$\left(1-\frac{1}{\gamma}\right)^{-1} = \frac{\gamma}{\gamma-1} = \frac{\sqrt{1+4r^{-1}}+1}{\sqrt{1+4r^{-1}}-1} = \frac{(\sqrt{1+4r^{-1}}+1)^2}{4r^{-1}} = \frac{2+r+\sqrt{r^2+4r}}{2}.$$

Therefore, we obtain $\frac{S_k}{S_{k+1}} \leq 1-\frac{1}{\gamma}$ by (51). Now the result follows by induction: If $\sum_{i=0}^k S_i/S_k \leq \gamma$ holds for some $k\geq 0$, we have

$$\frac{\sum_{i=0}^{k+1} S_i}{S_{k+1}} = 1+\frac{S_k}{S_{k+1}}\frac{\sum_{i=0}^k S_i}{S_k} \leq 1+\frac{\gamma-1}{\gamma}\cdot\gamma = \gamma.$$

This proves the first inequality; the second can be verified from $\sqrt{1+4r^{-1}} \leq 1+2\sqrt{r^{-1}}$. □

Note that the result of Lemma A.2 is the same as [11, Lemma 5] because $1+\frac{2\sqrt{r^{-1}}}{\sqrt{r}+\sqrt{r+4}} = \frac{1+\sqrt{1+4r^{-1}}}{2}$.

Lemma A.3. *Let $\{S_k\}_{k\geq 0}$ be defined as Lemma A.2 and $\{T_k\}_{k\geq 0}$ be defined by (48) with $r:=0$, namely $T_0 := 1$ and $T_{k+1} := \frac{1+2T_k+\sqrt{1+4T_k}}{2}$ for $k\geq 0$. Then, we have*

$$\frac{\sum_{i=0}^k S_i}{S_k} \leq \frac{\sum_{i=0}^k T_i}{T_k}, \quad \forall k \geq 0.$$

Proof. Due to the identity

$$\frac{\sum_{i=0}^k S_i}{S_k} = 1+\sum_{i=0}^{k-1}\frac{S_i}{S_k} = 1+\sum_{i=0}^{k-1}\prod_{j=i}^{k-1}\frac{S_j}{S_{j+1}}, \quad k\geq 0,$$

it is enough to show that $\frac{S_k}{S_{k+1}} \leq \frac{T_k}{T_{k+1}}$ for every $k\geq 0$. Notice that we have

$$\frac{S_{k+1}}{S_k} = \frac{\frac{1+rS_k}{S_k}+2+\sqrt{\left(\frac{1+rS_k}{S_k}+2\right)^2-4}}{2}, \quad \frac{T_{k+1}}{T_k} = \frac{\frac{1}{T_k}+2+\sqrt{\left(\frac{1}{T_k}+2\right)^2-4}}{2}, \tag{52}$$

which suggests us to prove $\frac{1+rS_k}{S_k} \geq \frac{1}{T_k}$ for $k\geq 0$. It is true for $k=0$ by $S_0=T_0$. If it holds for $k\geq 0$, then, writing $\alpha := \frac{1+rS_k}{S_k} \geq \beta := \frac{1}{T_k}$, we obtain

$$\begin{aligned}\frac{1+rS_{k+1}}{S_{k+1}} &\geq \frac{1+rS_k}{S_{k+1}} = \frac{S_k}{S_{k+1}}\alpha \overset{(52)}{=} \frac{2\alpha}{\alpha+2+\sqrt{(\alpha+2)^2-4}}\\ &\geq \frac{2\beta}{\beta+2+\sqrt{(\beta+2)^2-4}} \overset{(52)}{=} \frac{T_k}{T_{k+1}}\beta = \frac{1}{T_{k+1}}\end{aligned}$$

since $S_{k+1}\geq S_k$ and $x \mapsto \frac{2x}{x+2+\sqrt{(x+2)^2-4}} = \frac{2}{1+2x^{-1}+\sqrt{1+4x^{-1}}}$ is non-decreasing on $(0,\infty)$. Hence, we claim $\frac{1+rS_k}{S_k} \geq \frac{1}{T_k}$ for all $k\geq 0$ and therefore the proof is completed.

□

Lemma A.4. *Let $\{T_k\}_{k\geq 0}$ be a sequence defined by (48) with $r := 0$, namely $T_0 := 1$ and $T_{k+1} := \frac{1+2T_k+\sqrt{1+4T_k}}{2}$ for $k \geq 0$. Then, we have*

$$\frac{\sum_{i=0}^{k} T_i}{T_k} \leq \frac{1}{3}k + \frac{1}{6}\log(k+2) + 1, \quad \forall k \geq 0.$$

Proof. The case $k = 0$ is obvious. Assume that the assertion is true for some $k \geq 0$. Putting $U_k := \frac{1}{3}k + \frac{1}{6}\log(k+2) + 1$, we have

$$\frac{\sum_{i=0}^{k+1} T_i}{T_{k+1}} = 1 + \frac{T_k}{T_{k+1}} \frac{\sum_{i=0}^{k} T_i}{T_k} \leq 1 + \frac{T_k}{T_{k+1}} U_k.$$

Hence, it reminds to show $1 + \frac{T_k}{T_{k+1}} U_k \leq U_{k+1}$ for $k \geq 0$. For that, we analyze the sequence $t_0 := 1$, $t_{k+1} := T_{k+1} - T_k$ for $k \geq 0$ (namely, $T_k = \sum_{i=0}^{k} t_i$). The recurrence relation of T_k implies $t_k^2 = (T_k - T_{k-1})^2 = T_k$ and

$$t_{k+1} = T_{k+1} - T_k \overset{(49)}{=} \frac{1+\sqrt{1+4T_k}}{2} = \frac{1+\sqrt{1+4t_k^2}}{2}, \quad \forall k \geq 0.$$

Analyzing the difference $t_{k+1} - t_k$ shows for $k \geq 0$ that

$$t_{k+1} - t_k = \frac{1+\sqrt{1+4t_k^2} - 2t_k}{2} = \frac{1}{2} + \frac{1}{2\left(\sqrt{1+4t_k^2} + 2t_k\right)} \leq \frac{1}{2} + \frac{1}{2\left(\sqrt{4t_k^2} + 2t_k\right)} = \frac{1}{2} + \frac{1}{8t_k}.$$

Since Lemma A.1 yields $t_k = \sqrt{T_k} \geq \sqrt{(k+1)(k+4)/4} \geq (k+2)/2$ for $k \geq 0$, we obtain

$$t_{k+1} \leq t_0 + \frac{k+1}{2} + \frac{1}{8}\sum_{i=0}^{k} \frac{1}{t_i} \leq \frac{k}{2} + \frac{3}{2} + \frac{1}{8}\sum_{i=0}^{k} \frac{2}{i+2} \leq \frac{k}{2} + \frac{3}{2} + \frac{1}{4}\log(k+2) = \frac{3}{2}U_k$$

for all $k \geq 0$. Finally, this upper bound of t_k concludes that

$$\frac{U_k}{1 + U_k - U_{k+1}} = \frac{3U_k}{2 + \frac{1}{2}\log\frac{k+2}{k+3}} \geq \frac{3}{2}U_k \geq t_{k+1} = \frac{t_{k+1}^2}{t_{k+1}} = \frac{T_{k+1}}{T_{k+1} - T_k}.$$

Taking the inverse and multiplying by U_k for both sides yield $1 + \frac{T_k}{T_{k+1}} U_k \leq U_{k+1}$. □