ISSN 1342-2804

Research Reports on Mathematical and Computing Sciences

A Family of Subgradient-Based Methods for Convex Optimization Problems in a Unifying Framework

Masaru Ito and Mituhiro Fukuda

February 2014, B–477

Department of
Mathematical and
Computing Sciences
Tokyo Institute of Technology

SERIES B: Applied Mathematical Science

A Family of Subgradient-Based Methods for Convex Optimization Problems in a Unifying Framework

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

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

Research Report B-477
Department of Mathematical and Computing Sciences
Tokyo Institute of Technology
February 2014

Abstract

We consider subgradient- and gradient-based methods for convex optimization problems whose feasible region is simple enough. We unify the way of constructing the subproblems which are necessary to be solved at each iteration of these methods. Our unifying framework provides a novel analysis of (sub)gradient-based methods and permit us to propose a family of optimal complexity methods. For the non-smooth case, we propose an extension of the mirror-descent method originally proposed by Nemirovski and Yudin [8] and overcame its drawback on optimal convergence. Our analysis is also applied to the dual-averaging method proposed by Nesterov [14] using simpler arguments for its complexity analysis. Finally, we propose (inexact) gradient-based methods for structured optimization problems such as with composite structure or using an inexact oracle model. The proposed family of classical gradient methods and its accelerations generalizes some of primal/dual gradient and Tseng's accelerated proximal gradient methods [6, 13, 16, 17].

Keywords: non-smooth/smooth convex optimization, structured convex optimization, subgradient/gradient-based proximal method, mirror-descent method, dual-averaging method, complexity bounds.

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

1 Introduction

1.1 Background

The gradient-based method proposed by Nesterov in 1983 for smooth convex optimization problems brought a surprising class of "optimal complexity" methods with preeminent performance over classical gradient methods for the worst case instances [9]. A minimization of a smooth convex function, whose gradient is Lipschitz continuous with constant L, by these optimal complexity methods ensure an ε-solution for the objective value within $O(\sqrt{LR^2/\varepsilon})$ iterations, while the

*corresponding author

classical gradient methods require $O(LR^2/\varepsilon)$ iterations; R is the distance between an optimal solution and the initial point. It is important to observe that in all of those methods, the iteration complexity is with respect to the convergence rate of the approximate optimal values and not with respect to the approximate optimal solutions. The Nesterov's optimal complexity method, as well as further improvements and extensions [1, 10, 12], applied or extended for solving non-smooth convex problems [4, 11, 12, 15, 16, 17] with composite structure [3, 11, 12, 13] and the inexact oracle model [6], changed substantially the approach on how to solve large-scale structured convex optimization problems arising in compressed sensing, image processing, statistics, *etc.*

Many of those methods for smooth and structured convex problems were unified and generalized by Tseng [16, 17] and it was shown that they preserve the $O(\sqrt{LR^2/\varepsilon})$ iteration complexity. It is important to notice that the computational complexity of each iteration will strongly depend on the structure of the objective function, the feasible region, and the choice of the proximal operation to define the auxiliary subproblems solved at each iteration.

For a general non-smooth convex problem, the complexity analysis of those methods becomes quite different compared to the smooth case. The optimal complexity in the non-smooth case is $O(M^2R^2/\varepsilon^2)$ iterations for an ε-solution, where M is a Lipschitz constant for the objective function. A well known optimal method for this case is the Mirror-Descent Method (MDM) proposed by Nemirovski and Yudin [8] which was later related to the subgradient algorithm by Beck and Teboulle [2]. It can be proved that the MDM achieves the optimal complexity when we know in advance an upper bound for R. The Dual-Averaging Method (DAM) proposed by Nesterov [14], on the other hand, removed this requirement and still ensured the same complexity. The key contribution of the DAM was the introduction of a sequence which we call *scalling parameter* β_k in this paper.

The (approximate) gradient-based methods proposed as the *primal* and *dual gradient methods* in [6, 13] can be interpreted as particular cases of the MDM and the DAM for the corresponding smooth convex problems, respectively, as it will be clear along the article. They ensure the same complexity $O(LR^2/\varepsilon)$ as the classical gradient methods. Particular cases of Tseng's optimal methods [16, 17] can be also seen as accelerated versions of these methods.

1.2 Contribution

Our contribution will be a modest one over these methods, *i.e.*, identifying a common property intrinsic to the MDM and the DAM. This property will provide a unifying framework for analyzing (sub)gradient-based methods which can be applied for non-smooth and smooth cases as well as for structured problems.

Our unifying framework will permit us to propose a new family of optimal methods which includes the MDM and the DAM for the non-smooth case (Method 9).

We will introduce two models to update the auxiliary subproblems, which are necessary to be solved at each iteration, called the *extended MD model* (11) and the *DA model* (12). Based on them, our family of methods ensure the optimal complexity $O(M^2R^2/\varepsilon^2)$ for the non-smooth case. As a by-product, the proposed extended MDM overcomes the drawback of the original MDM which requires in advance the knowledge of an upper bound R to ensure the optimal complexity.

Furthermore, our unifying framework can be extended to structured problem whose objective functions are smooth, have composite structure, saddle structure, or their values as well as their subgradients depend on an inexact oracle [6]. For all these cases, we propose two general methods (Methods 16 and 17) whose complexity to obtain an ε-solution are $O(LR^2/\varepsilon)$ and $O(\sqrt{LR^2/\varepsilon})$, respectively (excepting for the inexact oracle case). The former method includes the classical gradient methods analyzed in [6, 13] and the latter one, which is optimal for smooth optimization, includes Tseng's second and third Accelerated Proximal Gradient (APG) methods [17] which are particular cases of Tseng's ones [16].

A unifying framework for structured optimization problems was already provided by Tseng [16, 17]. However, our result distinguishes in finding a common set of properties (Property 2) which the auxiliary functions should satisfy. This fact will simplify the subsequence convergence analysis of the proposed methods. In fact, our optimal complexity methods can be regarded as situated between the extended MDM and the DAM. Table 1 shows the relation between our family of (sub)gradient-based methods and other known methods.

Table 1: Relation between our family of (sub)gradient-based methods and other known methods. The star (*) corresponds to our result. "Complexity" indicates the number of iteration to obtain an ε-solution when the objective function has no inexactness for its oracle. [13] is included considering that its Lipschitz constant is known in advance.

problem class	complexity	known methods	generalized methods
non-smooth	optimal $O\left(M^2R^2/\varepsilon^2\right)$	mirror-descent [2, 8]	*Method 9 (a) with the model (11) $\equiv$ extended mirror-descent: Method 13
		dual-averaging [14]	*Method 9 (a) with the model (12)
structured/ smooth	classical $O(LR^2/\varepsilon)$	primal gradient [6, 7, 13]	*Method 16 with the model (11)
		dual gradient [6, 13]	*Method 16 with the model (12)
	optimal $O(\sqrt{LR^2/\varepsilon})$	FISTA [3]	Tseng's first APG [17]
		Nesterov's method [12]	Tseng's modified method; see [16, (35-36)]
		Tseng's second APG [17]	*Method 20 $\equiv$ Method 17 with the model (11) Tseng's method [16, Algorithm 1]
		Tseng's third APG [17]	*Method 21 $\equiv$ Method 17 with the model (12) Tseng's method [16, Algorithm 3]

At each iteration, all of the above methods need to solve one or two convex subproblems which can be easy or hard depending on the structure and/or the feasible region of the original problem. Nesterov's optimal methods [12, 13] require two subproblems at each iteration while Tseng's ones need only one subproblem and even preserve the same complexity. The proposed methods in this paper need only one subproblem at each iteration.

The structure of this article is as follows. First, we review some existing methods, in particular the MDM and the DAM for non-smooth and smooth objective functions (Section 2). In Section 3, we propose the Property 2 which will be required for all auxiliary functions of our methods, as well as some supporting lemmas. We then propose in Section 4 the unifying method and prove its convergence rate, in particular for the extended MDM and the DAM, and subsequently for the structured problems in Section 5.

1.3 Notations

In this paper, we consider a finite dimensional real vector space E endowed with a norm $\|\cdot\|$. The dual space of E is denoted by E^* endowed with the dual norm $\|\cdot\|_*$ defined by

$$\|s\|_* = \max_{\|x\|\leq 1} \langle s, x\rangle, \quad s \in E^*$$

where $\langle s, x\rangle$ denotes the value of $s \in E^*$ at $x \in E$. We consider subgradient-based and gradient-based methods to solve the following convex optimization problem :

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

where Q is a nonempty closed convex and possibly unbounded subset of E, and $f : E \to \mathbb{R}\cup\{+\infty\}$ is a proper lower semicontinuous convex function with $Q \subset \operatorname{dom} f := \{x \in E : f(x) < +\infty\}$. For each $x \in \operatorname{dom} f$, the subdifferential of f at x is denoted by $\partial f(x) := \{g \in E^* : f(y) \geq$

$f(x) + \langle g, y - x\rangle\,,\ \forall y \in E\}$. We assume throughout this paper that the problem (1) always has an *optimal solution* $x^* \in Q$, and the structure of Q is simple enough or has some special structure which permits one to solve a subproblem over it with moderate easiness. See [12] for some examples.

We additionally assume that there is a proper lower semi-continuous convex function $d : E \to \mathbb{R} \cup \{+\infty\}$ satisfying the following properties:

- $d(x)$ is a strongly convex function on Q with parameter $\sigma > 0$, *i.e.*,

$$d(\tau x + (1-\tau)y) \le \tau d(x) + (1-\tau)d(y) - \frac{1}{2}\sigma\tau(1-\tau)\|x-y\|^2, \quad \forall x, y \in Q,\ \forall \tau \in [0,1].$$

- $d(x)$ is continuously differentiable on Q.

We denote by $\xi(z, x)$ the Bregman distance [5] between z and x:

$$\xi(z, x) := d(x) - d(z) - \langle \nabla d(z), x - z\rangle, \quad z, x \in Q.$$

The Bregman distance satisfies $\xi(z, x) \ge \frac{\sigma}{2}\|x - z\|^2$ for any $x, z \in Q$ by the strong convexity of $d(x)$.

We also assume that $d(x_0) = \min_{x\in Q} d(x) = 0$ for $x_0 := \operatorname{argmin}_{x\in Q} d(x) \in Q$, which is used for the initial point of our methods.[1] Finally, we define R as $R := \sqrt{\frac{1}{\sigma}d(x^*)}$, $R := \sqrt{\frac{1}{\sigma}\xi(x_0, x^*)}$, or their upper bounds, which quantifies the distance between the optimal solution x^* and the initial point x_0 in view of properties $d(x_0) = 0$ and $d(x) \ge \frac{\sigma}{2}\|x - x_0\|^2$ for every $x \in Q$.

