At iteration k, the vector $\mathbf{x}^k$ is picked as an optimal solution of the problem $\min_{\mathbf{x}\in X}\{f(\mathbf{x}) + (\boldsymbol{\lambda}^k)^T\mathbf{g}(\mathbf{x})\}$, meaning

$$\begin{aligned}
\mathbf{x}^k &\in \operatorname{argmin}_{\mathbf{x}\in X}\left\{f(\mathbf{x}) + (\boldsymbol{\lambda}^k)^T\mathbf{g}(\mathbf{x})\right\} \\
&= \operatorname{argmin}_{\mathbf{x}\in X}\left\{-\sum_{s=1}^{S} u_s(x_s) + \sum_{\ell=1}^{L}\lambda_\ell^k\left[\sum_{s\in\mathcal{S}(\ell)} x_s - c_\ell\right]\right\} \\
&= \operatorname{argmin}_{\mathbf{x}\in X}\left\{-\sum_{s=1}^{S} u_s(x_s) + \sum_{\ell=1}^{L}\sum_{s\in\mathcal{S}(\ell)}\lambda_\ell^k x_s\right\} \\
&= \operatorname{argmin}_{\mathbf{x}\in X}\left\{-\sum_{s=1}^{S} u_s(x_s) + \sum_{s=1}^{S}\left[\sum_{\ell\in\mathcal{L}(s)}\lambda_\ell^k\right]x_s\right\}.
\end{aligned}$$

The above minimization problem is separable w.r.t. the decision variables $x_1, x_2, \ldots, x_S$. Therefore, the sth element of $\mathbf{x}^k$ can be chosen via the update rule (returning to the max form),

$$x_s^k \in \operatorname{argmax}_{x_s\in I_s}\left\{u_s(x_s) - \left[\sum_{\ell\in\mathcal{L}(s)}\lambda_\ell^k\right]x_s\right\}.$$

The dual projected subgradient method employed on problem (8.91) with stepsizes α_k and initialization $\boldsymbol{\lambda}^0 = \mathbf{0}$ therefore takes the form below. Note that we do not consider here a normalized stepsize (actually, in many practical scenarios, a constant stepsize is used).

Dual Projected Subgradient Method for Solving the NUM Problem (8.91)

Initialization: define $\lambda_\ell^0 = 0$ for all $\ell \in \mathcal{L}$.

(A) **Source-rate update:**

$$x_s^k = \operatorname{argmax}_{x_s\in I_s}\left\{u_s(x_s) - \left[\sum_{\ell\in\mathcal{L}(s)}\lambda_\ell^k\right]x_s\right\}, \qquad s \in \mathcal{S}. \tag{8.92}$$

(B) **Link-price update:**

$$\lambda_\ell^{k+1} = \left[\lambda_\ell^k + \alpha_k\left(\sum_{s\in\mathcal{S}(\ell)} x_s^k - c_\ell\right)\right]_+, \qquad \ell \in \mathcal{L}.$$

The multipliers λ_ℓ^k can actually be seen as prices that are associated with the links. The algorithm above can be implemented in a distributed manner in the following sense:

(a) Each source s needs to solve the optimization problem (8.92) involving only its own utility function u_s and the multipliers (i.e., prices) associated with the links that it uses, meaning $\lambda_\ell^k, \ell \in \mathcal{L}(s)$.

(b) The price (i.e., multiplier) at each link ℓ is updated according to the rates of the sources that use the link ℓ, meaning $x_s, s \in \mathcal{S}(\ell)$.

Therefore, the algorithm only requires *local* communication between sources and links and can be implemented in a decentralized manner by letting both the sources and the links cooperatively seek an optimal solution of the problem by following the source-rate/price-link update scheme described above. This is one example of a *distributed optimization* method.

Chapter 9
Mirror Descent

This chapter is devoted to the study of the mirror descent method and some of its variations. The method is essentially a generalization of the projected subgradient method to the non-Euclidean setting. Therefore, naturally, we will *not* assume in the chapter that the underlying space is Euclidean.

9.1 From Projected Subgradient to Mirror Descent

Consider the optimization problem

$$\text{(P)} \quad \min\{f(\mathbf{x}) : \mathbf{x} \in C\}, \tag{9.1}$$

where we assume the following.[49]

Assumption 9.1.

(A) $f : \mathbb{E} \to (-\infty, \infty]$ *is proper closed and convex.*

(B) $C \subseteq \mathbb{E}$ *is nonempty closed and convex.*

(C) $C \subseteq \operatorname{int}(\operatorname{dom}(f))$.

