論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	A Unifying Framework of Subgradient-Based Methods for Structured Convex Optimization Problems
著者(和文)	伊藤勝
Author(English)	Masaru Ito
出典(和文)	学位:博士(理学), 学位授与機関:東京工業大学, 報告番号:甲第10617号, 授与年月日:2017年9月20日, 学位の種別:課程博士, 審査員:福田 光浩,小島 定吉,三好 直人,山田 功,山下 真,鈴木 大慈
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第10617号, Conferred date:2017/9/20, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

A Unifying Framework of Subgradient-Based Methods for Structured Convex Optimization Problems

by

Masaru Ito

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Science

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

April 2017

To my parents

Acknowledgements

First of all, I would like to express my deep gratitude to my supervisor Prof. Mituhiro Fukuda. He have given me many instructive advises and opened valuable discussions since I belonged his laboratory and started to learn operations research. None of them can be lacked to arrive this thesis.

I would also like to thank Prof. Makoto Yamashita for having OR seminary and giving opportunities to learn other things from computer science. Moreover, I am grateful for having discussions with Prof. Taiji Suzuki who helped my view of machine learning aspects of first-order methods.

I would like to express a big thank to my supervisor in my bachelor, Prof. Noriko Hirata-Kohno, who inspired me the spirit of mathematics and gave me many constant supports.

During my research, I received insightful comments on related works. Especially, I am thankful to Prof. Nobuo Yamashita for indicating a relation with Tseng's work [58], to Prof. Yurii Nesterov for giving a comment on our approach and notifying a related work [56], to Prof. Guanghui Lan for pointing out a relation with his CGM [35]. I also thank Prof. Robert Freund and Prof. Paul Grigas for having valuable discussion.

My special thank goes to my colleague, Bruno Lourenço. I am really grateful to him for having constant discussions on optimization, mathematics, and many others. I also thank all the members of Fukuda lab. for sharing precious research time.

Finally, I would like to thank my family for their support and understanding throughout my study.

Masaru Ito, April 2017.

Contents

Symbols and Notations

APG	—	Accelerated Proximal Gradient
CGM	—	Conditional Gradient Method (Section 2.3.1)
$\operatorname{cl} A$	—	the closure of the set A in $(E, \|\cdot\|)$
$\operatorname{core} A$	—	the set $\{x \in A \mid \forall d \in E,\ \exists t_d > 0 \text{ s.t. } \forall t \in [0, t_d],\ x + td \in A\}$
$\mathcal{F}_M^\nu(Q)$	—	The class of convex functions satisfying the 'Hölder condition' on Q with coefficient M and exponent ν (Definition 2.1.5)
DA, DAM	—	Dual-Averaging, Dual-Averaging Method (Section 3.1.2)
$D_f(y, x)$	—	the Bregman distance associated with the function f between x and y (2.1.10)
$\operatorname{Diam}(Q)$	—	the diameter of the set Q in $(E, \|\cdot\|)$, *i.e.*, $\sup_{x,y \in Q} \|x - y\|$
$\operatorname{dom} f$	—	the domain of the function f, *i.e.*, the set $\{x \in E \mid f(x) < +\infty\}$ (2.1.1)
$\operatorname{dom}(\partial f)$	—	the set $\{x \in E \mid \partial f(x) \neq \emptyset\}$ for the function f (2.1.3)
EMD	—	Extended Mirror-Descent
E, $(E, \|\cdot\|)$	—	a finite dimensional real normed space endowed with the norm $\|\cdot\|$
$f'(x; d)$	—	the directional derivative of the function f at $x \in \operatorname{dom} f$ along $d \in E$ (2.1.4)
$\operatorname{int} A$	—	the interior of the set A in $(E, \|\cdot\|)$
lsc	—	lower semicontinuous
MD, MDM	—	Mirror-Descent, Mirror-Descent Method (Section 3.1.1)
$m_f(y; x)$	—	a lower approximation model of the function f at y (cf. Section 4.2.2)
$\mathcal{NSP}(g, \sigma)$	—	the class of non-smooth problems with gradient mapping g and convexity parameter $\sigma > 0$ (Definition 4.2.6)
PGM	—	Proximal Gradient Method (Section 2.3.1)
$\operatorname{ri} A$	—	the relative interior of the set A in $(E, \|\cdot\|)$
S_k	—	$\sum_{i=0}^k \lambda_i$ for weight parameters $\{\lambda_i\}_{i \geq 0}$
$\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$	—	the class of structured problems (Definition 4.2.7)
$\xi(y, x)$	—	the Bregman distance $D_d(y, x)$ associated with the prox-function d
$\sigma(f)$	—	the set $\{\sigma \geq 0 \mid f - \sigma d \text{ is convex on } Q\}$ where d is a prox-function on Q (4.2.1)
$\|s\|_*$	—	the dual norm of $s \in E^*$ of the norm $\|\cdot\|$ (2.0.1)
$\langle s, x \rangle$	—	the value $s(x)$ of $s \in E^*$ at $x \in E$
∂f	—	the subdifferential of the function f (2.1.2)

Chapter 1

Introduction

1.1 Purpose of the thesis

Subgradient- and gradient-based methods for convex optimization problems have been one of major interests in optimization in the last decades, providing efficient approaches for the recent demands to solve large-scale optimization problems which arise from image/signal processing, data mining, statistics, *etc.* Due to their cheap iteration cost, these methods can be an efficient solution to large-scale optimization rather than Newton-type methods when the desired accuracy for the solutions is moderate.

The (oracle-based) iteration complexity theory [46, 49] established a fundamental measure of efficiency for subgradient-based methods and consequently many methods were demonstrated to be of 'optimal' complexity for various classes of convex optimization problems. The major interest on the subgradient-based methods were paid to the so called *proximal gradient methods (PGMs)* whereas the *conditional gradient methods (CGMs)* have been another new focus in the past few years. The approaches to analyze the existing subgradient-based methods are often different for each method and for each class of problems.

This thesis is devoted to establish a methodology on developing efficient subgradient-based methods, the PGMs and the CGMs, unifying several existing methods. It provides a unifying view of known subgradient-based methods where some of them have different types or were originally analyzed in different ways. The unifying framework also yields new optimal complexity methods which sometimes improve and/or extend existing results. The idea of our unifying framework could be helpful for further developments and analysis of subgradient-based methods. The subsequent sections summarize details on our results.

1.2 Background

1.2.1 Subgradient-based methods

Throughout this thesis, we focus on *convex optimization problems*

$$(P)\begin{cases}\text{minimize} & f(x)\\ \text{subject to} & x\in Q\end{cases}$$

over a finite dimensional normed space $(E,\|\cdot\|)$ with its topological dual space $(E^*,\|\cdot\|_*)$ and the dual pairing $\langle\cdot,\cdot\rangle$. Here f and Q are called the objective function and the feasible set of (P), respectively, and assumed that Q is a closed convex subset of E and $f:E\to\mathbb{R}\cup\{+\infty\}$ is a lower semicontinuous proper convex function with $Q\subset\operatorname{dom} f=\{x\in E\mid f(x)<+\infty\}$. We assume that (P) has at least one optimal solution $x^*\in Q$. Our aim on solving the

problem (P) is to find an ε-solution for (P), *i.e.*, a point $\hat{x} \in Q$ with $f(\hat{x}) - f(x^*) < \varepsilon$, for a given absolute accuracy $\varepsilon > 0$.

A *subgradient-based method* of solving (P) is an iterative scheme generating a sequence of approximate solutions, which basically consists of two main computations at each iteration. The first one is the evaluation of the objective function f at some test point $x \in Q$ up to its first-order information, namely, the value $f(x)$ and a subgradient $g \in \partial f(x) = \{g \in E^* \mid f(y) \geq f(x) + \langle g, y - x \rangle, \ \forall y \in E\}$. Such information can be formulated as an *oracle* for f which plays a crucial role to define the 'complexity' of these methods. Note that if the convex objective function f is Gâteaux differentiable at x, then the only subgradient of f at x is the gradient $\nabla f(x)$. The second main computation is to solve a subproblem of minimizing some *auxiliary function* over the feasible set Q. The difficulty of the subproblem depends on the definition of the auxiliary function which affects the efficiency of the method.

The *iteration (or oracle-based) complexity* of a subgradient-based method for a given absolute accuracy $\varepsilon > 0$ is defined to be the minimal number of evaluations of the oracle in the method to find an ε-solution for (P). The iteration complexity ignores the computational cost per iteration. According to the types of subproblems at each iteration, subgradient-based methods of recent interest can be classified into two types:

- *Proximal Gradient Method (PGM)* solves subproblems of the form $\min_{x \in Q}\{\langle s, x \rangle + d(x)\}$ for some $s \in E^*$ at each iteration. The function $d(x)$ is a Gâteaux differentiable strongly convex function on Q called a prox-function.

- *Conditional Gradient Method (CGM)* solves subproblems of the form $\min_{x \in Q} \langle s, x \rangle$ for some $s \in E^*$ at each iteration. In this case, Q is assumed to be compact in order to ensure the existence of a solution to the subproblems.

The CGMs usually ensures worse iteration complexity than the PGMs while the former has cheaper cost per iteration than the latter, compensating the overall cost.

1.2.2 Classes of convex optimization problems

Given a feasible set Q and a family $\mathcal{F}$ of objective functions, a class of optimization problems $\min_{x \in Q} f(x), \ f \in \mathcal{F}$ is defined. Subgradient-based methods for the following two classes of convex optimization problems were particularly well studied in the literature.

- *Non-smooth problems.* The problems $\min_{x \in Q} f(x)$ of minimizing convex functions with bounded subgradients on Q (or minimizing Lipschitz convex functions). This corresponds to consider the class of convex functions f on Q such that $M_f(Q) := \sup\{\|g\|_* : g \in \partial f(x), \ x \in Q\} < +\infty$.

- *Smooth problems.* The problems $\min_{x \in Q} f(x)$ of minimizing f in the class $\mathcal{F} = \mathcal{F}_L^1(Q)$ of Gâteaux differentiable convex functions with L-Lipschitz continuous gradients on Q, namely, $\|\nabla f(x) - \nabla f(y)\|_* \leq L, \ \forall x, y \in Q$.

As their generalization, the following class of convex optimization problems received attention recently.

- *Weakly smooth problems.* The problems $\min_{x \in Q} f(x)$ with f in the class $\mathcal{F} = \mathcal{F}_M^\nu(Q)$ $(M \geq 0, \ \nu \in [0, 1])$, namely, f is a subdifferentiable convex functions on Q such that $\|g_1 - g_2\|_* \leq M \|x_1 - x_2\|^\nu, \quad \forall x_i \in Q, \ \forall g_i \in \partial f(x_i), \ i = 1, 2$. In the case $\nu \neq 0$, we have the Gâteaux differentiability of f on Q and so this condition is equivalent to

the Hölder condition $\|\nabla f(x_1) - \nabla f(x_2)\|_* \le M \|x_1 - x_2\|^\nu$, $\forall x_1, x_2 \in Q$. Non-smooth problems is included in this class because $f \in \mathcal{F}^0_{2M}(Q)$ holds whenever $M_f(Q) \le M$.

The *strong convexity* of the objective function f is also often concerned because it enables us to construct subgradient-based methods with better iteration complexity. We say f is σ_f-strongly convex on Q (with convexity parameter $\sigma_f \ge 0$) if $f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha)f(y) - \frac{1}{2}\sigma_f \alpha(1-\alpha)\|x-y\|^2$, $\forall x, y \in Q$, $\forall \alpha \in [0,1]$.

Tight lower bounds on the iteration complexity of subgradient-based methods for various classes of convex optimization problems have been established [27, 33, 45, 46, 49]. Table 1.1 summarizes the ones for the non-smooth, the smooth, and the weakly smooth problems when $(E, \|\cdot\|)$ is a Euclidean space. In fact, the tightness are attained by some (optimal) PGMs.

	non-smooth $M = M_f(Q)$	smooth $\mathcal{F}^1_L(Q)$	weakly smooth $\mathcal{F}^\nu_M(Q)$ $(\nu \ne 1)$
non strongly convex $(\sigma_f = 0)$	$\Theta\left(\frac{M^2R^2}{\varepsilon^2}\right)$	$\Theta\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$	$c_1(\nu)\left(\frac{MR^{1+\nu}}{\varepsilon}\right)^{\frac{2}{3\nu+1}}$
σ_f-strongly convex	$\Theta\left(\frac{M^2}{\sigma_f \varepsilon}\right)$	$\Theta\left(\sqrt{\frac{L}{\sigma_f}}\log\frac{1}{\varepsilon}\right)$	$c_2(\nu)\left(\frac{M^2}{\sigma_f^{1+\nu}}\frac{1}{\varepsilon^{1-\nu}}\right)^{\frac{1}{3\nu+1}}$

Table 1.1: Tight lower bounds of the iteration complexity of subgradient-based methods in the Euclidean setting. Here $R := \|x_0 - x^*\|$ for a starting point $x_0 \in Q$ and $c_1(\nu)$ and $c_2(\nu)$ are fixed continuous functions depending only on ν.

On the other hand, the following upper bound of the iteration complexity of existing CGMs can be ensured for the weakly smooth problems $\mathcal{F}^\nu_M(Q)$ (cf. [55]):

$$O\left(\left(\frac{M\mathrm{Diam}(Q)^{1+\nu}}{\varepsilon}\right)^{\frac{1}{\nu}}\right) \quad \text{where} \quad \mathrm{Diam}(Q) := \sup_{x,y\in Q}\|x-y\|.$$

This bound is known to be nearly optimal [27]. In particular, it is optimal if $\nu = 1$ in view of the complexity based on the linear optimization oracle [35].

1.2.3 Previous works

It have been proposed many subgradient-based methods for each classes of convex optimization problems mentioned above. We at first focus on the PGMs. The basic and important PGMs are the *Mirror-Descent Method (MDM)* and the *Dual-Averaging Method (DAM)* for the non-smooth problems. They support the basics of our study in this thesis and are related to many other existing subgradient-based methods.

The *Mirror-Descent Method (MDM)* was proposed by Nemirovski-Yudin [46] and iterates the procedure

$$x_{k+1} := \operatorname*{argmin}_{x\in Q}\Big\{\lambda_k[f(x_k) + \langle g_k, x - x_k\rangle] + \xi(x_k, x)\Big\}, \quad k = 0, 1, 2, \ldots,$$

starting from $x_0 \in Q$ where $g_k \in \partial f(x_k)$, $\lambda_k > 0$ is a weight parameter, and $\xi(y, x) := d(x) - d(y) - \langle \nabla d(y), x - y\rangle$ is the Bregman distance induced by a prox-function $d(x)$, a

Gâteaux differentiable and σ_d-strongly convex function on Q. In the non strongly convex case, the MDM ensures the optimal iteration complexity $O(M^2R^2/\varepsilon^2)$ with fixing the total iteration number and requiring to know upper bounds $R \geq \sqrt{d(x^*)/\sigma_d}$ and $M \geq M_f(Q)$ in advance. If we further assume the boundedness of Q, technical averages [43, 44] of $\{x_k\}$ provide an ε-solution with the iteration complexity $O(1/\varepsilon^2)$ without these requirements (Knowing M and the diameter D of Q further provides the optimal complexity $O(M^2D^2/\varepsilon^2)$).

The *Dual-Averaging Method (DAM)* proposed by Nesterov [52], on the other hand, iterates

$$x_{k+1} := \operatorname*{argmin}_{x \in Q} \left\{ \sum_{i=0}^{k} \lambda_i [f(x_i) + \langle g_i, x - x_i \rangle] + \beta_k d(x) \right\}, \quad k = 0, 1, 2, \ldots, \tag{1.2.1}$$

with the initial point $x_0 \in Q$ where $g_k \in \partial f(x_k)$, $\{\lambda_k\}_{k \geq 0}$ is the positive weight parameters, and $\{\beta_k\}_{k \geq -1}$ is a nondecreasing sequence of positive numbers called the scaling parameters. In contrast to the MDM, the DAM ensures the iteration complexity $O(1/\varepsilon^2)$ *without* fixing the total iteration number and knowing the above upper bounds R and M. Knowing R and M further ensures the optimal complexity $O(M^2R^2/\varepsilon^2)$. The same advantage is also valid for the variants [56] of the DAM.

For the non-smooth problems in the strongly convex case, the MDM ensures the optimal iteration complexity $O(M^2/(\sigma_f \varepsilon))$ [3, 42, 43]. The optimal complexity of the DAM can be achieved if we exploit a multistage procedure [33].

For the smooth problems ($f \in \mathcal{F}_L^1(Q)$), the first optimal PGM was established by Nesterov [47] in 1983. There are several variants [2, 4, 23, 38, 49, 50] and extensive consideration to some *structured problems* such as the composite structure [7, 53, 58, 59], the mixed smoothness structure [12, 25, 26, 34, 37], and the inexact oracle model [17, 18]. Tseng [58, 59] demonstrated a unified treatment of some of these methods separating into three algorithms. In particular, the second and the third Accelerated Proximal Gradient (APG) methods [59] solve quite similar subproblems as the MDM and the DAM, respectively. Furthermore, the Nesterov's modified method [50] can be seen as the hybrid version of the two Tseng's methods because it involves two subproblems in the Tseng's methods at each iteration. The Tseng's APG methods and the Nesterov's modified method ensure the optimal complexity in the non strongly convex case whereas the optimality in the strongly convex case is not known. In the strongly convex case, there are efficient PGMs for the smooth problems [47, 49], the composite structure [15, 26, 53], the mixed smoothness structure [12, 26], and the inexact oracle model [17]. They ensure the optimal complexity for the smooth problems.

For the weakly smooth problems ($f \in \mathcal{F}_M^\nu(Q)$), Nemirovski and Nesterov [45] established optimal PGMs for this class in both the non strongly and the strongly convex cases. In the strongly convex case, their method exploits a multistage procedure requiring a prior knowledge M, ν, a convexity parameter σ_f of f, and an upper bound $R \geq \|x_0 - x^*\|$. Recently, some *adaptive* PGMs [36, 54, 61] were proposed which ensure the optimal complexity in the non strongly convex case without knowing M and ν.

The CGMs, on the other hand, have received great attention in past few years due to its advantage for large-scale optimization [14, 28, 31, 32]. The CGM proposed by Frank and Wolfe [21] is the classical one and convergence properties of CGMs are well analyzed in particular for the smooth problems (see [16, 19, 22, 35, 55, 57] and references therein). It is interesting to see that the recent Lan's CGMs, Algorithms 4 and 5 in [35], have similar fashion to the Tseng's second and third APG methods, respectively, as clarified in this thesis. Some extensive concerns of the CGM to the composite structure [1, 3, 24], the inexact oracle model [22], and the weakly smooth problems [55] were also studied.

1.3 Our work and contributions

In this thesis, we establish a *unifying framework of subgradient-based methods*, that is, Methods I and II endowed with Properties A and B. Our framework provides a family of subgradient-based methods and a methodology to analyze them. As a consequence, several existing methods can be unified with some new extensions and improvements on their convergence properties. The development in this thesis covers the papers [29, 30] by the author and Fukuda.

We emphasis the crucial points of our approach and contributions dividing into three parts.

1.3.1 Unifying subgradient-based methods

Let us see how we unify existing methods. The essential motivation is to construct axiomatic properties (Property A) of the auxiliary functions $\{\varphi_k(x)\}$ in the subproblems $\min_{x\in Q}\varphi_k(x)$ solved at each iteration. For instance, the following construction of the auxiliary functions $\{\varphi_k(x)\}$ which we call the *Dual-Averaging (DA) model* satisfies Property A:

$$\varphi_k(x) := \sum_{i=0}^{k} \lambda_i[f(x_i) + \langle g_i, x - x_i\rangle] + \beta_k d(x) \quad \text{for the same parameters as the DAM (1.2.1)}$$

The DA model enables to handle the dual-averaging type methods such as the DAM, the Tseng's third APG method, and a particular instance of the Lan's CGM (Algorithm 5 in [35]). Moreover, we propose the *Extended Mirror-Descent (EMD) model* defined by $\varphi_{-1}(x) := \beta_{-1}d(x)$ and

$$\varphi_{k+1}(x) := \varphi_k(z_k)+\lambda_{k+1}[f(x_{k+1})+\langle g_{k+1}, x - x_{k+1}\rangle]+\beta_{k+1}d(x)-\beta_k[d(z_k)+\langle \nabla d(z_k), x - z_k\rangle]$$

where $z_k := \operatorname{argmin}_{x\in Q}\varphi_k(x)$. The EMD model satisfies Property A and permit to deal with the MDM, the Tseng's second APG method, and a particular instance of the Lan's CGM (Algorithm 4 in [35]). We also propose an extension, Property B, of Property A involving two kinds of auxiliary functions to handle the Nesterov's modified gradient method.

We proceed all of our analysis under Properties A and B. We propose two general subgradient-based methods (Methods I and II) and show their analysis under the properties. As a consequence, for each particular classes of convex optimization problems, we obtain a family of subgradient-based methods and their convergence estimates. Table 4.1 in Section 4.3.4 summarizes existing methods yielded from the proposed method.

We remark that Tseng [58, 59] already established a unified treatment on existing PGMs. Our unifying framework differs from Tseng's work in view of the following two points. First, the Tseng's second and third APG methods (unifying several known methods) require independent convergence analysis while we absorbed their difference via Property A. Second, our framework additionally includes not only PGMs but also CGMs. Moreover, we also focus on the weakly smooth problems and/or the strongly convex case which in particular generalize the original Tseng's APG methods.

It will turn out in Section 5.1 that Property A has a close relation to the Nesterov's *estimate sequence* framework [48, 49] which is a powerful principle for the construction of optimal PGMs especially for the smooth problems. Our approach further covers other types of convex optimization problems (*e.g.*, the non-smooth problems) and other types of methods (namely, the CGMs).

1.3.2 New convergence analysis and extended methods for non-smooth problems

We propose Method I for the non-smooth problems defined in the previous section. We prove that the PGMs yielded by Method I achieve the optimal iteration complexity (see Table 1.1).

In the non strongly convex case, we prove that Method I ensures the iteration complexity $O(1/\varepsilon^2)$ without fixing the total iteration number and knowing upper bounds $M = M_f(Q)$ and $R \geq \sqrt{d(x^*)/\sigma_d}$. Moreover, the optimal complexity $O(M^2R^2/\varepsilon^2)$ is ensured if we know M and R. In particular, Method I employing the DA model yields the Nesterov's DAM and its variant [56] with the same advantage.

As a byproduct, we obtain a novel extension of the MDM, the *extended MDM* (Method 4.3.5), exploiting the EMD model: Start from $x_0 \in Q$ and iterate

$$x_{k+1} := \operatorname*{argmin}_{x \in Q} \{\lambda_k [f(x_k) + \langle g_k, x - x_k \rangle] + \beta_k d(x) - \beta_{k-1} \ell_d(z_{k-1}; x)\}, \ k = 0, 1, 2, \ldots.$$

The original MDM is obtained by taking $\beta_k \equiv 1$. However, strategic choices of $\{\beta_k\}$ permit the same advantage as the DAM improving the original drawbacks on the MDM. In particular, in contrast to known averaging techniques [43, 44], the iteration complexity $O(1/\varepsilon^2)$ can be guaranteed even if the feasible set Q is *unbounded*.

We also show that Method I achieves the optimal iteration complexity in the strongly convex case. This yields a new strongly convex extension of the DAM which ensures the optimality without employing a multistage procedure in contrast to [33]. Moreover, it is interesting to see that Method I with the EMD model in this case yields the results of the Nedić-Lee's averaging [43] and the Bach's variant [3].

1.3.3 New convergence analysis and extended methods for structured problems

We also propose Method II for solving the *structured problems* which include the smooth and the weakly smooth problems as well as their extension such as the composite structure, the mixed smoothness structure, and the inexact oracle model. The class of structured problems is basically a generalization of the inexact oracle model which permits new analysis of PGMs and CGMs for the weakly smooth problems and the mixed smoothness structure.

Method II consists of two kinds of gradient-based methods, the *classical* and the *modified methods*. The classical method with the DA and the EMD models yields the *primal* and the *dual gradient methods* [17, 18, 53], respectively. The modified method includes the three particular PGMs: the Tseng's second APG method via the EMD model, the Tseng's third APG method via the DA model, and the Nesterov's modified method via their hybrid (which satisfies Property B). Moreover, the modified method with the EMD and DA models also includes particular instances of the two Lan's CGMs mentioned in the previous section. This explains a relation between the Tseng's APG methods and the Lan's CGMs. In fact, their general convergence analysis will be proceeded in a unified way.

The classical method of Method II is analyzed for the smooth/composite problems or the inexact oracle model. We prove its rate of convergence recovering results for the primal and dual gradient methods [17, 18, 53]. In the strongly convex case, our convergence estimate has a better linear convergence factor compared to [53] and a better coefficient factor compared to [17].

The modified method of Method II applied to the inexact oracle model yields a slight improvement on the convergence estimate by Devolder *et al.* [17]. Moreover, our result applied

to the smooth or the composite problems yields the same optimality result as the Nesterov's accelerated method [53]. As a consequence, we obtain strongly convex extensions of the above mentioned three particular PGMs of the modified method.

The modified method of Method II is also analyzed for the weakly smooth problems and the mixed smoothness structure. In particular, for weakly smooth problems we obtain new optimal PGMs for strongly convex case with less prior requirements than the existing method [45]. We also provide an analysis of CGMs including the Lan's CGM for the weakly smooth problems leading the same iteration complexity as the other CGMs [55]. We remark that our result in the *non* strongly convex case is less practical because the attainment of the optimality requires parameters in the Hölder condition while some recent *adaptive* PGMs [36, 54, 61] are freed from this requirement.

1.4 Outline of the thesis

This thesis is organized as follows.

We start Chapter 2 by fundamentals of convex analysis and optimization in particular for first-order methods. The prepared materials on the convex analysis are simple. For instance, we do not deal with duality theory of convex functions. Important notions are the strong convexity (with the notion of the Bregman distance), the Lipschitz functions, and the class $\mathcal{F}_M^\nu(Q)$ of weakly smooth functions, introduced in Section 2.1. The oracle-based complexity theory is described in Section 2.3. Then, important classes of convex optimization problems and their known complexity results are shown in Section 2.4.

Chapter 3 reviews existing PGMs and CGMs. The methods which will be unified in the next chapter are described in detail remarking their known convergence results. We focus on the MDM, the DAM, and variants of the DAM for the non-smooth problems in Section 3.1. In Section 3.2, we review gradient-based methods. In particular, we describe the primal and dual gradient methods [18, 53], the Tseng's APG methods, the Nesterov's modified method, and the Lan's CGMs as unified in Chapter 4.

Chapter 4 is the main part of the thesis. In Section 4.2, we at first introduce the notion of the strong convexity with respect to the prox-function, which generalizes the canonical strong convexity in the Euclidean setting. Using this notion, we define two classes of convex optimization problems, namely, the classes of the non-smooth and the structured problems.

Sections 4.3 and 4.4 are the core of the unifying framework. We introduce Properties A and B for auxiliary functions and we then propose Methods I and II. We show that several subgradient-based methods arises as particular instances of the proposed methods (Section 4.3.4). We also propose Method 4.3.5 as a novel extension of the MDM. Section 4.4 demonstrates a unified analysis of the proposed methods concluding general convergence estimates.

Finally, our general convergence estimate is used to establish optimal complexity results (or nearly optimal ones for CGMs) of the proposed methods for particular classes of convex optimization problems, namely, the non-smooth problems (Section 4.5), the smooth/composite problems as well as the inexact oracle model (Section 4.6), and the weakly smooth problems (Section 4.7). We compare our results with known ones reviewed in Chapter 3.

Chapter 5 discusses final concluding remarks on the unifying framework. Section 5.1 indicates a relationship with our unifying framework and the Nesterov's estimate sequence technique. We remark further considerable research directions in Section 5.2.

Chapter 2

Basic Theory

We collect basic tools of convex analysis and optimization, in Sections 2.1 and 2.2, necessary to our development in the subsequent chapters. See, *e.g.*, [5, 60] for fundamentals of convex analysis. In Section 2.3, we introduce the notion of oracle-based methods and their iteration complexity. We also define the proximal and the conditional gradient methods. Then, we review important classes of convex optimization problems and known iteration complexity results in Section 2.4.

Throughout the thesis, E denotes a finite dimensional real normed space endowed with the norm $\|\cdot\|$. The dual space of E is denoted by E^* equipped with the dual norm $\|\cdot\|_*$ on E^*:

$$\|s\|_* := \sup_{\|x\|\leq 1} \langle s, x\rangle \tag{2.0.1}$$

where $\langle s, x\rangle$ is the value of the functional $s \in E^*$ at $x \in E$.

For a set $A \subset E$, the interior, the closure, and the relative interior of A are denoted by $\operatorname{int} A$, $\operatorname{cl} A$, and $\operatorname{ri} A$, respectively. We denote the *core of* A by $\operatorname{core} A = \{x \in A \mid \forall d \in E,\ \exists t_d > 0 \text{ s.t. } \forall t \in [0, t_d],\ x + td \in A\}$.

2.1 Convex functions

A *convex set* is a set $C \subset E$ which satisfies $\alpha x + (1-\alpha)y \in C$ for all $x, y \in C$ and $\alpha \in [0, 1]$.

For a function $f : E \to \mathbb{R} \cup \{+\infty\}$, we define the *domain of* f by

$$\operatorname{dom} f := \{x \in E : f(x) < +\infty\}. \tag{2.1.1}$$

We call f *proper* if $\operatorname{dom} f \neq \emptyset$ (namely, $f \not\equiv +\infty$). A function $f : E \to \mathbb{R} \cup \{+\infty\}$ is said to be *convex* if we have

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y), \quad \forall x, y \in E,\ \forall \alpha \in [0, 1].$$

We say f is *convex on a convex set* Q if f is convex and $Q \subset \operatorname{dom} f$.

The *subdifferential of* $f : E \to \mathbb{R} \cup \{+\infty\}$ *at* $x \in E$ is defined by

$$\partial f(x) := \{g \in E^* : f(y) \geq f(x) + \langle g, x - y\rangle,\ \forall y \in E\} \tag{2.1.2}$$

and each element of $\partial f(x)$ is called a *subgradient of* f *at* $x \in E$. We say that f *is subdifferentiable at* $x \in E$ if $\partial f(x) \neq \emptyset$. The set of all points at where f is subdifferentiable is denoted by

$$\operatorname{dom}(\partial f) := \{x \in E \mid \partial f(x) \neq \emptyset\}. \tag{2.1.3}$$

For a function $f : E \to \mathbb{R} \cup \{+\infty\}$, the *directional derivative of* f *at* $x \in \operatorname{dom} f$ *along* $d \in E$ is defined by

$$f'(x;d) := \lim_{\alpha \downarrow 0} \frac{f(x+\alpha d) - f(x)}{\alpha} \tag{2.1.4}$$

if the limit exists.

Let f be a convex function. For $x \in \operatorname{dom} f$, the directional derivative $f'(x;d)$ exists for every $d \in E$ and the followings hold [60, Theorem 23.1]:

$$f'(x;d) = \inf_{\alpha > 0} \frac{f(x+\alpha d) - f(x)}{\alpha}, \tag{2.1.5}$$

$$f'(x;d) \geq -f'(x;-d), \quad \forall d \in E. \tag{2.1.6}$$

Taking $d = y - x$ and $\alpha = 1$ in the inequality (2.1.5) yields

$$f(y) \geq f(x) + f'(x; y-x), \quad \forall x \in \operatorname{dom} f, \ \forall y \in E. \tag{2.1.7}$$

We say that f is *(Gâteaux) differentiable at* $x \in \operatorname{dom} f$ if there exists $\nabla f(x) \in E^*$, *the gradient of* f *at* x, such that $f'(x;d) = \langle \nabla f(x), d \rangle$ holds for all $d \in E$. It is important to note that a convex function f is differentiable at $x \in \operatorname{dom} f$ if and only if $\partial f(x)$ is a singleton; at the same time, we have $\partial f(x) = \{\nabla f(x)\}$ [60, Theorem 25.1].