2 Existing optimal methods

In this section, we review some well-known subgradient-based and gradient-based methods. In particular, we focus on the Mirror-Descent Method (MDM) and Dual-Averaging Method (DAM) for non-smooth objective functions, and on Nesterov's accelerated gradient and Tseng's Accelerated Proximal Gradient (APG) methods for smooth objective functions (or for non-smooth ones with some special structures).

The purpose of this section is to unify the notation of these methods in order to introduce a unifying framework for them in Section 3. For that, we sometimes changed the variables' names, shifted their indices, and added constants in the objective functions of optimization subproblems compared to the original articles.

2.1 Optimal methods for the non-smooth case

Let us first assume that $f(x)$ in (1) is non-smooth. The MDM [8] in the form reinterpreted by Beck and Teboulle [2] generates the following iterates from the initial point $x_0 := \operatorname{argmin}_{x\in Q} d(x) \in Q$.

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

where $g_k \in \partial f(x_k)$ and $\lambda_k > 0$ is a *weight*. The parameter λ_k is also referred to as a *stepsize*; it is known that the procedure (2) reduces to the classical subgradient method $x_{k+1} := \pi_Q(x_k - \lambda_k g_k)$ when E is an Euclidean space, $\|\cdot\|$ is the norm of E induced by its inner product, $d(x) := \frac{1}{2}\|x - x_0\|^2$, and π_Q is the orthogonal projection onto Q (see also Auslender-Teboulle [1] and Fukushima-Mine [7] for some related works).

[1] We can always assume this requirement for an arbitrary point $x_0 \in Q$ by replacing $d(x)$ by $\xi(x_0, x)$.

The MDM produces the following estimate [2]:

$$\forall k \geq 0, \quad \min_{0\leq i\leq k} f(x_i) - f(x^*) \leq \frac{\sum_{i=0}^k \lambda_i f(x_i)}{\sum_{i=0}^k \lambda_i} - f(x^*) \leq \frac{\xi(x_0,x^*) + \frac{1}{2\sigma}\sum_{i=0}^k \lambda_i^2 \|g_i\|_*^2}{\sum_{i=0}^k \lambda_i} \tag{3}$$

where the right hand side can be bounded by $M\sqrt{2\sigma^{-1}\xi(x_0,x^*)}/\sqrt{k+1}$ if $M := \sup\{\|g\|_* : g \in \partial f(x),\ x \in Q\}$ is finite and if we choose the constant weights $\lambda_i := M^{-1}\sqrt{2\sigma\xi(x_0,x^*)}/\sqrt{k+1},\ i = 0,\ldots,k$ for a fixed $k \geq 0$. If we further know an upper bound $R \geq \sqrt{\frac{1}{\sigma}\xi(x_0,x^*)}$, this convergence result ensures an ε-solution in $O(M^2R^2/\varepsilon^2)$ iterations which provides the optimal complexity for the non-smooth case [2]. The above choice of weights, however, is impractical since it depends on the final iterate k and an upper bound for $\xi(x_0,x^*)$; a more practical choice $\lambda_i := r/\sqrt{i+1}$ for some $r > 0$ only ensures an upper bound $\frac{\xi(x_0,x^*)+(2\sigma)^{-1}r^2M^2(1+\log(k+1))}{2r(\sqrt{k+2}-1)} = O(\log k/\sqrt{k})$ for the right hand side of (3).

The DAM proposed by Nesterov [14] overcomes the dependence of weights of the MDM on k and even achieves the rate of convergence $O(1/\sqrt{k})$. This method employs non-decreasing positive *scaling parameters* $\{\beta_k\}_{k\geq -1}$ ($\beta_{k+1} \geq \beta_k > 0$) in addition to the weights $\{\lambda_k\}_{k\geq 0}$. From the initial point $x_0 := \operatorname{argmin}_{x\in Q} d(x) \in Q$, the DAM is performed as

$$x_{k+1} := \operatorname*{argmin}_{x\in Q}\left\{\sum_{i=0}^k \lambda_i[f(x_i) + \langle g_i, x - x_i\rangle] + \beta_k d(x)\right\},\ k = 0,1,2,\ldots \tag{4}$$

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

$$\hat{\beta}_{-1} := \hat{\beta}_0 := 1,\ \hat{\beta}_{k+1} := \hat{\beta}_k + \hat{\beta}_k^{-1}, \qquad \forall k \geq 0, \tag{5}$$

it yields the estimate

$$\forall k \geq 0, \quad \min_{0\leq i\leq k} f(x_i) - f(x^*) \leq \frac{\sum_{i=0}^k \lambda_i f(x_i)}{\sum_{i=0}^k \lambda_i} - f(x^*) \leq \left(\gamma d(x^*) + \frac{M^2}{2\sigma\gamma}\right)\frac{0.5 + \sqrt{2k+1}}{k+1},$$

which achieves the optimal complexity if we choose $\gamma := M/\sqrt{2\sigma d(x^*)}$.

A key of the analysis of the DAM in [14] is the use of dual approach such as the conjugate function of $\beta d(x)$ for $\beta > 0$. In this paper, we prove the same result with simpler arguments (in Section 4) for the DAM and (an extension of) the MDM without employing it.

Observe that for both methods, we do not need to evaluate any function value at any iteration and x_{k+1} is determined uniquely even if Q is unbounded since $d(x)$ is strongly convex [14, Lemma 6].

2.2 Optimal methods for the smooth case

Let us assume now that the function $f(x)$ in (1) is convex and continuously differentiable on Q. Nesterov [12] and Tseng [16, 17] proposed optimal methods whenever we could further assume that its gradient is Lipschitz continuous on Q. Let $L > 0$ be this Lipschitz constant, *i.e.*,

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

Given positive weights $\{\lambda_k\}_{k\geq 0}$, both methods depend on the following computation of optimal solutions $\hat{z}_k$ and/or z_k of auxiliary subproblems:

$$\begin{aligned} &\text{(a)} \quad \hat{z}_k := \operatorname{argmin}_{x\in Q}\left\{\lambda_k[f(x_k) + \langle \nabla f(x_k), x - x_k\rangle] + \tfrac{L}{\sigma}\xi(z_{k-1}, x)\right\}, \\ &\text{(b)} \quad z_k := \operatorname{argmin}_{x\in Q}\left\{\sum_{i=0}^k \lambda_i[f(x_i) + \langle \nabla f(x_i), x - x_i\rangle] + \tfrac{L}{\sigma}d(x)\right\} \end{aligned} \tag{6}$$

where $\{x_k\}_{k\geq 0} \subset Q$ is the sequence generated by those methods. Note that the subproblem (a) is closely related to the one of the MDM (2) and the subproblem (b) corresponds to the one of the DAM (4) with $\beta_k = L/\sigma$. Similarly to the non-smooth case, it is not necessary to evaluate the function values at x_k's and the minimums are uniquely defined.

The Nesterov's optimal method (see modified method in [12, Section 5.3]) with a particular choice for the weights λ_k is described as follow.

Nesterov's method: Set $\lambda_k := (k+1)/2$ for $k \geq 0$ and $x_0 := z_{-1} := \operatorname{argmin}_{x\in Q} d(x)$. Compute $\hat{z}_0$ by (a) and set $\hat{x}_0 := z_0 := \hat{z}_0$. For $k \geq 0$, iterate the following procedure:

$$\begin{array}{ll} \text{Set} & x_{k+1} := (1-\tau_k)\hat{x}_k + \tau_k z_k, \quad \text{where} \quad \tau_k := \frac{\lambda_{k+1}}{\sum_{i=0}^{k+1}\lambda_i}, \\ \text{Compute} & \hat{z}_{k+1} \text{ by (a)}, \\ \text{Set} & \hat{x}_{k+1} := (1-\tau_k)\hat{x}_k + \tau_k \hat{z}_{k+1}, \\ \text{Compute} & z_{k+1} \text{ by (b)}. \end{array} \tag{7}$$

On the other hand, Tseng's second and third Accelerated Proximal Gradient (APG) methods [17] which are particular cases of algorithms 1 and 3 in [16], only require the solution of either subproblem (a) or (b), respectively.

Tseng's second APG method: 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 $\hat{z}_0$ by (a) and set $\hat{x}_0 := \hat{z}_0$. For $k \geq 0$, iterate the following procedure:

$$\begin{array}{ll} \text{Set} & x_{k+1} := (1-\tau_k)\hat{x}_k + \tau_k \hat{z}_k, \quad \text{where} \quad \tau_k := \frac{\lambda_{k+1}}{\sum_{i=0}^{k+1}\lambda_i}, \\ \text{Compute} & \hat{z}_{k+1} \text{ by (a) replacing } z_k \text{ by } \hat{z}_k, \\ \text{Set} & \hat{x}_{k+1} := (1-\tau_k)\hat{x}_k + \tau_k \hat{z}_{k+1}. \end{array} \tag{8}$$

Tseng's third APG method: 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 z_0 by (b) and set $\hat{x}_0 := z_0$. For $k \geq 0$, iterate the following procedure:

$$\begin{array}{ll} \text{Set} & x_{k+1} := (1-\tau_k)\hat{x}_k + \tau_k z_k, \quad \text{where} \quad \tau_k := \frac{\lambda_{k+1}}{\sum_{i=0}^{k+1}\lambda_i}, \\ \text{Compute} & z_{k+1} \text{ by (b)}, \\ \text{Set} & \hat{x}_{k+1} := (1-\tau_k)\hat{x}_k + \tau_k z_{k+1}. \end{array} \tag{9}$$

Remark 1. To see the equivalence to Tseng's second APG method, notice that x_0 is not used at all in [17]. 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 $\sigma = 1$ 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.

It can be shown that both Nesterov's and Tseng's methods attain the optimal convergence rate; Nesterov's method (7) and Tseng's third APG method (9) satisfy

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

while Tseng's second APG method (8) satisfies

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

The convergence analysis of these three methods are performed in different ways. What we propose in Section 5 is a unified analysis for them using the following Nesterov's strategy [12]: Find a sequence $\{\hat{x}_k\} \subset Q$ such that

$$S_k f(\hat{x}_k) \leq \min_{x \in Q} \left\{ \sum_{i=0}^{k} \lambda_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] + \frac{L}{\sigma} d(x) \right\}$$

where $S_k = \sum_{i=0}^{k} \lambda_i$.

The above gradient-based methods for smooth problems can be generalized for non-smooth convex problems with special structures preserving the same iteration complexity. The structures of the *composite* objective function and the *inexact oracle model* are remarkably important since they have significant applications for compressed sensing, image processing, and statistics (see [3, 6, 17] for some examples). These structures will be detailed in Section 5. The Nesterov's method (7) was generalized for the composite structure [13] and the inexact oracle model [6]. Tseng's above methods were originally proposed for the composite objective function unifying some existing methods [1, 3, 12], while we only have described the particular ones for the smooth case.