(D) *The optimal set of* (P) *is nonempty and denoted by* X^*. *The optimal value of the problem is denoted by* f_{opt}.

The projected subgradient method for solving problem (P) was studied in Chapter 8. One of the basic assumptions made in Chapter 8, which was used throughout the analysis, is that the underlying space is Euclidean, meaning that $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. Recall that the general update step of the projected subgradient method has the form

$$\mathbf{x}^{k+1} = P_C(\mathbf{x}^k - t_k f'(\mathbf{x}^k)), \quad f'(\mathbf{x}^k) \in \partial f(\mathbf{x}^k), \tag{9.2}$$

for an appropriately chosen stepsize t_k. When the space is non-Euclidean, there is actually a "philosophical" problem with the update rule (9.2)—the vectors $\mathbf{x}^k$ and

[49] Assumption 9.1 is the same as Assumption 8.7 from Chapter 8.

$f'(\mathbf{x}^k)$ are in different spaces; one is in $\mathbb{E}$, while the other in $\mathbb{E}^*$. This issue is of course not really problematic since we can use our convention that the vectors in $\mathbb{E}$ and $\mathbb{E}^*$ are the same, and the only difference is in the norm associated with each of the spaces. Nonetheless, this issue is one of the motivations for seeking generalizations of the projected subgradient method better suited to the non-Euclidean setting.

To understand the role of the Euclidean norm in the definition of the projected subgradient method, we will consider the following reformulation of the update step (9.2):

$$\mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{x}\in C} \left\{ f(\mathbf{x}^k) + \langle f'(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{1}{2t_k} \|\mathbf{x} - \mathbf{x}^k\|^2 \right\}, \tag{9.3}$$

which actually shows that $\mathbf{x}^{k+1}$ is constructed by minimizing a linearization of the objective function plus a quadratic proximity term. The equivalence between the two forms (9.2) and (9.3) in the Euclidean case is evident by the following identity:

$$f(\mathbf{x}^k) + \langle f'(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{1}{2t_k} \|\mathbf{x} - \mathbf{x}^k\|^2 = \frac{1}{2t_k} \left\| \mathbf{x} - \left[\mathbf{x}^k - t_k f'(\mathbf{x}^k) \right] \right\|^2 + D,$$

where D is a constant (i.e., does not depend on $\mathbf{x}$).

Coming back to the non-Euclidean case, the idea will be to replace the Euclidean "distance" function $\frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2$ in (9.3) by a different distance, which is not based on the Euclidean norm. The non-Euclidean distances that we will use are *Bregman distances*.

Definition 9.2 (Bregman distance). *Let $\omega : \mathbb{E} \to (-\infty, \infty]$ be a proper closed and convex function that is differentiable over* $\operatorname{dom}(\partial\omega)$. *The* **Bregman distance** *associated with ω is the function $B_\omega : \operatorname{dom}(\omega) \times \operatorname{dom}(\partial\omega) \to \mathbb{R}$ given by*

$$B_\omega(\mathbf{x}, \mathbf{y}) = \omega(\mathbf{x}) - \omega(\mathbf{y}) - \langle \nabla\omega(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

The assumptions on ω (given a set C) are gathered in the following.

Assumption 9.3 (properties of ω).

- *ω is proper closed and convex.*
- *ω is differentiable over* $\operatorname{dom}(\partial\omega)$.
- $C \subseteq \operatorname{dom}(\omega)$.
- *$\omega + \delta_C$ is σ-strongly convex ($\sigma > 0$).*

A Bregman distance is actually not necessarily a distance. It is nonnegative and equal to zero if and only if its two arguments coincide, but other than that, in general it is not symmetric and does not satisfy the triangle inequality. The properties of Bregman distances that do hold are summarized in the following lemma.

Lemma 9.4 (basic properties of Bregman distances). *Suppose that $C \subseteq \mathbb{E}$ is nonempty closed and convex and that ω satisfies the properties in Assumption 9.3. Let B_ω be the Bregman distance associated with ω. Then*

(a) $B_\omega(\mathbf{x}, \mathbf{y}) \geq \frac{\sigma}{2}\|\mathbf{x} - \mathbf{y}\|^2$ *for all* $\mathbf{x} \in C, \mathbf{y} \in C \cap \operatorname{dom}(\partial\omega)$.