Lower semicontinuity

We say that a function $f : E \to \mathbb{R} \cup \{+\infty\}$ is *lower semicontinuous* (*lsc*, for short) at a point $x \in E$ if we have

$$f(x) \leq \liminf_{y \to x} f(y) := \sup_{\varepsilon > 0} \inf_{\|y-x\| < \varepsilon} f(y).$$

f is said to be *lower semicontinuous on* S for a subset $S \subset E$ if f is lsc at every point in S. It is well-known that the following conditions are equivalent for any function $f : E \to \mathbb{R} \cup \{+\infty\}$ [60, Theorem 7.1]:

(i) f is lsc on E,

(ii) the level set $\{x \in E \mid f(x) \leq \alpha\}$ is closed for every $\alpha \in \mathbb{R}$,

(iii) the epigraph $\{(x,t) \in E \times \mathbb{R} \mid f(x) \leq t\}$ of f is a closed subset of $E \times \mathbb{R}$.

Strong convexity

Let $f : E \to \mathbb{R} \cup \{+\infty\}$ be a convex function, $Q\,(\subset \operatorname{dom} f)$ be a convex set, and σ be a nonnegative real number. We say that f *is strongly convex on* Q *with parameter* σ (or σ*-strongly convex on* Q) if we have

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) - \frac{1}{2}\sigma\alpha(1-\alpha)\|x-y\|^2, \ \forall x, y \in Q, \ \forall \alpha \in [0,1]. \tag{2.1.8}$$

Therefore, f is 0-strongly convex on Q if and only if f is convex on Q.

Proposition 2.1.1. *Let $f : E \to \mathbb{R} \cup \{+\infty\}$ be a convex function and $Q\,(\subset \operatorname{dom} f)$ be a convex set. Then, f is σ-strongly convex on Q if and only if*

$$f(x) \geq f(y) + f'(y; x-y) + \frac{1}{2}\sigma\|x-y\|^2, \quad \forall x, y \in Q. \tag{2.1.9}$$

Proof. (2.1.8) $\Rightarrow$ (2.1.9). Pick $x, y \in Q$. For any $\alpha \in (0,1)$, the definition of the strong convexity of f implies

$$\begin{aligned} f(y) &\geq \frac{f(x+(1-\alpha)(y-x)) - \alpha f(x)}{1-\alpha} + \alpha\frac{\sigma}{2}\|x-y\|^2 \\ &= f(x) + \frac{f(x+(1-\alpha)(y-x)) - f(x)}{1-\alpha} + \alpha\frac{\sigma}{2}\|x-y\|^2. \end{aligned}$$

Taking $\alpha \uparrow 1$, we obtain the inequality in (2.1.9) (Note that the directional derivative $f'(x;\cdot)$ exists since $x \in Q \subset \operatorname{dom} f$).
(2.1.9) $\Rightarrow$ (2.1.8). Let $x, y \in Q$ and $\alpha \in (0,1)$. Set $z = \alpha x + (1-\alpha)y \in Q$. Since $x - z = (1-\alpha)(x-y)$ and $y - z = \alpha(y-x)$, we have

$$\begin{aligned} f(x) &\overset{(2.1.9)}{\geq} f(z) + f'(z; x-z) + \frac{\sigma}{2}\|x-z\|^2 \\ &= f(z) + (1-\alpha)f'(z; x-y) + (1-\alpha)^2\frac{\sigma}{2}\|x-y\|^2 \\ &\overset{(2.1.6)}{\geq} f(z) - (1-\alpha)f'(z; y-x) + (1-\alpha)^2\frac{\sigma}{2}\|x-y\|^2 \end{aligned}$$

and

$$f(y) \overset{(2.1.9)}{\geq} f(z) + f'(z; y-z) + \frac{\sigma}{2}\|y-z\|^2 = f(z) + \alpha f'(z; y-x) + \alpha^2\frac{\sigma}{2}\|y-x\|^2.$$

Thus,

$$\alpha f(x) + (1-\alpha)f(y) \geq f(z) + \left[\alpha(1-\alpha)^2 + (1-\alpha)\alpha^2\right]\frac{\sigma}{2}\|x-y\|^2 = f(z) + \alpha(1-\alpha)\frac{\sigma}{2}\|x-y\|^2.$$

□

Strongly convex functions have a coercivity condition:

Proposition 2.1.2. *Let f be a lsc σ-strongly convex on a convex set Q ($\subset \operatorname{dom} f$) for a positive constant $\sigma > 0$. Then, for any sequence $\{x_k\} \subset Q$ with $\|x_k\| \to +\infty$, we have $f(x_k) \to +\infty$. Therefore, for any $\alpha \in \mathbb{R}$, the level set $\{x \in Q \mid f(x) \leq \alpha\}$ is compact.*

Proof. Since the assertion is clear if $f \equiv +\infty$, suppose that f is proper. Then, there exist $s \in E^*$ and $\beta \in \mathbb{R}$ such that $f(x) \geq \langle s, x\rangle + \beta$ for every $x \in E$ (see [60, Corollary 12.1.2]). Fix a point $\bar{x} \in \operatorname{dom} f$. The strong convexity of f implies

$$f((x+\bar{x})/2) \leq \frac{1}{2}f(x) + \frac{1}{2}f(\bar{x}) - \frac{1}{2}\left(1 - \frac{1}{2}\right)\frac{\sigma}{2}\|x-\bar{x}\|^2.$$

Then, for any $x \in Q$, we have

$$\begin{aligned} f(x) &\geq -f(\bar{x}) + 2f((x+\bar{x})/2) + \frac{\sigma}{4}\|x-\bar{x}\|^2 \\ &\geq -f(\bar{x}) + 2\left(\left\langle s, \frac{x+\bar{x}}{2}\right\rangle + \beta\right) + \frac{\sigma}{4}\|x-\bar{x}\|^2 \\ &\geq -f(\bar{x}) - \|s\|_*\|x+\bar{x}\| - 2\beta + \frac{\sigma}{4}(\|x\| - \|\bar{x}\|)^2 \\ &\geq -f(\bar{x}) - \|s\|_*(\|x\| + \|\bar{x}\|) - 2\beta + \frac{\sigma}{4}(\|x\| - \|\bar{x}\|)^2. \end{aligned}$$

This inequality implies for any $\{x_k\} \subset Q$ with $\|x_k\| \to +\infty$ that $f(x_k) \to +\infty$. □

Bregman distance

Let Q be a convex set and f be a σ-strongly convex function on $Q \subset \operatorname{dom} f$ for a positive constant σ. Suppose that f is differentiable on Q. Now we define a distance-like function, called the *Bregman distance* [11] (or *Bregman divergence*) associated with f between $x, y \in Q$, by

$$D_f(y,x) := f(x) - f(y) - \langle \nabla f(y), x - y \rangle . \tag{2.1.10}$$

The domain of $D_f(y,x)$ is replaced by $\operatorname{ri} Q \times Q$ if f is *essentially smooth relative to* Q, that is, f is differentiable on $\operatorname{ri} Q$ and $\lim_{k\to\infty} \|\nabla f(x_k)\|_* = +\infty$ whenever $\{x_k\} \subset \operatorname{ri} Q,\ x_k \to Q \backslash \operatorname{ri} Q$.

For each $y \in Q$, the function $D_f(y,\cdot)$ is σ-strongly convex on Q with the derivative $\nabla_x D_f(y,x) = \nabla f(x) - \nabla f(y)$. The following invariances are sometimes useful:

$$D_{D_f(z,\cdot)}(y,x) = D_f(y,x), \quad D_{f(\cdot)-\alpha}(y,x) = D_f(y,x), \quad \forall x,y,z \in Q,\ \forall \alpha \in \mathbb{R}. \tag{2.1.11}$$

By Proposition 2.1.1 and the definition of $\nabla f(y)$, we have

$$D_f(y,x) \geq \frac{1}{2}\sigma \|x - y\|^2 , \quad \forall x, y \in Q. \tag{2.1.12}$$

In particular, the equality $D_f(y,x) = 0$ holds if and only if $x = y$. We say D_f *grows quadratically on* Q *with a constant* $A > 0$ if we have

$$D_f(y,x) \leq \frac{1}{2}A \|x - y\|^2 , \quad \forall x, y \in Q. \tag{2.1.13}$$

The following examples of Bregman distance are well-known.

Example 2.1.3.

(1) *Euclidean setting.* Suppose that E is a Euclidean space, the norm $\|\cdot\|$ is induced by its inner product, and $f(x) = \frac{1}{2}\|x\|^2$. Then, for $x, y \in E$, we have

$$D_f(y,x) = \frac{1}{2}\|x - y\|^2 .$$

(2) Let $E = \mathbb{R}^n$ be equipped with the ℓ_1-norm $\|x\| := \sum_{i=1}^n |x^{(i)}|$. Let Δ_n be the unit simplex in $\mathbb{R}^n$, that is, $\Delta_n = \{x = (x^{(1)}, \ldots, x^{(n)}) \in \mathbb{R}^n \mid \sum_{i=1}^n x^{(i)} = 1,\ x^{(i)} \geq 0\}$. The function $f(x) = \sum_{i=1}^n x^{(i)} \log x^{(i)}$ (where $0 \log 0 := 0$) is essentially smooth relative to Δ_n and 1-strongly convex on $\operatorname{ri} \Delta_n$ (see [6, Proposition 5.1]). The Bregman distance associated with f between $x \in \Delta_n$ and $y \in \operatorname{ri} \Delta_n$ is given by

$$D_f(y,x) = \sum_{i=1}^n x^{(i)} \log \frac{x^{(i)}}{y^{(i)}}.$$

□

Lipschitz functions

Let $f : E \to \mathbb{R} \cup \{+\infty\}$ and $Q \subset \operatorname{dom} f$. We say that f is *L-Lipschitz on* Q with constant $L \geq 0$ if

$$|f(x) - f(y)| \leq L \|x - y\| , \quad \forall x, y \in Q.$$

The constant L is called a *Lipschitz constant* of f on Q. The following fact shows a relation between the least Lipschitz constant and the greatest norm of subgradients.

Proposition 2.1.4. *Let $f : E \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function. Denote*

$$L_f(Q) := \sup_{\substack{x,y \in Q \\ x \neq y}} \frac{|f(x) - f(y)|}{\|x - y\|}, \quad M_f(Q) := \sup\{\|g\|_* \mid x \in Q,\ g \in \partial f(x)\}. \tag{2.1.14}$$

(i) For any set $Q \subset \operatorname{dom} f$, we have $M_f(\operatorname{core} Q) \leq L_f(Q)$.

(ii) For any set $Q \subset \operatorname{dom}(\partial f)$, we have $L_f(Q) \leq M_f(Q)$, that is, f is $M_f(Q)$-Lipschitz on Q.

In particular, we have $L_f(\operatorname{int}(\operatorname{dom} f)) = M_f(\operatorname{int}(\operatorname{dom} f))$.

Proof. (i) Take $x \in \operatorname{core} Q$ and $g \in \partial f(x)$. By the definition of the dual norm, it suffices to show

$$\langle g, y \rangle \leq L_f(Q), \quad \forall y \in E,\ \|y\| \leq 1.$$

For any $y \in E$ with $\|y\| \leq 1$, there exists $t > 0$ such that $x + ty \in Q$ (because $x \in \operatorname{core} Q$). Then,

$$\langle g, y \rangle = \frac{1}{t} \langle g, (x + ty) - x \rangle \leq \frac{f(x + ty) - f(x)}{t} \leq \frac{L_f(Q) \|(x + ty) - x\|}{t} = L_f(Q) \|y\| \leq L_f(Q).$$

(ii) Take $x, y \in Q \subset \operatorname{dom}(\partial f)$ and $g \in \partial f(x)$. Then,

$$f(x) - f(y) \leq \langle g, x - y \rangle \leq \|g\|_* \|x - y\| \leq M \|x - y\|.$$

Similarly, $f(y) - f(x) \leq L \|x - y\|$ follows by taking $g' \in \partial f(y)$. □

Hölder condition: Class $\mathcal{F}_M^\nu(Q)$

Here we introduce a class of convex functions satisfying the 'Hölder condition' on which we develop a complexity theory and efficient first-order methods.

Definition 2.1.5. Let $f : E \to \mathbb{R} \cup \{+\infty\}$. We say that f *belongs to the class $\mathcal{F}_M^\nu(Q)$ for coefficient $M \geq 0$ and exponent $\nu \in [0, 1]$* if f is a convex function with $Q \subset \operatorname{dom}(\partial f)$ such that

$$\|g_1 - g_2\|_* \leq M \|x_1 - x_2\|^\nu, \quad \forall x_i \in Q,\ \forall g_i \in \partial f(x_i). \tag{2.1.15}$$

When $\nu = 0$, the condition (2.1.15) becomes

$$\|g_1 - g_2\|_* \leq M,\ \forall x_i \in Q,\ \forall g_i \in \partial f(x_i).$$

For instance, if $M := M_f(Q)$ defined in (2.1.14) is finite, then we have $f \in \mathcal{F}_{2M}^0(Q)$ because $\|g_1 - g_2\|_* \leq \|g_1\|_* + \|g_2\|_* \leq 2M$ for any $x_i \in Q,\ g_i \in \partial f(x_i)$.

When $\nu > 0$, the condition (2.1.15) implies the differentiability of f on Q since $\partial f(x)$ becomes a singleton for all $x \in Q$. In this case, the class $\mathcal{F}_M^\nu(Q)$ is the one of differentiable convex functions on Q satisfying the *Hölder condition*:

$$\|\nabla f(x) - \nabla f(y)\|_* \leq M \|x - y\|^\nu, \quad \forall x, y \in Q. \tag{2.1.16}$$

The *Lipschitz condition* $\|\nabla f(x) - \nabla f(y)\|_* \leq M \|x - y\|$ $(x, y \in Q)$ is a special case $\nu = 1$ in the Hölder condition.

Proposition 2.1.6. *Let f be a convex function belongs to the class $\mathcal{F}_M^\nu(Q)$ for $M \geq 0$ and $\nu \in [0,1]$. Then, we have*

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

Proof. When $\nu = 0$, we have for any subgradient $g_x \in \partial f(x)$ that

$$f(x) - f(y) \leq -\langle g_x, y-x\rangle = \langle g_y, x-y\rangle + \langle g_x - g_y, x-y\rangle \leq \langle g_y, x-y\rangle + M\|x-y\|.$$

Here, the last inequality follows from the fact $\langle s, x\rangle \leq \|s\|_* \|x\|$ and the condition (2.1.15).

When $\nu > 0$, on the other hand, $f(x)$ is differentiable on Q and thus

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x-y\rangle &= \int_0^1 \langle \nabla f(y+\tau(x-y)) - \nabla f(y), x-y\rangle\, d\tau \\ &\leq \int_0^1 \|\nabla f(y+\tau(x-y)) - \nabla f(y)\|_* \|x-y\|\, d\tau \\ &\leq \int_0^1 M\tau^\nu \|x-y\|^{1+\nu} d\tau = \frac{M}{1+\nu}\|x-y\|^{1+\nu}. \end{aligned}$$

□

2.2 Convex optimization problems

Throughout this thesis, we focus on the *convex optimization problems* which is formally written by

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in Q \end{array} \tag{2.2.1}$$

or, simply, by $\min_{x\in Q} f(x)$, where $f : E \to \mathbb{R}\cup\{+\infty\}$ is a lsc convex function called the *objective function* and $Q \subset \operatorname{dom} f$ is a closed convex set called the *feasible set*. The objective of the problem (2.2.1) is to find the *optimal value* $\inf_{x\in Q} f(x)$ and/or an *optimal solution* x^* (the point in Q attaining the optimal value) if it exists. The set of optimal solutions is denoted by

$$\operatorname*{Argmin}_{x\in Q} f(x) := \{x^* \in Q : f(x) \geq f(x^*),\ \forall x \in Q\}.$$

We use the notation $\operatorname{argmin}_{x\in Q} f(x)$ as an (arbitrary) element in $\operatorname{Argmin}_{x\in Q} f(x)$ when the observation is independent of the choice of the element.

For given $\varepsilon > 0$, we say that a point $\hat{x} \in Q$ is an *ε-solution* to the problem (2.2.1) if $f(\hat{x}) - \inf_{x\in Q} f(x) < \varepsilon$. In this thesis, we study algorithms of constructing an ε-solution for given accuracy ε.

The following lemma gives a basic optimality condition of the convex optimization problem (2.2.1).

Proposition 2.2.1. *Let $f : E \to \mathbb{R}\cup\{+\infty\}$ be a proper convex function and Q be a convex subset of* $\operatorname{dom} f$. *Then a point $x^* \in Q$ is an optimal solution to the problem $\min_{x\in Q} f(x)$ if and only if*

$$f'(x^*; x-x^*) \geq 0, \quad \forall x \in Q. \tag{2.2.2}$$

Proof. Suppose that x^* satisfies (2.2.2). Since $x^* \in \operatorname{dom} f$, for every $x \in Q$, we have

$$f(x) \overset{(2.1.7)}{\geq} f(x^*) + f'(x^*; x - x^*) \overset{(2.2.2)}{\geq} f(x^*).$$

Namely, x^* is an optimal solution.

Now let us assume that the point x^* does not admit the condition (2.2.2). Namely, we assume the existence of a point $\bar{x} \in Q$ such that

$$0 > f'(x^*; \bar{x} - x^*) = \lim_{t \downarrow 0} \frac{f(x^* + t(x^* - \bar{x})) - f(x^*)}{t}.$$

Then, for sufficiently small $t \in (0, 1]$ we have $0 > \frac{f(x^*+t(x^*-\bar{x}))-f(x^*)}{t}$, and therefore $f(x^* + t(\bar{x} - x^*)) < f(x^*)$. Thus, x^* is not optimal since $x^* + t(\bar{x} - x^*) \in Q$. □

The compactness of the feasible set in convex optimization problems ensures the existence of an optimal solution as described in the following statements.

Proposition 2.2.2 (Proposition 2.8 in [5])**.** *Let X be a compact topological space. Then a real valued lsc function on X attains its minimum on X.*

Corollary 2.2.3. *Let $f : E \to \mathbb{R} \cup \{+\infty\}$ be a proper lsc convex function and $Q(\subset \operatorname{dom} f)$ be a compact convex set. Then, f attains its minimum on Q.*

The next property of a minimization of a strongly convex function is useful for our development.

Proposition 2.2.4. *Let f be a lsc σ-strongly convex function on a closed convex set $Q(\subset \operatorname{dom} f)$ for a positive constant $\sigma > 0$. Then, f attains a unique minimum x^* on Q. Moreover, we have*

$$f(x) \geq f(x^*) + \frac{\sigma}{2} \|x - x^*\|^2, \quad \forall x \in Q.$$

Proof. Take a point $x_0 \in Q \subset \operatorname{dom} f$. By Proposition 2.1.2, the level set $Q' := \{x \in Q \mid f(x) \leq f(x_0)\}$ is compact. The optimization problem $\min_{x \in Q} f(x)$ is equivalent to the one $\min_{x \in Q'} f(x)$ and thus the existence of an optimal solution x^* is guaranteed by Corollary 2.2.3. The optimal solution is unique because strongly convex functions are strictly convex.

Finally, using Proposition 2.1.1 and the optimality condition (Proposition 2.2.1) prove the assertion. □

2.3 First-order methods

In this section, we introduce the concept of *oracle* which is used to define *iterative methods* and their *iteration complexity*. We prepare the following objects.

- Let $Q \subset E$ be a closed convex set and $\mathcal{F}$ be a family of lsc convex functions. It defines the family of convex optimization problems $\{\min_{x \in Q} f(x) : f \in \mathcal{F}\}$.
- Let $\mathcal{O}$ be a mapping defined for points on $\mathcal{F} \times E$ which we refer an *oracle* for the class $\mathcal{F}$. The oracle $\mathcal{O}$ is said to be *local* if we have $\mathcal{O}(f, x) = \mathcal{O}(g, x)$ whenever $f, g \in \mathcal{F}$ satisfies $f \equiv g$ on a neighborhood of $x \in E$. An important example of an oracle is the *first-order oracle* $\mathcal{O}(f, x) = (f(x), g(x))$ where $g(x) \in \partial f(x)$. Remark that a first-order oracle is not necessarily local if there is a convex function $f \in \mathcal{F}$ such that $\partial f(x)$ is not a singleton.

With the above notations, an *iterative method* for the family $\{\min_{x\in Q} f(x) : f \in \mathcal{F}\}$ associated with the oracle $\mathcal{O}$ is a sequence $\mathcal{M} = \{X_k\}_{k\geq 0}$ of functions X_k (corresponding to k-th iteration); then, for each $f \in \mathcal{F}$, the iterative method $\mathcal{M}$ generates a sequence $\{x_k\} \subset Q$ defined by

$$x_k = X_k((x_i, \mathcal{O}(f, x_i))_{i=0}^{k-1}), \quad k \geq 0$$

where the initial point x_0 is fixed by the iterative method $\mathcal{M}$. In particular, an iterative method associated with a first-order oracle is called a *first-order method*.

We define the *iteration (or oracle-based) complexity* of the method $\mathcal{M}$ for an accuracy $\varepsilon > 0$ by the smallest integer k such that the k-th result of $\mathcal{M}$ is an ε-solution of $\min_{x\in Q} f(x)$ for every $f \in \mathcal{F}$. The *iteration (or oracle-based) complexity* of the class of problems $\{\min_{x\in Q} f(x) : f \in \mathcal{F}\}$ associated with an oracle $\mathcal{O}$ and an accuracy $\varepsilon > 0$ is the least integer k among iterative methods $\mathcal{M}$ which finds an ε-solution within k calls of oracle for every problems $\min_{x\in Q} f(x),\ f \in \mathcal{F}$.

2.3.1 Proximal and conditional gradient methods

The iteration complexity measures the performance of methods by counting the number of calls of an oracle. Therefore, it is independent of the 'cost' of the computation at each iteration. Let us see a basic iteration of first-order methods and observe the cost per iteration. Many of existing first-order methods solving the convex optimization problem $\min_{x\in Q} f(x)$ are variations of the following basic iteration.

1. Obtain the result of the (first-order) oracle at the test point x_k.
2. Construct an *auxiliary function* $\varphi_k(x)$ based on the oracle's answer and the information until the previous iteration.
3. Solve the subproblem $\min_{x\in Q} \varphi_k(x)$ and find a solution $z_k \in \operatorname{Argmin}_{x\in Q} \varphi_k(x)$.
4. Update the next test point x_{k+1} based on the previous results.

The auxiliary function $\varphi_k(x)$ (or the subproblem $\min_{x\in Q} \varphi_k(x)$) will be a kind of approximation of the objective function $f(x)$ (or the problem $\min_{x\in Q} f(x)$) based on the previous information. In some actual methods, the steps 2 and 3 may involve multiple subproblems using several auxiliary functions.

The construction of $\varphi_k(x)$ is an important factor which affects the difficulty of solving the subproblem $\min_{x\in Q} \varphi_k(x)$ at each iteration. According to the construction of auxiliary functions, there are two major kinds of first-order methods.

- *Proximal (sub)Gradient Method (PGM)* is an iterative method which solves subproblems of the form

$$\min_{x\in Q}\{\langle s, x\rangle + d(x)\}, \quad s \in E^* \tag{2.3.1}$$

  at each iteration; the function $d : E \to \mathbb{R} \cup \{+\infty\}$ is called the *prox-function* and assumed to satisfy:

  - d is differentiable and σ_d-strongly convex on Q (with $\sigma_d > 0$), and
  - we have $\min_{x\in Q} d(x) = 0$.

The second condition is not restrictive because we can replace $d(x)$ by $d(x) - \min_{z \in Q} d(z)$ or by $D_d(z, x)$ for arbitrary $z \in Q$; this replacement change neither the convexity parameter σ_d nor the Bregman distance $D_d(y, x)$ (recall (2.1.11)).

It is preferable to choose the prox-function so that the subproblem (2.3.1) can be easily solvable. The choice will depend on the structure of the feasible set. See [50] for some examples.

An illustrative PGM is the *projected subgradient method* for a convex function f on Q in the Euclidean setting: Start from an initial point $x_0 \in Q$ and iterate $x_{k+1} := \pi_Q(x_k - \lambda_k g_k)$ with $g_k \in \partial f(x_k)$ and $\lambda_k > 0$ (a weight parameter). Here $\pi_Q(x) := \operatorname{argmin}_{y \in Q} \|x - y\|_2$ is the orthogonal projection onto Q. The projected subgradient method can be rewritten as

$$x_{k+1} = \operatorname*{argmin}_{x \in Q} \|x - (x_k - \lambda_k g_k)\|_2^2 = \operatorname*{argmin}_{x \in Q} \left\{ f(x_k) + \langle g_k, x - x_k \rangle + \frac{1}{2\lambda_k} \|x - x_k\|_2^2 \right\} \tag{2.3.2}$$

which becomes of the form (2.3.1) with $s = \lambda_k g_k - x_k + x_0$ and the prox-function $d(x) := \frac{1}{2} \|x - x_0\|_2^2$.

- *Conditional Gradient Method (CGM)* is an iterative method which solves a linear optimization of the form

$$\min_{x \in Q} \langle s, x \rangle , \quad s \in E^* \tag{2.3.3}$$

at each iteration. In this case, we assume the boundedness of Q to ensure the existence of an optimal solution of the subproblem. The following CGM, the *classical CGM*, proposed by Frank and Wolfe [21] is the most basic one: For a differentiable convex function f on Q, start from an initial point $x_0 \in Q$ and iterate

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

where $\tau_k \in [0, 1)$.

The CGMs have received great attention in past few years due to its advantage such as the cheap iteration cost and sparsity of the approximate solution (see, *e.g.*, [14, 28, 31, 32]).

A remarkable difference between the above two first-order methods is that the CGMs would have worse iteration complexity than the PGMs while the computational cost of each iteration of the former can be cheaper, compensating the overall cost. Therefore, it is important to choose between the PGM or the CGM depending on the structure of the problem to solve.

The PGM and the CGM can be extended to be *composite-type* when the objective function $f(x)$ has a *composite structure*, as we introduce next. Because our definition does not allow the composite-type PGM/CGM to be a first-order method in general, we refer comprehensively to all these kinds of methods as *(sub)gradient-based methods*.

2.4 Classes of convex optimization problems

Here we collect important classes of the convex optimization problem (2.2.1) below mentioning known subgradient-based methods and their complexity guarantees. We will review further details of some existing methods in Chapter 3.

Non-smooth problems

Consider a convex optimization problem (2.2.1) where the objective function f is subdifferentiable on Q. Let us employ the first-order oracle $\mathcal{O} : x \mapsto (f(x), g(x)),\ g(x) \in \partial f(x)$.

Important (optimal) PGMs are the Mirror-Descent Method (MDM) proposed by Nemirovski-Yudin [46] and the Dual-Averaging Method (DAM) proposed by Nesterov [52]. If we further assume the boundedness condition

$$\|g\|_* \leq M,\ \forall g \in \partial f(x),\ \forall x \in Q \tag{2.4.1}$$

for a constant $M > 0$, the MDM and the DAM for this class of problems have the following iteration complexity:

$$O\left(\frac{M^2 R^2}{\varepsilon^2}\right) \tag{2.4.2}$$

where $R := \sqrt{\frac{d(x^*)}{\sigma_d}}$. This upper bound is *optimal* in the sense that the iteration complexity for this class of problems has the lower bound $\min\{n-1, \Theta(M^2R^2/\varepsilon^2)\}$ due to a worst case analysis (see, *e.g.*, [49, Theorem 3.2.1]).

When the objective function f is further strongly convex on Q with constant $\sigma_f > 0$, the MDM [43] ensures the iteration complexity

$$O\left(\frac{M^2}{\sigma_f \varepsilon}\right) \tag{2.4.3}$$

which is optimal for the strongly convex case [33].

We remark that the class of non-smooth problems is a subclass of weakly smooth problems introduced later.

Smooth problems

Let us consider the class of convex optimization problems $\min_{x \in Q} f(x)$ with objective functions $f \in \mathcal{F}_L^1(Q)$; that is, $f(x)$ is differentiable on Q and $\nabla f(x)$ satisfies the Lipschitz condition $\|\nabla f(x) - \nabla f(y)\|_* \leq L \|x - y\|,\ \forall x, y \in Q$ with constant $L > 0$. Again, we employ the first-order oracle $\mathcal{O} : x \mapsto (f(x), \nabla f(x))$ for this class.

In the Euclidean setting, the smooth problem is the most basic one in the examples here. In this case, the iteration complexity for the smooth problems has the lower bound

$$\Theta\left(\sqrt{\frac{LR^2}{\varepsilon}}\right) \tag{2.4.4}$$

where $R = \|x_0 - x^*\|_2$ (if n is large enough; see [49, Theorem 2.1.7]). The first optimal PGM were established by Nesterov [47] for this case and many other optimal PGMs are known (*e.g.*, [2, 4, 38, 49, 50] and see known methods for extended classes below).

When we further restrict the objective functions to be σ_f-strongly convex for a constant $\sigma_f > 0$, the following lower complexity bound is known [49, Theorem 2.1.13]:

$$\Theta\left(\sqrt{\frac{L}{\sigma_f}} \log \frac{1}{\varepsilon}\right). \tag{2.4.5}$$

Nesterov established optimal PGMs [47, 49, 53] and several other optimal PGMs are known which are available for further extended problems [12, 15, 17, 26].

For CGMs, there are many considerations on the Frank-Wolfe method (see, *e.g.*, [16, 19, 21, 22, 35, 55, 57]) whereas several new variants were investigated recently [24, 35, 55]. They ensure the iteration complexity

$$O\left(\frac{L\mathrm{Diam}(Q)}{\varepsilon}\right) \tag{2.4.6}$$

for the smooth problems where $\mathrm{Diam}(Q) := \sup_{x,y\in Q} \|x-y\|$ (for any norm $\|\cdot\|$). This upper bound is known to be optimal in view of the iteration complexity based on the *linear optimization oracle* [35].

Weakly smooth problems

Consider the convex optimization problems $\min_{x\in Q} f(x)$ with objective functions $f \in \mathcal{F}_M^{\rho-1}(Q)$, $\rho \in [1,2]$. Notice that the above introduced class of problems, the non-smooth and the smooth ones, are subclasses of this case by setting $\nu = 0$ and $\nu = 1$, respectively.

For the weakly smooth problems, Nemirovski and Nesterov [45] (see also [20, Section 2.3]) proposed an optimal PGM with the optimal iteration complexities

$$c_1(\rho)\left(\frac{MR^\rho}{\varepsilon}\right)^{\frac{2}{3\rho-2}} \quad \text{and} \quad c_2(\rho)\left(\frac{M^2}{\sigma^\rho}\frac{1}{\varepsilon^{2-\rho}}\right)^{\frac{1}{3\rho-2}}, \tag{2.4.7}$$

for non strongly and strongly convex cases, respectively, where $R := \sqrt{\frac{d(x^*)}{\sigma_d}}$, $\rho := 1+\nu \in [1,2)$, $c_1(\cdot), c_2(\cdot)$ 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 [54] proposed the PGM, called the *universal gradient method*, for the non strongly convex case which ensures the optimal complexity even if we do not know M and ν, that is, the method adapts these parameters. The works [36, 61] also proposed adaptive PGMs.

The studies [17, 18] of the inexact oracle model are also important. They proposed an optimal method for weakly smooth problems in the non strongly convex case and a sub-optimal one in the strongly convex case (PGMs for uniformly convex functions are also discussed).

There are analysis of CGMs for the weakly smooth problems and the following iteration complexity is known (see [14, Proposition 1.1] and [55])

$$O\left(\left(\frac{M\mathrm{Diam}(Q)^{1+\nu}}{\varepsilon}\right)^{\frac{1}{\nu}}\right). \tag{2.4.8}$$

When E is the vector space $\mathbb{R}^n$ equipped with the ℓ_∞-norm $\|\cdot\| := \|\cdot\|_\infty$, this bound is known to be nearly optimal [27, Corollary 1].

Composite structure

Let us fix a lsc convex function $\Psi(x)$ on Q. Consider an objective function $f(x)$ with a *composite structure*:

$$f(x) := f_0(x) + \Psi(x) \tag{2.4.9}$$

where $f_0(x)$ is differentiable on Q.