3 General conditions for the auxiliary functions in the unifying framework

For all methods we reviewed for non-smooth or smooth objective functions, we need to form one or two *auxiliary functions* $\psi_k(x)$ and solve the corresponding subproblem(s) $\min_{x \in Q} \psi_k(x)$ at each iteration. In this section, we will propose general conditions which these auxiliary functions should satisfy in order to provide a unifying analysis. In particular, we can show that these auxiliary functions are derived from the extended MD model (11), the DA model (12), or a mixture of them. Based on these results, we will propose a family of methods in a unifying framework for non-smooth functions in Section 4 and for structured convex problems in Section 5 which includes the smooth functions.

We use the following notations for the description and the analysis of our methods. For a point $y \in Q$, denote by $l_f(y; x) : E \to \mathbb{R} \cup \{+\infty\}$ a proper lower semicontinuous convex function with $f(x) \geq l_f(y; x)$, $\forall x \in E$, *i.e.*, a lower approximation of $f(x)$ at $y \in Q$. The explicit description of the function $l_f(y; x)$ will be given in Sections 4 and 5 and will vary according to the property of $f(x)$. For the function $d(x)$, we denote $l_d(y; x) := d(y) + \langle \nabla d(y), x - y \rangle$. Note that $d(x) \geq l_d(y; x)$ and $\xi(y, x) = d(x) - l_d(y; x)$ for any $x, y \in Q$.

We introduce two kinds of "parameters" for our methods, namely, the *weight parameter* $\{\lambda_k\}_{k \geq 0}$ and the *scaling parameter* $\{\beta_k\}_{k \geq -1}$. We define $S_k := \sum_{i=0}^{k} \lambda_i$. Moreover, we use $\{\hat{x}_k\}_{k \geq 0} \subset Q$ and $\{x_k\}_{k \geq 0} \subset Q$ for sequences of approximate solutions and test points (for which we compute the (sub)gradients), respectively (recall that $x_0 := \operatorname{argmin}_{x \in Q} d(x)$).

Finally, we consider auxiliary functions $\psi_k(x)$ whose unique minimizers on Q are denoted by $z_k := \operatorname{argmin}_{x \in Q} \psi_k(x)$. The function $\psi_k(x)$ is assumed to be defined by $\{\lambda_i\}_{i=0}^{k}$, $\{\beta_i\}_{i=-1}^{k}$, $\{x_i\}_{i=0}^{k}$ and $\{z_i\}_{i=0}^{k-1}$ for each $k \geq 0$. We also consider $\psi_{-1}(x)$ (and $z_{-1} := \operatorname{argmin}_{x \in Q} \psi_{-1}(x)$) for convenience.

The following property will be a fundamental one for the construction of auxiliary functions $\{\psi_k(x)\}_{k \geq -1}$ in our unifying framework.

Property 2. *Let $\{\lambda_k\}_{k \geq 0}$ be a sequence of positive weight parameters, $\{\beta_k\}_{k \geq -1}$ be a sequence of positive and non-decreasing scaling parameters, and $\{x_k\}_{k \geq 0}$ be a sequence of test points. Let $\psi_k(x)$ be auxiliary functions which are determined by $\{\lambda_i\}_{i=0}^{k}$, $\{\beta_i\}_{i=-1}^{k}$, $\{x_i\}_{i=0}^{k}$, and $\{z_i\}_{i=0}^{k-1}$ where $z_i := \operatorname{argmin}_{x \in Q} \psi_i(x)$ for each $k \geq -1$. Then the following conditions hold:*

(i) $\min_{x \in Q} \psi_{-1}(x) = 0$ *and* $z_{-1} = x_0$.

(ii) The following inequality holds for every $k \geq -1$:

$$\forall x \in Q,\ \psi_{k+1}(x) \geq \min_{z \in Q} \psi_k(z) + \lambda_{k+1} l_f(x_{k+1}; x) + \beta_{k+1} d(x) - \beta_k l_d(z_k; x).$$

(iii) The following inequality holds for every $k \geq 0$:

$$\min_{x \in Q} \psi_k(x) \leq \min_{x \in Q} \left\{ \sum_{i=0}^{k} \lambda_i l_f(x_i; x) + \beta_k l_d(z_k; x) \right\}.$$

On the construction of an auxiliary function, the following lemma [16, Property 2] is useful.

Lemma 3. *Let $h : E \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous convex function with $Q \subset \mathrm{dom}\, h$ and β be a positive number. Denote $\psi(x) = h(x) + \beta d(x)$. Then the minimization problem $\min_{x \in Q} \psi(x)$ has a unique solution $z^* \in Q$ and it satisfies*

$$\psi(x) \geq \psi(z^*) + \beta \xi(z^*, x), \quad \forall x \in Q.$$

We now propose a family of auxiliary functions which satisfy Property 2.

$$\left.\begin{array}{l} \text{(0) Define } \psi_{-1}(x) := \beta_{-1} d(x). \\ \text{(1) For each } k \geq -1, \text{ define } \psi_{k+1}(x) \text{ by either the } \textit{extended Mirror-Descent (MD)} \\ \quad \textit{model} \text{ (11) or the } \textit{Dual-Averaging (DA) model} \text{ (12).} \end{array}\right\} \tag{10}$$

Extended MD model:

$$\psi_{k+1}(x) := \min_{z \in Q} \psi_k(z) + \lambda_{k+1} l_f(x_{k+1}; x) + \beta_{k+1} d(x) - \beta_k l_d(z_k; x). \tag{11}$$

DA model:

$$\psi_{k+1}(x) := \psi_k(x) + \lambda_{k+1} l_f(x_{k+1}; x) + \beta_{k+1} d(x) - \beta_k d(x). \tag{12}$$

In both cases, $\psi_{k+1}(x)$ is proper lower semicontinuous and strongly convex on Q.

Proposition 4. *Any sequence of auxiliary functions $\{\psi_k(x)\}$ constructed by (10) satisfies Property 2.*

Proof. Since $\min_{x \in Q} d(x) = d(x_0) = 0$, $\psi_{-1}(x) = \beta_{-1} d(x)$ satisfies condition (i).

If we construct $\psi_{k+1}(x)$ by (11), then it is clear that the condition (ii) holds. Let us consider the case (12). Notice that on the construction (10), we can show by induction that the functions $h_k(x) := \psi_k(x) - \beta_k d(x)$ are always proper lower semicontinuous and convex. Thus Lemma 3 implies that $\psi_k(x) \geq \min_{z \in Q} \psi_k(z) + \beta_k \xi(z_k, x)$ for every $x \in Q$. Therefore, we obtain

$$\begin{aligned} \psi_{k+1}(x) &= \psi_k(x) + \lambda_{k+1} l_f(x_{k+1}; x) + \beta_{k+1} d(x) - \beta_k d(x) \\ &\geq [\min_{z \in Q} \psi_k(z) + \beta_k \xi(z_k, x)] + \lambda_{k+1} l_f(x_{k+1}; x) + \beta_{k+1} d(x) - \beta_k d(x) \\ &= \min_{z \in Q} \psi_k(z) + \lambda_{k+1} l_f(x_{k+1}; x) + \beta_{k+1} d(x) - \beta_k l_d(z_k; x) \end{aligned}$$

for all $x \in Q$.

Let us finally prove the condition (iii) by induction. We actually show that it is also valid for $k \geq -1$. The case $k = -1$ is due to the optimality condition for $z_{-1} = \mathrm{argmin}_{x \in Q} \psi_{-1}(x) =$

$\mathrm{argmin}_{x\in Q}\,\beta_{-1}d(x)$, that is, $\min_{x\in Q}\,\beta_{-1}d(x) = \min_{x\in Q}\,\beta_{-1}l_d(z_{-1};x)$ holds. Suppose that the condition (iii) holds for some $k \geq -1$. Consider the auxiliary function $\psi_{k+p}(x)$ for a positive integer p defined as follows. Define $\psi_{k+1}(x)$ by (11) and define $\psi_{k+i+1}(x)$ by (12) for $i = 1, \ldots, p-1$. Then

$$\psi_{k+p}(x) = \psi_k(z_k) + \sum_{i=k+1}^{k+p} \lambda_i l_f(x_i;x) + \beta_{k+p}d(x) - \beta_k l_d(z_k;x), \tag{13}$$

and using Lemma 3 we have

$$\begin{aligned}
\min_{x\in Q}\psi_{k+p}(x) &\leq \psi_{k+p}(x) - \beta_{k+p}\xi(z_{k+p},x) \\
&= \left[\psi_k(z_k) + \sum_{i=k+1}^{k+p} \lambda_i l_f(x_i;x) + \beta_{k+p}d(x) - \beta_k l_d(z_k;x)\right] - \beta_{k+p}\xi(z_{k+p},x) \\
&\overset{\text{(iii)}}{\leq} \left[\sum_{i=0}^{k} \lambda_i l_f(x_i;x) + \beta_k l_d(z_k;x)\right] \\
&\qquad + \sum_{i=k+1}^{k+p} \lambda_i l_f(x_i;x) + \beta_{k+p}d(x) - \beta_k l_d(z_k;x) - \beta_{k+p}\xi(z_{k+p},x) \\
&= \sum_{i=0}^{k+p} \lambda_i l_f(x_i;x) + \beta_{k+p} l_d(z_{k+p};x)
\end{aligned}$$

for all $x \in Q$. This proves the condition (iii) for $\psi_{k+p}(x)$. It is, therefore, enough to prove the condition (iii) in the case when the auxiliary function $\psi_k(x)$ is defined only by (12) updates. We have $\psi_k(x) = \sum_{i=0}^{k} \lambda_i l_f(x_i;x) + \beta_k d(x)$ in this case and again Lemma 3 implies that

$$\min_{x\in Q}\psi_k(x) \leq \psi_k(x) - \beta_k\xi(z_k,x) = \sum_{i=0}^{k} \lambda_i l_f(x_i;x) + \beta_k l_d(z_k;x)$$

for every $x \in Q$. □

Remark 5. Proposition 4 proves that the following auxiliary functions satisfy Property 2 for appropriate choices of $l_f(x_i;x)$'s.

- Constructing $\{\psi_k(x)\}$ by (10) with only extended MD model updates (11) yields

$$\begin{aligned}
\psi_k(x) &= \min_{x\in Q}\,\psi_{k-1}(x) + \lambda_k l_f(x_k;x) + \beta_k d(x) - \beta_{k-1} l_d(z_{k-1};x), \\
z_k &= \mathrm{argmin}_{x\in Q}\left\{\lambda_k l_f(x_k;x) + \beta_k d(x) - \beta_{k-1} l_d(z_{k-1};x)\right\}
\end{aligned} \tag{14}$$

  which coincides with the MDM (2) for $\beta_k = 1$ and $x_k = z_{k-1}$, and the subproblems of Tseng's second APG method (8) for $\beta_k = L/\sigma$.