(b) *Let* $\mathbf{x} \in C$ *and* $\mathbf{y} \in C \cap \operatorname{dom}(\partial\omega)$. *Then*

- $B_\omega(\mathbf{x}, \mathbf{y}) \geq 0$;
- $B_\omega(\mathbf{x}, \mathbf{y}) = 0$ *if and only if* $\mathbf{x} = \mathbf{y}$.

Proof. Part (a) follows by the first-order characterization of strongly convex functions described in Theorem 5.24(ii). Part (b) is a direct consequence of part (a). ∎

Assume that $\mathbf{x}^k \in C \cap \operatorname{dom}(\partial\omega)$. Replacing the term $\frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|^2$ in formula (9.3) by a Bregman distance $B_\omega(\mathbf{x}, \mathbf{x}^k)$ leads to the following update step:

$$\mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{x} \in C} \left\{ f(\mathbf{x}^k) + \langle f'(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{1}{t_k} B_\omega(\mathbf{x}, \mathbf{x}^k) \right\}. \tag{9.4}$$

Omitting constant terms, (9.4) becomes

$$\mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{x} \in C} \left\{ \langle f'(\mathbf{x}^k), \mathbf{x} \rangle + \frac{1}{t_k} B_\omega(\mathbf{x}, \mathbf{x}^k) \right\}. \tag{9.5}$$

Further simplification of the update formula can be achieved by noting the following simple identity:

$$\begin{aligned} &\langle f'(\mathbf{x}^k), \mathbf{x} \rangle + \frac{1}{t_k} B_\omega(\mathbf{x}, \mathbf{x}^k) \\ &= \frac{1}{t_k} \left[\langle t_k f'(\mathbf{x}^k) - \nabla\omega(\mathbf{x}^k), \mathbf{x} \rangle + \omega(\mathbf{x}) \right] \underbrace{- \frac{1}{t_k}\omega(\mathbf{x}^k) + \frac{1}{t_k}\langle \nabla\omega(\mathbf{x}^k), \mathbf{x}^k \rangle}_{\text{constant}}. \end{aligned}$$

Therefore, the update formula in its most simplified form reads as

$$\mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{x} \in C} \left\{ \langle t_k f'(\mathbf{x}^k) - \nabla\omega(\mathbf{x}^k), \mathbf{x} \rangle + \omega(\mathbf{x}) \right\}.$$

We are now ready to define the mirror descent method.

The Mirror Descent Method

Initialization: pick $\mathbf{x}^0 \in C \cap \operatorname{dom}(\partial\omega)$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick a stepsize $t_k > 0$ and a subgradient $f'(\mathbf{x}^k) \in \partial f(\mathbf{x}^k)$;

(b) set

$$\mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{x} \in C} \left\{ \langle t_k f'(\mathbf{x}^k) - \nabla\omega(\mathbf{x}^k), \mathbf{x} \rangle + \omega(\mathbf{x}) \right\}. \tag{9.6}$$

Remark 9.5. *Although* (9.6) *is the most simplified form of the update step of the mirror descent method, the formula* (9.5)*, which can also be written as*

$$\mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{x}\in C} \left\{ \langle t_k f'(\mathbf{x}^k), \mathbf{x} \rangle + B_\omega(\mathbf{x}, \mathbf{x}^k) \right\}, \tag{9.7}$$

will also prove itself to be useful.

Remark 9.6. *Defining* $\tilde{\omega} = \omega + \delta_C$*, we can write the step* (9.6) *as*

$$\mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{x}\in \mathbb{E}} \left\{ \langle t_k f'(\mathbf{x}^k) - \nabla\omega(\mathbf{x}^k), \mathbf{x} \rangle + \tilde{\omega}(\mathbf{x}) \right\}. \tag{9.8}$$

Since $\nabla\omega(\mathbf{x}^k) \in \partial\tilde{\omega}(\mathbf{x}^k)$*, we can write it as* $\tilde{\omega}'(\mathbf{x}^k)$*, so* (9.8) *becomes*

$$\mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{x}\in \mathbb{E}} \left\{ \langle t_k f'(\mathbf{x}^k) - \tilde{\omega}'(\mathbf{x}^k), \mathbf{x} \rangle + \tilde{\omega}(\mathbf{x}) \right\}. \tag{9.9}$$

Finally, by the conjugate correspondence theorem (Theorem 5.26*), whose assumptions hold (properness, closedness, and strong convexity of* $\tilde{\omega}$*),* $\tilde{\omega}^*$ *is differentiable, which, combined with the conjugate subgradient theorem (Corollary* 4.21*), yields that* (9.9) *is equivalent to the following known formula for the mirror descent method:*

$$\mathbf{x}^{k+1} = \nabla\tilde{\omega}^*(\tilde{\omega}'(\mathbf{x}^k) - t_k f'(\mathbf{x}^k)).$$