As we fixed the function $\Psi(x)$, we consider a special kind of oracle and iterative method. Let us employ the oracle $\mathcal{O} : x \mapsto (f_0(x), \nabla f_0(x))$, which is not necessarily a first-order oracle

for f. We consider a generalization of the PGM and the CGM as follows. A *composite-type PGM* is an iterative method whose iterations involve subproblems of the form

$$\min_{x\in Q}\{\langle s, x\rangle + \alpha d(x) + \Psi(x)\}, \quad s \in E^*, \ \alpha > 0 \tag{2.4.10}$$

while a *composite-type CGM* equips the subproblems

$$\min_{x\in Q}\{\langle s, x\rangle + \Psi(x)\}, \quad s \in E^*. \tag{2.4.11}$$

The smooth problems are included in this class with the case $\Psi(x) \equiv 0$. Another illustrative example is the so called Lasso regularization, *i.e.*, $f_0(x) = \|Ax + b\|_2^2$ and $\Psi(x) = \|x\|_1 = \sum_{i=1}^n |x^{(i)}|$ (for $x = (x^{(i)})_{i=1}^n \in E := \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m$), which arises from image/signal processing, compressed sensing, statistics, and so on. The corresponding subproblem (2.4.10) for PGMs in the Euclidean setting is easily solvable with the cost of $O(n)$. See [10] for examples of composite structure and the solvability of subproblems.

Mathematical fundamentals of the composite structure and the composite-type PGMs were investigated by Fukushima and Mine [23] and also by Nesterov [53] without assuming the convexity of $f_0(x)$.

Under the assumption $f_0(x) \in \mathcal{F}_L^1(Q)$, there are many composite-type PGMs for this problem [7, 23, 53, 58, 59] and they provide the same iteration complexity as the optimal one for the smooth problems in the non strongly convex case. Nesterov [53] further proposed an optimal method for strongly convex composite problems in the Euclidean setting. The PGMs [12, 15, 26] are also applicable to this problem ensuring the optimal complexity.

Nesterov's universal gradient method [54] is a composite-type PGM for the case $f_0 \in \mathcal{F}_M^\nu(Q)$ which ensures the same complexity as the optimal one for the weakly smooth problems in the non strongly convex case.

The smoothing technique proposed by Nesterov [50] and its extension [9] by Beck and Teboulle 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 [51].

Composite-type CGMs were investigated in [1, 3, 24, 55]. A duality relationship to the MDM were shown in [1, 3].

Mixed smoothness

Suppose that $f(x)$ has the form

$$f(x) = \varphi(x) + \psi(x), \quad \varphi(x) \in \mathcal{F}_L^1(Q), \quad \psi(x) \in \mathcal{F}_M^0(Q). \tag{2.4.12}$$

This class of convex functions covers the classes of the non-smooth and the smooth problems with applications as the composite structure. Lan [34, 37] proposed PGMs for this class in the non strongly convex case (with further stochastic settings). The works by Ghadimi and Lan [25, 26] employed a more general assumption

$$f(y) \le f(x) + \langle g(x), y - x\rangle + \frac{1}{2}L\|x - y\|^2 + M\|x - y\|, \quad \forall x, y \in Q, \tag{2.4.13}$$

where $g(x) \in \partial f(x)$ is a subgradient mapping of f (In the original papers [25, 26], this class was considered with the composite structure and a stochastic setting).

In the Euclidean setting, it turns out that the iteration complexity of PGMs for the class of convex functions satisfying (2.4.13) (or (2.4.12)) cannot be better than

$$O\left(\sqrt{\frac{LR^2}{\varepsilon}} + \frac{M^2R^2}{\varepsilon^2}\right) \text{ and } O\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon} + \frac{M^2}{\mu\varepsilon}\right)$$

in the non strongly and strongly convex cases, respectively (again $R := \|x_0 - x^*\|_2$). Optimal PGMs were presented in [12, 25, 26, 34, 37].

Inexact oracle model

In the Euclidean setting $\|\cdot\| = \|\cdot\|_2$, suppose that $f(x)$ is equipped with a *first-order* (δ, L, μ)-*oracle* [17], *i.e.*, for each $y \in Q$, we can compute $(\bar{f}(y), \bar{g}(y)) \in \mathbb{R} \times E^*$ such that

$$\frac{\mu}{2}\|x-y\|_2^2 \le f(x) - (\bar{f}(y) + \langle \bar{g}(y), x-y\rangle) \le \frac{L}{2}\|x-y\|_2^2 + \delta, \quad \forall x \in Q, \tag{2.4.14}$$

where $\delta \ge 0$ and $L \ge \mu \ge 0$.

The inexact oracle model was firstly studied by Devolder *et al.* [18] with $\mu = 0$ and they proposed the classical and the fast (proximal) gradient methods which were extended to the strongly convex case in [17]. A CGM for this model in the case $\mu = 0$ was analyzed by [22].

The inexact oracle model can be applicable to situations when an oracle is computed by an auxiliary optimization problem (*e.g.*, saddle point problems, Augmented Lagrangians, and Moreau-Yoshida regularization; see [18]). Another interesting application is an approximation of the weakly smooth problems via the inexact oracle model. This enables to give an optimal and a nearly optimal PGMs for the weakly smooth problems in the non strongly and the strongly convex cases, respectively [17, 18].

Chapter 3

Subgradient-Based Methods for Convex Optimization Problems

In this chapter, we review particular existing subgradient-based methods for convex optimization problems and discuss their iteration complexities. Many of these methods will be unified in the framework investigated in Chapter 4.

We review details of existing methods for which we have particular interest to compare with the contribution of this thesis. See Section 2.4 for more general backgrounds.

Section 3.1 focuses on the non-smooth problems. We in particular review the mirror-descent method (Section 3.1.1) and the dual-averaging method (Section 3.1.2). Section 3.2 reviews PGMs for the smooth or further structured problems. Staring from the so called classical PGM in Section 3.2.1, we deal with their accelerated methods in Section 3.2.2. We finally focus on existing CGMs in Section 3.2.3.

3.1 Proximal subgradient methods for non-smooth problems

For a closed convex set $Q \subset E$, consider the non-smooth problem

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

where $f(x)$ is a subdifferentiable convex function on Q. We assume that there exists an optimal solution $x^* \in \operatorname{Argmin}_{x \in Q} f(x)$.

In this section, we review two important (optimal) PGMs, the mirror-descent and the dual-averaging methods. We assume that we have a subgradient mapping $g : Q \to E^*$, $g(x) \in \partial f(x)$ and a prox-function $d(x)$ on Q. We denote the Bregman distance associated with d as

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

3.1.1 Mirror-descent method

The Mirror-Descent Method (MDM) is a PGM proposed by Nemirovski and Yudin [46] and reinterpreted by Beck and Teboulle [6] in the form as follow: Generate $\{x_k\}_{k \geq 0} \subset Q$ by $x_0 := \operatorname{argmin}_{x \in Q} d(x)$ and the iteration

$$\begin{aligned} g_k &:= g(x_k) \in \partial f(x_k) \\ z_k &:= \operatorname{argmin}_{x \in Q} \{\lambda_k [f(x_k) + \langle g_k, x - x_k \rangle] + \xi(x_k, x)\} \\ x_{k+1} &:= z_k \end{aligned} \tag{3.1.1}$$

for each $k \geq 0$, where $\lambda_k > 0$ is a *weight parameter*. The notation z_k is redundant here, but it is important to connect with our unifying framework developed in Chapter 4. Notice that, by the definition of the Bregman distance, the computation of z_k reduces to the form of (2.3.1) so that it is a PGM.

The parameter λ_k is also referred to as a *stepsize*; the MDM in the Euclidean setting $\|\cdot\| = \|\cdot\|_2$, $d(x) := \frac{1}{2}\|x - x_0\|_2^2$ yields the projected subgradient method $x_{k+1} := \pi_Q(x_k - \lambda_k g_k)$ in view of (2.3.2) (see also Auslender-Teboulle [2] and Fukushima-Mine [23] for some related works).

The MDM produces the following estimate [6]:

$$\forall k \geq 0, \quad \Delta_k := \frac{\sum_{i=0}^k \lambda_i f(x_i)}{\sum_{i=0}^k \lambda_i} - f(x^*) \leq \frac{\xi(x_0, x^*) + \frac{1}{2\sigma_d}\sum_{i=0}^k \lambda_i^2 \|g_i\|_*^2}{\sum_{i=0}^k \lambda_i}. \tag{3.1.2}$$

It is important to note that, by the convexity of f, the quantity Δ_k provides an approximate solution

$$\hat{x}_k := \frac{\sum_{i=0}^k \lambda_i x_i}{\sum_{i=0}^k \lambda_i} \tag{3.1.3}$$

yielding the estimate $f(\hat{x}_k) - f(x^*) \leq \Delta_k$. We can also obtain the estimate $\min_{0\leq i\leq k} f(x_i) - f(x^*) \leq \Delta_k$. Therefore, let us focus on the right hand side of (3.1.2).

Suppose that $M := \sup\{\|g\|_* : g \in \partial f(x),\ x \in Q\}$ is finite and we know an upper bound $R \geq \sqrt{\frac{1}{\sigma_d}\xi(x_0, x^*)}$. Then, the inequality (3.1.2) yields $\Delta_k \leq \sqrt{2}MR/\sqrt{k+1}$ if we choose the constant weight parameters

$$\lambda_0 = \lambda_1 = \cdots = \lambda_k := \frac{\sqrt{2}\sigma_d R}{M\sqrt{k+1}} \tag{3.1.4}$$

for a fixed $k \geq 0$. In this case, the MDM ensures an ε-solution with at most $O(M^2R^2/\varepsilon^2)$ iterations which provides the optimal complexity for the non-smooth case.

The above choice (3.1.4) of weight parameters, however, is impractical since it depends on the final iterate k and an upper bound for $\sqrt{\xi(x_0, x^*)/\sigma_d}$. A more practical choice $\lambda_i := \gamma/\sqrt{i+1}$ for some $r > 0$ only ensures an upper bound

$$\frac{\xi(x_0, x^*) + (2\sigma_d)^{-1}\gamma^2 M^2(1 + \log(k+1))}{2\gamma(\sqrt{k+2} - 1)} = O(\log k/\sqrt{k})$$

for the right hand side of (3.1.2). It is important to note that, however, when the feasible set Q is compact, the weight parameters $\lambda_i := \gamma/\sqrt{i+1}$ ($\gamma > 0$) ensure the convergence $f(\widetilde{x}_k) - f(x^*) \leq O(1/\sqrt{k})$ for special weighted averages $\widetilde{x}_k$ of $x_0, \ldots, x_k$ [43, 44]. For instance, with the Nedić-Lee's averaging [43, eq. (17)]

$$\widetilde{x}_k := \frac{1}{\sum_{i=0}^k \lambda_i^{-1}} \sum_{i=0}^k \lambda_i^{-1} x_i, \tag{3.1.5}$$

the MDM with the weight parameters $\lambda_i := \gamma/\sqrt{i+1}$ ($\gamma > 0$) ensures the estimate

$$f(\widetilde{x}_k) - f(x^*) \leq \left(\frac{D_\xi^2}{\gamma} + \frac{\gamma M^2}{\sigma_d}\right)\frac{3}{2\sqrt{k+1}}$$

for all $k \geq 0$, where $D_\xi^2 := \sup_{x,y\in Q} \xi(x, y)$ [43, Corollary 2]. The choice $\gamma := D_\xi\sqrt{\sigma_d}/M$ leads this upper bound to its minimum $\frac{3}{2\sqrt{\sigma_d}}\frac{MD_\xi}{\sqrt{k+1}}$ (with respect to γ).

The MDM for strongly convex case

The MDM also attains the optimal iteration complexity in the strongly convex case. Let us further assume that we know a convexity parameter $\sigma_f > 0$ of $f(x)$ on Q and that the quadratic growth condition $\xi(y,x) \leq \frac{1}{2}\|x-y\|^2$, $\forall x, y \in Q$ holds. Then, [43, Lemma 3] shows that the Nedić-Lee's averaging (3.1.5) of the MDM with weight parameters $\{\lambda_k\}_{k\geq 0}$ satisfying

$$\lambda_k = \frac{\alpha_k}{\sigma_f}, \quad \alpha_0 = 1, \quad \alpha_k \in (0,1], \quad \frac{1-\alpha_{k+1}}{\alpha_{k+1}^2} \leq \frac{1}{\alpha_k^2}, \quad \forall k \geq 0$$

ensures the estimate

$$f(\widetilde{x}_k) - f(x^*) \leq (k+1)\alpha_k^2 \frac{M^2}{2\sigma_f \sigma_d}, \quad \forall k \geq 0. \tag{3.1.6}$$

For instance, the choice $\lambda_k := \frac{1}{\sigma_f t_k}$ where $t_0 := 1$, $t_{k+1} := \frac{1+\sqrt{1+4t_k^2}}{2}$ $(k \geq 0)$ leads the estimate (3.1.6) to

$$f(\widetilde{x}_k) - f(x^*) \leq \frac{2M^2}{\sigma_d \sigma_f (k+1)}, \quad \|\widetilde{x}_k - x^*\|^2 \leq \frac{4M^2}{\sigma_d \sigma_f^2 (k+1)}, \tag{3.1.7}$$

for all $k \geq 0$. Therefore, the MDM guarantees the optimal iteration complexity $O(M^2/(\sigma_d \sigma_f \varepsilon))$ in the strongly convex case (see also [42, Proposition 2.8] for a related result). Bach also analyzed the choice $\lambda_k = \frac{2}{\sigma_f(k+2)}$ of the MDM for a special form of strongly convex objective function [3, Proposition 3.1]. He proved almost the same estimate as (3.1.7) for the approximate solution

$$\widetilde{x}_k := \frac{2}{(k+1)(k+2)} \sum_{i=0}^{k} (i+1) x_i \tag{3.1.8}$$

which is slightly different from the Nedić-Lee's averaging $\frac{2}{(k+2)(k+3)} \sum_{i=0}^{k} (i+2) x_i$ (3.1.5).

3.1.2 Dual-averaging method and its variants

The Dual-Averaging Method (DAM) proposed by Nesterov [52] is an optimal PGM which overcomes the dependence of weight parameters of the MDM on the iteration counter k and achieves the rate of convergence $O(1/\sqrt{k})$ even if Q is unbounded. This method employs a non-decreasing sequence of positive numbers $\{\beta_k\}_{k\geq -1}$ $(\beta_{k+1} \geq \beta_k > 0)$, the *scaling parameters*, in addition to the weight parameters $\{\lambda_k\}_{k\geq 0}$.

From the initial point $x_0 := \operatorname{argmin}_{x\in Q} d(x) \in Q$, the DAM is performed by the iteration

$$\begin{aligned} g_k &:= g(x_k) \in \partial f(x_k) \\ z_k &:= \operatorname{argmin}_{x\in Q} \left\{ \textstyle\sum_{i=0}^{k} \lambda_i [f(x_i) + \langle g_i, x - x_i \rangle] + \beta_k d(x) \right\} \\ x_{k+1} &:= z_k \end{aligned} \tag{3.1.9}$$

for each $k \geq 0$.

It is important to note that the difference between the MDM and the DAM is the construction of the subproblems. They solves subproblems of the form $z_k := \operatorname{argmin}_{x\in Q} \varphi_k(x)$ defining *auxiliary functions* in their methods, namely, $\varphi_k(x)$ is defined by

$$\varphi_k(x) := \lambda_k (f(x_k) + \langle g_k, x - x_k \rangle) + \xi(x_k, x) \tag{3.1.10}$$

in the MDM and

$$\varphi_k(x) := \sum_{i=0}^{k} \lambda_i (f(x_i) + \langle g_i, x - x_i \rangle) + \beta_k d(x) \tag{3.1.11}$$

in the DAM, where $g_k := g(x_k) \in \partial f(x_k)$.

Nesterov proved that the following general estimate for the DAM (set $D = d(x^*)$ in [52, Theorem 1 and (3.2)]):

$$\forall k \geq 0, \quad \Delta_k := \frac{\sum_{i=0}^{k} \lambda_i f(x_i)}{\sum_{i=0}^{k} \lambda_i} - f(x^*) \leq \frac{\beta_k d(x^*) + \frac{1}{2\sigma_d} \sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_{i-1}} \|g_i\|_*^2}{\sum_{i=0}^{k} \lambda_i} \tag{3.1.12}$$

In order to ensure the rate $O(1/\sqrt{k})$ of convergence, we do not even need a prior knowledge of an upper bound for $\sqrt{\xi(x_0, x^*)/\sigma_d}$ in contrast to the MDM; for instance, choosing $\lambda_k := 1$ and $\beta_k := \gamma \hat{\beta}_k$ where $\gamma > 0$ and

$$\hat{\beta}_{-1} := \hat{\beta}_0 := 1, \ \hat{\beta}_{k+1} := \hat{\beta}_k + \hat{\beta}_k^{-1}, \qquad \forall k \geq 0, \tag{3.1.13}$$

the estimate (3.1.12) yields

$$\forall k \geq 0, \quad \Delta_k \leq \left(\gamma d(x^*) + \frac{M^2}{2\sigma_d \gamma} \right) \frac{0.5 + \sqrt{2k+1}}{k+1}.$$

If we further know $R \geq \sqrt{\frac{1}{\sigma_d} d(x^*)}$, the choice $\gamma := \frac{M}{\sqrt{2}\sigma_d R}$ yields $\Delta_k \leq \frac{MR}{\sqrt{2}} \frac{0.5+\sqrt{2k+1}}{k+1}$ achieving the optimal iteration complexity $O(M^2 R^2/\varepsilon^2)$ to obtain an ε-solution.

Nesterov and Shikhman [56] further proposed variants of the DAM, the double and triple averaging methods, in order to obtain convergence results for the sequence $\{x_k\}$. The double averaging method [56, eq. (28)] iterates starting from $x_0 := \operatorname{argmin}_{x \in Q} d(x) \in Q$ as follows:

$$z_k := \operatorname*{argmin}_{x \in Q} \varphi_k(x), \quad x_{k+1} := (1 - \tau_k) x_k + \tau_k z_k, \quad k = 0, 1, 2, \ldots \tag{3.1.14}$$

where $\tau_k := \lambda_{k+1} / \sum_{i=0}^{k+1} \lambda_i$ and $\varphi_k(x)$ is defined by the auxiliary function (3.1.11) used in the DAM. This method bounds the difference $f(x_k) - f(x^*)$ by the same value as the right hand side of (3.1.12) [56, Theorem 3.1] for all $k \geq 0$. Hence, it achieves the optimality. The triple averaging, which is a modification of (3.1.14), allows further flexibility on the choices for $\{\lambda_k\}$ and $\{\beta_k\}$ [56, Theorem 3.3].

Now, suppose that the objective function $f(x)$ is strongly convex with constant $\sigma_f > 0$ on Q. Juditsky and Nesterov [33] proposed a *multistage procedure* (or restarting technique) of the DAM which can be applied to this case (The original method is further applicable to *uniformly convex functions*, a generalization of strongly convex ones). This procedure ensures the optimal iteration complexity $O(M^2/(\sigma_d \sigma_f \varepsilon))$ in the strongly convex case using the prior knowledge of M, σ_f, and $R \geq \|x_0 - x^*\|$.

3.2 Gradient-based methods for smooth/structured problems

In this section, we review gradient-based methods for smooth problems or further structured ones (namely, weakly smooth problems, composite structure, and inexact oracle model). Let

us consider a closed convex set $Q \subset E$ equipped with a prox-function $d(x)$. We consider gradient-based methods for the convex optimization problem

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

where $f(x)$ is a lsc convex function on Q. Additional assumptions ((weak) smoothness, composite structure, and so on) will be specified corresponding to each method.

3.2.1 Classical proximal gradient methods

Let us begin by the most basic gradient method, the *steepest descent method* in the Euclidean setting ($\|\cdot\| = \|\cdot\|_2$): For an unconstrained minimization $\min_{x \in \mathbb{R}^n} f(x)$, $f \in \mathcal{F}^1_L(\mathbb{R}^n)$, start from $x_0 \in \mathbb{R}^n$ and iterate $x_{k+1} := x_k - \frac{\lambda_k}{L} \nabla f(x_k)$. Since $x_{k+1} = x_0 - \frac{1}{L} \sum_{i=0}^k \lambda_i \nabla f(x_i)$, it can be rewritten into two ways:

$$\begin{aligned} x_{k+1} &:= \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \lambda_k [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle] + \frac{L}{2} \|x - x_k\|_2^2 \right\} \\ &= \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \sum_{i=0}^k \lambda_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] + \frac{L}{2} \|x - x_0\|_2^2 \right\}. \end{aligned}$$

Notice that the first and the second expressions of x_{k+1} are the iterations of the MDM (3.1.1) and the DAM (3.1.9) (with replacing λ_k by λ_k/L and letting $\beta_k \equiv 1$), respectively.

The steepest descent method were extensively considered with a composite structure or an inexact oracle model as explained below.

Primal and dual gradient methods for composite problems

Consider the composite structure (2.4.9), namely, $f(x) := f_0(x) + \Psi(x)$ where $f \in \mathcal{F}^1_L(Q)$ and $\Psi(x)$ is a lsc convex function on Q. Nesterov [53] proposed the following (composite-type) PGMs, the *primal* and the *dual gradient methods* (for known Lipschitz constant L), in the Euclidean setting: Start from $x_0 \in Q$ and generate $\{x_k\}_{k \geq 0}$ by

$$\text{primal gradient method}: \quad x_{k+1} := \operatorname*{argmin}_{x \in Q} \left\{ f_0(x_k) + \langle \nabla f_0(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|_2^2 + \Psi(x) \right\} \tag{3.2.1}$$

or by

$$\text{dual gradient method}: \quad x_{k+1} := \operatorname*{argmin}_{x \in Q} \left\{ \sum_{i=0}^k [f_0(x_i) + \langle \nabla f_0(x_i), x - x_i \rangle + \Psi(x)] + \frac{L}{2} \|x - x_0\|_2^2 \right\}. \tag{3.2.2}$$

These updates do not involve weight parameter λ_k while the original methods involve the λ_k if we employ a line-search procedure to estimate an (unknown) Lipschitz constant L.

Remark 3.2.1. In the original (primal) gradient method (3.3) in [53], replacing (y_k, M_k, L_k) by (x_k, L, L) yields the above description of the primal gradient method. Similarly, the dual one is obtained by replacing (v_k, M_k, L_k) of (4.4) by (x_k, L, L). Moreover, the notation y_k in (4.4) is referred to as w_k below.

□

Nesterov showed the following estimate of the dual gradient method (One can take $\gamma_u \to 1$ in [53, eq. (4.8)]):

$$\min_{0\le i\le k} f(w_i) - f(x^*) \le \frac{L\|x_0 - x^*\|_2^2}{2(k+1)} \tag{3.2.3}$$

for all $k \ge 0$, where $w_k := \operatorname{argmin}_{x\in Q}\left\{f_0(x_k) + \langle \nabla f_0(x_k), x - x_k\rangle + \frac{L}{2}\|x - x_k\|_2^2 + \Psi(x)\right\}$.

The primal gradient method were analyzed by Beck and Teboulle [7, 8] in the case when $Q = \mathbb{R}^n$ and Ψ is subdifferentiable on $\operatorname{dom}\Psi$. In this case, the primal gradient method generates $\{x_k\}_{k\ge 0}$ satisfying

$$f(x_{k+1}) - f(x^*) \le \frac{L\|x_0 - x^*\|_2^2}{2(k+1)}$$

for all $k \ge 0$ [8, Theorem 1.1]. Note that the primal gradient method generates the test points $\{x_k\}$ such that $f(x_{k+1}) \le f(x_k)$ due to Proposition 2.1.6. Therefore, using the same notation $w_k := \operatorname{argmin}_{x\in Q}\left\{f_0(x_k) + \langle \nabla f_0(x_k), x - x_k\rangle + \frac{L}{2}\|x - x_k\|_2^2 + \Psi(x)\right\}$ $(= x_{k+1})$ as above, we see that both the primal and the dual gradient methods admit the estimate (3.2.3).

Nesterov showed the following linear convergence of the primal gradient method in the strongly convex case (Again, take $\gamma_u \to 1$ in [53, Theorem 5]). Suppose that $f(x)$ is strongly convex with constant $\sigma_f > 0$ on Q. Then the sequence $\{x_k\}_{k\ge 0}$ generated by the primal gradient method satisfies

$$f(x_k) - f(x^*) \le \begin{cases} \left(\frac{L}{\sigma_f}\right)^k (f(x_0) - f(x^*)) & : \ L/\sigma_f \le 1/2, \\ \left(1 - \frac{\sigma_f}{4L}\right)^k (f(x_0) - f(x^*)) & : \ \text{otherwise}, \end{cases} \tag{3.2.4}$$

for all $k \ge 0$. It is important to note that we do not need to know $\sigma_f > 0$ in the primal gradient method to ensure this result. The linear convergence of the dual gradient method was firstly shown by Devolder *et al.* [18] for the inexact oracle model, which we discuss next.

Primal and dual gradient methods with inexact oracle model

In the Euclidean setting, let us consider the inexact oracle model, *i.e.*, suppose that we can compute $(\bar{f}(y), \bar{g}(y)) \in \mathbb{R}\times E^*$ satisfying oracle inexactness (2.4.14) for parameters $L \ge \mu \ge 0$, $\delta \ge 0$. Devolder *et al.* [17] analyzed the primal and the dual gradient method in this setting. Starting from $x_0 \in Q$, they are described as

$$x_{k+1} := \operatorname*{argmin}_{x\in Q}\left\{\bar{f}(x_k) + \langle \bar{g}(x_k), x - x_k\rangle + \frac{L}{2}\|x - x_k\|_2^2\right\} \tag{3.2.5}$$

for the primal gradient method and

$$x_{k+1} := \operatorname*{argmin}_{x\in Q}\left\{\sum_{i=0}^{k} \lambda_i[\bar{f}(x_i) + \langle \nabla \bar{g}(x_i), x - x_i\rangle + \frac{\mu}{2}\|x - x_k\|_2^2] + \frac{L}{2}\|x - x_0\|_2^2\right\} \tag{3.2.6}$$

for the dual gradient method where $\{\lambda_k\}$ is a sequence of weight parameters.

The primal gradient method admits the following convergence result (see [18, Theorem 2] and the proof of [17, Theorem 4]):

$$\min_{0\le i\le k+1} f(x_i) - f(x^*) \le \frac{L\|x_0 - x^*\|_2^2}{2}\min\left\{\left(1 - \frac{\mu}{L}\right)^k, \frac{1}{k+1}\right\} + \delta. \tag{3.2.7}$$

The dual gradient method with

$$\lambda_0 := \frac{L}{L-\mu},\ \lambda_{k+1} := \frac{L+\mu S_k}{L-\mu} \tag{3.2.8}$$

also satisfies this estimate replacing the left hand side by $\min_{0\le i\le k} f(w_i) - f(x^*)$, where $w_k := \operatorname{argmin}_{x\in Q}\left\{\bar{f}(x_k) + \langle \bar{g}(x_k), x - x_k\rangle + \frac{L}{2}\|x - x_k\|_2^2\right\}$ [17, Theorem 5 and Remark 7]. Notice that this choice of weight parameters in the case $\mu = 0$ becomes $\lambda_k \equiv 1$.

Remark that the right hand side of the estimate (3.2.7) tends to δ as $k \to \infty$. In particular, whenever $\delta < \varepsilon$, we ensure an ε-solution with the iteration complexity

$$\min\left\{\Theta\left(\frac{LR^2}{\varepsilon-\delta}\right), \Theta\left(\frac{L}{\mu}\log\frac{LR^2}{\varepsilon-\delta}\right)\right\}$$

where $R := \|x_0 - x^*\|_2$. This complexity with $\delta = 0$ gives a well-known one of the steepest descent method or the projected gradient method for the class $\mathcal{F}_L^1(Q)$. Comparing with the lower bounds for the smooth problems (2.4.4), we see that the primal and the dual gradient methods does not ensure the optimal complexity.

3.2.2 Fast proximal gradient methods

Now we review PGMs, so called fast or accelerated PGMs, ensuring much better iteration complexity than the classical ones. In particular, they guarantees the optimal iteration complexity for the smooth problems.

For the smooth problems, we review three fast PGMs, the Nesterov's modified method [50] and the Tseng second/third accelerated proximal gradient methods [58, 59]. They are optimal in the non strongly convex case. We generalize the three methods in our unifying framework later.

We further review two fast PGMs, the Nesterov's accelerated method [53] for composite structure and the fast gradient method [17, 18] for inexact oracle model. They ensure the optimal iteration complexity for the smooth problems even in the strongly convex case.

Fast PGMs for smooth problems

Suppose that the objective function $f(x)$ belongs to the class $\mathcal{F}_L^1(Q)$. The first optimal complexity PGM in this case was proposed by Nesterov [47] and many variants or extensions were investigated (refer Section 2.4). Here we recall the *modified method* (with particular choice of the weight parameters λ_k) proposed in [50, Section 5.3]:

Nesterov's modified method [50]: Set $\lambda_k := (k+1)/2$ for $k \ge 0$ and $x_0 := z_{-1} := \operatorname{argmin}_{x\in Q} d(x)$. Compute the solution $w_0 := \operatorname{argmin}_{x\in Q}\{\lambda_0[f(x_0) + \langle \nabla f(x_0), x - x_0\rangle] + \frac{L}{\sigma_d}d(x)\}$ and set $\hat{x}_0 := z_0 := w_0$. For $k \ge 0$, iterate the following procedure:

$$\begin{aligned}
&\text{Set} && x_{k+1} := (1-\tau_k)\hat{x}_k + \tau_k z_k, \quad \text{where } \tau_k := \tfrac{\lambda_{k+1}}{\sum_{i=0}^{k+1}\lambda_i},\\
&\text{Compute} && w_{k+1} := \operatorname{argmin}_{x\in Q}\left\{\lambda_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle] + \tfrac{L}{\sigma}\xi(z_k, x)\right\},\\
&\text{Set} && \hat{x}_{k+1} := (1-\tau_k)\hat{x}_k + \tau_k w_{k+1},\\
&\text{Compute} && z_{k+1} := \operatorname{argmin}_{x\in Q}\left\{\textstyle\sum_{i=0}^{k+1}\lambda_i[f(x_i) + \langle \nabla f(x_i), x - x_i\rangle] + \tfrac{L}{\sigma_d}d(x)\right\}.
\end{aligned} \tag{3.2.9}$$

In comparison, the Tseng's second and third Accelerated Proximal Gradient (APG) methods [59], which are particular cases of algorithms 1 and 3 in [58], only require the computation

of either z_k or w_k of the Nesterov's method, respectively.