- Constructing $\{\psi_k(x)\}$ by (10) with only DA model updates (12) yields

$$\begin{aligned}
\psi_k(x) &= \textstyle\sum_{i=0}^{k} \lambda_i l_f(x_i;x) + \beta_k d(x), \\
z_k &= \mathrm{argmin}_{x\in Q}\left\{\textstyle\sum_{i=0}^{k} \lambda_i l_f(x_i;x) + \beta_k d(x)\right\}
\end{aligned} \tag{15}$$

  which coincides with the DAM (4) and the subproblems of Tseng's third APG method (9) with $\beta_k = L/\sigma$.

Notice that a pure extended MD model updates (14) considers only the previous $l_f(x_k; x)$ while the DA model updates (15) accumulates all $l_f(x_i; x)$'s. Moreover, Proposition 4 shows that Property 2 is satisfied even if we mix the strategies (14) and (15) which correspond in selecting *some* of previous $l_f(x_i; x)$'s to define the subproblem as shown in (13). Note that, for a fixed $\psi_k(x)$, the construction (11) of $\psi_{k+1}(x)$ is the minimalist choice which satisfies Property 2; according to (ii), any auxiliary function $\psi_{k+1}(x)$ majorizes the one defined by (11) on the set Q.

To conclude this section, we define the following relation based on the Nesterov's approach [12]; we propose (sub)gradient-based methods which generates approximate solutions $\{\hat{x}_k\} \subset Q$ satisfying the following relation for every $k \geq 0$:

$$(R_k) \quad S_k f(\hat{x}_k) \leq \min_{x \in Q} \psi_k(x) + C_k \tag{16}$$

where C_k is defined according to the problem structure.

This relation yields the following lemma which provides a convergence rate of all methods.

Lemma 6. *Let $\{\psi_k(x)\}$ be a sequence of auxiliary functions satisfying Property 2 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 a sequence $\{\hat{x}_k\} \subset Q$ satisfies the relation (R_k) for some $k \geq 0$, then we have*

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

where $z_k := \operatorname{argmin}_{x \in Q} \psi_k(x)$.

Proof. Since $\sum_{i=0}^{k} \lambda_i l_f(x_i; x) \leq S_k f(x)$ for all $x \in Q$, using the condition (iii) of Property 2 yields

$$\min_{x \in Q} \psi_k(x) \leq \min_{x \in Q} \left\{ \sum_{i=0}^{k} \lambda_i l_f(x_i; x) + \beta_k l_d(z_k; x) \right\} \leq \min_{x \in Q} \{S_k f(x) + \beta_k l_d(z_k; x)\} \leq S_k f(x^*) + \beta_k l_d(z_k; x^*).$$

Therefore, the relation (R_k) implies

$$S_k f(\hat{x}_k) \leq \min_{x \in Q} \psi_k(x) + C_k \leq S_k f(x^*) + \beta_k l_d(z_k; x^*) + C_k.$$

□

4 A family of subgradient-based methods in the unifying framework

4.1 The unifying framework

In this section, we propose novel subgradient-based methods for solving problem (1). Throughout this section, we assume that subgradients of the objective function f, $g(y) \in \partial f(y)$, are computable at any point $y \in Q$ and a lower approximation $l_f(y; \cdot)$ at the same point is defined by

$$l_f(y; x) := f(y) + \langle g(y), x - y \rangle, \qquad \forall x \in Q.$$

For a test point $x_k \in Q$, we denote $g_k = g(x_k) \in \partial f(x_k)$. In this case, the subproblems $z_k = \operatorname{argmin}_{x \in Q} \psi_k(x)$ constructed from (10) are of the form

$$\min_{x \in Q} \{\langle s, x \rangle + \beta d(x)\} \tag{17}$$

for some $s \in E^*$ and $\beta > 0$.

We use the following lemma for our analysis.

Lemma 7. *Let $\{x_k\}_{k\geq 0} \subset Q$ and $g_k \in \partial f(x_k),\ k \geq 0$. Then, for $\lambda \in \mathbb{R},\ \beta > 0$ and $x, z \in Q$, we have*

$$\langle \lambda g_k, x - z\rangle + \beta\xi(z, x) + \frac{1}{2\sigma\beta}\|\lambda g_k\|_*^2 \geq 0, \quad \forall k \geq 0,$$

and, in particular,

$$\lambda l_f(x_k, x) + \beta\xi(x_k, x) + \frac{\lambda^2}{2\sigma\beta}\|g_k\|_*^2 \geq \lambda f(x_k), \quad \forall k \geq 0.$$

Proof. Since for every $x \in E$ and $s \in E^*$ the inequality $\frac{1}{2}\|x\|^2 + \frac{1}{2}\|s\|_*^2 \geq \langle s, x\rangle$ holds, we have

$$\langle \lambda g_k, x - z\rangle + \beta\xi(z, x) + \frac{1}{2\sigma\beta}\|\lambda g_k\|_*^2 \geq \langle \lambda g_k, x - z\rangle + \frac{\sigma\beta}{2}\|x - z\|^2 + \frac{1}{2\sigma\beta}\|\lambda g_k\|_*^2 \geq 0.$$

Substituting $z = x_k$ for this inequality and adding $\lambda f(x_k)$ to both sides, we obtain the second assertion. □

Let us consider the relation (R_k) defined at the previous section with

$$C_k = \frac{1}{2\sigma}\sum_{i=0}^{k}\frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2. \tag{18}$$

We also use the following alternative relation:

$$(\hat{R}_k) \quad \sum_{i=0}^{k}\lambda_i f(x_i) \leq \min_{x\in Q}\psi_k(x) + C_k. \tag{19}$$

Note that the relation $(\hat{R}_k)$ provides an alternative to Lemma 6 which can be proven in the same way: If $\{\psi_k(x)\}$ admits Property 2 and the relation $(\hat{R}_k)$ is satisfied for some $k \geq 0$, then we have

$$\frac{1}{S_k}\sum_{i=0}^{k}\lambda_i f(x_i) - f(x^*) \leq \frac{\beta_k l_d(z_k; x^*) + C_k}{S_k}. \tag{20}$$

Now, let us show the following key result which will provide efficient subgradient-based methods in a straightforward way.

Theorem 8. *Let $\{\psi_k(x)\}$ be a sequence of auxiliary functions satisfying Property 2 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 $z_k = \operatorname{argmin}_{x\in Q}\psi_k(x)$ and define C_k by (18). Then the following assertions hold.*

(a) The relations (R_0) and $(\hat{R}_0)$ are satisfied by setting $\hat{x}_0 := x_0$.

(b) Suppose that the relation (R_k) is satisfied for some integer $k \geq 0$. If the relation $x_{k+1} = z_k$ holds, then the relation (R_{k+1}) is satisfied by setting

$$\hat{x}_{k+1} := \frac{S_k\hat{x}_k + \lambda_{k+1}x_{k+1}}{S_{k+1}}.$$

Moreover, if the relations $(\hat{R}_k)$ is satisfied for some $k \geq 0$ and $x_{k+1} = z_k$ holds, then $(\hat{R}_{k+1})$ is satisfied.

(b') Suppose that the relation (R_k) is satisfied for some integer $k \geq 0$. 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 by setting $\hat{x}_{k+1} := x_{k+1}$.

Proof. We remark that using condition (ii) of Property 2 we obtain the inequality

$$\forall k \geq -1, \quad \min_{x \in Q} \psi_{k+1}(x) \geq \min_{x \in Q} \psi_k(x) + \lambda_{k+1} l_f(x_{k+1}, z_{k+1}) + \beta_k \xi(z_k, z_{k+1})$$

by setting $x = z_{k+1} = \mathrm{argmin}_{x \in Q} \psi_{k+1}(x)$ (recall that $d(x) \geq 0$ $(x \in Q)$ and $\beta_{k+1} \geq \beta_k$).
(a) Letting $k = -1$ in the condition (ii) and using the condition (i) of Property 2, we have

$$\begin{aligned}
\min_{x \in Q} \psi_0(x) + \frac{\lambda_0^2}{2\sigma\beta_{-1}} \|g_0\|_*^2 &\geq \left[\min_{x \in Q} \psi_{-1}(x) + \lambda_0 l_f(x_0; z_0) + \beta_{-1}\xi(z_{-1}, z_0)\right] + \frac{\lambda_0^2}{2\sigma\beta_{-1}} \|g_0\|_*^2 \\
&= \lambda_0 l_f(x_0; z_0) + \beta_{-1}\xi(z_{-1}, z_0) + \frac{\lambda_0^2}{2\sigma\beta_{-1}} \|g_0\|_*^2 \\
&= \lambda_0 l_f(x_0; z_0) + \beta_{-1}\xi(x_0, z_0) + \frac{\lambda_0^2}{2\sigma\beta_{-1}} \|g_0\|_*^2 \\
&\geq \lambda_0 f(x_0) \\
&= S_0 f(\hat{x}_0),
\end{aligned}$$

where the last inequality is due to Lemma 7.
(b) By the condition (ii) of Property 2 and the assumptions for x_{k+1} and $\hat{x}_{k+1}$, we obtain that

$$\begin{aligned}
&\min_{x \in Q} \psi_{k+1}(x) + \frac{1}{2\sigma} \sum_{i=0}^{k+1} \frac{\lambda_i^2}{\beta_{i-1}} \|g_i\|_*^2 \\
&\geq \left[\min_{x \in Q} \psi_k(x) + \lambda_{k+1} l_f(x_{k+1}; z_{k+1}) + \beta_k \xi(z_k, z_{k+1})\right] + \frac{\lambda_{k+1}^2}{2\sigma\beta_k} \|g_{k+1}\|_*^2 + \frac{1}{2\sigma} \sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_{i-1}} \|g_i\|_*^2 \\
&= \min_{x \in Q} \psi_k(x) + \frac{1}{2\sigma} \sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_{i-1}} \|g_i\|_*^2 + \left[\lambda_{k+1} l_f(x_{k+1}; z_{k+1}) + \beta_k \xi(x_{k+1}, z_{k+1}) + \frac{\lambda_{k+1}^2}{2\sigma\beta_k} \|g_{k+1}\|_*^2\right] \\
&\geq \left[\min_{x \in Q} \psi_k(x) + \frac{1}{2\sigma} \sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_{i-1}} \|g_i\|_*^2\right] + \lambda_{k+1} f(x_{k+1}) \\
&\geq S_k f(\hat{x}_k) + \lambda_{k+1} f(x_{k+1}) \\
&\geq S_{k+1} f\left(\frac{S_k \hat{x}_k + \lambda_{k+1} x_{k+1}}{S_{k+1}}\right) \\
&= S_{k+1} f(\hat{x}_{k+1}),
\end{aligned}$$

where we used Lemma 7, the relation (R_k), and the convexity of f in the last three inequalities, respectively. This implies that the relation (R_{k+1}) holds. Moreover, replacing the use of (R_k) by $(\hat{R}_k)$ in the above inequality, we obtain the relation $(\hat{R}_{k+1})$, which proves the latter assertion.
(b') Denote $x'_{k+1} = \dfrac{S_k \hat{x}_k + \lambda_{k+1} z_{k+1}}{S_{k+1}}$. Then the relation $x_{k+1} = \dfrac{S_k \hat{x}_k + \lambda_{k+1} z_k}{S_{k+1}}$ yields