The basic step of the mirror descent method (9.6) is of the form

$$\min_{\mathbf{x}\in C} \{ \langle \mathbf{a}, \mathbf{x} \rangle + \omega(\mathbf{x}) \} \tag{9.10}$$

for some $\mathbf{a} \in \mathbb{E}^*$. To show that the method is well defined, Theorem 9.8 below establishes the fact that the minimum of problem (9.10) is uniquely attained at a point in $C \cap \operatorname{dom}(\partial\omega)$. The reason why it is important to show that the minimizer is in $\operatorname{dom}(\partial\omega)$ is that the method requires computing the gradient of ω at the new iterate vector (recall that ω is assumed to be differentiable over $\operatorname{dom}(\partial\omega)$). We will prove a more general lemma that will also be useful in other contexts.

Lemma 9.7. *Assume the following:*

- $\omega : \mathbb{E} \to (-\infty, \infty]$ *is a proper closed and convex function differentiable over* $\operatorname{dom}(\partial\omega)$*.*
- $\psi : \mathbb{E} \to (-\infty, \infty]$ *is a proper closed and convex function satisfying* $\operatorname{dom}(\psi) \subseteq \operatorname{dom}(\omega)$*.*
- $\omega + \delta_{\operatorname{dom}(\psi)}$ *is* σ*-strongly convex (*$\sigma > 0$*).*

Then the minimizer of the problem

$$\min_{\mathbf{x}\in \mathbb{E}} \{ \psi(\mathbf{x}) + \omega(\mathbf{x}) \} \tag{9.11}$$

is uniquely attained at a point in $\operatorname{dom}(\psi) \cap \operatorname{dom}(\partial\omega)$*.*

Proof. Problem (9.11) is the same as

$$\min_{\mathbf{x}\in \mathbb{E}} \varphi(\mathbf{x}), \tag{9.12}$$

where $\varphi = \psi + \omega$. The function φ is closed since both ψ and ω are closed; it is proper by the fact that $\mathrm{dom}(\varphi) = \mathrm{dom}(\psi) \neq \emptyset$. Since $\omega + \delta_{\mathrm{dom}(\psi)}$ is σ-strongly convex and ψ is convex, their sum $\psi + \omega + \delta_{\mathrm{dom}(\psi)} = \psi + \omega = \varphi$ is σ-strongly convex. To conclude, φ is proper closed and σ-strongly convex, and hence, by Theorem 5.25(a), problem (9.12) has a unique minimizer $\mathbf{x}^*$ in $\mathrm{dom}(\varphi) = \mathrm{dom}(\psi)$. To show that $\mathbf{x}^* \in \mathrm{dom}(\partial\omega)$, note that by Fermat's optimality condition (Theorem 3.63), $\mathbf{0} \in \partial\varphi(\mathbf{x}^*)$, and in particular $\partial\varphi(\mathbf{x}^*) \neq \emptyset$. Therefore, since by the sum rule of subdifferential calculus (Theorem 3.40), $\partial\varphi(\mathbf{x}^*) = \partial\psi(\mathbf{x}^*) + \partial\omega(\mathbf{x}^*)$, it follows in particular that $\partial\omega(\mathbf{x}^*) \neq \emptyset$, meaning that $\mathbf{x}^* \in \mathrm{dom}(\partial\omega)$. ∎

The fact that the mirror descent method is well defined can now be easily deduced.

Theorem 9.8 (mirror descent is well defined). *Suppose that Assumptions* 9.1 *and* 9.3 *hold. Let* $\mathbf{a} \in \mathbb{E}^*$. *Then the problem*

$$\min_{\mathbf{x}\in C}\{\langle \mathbf{a}, \mathbf{x}\rangle + \omega(\mathbf{x})\}$$

has a unique minimizer in $C \cap \mathrm{dom}(\partial\omega)$.

Proof. The proof follows by invoking Lemma 9.7 with $\psi(\mathbf{x}) \equiv \langle \mathbf{a}, \mathbf{x}\rangle + \delta_C(\mathbf{x})$. ∎

Two very common choices of strongly convex functions are described below.