Tseng's second APG method [59]: Set $\lambda_0 := 1$, $\lambda_{k+1} := \frac{1+\sqrt{1+4\lambda_k^2}}{2}$ for $k \geq 0$, and $x_0 := z_{-1} := \operatorname{argmin}_{x\in Q} d(x)$. Compute the solution $z_0 := \operatorname{argmin}_{x\in Q}\{\lambda_0[f(x_0) + \langle \nabla f(x_0), x - x_0\rangle] + \frac{L}{\sigma_d}\xi(x_0, x)\}$ and set $\hat{x}_0 := z_0$. For $k \geq 0$, iterate the following procedure:

$$\begin{array}{ll} \text{Set} & x_{k+1} := (1-\tau_k)\hat{x}_k + \tau_k z_k, \quad \text{where } \tau_k := \frac{\lambda_{k+1}}{\sum_{i=0}^{k+1}\lambda_i}, \\ \text{Compute} & z_{k+1} := \operatorname{argmin}_{x\in Q}\left\{\lambda_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle] + \frac{L}{\sigma_d}\xi(z_k, x)\right\}, \\ \text{Set} & \hat{x}_{k+1} := (1-\tau_k)\hat{x}_k + \tau_k z_{k+1}. \end{array} \tag{3.2.10}$$

Tseng's third APG method [59]: Set $\lambda_0 := 1$, $\lambda_{k+1} := \frac{1+\sqrt{1+4\lambda_k^2}}{2}$ for $k \geq 0$, and $x_0 := z_{-1} := \operatorname{argmin}_{x\in Q} d(x)$. Compute the solution $z_0 := \operatorname{argmin}_{x\in Q}\{\lambda_0[f(x_0) + \langle \nabla f(x_0), x - x_0\rangle] + \frac{L}{\sigma_d}d(x)\}$ and set $\hat{x}_0 := z_0$. For $k \geq 0$, iterate the following procedure:

$$\begin{array}{ll} \text{Set} & x_{k+1} := (1-\tau_k)\hat{x}_k + \tau_k z_k, \quad \text{where } \tau_k := \frac{\lambda_{k+1}}{\sum_{i=0}^{k+1}\lambda_i}, \\ \text{Compute} & z_{k+1} := \operatorname{argmin}_{x\in Q}\left\{\sum_{i=0}^{k+1}\lambda_i[f(x_i) + \langle \nabla f(x_i), x - x_i\rangle] + \frac{L}{\sigma_d}d(x)\right\}, \\ \text{Set} & \hat{x}_{k+1} := (1-\tau_k)\hat{x}_k + \tau_k z_{k+1}. \end{array} \tag{3.2.11}$$

Remark 3.2.2. To see the equivalence to the Tseng's second APG method, notice that x_0 is not used at all in [59]. Then defining $d(x) := D(x, z_0) = \eta(x) - \eta(z_0) - \langle \nabla\eta(z_0), x - z_0\rangle$ for an arbitrary $z_0 \in Q$, we have $\sigma_d = 1$ in (a). Finally, making the correspondence $z_k \to z_{k-1}$, $y_k \to x_k$, $x_k \to \hat{x}_k$, and $\theta_k \to \frac{1}{\lambda_k}$, it will result in our notation. For the Tseng's third APG method, identical observations are valid, excepting that we define $d(x) := \eta(x) - \eta(z_0)$ instead. □

Remark 3.2.3. Tseng's APG methods were originally proposed for the composite problem $\min_{x\in\operatorname{dom}\Psi}[f(x) \equiv f_0(x) + \Psi(x)]$ where $f_0 \in \mathcal{F}_L^1(\operatorname{dom}\Psi)$ and $\Psi(x)$ is a lsc convex function with closed domain. The above description is obtained by letting Ψ the indicator function of Q. □

Similar to the comparison of the MDM and the DAM, the difference among the above three methods is basically the subproblems at each iteration. These subproblems has the form $z_k := \operatorname{argmin}_{x\in Q}\varphi_k(x)$ with the auxiliary functions

$$\varphi_k(x) := \lambda_k[f(x_k) + \langle \nabla f(x_k), x - x_k\rangle] + \frac{L}{\sigma_d}\xi(z_{k-1}, x) \tag{3.2.12}$$

for the Tseng's second APG method and

$$\varphi_k(x) := \sum_{i=0}^{k}\lambda_i[f(x_i) + \langle \nabla f(x_i), x - x_i\rangle] + \frac{L}{\sigma_d}d(x) \tag{3.2.13}$$

for the Tseng's third APG method; the Nesterov's method can be seen as their hybrid. Notice that the auxiliary functions (3.2.12) and (3.2.13) correspond to the one of the MDM (3.1.10) except the factor L/σ_d, and the one of the DAM (3.1.11) with $\beta_k \equiv L/\sigma_d$, respectively.

It can be shown that the Nesterov's and the Tseng's methods attain the optimal iteration complexity for smooth problems in the non strongly convex case. The Nesterov's method (3.2.9) and the Tseng's third APG method (3.2.11) satisfy

$$\forall k \geq 0, \quad f(\hat{x}_k) - f(x^*) \leq \frac{4Ld(x^*)}{\sigma_d(k+1)(k+2)} \tag{3.2.14}$$

while the Tseng's second APG method (3.2.10) satisfies

$$\forall k \geq 0, \quad f(\hat{x}_k) - f(x^*) \leq \frac{4L\xi(x_0, x^*)}{\sigma_d (k+2)^2} \tag{3.2.15}$$

(see [50, Theorem 2] and [58, Corollaries 1,3]). Therefore, they ensure an ε-solution with the optimal iteration complexity $O(\sqrt{LR^2/\varepsilon})$ where $R = \sqrt{d(x^*)/\sigma_d}$ for the first estimate and $R = \sqrt{\xi(x_0, x^*)/\sigma_d}$ for the second one, respectively.

In the Euclidean setting $d(x) = \frac{1}{2}\|x - x_0\|_2^2$, we can apply a multistage procedure [47, 53] to the above methods to achieve the optimal complexity in the strongly convex case if we know a convexity parameter $\sigma_f > 0$ of f on Q. Indeed, since the above estimates have the form $f(\hat{x}_k) - f(x^*) \leq \frac{cL\|x_0 - x^*\|_2^2}{2k^2}$ for some $c > 0$, after $k \geq \sqrt{2cL/\sigma_f}$ iterations, we have $f(\hat{x}_k) - f(x^*) \leq \frac{\sigma_f}{4}\|x_0 - x^*\|_2^2 \leq \frac{1}{2}(f(x_0) - f(x^*))$ by the strong convexity of f and the optimality of x^*. Then, one can restart the method setting $\hat{x}_k$ as the new initial point, which ensures an ε-solution repeating $\left\lceil \log_2 \frac{f(x_0) - f(x^*)}{\varepsilon} \right\rceil$ times of restarting; the total iteration complexity $O\left(\sqrt{\frac{L}{\sigma_f}} \log \frac{1}{\varepsilon}\right)$ is optimal for the smooth problems.

One can also apply PGMs introduced next which ensures the optimal complexity without multistage procedure. Note that the optimality of the above Nesterov's and Tseng's methods without a multistage procedure in the strongly convex case is not known.

Fast PGMs for convex optimization problems with composite structure

Consider a convex optimization problems with a composite structure: $\min_{x \in Q}[f(x) \equiv f_0(x) + \Psi(x)]$ where $f_0 \in \mathcal{F}_L^1(Q)$ and $\Psi(x)$ is a lsc convex function on Q. Tseng's second and third APG methods [58, 59] were originally proposed for this case which can be described by replacing the first-order approximation $f(x_i) + \langle \nabla f(x_i), x - x_i \rangle$ by its composite version $f_0(x_i) + \langle \nabla f_0(x_i), x - x_i \rangle + \Psi(x)$ in the subproblems in (3.2.10) and (3.2.11), preserving the efficiency estimates (3.2.15) and (3.2.14), respectively.

In the Euclidean setting $d(x) := \frac{1}{2}\|x - x_0\|_2^2$, Nesterov's *accelerated method* [53] further ensures a linear convergence in the strongly convex case. When we know the Lipschitz constant L and a convexity parameter σ_Ψ of $\Psi(x)$ on Q, the Nesterov's accelerated method equips two subproblems at each iteration (collaborating the ones of the primal and the dual PGMs in Section 3.2.1) and generates points $\{\hat{x}_k\}_{k \geq 0} \subset Q$ satisfying the following estimate (let $\gamma_u \to 1$ in [53, Theorem 6]):

$$\forall k \geq 1, \quad f(\hat{x}_k) - f(x^*) \leq \frac{L\|x_0 - x^*\|_2^2}{4} \min\left\{\frac{4}{k^2}, \left(1 + \sqrt{\frac{\sigma_\Psi}{2L}}\right)^{-2(k-1)}\right\}. \tag{3.2.16}$$

We remark that the Nesterov's accelerated method does not involve a multistage procedure.

In the case when we further know a convexity parameter σ_{f_0} of $f_0(x)$ on Q, one can improve the estimate (3.2.16) by reallocating the function $\Psi(x)$ as we briefly explain here (see [53, Section 5]). In fact, we can rewrite the objective function as $f(x) = g_0(x) + \Phi(x)$ where

$$g_0(x) := f_0(x) - \frac{\sigma_{f_0}}{2}\|x - x_0\|_2^2, \quad \Phi(x) := \Psi(x) + \frac{\sigma_{f_0}}{2}\|x - x_0\|_2^2.$$

The convexity parameter of $\Phi(x)$ is $\sigma_{f_0} + \sigma_\Psi$ and the Lipschitz constant of $\nabla g_0(x)$ on Q is $L - \sigma_{f_0}$ where L is the one of $f_0(x)$. Therefore, the Nesterov's accelerated method for this

reallocation leads the above estimate (3.2.16) to

$$\forall k \geq 1, \quad f(\hat{x}_k) - f(x^*) \leq \frac{(L - \sigma_{f_0}) \|x_0 - x^*\|_2^2}{4} \min \left\{ \frac{4}{k^2}, \left(1 + \sqrt{\frac{\sigma_f}{2(L - \sigma_{f_0})}}\right)^{-2(k-1)} \right\} \tag{3.2.17}$$

where we denote $\sigma_f := \sigma_{f_0} + \sigma_\Psi$.

Fast PGMs for convex optimization problems with inexact oracle model

Consider a convex optimization problem $\min_{x \in Q} f(x)$ equipped with a (δ, L, μ) oracle (2.4.14). Similar to the Nesterov's accelerated method, Devolder *et al.* proposed the *fast gradient method* [17, Algorithm 3] collaborating the primal and the dual methods for this case. The fast gradient method generates $\{\hat{x}_k\}_{k \geq 0} \subset Q$ ensuring the following estimate [17, Theorem 7]:

$$\forall k \geq 1, \quad f(\hat{x}_k) - f(x^*) \leq \frac{L \|x_0 - x^*\|_2^2}{2} \min \left\{ \frac{4}{k^2}, \left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{-2k} \right\} + E_k \tag{3.2.18}$$

where $E_k = \Theta\left(\delta \cdot \min\{k, \sqrt{L/\mu}\}\right)$. Due to the error term E_k, ensuring an ε-solution and the iteration complexity depend on the parameters L, μ, δ. One can see further discussions in [17, Sections 5.3, 5.4] and [18, Section 5.2].

3.2.3 Conditional gradient methods

We finally discuss on the CGMs for some structured problems.

Suppose that Q is bounded and the objective function $f(x)$ is differentiable on Q. The CGM proposed by Frank and Wolfe [21], which we refer to as the *classical CGM*, is the most basic one: Start from $x_0 \in Q$ and, for each $k \geq 0$, iterate

$$z_k \in \operatorname*{Argmin}_{x \in Q} [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle], \quad x_{k+1} := (1 - \tau_k) x_k + \tau_k z_k \tag{3.2.19}$$

where $\tau_k \in [0, 1]$. A popular choice $\tau_k := \frac{2}{k+2}$ ensures the following estimate when $f \in \mathcal{F}_L^1(Q)$ [22, Bound 3.1][1]:

$$\forall k \geq 1, \quad f(x_{k+1}) - f(x^*) \leq \frac{2L\mathrm{Diam}(Q)^2}{k+4}. \tag{3.2.20}$$

In the weakly smooth case $f \in \mathcal{F}_M^{\rho-1}(Q)$ ($\rho \in (1, 2]$), the same choice $\tau_k := 2/(k+2)$ ensures an estimate $f(x_k) - f(x^*) \leq O(M\mathrm{Diam}(Q)^\rho / k^{\rho-1})$ (see [55]). The same convergence rate holds for a composite-type CGM extending the classical CGM to composite structured problems [1, 3, 55].

The classical CGM for the inexact oracle model was analyzed by Freund and Grigas (see [22, Section 5.2.1]). If the objective function f is equipped with a $(\delta, L, 0)$-oracle $(\bar{f}, \bar{g})$, then the classical CGM with $\tau_k = 2/(k+2)$ replacing $(f, \nabla f)$ by $(\bar{f}, \bar{g})$ satisfies $f(x_k) - f^* \leq O(L\mathrm{Diam}(Q)^2/k) + O(\delta k)$.

There are some variants [35, 55] of the classical CGM. In particular, Nesterov [55] demonstrated the iteration complexity (2.4.8) for the weakly smooth problems. Here we describe particular instances of Lan's variants [35] for the smooth problems which are discussed in the

[1] Replace $(h(\cdot), \lambda_k, \tilde{\lambda}_k, \bar{\alpha}_k)$ in [22] by $(-f(\cdot), x_k, z_k, \tau_k)$.

unifying framework in Chapter 4. The *primal averaging CGM* [35, Algorithm 4] with $\alpha_k = 2/(k+1)$ is described as follow: Initializing $x_0 := z_{-1} \in Q$, $\hat{x}_0 := z_0 \in \operatorname{Argmin}_{x\in Q} \langle \nabla f(x_0), x\rangle$, iterate the following procedure

$$\begin{aligned} x_{k+1} &:= \frac{k+1}{k+3}\hat{x}_k + \frac{2}{k+3}z_k, \\ z_{k+1} &\in \operatorname*{Argmin}_{x\in Q} \langle \nabla f(x_{k+1}), x\rangle, \\ \hat{x}_{k+1} &:= \frac{k+1}{k+3}\hat{x}_k + \frac{2}{k+3}z_{k+1}, \end{aligned} \tag{3.2.21}$$

for each $k \geq 0$. Lan also proposed another variant, the *primal dual averaging CGM*, where we give its particularization by $\alpha_k = 2/(k+1)$, $\theta_k = k$ in [35, Algorithm 5] here: Initialize $x_0 := z_{-1} \in Q$, $\hat{x}_0 := z_0 \in \operatorname{Argmin}_{x\in Q} \langle \nabla f(x_0), x\rangle$ and iterate

$$\begin{aligned} x_{k+1} &:= \frac{k+1}{k+3}\hat{x}_k + \frac{2}{k+3}z_k, \\ z_{k+1} &\in \operatorname*{Argmin}_{x\in Q} \left\langle \sum_{i=0}^{k+1} (i+1)\nabla f(x_i), x \right\rangle, \\ \hat{x}_{k+1} &:= \frac{k+1}{k+3}\hat{x}_k + \frac{2}{k+3}z_{k+1} \end{aligned} \tag{3.2.22}$$

for each $k \geq 0$. These two CGMs satisfy the following convergence rate [35, Theorems 7,8].

$$\forall k \geq 0, \quad f(\hat{x}_k) - f(x^*) \leq \frac{2L\operatorname{Diam}(Q)^2}{k+2}. \tag{3.2.23}$$

Chapter 4

A Unifying Framework of Subgradient-Based Methods for Structured Convex Optimization Problems

4.1 Overview

In this chapter, we establish a methodology of generating optimal or nearly optimal complexity subgradient-based methods for several classes of convex optimization problems. After some preliminaries in Section 4.2, the core notion of the thesis will be introduced in Section 4.3. The remaining sections demonstrate how our notion works as a unifying framework.

We at first introduce two classes of convex optimization problems, the *non-smooth* and the *structured problems*. The former was already introduced while the latter is a large class of problems including the (weakly) smooth problems, the mixed smoothness structure, the composite structure, and the inexact oracle model. We additionally consider the 'strong convexity' with respect to the prox-function, generalizing the one in the Euclidean setting.

The unifying framework is introduced in Section 4.3. Recall, for instance, that both the MDM and the DAM solves subproblems $z_k := \operatorname{argmin}_{x \in Q} \varphi_k(x)$ where $\varphi_k(x)$ is defined by (3.1.10) and (3.1.11), respectively. In order to discuss them in a unified way, we define Properties A (and B) as axioms for the auxiliary functions $\{\varphi_k(x)\}$ necessary to develop efficient subgradient-based methods. Based on these properties, we then propose two general methods, Methods I and II, of solving the non-smooth and the structured problems, respectively. We totally propose four kinds of methods since both Methods I and II consist of the *classical* and the *modified methods* which sometimes have different rates of convergence. We demonstrate in Section 4.3.4 that particular instances of the proposed methods yield existing subgradient-based methods reviewed in Chapter 3.

The remaining sections correspond to unified analysis of subgradient-based methods. We develop general convergence estimates in Section 4.4. The development here exploits the Nesterov's approach [50] using the relation (R_k). We then particularize our general estimate to the non-smooth problems (Section 4.5), the smooth/composite problems as well as the inexact oracle model (Section 4.6), and the weakly smooth problems (Section 4.7). We compare our results with known ones reviewed in Chapter 3. We summarize remarkable results below.

- *Results for the non-smooth problems (non strongly convex case).*
  Theorem 4.5.1 shows that Method I ensures the optimal iteration complexity for the non-smooth problems with the same advantage as the Nesterov's DAM. As a byproduct, the extended MDM proposed in Method 4.3.5 does not require the boundedness

assumption of the feasible set to attain the optimality in contrast to the existing averaging techniques for the original MDM.

- *Results for the non-smooth problems (strongly convex case).*
  We show an optimal convergence result of Method I in Section 4.5.2. It recovers the optimality of the MDM for the Nedić-Lee's averaging (3.1.5) and the Bach's variant (3.1.8). Moreover, a new extension of the DAM to the strongly convex case is obtained.

- *Results for the inexact oracle model and the composite problems.*
  In Section 4.6, we analyze Method II and show the so called classical convergence rate (for the smooth problems) of the classical method and the optimal convergence of the modified method. For particular cases, we obtain the same convergence results as the known ones where some of them have slight improvements. In particular, we obtain extensions of the Tseng's APG methods and the Nesterov's modified method to the strongly convex case ensuring the optimal iteration complexity.

- *Results for the weakly smooth problems and the mixed smoothness structure.*
  In Section 4.7 we analyze the modified methods of Method II for a class of problems including the weakly smooth problems and the mixed smoothness structure. In particular, we obtain the optimal iteration complexity for the weakly smooth problems. We ensure the optimality in the strongly convex case with less prior requirements compared with the existing method, while the result in the non strongly convex case may be restrictive because it does not 'adapt' the Hölder condition.

  Nearly optimal iteration complexity for CGMs is also obtained in the non strongly convex case.

4.1.1 Notations and settings

Here we collect common notations in this chapter.

Let E be a finite dimensional real vector space equipped with a norm $\|\cdot\|$.
Throughout this chapter, we fix a prox-function $d(x)$ on Q, that is,

- $d : E \to \mathbb{R} \cup \{+\infty\}$ is a differentiable and strongly convex function on Q with constant $\sigma_d > 0$,
- $d(x) \geq 0, \forall x \in Q$ and $d(x_0) = 0$ for $x_0 := \operatorname{argmin}_{x \in Q} d(x)$.

As we fixed the prox-function, we simply denote the associated Bregman distance as

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

Consider the following convex optimization problem:

$$\min_{x \in Q} f(x) \tag{4.1.1}$$

where Q is a closed convex subset of E and f is a lsc convex function on E. We introduce additional assumptions on this problem in the next section. We mainly focus, in particular, on the problem (4.1.1) in two categories, the non-smooth and the structured problems introduced in Section 4.2.2.

4.2 Non-smooth and structured convex problems

In this section, we introduce two kinds convex problems, the non-smooth and the structured ones, which we apply our methodology to obtain efficient subgradient-based methods. These convex problems cover several classes of known convex problems as clarified later (see Example 4.2.8).

We firstly prepare the notion of a generalization of the strong convexity in the Euclidean setting in Section 4.2.1 which is used to define the non-smooth and the structured problems in Section 4.2.2.

4.2.1 Strong convexity with respect to prox-function

Development of subgradient methods for convex problems with strongly convex objective functions often assume the Euclidean setting or the quadratic growth condition (2.1.13) for $\xi(y, x)$. We consider the following notion of strong convexity to handle the both cases.

Definition 4.2.1 (strong convexity with respect to prox-function)**.** Let $\varphi : E \to \mathbb{R} \cup \{+\infty\}$ be a lsc convex function with $Q \subset \operatorname{dom}\varphi$. For a nonnegative constant σ, we say φ is *σ-strongly convex with respect to the prox-function d on Q* if $\varphi - \sigma d$ is convex on Q. Then we call σ a *convexity parameter of φ with respect to d on Q*. The set of the convexity parameters is written by $\sigma(\varphi)$, namely,

$$\sigma(\varphi) := \{\sigma \geq 0 \mid \varphi - \sigma d \text{ is convex on } Q\}. \tag{4.2.1}$$

Remark that we omitted the dependence to the prox-function d and the feasible set Q from the notation $\sigma(\varphi)$ for simplicity. This notion is also discussed in [41] when $\varphi(x)$ is differentiable on Q.

When E is a Euclidean space and $\|\cdot\|_2$ is the norm induced by the inner product on E, the strong convexity with respect to the norm $\|\cdot\|_2$ is equivalent to the one with respect to the prox-function $d(x) := \frac{1}{2}\|x\|_2^2$. This fact is a particular one of the following characterization.

Proposition 4.2.2. *Let $\varphi : E \to \mathbb{R} \cup \{+\infty\}$ be a lsc convex function with $Q \subset \operatorname{dom}\varphi$. Then the followings are equivalent.*

(i) $\sigma \in \sigma(\varphi)$.

(ii) For every $x, y \in Q$ ($\subset \operatorname{dom}\varphi$), we have

$$\varphi(x) \geq \varphi(y) + \varphi'(y; x - y) + \sigma\xi(y, x).$$

Proof. Remark that, in general, the function $\psi(x) := \varphi(x) - \sigma d(x)$ satisfies

$$\psi'(y; x - y) = \varphi'(y; x - y) - \sigma \langle \nabla d(y), x - y \rangle, \quad \forall x, y \in Q. \tag{4.2.2}$$

Therefore, by the definition of the Bregman distance, the condition (ii) is equivalent to

$$\text{(ii)'} \qquad \psi(x) \geq \psi(y) + \psi'(y; x - y), \quad \forall x, y \in Q$$

Suppose that the condition (i) holds. Then, since ψ is convex on Q, we have the condition (ii)', that is, the condition (ii) holds.

Conversely, suppose that the condition (ii) or, equivalently, (ii)' holds. Since φ is convex on Q, $\varphi'(y; x-y) \geq -\varphi'(y; y-x)$ holds by (2.1.6) and so is true for $\psi(\cdot)$ by (4.2.2). Therefore,

we obtain two inequalities $\psi(y) \geq \psi(z) + \psi'(z; y - z)$ and $\psi(x) \geq \psi(z) - \psi'(z; z - x)$ for all $x, y, z \in Q$. Since $\psi'(y; \cdot)$ is positively homogeneous, the convexity of $\psi(\cdot)$ on Q follows by taking a convex combination of the two with $z = \alpha x + (1 - \alpha)y,\ \alpha \in [0, 1],\ x, y \in Q$. □

As an immediate consequence of Proposition 4.2.2 combining with (2.1.12) and Proposition 2.1.1, we have the following.

Corollary 4.2.3. *Let $\varphi : E \to \mathbb{R} \cup \{+\infty\}$ be a lsc convex function with $Q \subset \operatorname{dom} \varphi$.*

(i) For any $\sigma \in \sigma(\varphi)$, the constant $\sigma_d \sigma$ is a convexity parameter of $\varphi(x)$ with respect to the norm $\|\cdot\|$.

(ii) Suppose that the Bregman distance $\xi(y, x)$ grows quadratically on Q with a constant $A > 0$, that is, $\xi(y, x) \leq \frac{A}{2} \|x - y\|^2$ holds for every $x, y \in Q$. If $\varphi(x)$ is τ-strongly convex on Q with respect to the norm $\|\cdot\|$, then we have $\tau / A \in \sigma(\varphi)$.

Finally, we show that the minimizer of a strongly convex function with respect to the prox-function has the following property which is a key in the analysis of subgradient-based methods (see also [13, Lemma 3.2], [38, Lemma 1], [54, Lemma 3], and [58, Property 2]).

Lemma 4.2.4. *For a positive constant $\sigma > 0$, let φ be a σ-strongly convex function with respect to d on Q. Then, φ has a unique minimizer z^* on Q satisfying the following inequality for every $z \in Q$:*

$$\varphi(z) \geq \varphi(z^*) + \sigma \xi(z^*, z).$$

Proof. Since φ becomes $\sigma\sigma_d$-strongly convex on Q with respect to the norm $\|\cdot\|$, it has a unique minimizer z^* on Q (Proposition 2.2.4). Then, by the optimality condition $\varphi'(z^*; x - z^*) \geq 0,\ \forall x \in Q$ for the minimizer z^* (Lemma 2.2.1), the assertion follows from Proposition 4.2.2 (ii). □

4.2.2 Non-smooth and structured convex problems

This section introduces the classes of the non-smooth and the structured convex problems (Definitions 4.2.6 and 4.2.7, respectively) unifying some particular classes. We at first define Assumption 4.2.5 which is assumed for the both classes. It introduces the notation $m_f(y; x)$ which we refer a *lower approximation model of $f(x)$ (at y)*.

Assumption 4.2.5. The convex optimization problem (4.1.1) is equipped with a function $m_f(y; x)$ and a parameter $\sigma_f \geq 0$ satisfying the following conditions.

(i) $m_f(y; x)$ is defined for each $x, y \in Q$ and, for every $y \in Q$, the function $m_f(y; \cdot)$ is lsc and convex on Q satisfying $f(x) \geq m_f(y; x),\ \forall x \in Q$.

(ii) The parameter σ_f satisfies

$$\sigma_f \in \sigma(f) \cap \bigcap_{y \in Q} \sigma(m_f(y; \cdot)). \tag{4.2.3}$$

For the problem (4.1.1) satisfying Assumption 4.2.5, we refer to the case $\sigma_f > 0$ as the *strongly convex case* while the *non strongly convex case* is referred to as the one $\sigma_f = 0$ which corresponds to assume the item (i) only.

Now we describe the classes of the non-smooth and the structured problems.

Definition 4.2.6 (class of non-smooth problems)**.** *The class of non-smooth problems* consists of convex optimization problems (4.1.1) where we assume for each problem that we know a subgradient mapping $g(x) \in \partial f(x)$, $x \in Q$ and a convexity parameter $\sigma_f \in \sigma(f)$. Then, we can naturally define its lower approximation model $m_f(y; x)$ by

$$m_f(y; x) := f(y) + \langle g(y), x - y \rangle + \sigma_f \xi(y, x), \quad x, y \in Q. \tag{4.2.4}$$

Furthermore, we assume that the following optimization problem is solvable for every $s \in E^*$ and $\beta > 0$:

$$\min_{x \in Q} \{\langle s, x \rangle + \beta d(x)\}. \tag{4.2.5}$$

This class of problems is denoted by $\mathcal{NSP}(g, \sigma_f)$.

Notice that each problem of $\mathcal{NSP}(g, \sigma_f)$ satisfies Assumption 4.2.5 because $m_f(y; x) - \sigma_f d(x)$ is an affine function so that (4.2.3) follows.

The class $\mathcal{NSP}(g, \sigma_f)$ formalizes the non-smooth problems introduced in Section 2.4. Therefore, under the boundedness assumption (2.4.1) of subgradients, the optimal iteration complexities of the classes $\mathcal{NSP}(g, 0)$ and $\mathcal{NSP}(g, \sigma_f)$ for $\sigma_f > 0$ are given by (2.4.2) and (2.4.3), respectively.

Definition 4.2.7 (class of structured problems)**.** *The class of structured problems* consists of convex optimization problems (4.1.1) where we assume for each problem that there exists $(m_f(\cdot; \cdot), \sigma_f, \bar{\sigma}_f, L(\cdot), \delta(\cdot, \cdot))$, *i.e.*, functions and constants, satisfying the inequality

$$f(x) \leq [m_f(y; x) - \bar{\sigma}_f \xi(y, x)] + \frac{L(y)}{2} \|y - x\|^2 + \delta(y, x), \quad \forall x, y \in Q, \tag{4.2.6}$$

where $m_f(y; x)$ is a lower approximation model of $f(x)$ which admits Assumption 4.2.5 for a convexity parameter $\sigma_f \geq 0$, $\delta(y, \cdot)$ is a nonnegative convex function on Q for each $y \in Q$, $L(\cdot) \geq 0$, and $\bar{\sigma}_f \in [0, \sigma_f]$. We further assume that the following optimization problem is efficiently solvable for every $\beta \geq 0$, $y \in E$, and $s \in E^*$:

$$\min_{x \in Q} \{m_f(y; x) + \langle s, x \rangle + \beta d(x)\}. \tag{4.2.7}$$

This class of problems is denoted by $\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$.

We explain the role of parameters in this definition (Example 4.2.8 will show further detail).

Although the class $\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$ is quite general, we are particularly interested in the following special cases for our purpose:

- the constant case $L(y) \equiv L$ and $\delta(y, x) \equiv \delta$ which corresponds to the inexact oracle model;
- the case $\delta(y, x) = \frac{M}{\rho} \|y - x\|^\rho$ for some $M \geq 0$, $\rho \in [1, 2)$ which includes the weakly smooth problems.

We will focus on these cases when we analyze the concrete convergence rates of our methods.

The parameter $\bar{\sigma}_f$ represents 'the coefficient of $\xi(y, x)$ in $m_f(y; x)$.' The $\bar{\sigma}_f$ may take a different value from σ_f when we consider the composite structure. Remark that $m_f(y; \cdot) - \bar{\sigma}_f \xi(y, \cdot)$ is a convex function because $\sigma_f \in \sigma(m_f(y; \cdot))$ and $\bar{\sigma}_f \in [0, \sigma_f]$.

The assumption of the solvability of (4.2.7) can be reduced depending on the implementation of the methods. For instance, the implementation of Theorem 4.6.4 (ii) will require only the subproblems (4.2.7) with $\beta = 0$ which corresponds to consider a generalization of the CGM.

The class of structured problems enables us to generalize several classes of convex optimization problems as shown next.

Example 4.2.8 (examples of structured problems)**.** Let us consider the convex optimization problem (4.1.1). We demonstrate that the classes of convex optimization problems introduced in Section 2.4 belong to $\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$ with appropriate constants and functions.

(i) *Smooth problems.* Let $f \in \mathcal{F}_L^1(Q)$ and $\sigma_f \in \sigma(f)$. Define the lower approximation model $m_f(y; x)$ by

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

Due to Proposition 2.1.6 with $\nu = 1$, we see that the inequality (4.2.6) follows with

$$\bar{\sigma}_f := \sigma_f, \quad L(\cdot) := L, \quad \delta(\cdot, \cdot) := 0.$$

We remark that the corresponding subproblem (4.2.7) in the cases $\beta > 0$ and $\beta = 0$ reduces to the one (2.3.1) of the PGMs and the one (2.3.3) of the CGMs, respectively.

(ii) *Weakly smooth problems.* Let $f \in \mathcal{F}_M^\nu(Q)$ for $M \geq 0$ and $\nu \in [0, 1)$ (the case $\nu = 1$ for the smooth problems was separately discussed above). By Proposition 2.1.6, we have

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

where $g(y) \in \partial f(y)$ is any subgradient mapping of f (Recall that $g(y) = \nabla f(y)$ whenever $\nu > 0$). For a convexity parameter $\sigma_f \in \sigma(f)$, let us define $m_f(y; x)$ by (4.2.4). Then, the inequality (4.2.6) follows by letting

$$\bar{\sigma}_f := \sigma_f, \quad L(\cdot) := 0, \quad \delta(y, x) := \frac{M}{1+\nu} \|x - y\|^{1+\nu}.$$

The subproblem (4.2.7) has the same form as the smooth problems (i).

(iii) *Composite structure.* Suppose that the objective function f is of the form $f(x) = f_0(x) + \Psi(x)$ where $f_0 \in \mathcal{F}_M^\nu(Q)$ for some $\nu \in [0, 1]$ and Ψ is a lsc convex function on Q. Take convexity parameters $\sigma_{f_0} \in \sigma(f_0)$ and $\sigma_\Psi \in \sigma(\Psi)$. Applying (i) or (ii) to $f_0(x)$, we can define $(m_{f_0}, \bar{\sigma}_{f_0}, L(\cdot), \delta(\cdot, \cdot))$ so that the inequality (4.2.6) holds for the function $f_0(x)$. Now let us define

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

Then, the inequality (4.2.6) holds with

$$\sigma_f := \sigma_{f_0} + \sigma_\Psi, \quad \bar{\sigma}_f := \sigma_{f_0}$$

and with the above $(L(\cdot), \delta(\cdot, \cdot))$ for $f_0(x)$. The corresponding subproblem (4.2.7) in the cases $\beta > 0$ and $\beta = 0$ is equivalent to the one (2.4.10) of composite-type PGMs and the one (2.4.11) of composite-type CGMs, respectively.