$$z_{k+1} - z_k = \frac{S_{k+1}}{\lambda_{k+1}} (x'_{k+1} - x_{k+1}).$$

Thus the condition (ii) of Property 2 and the relation (R_k) imply that

$$
\begin{aligned}
&\min_{x\in Q}\psi_{k+1}(x)+\frac{1}{2\sigma}\sum_{i=0}^{k+1}\frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2\\
&\quad\geq \min_{x\in Q}\psi_k(x)+\lambda_{k+1}l_f(x_{k+1};z_{k+1})+\beta_k\xi(z_k,z_{k+1})+\frac{\lambda_{k+1}^2}{2\sigma\beta_k}\|g_{k+1}\|_*^2+\frac{1}{2\sigma}\sum_{i=0}^{k}\frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2\\
&\quad\geq S_kf(\hat{x}_k)+\lambda_{k+1}l_f(x_{k+1};z_{k+1})+\beta_k\xi(z_k,z_{k+1})+\frac{\lambda_{k+1}^2}{2\sigma\beta_k}\|g_{k+1}\|_*^2\\
&\quad\geq S_kl_f(x_{k+1};\hat{x}_k)+\lambda_{k+1}l_f(x_{k+1};z_{k+1})+\beta_k\xi(z_k,z_{k+1})+\frac{\lambda_{k+1}^2}{2\sigma\beta_k}\|g_{k+1}\|_*^2\\
&\quad= S_{k+1}l_f\left(x_{k+1};\frac{S_k\hat{x}_k+\lambda_{k+1}z_{k+1}}{S_{k+1}}\right)+\beta_k\xi(z_k,z_{k+1})+\frac{\lambda_{k+1}^2}{2\sigma\beta_k}\|g_{k+1}\|_*^2\\
&\quad= S_{k+1}l_f(x_{k+1};x'_{k+1})+\beta_k\xi(z_k,z_{k+1})+\frac{\lambda_{k+1}^2}{2\sigma\beta_k}\|g_{k+1}\|_*^2\\
&\quad= S_{k+1}f(x_{k+1})+\left\langle g_{k+1},S_{k+1}(x'_{k+1}-x_{k+1})\right\rangle\\
&\qquad\qquad +\beta_k\xi(z_k,z_{k+1})+\frac{\lambda_{k+1}^2}{2\sigma\beta_k}\|g_{k+1}\|_*^2\\
&\quad= S_{k+1}f(x_{k+1})+\langle\lambda_{k+1}g_{k+1},z_{k+1}-z_k\rangle+\beta_k\xi(z_k,z_{k+1})+\frac{\lambda_{k+1}^2}{2\sigma\beta_k}\|g_{k+1}\|_*^2\\
&\quad\geq S_{k+1}f(x_{k+1})\\
&\quad= S_{k+1}f(\hat{x}_{k+1})
\end{aligned}
$$

where the last inequality is due to Lemma 7. □

Now we are ready to propose the following two novel subgradient-based methods for the non-smooth case.

Method 9 (Unifying framework)**.** *Choose weight parameters* $\{\lambda_k\}_{k\geq 0}$ *and scaling parameters* $\{\beta_k\}_{k\geq -1}$*. Generate sequences* $\{(z_{k-1},x_k,g_k,\hat{x}_k)\}_{k\geq 0}$ *by*

$$
(a)\qquad x_k:=z_{k-1}:=\operatorname*{argmin}_{x\in Q}\psi_{k-1}(x),\quad \hat{x}_k:=\frac{1}{S_k}\sum_{i=0}^{k}\lambda_ix_i,\quad g_k\in\partial f(x_k),\quad \text{for } k\geq 0 \tag{21}
$$

or by

$$
(b)\qquad z_{k-1}:=\operatorname*{argmin}_{x\in Q}\psi_{k-1}(x),\quad \hat{x}_k:=x_k:=\frac{1}{S_k}\sum_{i=0}^{k}\lambda_iz_{i-1},\quad g_k\in\partial f(x_k),\quad \text{for } k\geq 0 \tag{22}
$$

where $\{\psi_k(x)\}_{k\geq -1}$ *is defined using the construction (10) as well as any construction which admits Property 2.*

Notice that the sequences $\{z_k\}_{k\geq -1}$ and $\{x_k\}_{k\geq 0}$ are dummy ones for the methods (a) and (b), respectively, but we kept them to preserve the notation.

4.2 Convergence analysis of the unifying framework method

Corollary 10. *Given the weight parameter* $\{\lambda_k\}_{k\geq 0}$*, the scaling parameter* $\{\beta_k\}_{k\geq -1}$*, and any sequence* $\{(z_{k-1},x_k,g_k,\hat{x}_k)\}_{k\geq 0}$ *generated by*

(a) the first procedure (21) in Method 9, we have:

$$f(\hat{x}_k) - f(x^*) \leq \frac{1}{S_k}\sum_{i=0}^{k}\lambda_i f(x_i) - f(x^*) \leq \frac{\beta_k l_d(z_k;x^*) + \frac{1}{2\sigma}\sum_{i=0}^{k}\frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2}{S_k} \tag{23}$$

for all $k \geq 0$; or

(b) the second procedure (22) in Method 9, we have:

$$f(\hat{x}_k) - f(x^*) \leq \frac{\beta_k l_d(z_k;x^*) + \frac{1}{2\sigma}\sum_{i=0}^{k}\frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2}{S_k} \tag{24}$$

for all $k \geq 0$.

Proof. The first inequality in (23) is from the convexity of $f(x)$. Proposition 4 and Theorem 8 show that the sequences generated by the procedures (21) and (22) satisfy the relation (R_k); futhermore, the former construction (21) also satisfies $(\hat{R}_k)$. Thus, Lemma 6 and the alternative (20) of Lemma 6 for $(\hat{R}_k)$ prove the assertion. □

In [14], Nesterov proposed to use of the auxiliary sequence (5) to ensure an efficient convergence of the DAM (4). This sequence also satisfies the identity

$$\hat{\beta}_k = \sum_{i=-1}^{k-1}\frac{1}{\hat{\beta}_i} \quad (k \geq 0) \tag{25}$$

and the inequality

$$\forall k \geq 0, \quad \sqrt{2k+1} \leq \hat{\beta}_k \leq \frac{1}{1+\sqrt{3}} + \sqrt{2k+1}. \tag{26}$$

Corollary 11 (see also [14])**.** *Consider the following two choices for the parameters.*

(Simple Averages) *Let $\{(z_{k-1}, x_k, g_k, \hat{x}_k)\}_{k\geq 0}$ be generated by Method 9 with parameters $\lambda_k := 1$ and $\beta_k := \gamma\hat{\beta}_k$ for some $\gamma > 0$. Then we have*

$$\forall k \geq 0, \quad f(\hat{x}_k) - f(x^*) \leq \left(\gamma l_d(z_k;x^*) + \frac{M_k^2}{2\sigma\gamma}\right)\frac{0.5+\sqrt{2k+1}}{k+1} \tag{27}$$

and

$$\forall k \geq -1, \quad z_k, x_{k+1}, \hat{x}_{k+1} \in \left\{x \in Q \ : \ \|x - x^*\|^2 \leq \frac{2d(x^*)}{\sigma} + \frac{M_k^2}{\sigma^2\gamma^2}\right\} \tag{28}$$

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

(Weighted Averages) *Let $\{(z_{k-1}, x_k, g_k, \hat{x}_k)\}_{k\geq 0}$ be generated by Method 9 with parameters $\lambda_k := \frac{1}{\|g_k\|_*}$ and $\beta_k := \frac{\hat{\beta}_k}{\rho\sqrt{\sigma}}$ for some $\rho > 0$. Then we have*

$$\forall k \geq 0, \quad f(\hat{x}_k) - f(x^*) \leq M_k\frac{1}{\sqrt{\sigma}}\left(\frac{l_d(z_k;x^*)}{\rho} + \frac{\rho}{2}\right)\frac{0.5+\sqrt{2k+1}}{k+1} \tag{29}$$

and

$$\forall k \geq -1, \quad z_k, x_{k+1}, \hat{x}_{k+1} \in \left\{x \in Q \ : \ \|x - x^*\|^2 \leq \frac{2d(x^*) + \rho^2}{\sigma}\right\}. \tag{30}$$

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 first procedure (21) in Method 9. In this case, the left hand side of the inequality can be replaced by $\min\{f(\hat{x}_k)-f(x^*), \min_{0\le i\le k} f(x_i)-f(x^*)\}$.

Proof. Substituting the specified λ_k and β_k into the estimations in Corollary 10 and using the properties (25) and (26) of $\hat{\beta}_k$, we obtain (27) and (29), respectively. Denote by B_k the ball on the right hand side of (28) for $k \ge -1$. Then $B_k \subset B_{k+1}$ for each $k \ge -1$. The inequality (27) implies that $\gamma l_d(z_k;x^*) + (2\sigma\gamma)^{-1}M_k^2 \ge 0$ for all $k \ge 0$. Using the strong convexity, $d(x^*) \ge l_d(z_k;x^*) + \frac{\sigma}{2}\|x^*-z_k\|^2$, and therefore, shows that $z_k \in B_k$ for each $k \ge 0$. 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 $d(x^*) \ge l_d(z_{-1};x^*) + \frac{\sigma}{2}\|z_{-1}-x^*\|^2 \ge \frac{\sigma}{2}\|z_{-1}-x^*\|^2$. Finally, we conclude that $x_{k+1}, \hat{x}_{k+1} \in B_k$ for all $k \ge -1$ because they are convex combinations of $\{z_i\}_{i=-1}^k$. The proof of (30) is similar. □

Remark 12. Notice that in our approach, the bounds in (27) and (29) are slightly smaller than the ones in (3.3) and (3.5) in [14], respectively, since $l_d(z_k;x^*) \le d(x^*) \le D$. However, essentially, Nesterov's original argument also arrives to the same one when $d(x)$ is continuously differentiable on Q (note that the argument in [14] does not impose the differentiability for $d(x)$). In fact, in [14], Theorems 2 and 3 rely on the estimate (2.15) which is implied from (2.18). Notice in (2.18) that we have

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

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), it yields

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

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

A consequence of Corollary 11 is that if $M := \sup\{\|g\|_* : g \in \partial f(x),\ x \in Q\}$ is finite, Method 9 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. Therefore, the estimates (27) and (29) achieve the optimal complexity for the non-smooth case when we choose $\gamma := M/\sqrt{2\sigma d(x^*)}$ and $\rho := \sqrt{2d(x^*)}$, respectively. Also Method 9 with the parameters suggested in Corollary 11 produces bounded sequences $\{x_k\}$, $\{\hat{x}_k\}$, and $\{z_k\}$ (even if $M = +\infty$ for the Weighted Averages case).