Example 9.9 (squared Euclidean norm). Suppose that Assumption 9.1 holds and that $\mathbb{E}$ is Euclidean, meaning that its norm satisfies $\|\cdot\| = \sqrt{\langle \cdot, \cdot\rangle}$. Define

$$\omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2.$$

Then ω obviously satisfies the properties listed in Assumption 9.3—it is proper closed and 1-strongly convex. Since $\nabla\omega(\mathbf{x}) = \mathbf{x}$, then the general update step of the mirror descent method reads as

$$\mathbf{x}^{k+1} = \mathrm{argmin}_{\mathbf{x}\in C}\left\{\langle t_k f'(\mathbf{x}^k) - \mathbf{x}^k, \mathbf{x}\rangle + \frac{1}{2}\|\mathbf{x}\|^2\right\},$$

which is the same as the projected subgradient update step: $\mathbf{x}^{k+1} = P_C(\mathbf{x}^k - t_k f'(\mathbf{x}^k))$. This is of course not a surprise since the method was constructed as a generalization of the projected subgradient method. ∎

Example 9.10 (negative entropy over the unit simplex). Suppose that Assumption 9.1 holds with $\mathbb{E} = \mathbb{R}^n$ endowed with the l_1-norm and $C = \Delta_n$. We will take ω to be the negative entropy over the nonnegative orthant:

$$\omega(\mathbf{x}) = \begin{cases} \sum_{i=1}^n x_i \log x_i, & \mathbf{x} \in \mathbb{R}^n_+, \\ \infty & \text{else.} \end{cases}$$

As usual, we use the convention that $0 \log 0 = 0$. By Example 5.27, $\omega + \delta_{\Delta_n}$ is 1-strongly convex w.r.t. the l_1-norm. In this case,

$$\mathrm{dom}(\partial\omega) = \mathbb{R}^n_{++},$$

on which ω is indeed differentiable. Thus, all the properties of Assumption 9.3 hold. The associated Bregman distance is given for any $\mathbf{x} \in \Delta_n$ and $\mathbf{y} \in \Delta_n^+ \equiv \{\mathbf{x} \in \mathbb{R}^n_{++} : \mathbf{e}^T\mathbf{x} = 1\}$ by

$$\begin{aligned} B_\omega(\mathbf{x}, \mathbf{y}) &= \sum_{i=1}^n x_i \log x_i - \sum_{i=1}^n y_i \log y_i - \sum_{i=1}^n (\log(y_i) + 1)(x_i - y_i) \\ &= \sum_{i=1}^n x_i \log(x_i/y_i) + \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \\ &= \sum_{i=1}^n x_i \log(x_i/y_i), \end{aligned} \tag{9.13}$$

which is the so-called Kullback-Leibler divergence distance measure. The general update step of the mirror descent method has the form ($f_i'(\mathbf{x}^k)$ is the ith component of $f'(\mathbf{x}^k)$),

$$\mathbf{x}^{k+1} = \text{argmin}_{\mathbf{x} \in \Delta_n} \left\{ \sum_{i=1}^n (t_k f_i'(\mathbf{x}^k) - 1 - \log(x_i^k)) x_i + \sum_{i=1}^n x_i \log x_i \right\}. \tag{9.14}$$

By Example 3.71, the optimal solution of problem (9.14) is

$$x_i^{k+1} = \frac{e^{\log(x_i^k) + 1 - t_k f_i'(\mathbf{x}^k)}}{\sum_{j=1}^n e^{\log(x_j^k) + 1 - t_k f_j'(\mathbf{x}^k)}}, \quad i = 1, 2, \ldots, n,$$

which can be simplified into the following:

$$x_i^{k+1} = \frac{x_i^k e^{-t_k f_i'(\mathbf{x}^k)}}{\sum_{j=1}^n x_j^k e^{-t_k f_j'(\mathbf{x}^k)}}, \quad i = 1, 2, \ldots, n. \qquad \blacksquare$$

The natural question that arises is how to choose the stepsizes. The convergence analysis that will be developed in the next section will reveal some possible answers to this question.

9.2 Convergence Analysis

9.2.1 The Toolbox

The following identity, also known as the *three-points lemma*, is essential in the analysis of the mirror descent lemma.

Lemma 9.11 (three-points lemma).[50] *Suppose that $\omega : \mathbb{E} \to (-\infty, \infty]$ is proper closed and convex. Suppose in addition that ω is differentiable over $\text{dom}(\partial\omega)$. Assume that $\mathbf{a}, \mathbf{b} \in \text{dom}(\partial\omega)$ and $\mathbf{c} \in \text{dom}(\omega)$. Then the following equality holds:*

$$\langle \nabla\omega(\mathbf{b}) - \nabla\omega(\mathbf{a}), \mathbf{c} - \mathbf{a} \rangle = B_\omega(\mathbf{c}, \mathbf{a}) + B_\omega(\mathbf{a}, \mathbf{b}) - B_\omega(\mathbf{c}, \mathbf{b}).$$

[50] The three-points lemma was proven by Chen and Teboulle in [43].