(iv) *Mixed smoothness.* Suppose that the objective function f satisfies

$$f(x) \leq f(y) + \langle g(y), x - y \rangle + \frac{1}{2} L \|x - y\|^2 + \frac{M}{1+\nu} \|x - y\|^{1+\nu}, \quad \forall x, y \in Q$$

for a subgradient mapping $g(x) \in \partial f(x)$, constants $L, M \geq 0$, and $\nu \in [0, 1)$. Notice that this class of convex functions covers the one of both smooth and weakly smooth problems. The case $\nu = 0$ corresponds to the deterministic version of the Ghadimi-Lan's model [25, 26] (recall (2.4.13)).

For a convexity parameter $\sigma_f \in \sigma(f)$, define $m_f(y; x)$ by (4.2.4). Then we obtain the inequality (4.2.6) with

$$\bar{\sigma}_f := \sigma_f, \quad L(\cdot) := L, \quad \delta(y, x) := \frac{M}{1+\nu} \|x - y\|^{1+\nu}.$$

The subproblem (4.2.7) has the same form as the smooth problems (i).

(v) *Inexact oracle model.* Consider the Euclidean setting $\|\cdot\| = \|\cdot\|_2$, $d(x) = \frac{1}{2}\|x - x_0\|_2^2$. Suppose that $f(x)$ is equipped with a first-order (δ, L, μ)-oracle, that is, we can compute $(\bar{f}(y), \bar{g}(y)) \in \mathbb{R} \times E^*$ for each $y \in Q$ such that (2.4.14) holds. Then, we can define the lower approximation model

$$m_f(y; x) := \bar{f}(y) + \langle \bar{g}(y), x - y \rangle + \frac{\mu}{2} \|x - y\|_2^2$$

of $f(x)$ to obtain (4.2.6) with

$$\sigma_f := \bar{\sigma}_f := \mu, \quad L(\cdot) := L, \quad \delta(y, x) := \delta.$$

The subproblem (4.2.7) has the same form as the item (i).

□

4.3 Unifying framework for (sub)gradient-based methods

In this section, we introduce Properties A and B as a part of our unifying framework which auxiliary functions are assumed to satisfy. We at first introduce common notations below. They are compatible with the review of existing methods (Chapter 3).

The proposed methods are associated with the parameters $\{(\lambda_k, \beta_{k-1})\}_{k \geq 0}$ and functions $\{(\varphi_k(x), \psi_k(x))\}_{k \geq -1}$ where

- $\{\lambda_k\}_{k \geq 0}$ is a sequence of positive real numbers which we call the *weight parameters*.
- $\{\beta_k\}_{k \geq -1}$ is a nondecreasing sequence of nonnegative real numbers which we call the *scaling parameters*.
- $\{(\varphi_k(x), \psi_k(x))\}_{k \geq -1}$ is a *coupled sequence of auxiliary functions* which are minimized as subproblems at each iteration.

These parameters and functions can be determined in a recursive way during the method. Then the methods generate the sequence $\{(x_k, \hat{x}_k, z_{k-1}, w_{k-1})\}_{k \geq 0}$ where

- $x_k \in Q$ is a test point at which we evaluate $m_f(x_k; x)$ for $k \geq 0$.
- $z_k \in Q$ is a solution of the subproblem $\min_{x \in Q} \varphi_k(x)$ for $k \geq -1$.
- $w_k \in Q$ is a solution of the subproblem $\min_{x \in Q} \psi_k(x)$ for $k \geq -1$.
- $\hat{x}_k \in Q$ is an approximate solution to the problem (4.1.1) for $k \geq 0$.

We sometimes consider the case of a single sequence $\{\varphi_k(x)\}_{k \geq -1}$ of auxiliary functions regarding as $\psi_k(x) \equiv \varphi_k(x)$.

We also define

$$S_k := \sum_{i=0}^{k} \lambda_i \quad (k \geq 0), \quad S_{-1} := 0.$$

4.3.1 General properties for the construction of auxiliary functions in the unifying framework

Now we introduce the following framework for the auxiliary functions (We define $\sum_{i=0}^{-1}(\cdot) := 0$).

Property A (in the unifying framework)**.** *Suppose that the convex optimization problem (4.1.1) admits Assumption 4.2.5 with a lower approximation model $m_f(y; x)$ of $f(x)$ and a convexity parameter $\sigma_f \geq 0$. 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 $\{x_k\}_{k \geq 0}$. Denote $z_k := \operatorname{argmin}_{x \in Q} \varphi_k(x)$. Then, the following conditions hold:*

(A1) $\varphi_{-1}(z_{-1}) = 0$ *and* $z_{-1} = x_0$ $(:= \operatorname{argmin}_{x \in Q} d(x))$.

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

$$\varphi_{k+1}(x) \geq \varphi_k(z_k) + \lambda_{k+1} m_f(x_{k+1}; x) + \beta_{k+1} d(x) - \beta_k \ell_d(z_k; x) + S_k \sigma_f \xi(z_k, x). \quad (4.3.1)$$

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

We further consider a generalization of Property A to a coupled sequence of auxiliary functions as follow.

Property B (in the unifying framework)**.** *Suppose that the convex optimization problem (4.1.1) admits Assumption 4.2.5 with a lower approximation model $m_f(y; x)$ of $f(x)$ and a convexity parameter $\sigma_f \geq 0$. 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 $w_k := \operatorname{argmin}_{x \in Q} \psi_k(x)$. 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* $z_{-1} = w_{-1} = x_0$ $(:= \operatorname{argmin}_{x \in Q} d(x))$.

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

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

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

Notice that letting $\varphi_k(x) \equiv \psi_k(x)$ in Property B exactly yields Property A.

4.3.2 (Sub)gradient-based methods in the unifying framework

Under Properties A and B, we propose the following (sub)gradient-based methods for non-smooth problems $\mathcal{NSP}(g, \sigma_f)$ and the structured problems $\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$, respectively.

Method I (Subgradient-based methods for non-smooth problems in the unifying framework)**.**

Suppose that the convex optimization problem (4.1.1) belongs to the class $\mathcal{NSP}(g, \sigma_f)$. Let $\{\lambda_k\}_{k \geq 0}$ and $\{\beta_k\}_{k \geq -1}$ be sequences of weight and scaling parameters, respectively. 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 $g_k := g(x_k) \in \partial f(x_k)$ and compute $z_k, x_{k+1}, \hat{x}_{k+1}$ by

$$\begin{aligned} \text{Classical method} \quad &: \quad x_{k+1} := z_k := \operatorname{argmin}_{x \in Q} \varphi_k(x), \quad \hat{x}_{k+1} := \frac{S_k \hat{x}_k + \lambda_{k+1} z_k}{S_{k+1}}, \\ \text{or} & \\ \text{Modified method} \quad &: \quad z_k := \operatorname{argmin}_{x \in Q} \varphi_k(x), \quad \hat{x}_{k+1} := x_{k+1} := \frac{S_k \hat{x}_k + \lambda_{k+1} z_k}{S_{k+1}}. \end{aligned}$$

Here, $\{\varphi_k(x)\}_{k \geq -1}$ is a single sequence of auxiliary functions satisfying Property A. □

In Method I, the sequences $\{z_k\}_{k \geq -1}$ and $\{\hat{x}_k\}_{k \geq 0}$ can be reduced from the classical and the modified methods, respectively, where we kept them to preserve the notation. Notice that the update of $\{\hat{x}_k\}_{k \geq 0}$ has the following alternative expressions:

$$\hat{x}_{k+1} = \frac{S_k \hat{x}_k + \lambda_{k+1} z_k}{S_{k+1}} = \frac{1}{S_{k+1}} \sum_{i=-1}^{k} \lambda_{i+1} z_i = (1 - \tau_k) \hat{x}_k + \tau_k z_k$$

for $k \geq 0$, where $\tau_k := \lambda_{k+1} / S_{k+1}$. In particular, for the classical method, we also have

$$\hat{x}_k = \frac{1}{S_k} \sum_{i=0}^{k} \lambda_i x_i, \quad k \geq 0.$$

Method II (Gradient-based methods for structured problems in the unifying framework)**.** Suppose that the convex optimization problem (4.1.1) belongs to the class $\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$. Let $\{\lambda_k\}_{k \geq 0}$ and $\{\beta_k\}_{k \geq -1}$ be sequences of weight and scaling parameters, respectively. 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

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

Here, $\{(\varphi_k(x), \psi_k(x))\}_{k \geq 0}$ is a coupled sequence of auxiliary functions satisfying Property B. □

Again, we remark that $\{\hat{x}_k\}_{k \geq 0}$ can be expressed as the following alternative ways:

$$\hat{x}_{k+1} = \frac{S_k \hat{x}_k + \lambda_{k+1} w_{k+1}}{S_{k+1}} = \frac{1}{S_{k+1}} \sum_{i=0}^{k+1} \lambda_{i+1} w_{i+1} = (1 - \tau_k)\hat{x}_k + \tau_k w_{k+1} \tag{4.3.2}$$

for $k \geq 0$, where $\tau_k := \lambda_{k+1}/S_{k+1}$.

In order to obtain a particular instance of these methods, we need to specify the auxiliary functions $\{(\varphi_k(x), \psi_k(x))\}$ as well as the weight and scaling parameters. We discuss a concrete formula of the construction of the auxiliary functions next.

4.3.3 Concrete constructions of auxiliary functions

Here we develop recursive formulas to generate auxiliary functions satisfying Property A or B which can be used to obtain particular instances of Methods I and II.

The following result is crucial for the main consequences of our unifying framework.

Theorem 4.3.1. *Suppose that the convex optimization problem (4.1.1) admits Assumption 4.2.5 with a lower approximation model $m_f(y; x)$ of $f(x)$ and a convexity parameter $\sigma_f \geq 0$. Given the weight parameters $\{\lambda_k\}_{k \geq 0}$, the scaling parameters $\{\beta_k\}_{k \geq -1}$, and the test points $\{x_k\}_{k \geq 0}$, construct the sequence $\{\varphi_k(x)\}_{k \geq -1}$ of auxiliary functions by $\varphi_{-1}(x) := \beta_{-1} d(x)$, $z_{-1} := x_0$ and, for each $k \geq -1$,*

$$\varphi_{k+1}(x) := \theta_k \varphi_{k+1}^{\text{lower}}(x) + (1 - \theta_k)\varphi_{k+1}^{\text{upper}}(x)$$

where $\theta_k \in [0, 1]$ is arbitrary and

$$\begin{aligned} \varphi_{k+1}^{\text{lower}}(x) &:= \varphi_k(z_k) + \lambda_{k+1} m_f(x_{k+1}; x) + \beta_{k+1} d(x) - \beta_k \ell_d(z_k; x) + S_k \sigma_f \xi(z_k, x), \\ \varphi_{k+1}^{\text{upper}}(x) &:= \varphi_k(x) + \lambda_{k+1} m_f(x_{k+1}; x) + (\beta_{k+1} - \beta_k) d(x). \end{aligned} \tag{4.3.3}$$

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

Proof. The definitions $\varphi_{-1}(x) := \beta_{-1} d(x)$, $z_{-1} := x_0 := \operatorname{argmin}_{x \in Q} d(x)$ clearly ensures (A1) because of $d(x_0) = 0$.

Since $\sigma_f \in \sigma(m_f(x_i; \cdot))$, $\forall i \geq 0$ holds by Assumption 4.2.5, one can verify by induction that $\beta_k + S_k \sigma_f$ belongs to $\sigma(\varphi_k^{\text{lower}})$, $\sigma(\varphi_k^{\text{upper}})$, and so to $\sigma(\varphi_k)$ for all $k \geq -1$. Therefore, in view of Lemma 4.2.4 for the minimizer $z_k = \operatorname{argmin}_{x \in Q} \varphi_k(x)$, we obtain

$$\varphi_k(x) \geq \varphi_k(z_k) + (\beta_k + S_k \sigma_f)\xi(z_k, x) \tag{4.3.4}$$

for all $x \in Q$ and $k \geq -1$.

Showing the condition (A2) is equivalent to prove the inequality $\varphi_{k+1}(x) \geq \varphi_{k+1}^{\text{lower}}(x)$ for all $x \in Q$ and $k \geq -1$. Indeed, we obtain $\varphi_{k+1}^{\text{lower}}(x) \leq \varphi_{k+1}(x) \leq \varphi_{k+1}^{\text{upper}}(x)$, $\forall x \in Q$, $\forall k \geq -1$ because

$$\varphi_{k+1}^{\text{upper}}(x) - \varphi_{k+1}^{\text{lower}}(x) = \varphi_k(x) - \left[\varphi_k(z_k) + (\beta_k + S_k\sigma_f)\xi(z_k, x)\right] \overset{(4.3.4)}{\geq} 0.$$

Let us finally show the condition (A3). One can verify by induction the following inequality for each $k \geq -1$:

$$\varphi_k(x) \leq \sum_{i=0}^{k} \lambda_i m_f(x_i; x) + \beta_k d(x), \quad \forall x \in Q. \tag{4.3.5}$$

In fact, the right hand side of (4.3.5) is exactly the k-th auxiliary function constructed by the formula of the theorem with $\theta_k \equiv 0$ (that is, the one updated as $\varphi_{k+1}(x) := \varphi_{k+1}^{\text{upper}}(x)$ for all $k \geq -1$). As a result, we conclude that

$$\begin{aligned}
\varphi_k(z_k) &\overset{(4.3.4)}{\leq} \varphi_k(x) - (\beta_k + S_k\sigma_f)\xi(z_k, x) \\
&\overset{(4.3.5)}{\leq} \sum_{i=0}^{k} \lambda_i m_f(x_i; x) + \beta_k d(x) - (\beta_k + S_k\sigma_f)\xi(z_k, x) \\
&= \sum_{i=0}^{k} \lambda_i m_f(x_i; x) + \beta_k \ell_d(z_k; x) - S_k\sigma_f\xi(z_k, x)
\end{aligned}$$

for all $x \in Q$ and $k \geq -1$, which shows the condition (A3). □

As a simple consequence of Theorem 4.3.1, we obtain the following construction of a coupled sequence of auxiliary functions satisfying Property B.

Corollary 4.3.2. *Under the assumption in Theorem 4.3.1, define the sequence* $\{\psi_k(x)\}_{k\geq -1}$ *by* $\psi_{-1}(x) := \varphi_{-1}(x)$ *and the recurrence*

$$\psi_{k+1}(x) := \vartheta_k\varphi_{k+1}(x) + (1 - \vartheta_k)\varphi_{k+1}^{\text{lower}}(x) \tag{4.3.6}$$

for an arbitrary $\vartheta_k \in [0, 1]$. *Then, the sequence* $\{(\varphi_k(x), \psi_k(x))\}_{k\geq -1}$ *satisfies Property B.*

Proof. By Theorem 4.3.1, $\{\varphi_k(x)\}_{k\geq -1}$ satisfies Property A. Notice that we have

$$\varphi_{k+1}^{\text{lower}}(x) \leq \psi_{k+1}(x) \leq \varphi_{k+1}(x) \leq \varphi_{k+1}^{\text{upper}}(x) \quad \forall x \in Q, \ \forall k \geq -1$$

by the proof of Theorem 4.3.1. Therefore, (B0) to (B3) immediately follow (use (A3) to obtain (B3)). □

We particularly consider three special cases of Theorem 4.3.1 and Corollary 4.3.2 below, which are important to relate to existing (sub)gradient-based methods.

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

  $$\varphi_{k+1}(x) := \varphi_k(z_k) + \lambda_{k+1}m_f(x_{k+1}; x) + \beta_{k+1}d(x) - \beta_k\ell_d(z_k; x) + S_k\sigma_f\xi(z_k, x) \tag{4.3.7}$$

  for $k \geq -1$. Then, Property A follows from Theorem 4.3.1 with $\theta_k \equiv 1$ (namely, $\varphi_{k+1} \equiv \varphi_{k+1}^{\text{lower}}$, $\forall k \geq -1$).

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

$$\varphi_k(x) := \sum_{i=0}^{k} \lambda_i m_f(x_i; x) + \beta_k d(x) \tag{4.3.8}$$

  for $k \geq 0$. Then, Property A follows from Theorem 4.3.1 with $\theta_k \equiv 0$ (namely, $\varphi_{k+1} \equiv \varphi_{k+1}^{\mathrm{upper}}$, $\forall k \geq -1$).

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

$$\begin{array}{rcl} \varphi_k(x) & := & \sum_{i=0}^{k} \lambda_i m_f(x_i; x) + \beta_k d(x), \\ \psi_{k+1}(x) & := & \varphi_k(z_k) + \lambda_{k+1} m_f(x_{k+1}; x) + \beta_{k+1} d(x) - \beta_k \ell_d(z_k; x) + S_k \sigma_f \xi(z_k, x) \end{array} \tag{4.3.9}$$

  for $k \geq -1$. Then, Property B follows from Corollary 4.3.2 with $\theta_k \equiv \vartheta_k \equiv 0$ (namely, $\varphi_{k+1} \equiv \varphi_{k+1}^{\mathrm{upper}}$ and $\psi_{k+1} \equiv \varphi_{k+1}^{\mathrm{lower}}$ for $k \geq -1$). As an alternative, Property B is also satisfied when we use the hybrid update (4.3.9) except initializing $\varphi_0(x) = \psi_0(x)$ with the DA update (4.3.8) (take $\vartheta_{-1} = 1$ and $\vartheta_k = 0$ for $k \geq 0$ in Corollary 4.3.2).

The EMD and the DA models yield four particularizations of Method I combining with the classical and the modified updates. Notice that, in this case, the subproblem $z_k := \mathrm{argmin}_{x\in Q} \varphi_k(x)$ at each iteration is of the form (4.2.5) because of the definition (4.2.4) of $m_f(y; x)$. In particular, if $\beta_k \equiv 0$, $\sigma_f = 0$, and if $m_f(y, \cdot)$ is an affine function, then the subproblem $z_k := \mathrm{argmin}_{x\in Q} \varphi_k(x)$ as well as $w_k := \mathrm{argmin}_{x\in Q} \psi_k(x)$ with the above models becomes a minimization of an affine function which will yield an instance of CGM.

Method II gives six particularizations due to the additional choice of the hybrid model. Remark that employing the EMD and the DA models in Method II reduces the number of subproblems at each iteration since $z_k \equiv w_k$. Note that only the EMD model turns the subproblem $w_k = z_k := \mathrm{argmin}_{x\in Q} \varphi_k(x)$ of the form (4.2.7) among the above three models; the others require the solution of the subproblem of minimizing the function (4.3.8). However, the subproblems with these three models have the same computational difficulty for all the examples in Example 4.2.8.

4.3.4 Particular instances of general methods in the unifying framework

We demonstrate that Methods I and II equipped with the above models of auxiliary functions for particular classes of optimization problems yield existing methods reviewed in Chapter 3.

Example 4.3.3. Consider a non-smooth problem in the class $\mathcal{NSP}(g, \sigma_f)$. Let us see that Method I with the EMD and the DA models yield the MDM, the DAM, and a variant of the DAM.

(1) Let us consider the non strongly convex case $\sigma_f = 0$.

(1a) *Mirror-descent method.* Consider the auxiliary functions $\{\varphi_k(x)\}_{k\geq -1}$ defined by the EMD model (4.3.7). If $\beta_k \equiv 1$, then $\{\varphi_k(x)\}_{k\geq -1}$ is defined by $\varphi_{-1}(x) = d(x)$ and

$$\varphi_{k+1}(x) = \varphi_k(z_k) + \lambda_{k+1}[f(x_k) + \langle g_k, x - x_k\rangle] + \xi(z_k, x)$$

for $k \geq -1$. Therefore, the sequence $\{x_k\}_{k\geq 0}$ of test points generated by the classical method in Method I in this case is exactly the one generated by the MDM (3.1.1).

Therefore, the classical method in Method I associated with the EMD model (4.3.7) can be seen as a generalization of the MDM introducing the scaling parameters $\{\beta_k\}_{k\geq -1}$. We call this method the *extended mirror-descent method*.

(1b) *Dual-averaging method.* Consider Method I with the auxiliary functions $\{\varphi_k(x)\}_{k\geq -1}$ defined by the DA model (4.3.8). Then, the classical method yields the Nesterov's DAM (3.1.9) and the modified method yields the Nesterov-Shikhman's double averaging method (3.1.14).

(2) *Mirror-descent method (strongly convex case).* In the strongly convex case $\sigma_f > 0$, consider the classical method of Method I with the auxiliary functions $\{\varphi_k(x)\}_{k\geq -1}$ defined by the EMD model (4.3.7). Then, the sequence $\{x_k\}_{k\geq 0}$ is computed as follow.

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

This iteration corresponds to the MDM (3.1.1) with the weight parameters $\{\tilde{\lambda}_k\}_{k\geq 0}$ defined by $\tilde{\lambda}_k := \frac{\lambda_k}{\sigma_f S_k}$. The approximate solution $\hat{x}_k = \frac{1}{S_k}\sum_{i=0}^k \lambda_i x_i$ of Method I coincides with the Nedić-Lee's averaging $\widetilde{x}_k = \left(\sum_{i=0}^k \tilde{\lambda}_k^{-1}\right)^{-1}\sum_{i=0}^k \tilde{\lambda}_k^{-1} x_i$ (3.1.5) when $\lambda_0 := 1$ and $\lambda_{k+1} := \frac{1+\sqrt{1+4\lambda_k^2}}{2}$ $(k \geq 0)$ because we have $\lambda_k^2 = S_k$ and so $\tilde{\lambda}_k = \frac{1}{\sigma_f\lambda_k}$ (see Lemma A.4 (i) in Appendix). With the another choice $\lambda_k := \frac{k+1}{2}$, we have $\tilde{\lambda}_k = \frac{1}{\sigma_f(k+2)}$ and the approximate solution $\hat{x}_k = \frac{2}{(k+1)(k+2)}\sum_{i=0}^k (i+1)x_i$ which coincides with the Bach's weighted average (3.1.8).

(3) *Dual-averaging method (strongly convex case).* Consider the auxiliary functions $\{\varphi_k(x)\}_{k\geq -1}$ defined by the DA model (4.3.8). The corresponding subproblems become

$$z_k := \operatorname*{argmin}_{x\in Q}\left\{\sum_{i=0}^k \lambda_i[f(x_i) + \langle g(x_i), x - x_i\rangle + \sigma_f\xi(x_i, x)] + \beta_k d(x)\right\}.$$

Therefore, Method I in this case yields extensions of the DAM (3.1.9) and its variant (3.1.14) to the strongly convex case $\sigma_f > 0$.

□

Example 4.3.4. Let us see that Method II with the models (4.3.7), (4.3.8), and (4.3.9) yield some existing methods for particular structured problems.

(1) Consider the smooth problem as Example 4.2.8 (i). Let us consider the modified method of Method II with $\beta_k \equiv L/\sigma_d$. If we equip the EMD model (4.3.7), the corresponding auxiliary function is given by $\varphi_{-1}(x) := \frac{L}{\sigma_d}d(x)$ and

$$\begin{aligned} \varphi_{k+1}(x) &:= \varphi(z_k) + \lambda_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle + \sigma_f\xi(x_{k+1}, x)] \\ &\quad + \left(\frac{L}{\sigma_d} + S_k\sigma_f\right)\xi(z_k, x). \end{aligned}$$

Therefore, in the non strongly convex case $\sigma_f = 0$, we see that taking $\lambda_0 := 1$, $\lambda_{k+1} := \frac{1+\sqrt{1+4\lambda_k^2}}{2}$ for $k \geq 0$ yields the Tseng's second APG method (3.2.10). In the strongly convex case $\sigma_f > 0$, the corresponding algorithm can be seen as an extension of the Tseng's second APG method. The Nesterov's modified method, the Tseng's third APG method, and the Lan's variants of CGMs can be obtained in a similar way as we summarize their correspondences in items (1a) to (1e) below:

(1a) Define the auxiliary function $\{(\varphi_k(x), \psi_k(x))\}_{k\geq -1}$ by the hybrid model (4.3.9) except initializing $\varphi_0(x) = \psi_0(x)$ by the DA model (4.3.8). The modified method of Method II with this auxiliary functions in the case $\sigma_f = 0$, $\beta_k \equiv L/\sigma_d$, $\lambda_k := \frac{k+1}{2}$ yields the Nesterov's modified method (3.2.9).

(1b) The modified method of Method II with the EMD model (4.3.7) in the case $\sigma_f = 0$, $\beta_k \equiv L/\sigma_d$, $\lambda_0 := 1$, $\lambda_{k+1} := \frac{1+\sqrt{1+4\lambda_k^2}}{2}$ yields the Tseng's second APG method (3.2.10).

(1c) The modified method of Method II with the DA model (4.3.8) in the case $\sigma_f = 0$, $\beta_k \equiv L/\sigma_d$, $\lambda_0 := 1$, $\lambda_{k+1} := \frac{1+\sqrt{1+4\lambda_k^2}}{2}$ yields the Tseng's third APG method (3.2.11).

(1d) The modified method of Method II with the EMD model (4.3.7) in the case $\sigma_f = 0$, $\beta_k \equiv 0$, $\lambda_{k+1} := \frac{k+1}{2}$ yields the Lan's primal averaging CGM (3.2.21). Notice in (4.3.2) that we have $S_k = \frac{(k+1)(k+2)}{2}$ and so $\tau_k = \lambda_{k+1}/S_{k+1} = \frac{2}{k+3}$.

(1e) The modified method of Method II with the DA model (4.3.8) in the case $\sigma_f = 0$, $\beta_k \equiv 0$, $\lambda_{k+1} := \frac{k+1}{2}$ yields the Lan's primal dual averaging CGM (3.2.22).

(2) Consider the composite problem $\min_{x\in Q}[f(x) \equiv f_0(x) + \Psi(x)]$ as Example 4.2.8 (iii).

(2a) Let us see that the classical method of Method II with the EMD model (4.3.7) in the case

$$\beta_k \equiv \frac{L - \bar{\sigma}_f \sigma_d}{\sigma_d}, \quad \lambda_0 := 1, \quad \lambda_{k+1} := \frac{\beta_k + S_k \sigma_f}{\beta_k} \tag{4.3.10}$$

includes the primal gradient method (3.2.1). In fact, since $\frac{L}{\sigma_d}\lambda_{k+1} = \bar{\sigma}_f \lambda_{k+1} + \beta_k + S_k\sigma_f$ hold for $k \geq -1$, the auxiliary function with the EMD model is given by

$$\begin{aligned}
\varphi_k(x) &= \varphi_{k-1}(z_{k-1}) + \lambda_k[f_0(x_k) + \langle \nabla f_0(x_k), x - x_k\rangle + \bar{\sigma}_f \xi(x_k, x) + \Psi(x)] \\
&\quad + (\beta_{k-1} + S_{k-1}\sigma_f)\xi(x_k, x) \\
&= \varphi_{k-1}(z_{k-1}) + \lambda_k \left(f_0(x_k) + \langle \nabla f_0(x_k), x - x_k\rangle + \Psi(x) + \frac{L}{\sigma_d}\xi(x_k, x) \right)
\end{aligned}$$

from which the update formula $x_{k+1} := z_k = \operatorname{argmin}_{x\in Q} \varphi_k(x)$ yields the primal gradient method (3.2.1) in the Euclidean setting $d(x) = \frac{1}{2}\|x - x_0\|_2^2$.
It is interesting to see that the update $x_{k+1} = \operatorname{argmin}_{x\in Q}\{f_0(x_k) + \langle \nabla f_0(x_k), x - x_k\rangle + \Psi(x) + \frac{L}{\sigma_d}\xi(x_k, x)\}$ does not require to know parameters $\bar{\sigma}_f$ and σ_f while the parameters β_k and λ_k in (4.3.10) involve them.

(2b) Let $\{(\varphi_k(x), \psi_k(x))\}_{k\geq -1}$ be generated by the hybrid model (4.3.9) with the choice (4.3.10) of parameters. Then, for $k \geq 0$, we have

$$\varphi_k(x) = \sum_{i=0}^{k} \lambda_i [f_0(x_i) + \langle \nabla f_0(x_i), x - x_i \rangle + \bar{\sigma}_f \xi(x_i, x) + \Psi(x)] + \beta_k d(x).$$

In the Euclidean setting $d(x) = \frac{1}{2}\|x - x_0\|_2^2$ and the non strongly convex case $\bar{\sigma}_f = \sigma_f = 0$, the classical method of Method II taking $\lambda_k \equiv 1$ and $\beta_k \equiv L$ yields the same sequence $\{x_k\}_{k\geq 0}$ as the dual gradient method (3.2.2). As the same way as (2a), we also have $w_k = \operatorname{argmin}_{x\in Q}\{f_0(x_k) + \langle \nabla f_0(x_k), x - x_k\rangle + \frac{L}{2}\|x - x_k\|_2^2 + \Psi(x)\}$ in this case which is compatible with the notation w_k in (3.2.3).

The DA model (4.3.8) also generates the same $\{\varphi_k(x)\}$ as above. Then, the sequence $\{x_k\}_{k\geq 0}$ of test points generated by the classical method in Method II are the same as the one of the dual gradient method (3.2.2) while the w_k does not coincides with the one (3.2.3) since $w_k = z_k = \operatorname{argmin}_{x\in Q} \varphi_k(x)$. Therefore, in this case, the DA model reduces the number of subproblems from two to one compared with the hybrid model (but they generate different approximate solutions).

(3) Consider the convex optimization problem with the inexact oracle model as Example 4.2.8 (v) (then, $\sigma_f = \bar{\sigma}_f = \mu,\ \sigma_d = 1$).

(3a) Similar to (2a), the classical method of Method II with EMD model (4.3.7) and with the choice (4.3.10) of parameters yields the primal gradient method (3.2.5).

(3b) The classical method of Method II with the hybrid model (4.3.9) yields the dual gradient method (3.2.6) when we choose constant $\beta_k \equiv \beta > 0$: in fact, the corresponding auxiliary function $\varphi_k(x)$ is given by

$$\varphi_k(x) = \sum_{i=0}^{k} \lambda_i \left[\bar{f}(x_i) + \langle \bar{g}(x_i), x - x_i\rangle + \frac{\mu}{2}\|x - x_i\|_2^2\right] + \frac{\beta}{2}\|x - x_0\|_2^2$$

from which the update $x_{k+1} := z_k = \operatorname{argmin}_{x\in Q} \varphi_k(x)$ is compatible with the one (3.2.6) (the weight parameters correspond up to multiplication). If we further choose the parameters by (4.3.10), we obtain the compatibility with the notation $w_k = \operatorname{argmin}_{x\in Q}\{\bar{f}(x_k) + \langle \bar{g}(x_k), x - x_k\rangle + \frac{L}{2}\|x - x_k\|_2^2\}$ as (2a). We remark that, in general, the choice (4.3.10) of parameters is not equivalent to the one (3.2.8) analyzed in [17]; we will suggest (4.3.10) as an alternative choice (Theorem 4.6.1). □

As discussed in Example 4.3.3 (1a) above, we propose the following new extension of the MDM in the non strongly convex case.

Method 4.3.5 (extended mirror-descent method for $\mathcal{NSP}(g, 0)$)**.** Suppose that we know a subgradient mapping $g(x) \in \partial f(x),\ x \in Q$ in the convex optimization problem (4.1.1). Let $\{\lambda_k\}_{k\geq 0}$ and $\{\beta_k\}_{k\geq -1}$ be sequences of weight and scaling parameters, respectively. Starting from $x_0 := z_{-1} := \operatorname{argmin}_{x\in Q} d(x)$, iterate

$$g_k := g(x_k) \in \partial f(x_k), \quad x_{k+1} := \operatorname*{argmin}_{x\in Q}\{\lambda_k[f(x_k) + \langle g_k, x - x_k\rangle] + \beta_k d(x) - \beta_{k-1}\ell_d(z_{k-1}; x)\}$$

for $k \geq 0$. Define the sequence $\{\hat{x}_k\}_{k\geq 0}$ of approximate solutions by

$$\hat{x}_k := \frac{1}{S_k} \sum_{i=0}^{k} \lambda_i x_i, \quad k \geq 0.$$

□

In particular, the original MDM (3.1.1) is obtained by letting $\beta_k \equiv 1$ in the extended MDM.

Due to Theorem 4.3.1 and Corollary 4.3.2, we can provide infinitely many instances of PGMs and CGMs via Methods I and II. In Table 4.1 below, We summarize important cases given by the EMD, the DA, and the hybrid models.