4.3 Particular cases: The extended MD and the DA models

Restricting to the extended MD model (11) in Method 9, the first procedure (21) provides the following extension of the MDM.

Method 13 (Extended Mirror-Descent)**.** *Set $x_0 := \operatorname{argmin}_{x\in Q} d(x)$. Choose weight parameters $\{\lambda_k\}_{k\ge 0}$ and scaling parameters $\{\beta_k\}_{k\ge -1}$. Generate sequences $\{(x_k, g_k, \hat{x}_k)\}_{k\ge 0}$ by*

$$\begin{aligned} g_k &\in \partial f(x_k), \\ x_{k+1} &:= \operatorname*{argmin}_{x\in Q}\left\{\lambda_k[f(x_k)+\langle g_k, x-x_k\rangle] + \beta_k d(x) - \beta_{k-1}l_d(x_k;x)\right\}, \\ \hat{x}_k &:= \frac{1}{S_k}\sum_{i=0}^k \lambda_i x_i \end{aligned}$$

[2]Notice that x_{k+1} and β_{k+1} in [14] are called z_k and β_k here, respectively.

for $k \geq 0$.

The iteration updates described by (2) of the original MDM corresponds to Method 13 with $\beta_k := 1$. Corollary 11 shows, in particular, that this extended MDM has a better complexity bound for the objective function compared to the original MDM described in Section 2.

On the other hand, restricting to the DA model (12) in Method 9, the first procedure (21) yields the Nesterov's DAM (4) described in Section 2. In particular, Corollary 10 and subsequently Corollary 11 provide a small improvement over the original result assuming the differentiability of $d(x)$ as pointed out before. Since our analysis does not introduce the dual space, the argments are more straighforward than the original one.

We can also obtain variants of the extended MDM and the DAM from the second procedure (22). An upper bound of $f(\hat{x}_k) - f(x^*)$ for the sequence $\{\hat{x}_k\}$ generated by these methods can be derived from Corollaries 10 or 11.

5 A family of (inexact) gradient-based methods for structured problems in the unifying framework

The framework discussed in Section 3 can be also applied to develop efficient (inexact) gradient-based methods for structured convex problems.

In this section, we assume that the objective function $f(x)$ of the problem (1) has the following structure; for any $y \in Q$, there exists a lower approximation $l_f(y;x)$ of $f(x)$, which is convex in x, and satisfies the inequalities

$$l_f(y;x) \leq f(x) \leq l_f(y;x) + \frac{L(y)}{2}\|x-y\|^2 + \delta(y), \quad \forall x \in Q, \tag{31}$$

for some $L(y) > 0$ and $\delta(y) \geq 0$. We also assume that for any $y \in Q$, $s \in E^*$, and $\beta > 0$, we can compute the optimal solution of the (sub)problem

$$\min_{x \in Q}\{l_f(y;x) + \langle s, x\rangle + \beta d(x)\}. \tag{32}$$

Let us see some examples which admit these assumptions.

Example 14. The first four cases were already considered in the literature.

(i) *Smooth case.* If the convex objective function $f(x)$ is continuously differentiable on Q and its gradient $\nabla f(x)$ is Lipschitz continuous on Q with a constant $L > 0$, defining $l_f(y;x) := f(y) + \langle \nabla f(y), x - y\rangle$ yields the condition (31) with $L(\cdot) \equiv L$ and $\delta(\cdot) \equiv 0$. Then subproblem (32) is of the form

$$\min_{x \in Q}\{f(y) + \langle s + \nabla f(y), x - y\rangle + \beta d(x)\}. \tag{33}$$

(ii) *Composite structure.* Let the objective function $f(x)$ has the form

$$f(x) = f_0(x) + \Psi(x) \tag{34}$$

where $f_0(x) : E \to \mathbb{R} \cup \{+\infty\}$ is convex and continuously differentiable on Q with Lipschitz continuous gradient and $\Psi(x) : E \to \mathbb{R} \cup \{+\infty\}$ is a closed convex function with $Q \subset \operatorname{dom} \Psi$. Letting $L > 0$ be the Lipschitz constant of ∇f_0 on Q, we can define $l_f(y;x) := f_0(y) + \langle \nabla f_0(y), x - y\rangle + \Psi(x)$ so that we have (31) with $L(\cdot) \equiv L$ and $\delta(\cdot) \equiv 0$. The corresponding (sub)problem has the form

$$\min_{x \in Q}\{f_0(y) + \langle s + \nabla f_0(y), x - y\rangle + \beta d(x) + \Psi(x)\}.$$

A generalization of classical methods such as proximal gradient method for this model was proposed by Fukushima and Mine [7] (without assuming convexity for $f_0(x)$). Nesterov's optimal method (7) can be also generalized to this case [13]. Smoothing techniques are also an important approach for this example. Nesterov [12] showed a significant improvement on the convergence rate for a particular class and Beck and Teboulle [4] proposed an unifying generalization.

(iii) *Inexact oracle model.* Let us assume that our oracle for $f(x)$ has *inexactness* [6], that is, we can compute $(\bar{f}(y), \bar{g}(y)) \in \mathbb{R} \times E^*$ at each $y \in Q$ such that

$$0 \leq f(x) - (\bar{f}(y) + \langle \bar{g}(y), x - y \rangle) \leq \frac{L_y}{2}\|x - y\|^2 + \delta_y, \quad \forall x \in Q \tag{35}$$

is satisfied for some $L_y > 0$ and $\delta_y \geq 0$. Then defining $l_f(y; x) := \bar{f}(y) + \langle \bar{g}(y), x - y \rangle, L(y) := L_y$, and $\delta(y) := \delta_y$ we have exactly (31). This model was investigated in [6] and the *primal, dual,* and *fast gradient method* was proposed. These methods were also implemented in [15] for a particular class of this model, equipped by an iterative scheme to estimate the Lipschitz constants L_y at each iteration. The fast gradient methods can be seen as generalizations of Nesterov's optimal method (7) to those cases.

(iv) *Saddle structure.* Let us consider an objective function with the following structure:

$$f(x) = \sup_{u \in U} \phi(u, x)$$

where U is a compact convex set of a finite dimensional real vector space E' and $\phi : U \times E \to \mathbb{R} \cup \{+\infty\}$ is a concave-convex function satisfying the following conditions.

- $\phi(\cdot, x)$ is a closed concave function for all $x \in Q$.
- $\phi(u, \cdot)$ is a closed convex function with $Q \subset \operatorname{dom} \phi(u, \cdot)$ for all $u \in U$.
- For all $u \in U$, $\phi(u, \cdot)$ is a continuously differentiable on Q and its gradient is Lipschitz continuous on Q, *i.e.*, there exists a constant $L_u \geq 0$ such that

$$\|\nabla_x \phi(u, x_1) - \nabla_x \phi(u, x_2)\|_* \leq L_u \|x_1 - x_2\|, \quad \forall x_1, x_2 \in Q.$$

- $L := \max_{u \in U} L_u$ is finite and positive.

Then defining

$$l_f(y; x) := \max_{u \in U} \{\phi(u, y) + \langle \nabla_x \phi(u, y), x - y \rangle\}, \tag{36}$$

it satisfies condition (31) with $L(\cdot) \equiv L$, $\delta(\cdot) \equiv 0$, and we will have the following subproblem:

$$\min_{x \in Q} \left\{ \max_{u \in U} \{\phi(u, y) + \langle s + \nabla_x \phi(u, y), x - y \rangle\} + \beta d(x) \right\}.$$

This case is a generalization of the structured convex problem discussed in [9], namely, $E' \equiv \mathbb{R}^m$ and, for each $u = (u^{(1)}, \ldots, u^{(m)}) \in U$, defining $\phi(u, x) = \sum_{i=0}^m u^{(i)} f_i(x)$ for given differentiable convex functions $f_1(x), \ldots, f_m(x)$ on E with Lipschitz continuous gradient. The convexity of $\phi(u, \cdot)$ is satisfied by imposing the following assumption as in [9]: If there exists $u \in U$ such that $u^{(i)} < 0$, then $f_i(x)$ is a linear function. Letting $L^{(i)}$ be a Lipchitz constant of $\nabla f_i(x)$ for $i = 1, \ldots, m$, we have $L = \max_{u \in U} L_u = \max_{u \in U} \sum_{i=1}^m u^{(i)} L^{(i)}$.

The definition of $l_f(y;x)$ can be simplified when $Q \subset \mathrm{int}(\mathrm{dom}\, f)$ and $\phi(\cdot, x)$ is strictly concave for all $x \in Q$. In this case, denoting $u_x = \mathrm{argmax}_{u\in U}\, \phi(u,x)$, we have $\nabla f(x) = \nabla_x \phi(u_x, x)$ and therefore we can define

$$l_f(y;x) := \phi(u_y, y) + \langle \nabla_x \phi(u_y, y), x - y\rangle$$

which satisfies (31) with $L(\cdot) \equiv L$ and $\delta(\cdot) \equiv 0$. Its subproblem is of the form (33). This situation is also discussed in Tseng's methods [16].

(v) *Mixed structure.* The above examples can be combined with each other; for instance, considering the function $f_0(x)$ in (ii) with inexactness (iii) or with the saddle structure (iv), or considering the function $\phi(u,x)$ in (iv) with inexactness (iii) or with the composite structure (ii) satisfies our requirement (31).

5.1 The unifying framework

We will propose (inexact) gradient-based methods for structured convex optimization problems which satisfies (31) and admits computable solutions for (32) highlighted by Example 14. These methods generate approximate solutions $\{\hat{x}_k\} \subset Q$ satisfying the relation (R_k). We also consider, in this section, the following alternative of this relation (R_k) for some constant C_k:

$$(\hat{R}'_k) \quad \sum_{i=0}^{k} \lambda_i f(x_{i+1}) \le \min_{x\in Q} \psi_k(x) + C_k. \tag{37}$$

Notice that the relation $(\hat{R}_k)$ is sightly different from that of the non-smooth case (19). We use the following alternative of Lemma 6 for this relation; if $\{\psi_k(x)\}$ satisfies Property 2 and the relation $(\hat{R}'_k)$ is satisfied for some $k \ge 0$, then we have

$$\frac{1}{S_k}\sum_{i=0}^{k} \lambda_i f(x_{i+1}) - f(x^*) \le \frac{\beta_k l_d(z_k, x^*) + C_k}{S_k}. \tag{38}$$

The following theorem validates our methods.