Proof. By definition of B_ω,

$$\begin{aligned} B_\omega(\mathbf{c},\mathbf{a}) &= \omega(\mathbf{c}) - \omega(\mathbf{a}) - \langle \nabla\omega(\mathbf{a}), \mathbf{c}-\mathbf{a}\rangle, \\ B_\omega(\mathbf{a},\mathbf{b}) &= \omega(\mathbf{a}) - \omega(\mathbf{b}) - \langle \nabla\omega(\mathbf{b}), \mathbf{a}-\mathbf{b}\rangle, \\ B_\omega(\mathbf{c},\mathbf{b}) &= \omega(\mathbf{c}) - \omega(\mathbf{b}) - \langle \nabla\omega(\mathbf{b}), \mathbf{c}-\mathbf{b}\rangle. \end{aligned}$$

Hence,

$$\begin{aligned} B_\omega(\mathbf{c},\mathbf{a}) + B_\omega(\mathbf{a},\mathbf{b}) - B_\omega(\mathbf{c},\mathbf{b}) &= -\langle \nabla\omega(\mathbf{a}), \mathbf{c}-\mathbf{a}\rangle - \langle \nabla\omega(\mathbf{b}), \mathbf{a}-\mathbf{b}\rangle + \langle \nabla\omega(\mathbf{b}), \mathbf{c}-\mathbf{b}\rangle \\ &= \langle \nabla\omega(\mathbf{b}) - \nabla\omega(\mathbf{a}), \mathbf{c}-\mathbf{a}\rangle. \quad \square \end{aligned}$$

Another key lemma is an extension of the second prox theorem (Theorem 6.39) to the case of non-Euclidean distances.

Theorem 9.12 (non-Euclidean second prox theorem). *Let*

- $\omega : \mathbb{E} \to (-\infty, \infty]$ *be a proper closed and convex function differentiable over* $\mathrm{dom}(\partial\omega)$*;*
- $\psi : \mathbb{E} \to (-\infty, \infty]$ *be a proper closed and convex function satisfying* $\mathrm{dom}(\psi) \subseteq \mathrm{dom}(\omega)$*;*
- $\omega + \delta_{\mathrm{dom}(\psi)}$ *be* σ*-strongly convex* ($\sigma > 0$)*.*

Assume that $\mathbf{b} \in \mathrm{dom}(\partial\omega)$*, and let* $\mathbf{a}$ *be defined by*

$$\mathbf{a} = \mathrm{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ \psi(\mathbf{x}) + B_\omega(\mathbf{x},\mathbf{b}) \right\}. \tag{9.15}$$

Then $\mathbf{a} \in \mathrm{dom}(\partial\omega)$ *and for all* $\mathbf{u} \in \mathrm{dom}(\psi)$*,*

$$\langle \nabla\omega(\mathbf{b}) - \nabla\omega(\mathbf{a}), \mathbf{u}-\mathbf{a}\rangle \le \psi(\mathbf{u}) - \psi(\mathbf{a}). \tag{9.16}$$

Proof. Using the definition of B_ω, (9.15) can be rewritten as

$$\mathbf{a} = \mathrm{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ \psi(\mathbf{x}) - \langle \nabla\omega(\mathbf{b}), \mathbf{x}\rangle + \omega(\mathbf{x}) \right\}. \tag{9.17}$$

The fact that $\mathbf{a} \in \mathrm{dom}(\partial\omega)$ follows by invoking Lemma 9.7 with $\psi(\mathbf{x}) - \langle \nabla\omega(\mathbf{b}), \mathbf{x}\rangle$ taking the role of $\psi(\mathbf{x})$. Using Fermat's optimality condition (Theorem 3.63), it follows by (9.17) that there exists $\psi'(\mathbf{a}) \in \partial\psi(\mathbf{a})$ for which

$$\psi'(\mathbf{a}) + \nabla\omega(\mathbf{a}) - \nabla\omega(\mathbf{b}) = \mathbf{0}.$$

Hence, by the subgradient inequality, for any $\mathbf{u} \in \mathrm{dom}(\psi)$,

$$\langle \nabla\omega(\mathbf{b}) - \nabla\omega(\mathbf{a}), \mathbf{u}-\mathbf{a}\rangle = \langle \psi'(\mathbf{a}), \mathbf{u}-\mathbf{a}\rangle \le \psi(\mathbf{u}) - \psi(\mathbf{a}),$$

proving the desired result. $\square$

Using the non-Euclidean second prox theorem and the three-points lemma, we can now establish a fundamental inequality satisfied by the sequence generated

by the mirror descent method. The inequality can be seen as a generalization of Lemma 8.11.