Table 4.1: Particular instances of the proposed methods. The column 'Aux. func.' corresponds to the model of auxiliary functions defined in Section 4.3.3. The star (∗) is attached for new methods. In particular, '*an extension' means a new extension of the left to the strongly convex case. The dagger symbol (†) means that our method and the existing one *shares* a particular instance.

Method type	Aux. func.	Non strongly convex case	Strongly convex case
classical method of Method I	DA	dual-averaging [52]	*an extension
	EMD	mirror-descent [46] *extended MDM (Method 4.3.5)	†Nedić-Lee's averaging [43] Bach's averaging [3]
modified method of Method I	DA	double averaging [56]	*an extension
	EMD	*double averaging for the MDM	*an extension
classical method of Method II	EMD	primal gradient method [18, 53] (without line search)	
	Hybrid	dual gradient method [18, 53] (without line search)	
	DA	*a variant of the dual gradient method [18, 53]	
modified method of Method II	EMD	†Tseng's second APG [58]	*an extension
		†Lan's primal averaging CGM [35]	—
	DA	†Tseng's third APG [58]	*an extension
		†Lan's primal dual averaging CGM [35]	—
	Hybrid (+DA)	Nesterov's modified method [50]	*an extension

4.4 General convergence estimates of subgradient-based methods in the unifying framework

In this section, we prove a general convergence estimates of Methods I and II for the non-smooth and the structured problems. These results will be used to derive particular rates of convergence in the next sections.

We will obtain different convergence estimates for the classical and the modified methods. We show in Section 4.5 that they have the same rate of convergence for the non-smooth problems but, for the smooth problems, the modified method gives much better efficiency than the classical method as discussed in Sections 4.6.

In the remainder of this section, we aim to prove the following general estimates, Theorems 4.4.1 and 4.4.2, after which we can focus on the choice of parameters $\{\lambda_k\}$ and $\{\beta_k\}$ to ensure an efficient convergence rate.

Theorem 4.4.1. *Consider a non-smooth problem in the class $\mathcal{NSP}(g, \sigma_f)$. Let $\{(z_{k-1}, x_k, g_k, \hat{x}_k)\}_{k\geq 0}$ be generated by Method I associated with weight parameters $\{\lambda_k\}_{k\geq 0}$ and scaling parameters $\{\beta_k\}_{k\geq -1}$. Then, for every $k \geq 0$, the estimate*

$$f(\hat{x}_k) - f(x^*) + \sigma_f \xi(z_k, x^*) \leq \frac{\beta_k \ell_d(z_k; x^*) + C_k}{S_k} \tag{4.4.1}$$

holds, where

$$C_k := \begin{cases} \frac{1}{2\sigma_d} \sum_{i=0}^k \frac{\lambda_i^2}{\beta_{i-1} + S_i \sigma_f} \|g_i\|_*^2 & \text{for the classical method; and} \\ \frac{1}{2\sigma_d} \sum_{i=0}^k \frac{\lambda_i^2 S_i}{\lambda_i^2 \sigma_f + S_i(\beta_{i-1} + S_{i-1}\sigma_f)} \|g_i\|_*^2 & \text{for the modified method.} \end{cases} \tag{4.4.2}$$

Furthermore, for every $k \geq 0$, 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 $\min_{0\leq i\leq k} f(x_i) - f(x^*) + \sigma_f \xi(z_k, x^*)$ for the classical method.

Theorem 4.4.2. *Consider a structured problem in the class $\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$. Let $\{(z_{k-1}, w_{k-1}, x_k, \hat{x}_k)\}_{k\geq 0}$ be generated by Method II associated with weight parameters $\{\lambda_k\}_{k\geq 0}$ and scaling parameters $\{\beta_k\}_{k\geq -1}$. Then, for every $k \geq 0$, the estimate*

$$f(\hat{x}_k) - f(x^*) + \sigma_f \xi(z_k, x^*) \leq \frac{\beta_k \ell_d(z_k; x^*) + C_k}{S_k} \tag{4.4.3}$$

holds, where

$$C_k := \begin{cases} \frac{1}{2} \sum_{i=0}^k \lambda_i \left(L(x_i) - \sigma_d \left(\bar{\sigma}_f + \frac{\beta_{i-1} + S_{i-1}\sigma_f}{\lambda_i} \right) \right) \|w_i - x_i\|^2 + \sum_{i=0}^k \lambda_i \delta(x_i, w_i) \\ \qquad\qquad \text{for the classical method; and} \\ \frac{1}{2} \sum_{i=0}^k S_i \left(L(x_i) - \sigma_d \left(\bar{\sigma}_f + \frac{S_i(\beta_{i-1} + S_{i-1}\sigma_f)}{\lambda_i^2} \right) \right) \|\hat{x}_i - x_i\|^2 + \sum_{i=0}^k S_i \delta(x_i, \hat{x}_i) \\ \qquad\qquad \text{for the modified method.} \end{cases} \tag{4.4.4}$$

Furthermore, for every $k \geq 0$, 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 $\min_{0\leq i\leq k} f(w_i) - f(x^*) + \sigma_f \xi(z_k, x^*)$ for the classical method.

Remark 4.4.3. Method II with $\sigma_f = \bar{\sigma}_f = 0$ and $\beta_k \equiv 0$ yields several versions of CGMs because the constructed auxiliary functions are non-negative linear combinations of constants and $\{m_f(x_i; x)\}_{i=0}^k$. In this case, Theorem 4.4.2 implies that the modified method ensures

$$f(\hat{x}_k) - f(x^*) \leq \frac{C_k}{S_k} \leq \frac{\frac{1}{2}\mathrm{Diam}(Q)^2 \sum_{i=0}^k L(x_i) \frac{\lambda_i^2}{S_i}}{S_k} + \frac{\sum_{i=0}^k S_i \delta(x_i, \hat{x}_i)}{S_k} \tag{4.4.5}$$

for all $k \geq 0$, because $\|\hat{x}_i - x_i\|^2 = \frac{\lambda_i^2}{S_i^2} \|w_i - z_{i-1}\|^2 \leq \frac{\lambda_i^2}{S_i^2} \mathrm{Diam}(\mathrm{Q})^2$. This bound resembles with the result [22, Theorem 5.3] of the classical CGM for the inexact oracle model. In fact, if $m_f(y; \cdot)$ is affine for each $y \in Q$ (say, $m_f(y; x) = \bar{f}(y) + \langle \bar{g}(y), x - y \rangle$), then the classical

CGM (3.2.19) can be arranged by replacing ∇f with $\bar{g}$ so that it is applicable to structured problems in $\mathcal{SP}(m_f, 0, 0, L, \delta)$. Then, taking parameters $\tau_k := \lambda_{k+1}/S_{k+1}$ and $\hat{x}_k := x_k$, the arranged classical CGM admits the estimate[1]

$$f(\hat{x}_k) - f(x^*) \leq \frac{\lambda_0[f(x_0) - f(x^*)]}{S_k} + \frac{\frac{1}{2}\mathrm{Diam}(Q)^2 \sum_{i=1}^k L(x_{i-1})\frac{\lambda_i^2}{S_i}}{S_k} + \frac{\sum_{i=1}^k S_i \delta(x_{i-1}, x_i)}{S_k} \tag{4.4.6}$$

for all $k \geq 0$. □

4.4.1 Key strategy of the proof

Although we have firstly shown the descriptions of Methods I and II, they can be derived as a consequence of the discussion in this section. Moreover, we simultaneously obtain their general convergence estimate shown in Theorems 4.4.1 and 4.4.2. Therefore, our observation is taken under a general assumption rather than the ones in Theorems 4.4.1 and 4.4.2.

For non-smooth or structured problems, we generally consider a coupled sequence $\{(\varphi_k(x), \psi_k(x))\}_{k\geq -1}$ of auxiliary functions satisfying Property B associated with weight parameters $\{\lambda_k\}_{k\geq 0}$, scaling parameters $\{\beta_k\}_{k\geq -1}$, and test points $\{x_k\}_{k\geq 0}$. Then, we try to find a sequence $\{\hat{x}_k\} \subset Q$ and constants $\{C_k\}_{k\geq 0}$ satisfying the following relation

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

for each $k \geq 0$. We use this relation to prove the estimates (4.4.1) and (4.4.3).

We also consider the alternative relations

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

to prove the latter assertion of Theorems 4.4.1 and 4.4.2, respectively.

These relations yield the following estimate.

Lemma 4.4.4. *Suppose that the convex optimization problem (4.1.1) admits Assumption 4.2.5 with a lower approximation model $m_f(y; x)$ of $f(x)$ and a convexity parameter $\sigma_f \geq 0$. Suppose further 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, then we have*

$$f(\hat{x}_k) - f(x) + \sigma_f \xi(z_k, x) \leq \frac{\beta_k \ell_d(z_k; x) + C_k}{S_k}, \quad \forall x \in Q. \tag{4.4.7}$$

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

$$\begin{aligned} S_k f(\hat{x}_k) &\leq \sum_{i=0}^k \lambda_i m_f(x_i; x) + \beta_k \ell_d(z_k; x) - S_k \sigma_f \xi(z_k, x) + C_k \\ &\leq S_k f(x) + \beta_k \ell_d(z_k; x) - S_k \sigma_f \xi(z_k, x) + C_k. \end{aligned}$$

□

[1] The proof of [22, Theorem 5.3] (with $B_0 = h^*$ so that $B_k = h^*$) replacing the notation $(h(\cdot), \lambda_{k+1}, \tilde{\lambda}_{k+1}, L_{k+1}, \delta_{k+1}, \bar{\alpha}_{k+1}, \beta_{k+1}, \alpha_k)$ of [22] 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 [22, 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 (4.2.6) at $(x, y) = (x_k, x_{k+1})$.

Remark 4.4.5. (1) Analogues of Lemma 4.4.4 easily show that (P_k) and (B3) imply the inequality

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

for $x \in Q$. The conditions (Q_k) and (B3) also conclude the same replacing x_i by w_i.
(2) When $\sigma_f > 0$, (4.4.7) 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 (4.4.7) implies

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

□

Lemma 4.4.4 and Remark 4.4.5 (1) suggest us to prove (R_k) and its variants (P_k) or (Q_k) in order to complete Theorems 4.4.1 and 4.4.2 (as detailed in Section 4.4.5). We now turn an induction to establish them.

4.4.2 Validity of (R_k), (P_k), and (Q_k) when $k = 0$

We start our induction in the case $k = 0$. Note that the settings of (i) and (ii) in the following lemma are exactly the situations of the initialization step (0) in Methods I and II, respectively.

Lemma 4.4.6. *(i) Consider a non-smooth problem in the class $\mathcal{NSP}(g, \sigma_f)$ and let $\{(\varphi_k(x), \psi_k(x))\}_{k\ge -1}$ be a coupled sequence of auxiliary functions satisfying Property B associated with weight parameters $\{\lambda_k\}_{k\ge 0}$, scaling parameters $\{\beta_k\}_{k\ge -1}$, and test points $\{x_k\}_{k\ge 0}$. Then, the relation $(R_0) \equiv (P_0)$ is satisfied with $\hat{x}_0 := x_0$ and*

$$C_0 := \frac{1}{2}\frac{\lambda_0^2}{\sigma_d(\lambda_0\sigma_f + \beta_{-1})}\|g_0\|_*^2. \tag{4.4.8}$$

(ii) Consider a structured problem in the class $\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$ and let $\{(\varphi_k(x), \psi_k(x))\}_{k\ge -1}$ be a coupled sequence of auxiliary functions satisfying Property B associated with weight parameters $\{\lambda_k\}_{k\ge 0}$, scaling parameters $\{\beta_k\}_{k\ge -1}$, and test points $\{x_k\}_{k\ge 0}$. Then, the relation $(R_0) \equiv (Q_0)$ is satisfied with $\hat{x}_0 := w_0$ and

$$C_0 := \lambda_0\left(\frac{L(x_0)}{2} - \frac{\sigma_d}{2}\left(\bar{\sigma}_f + \frac{\beta_{-1}}{\lambda_0}\right)\right)\|w_0 - x_0\|^2 + \lambda_0\delta(x_0, \hat{x}_0). \tag{4.4.9}$$

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

$$\begin{aligned}\psi_{k+1}(w_{k+1}) &\ge \varphi_k(z_k) + \lambda_{k+1} m_f(x_{k+1}; w_{k+1}) + (\beta_k + S_k\sigma_f)\xi(z_k, w_{k+1}) \\ &\ge \psi_k(w_k) + \lambda_{k+1} m_f(x_{k+1}; w_{k+1}) + (\beta_k + S_k\sigma_f)\xi(z_k, w_{k+1}) \end{aligned} \tag{4.4.10}$$

for every $k \ge -1$. Let $\sigma \ge 0$ be an arbitrary nonnegative number. Then, the conditions (B1) and $S_{-1} = 0$ lead (4.4.10) with $k = -1$ to

$$\begin{aligned}\psi_0(w_0) &\ge \lambda_0\left[m_f(x_0; w_0) - \sigma\xi(x_0, w_0) + \left(\sigma + \frac{\beta_{-1}}{\lambda_0}\right)\xi(x_0, w_0)\right] \\ &\ge \lambda_0\left[m_f(x_0; w_0) - \sigma\xi(x_0, w_0) + \frac{\sigma_d}{2}\left(\sigma + \frac{\beta_{-1}}{\lambda_0}\right)\|w_0 - x_0\|^2\right]. \end{aligned} \tag{4.4.11}$$

Let us firstly show (ii). Letting $\sigma := \bar{\sigma}_f$, the settings $\hat{x}_0 = w_0$ and (4.4.9) yields

$$\psi_0(w_0) + C_0 \overset{(4.4.11)}{\geq} \lambda_0 \left[m_f(x_0; w_0) - \bar{\sigma}_f \xi(x_0, \hat{x}_0) + \frac{L(x_0)}{2} \|\hat{x}_0 - x_0\|^2 + \delta(x_0, \hat{x}_0) \right] \geq \lambda_0 f(\hat{x}_0)$$

which proves the relation (R_0).

It remains to prove (i). By the definition (4.2.4) of $m_f(\cdot;\cdot)$ for the non-smooth case, the inequality (4.4.11) with $\sigma := \sigma_f$ implies

$$\begin{aligned} \psi_0(w_0) &\overset{(4.4.11)}{\geq} \lambda_0 \left[f(x_0) + \langle g_0, w_0 - x_0 \rangle + \frac{\sigma_d}{2} \left(\sigma_f + \frac{\beta_{-1}}{\lambda_0} \right) \|w_0 - x_0\|^2 \right] \\ &= \lambda_0 f(x_0) + \langle \lambda_0 g_0, w_0 - x_0 \rangle + \frac{\sigma_d}{2} (\lambda_0 \sigma_f + \beta_{-1}) \|w_0 - x_0\|^2 \\ &\geq \lambda_0 f(x_0) - \frac{1}{2} \frac{\lambda_0^2}{\sigma_d (\lambda_0 \sigma_f + \beta_{-1})} \|g_0\|_*^2, \end{aligned}$$

where the last inequality is due to the basic fact

$$\frac{1}{2} \|x\|^2 + \frac{1}{2} \|s\|_*^2 \geq \langle s, x \rangle \quad \text{for } x \in E, \ s \in E^*. \tag{4.4.12}$$

This means that the relation (R_0) is satisfied with the setting $\hat{x}_0 = x_0$ and (4.4.8). □

4.4.3 Validity of (R_k), (P_k), and (Q_k) for the classical method when $k > 0$

Let us complete our induction for the classical method. The items (i) and (ii) in the following lemma correspond to the k-th iteration of the classical method in Methods I and II, respectively.

Lemma 4.4.7. *(i) Consider a non-smooth problem in the class $\mathcal{NSP}(g, \sigma_f)$ and let $\{(\varphi_k(x), \psi_k(x))\}_{k \geq -1}$ be a coupled sequence of auxiliary functions satisfying Property B associated with weight parameters $\{\lambda_k\}_{k \geq 0}$, scaling parameters $\{\beta_k\}_{k \geq -1}$, and test points $\{x_k\}_{k \geq 0}$. Suppose for $k \geq 0$ that the relation (R_k) is satisfied for some $\hat{x}_k \in Q$, $C_k \geq 0$. If the relation $x_{k+1} = z_k$ holds, then the relation (R_{k+1}) is satisfied with $\hat{x}_{k+1} := \frac{S_k \hat{x}_k + \lambda_{k+1} x_{k+1}}{S_{k+1}}$ and*

$$C_{k+1} := C_k + \frac{1}{2\sigma_d} \frac{\lambda_{k+1}^2}{\beta_k + S_{k+1} \sigma_f} \|g_{k+1}\|_*^2. \tag{4.4.13}$$

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) Consider a structured problem in the class $\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$ and let $\{(\varphi_k(x), \psi_k(x))\}_{k \geq -1}$ be a coupled sequence of auxiliary functions satisfying Property B associated with weight parameters $\{\lambda_k\}_{k \geq 0}$, scaling parameters $\{\beta_k\}_{k \geq -1}$, and test points $\{x_k\}_{k \geq 0}$. Suppose for $k \geq 0$ that the relation (R_k) is satisfied for some $\hat{x}_k \in Q$, $C_k \geq 0$. If the relation $x_{k+1} = z_k$ holds, then the relation (R_{k+1}) is satisfied with $\hat{x}_{k+1} := \frac{S_k \hat{x}_k + \lambda_{k+1} w_{k+1}}{S_{k+1}}$ and

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

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

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

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.

$$\begin{aligned}
\psi_{k+1}(w_{k+1}) &\geq \psi_k(w_k) + \lambda_{k+1} f(x_{k+1}) \\
&\quad + \langle \lambda_{k+1} g_{k+1}, w_{k+1} - x_{k+1}\rangle + \frac{\sigma_d}{2}(\beta_k + S_{k+1}\sigma_f)\|w_{k+1} - x_{k+1}\|^2 \\
&\overset{(4.4.12)}{\geq} \psi_k(w_k) + \lambda_{k+1} f(x_{k+1}) - \frac{1}{2}\frac{\lambda_{k+1}^2}{\sigma_d(\beta_k + S_{k+1}\sigma_f)}\|g_{k+1}\|_*^2.
\end{aligned}$$

Hence, the definition (4.4.13) 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). □

4.4.4 Validity of (R_k) for the modified method when $k > 0$

The following lemma completes our induction for the modified method. In a similar manner as Lemma 4.4.7, the items (i) and (ii) below correspond to the k-th iteration of the modified method in Methods I and II, respectively.

Lemma 4.4.8. *(i) Consider a non-smooth problem in the class $\mathcal{NSP}(g, \sigma_f)$ and let $\{(\varphi_k(x), \psi_k(x))\}_{k\geq -1}$ be a coupled sequence of auxiliary functions satisfying Property B associated with weight parameters $\{\lambda_k\}_{k\geq 0}$, scaling parameters $\{\beta_k\}_{k\geq -1}$, and test points $\{x_k\}_{k\geq 0}$. Suppose for $k \geq 0$ that the relation (R_k) is satisfied for some $\hat{x}_k \in Q$, $C_k \geq 0$. If the relation $x_{k+1} = \frac{S_k\hat{x}_k + \lambda_{k+1}z_k}{S_{k+1}}$ holds, then the relation (R_{k+1}) is satisfied with $\hat{x}_{k+1} := x_{k+1}$ and*

$$C_{k+1} := C_k + \frac{1}{2\sigma_d}\frac{\lambda_{k+1}^2 S_{k+1}}{\lambda_{k+1}^2\sigma_f + S_{k+1}(\beta_k + S_k\sigma_f)}\|g_{k+1}\|_*^2. \tag{4.4.14}$$

(ii) Consider a structured problem in the class $\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$ *and let* $\{(\varphi_k(x), \psi_k(x))\}_{k\geq -1}$ *be a coupled sequence of auxiliary functions satisfying Property B associated with weight parameters* $\{\lambda_k\}_{k\geq 0}$*, scaling parameters* $\{\beta_k\}_{k\geq -1}$*, and test points* $\{x_k\}_{k\geq 0}$*. Suppose for* $k \geq 0$ *that the relation* (R_k) *is satisfied for some* $\hat{x}_k \in Q$*,* $C_k \geq 0$*. 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*

$$\begin{aligned} C_{k+1} &:= C_k + S_{k+1}\left(\frac{L(x_{k+1})}{2} - \frac{\sigma_d}{2}\left(\bar{\sigma}_f + \frac{S_{k+1}(\beta_k + S_k\sigma_f)}{\lambda_{k+1}^2}\right)\right)\|\hat{x}_{k+1} - x_{k+1}\|^2 \\ &\quad + S_{k+1}\delta(x_{k+1}, \hat{x}_{k+1}). \end{aligned} \tag{4.4.15}$$

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 (4.4.10) and the relation (R_k), we have

$$\begin{aligned} \psi_{k+1}(w_{k+1}) + C_k &\geq \psi_k(w_k) + C_k + \lambda_{k+1}m_f(x_{k+1}; w_{k+1}) + (\beta_k + S_k\sigma_f)\xi(z_k, w_{k+1}) \\ &\geq S_k f(\hat{x}_k) + \lambda_{k+1}m_f(x_{k+1}; w_{k+1}) + (\beta_k + S_k\sigma_f)\xi(z_k, w_{k+1}) \\ &\geq S_k m_f(x_{k+1}; \hat{x}_k) + \lambda_{k+1}m_f(x_{k+1}; w_{k+1}) + (\beta_k + S_k\sigma_f)\xi(z_k, w_{k+1}) \\ &\geq S_{k+1}m_f(x_{k+1}; x'_{k+1}) + (\beta_k + S_k\sigma_f)\xi(z_k, w_{k+1}), \end{aligned} \tag{4.4.16}$$

where we used $f(x) \geq m_f(y; x), \forall x, y \in Q$ and the convexity of $m_f(x_{k+1}; \cdot)$ for the last two inequalities, respectively. 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

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

hold for any $\sigma \geq 0$, the inequality (4.4.16) implies that

$$\begin{aligned} \psi_{k+1}(w_{k+1}) + C_k &\geq S_{k+1}[m_f(x_{k+1}; x'_{k+1}) - \sigma\xi(x_{k+1}, x'_{k+1})] \\ &\quad + \frac{\sigma_d}{2}S_{k+1}\left(\sigma + \frac{S_{k+1}(\beta_k + S_k\sigma_f)}{\lambda_{k+1}^2}\right)\|x'_{k+1} - x_{k+1}\|^2. \end{aligned} \tag{4.4.17}$$

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 (4.4.17) with $\sigma := \bar{\sigma}_f$ and using the inequality (4.2.6) imply the relation (R_{k+1}) with C_{k+1} defined by (4.4.15).

To prove (i), on the other hand, letting $\sigma := \sigma_f$ and using $m_f(x_{k+1}; x'_{k+1}) - \sigma\xi(x_{k+1}, x'_{k+1}) = f(x_{k+1}) + \langle g_{k+1}, x'_{k+1} - x_{k+1}\rangle$ leads (4.4.17) to

$$\begin{aligned} \psi_{k+1}(w_{k+1}) + C_k &\geq S_{k+1}f(x_{k+1}) + \langle S_{k+1}g_{k+1}, x'_{k+1} - x_{k+1}\rangle \\ &\quad + \frac{\sigma_d}{2}S_{k+1}\left(\sigma_f + \frac{S_{k+1}(\beta_k + S_k\sigma_f)}{\lambda_{k+1}^2}\right)\|x'_{k+1} - x_{k+1}\|^2 \\ &\overset{(4.4.12)}{\geq} S_{k+1}f(x_{k+1}) - \frac{1}{2}\frac{S_{k+1}^2}{\sigma_d S_{k+1}\left(\sigma_f + \frac{S_{k+1}(\beta_k + S_k\sigma_f)}{\lambda_{k+1}^2}\right)}\|g_{k+1}\|_*^2 \\ &= S_{k+1}f(x_{k+1}) - \frac{1}{2\sigma_d}\frac{\lambda_{k+1}^2 S_{k+1}}{\lambda_{k+1}^2\sigma_f + S_{k+1}(\beta_k + S_k\sigma_f)}\|g_{k+1}\|_*^2. \end{aligned}$$

This means that the relation (R_{k+1}) is obtained with C_{k+1} defined by (4.4.14). □

4.4.5 Proof of Theorems 4.4.1 and 4.4.2

Let us complete the proof of Theorem 4.4.1.

Recall that Method I is equipped with a single sequence $\{\varphi_k(x)\}_{k\geq -1}$ of auxiliary functions satisfying Property A. Let $\{\psi_k(x)\}_{k\geq -1}$ be any sequence so that the coupled sequence $\{(\varphi_k(x), \psi_k(x))\}_{k\geq -1}$ satisfies Property B (*e.g.*, take $\psi_k := \varphi_k$). By the description of Method I, we can apply part (i) of each Lemmas 4.4.6, 4.4.7, and 4.4.8 to show that the relation (R_k) holds for every $k \geq 0$ with C_k defined by (4.4.2); for the classical method, the relation (P_k) can also be verified. The assertion follows from Lemma 4.4.4 and its analogue for the relation (P_k) (see Remark 4.4.5 (1)). □

Remark that this proof additionally introduced $\{\psi_k(x)\}_{k\geq -1}$ but it did not affect our conclusion because its dependence appears only in the relations (R_k) and (P_k).

Theorem 4.4.2 can be proved as an analogue replacing (P_k) with (Q_k) and the part (i) with (ii) in Lemmas 4.4.6, 4.4.7, 4.4.8. Note that, in this case, we do not need to introduce an additional $\{\psi_k(x)\}_{k\geq -1}$ since it is already in our assumption.

4.5 Optimal rate of convergence for non-smooth problems

From this section to Section 4.7, we discuss rates of convergence of the proposed methods for specific classes of problems. We firstly focus on Method I for the non-smooth problems in the class $\mathcal{NSP}(g, \sigma_f)$. Our aim is to find explicit choices of weight parameters $\{\lambda_k\}_{k\geq 0}$ and scaling parameters $\{\beta_k\}_{k\geq -1}$ which ensure an efficient convergence. Recall that the optimal complexity for the non-smooth problems is given by

$$O\left(\frac{M^2R^2}{\varepsilon}\right) \quad \text{and} \quad O\left(\frac{M^2}{\sigma_f \varepsilon}\right)$$

for the non strongly and strongly convex cases, respectively, where $M = \sup\{\|g\|_* \mid x \in Q,\ g \in \partial f(x)\}$ and $R = \frac{1}{\sigma_d}\sqrt{d(x^*)}$.

We divide our discussion into the non strongly convex and the strongly convex cases in Sections 4.5.1 and 4.5.2, respectively.

4.5.1 Optimal rate of convergence in the non strongly convex case

Let us consider Method I in the non strongly convex case $\sigma_f = 0$. In the next theorem, we analyze two choices of parameters $\{\lambda_k\}_{k\geq 0}$ and $\{\beta_k\}_{k\geq -1}$, called the *simple* and the *weighted* averages [52, eq. (2.21) and (2.22)], which ensure the optimal convergence rate. These choices utilize the sequence $\{\hat{\beta}_k\}_{k\geq -1}$ defined in (3.1.13) where we use the identity

$$\forall k \geq 0, \quad \hat{\beta}_k = \sum_{i=-1}^{k-1} \frac{1}{\hat{\beta}_i} \tag{4.5.1}$$

and the inequality [52, Lemma 3]

$$\forall k \geq 0, \quad \sqrt{2k+1} \leq \hat{\beta}_k \leq \frac{1}{1+\sqrt{3}} + \sqrt{2k+1}. \tag{4.5.2}$$

Theorem 4.5.1 (see also [52]). *Consider a non strongly convex and non-smooth problem in the class $\mathcal{NSP}(g,0)$. Let $\{\hat{\beta}_k\}_{k\geq -1}$ be the sequence defined by (3.1.13).*

(Simple Averages) *Let $\{(z_{k-1}, x_k, g_k, \hat{x}_k)\}_{k\geq 0}$ be generated by Method I with parameters $\lambda_k := 1$ and $\beta_k := \gamma\hat{\beta}_k$ for some $\gamma > 0$. Then we have*

$$\forall k \geq 0, \quad f(\hat{x}_k) - f(x^*) \leq \left(\gamma\ell_d(z_k; x^*) + \frac{M_k^2}{2\sigma_d\gamma}\right)\frac{0.5 + \sqrt{2k+1}}{k+1} \tag{4.5.3}$$

and

$$\forall k \geq -1, \quad z_k, x_{k+1}, \hat{x}_{k+1} \in \left\{x \in Q \;:\; \|x - x^*\|^2 \leq \frac{2d(x^*)}{\sigma_d} + \frac{M_k^2}{\sigma_d^2\gamma^2}\right\} \tag{4.5.4}$$

where $M_{-1} = 0$ and $M_k = \max\limits_{0\leq i\leq k} \|g_i\|_*$ for $k \geq 0$.

(Weighted Averages) *Let $\{(z_{k-1}, x_k, g_k, \hat{x}_k)\}_{k\geq 0}$ be generated by Method I with parameters $\lambda_k := \dfrac{1}{\|g_k\|_*}$ and $\beta_k := \dfrac{\hat{\beta}_k}{\rho\sqrt{\sigma_d}}$ for some $\rho > 0$. Then we have*

$$\forall k \geq 0, \quad f(\hat{x}_k) - f(x^*) \leq M_k\frac{1}{\sqrt{\sigma_d}}\left(\frac{\ell_d(z_k; x^*)}{\rho} + \frac{\rho}{2}\right)\frac{0.5 + \sqrt{2k+1}}{k+1} \tag{4.5.5}$$

and

$$\forall k \geq -1, \quad z_k, x_{k+1}, \hat{x}_{k+1} \in \left\{x \in Q \;:\; \|x - x^*\|^2 \leq \frac{2d(x^*) + \rho^2}{\sigma_d}\right\}. \tag{4.5.6}$$

Moreover, for both simple and weighted averages, the above $f(\hat{x}_k) - f(x^*)$'s can be replaced by its upper bound $\frac{1}{S_k}\sum_{i=0}^k \lambda_i f(x_i) - f(x^*)$ when we use the classical method in Method I. In this case, the left hand side of the inequality can be replaced by $\min\{f(\hat{x}_k) - f(x^*), \min_{0\leq i\leq k} f(x_i) - f(x^*)\}$.

Proof. Because $\sigma_f = 0$, both the classical and the modified methods yield the same estimate of Theorem 4.4.1:

$$\forall k \geq 0, \quad f(\hat{x}_k) - f(x^*) \leq \frac{\beta_k\ell_d(z_k; x^*) + \frac{1}{2\sigma_d}\sum_{i=0}^k \frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2}{S_k}.$$

Substituting the specified (λ_k, β_k) in the simple and the weighted averages shows (4.5.3) and (4.5.5), respectively, thanks to the properties (4.5.1) and (4.5.2) of $\hat{\beta}_k$.