Theorem 15. *Let $\{\psi_k(x)\}_{k\ge -1}$ be a sequence of auxiliary functions satisfying Property 2 associated with weight parameters $\{\lambda_k\}_{k\ge 0}$, scaling parameters $\{\beta_k\}_{k\ge -1}$, and test points $\{x_k\}_{k\ge 0}$. Denote $z_k = \mathrm{argmin}_{x\in Q}\, \psi_k(x)$. Then the following assertions hold.*

(a) If $\sigma\beta_{-1}/\lambda_0 \ge L(x_0)$ holds, then relation (R_0) is satisfied with $\hat{x}_0 := z_0$ and $C_0 := \lambda_0\delta(x_0)$.

(b) Suppose that the relation (R_k) is satisfied for some integer $k \ge 0$. If the relations $x_{k+1} = z_k$ and $\sigma\beta_k/\lambda_{k+1} \ge L(x_{k+1})$ hold, then the relation (R_{k+1}) is satisfied with

$$\hat{x}_{k+1} := \frac{S_k\hat{x}_k + \lambda_{k+1}z_{k+1}}{S_{k+1}}, \quad C_{k+1} := C_k + \lambda_{k+1}\delta(x_{k+1}). \tag{39}$$

Moreover, if the relations $(\hat{R}'_k)$ is satisfied for some integer $k \ge 0$, and the relations $\sigma\beta_k/\lambda_{k+1} \ge L(x_{k+1})$ and $x_{k+1} = z_k$ hold, then the relation $(\hat{R}'_{k+1})$ is satisfied with $C_{k+1} := C_k + \lambda_{k+1}\delta(x_{k+1})$.

(b') Suppose that the relation (R_k) is satisfied for some integer $k \ge 0$. If the relations

$$x_{k+1} = \frac{S_k\hat{x}_k + \lambda_{k+1}z_k}{S_{k+1}} \quad \text{and} \quad \sigma\beta_k S_{k+1}/\lambda_{k+1}^2 \ge L(x_{k+1})$$

hold, then the relation (R_{k+1}) *is satisfied with*

$$\hat{x}_{k+1} := \frac{S_k \hat{x}_k + \lambda_{k+1} z_{k+1}}{S_{k+1}}, \quad C_{k+1} := C_k + S_{k+1}\delta(x_{k+1}).$$

Proof. Denote $L_k = L(x_k)$ and $\delta_k = \delta(x_k)$.
(a) Condition (ii) with $k = -1$ and condition (i) of Property 2 yields that

$$\begin{aligned}
\min_{x\in Q} \psi_0(x) + \lambda_0\delta_0 &\geq \min_{x\in Q} \psi_{-1}(x) + \lambda_0 l_f(x_0; z_0) + \beta_{-1}\xi(z_{-1}, z_0) + \lambda_0\delta_0 \\
&= \lambda_0 \left(l_f(x_0; z_0) + \frac{\beta_{-1}}{\lambda_0}\xi(x_0, z_0) + \delta_0 \right) \\
&\geq \lambda_0 \left(l_f(x_0; z_0) + \frac{\sigma\beta_{-1}}{\lambda_0}\frac{1}{2}\|z_0 - x_0\|^2 + \delta_0 \right) \\
&\geq \lambda_0 \left(l_f(x_0; z_0) + \frac{L_0}{2}\|z_0 - x_0\|^2 + \delta_0 \right) \\
&\geq \lambda_0 f(z_0) = S_0 f(\hat{x}_0)
\end{aligned}$$

where the last inequality is due to (31).
(b) Condition (ii) of Property 2 implies that

$$\begin{aligned}
\min_{x\in Q} \psi_{k+1}(x) + C_{k+1} &\geq \min_{x\in Q} \psi_k(x) + C_k + \lambda_{k+1} l_f(x_{k+1}; z_{k+1}) + \beta_k\xi(z_k, z_{k+1}) + \lambda_{k+1}\delta_{k+1} \\
&= \min_{x\in Q} \psi_k(x) + C_k + \lambda_{k+1} l_f(x_{k+1}; z_{k+1}) + \beta_k\xi(x_{k+1}, z_{k+1}) + \lambda_{k+1}\delta_{k+1} \\
&\geq \min_{x\in Q} \psi_k(x) + C_k + \lambda_{k+1} \left(l_f(x_{k+1}; z_{k+1}) + \frac{\sigma\beta_k}{2\lambda_{k+1}}\|z_{k+1} - x_{k+1}\|^2 + \delta_{k+1} \right) \\
&\geq \min_{x\in Q} \psi_k(x) + C_k + \lambda_{k+1} \left(l_f(x_{k+1}; z_{k+1}) + \frac{L_{k+1}}{2}\|z_{k+1} - x_{k+1}\|^2 + \delta_{k+1} \right) \\
&\geq \min_{x\in Q} \psi_k(x) + C_k + \lambda_{k+1} f(z_{k+1}) \qquad (40) \\
&\geq S_k f(\hat{x}_k) + \lambda_{k+1} f(z_{k+1}) \qquad (41) \\
&\geq S_{k+1} f\left(\frac{S_k\hat{x}_k + \lambda_{k+1} z_{k+1}}{S_{k+1}} \right) = S_{k+1} f(\hat{x}_{k+1})
\end{aligned}$$

where the inequalities (40) and (41) are due to (31) and (R_k), respectively. When we use $(\hat{R}'_k)$ at (41), it yields the relation $(\hat{R}'_{k+1})$.
(b') The assumptions for x_{k+1} and $\hat{x}_{k+1}$ implies $z_{k+1} - z_k = \frac{S_{k+1}}{\lambda_{k+1}}(\hat{x}_{k+1} - x_{k+1})$. Thus, from

condition (ii) of Property 2 and relation (R_k), we obtain

$$\begin{aligned}
\min_{x\in Q}\psi_{k+1}(x)+C_{k+1} &\geq \min_{x\in Q}\psi_k(x)+C_k+\lambda_{k+1}l_f(x_{k+1};z_{k+1})+\beta_k\xi(z_k,z_{k+1})+S_{k+1}\delta_{k+1}\\
&\geq S_kf(\hat{x}_k)+\lambda_{k+1}l_f(x_{k+1};z_{k+1})+\beta_k\xi(z_k,z_{k+1})+S_{k+1}\delta_{k+1}\\
&\geq S_kl_f(x_{k+1};\hat{x}_k)+\lambda_{k+1}l_f(x_{k+1};z_{k+1})+\beta_k\xi(z_k,z_{k+1})+S_{k+1}\delta_{k+1}\\
&\geq S_{k+1}l_f\left(x_{k+1};\frac{S_k\hat{x}_k+\lambda_{k+1}z_{k+1}}{S_{k+1}}\right)+\beta_k\xi(z_k,z_{k+1})+S_{k+1}\delta_{k+1}\\
&\geq S_{k+1}l_f(x_{k+1};\hat{x}_{k+1})+\frac{\sigma\beta_k}{2}\|z_{k+1}-z_k\|^2+S_{k+1}\delta_{k+1}\\
&= S_{k+1}\left(l_f(x_{k+1};\hat{x}_{k+1})+\frac{\sigma\beta_kS_{k+1}}{2\lambda_{k+1}^2}\|\hat{x}_{k+1}-x_{k+1}\|^2+\delta_{k+1}\right)\\
&\geq S_{k+1}\left(l_f(x_{k+1};\hat{x}_{k+1})+\frac{L_{k+1}}{2}\|\hat{x}_{k+1}-x_{k+1}\|^2+\delta_{k+1}\right)\\
&\geq S_{k+1}f(\hat{x}_{k+1}).
\end{aligned}$$

□

Now we are ready to propose the following two unifying framework methods.

Method 16 (Classical Gradient Method (CGM))**.** *Choose weight parameters $\{\lambda_k\}_{k\geq 0}$ and scaling parameters $\{\beta_k\}_{k\geq -1}$. Generate sequences $\{(z_{k-1},x_k,\hat{x}_k)\}_{k\geq 0}$ by setting*

$$x_k:=z_{k-1}:=\operatorname*{argmin}_{x\in Q}\psi_{k-1}(x),\quad \hat{x}_k:=\frac{1}{S_k}\sum_{i=0}^{k}\lambda_ix_{i+1},$$

for $k\geq 0$, where $\{\psi_k(x)\}_{k\geq -1}$ is defined using the construction (10) as well as any construction which admits Property 2.

Method 17 (Fast Gradient Method (FGM))**.** *Choose weight parameters $\{\lambda_k\}_{k\geq 0}$ and scaling parameters $\{\beta_k\}_{k\geq -1}$. Set $x_0:=z_{-1}:=\operatorname{argmin}_{x\in Q}d(x)$ and $\hat{x}_0:=z_0:=\operatorname{argmin}_{x\in Q}\psi_0(x)$. Generate sequences $\{(z_{k-1},x_k,\hat{x}_k)\}_{k\geq 0}$ by setting*

$$\begin{aligned}
x_{k+1} &:= \frac{\sum_{i=0}^{k}\lambda_iz_i+\lambda_{k+1}z_k}{S_{k+1}},\\
z_{k+1} &:= \operatorname*{argmin}_{x\in Q}\psi_{k+1}(x),\\
\hat{x}_{k+1} &:= \frac{1}{S_{k+1}}\sum_{i=0}^{k+1}\lambda_iz_i,
\end{aligned}$$

for $k\geq 0$ where $\{\psi_k(x)\}_{k\geq -1}$ is defined using the construction (10) as well as any construction which admits Property 2.

Notice that in this case, only the sequence $\{z_k\}_{k\geq -1}$ is a dummy one for the CGM.

5.2 Convergence analysis of unifying framework methods

By the same observation as Corollary 10, combining Theorem 15 and Lemma 6 (or the alternative (38) of Lemma 6), we arrive at the following estimates.

Corollary 18. *(a) Let $\{(z_{k-1}, x_k, \hat{x}_k)\}_{k\geq 0}$ be generated by the CGM associated with weight parameters $\{\lambda_k\}_{k\geq 0}$ and scaling parameters $\{\beta_k\}_{k\geq -1}$. If $\sigma\beta_{k-1}/\lambda_k \geq L(x_k)$ holds for all $k \geq 0$, then we have*

$$\forall k \geq 0, \quad f(\hat{x}_k) - f(x^*) \leq \frac{1}{S_k}\sum_{i=0}^{k} \lambda_i f(x_{i+1}) - f(x^*) \leq \frac{\beta_k l_d(z_k; x^*) + \sum_{i=0}^{k} \lambda_i \delta(x_i)}{S_k}.$$

(b) Let $\{(z_{k-1}, x_k, \hat{x}_k)\}_{k\geq 0}$ be generated by the FGM associated with weight parameters $\{\lambda_k\}_{k\geq 0}$ and scaling parameters $\{\beta_k\}_{k\geq -1}$. If $\sigma\beta_{k-1}S_k/\lambda_k^2 \geq L(x_k)$ holds for all $k \geq 0$, then we have

$$\forall k \geq 0, \quad f(\hat{x}_k) - f(x^*) \leq \frac{\beta_k l_d(z_k; x^*) + \sum_{i=0}^{k} S_i \delta(x_i)}{S_k}.$$

Particular choices for the parameters λ_k and β_k in the above estimates simplify the situation.