Lemma 9.13 (fundamental inequality for mirror descent). *Suppose that Assumptions* 9.1 *and* 9.3 *hold. Let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by the mirror descent method with positive stepsizes* $\{t_k\}_{k\geq 0}$*. Then for any* $\mathbf{x}^* \in X^*$ *and* $k \geq 0$*,*

$$t_k(f(\mathbf{x}^k) - f_{\text{opt}}) \leq B_\omega(\mathbf{x}^*, \mathbf{x}^k) - B_\omega(\mathbf{x}^*, \mathbf{x}^{k+1}) + \frac{t_k^2}{2\sigma}\|f'(\mathbf{x}^k)\|_*^2.$$

Proof. By the update formula (9.7) for $\mathbf{x}^{k+1}$ and the non-Euclidean second prox theorem (Theorem 9.12) invoked with $\mathbf{b} = \mathbf{x}^k$ and $\psi(\mathbf{x}) \equiv t_k\langle f'(\mathbf{x}^k), \mathbf{x}\rangle + \delta_C(\mathbf{x})$ (and hence $\mathbf{a} = \mathbf{x}^{k+1}$), we have for any $\mathbf{u} \in C$,

$$\langle \nabla\omega(\mathbf{x}^k) - \nabla\omega(\mathbf{x}^{k+1}), \mathbf{u} - \mathbf{x}^{k+1}\rangle \leq t_k\langle f'(\mathbf{x}^k), \mathbf{u} - \mathbf{x}^{k+1}\rangle. \tag{9.18}$$

By the three-points lemma (with $\mathbf{a} = \mathbf{x}^{k+1}, \mathbf{b} = \mathbf{x}^k$, and $\mathbf{c} = \mathbf{u}$),

$$\langle \nabla\omega(\mathbf{x}^k) - \nabla\omega(\mathbf{x}^{k+1}), \mathbf{u} - \mathbf{x}^{k+1}\rangle = B_\omega(\mathbf{u}, \mathbf{x}^{k+1}) + B_\omega(\mathbf{x}^{k+1}, \mathbf{x}^k) - B_\omega(\mathbf{u}, \mathbf{x}^k),$$

which, combined with (9.18), gives

$$B_\omega(\mathbf{u}, \mathbf{x}^{k+1}) + B_\omega(\mathbf{x}^{k+1}, \mathbf{x}^k) - B_\omega(\mathbf{u}, \mathbf{x}^k) \leq t_k\langle f'(\mathbf{x}^k), \mathbf{u} - \mathbf{x}^{k+1}\rangle.$$

Therefore,

$$\begin{aligned}
&t_k\langle f'(\mathbf{x}^k), \mathbf{x}^k - \mathbf{u}\rangle \\
&\leq B_\omega(\mathbf{u}, \mathbf{x}^k) - B_\omega(\mathbf{u}, \mathbf{x}^{k+1}) - B_\omega(\mathbf{x}^{k+1}, \mathbf{x}^k) + t_k\langle f'(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^{k+1}\rangle \\
&\overset{(*)}{\leq} B_\omega(\mathbf{u}, \mathbf{x}^k) - B_\omega(\mathbf{u}, \mathbf{x}^{k+1}) - \frac{\sigma}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + t_k\langle f'(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^{k+1}\rangle \\
&= B_\omega(\mathbf{u}, \mathbf{x}^k) - B_\omega(\mathbf{u}, \mathbf{x}^{k+1}) - \frac{\sigma}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \left\langle \frac{t_k}{\sqrt{\sigma}} f'(\mathbf{x}^k), \sqrt{\sigma}(\mathbf{x}^k - \mathbf{x}^{k+1})\right\rangle \\
&\overset{(**)}{\leq} B_\omega(\mathbf{u}, \mathbf{x}^k) - B_\omega(\mathbf{u}, \mathbf{x}^{k+1}) - \frac{\sigma}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \frac{t_k^2}{2\sigma}\|f'(\mathbf{x}^k)\|_*^2 + \frac{\sigma}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \\
&= B_\omega(\mathbf{u}, \mathbf{x}^k) - B_\omega(\mathbf{u}, \mathbf{x}^{k+1}) + \frac{t_k^2}{2\sigma}\|f'(\mathbf{x}^k)\|_*^2,
\end{aligned}$$