Denote by B_k the ball on the right hand side of (4.5.4) for $k \geq -1$. Then $B_k \subset B_{k+1}$ for each $k \geq -1$. The inequality (4.5.3) implies that $\gamma\ell_d(z_k; x^*) + (2\sigma_d\gamma)^{-1}M_k^2 \geq 0$ for all $k \geq 0$, and using the strong convexity, $d(x^*) \geq \ell_d(z_k; x^*) + \frac{\sigma_d}{2}\|x^* - z_k\|^2$, we can obtain that $z_k \in B_k$ for each $k \geq 0$. We also have $z_{-1} \in B_{-1}$ since $z_{-1} = x_0 = \operatorname{argmin}_{x\in Q} d(x)$, $d(z_{-1}) = d(x_0) = 0$, and $d(x^*) \geq \ell_d(z_{-1}; x^*) + \frac{\sigma_d}{2}\|z_{-1} - x^*\|^2 \geq \frac{\sigma_d}{2}\|z_{-1} - x^*\|^2$. Finally, we conclude that $x_{k+1}, \hat{x}_{k+1} \in B_k$ for all $k \geq -1$ because they are convex combinations of $\{z_i\}_{i=-1}^k$. The proof of (4.5.6) is similar. □

Remark 4.5.2. The bounds (4.5.3) and (4.5.5) are slightly smaller than the ones in (3.3) and (3.5) in [52], respectively, because of $\ell_d(z_k; x^*) \le d(x^*) \le D$. However, essentially, Nesterov's original argument also arrives to the same bound when $d(x)$ is continuously differentiable on Q (note that [52] does not impose the differentiability on $d(x)$). In fact, in [52], Theorems 2 and 3 rely on the estimate (2.15) which is implied from (2.18). Notice in (2.18) that we have

$$-V_{\beta_{k+1}}(-s_{k+1}) = \min_{x \in Q}\{\langle s_{k+1}, x - x_0\rangle + \beta_{k+1} d(x)\} = \min_{x \in Q}\{\langle s_{k+1}, x - x_0\rangle + \beta_{k+1} \ell_d(x_{k+1}; x)\}$$

by the optimality of $x_{k+1} = \pi_{\beta_{k+1}}(-s_{k+1})$. Then adding $\sum_{i=0}^{k} \lambda_i [f(x_i) + \langle g_i, x_0 - x_i\rangle]$ and using $s_{k+1} = \sum_{i=0}^{k} \lambda_i g_i$ in (2.18), we obtain

$$\sum_{i=0}^{k} \lambda_i f(x_i) \le \min_{x \in Q} \left\{ \sum_{i=0}^{k} \lambda_i [f(x_i) + \langle g_i, x - x_i\rangle] + \beta_{k+1} \ell_d(x_{k+1}; x) \right\} + \frac{1}{2\sigma_d} \sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_i} \|g_i\|_*^2$$

which corresponds to the relation $(\hat{R}_k)$[2]. This yields the same bound as our analysis for the DA model. □

A consequence of Corollary 4.5.1 is that if $M := \sup\{\|g\|_* : g \in \partial f(x),\ x \in Q\}$ is finite, Method I generates a sequence $\{\hat{x}_k\}$ such that $f(\hat{x}_k) \to f(x^*)$ with a rate $O(1/\sqrt{k})$ in the number k of iterations. In particular, if we know an upper bound $R \ge \sqrt{\frac{1}{\sigma_d} d(x^*)}$ and for the single averages case additionally the M, the choices $\gamma := \frac{M}{\sqrt{2\sigma_d} R}$ and $\rho := \sqrt{2\sigma_d} R$ make the estimates (4.5.3) and (4.5.5) optimal, respectively, giving the optimal iteration complexity $O(M^2 R^2/\varepsilon^2)$ for the non-smooth problems. Also Method I with the parameters suggested in Corollary 4.5.1 produces bounded sequences $\{x_k\}$, $\{\hat{x}_k\}$, and $\{z_k\}$ (even if $M = +\infty$ for the weighted averages case).

These features are similar to the DAM. We can obtain the optimal convergence rate if we know an upper bound for $d(x^*)$, but without assuming the compactness of Q and fixing the required number of iterations.

It is important to note that, according to Corollary 4.5.1, the extended MDM (Method 4.3.5) ensures the rate $O(1/\sqrt{k})$ of convergence without fixing a priori the total number of iterations and knowing an upper bound of $d(x^*)$ required for the weight parameters (3.1.4) of the original MDM. Furthermore, this advantage holds *even if* the feasible set Q is unbounded. The existing averaging techniques [43, 44] of the MDM assume the compactness of Q to achieve the same complexity.

Method I with the DA model (4.3.8) recovers the convergence result for the Nesterov's DAM (3.1.9) and its variant (3.1.14). In particular, Theorem 4.4.1 and Corollary 4.5.1 provide a small improvement over the original result assuming the differentiability of $d(x)$ (see Remark 4.5.2).

4.5.2 Optimal rate of convergence in the strongly convex case

Let us consider the strongly convex case $\sigma_f > 0$ in Method I. In the general convergence estimate (4.4.1), we remark that

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

[2] Notice that x_{k+1} and β_{k+1} in [52] are called z_k and β_k here, respectively.

holds since $\lambda_i/S_i \leq 1$. In this case, theoretically, the classical method ensures not a worse convergence rate than the modified 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 Theorem 4.3.1).

Theorem 4.5.3. *Consider a non-smooth problem in the class $\mathcal{NSP}(g, \sigma_f)$. Let $\{(z_{k-1}, x_k, g_k, \hat{x}_k)\}_{k\geq 0}$ be generated by Method I 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 4.4.1 implies the estimate

$$f(\hat{x}_k) - f(x^*) + \sigma_f\xi(z_k, x^*) \leq \frac{C_k}{S_k} = \frac{4C_k}{(k+1)(k+2)} \tag{4.5.7}$$

with C_k defined by (4.4.2). The classical method also admits the same estimate replacing $f(\hat{x}_k) - f(x^*)$ by $\min_{0\leq i\leq k} f(x_i) - f(x^*)$.

For the classical method, we have

$$C_k = \frac{1}{2\sigma_d}\sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_{i-1} + S_i\sigma_f}\|g_i\|_*^2 \leq \frac{M_f^2}{2\sigma_d\sigma_f}\sum_{i=0}^{k}\frac{\lambda_i^2}{S_i}. \tag{4.5.8}$$

because of $\beta_k \equiv 0$ and $\|g_i\|_* \leq M_f$. Using the inequality

$$\sum_{i=0}^{k}\frac{\lambda_i^2}{S_i} = \sum_{i=0}^{k}\frac{i+1}{i+2} \leq \frac{(k+1)(k+2)}{k+4} \tag{4.5.9}$$

(see [22, Proposition 7.3]), we obtain the first assertion.

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 (4.5.7) to the second assertion. □

Let us see another choice of weight parameters ensuring the optimal iteration complexity. Due to the bound (4.5.8), the classical method with $\beta_k \equiv 0$ provides the estimate

$$f(\hat{x}_k) - f(x^*) + \sigma_f \xi(z_k, x^*) \leq \frac{C_k}{S_k} \leq \frac{M_f^2 \sum_{i=0}^k \frac{\lambda_i^2}{S_i}}{2\sigma_d \sigma_f S_k}.$$

For instance, the choice $\lambda_0 := 1$, $\lambda_{k+1} := \frac{1+\sqrt{1+4\lambda_k^2}}{2}$ $(k \geq 0)$ ensures that

$$f(\hat{x}_k) - f(x^*) + \sigma_f \xi(z_k, x^*) \leq \frac{2M_f^2}{\sigma_d \sigma_f (k+4)} \tag{4.5.10}$$

since we have $\lambda_k^2/S_k = 1$ and $S_k \geq (k+1)(k+4)/4$ by Lemma A.4 given at Appendix.

Let us consider the particular case presented in Example 4.3.3 (2), that is, the classical method in Method I with the auxiliary functions define by the EMD model (4.3.7). Theorem 4.5.3 above recovers the convergence for the Bach's averaging (3.1.8) because it coincides with our approximate solution $\hat{x}_k$. Moreover, the approximate solution $\hat{x}_k$ coincides with the Nedić-Lee's averaging $\widetilde{x}_k$ (3.1.5) with the choice $\lambda_0 := 1$, $\lambda_{k+1} := \frac{1+\sqrt{1+4\lambda_k^2}}{2}$ $(k \geq 0)$. Under the same assumption as [43] that f is σ_f-strongly convex on Q and $\xi(y, x) \leq \frac{1}{2}\|x - y\|^2$ holds for $x, y \in Q$ (recall Section 3.1.1), the estimate (4.5.10) provides the same rate of convergence as the estimate (3.1.7) (notice that we have $\sigma_f \in \sigma(f)$ by Corollary 4.2.3).

When we apply our result for Method I with the auxiliary functions generated by the DA model (4.3.8), we obtain a new convergence result on the extensions of the DAM and its variant to the strongly convex case. Note that we do not exploit a multistage procedure and do not require an upper bound of $d(x^*)$ in contrast to [33].

4.6 Convergence results for structured problems with constants L and δ

In this section, we focus on structured problems in the class $\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$ with the particular case $L(\cdot) \equiv L \geq 0$, $\delta(\cdot, \cdot) \equiv \delta \geq 0$. In this case, we additionally assume that $L \geq \bar{\sigma}_f \sigma_d$; notice that, in view of $m_f(y; x) \leq f(x)$ and $\xi(y, x) \geq \frac{\sigma_d}{2}\|x - y\|^2$ for $x, y \in Q$, the inequality (4.2.6) yields $0 \leq \frac{1}{2}(L - \bar{\sigma}_f \sigma_d)\|y - x\|^2 + \delta$ for every $x, y \in Q$ which forces $\mathrm{Diam}(Q) \leq \sqrt{2\delta/(\bar{\sigma}_f \sigma_d - L)}$ if $L < \bar{\sigma}_f \sigma_d$. Note that this case includes the smooth problems, the composite model, and the inexact oracle model (with constant $L(\cdot)$ and $\delta(\cdot)$) in Example 4.2.8.

4.6.1 Convergence rate of the classical method

Let us see the convergence result of PGMs yielded from the classical method of Method II. The obtained rate of convergence does not ensure the optimality for smooth problems in the class $\mathcal{F}_L^1(Q)$.

Theorem 4.6.1. *Consider a structured problem in the class $\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$. Assume additionally that $L(\cdot) = L \geq 0$, $\delta(\cdot, \cdot) = \delta \geq 0$, and $L \geq \bar{\sigma}_f \sigma_d$. Let $\{(z_{k-1}, w_{k-1}, x_k, \hat{x}_k)\}_{k \geq 0}$ be generated by the classical method of Method II with*

$$\beta_k \equiv \frac{L - \bar{\sigma}_f \sigma_d}{\sigma_d}, \quad \lambda_0 = 1, \quad \lambda_{k+1} = \frac{\beta_k + S_k \sigma_f}{\beta_k}. \tag{4.6.1}$$

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

$$f(\hat{x}_k)-f(x^*)+\sigma_f\xi(z_k,x^*) \leq \frac{L-\bar{\sigma}_f\sigma_d}{\sigma_d}\ell_d(z_k;x^*)\min\left\{\left(1-\frac{\sigma_f\sigma_d}{L-\bar{\sigma}_f\sigma_d+\sigma_f\sigma_d}\right)^k, \frac{1}{k+1}\right\}+\delta. \tag{4.6.2}$$

Furthermore, the left hand side of (4.6.2) 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. By Theorem 4.4.2, the classical method satisfies the estimate (4.4.3) with

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

Remark that the definitions of λ_k and β_k eliminate the first summation of C_k so that $C_k = \sum_{i=0}^k\lambda_i\delta = S_k\delta$ (since $\frac{\beta_{i-1}+S_{i-1}\sigma_f}{\lambda_i} = \beta_{i-1} = \frac{L-\bar{\sigma}_f\sigma_d}{\sigma_d}$). Moreover, we have for all $k\geq 0$ that

$$\begin{aligned} S_k = 1+\left(1+\frac{\sigma_f}{\beta_{-1}}\right)S_{k-1} \quad &\Rightarrow \quad S_k \geq \max\left(1+S_{k-1}, \left(1+\frac{\sigma_f}{\beta_{-1}}\right)S_{k-1}\right) \\ &\Rightarrow \quad S_k \geq \max\left\{k+1, \left(1-\frac{\sigma_f}{\beta_{-1}+\sigma_f}\right)^{-k}\right\}. \end{aligned}$$

Therefore, the assertion follows from Theorem 4.4.2. □

In the case $L > \bar{\sigma}_f\sigma_d$, the right hand side of the estimate (4.6.2) converges to δ. When $L = \bar{\sigma}_f\sigma_d$ (*i.e.*, $\beta_k \equiv 0$) and $\delta = 0$, the estimate (4.6.2) says that $\hat{x}_0$ is an optimal solution. This is an obvious assertion because $f(x) = m_f(y,x)$, $x,y\in Q$ follows from (4.2.6) and thus the conditions (B2) and (B3) imply $\hat{x}_0 \in \operatorname{Argmin}_{x\in Q}\psi_0(x) = \operatorname{Argmin}_{x\in Q}\lambda_0 m_f(x_0,x) = \operatorname{Argmin}_{x\in Q} f(x)$.

Let us see our result in particular examples.

Example 4.6.2 (Composite structure)**.** Consider the composite problem $\min_{x\in Q}[f(x)\equiv f_0(x)+\Psi(x)]$ with $f_0\in\mathcal{F}_L^1(Q)$ in the Euclidean setting $d(x)=\frac{1}{2}\|x-x_0\|_2^2$. This corresponds to the case $\sigma_d=1$, $\bar{\sigma}_f=\sigma_{f_0}$, $\sigma_f=\sigma_{f_0}+\sigma_\Psi$, and $\delta=0$ as Example 4.2.8 (iii).

Consider the classical method of Method II with the choice of parameters in Theorem 4.6.1. In this case, the EMD model (4.3.7) and the hybrid model (4.3.9) yields the primal gradient method (3.2.1) and the dual one (3.2.2), respectively (Example 4.3.4 (2b)). In the non strongly convex case $\sigma_f=0$, Theorem 4.6.1 gives the estimate

$$\min_{0\leq i\leq k} f(w_i)-f(x^*) \leq \frac{L\ell_d(z_k;x^*)}{k+1} \leq \frac{L\|x_0-x^*\|_2^2}{2(k+1)}$$

recovering the known estimate (3.2.3) for them. This estimate also holds for the DA model (4.3.8) reducing the number of subproblems compared with the hybrid model (cf. Example 4.3.4 (2b)).

In the strongly convex case $\sigma_f>0$, Theorem 4.6.1 yields the linear convergence

$$\begin{aligned} \min\left\{f(\hat{x}_k)-f(x^*), \min_{0\leq i\leq k} f(w_i)-f(x^*)\right\}+\sigma_f\xi(z_k,x^*) \quad &\leq \quad (L-\sigma_{f_0})\ell_d(z_k;x^*)\left(1-\frac{\sigma_f}{L+\sigma_\Psi}\right)^k \\ &\leq \quad \frac{L-\sigma_{f_0}}{2}\left(1-\frac{\sigma_f}{L+\sigma_\Psi}\right)^k\|x_0-x^*\|_2^2. \end{aligned} \tag{4.6.3}$$

In particular the primal gradient method ensures this estimate without knowing $\bar{\sigma}_f$ and σ_f. For the primal gradient method, we have another estimate (3.2.4) by Nesterov. Here we show that our linear convergence factor $1 - \frac{\sigma_f}{L+\sigma_\Psi}$ is better than the one in (3.2.4).[3] Now since we have $L \geq \sigma_{f_0}$ and $\sigma_f = \sigma_{f_0} + \sigma_\Psi$, we always have

$$1 - \frac{\sigma_f}{L+\sigma_\Psi} = \frac{L - \sigma_{f_0}}{L+\sigma_\Psi} \leq \frac{L-\sigma_{f_0}}{\sigma_{f_0}+\sigma_\Psi} \leq \frac{L}{\sigma_f}.$$

Therefore, it suffices to consider the case $L/\sigma_f \geq 1/2$, that is, $2L \geq \sigma_f$. In this case, we have

$$4L \geq L + 2L \geq L + \sigma_f \geq L + \sigma_\Psi$$

and thus $1 - \frac{\sigma_f}{L+\sigma_\Psi} \leq 1 - \frac{\sigma_f}{4L}$ holds. This shows the claim. □

Example 4.6.3 (Inexact oracle model)**.** Suppose that the objective function f is equipped with a (δ, L, μ)-oracle in the Euclidean setting $d(x) = \frac{1}{2}\|x - x_0\|_2^2$. In this case, we have $\sigma_f = \bar{\sigma}_f = \mu$ and $\delta(y, x) \equiv \delta$ as Example 4.2.8 (v). Theorem 4.6.1 states that the classical method of Method II setting

$$\beta_k \equiv L - \mu, \quad \lambda_0 = 1, \quad \lambda_{k+1} = \frac{L - \mu + S_k\mu}{L-\mu}$$

yields the estimate

$$\min_{0\leq i\leq k+1} f(w_i) - f(x^*) + \sigma_f\xi(z_k, x^*) \leq (L-\mu)\ell_d(z_k; x^*)\min\left\{\left(1 - \frac{\mu}{L}\right)^k, \frac{1}{k+1}\right\} + \delta.$$

Recall from Example 4.3.4 (3) that the EMD model (4.3.7) and the hybrid model (4.3.9) yields the primal gradient method (3.2.5) and the dual gradient method (3.2.6), respectively. Also recall that our choices of $\{\lambda_k\}$ and $\{\beta_k\}$ were not equivalent to (3.2.8) analyzed in [17] for the dual gradient method. Comparing the above estimate with (3.2.7), our choice of parameters ensure smaller upper bound in view of $L - \mu \leq L$ and $\ell_d(z_k; x^*) \leq \frac{1}{2}\|x_0 - x^*\|_2^2$. Again the primal gradient method ensure our result without knowing μ. Using the DA model, we can reduce the number of subproblems of the dual gradient method preserving the convergence property as remarked for the composite structure (Example 4.3.4 (2b)). □

4.6.2 Optimal rate of convergence for the modified method

The modified method of Method II for the structured problem in the particular case $L(\cdot) = L \geq 0$, $\delta(\cdot,\cdot) = \delta \geq 0$ can be analyzed as follows. Differently from the classical method, it achieves the optimal convergence rate for the class $\mathcal{F}_L^1(Q)$. The result below further implies efficient rates for the CGMs, too.

Theorem 4.6.4. *Consider a structured problem in the class $\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$. Assume in addition that $L(\cdot) = L \geq 0$, $\delta(\cdot,\cdot) = \delta \geq 0$, and $L > \bar{\sigma}_f\sigma_d$.*
(i) Let $\{(z_{k-1}, w_{k-1}, x_k, \hat{x}_k)\}_{k\geq 0}$ be generated by the modified method of Method II with

$$\beta_k \equiv \frac{L - \bar{\sigma}_f\sigma_d}{\sigma_d}, \quad \lambda_0 = 1, \quad \lambda_{k+1}^2 = \left(1 + S_k\frac{\sigma_f\sigma_d}{L - \bar{\sigma}_f\sigma_d}\right)(\lambda_{k+1} + S_k) \ (k \geq 0) \tag{4.6.4}$$

[3] Remark that this improvement does not imply that our estimate (4.6.3) is better than the Nesterov's one (3.2.4); their comparison will depend on the parameters.

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

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

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

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

Proof. By Theorem 4.4.2, we have the estimate (4.4.3) with

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

(i) The recurrence of $\{\lambda_k\}$ can be rewritten as

$$\begin{aligned} \lambda_{k+1}^2 = \beta_k^{-1}(S_k\sigma_f + \beta_k)S_{k+1} &\iff L - \bar{\sigma}_f\sigma_d = \frac{\sigma_d}{\lambda_{k+1}^2}(S_k\sigma_f + \beta_k)S_{k+1} \\ &\iff L = \sigma_d \left(\bar{\sigma}_f + \frac{S_{k+1}(\beta_k + S_k\sigma_f)}{\lambda_{k+1}^2} \right). \end{aligned}$$

This eliminates the first summation in C_k above so that we have $C_k = \sum_{i=0}^{k} S_i\delta$. Therefore, the estimate (4.4.3) implies

$$f(\hat{x}_k) - f(x^*) + \sigma_f \xi(z_k, x^*) \leq \frac{L - \bar{\sigma}_f\sigma_d}{\sigma_d} \ell_d(z_k; x^*) \cdot \frac{1}{S_k} + \frac{\sum_{i=0}^{k} S_i\delta}{S_k}.$$

It remains to estimate $1/S_k$ and $\sum_{i=0}^{k} S_i/S_k$ which is left to Lemmas A.1 to A.4, given at Appendix.

(ii) Letting $\beta_k = 0$, and $\sigma_f = 0$ in Theorem 4.4.2 with C_k described above and using the inequality (4.5.9) establish

$$f(\hat{x}_k) - f(x^*) \leq \frac{C_k}{S_k} = \frac{L\sum_{i=0}^{k} \frac{\lambda_i^2}{S_i} \|w_i - z_{i-1}\|^2}{2S_k} + \frac{\sum_{i=0}^{k} S_i\delta}{S_k}. \tag{4.6.5}$$

Therefore, the choice $\lambda_k = (k+1)/2$ yields the estimate

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

□

Let us discuss the above result for particular classes of structured problems.

Example 4.6.5 (PGMs for composite/smooth problems)**.** Consider the composite problem $\min_{x\in Q}[f(x) \equiv f_0(x) + \Psi(x)]$ in the Euclidean setting $d(x) = \frac{1}{2}\|x - x_0\|_2^2$. We have $\sigma_d = 1$, $\bar{\sigma}_f = \sigma_{f_0}$, $\sigma_f = \sigma_{f_0} + \sigma_\Psi$, and $\delta = 0$ (Example 4.2.8 (iii)). By Theorem 4.6.4 (i), the modified method of Method II with the parameters (4.6.4) ensures the estimate

$$f(\hat{x}_k) - f(x^*) + \frac{\sigma_f}{2}\|z_k - x^*\|_2^2 \le \frac{(L-\sigma_{f_0})\|x_0 - x^*\|_2^2}{2} \min\left\{\frac{4}{(k+2)^2}, \left(1 + \frac{1}{2}\sqrt{\frac{\sigma_f}{L-\sigma_{f_0}}}\right)^{-2k}\right\}.$$

This estimate resembles the one (3.2.17) of the Nesterov's accelerated method. We remark that Method II does not include the Nesterov's accelerated method [53] as a particular instance in general.

When $\Psi(x) \equiv 0$ (then $\sigma_\Psi = 0$ and $\sigma_f = \sigma_{f_0}$), our method attains the optimal iteration complexity for the smooth problems. In fact, since $\log(1+x) \ge \frac{1}{2}x$ holds for $x \in [0,1]$, we obtain

$$f(\hat{x}_k) - f(x^*) \le \frac{(L-\sigma_f)\|x_0 - x^*\|_2^2}{2} \min\left\{\frac{4}{(k+2)^2}, \exp\left(-\frac{k}{2}\sqrt{\frac{\sigma_f}{L}}\right)\right\}.$$

Therefore, $\hat{x}_k$ is an ε-solution whenever

$$k \ge \min\left\{\sqrt{\frac{2(L-\sigma_f)\|x_0 - x^*\|_2^2}{\varepsilon}} - 2, \quad 2\sqrt{\frac{L}{\sigma_f}}\log\left(\frac{(L-\sigma_f)\|x_0 - x^*\|_2^2}{2\varepsilon}\right)\right\}.$$

Recall from Example 4.3.4 (1) that particular instances of Method II include the Nesterov's modified method (3.2.9) and Tseng's APG methods (3.2.10), (3.2.11). Our result therefore gives extensions of them to the strongly convex case ensuring the optimal iteration complexity. □

Example 4.6.6 (PGMs for inexact oracle model)**.** Suppose that the objective function $f(x)$ is equipped with a (δ, L, μ)-oracle (2.4.14) in the Euclidean setting $d(x) = \frac{1}{2}\|x - x_0\|_2^2$, $\sigma_d = 1$. Theorem 4.6.4 (i) with the correspondence $\sigma_f = \bar{\sigma}_f = \mu$ yields the estimate

$$\begin{aligned} f(\hat{x}_k) - f(x^*) + \frac{\sigma_f}{2}\|z_k - x^*\|_2^2 \quad \le \quad & (L-\mu)\ell_d(z_k; x^*) \min\left\{\frac{4}{(k+2)^2}, \left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L-\mu}}\right)^{-2k}\right\} \\ & + \min\left\{\frac{1}{3}k + \frac{1}{6}\log(k+2) + 1, \ 1 + \sqrt{\frac{L-\mu}{\mu}}\right\}\delta \end{aligned}$$

for all $k \ge 0$, which is slightly better than the estimate (3.2.18) for the fast gradient method [17, Algorithm 3] in view of $(L-\mu)\ell_d(z_k; x^*) \le Ld(x^*)$ and $\frac{\mu}{L} \le \frac{\mu}{L-\mu}$. We remark that the fast gradient method does not arise as a particular instance of Method II in general. □

Example 4.6.7 (Convergence results for CGMs)**.** Let us compare Theorem 4.6.4 (ii) with known estimates for existing CGMs reviewed in Section 3.2.3. Note that Theorem 4.6.4 (ii) yields the estimate

$$f(\hat{x}_k) - f(x^*) \le \frac{2L\mathrm{Diam}(Q)^2}{k+4} + \frac{k+3}{3}\delta, \quad \forall k \ge 0.$$

For the inexact oracle model, it is similar to the estimate of the classical CGM (3.2.19) shown by Freund and Grigas [22, Section 5.2.1]. In fact, the general bound (4.6.5) of the CGM resembles the bound (53) in [22].

Our result for the composite problems as well as the smooth ones yields the the same upper bound $\frac{2L\mathrm{Diam}(Q)^2}{k+4}$ as the one (3.2.20) of the classical CGM. For the smooth problems, Method II with the choice $\beta_k \equiv 0$ and $\lambda_k = \frac{k+2}{2}$ includes the Lan's CGMs (3.2.21) and (3.2.22) (cf. Example 4.3.4 (1)) recovering the estimate (3.2.23). □

4.7 Optimal/nearly optimal rates of convergence for weakly smooth problems

In this section, we consider structured problems in the class $\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$ in the particular case

$$\delta(y,x) = \frac{M(y)}{\rho}\|y-x\|^\rho \text{ where } M(\cdot) \geq 0,\ \rho \in [1,2),$$

which include convex problems involving weakly smooth functions in the class $\mathcal{F}_M^{\rho-1}(Q)$ (Example 4.2.8 (ii)) or involving a mixed smoothness (Example 4.2.8 (iv)). We excluded the case $\rho = 2$ since it reduces the situation $\delta(y,x) = 0$ which has been already discussed in the previous section.

In particular, we aim to establish the iteration complexities (2.4.7) for the PGMs and (2.4.8) for the CGMs via the modified method of Method II for weakly smooth problems. Complexity results for PGMs in the non strongly and the strongly convex cases will be given in Sections 4.7.1 and 4.7.2, respectively. In Section 4.7.3, we finally prove optimal/nearly optimal convergence results for CGMs.

We at first prepare the following lemma for the analysis of PGMs.

Lemma 4.7.1. *Consider a structured problem in the class $\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$. Assume that $\delta(y,x) = \frac{M(y)}{\rho}\|y-x\|^\rho$, $\rho \in [1,2)$, $M(\cdot) \geq 0$. Let $\{(z_{k-1}, w_{k-1}, x_k, \hat{x}_k)\}_{k\geq 0}$ be generated by the modified method of Method II with weight parameters $\{\lambda_k\}_{k\geq 0}$ and scaling parameters $\{\beta_k\}_{k\geq -1}$. 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*

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

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 4.4.2 concludes that

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

which proves the assertion. □

4.7.1 Optimal rate of convergence in the non strongly convex case

Let us deduce a convergence result of PGMs given by the modified method of Method II 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 [25, Proposition 8] and [12, Corollary 1].

Theorem 4.7.2. *Consider a structured problem in the class $\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$. Assume additionally that $L(\cdot) = L \geq 0$, $\sigma_f = \bar{\sigma}_f = 0$, and $\delta(y, x) = \frac{M(y)}{\rho}\|y - x\|^\rho$ for $\rho \in [1, 2)$, $M(\cdot) \geq 0$. Let $\{(z_{k-1}, w_{k-1}, x_k, \hat{x}_k)\}_{k\geq 0}$ be generated by the modified method of Method II 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

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

Proof. We apply Lemma 4.7.1 to prove the assertion. Note that

$$\frac{\beta_k}{S_k} = \frac{4L}{\sigma_d(k+1)(k+2)} + \frac{4\gamma(k+3)^{\frac{3}{2}(2-\rho)}}{\sigma_d(k+1)(k+2)} \tag{4.7.1}$$

and α_k in Lemma 4.7.1 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

$$\begin{aligned}\frac{1}{S_k}\sum_{i=0}^{k}\frac{S_i}{(-\alpha_i)^{\frac{\rho}{2-\rho}}} &\leq \frac{1}{S_k}\sum_{i=0}^{k}\frac{(i+1)^{\frac{\rho}{2-\rho}+1}}{4\gamma^{\frac{\rho}{2-\rho}}(i+2)^{\frac{3}{2}\rho+\frac{\rho}{2-\rho}-1}} \leq \frac{1}{4\gamma^{\frac{\rho}{2-\rho}}S_k}\sum_{i=0}^{k}(i+2)^{2-\frac{3}{2}\rho} \\ &\leq \frac{1}{4\gamma^{\frac{\rho}{2-\rho}}S_k}\frac{2}{3(2-\rho)}(k+3)^{3-\frac{3}{2}\rho} = \frac{2(k+3)^{\frac{3}{2}(2-\rho)}}{3(2-\rho)\gamma^{\frac{\rho}{2-\rho}}(k+1)(k+2)},\end{aligned} \tag{4.7.2}$$

where the second and the third inequalities are due to $i+1 \leq i+2$ and the fact $\sum_{i=0}^{k}(i+2)^q \leq \frac{1}{1+q}(k+3)^{1+q}$, $\forall q > -1$, respectively. Consequently, the theorem follows by applying Lemma 4.7.1 with the inequalities (4.7.1) and (4.7.2). □

Notice that we need the parameter ρ to define β_k but not the $M(\cdot)$. Now let us observe an efficient choice for γ. Suppose that $M(\cdot) \equiv M$. Using $\ell_d(z_k; x^*) \leq d(x^*)$ 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

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

makes the estimate of Theorem 4.7.2 as follows:

$$\begin{aligned} f(\hat{x}_k) - f(x^*) &\leq \frac{4Ld(x^*)}{\sigma_d(k+1)(k+2)} \\ &\quad + \frac{2}{2-\rho}\left(\frac{\rho}{2-\rho}\right)^{-\frac{\rho}{2}}\left(\frac{4d(x^*)}{\sigma_d}\right)^{\frac{\rho}{2}}\left(\frac{M^{\frac{2}{2-\rho}}}{3\rho}\right)^{\frac{2-\rho}{2}}\frac{(k+3)^{\frac{3}{2}(2-\rho)}}{(k+1)(k+2)} \\ &= \frac{4Ld(x^*)}{\sigma_d(k+1)(k+2)} + \frac{2(2\sqrt{3})^\rho}{3\rho(2-\rho)^{\frac{2-\rho}{2}}}M\left(\frac{d(x^*)}{\sigma_d}\right)^{\frac{\rho}{2}}\frac{(k+3)^{\frac{3}{2}(2-\rho)}}{(k+1)(k+2)}. \end{aligned} \tag{4.7.3}$$

The case $M = 0$ matches the optimal convergence rate for the smooth problems.

Let us see that the case $L = 0$ in (4.7.3) attains the optimal iteration complexity (2.4.7) for the weakly smooth problems ($f \in \mathcal{F}_M^{\rho-1}(Q)$) in the non strongly convex case. Relaxing $k + 3 \leq 3(k + 1)$, the estimate (4.7.3) with $L = 0$ yields

$$f(\hat{x}_k) - f(x^*) \leq \frac{2^{1+\rho}\cdot 3^{2-\rho}}{\rho(2-\rho)^{\frac{2-\rho}{2}}}M\left(\frac{d(x^*)}{\sigma_d}\right)^{\frac{\rho}{2}}(k+1)^{-\frac{3\rho-2}{2}}.$$

Therefore, we obtain $f(\hat{x}_k) - f(x^*) \leq \varepsilon$ whenever

$$k + 1 \geq c(\rho)\left(\frac{d(x^*)}{\sigma_d}\right)^{\frac{\rho}{3\rho-2}}\left(\frac{M}{\varepsilon}\right)^{\frac{2}{3\rho-2}}$$

where

$$c(\rho) = \left(\frac{2^{1+\rho}\cdot 3^{2-\rho}}{\rho(2-\rho)^{\frac{2-\rho}{2}}}\right)^{\frac{2}{3\rho-2}}.$$

It matches the optimal iteration complexity (2.4.7). One can verify that $c(\rho)$ is decreasing on $[1, 2)$, $c(1) = 144$, and $\lim_{\rho\to 2} c(\rho) = 2$. In fact, $c'(\rho)$ is given by

$$c'(\rho) = -\frac{6c(\rho)}{(3\rho-2)^2}\log a(\rho) - \frac{c(\rho)}{(3\rho-2)\rho}\left(2 - \rho + 2\rho\log\frac{3}{2} - \rho\log(2-\rho)\right)$$

where $a(\rho) = \frac{2^{\rho+1}3^{2-\rho}}{\rho(2-\rho)^{\frac{2-\rho}{2}}}$. It is easy to see $2 - \rho + 2\rho\log\frac{3}{2} > 0$, $\log(2-\rho) \leq 0$, and $\log a(\rho) \geq 0$ for $\rho \in [1, 2)$ showing that $c'(\rho) < 0$.