Corollary 19. *Suppose that $\delta(\cdot) \equiv \delta$ and $L(\cdot) \equiv L$ are constants.*

(a) Any sequence $\{(z_{k-1}, x_k, \hat{x}_k)\}_{k\geq 0}$ generated by the CGM with $\lambda_k := 1$ and $\beta_k := L/\sigma$ satisfies

$$\forall k \geq 0, \quad f(\hat{x}_k) - f(x^*) \leq \frac{1}{k+1}\sum_{i=0}^{k} f(x_{i+1}) - f(x^*) \leq \frac{L l_d(z_k; x^*)}{\sigma(k+1)} + \delta \tag{42}$$

and

$$\forall k \geq -1, \quad z_k, x_{k+1}, \hat{x}_k \in \left\{ x \in Q \ : \ \|x - x^*\|^2 \leq \frac{2d(x^*)}{\sigma} + \frac{2\delta}{L}(k+1) \right\}. \tag{43}$$

(b) Any sequence $\{(z_{k-1}, x_k, \hat{x}_k)\}_{k\geq 0}$ generated by the FGM with $\lambda_k := \frac{k+1}{2}$ and $\beta_k := L/\sigma$ satisfies

$$\forall k \geq 0, \quad f(\hat{x}_k) - f(x^*) \leq \frac{4L l_d(z_k; x^*)}{\sigma(k+1)(k+2)} + \frac{k+3}{3}\delta \tag{44}$$

and

$$\forall k \geq -1, \quad z_k, x_{k+1}, \hat{x}_k \in \left\{ x \in Q \ : \ \|x - x^*\|^2 \leq \frac{2d(x^*)}{\sigma} + \frac{\delta}{6L}(k+3)(k+1)^2 \right\}. \tag{45}$$

Proof. The estimations (42) and (44) can be obtained by substituting the specified parameters to Corollary 18. By a similar argument as the proof of Corollary 11, remarking $x_k \in \mathrm{conv}\{z_i\}_{i=-1}^{k-1}$ and $\hat{x}_k \in \mathrm{conv}\{z_i\}_{i=0}^{k}$, we have (43) and (45). □

Let us consider the case $\delta = 0$ in Corollary 19. This includes the case of minimization of a convex function with a Lipschitz continuous gradient. Then the FGM ensures the optimal convergence rate $f(\hat{x}_k) - f(x^*) \leq O\left(\frac{LR^2}{k^2}\right)$ where $R = \sqrt{\frac{1}{\sigma}d(x^*)}$ which is faster than the rate $O\left(\frac{LR^2}{k}\right)$ guaranteed by the CGM. Corollary 19 also ensures that the generated sequences $\{z_k\}$, $\{x_k\}$, and $\{\hat{x}_k\}$ are bounded when $\delta = 0$.

In the case $\delta > 0$, a comparison between the CGM and the FGM is not obvious; an immediate fact is that the upper bound in (44) diverges while the one in (42) converges to δ. There is a detailed discussion about different situations in [6, Section 6].

5.3 Particular cases: The extended MD and the DA models

The unifying framework methods, the CGM and the FGM, yield some existing methods by adopting particular choices for the auxiliary functions $\{\psi_k(x)\}$.

When we apply the extended MD model (11) and the DA model (12) to the CGM, it yields the iteration updates

$$x_{k+1} := \operatorname*{argmin}_{x\in Q} \left\{\lambda_k l_f(x_k; x) + \beta_k d(x) - \beta_{k-1} l_d(x_k; x)\right\} \tag{46}$$

and

$$x_{k+1} := \operatorname*{argmin}_{x\in Q} \left\{\sum_{i=0}^{k} \lambda_i l_f(x_i; x) + \beta_k d(x)\right\}, \tag{47}$$

respectively.

In the composite structure (34), these updates with the choice of λ_k's and β_k's as in Corollary 19 (a) yield the *primal* and *dual gradient methods* analyzed by Nesterov [13] with known Lipschitz constants. In the Euclidean setting (*i.e.*, E is a Euclidean space, the norm $\|\cdot\|$ is induced by its inner product, and $d(x) = \frac{1}{2}\|x - x_0\|^2$), the extended MD update (46) is also closely related to the proximal point method proposed by Fukushima and Mine [7]. In fact, assuming the same conditions in [7, Corollary at p.996], this method is equivalent to the CGM with $\lambda_k := 1/c_k$ and $\beta_k := 1$.

The above updates in the Euclidean setting also correspond to the *primal* and *dual gradient methods* [6] for the inexact oracle model (35) by choosing $\lambda_k := 1/L(x_k)$, and $\beta_k := 1/\sigma$. Since $l_d(z_k; x^*) \le d(x^*)$, Corollary 18 for this case provide estimates for the optimal values with smaller upper bounds than those of [6, Section 4]; for the dual gradient method, in particular, our estimate does not require the computation of the solution (y_k in [6, Theorem 3]) of another auxiliary subproblem.

The FGM, on the other hand, provides accelerated versions of the above ones derived from the CGM. Using the extended MD model (11) for the FGM, it yields the following algorithm.

Method 20. *Choose weight parameters* $\{\lambda_k\}_{k\ge 0}$ *and scaling parameters* $\{\beta_k\}_{k\ge -1}$. *Set* $x_0 := \operatorname{argmin}_{x\in Q} d(x)$ *and* $\hat{x}_0 := z_0 := \operatorname{argmin}_{x\in Q} \left\{\lambda_0 l_f(x_0; x) + \beta_0 d(x) - \beta_{-1} l_d(x_0; x)\right\}$. *Generate sequences* $\{(z_k, x_k, \hat{x}_k)\}_{k\ge 0}$ *by setting*

$$\begin{aligned} x_{k+1} &:= \frac{\sum_{i=0}^{k} \lambda_i z_i + \lambda_{k+1} z_k}{S_{k+1}}, \\ z_{k+1} &:= \operatorname{argmin}_{x\in Q} \left\{\lambda_{k+1} l_f(x_{k+1}; x) + \beta_{k+1} d(x) - \beta_k l_d(z_k; x)\right\}, \\ \hat{x}_{k+1} &:= \frac{1}{S_{k+1}} \sum_{i=0}^{k+1} \lambda_i z_i. \end{aligned}$$

for $k \ge 0$.

The DA model (12), on the other hand, yields the following algorithm.

Method 21. *Choose weight parameters* $\{\lambda_k\}_{k\ge 0}$ *and scaling parameters* $\{\beta_k\}_{k\ge -1}$. *Set* $x_0 := \operatorname{argmin}_{x\in Q} d(x)$ *and* $\hat{x}_0 := z_0 := \operatorname{argmin}_{x\in Q} \left\{\lambda_0 l_f(x_0; x) + \beta_0 d(x)\right\}$. *Generate sequences* $\{(z_k, x_k, \hat{x}_k)\}_{k\ge 0}$

by setting

$$\begin{aligned} x_{k+1} &:= \frac{\sum_{i=0}^{k} \lambda_i z_i + \lambda_{k+1} z_k}{S_{k+1}}, \\ z_{k+1} &:= \operatorname{argmin}_{x \in Q} \left\{ \sum_{i=0}^{k+1} \lambda_i l_f(x_i; x) + \beta_{k+1} d(x) \right\}, \\ \hat{x}_{k+1} &:= \frac{1}{S_{k+1}} \sum_{i=0}^{k+1} \lambda_i z_i. \end{aligned}$$

for $k \geq 0$.

Apparently, Method 21 seems to demand a computation proportional to k to solve the auxiliary subproblems to obtain each z_{k+1} due to the weighted summation of $l_f(x_i; x)$'s. However, for all cases considered in Example 14, excepting (36), the auxiliary subproblems can be simplified to the form (32).

When $\{\beta_k\}$ is constant, $L(\cdot) \equiv L$, and $\delta(\cdot) \equiv 0$, the above two methods are very similar to Tseng's methods [16, 17]. In particular, choosing $\beta_k := L/\sigma$, $\lambda_0 := 1$ and $\lambda_{k+1} := \frac{1+\sqrt{1+4\lambda_k^2}}{2}$ in Methods 20 and 21, they yield Tseng's second and third APG methods (8) and (9), respectively. We provide a unified way to analyze these methods while Tseng's methods require slightly different approaches for each case.

For the inexact oracle model, Methods 20 and 21 can be seen as accelerated versions of primal and dual gradient methods in [6]. *The fast gradient method* in [6] corresponds to a hybrid of these accelerations. Methods 20 and 21 solve only one subproblem at each iteration preserving the same complexity as the fast gradient method.

6 Concluding remarks

We have proposed a new family of (sub)gradient-based methods for some classes of convex optimization problems, and also provided a unifying way of analyzing these methods which were performed separately and independently in the past. We have identified a general relation (Property 2) which the auxiliary functions of the mirror-descent and dual-averaging methods should satisfy.

There are infinitely many ways of implementing our methods since Proposition 4 shows that we can freely select from the extended MD model (11) or the DA model (12) the $l_f(x_i; x)$'s and the scaled proximal function to construct each subproblem at each iteration. All of them achieve the optimal complexity. These methods require a solution of only one subproblem per iteration. From the viewpoint of the relation (16), which we call (R_k), the extended mirror-descent model (11) has a "greedy" feature in the following sense; at each iteration, it attains the smallest upper bound $f(\hat{x}_k) \leq \psi_k(z_k)/S_k$ among those bounds for auxiliary functions satisfying Property 2 given the previous $\psi_{k-1}(x)$.

We list some further consideration to extend our approach as follows:

- In order to ensure optimal convergence, our methods require knowing the Lipschitz constant of the gradient of the objective function for the structured case (Section 5). There are, however, some approaches which remove this requirement as observed in [3, 9, 13, 15]. One can expect to obtain similar result applying these techniques for the proposed methods.

- For the case of convex problems with composite structure considered in Beck and Teboulle [4], it is possible to obtain a family of *smoothing-based first order methods* since our methods correspond to the *fast iterative method*.

Acknowledgements

We thank Nobuo Yamashita for pointing out Paul Tseng's earlier work [16, 17].

References

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

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

[3] 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.

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

[5] 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.

[6] O. Devolder, F. Glineur and Y. Nesterov, First-order methods of smooth convex optimization with inexact oracle, *Mathematical Programming*, Online first, DOI 10.1007/s10107-013-0677-5.

[7] 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.

[8] A. Nemirovski and D. Yudin, *Problem complexity and method efficiency in optimization*, Nauka Publishers, 1978 (in Russian); English translation: John Wiley & Sons, 1983.

[9] 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.

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

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

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

[13] Y. Nesterov, Gradient methods for minimizing composite objective function, *CORE Discussion Paper*, **76**, 2007.

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

[15] Y. Nesterov, Universal gradient methods for convex optimization problems, *CORE Discussion Paper*, **26**, 2013.

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

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