This result is more of theoretical interest only because the attainment of the optimal iteration complexity above requires to know M and ρ to determine the parameters $\{(\beta_k, \lambda_k)\}$, in contrast to the Nesterov's universal gradient method [54]. One of their differences is that our result ensures the convergence $f(\hat{x}_k) \to f(x^*)$ with the optimal rate while the Nesterov's method [54] ensures only $\limsup_{k\to\infty}(f(\hat{x}_k) - f(x^*)) \leq \frac{\varepsilon}{2}$ for a tolerance parameter $\varepsilon > 0$ *fixed* in the method.

4.7.2 Optimal rate of convergence in the strongly convex case

Now we show a convergence result of PGMs in the strongly convex case $\sigma_f > 0$. We use the following notation for our claim

$$P(k) := \begin{cases} \left(p+2-\frac{2\rho}{2-\rho}\right)^{-1}(k+1)^{-\frac{3\rho-2}{2-\rho}} & : p+1 > \frac{3\rho-2}{2-\rho}, \\ \dfrac{1+\log k}{(k+1)^{p+1}} & : p+1 = \frac{3\rho-2}{2-\rho}, \\ \dfrac{1-\left(p+2-\frac{2\rho}{2-\rho}\right)^{-1}}{(k+1)^{p+1}} & : p+1 < \frac{3\rho-2}{2-\rho}. \end{cases} \tag{4.7.4}$$

Theorem 4.7.3. *Consider a structured problem in the class $\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$. Assume additionally that $L(\cdot) = L \geq 0$, $\sigma_f > 0$, and $\delta(y,x) = \frac{M(y)}{\rho}\|y-x\|^\rho$ for $\rho \in [1,2)$, $M(\cdot) \geq 0$. Let $\{(z_{k-1}, w_{k-1}, x_k, \hat{x}_k)\}_{k\geq 0}$ be generated by the modified method of Method II 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 $p \geq 1$ and $\beta \geq 0$ with $\sigma_d\bar{\sigma}_f + pL + (p+1)\sigma_d\beta > 0$. Then, for every $k \geq 0$, we have

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

where $P(k)$ is defined by (4.7.4).

Proof. Let us apply Lemma 4.7.1. Firstly, note that β_k is non-decreasing and $\frac{1}{(p+1)^2}(k+1)^{p+1} \leq S_k \leq \frac{1}{(p+1)^2}(k+2)^{p+1}$. We then have

$$\frac{\beta_k}{S_k} \leq \left(\frac{L}{\sigma_d}+\beta\right)(p+1)^2\frac{(k+2)^{p-1}}{(k+1)^{p+1}} = O(k^{-2}). \tag{4.7.5}$$

Secondly, 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

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

Therefore, 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

$$\frac{1}{S_k}\sum_{i=0}^{k}\frac{S_i}{(-\alpha_i)^{\frac{\rho}{2-\rho}}} \leq \frac{p+1}{(\sigma_d\bar{\sigma}_f+pL+(p+1)\sigma_d\beta)^{\frac{\rho}{2-\rho}}}\frac{1}{(k+1)^{p+1}} + 3^{p+1}\left(\frac{2^{p-1}(p+1)^2}{\sigma_d\sigma_f}\right)^{\frac{\rho}{2-\rho}}P(k), \tag{4.7.6}$$

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

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

Consequently, the assertion follows from Lemma 4.7.1 with the inequalities (4.7.5) and (4.7.6). □

Notice that we do not need ρ and $M(\cdot)$ in the definition of 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 $L=0,\ \sigma_f = \bar{\sigma}_f > 0,\ M(\cdot) = M,\ \beta = 0,\ p+1 > \frac{3\rho-2}{2-\rho}$:

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

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 complexity (2.4.7) for the strongly convex case. In contrast to the optimal method in [45], we do not employ restarting the method and do not require constants M and $R \ge d(x^*)$ in advance[4] to ensure the optimality.

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

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

This result resembles the ones [25, Proposition 9] and [12, Corollary 2] in the deterministic case.

4.7.3 Optimal/nearly optimal rate of convergence 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 4.7.1.

Theorem 4.7.4. *Consider a structured problem in the class $\mathcal{SP}(m_f, \sigma_f, \bar{\sigma}_f, L, \delta)$. Assume additionally that $L(\cdot) = L \ge 0$, $\sigma_f = \bar{\sigma}_f = 0$, and $\delta(y,x) = \frac{M}{\rho}\|y-x\|^\rho$ for $\rho \in [1,2)$, $M \ge 0$. Then, the modified method of Method II for the problem with $\lambda_k = (k+1)/2$ and $\beta_k \equiv 0$ generates a sequence $\{\hat{x}_k\}_{k\ge 0} \subset Q$ satisfying*

$$f(\hat{x}_k) - f(x^*) \le \frac{2L\mathrm{Diam}(Q)^2}{k+4} + \frac{2^{\rho+1} M \mathrm{Diam}(Q)^\rho}{\rho(3-\rho)(k+2)^{\rho-1}} \tag{4.7.7}$$

for every $k \ge 0$.

[4] As is indicated in [45], an obvious upper bound of $d(x^*)$ can be obtained if $\nabla f(x^*) = 0$ and we know M for the weakly smooth problems ($f \in \mathcal{F}_M^{\rho-1}(Q)$) in the Euclidean setting $d(x) = \frac{1}{2}\|x - x_0\|_2^2$: The inequality $d(x^*) \le \frac{1}{2}(\frac{2M}{\rho\sigma_f})^{2/(2-\rho)}$ follows since we have $\frac{\sigma_f}{2}\|x_0 - x^*\|_2^2 \le f(x_0) - f(x^*) \le \frac{M}{\rho}\|x_0 - x^*\|_2^\rho$ (recall the strong convexity and (4.2.6)).

Proof. Theorem 4.4.2 yields that $f(\hat{x}_k) - f(x^*) \le C_k/S_k$ with $S_k = (k+1)(k+2)/4$ and

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

(see Remark 4.4.3). Using the inequality (4.5.9) and

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

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

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

The estimate (4.7.7) now follows from $\frac{1}{k+1} \le \frac{2}{k+2}$ for $k \ge 0$. □

Remark 4.7.5. We can also obtain the estimate (4.7.7) for the classical CGM arranged for the class $\mathcal{SP}(m_f, 0, 0, L, \delta)$ with affine $m_f(y;\cdot)$'s (refer Remark 4.4.3). In fact, taking the initial point $x_0 \in \mathrm{Argmin}_{x\in Q} m_f(x_{-1}; x)$ for an arbitrary $x_{-1} \in Q$, the (arranged) classical CGM admits the estimate (4.4.6). Then, using $f(x_0) - f(x^*) \le f(x_0) - m_f(x_{-1}; x_0) \overset{(4.2.6)}{\le} \frac{L}{2}\mathrm{Diam}(Q)^2 + \frac{M}{\rho}\mathrm{Diam}(Q)^\rho$ and $\delta(x_{k-1}, x_k) = \frac{M}{\rho}\|x_k - x_{k-1}\|^\rho \overset{(3.2.19)}{=} \frac{M}{\rho}\frac{\lambda_k^\rho}{S_k^\rho}\|x_{k-1} - z_{k-1}\|^\rho \le \frac{M}{\rho}\frac{\lambda_k^\rho}{S_k^\rho}\mathrm{Diam}(Q)^\rho$ for $k \ge 1$, we arrive at

$$f(\hat{x}_k) - f(x^*) \le \frac{1}{S_k} \left(\frac{L}{2}\mathrm{Diam}(Q)^2 \sum_{i=0}^{k} \frac{\lambda_i^2}{S_i} + \frac{M}{\rho}\mathrm{Diam}(Q)^\rho \sum_{i=0}^{k} \frac{\lambda_i^\rho}{S_i^{\rho-1}} \right).$$

Hence, as the same way as the proof of Theorem 4.7.4, the estimate (4.7.7) holds when $\lambda_k = \frac{k+1}{2}$ (namely, $\tau_k = \frac{2}{k+3}$). This result in the case $L = 0$ is very similar to a known result for the classical CGM (see [14, Proposition 1.1] and [55]). □

Notice that the choice $\lambda_k = (k+1)/2$ and $\beta_k \equiv 0$ are independent of L, M, and ρ. Theorem 4.7.4 applied to the weakly smooth problem $\min_{x\in Q} f(x)$, $f \in \mathcal{F}_M^\nu(Q)$, $\nu \in (0, 1]$ ensures the convergence

$$f(\hat{x}_k) - f(x^*) \le \frac{2^{2+\nu} M \mathrm{Diam}(Q)^{1+\nu}}{(1+\nu)(2-\nu)(k+2)^\nu}, \quad \forall k \ge 0.$$

Therefore, we obtain an ε-solution whenever

$$k + 2 \ge \left(\frac{2^{2+\nu} M \mathrm{Diam}(Q)^{1+\nu}}{(1+\nu)(2-\nu)\varepsilon} \right)^{\frac{1}{\nu}}$$

which matches the known iteration complexity (2.4.8) of the classical CGM for weakly smooth problems. Recall that this is optimal for the smooth problems ($\nu = 1$) in view of the linear optimization oracle [35] and nearly optimal for the weakly smooth problems (cf. Section 2.4).

When we employ the EMD model (4.3.7) or the DA model (4.3.8) in Theorem 4.7.4, the obtained CGMs match particular cases of Lan's CGMs as mentioned in Example 4.3.4 (1). Since the convergence rates for Lan's CGMs was analyzed only for smooth problems in [35], our result provides a generalization of them for the weakly smooth problems.

Chapter 5

Conclusion and Further Remarks

In this thesis, we proposed Methods I and II based on Properties A and B as a unifying framework of subgradient-based methods for 'structured' convex optimization problems, namely, for the non-smooth and the structured problems introduced in Section 4.2.2. We demonstrated a general scheme of construction of PGMs and CGMs where some particular instances yield existing methods. Our analysis was performed in a unified way and derived optimal complexity results of the PGMs and nearly optimal one of the CGMs for various classes of convex optimization problems.

Our unification comes essentially from Property A (and B). In the next section, we observe a connection between Property A and Nesterov's estimate sequence technique [48, 49]. This connection, perhaps, enrich the understanding of Property A.

We finally discuss further considerable research directions based on our unifying framework in Section 5.2.

5.1 Relation to Nesterov's estimate sequence

Nesterov's *estimate sequence* approach [48, 49] is a powerful methodology to construct efficient PGMs especially for the smooth problems (see also [2, 4] for related or extensive works). A variant of this approach [50, 53], based on the relation (R_k), was exploited in this thesis. Here we observe that Property A is closely related to the Nesterov's estimate sequence framework.

Consider a convex optimization problem $\min_{x\in Q} f(x)$ (It will be helpful to consider the smooth problems so that $m_f(y;x) = f(y) + \langle \nabla f(y), x-y\rangle + \sigma_f \xi(y,x)$). For a sequence of (auxiliary) functions $\{\Phi_k(x)\}_{k\geq 0}$ and positive real numbers $\{T_k\}_{k\geq 0}$, the sequence $\{(\Phi_k(x), T_k)\}_{k\geq 0}$ is called an *estimate sequence* [49, Definition 2.2.1] if $T_k \to 0$ and

$$\Phi_k(x) \leq (1-T_k)f(x) + T_k\Phi_0(x), \quad \forall x \in E,\ \forall k \geq 0. \tag{5.1.1}$$

For a sequence $\{\varphi_k(x)\}_{k\geq 0}$ of (auxiliary) functions, let us consider Property A to deal with a relation to the estimate sequence. For simplicity, we consider weight parameters $\{\lambda_k\}_{k\geq 0}$ and scaling parameters $\{\beta_k\}_{k\geq -1}$ such that

$$\lambda_0 = 1, \quad \beta_k \equiv 1.$$

Given weight parameters $\{\lambda_k\}_{k\geq 0}$, denote $\tau_k := \lambda_{k+1}/S_{k+1}$, $T_0 := 1$, $T_k := \prod_{i=0}^{k-1}(1-\tau_i)$ and $\Phi_k(x) := \varphi_k(x)/S_k$. Then we have $1-\tau_k = S_k/S_{k+1}$ and $T_k = 1/S_k$ because of the recurrence $(1-\tau_k)/S_k = 1/S_{k+1}$.

We observe that, for the sequence $\{(\Phi_k(x), T_k)\} = \{(\varphi_k(x)/S_k, 1/S_k)\}$, the condition (A3) is closely related to the estimate sequence (5.1.1) and the condition (A2) is connected to the formula of the construction of estimate sequences.

The DA update (4.3.3), namely, $\varphi_{-1}(x) = \beta_{-1}d(x)$, $\varphi_{k+1}(x) = \varphi_k(x) + \lambda_{k+1}m_f(x_{k+1};x) + \beta_{k+1}d(x) - \beta_k d(x)$, is now equivalent to

$$\begin{aligned}\Phi_0(x) &= m_f(x_0;x) + d(x),\\ \Phi_{k+1}(x) &= (1-\tau_k)\Phi_k(x) + \tau_k m_f(x_{k+1};x).\end{aligned}$$

The update formula corresponds to the Nesterov's construction (see eq. (2.2.3) in [49]) of estimate sequence. Moreover, the EMD update (4.3.7), namely, $\varphi_{-1}(x) = \beta_{-1}\ell_d(z_{-1};x)$, $\varphi_{k+1}(x) = \varphi_k(z_k) + \lambda_{k+1}m_f(x_{k+1};x) + \beta_{k+1}d(x) - \beta_k\ell_d(z_k;x) + S_k\sigma_f\xi(z_k,x)$ can be rewritten as

$$\begin{aligned}\Phi_0(x) &= m_f(x_0;x) + \xi(z_{-1},x), \qquad (5.1.2)\\ \Phi_{k+1}(x) &= (1-\tau_k)\Phi_k(z_k) + \tau_k m_f(x_{k+1};x) + (T_{k+1} + (1-\tau_k)\sigma_f)\xi(z_k,x)\\ &= (1-\tau_k)[\Phi_k(z_k) + \gamma_k\xi(z_k,x)] + \tau_k m_f(x_{k+1};x), \quad \gamma_k := T_k + \sigma_f,\end{aligned}$$

which resembles with the second equation in [49, p. 74] (observe further that $\gamma_{k+1} = (1-\tau_k)\gamma_k + \tau_k\sigma_f$ holds). A relation between the condition (A2) and the estimate sequence can be seen in a similar manner as the EMD case.

The relation (R_k) $S_k f(\hat{x}_k) \le \min_{x\in Q}\varphi_k(x)$ is equivalent to $f(\hat{x}_k) \le \min_{x\in Q}\Phi_k(x)$ which corresponds to [49, eq. (2.2.2)].

Let us finally see the condition (A3). We have

$$\begin{aligned}\min_{x\in Q}\Phi_k(x) &\overset{(A3)}{\le} \min_{x\in Q}\left\{\frac{1}{S_k}\sum_{i=0}^{k}\lambda_i m_f(x_i;x) + T_k\ell_d(z_k;x)\right\}\\ &\le \min_{x\in Q}\left\{\frac{1}{S_k}\left[\sum_{i=1}^{k}\lambda_i f(x) + m_f(x_0;x)\right] + T_k d(z_k;x)\right\} \quad (\because\ f(x) \ge m_f(y;x))\\ &= \min_{x\in Q}\left\{\frac{\sum_{i=1}^{k}\lambda_i}{S_k}f(x) + T_k[m_f(x_0;x) + \ell_d(z_k;x)]\right\}\\ &= \min_{x\in Q}\{(1-T_k)f(x) + T_k[m_f(x_0;x) + \ell_d(z_k;x)]\}\\ &\le \min_{x\in Q}\{(1-T_k)f(x) + T_k[m_f(x_0;x) + d(x)]\}\\ &\le \min_{x\in Q}\{(1-T_k)f(x) + T_k[m_f(x_0;x) + \xi(z_{-1},x)]\} \quad (\because\ \ell_d(z_{-1},x) \ge 0, \forall x\in Q)\\ &\le \min_{x\in Q}\{(1-T_k)f(x) + T_k\Phi_0(x)]\},\end{aligned}$$

where the last inequality follows from (A2) with $k = -1$ (cf. (5.1.2)). Consequently, (A3) yields

$$\min_{x\in Q}\Phi_k(x) \le \min_{x\in Q}\{(1-T_k)f(x) + T_k\Phi_0(x)\}$$

which is closely related to the definition (5.1.1) of the estimate sequence.

Our unified analysis taken in Section 4.4 was based on the Nesterov's variant [50, 53] of the estimate sequence. The crucial contribution in view of the unifying framework would be Theorem 4.3.1 which allowed not only the DA model but also the EMD model to be handled via Property A.

5.2 Further research directions

The methodology used in our unifying framework could be a useful tool for further development of subgradient-based methods. We show some considerable topics for further application and research directions.

In general, the methodology in this thesis could be modified for other types of methods or problems. For instance, one can examine to consider modifications of Property A (as well as the definition of the structured problems, and so on) for such situations.

There are important classes of convex optimization problems which were not addressed in this thesis. It could be interesting to consider the possibility of application of our methodology to them.

- For instance, the PGMs proposed in [12, 25, 26, 33] addressed the *stochastic optimization*, namely, the oracle has an inexactness in a stochastic manner, while this thesis employed the deterministic setting. It will be interesting to examine the stochastic setting in our unifying framework.

- Another example is the class of *uniformly convex functions* which is a generalization of the strong convexity: f is said to be uniformly convex on Q with coefficient $\sigma \geq 0$ and exponent $\tau \geq 2$ if

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{\sigma}{\tau}\alpha(1-\alpha)[\alpha^{\tau-1} + (1-\alpha)^{\tau-1}] \|x-y\|^{\tau}$$

  for all $x, y \in Q$ and $\alpha \in [0,1]$. See [62, Section 3.5] for an elegant treatment on uniformly convex functions. Some optimal PGMs [33, 45] are known using the multistage procedure. The inexact oracle model [17] in the strongly convex case can also be applied to the uniformly convex case giving nearly optimal PGMs.

- The *smoothable functions* investigated by Nesterov [50] is an important class of (non-smooth) convex functions which can be efficiently minimized via the so called smoothing technique. According to an extensive study by Beck and Teboulle [9], Method II applied for the smooth problems belongs to the class of *fast iterative methods*, so that it can be used to construct *smoothing-based first-order methods*.

For the weakly smooth problems ($f \in \mathcal{F}_M^{\nu}(Q)$), the obtained convergence results (Theorems 4.7.2 and 4.7.3) may be less practical since we will need to know M, ν, and $\sigma_f \in \sigma(f)$ in advance. In fact, there are adaptive PGMs [36, 54, 61] in the non strongly convex case. An adaptive approach for parameters (M, ν, σ_f) in the strongly convex case will valuable for future study. For instance, one can examine various choices of weight and scaling parameters tuning the general bound in Theorems 4.4.1 and 4.4.2. The presented choices shown in our convergence results are just examples to ensure the optimality.

Some approaches for PGMs which were not employed in this thesis could be useful to improve practicality. For instance, we did not discuss about the so called *backtracking* or *line search* procedure [7, 47, 53, 54]. For smooth problems ($f \in \mathcal{F}_L^1(Q)$), this procedure can adapt (unknown) Lipschitz constant L. For weakly smooth problems ($f \in \mathcal{F}_M^{\nu}(Q)$), we can also expect to adapt parameters (M, ν) as the Nesterov's universal gradient method [54]. Another considerable approach is the *multistage* or *restarting* procedure [33, 45, 47, 53]. This is useful to construct optimal complexity methods in the strongly or the uniformly convex case.

The *gradient sliding* technique [37, 39, 40] for convex optimization problems, say, $\min_{x \in Q}\{f(x) \equiv g(x) + h(x)\}$, enables to reduce the iteration complexity by distinguishing the

call of two oracles for $g(x)$ and $h(x)$ instead of the one for $f(x)$. Recall that the complexity result in Section 4.7 holds in particular for the mixed smoothness structure (that is, $g \in \mathcal{F}_L^1(Q)$ and $h \in \mathcal{F}_M^\nu(Q)$). It can be considered the gradient sliding technique for our methods in this case.

Bibliography

[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, 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] M. Baes, Estimate sequence methods: Extensions and approximations, Technical Report, Institute of Operations Research, ETH Zürich, 2009. Available at `http://www.optimization-online.org/DB_HTML/2009/08/2372.html`

[5] V. Barbu and T. Precupanu, *Convexity and optimization in Banach spaces*, Forth edition, Springer, Dordrecht, 2012.

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

[7] 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.

[8] A. Beck and M. Teboulle, Gradient-Based Algorithms with Applications to Signal Recovery Problems, in *Convex Optimization in Signal Processing and Communications* (D. Palomar, Y. Eldar, eds.), pp. 33–88, Combridge University Press, New York, 2010.

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

[10] S. Boyd and N. Parikh, Proximal algorithms, *Foundations and Trends in Optimization*, **1**, pp. 123–231, 2011.

[11] 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.

[12] 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.

[13] G. Chen and M. Teboulle, Convergence analaysis of a proximal-like minimization algorithm using Bregman functions, *SIAM Journal on Optimization*, **3**, pp. 538–543, 1993.

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

[15] C. D. Dang, K. Dai, and G. Lan, A linearly convergent first-order algorithm for total variation minimisation in image processing, *International Journal of Bioinformatics Research and Applications*, **10**, pp. 4–26, 2014.

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

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

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

[19] 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.

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

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

[22] R. M. Freund and P. Grigas, New analysis and results for the Frank-Wolfe method, *Mathematical Programming*, **155**, pp. 199–230, 2016.

[23] 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.

[24] S. Ghadimi, Conditional gradient type methods for composite nonlinear and stochastic optimization, *arXiv:1602.00961*, 2016.

[25] 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.

[26] 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.

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

[28] Z. Harchaoui, A. Juditsky, and A. Nemirovski, Conditional gradient algorithms for norm-regularized smooth convex optimization, *Mathematical Programming*, **152**, pp. 75–112, 2015.

[29] M. Ito, New results on subgradient methods for strongly convex optimization problems with a unified analysis, *Computational Optimization and Applications*, **65**, pp. 127–172, 2016.

[30] M. Ito and M. Fukuda, A family of subgradient-based methods for convex optimization problems in a unifying framework, *Optimization Methods and Software*, **31**, pp. 952–982, 2016.

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

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

[33] A. Juditsky and Y. Nesterov, Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization, *Stochastic Systems*, **4**, pp. 44–80, 2014.

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

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

[36] G. Lan, Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization, *Mathematical Programming*, **149**, pp. 1–45, 2015.

[37] G. Lan, Gradient sliding for composite optimization, *Mathematical Programming*, **159**, pp. 201–235, 2016.

[38] G. Lan, Z. Lu, and R.D.C. Monteiro, Primal-dual first-order methods with $\mathcal{O}(1/\varepsilon)$ iteration-complexity for cone programming, *Mathematical Programming*, **126**, pp. 1–29, 2011.

[39] G. Lan and Y. Ouyang, Accelerated gradient sliding for structured convex optimization, *arXiv:1609.04905v1*, 2016.

[40] G. Lan and Y. Zhou, Conditional gradient sliding for convex optimization, *SIAM Journal on Optimization*, **26**, pp. 1379–1409, 2016.

[41] H. Lu, R. M. Freund, and Y. Nesterov, Relatively-Smooth Convex Optimization by First-Order Methods, and Applications, *arXiv:1610.05708v1*, 2016.

[42] 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, 2000.

[43] A. Nedić and S. Lee, On stochastic subgradient mirror-descent algorithm with weighted averaging, *SIAM Journal on Optimization*, **24**, pp. 84–107, 2014.

[44] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, Robust stochastic approximation approach to stochastic programming, *SIAM Journal on Optimimization* **19**, pp. 1574–1609, 2009.

[45] A. Nemirovski and Y. Nesterov, Optimal methods for smooth convex minimization, *Zh. Vychishl. Mat. i Mat. Fiz.*, **25**, pp. 356–369, 1985 (in Russian); English translation: *USSR Computational Mathematics and Mathematical Physics*, **24**, pp. 80–82, 1984.

[46] 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, 1983.

[47] 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.

[48] Y. Nesterov, On an approach to the construction of optimal methods of minimization of smooth convex functions, *Ekonomika i Mathematicheskie Metody*, **24**,pp. 509–517, 1988.

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

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

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

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

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

[54] Y. Nesterov, Universal gradient methods for convex optimization problems, *Mathematical Programming*, **152**, pp. 381–404, 2015.

[55] Y. Nesterov, Complexity bounds for primal-dual methods minimizing the model of objective function, *CORE Discussion Paper*, **2015/3**, 2015.

[56] Y. Nesterov and V. Shikhman, Quasi-monotone subgradient methods for nonsmooth convex minimization, *Journal of Optimization Theory and Applications* **165**, pp. 917–940, 2015.

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

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

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

[60] R. T. Rockafellar, *Convex analysis*, Princeton University Press, New Jersey, 1970.

[61] A. Yurtsever, Q. Tran-Dinh, and V. Cevher, A universal primal-dual convex optimization framework, in *Advances in Neural Information Processing Systems 28*, pp. 3150–3158, 2015.

[62] C. Zălinescu, *Convex analysis in general vector spaces*, World Scientific Publishing Co. Inc., New Jersey, 2002.

Chapter 6

Appendix

6.1 Lemmas for the proof of Theorem 4.6.4

In order to complete the proof of Theorem 4.6.4, we need to estimate upper bounds of $1/S_k$ and $\sum_{i=0}^{k} S_i/S_k$ for the sequence $\{\lambda_k\}_{k\geq 0}$ defined by (4.6.4):

$$\lambda_0 = 1, \quad \lambda_{k+1}^2 = \left(1 + S_k \frac{\sigma_f \sigma_d}{L - \bar{\sigma}_f \sigma_d}\right)(\lambda_{k+1} + S_k) \ (k \geq 0).$$

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 $\{S_k\}_{k\geq 0}$ is determined by the recurrence

$$S_0 = 1, \quad (S_{k+1} - S_k)^2 = S_{k+1}(1 + rS_k), \quad k \geq 0 \tag{6.1.1}$$

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

$$S_{k+1} = \frac{1 + (2+r)S_k + \sqrt{(1 + (2+r)S_k)^2 - 4S_k^2}}{2}. \tag{6.1.2}$$

The essentials of lemmas below are the same as [17, 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 (6.1.1) 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

$$\sqrt{S_{k+1}} - \sqrt{S_k} = \frac{S_{k+1} - S_k}{\sqrt{S_{k+1}} + \sqrt{S_k}} \geq \frac{S_{k+1} - S_k}{2\sqrt{S_{k+1}}} \overset{(6.1.1)}{=} \frac{1}{2}\sqrt{1 + rS_k} \geq \frac{1}{2} \tag{6.1.3}$$

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

$$S_k - S_0 = \sum_{i=0}^{k-1}(S_{i+1} - S_i) \overset{(6.1.1)}{=} \sum_{i=0}^{k-1}\sqrt{S_{i+1}(1 + rS_i)} \geq \sum_{i=0}^{k-1}\sqrt{S_{i+1}} \geq \sum_{i=0}^{k-1}\frac{i+3}{2} = \frac{k(k+5)}{4}$$

which gives $S_k \geq S_0 + \frac{k(k+5)}{4} = \frac{(k+1)(k+4)}{4}$. On the other hand, using (6.1.2) yields that

$$\frac{S_{k+1}}{S_k} = \frac{\frac{1}{S_k} + 2 + r + \sqrt{\left(\frac{1}{S_k} + (2+r)\right)^2 - 4}}{2} \geq \frac{2 + r + \sqrt{(2+r)^2 - 4}}{2} = \frac{2 + r + \sqrt{r^2 + 4r}}{2} \tag{6.1.4}$$

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

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

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

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

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 $\{S_k\}_{k\ge0}$ defined by (6.1.1) for $r>0$ satisfies*

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

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 $\frac{S_k}{S_{k+1}} \le 1-\frac{1}{\gamma}$ by (6.1.4). Now the result follows by induction: If $\sum_{i=0}^k S_i/S_k \le \gamma$ holds for some $k \ge 0$, we have

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

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

Note that the result of Lemma A.2 is the same as [17, 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 $\{S_k\}_{k\ge0}$ be defined as Lemma A.2 and $\{T_k\}_{k\ge0}$ be defined by (6.1.1) with $r:=0$, namely $T_0:=1$ and $T_{k+1} := \frac{1+2T_k+\sqrt{1+4T_k}}{2}$ for $k\ge0$. Then, we have*

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

Proof. Due to the identity

$$\frac{\sum_{i=0}^k S_i}{S_k} = 1+\sum_{i=0}^{k-1}\frac{S_i}{S_k} = 1+\sum_{i=0}^{k-1}\prod_{j=i}^{k-1}\frac{S_j}{S_{j+1}}, \quad k\ge0,$$

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

$$\frac{S_{k+1}}{S_k} = \frac{\frac{1+rS_k}{S_k} + 2 + \sqrt{\left(\frac{1+rS_k}{S_k}+2\right)^2 - 4}}{2}, \quad \frac{T_{k+1}}{T_k} = \frac{\frac{1}{T_k} + 2 + \sqrt{\left(\frac{1}{T_k}+2\right)^2 - 4}}{2}, \tag{6.1.5}$$

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

$$\begin{aligned} \frac{1+rS_{k+1}}{S_{k+1}} &\ge \frac{1+rS_k}{S_{k+1}} = \frac{S_k}{S_{k+1}}\alpha \overset{(6.1.5)}{=} \frac{2\alpha}{\alpha + 2 + \sqrt{(\alpha+2)^2-4}} \\ &\ge \frac{2\beta}{\beta + 2 + \sqrt{(\beta+2)^2-4}} \overset{(6.1.5)}{=} \frac{T_k}{T_{k+1}}\beta = \frac{1}{T_{k+1}} \end{aligned}$$

since $S_{k+1} \ge S_k$ 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 $\frac{1+rS_k}{S_k} \ge \frac{1}{T_k}$ for all $k \ge 0$ and therefore the proof is completed. □

Lemma A.4. *Let $\{T_k\}_{k\ge 0}$ be a sequence defined by (6.1.1) with $r := 0$, namely,*

$$T_0 := 1, \quad T_{k+1} := \frac{1 + 2T_k + \sqrt{1+4T_k}}{2} \quad (k \ge 0).$$

Define $\{t_k\}_{k\ge 0}$ by $t_0 := 1$ and $t_{k+1} := T_{k+1} - T_k$ for $k \ge 0$. Then, the followings hold.

(i) $t_k^2 = \sum_{i=0}^k t_i = T_k$ and $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ hold for $k \ge 0$.

(ii) We have

$$\frac{k+2}{2} \le \frac{\sqrt{(k+1)(k+4)}}{2} \le t_k \le \frac{k+2}{2} + \frac{1}{4}\log(k+1), \quad \forall k \ge 0.$$

(iii) We have

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

Proof. (i) The definition of t_k yields $T_k = \sum_{i=0}^k t_i$. The recurrence relation of T_k implies $t_k^2 = (T_k - T_{k-1})^2 = T_k$ and

$$t_{k+1} = T_{k+1} - T_k \overset{(6.1.2)}{=} \frac{1+\sqrt{1+4T_k}}{2} = \frac{1+\sqrt{1+4t_k^2}}{2}, \quad \forall k \ge 0.$$

(ii) Lemma A.1 yields $t_k = \sqrt{T_k} \ge \sqrt{(k+1)(k+4)/4} \ge (k+2)/2$ for $k \ge 0$. The right hand side for $k = 0$ is obvious. Analyzing the difference $t_{k+1} - t_k$ shows for $k \ge 0$ that

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

Therefore, we obtain

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

for all $k \geq 0$.
(iii) 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 remains to show $1 + \frac{T_k}{T_{k+1}}U_k \leq U_{k+1}$ for $k \geq 0$. In fact, (ii) 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 $1 + \frac{T_k}{T_{k+1}}U_k \leq U_{k+1}$. □