where the inequality $(*)$ follows by Lemma 9.4(a) and $(**)$ by Fenchel's inequality (Theorem 4.6) employed on the function $\frac{1}{2}\|\mathbf{x}\|^2$ (whose conjugate is $\frac{1}{2}\|\mathbf{y}\|_*^2$—see Section 4.4.15). Plugging in $\mathbf{u} = \mathbf{x}^*$ and using the subgradient inequality, we obtain

$$t_k(f(\mathbf{x}^k) - f_{\text{opt}}) \leq B_\omega(\mathbf{x}^*, \mathbf{x}^k) - B_\omega(\mathbf{x}^*, \mathbf{x}^{k+1}) + \frac{t_k^2}{2\sigma}\|f'(\mathbf{x}^k)\|_*^2. \qquad \square$$

Under a boundedness assumption on $B_\omega(\mathbf{x}, \mathbf{x}^0)$ over C, we can deduce a useful bound on the sequence of best achieved function values defined by

$$f_{\text{best}}^k \equiv \min_{n=0,1,\ldots,k} f(\mathbf{x}^n). \tag{9.19}$$

Lemma 9.14. *Suppose that Assumptions* 9.1 *and* 9.3 *hold and that* $\|f'(\mathbf{x})\|_* \le L_f$ *for all* $\mathbf{x} \in C$*, where* $L_f > 0$*. Suppose that* $B_\omega(\mathbf{x}, \mathbf{x}^0)$ *is bounded over* C*, and let* $\Theta(\mathbf{x}^0)$ *satisfy*

$$\Theta(\mathbf{x}^0) \ge \max_{\mathbf{x}\in C} B_\omega(\mathbf{x}, \mathbf{x}^0).$$

Let $\{\mathbf{x}^k\}_{k\ge 0}$ *be the sequence generated by the mirror descent method with positive stepsizes* $\{t_k\}_{k\ge 0}$*. Then for any* $N \ge 0$*,*

$$f_{\text{best}}^N - f_{\text{opt}} \le \frac{\Theta(\mathbf{x}^0) + \frac{L_f^2}{2\sigma}\sum_{k=0}^N t_k^2}{\sum_{k=0}^N t_k}, \tag{9.20}$$

where f_{best}^N *is defined in* (9.19)*.*

Proof. Let $\mathbf{x}^* \in X^*$. By Lemma 9.13 it follows that for any $k \ge 0$,

$$t_k(f(\mathbf{x}^k) - f_{\text{opt}}) \le B_\omega(\mathbf{x}^*, \mathbf{x}^k) - B_\omega(\mathbf{x}^*, \mathbf{x}^{k+1}) + \frac{t_k^2}{2\sigma}\|f'(\mathbf{x}^k)\|_*^2. \tag{9.21}$$

Summing (9.21) over $k = 0, 1, \ldots, N$, we obtain

$$\begin{aligned}\sum_{k=0}^N t_k(f(\mathbf{x}^k) - f_{\text{opt}}) &\le B_\omega(\mathbf{x}^*, \mathbf{x}^0) - B_\omega(\mathbf{x}^*, \mathbf{x}^{N+1}) + \sum_{k=0}^N \frac{t_k^2}{2\sigma}\|f'(\mathbf{x}^k)\|_*^2 \\ &= \Theta(\mathbf{x}^0) + \frac{L_f^2}{2\sigma}\sum_{k=0}^N t_k^2,\end{aligned}$$

which, combined with the inequality $(\sum_{k=0}^n t_k)(f_{\text{best}}^N - f_{\text{opt}}) \le \sum_{k=0}^N t_k(f(\mathbf{x}^k) - f_{\text{opt}})$, yields the result (9.20). ∎

9.2.2 Fixed Number of Iterations

Let us begin by fixing the number of iterations N and deduce what the "optimal" stepsizes are in the sense that they bring the right-hand side of (9.20) to a minimum. For that, we will prove the following technical lemma.

Lemma 9.15. *The optimal solution of the problem*

$$\min_{t_1,\ldots,t_m>0} \frac{\alpha + \beta\sum_{k=1}^m t_k^2}{\sum_{k=1}^m t_k}, \tag{9.22}$$

where $\alpha, \beta > 0$*, is given by* $t_k = \sqrt{\frac{\alpha}{\beta m}}$*,* $k = 1, 2, \ldots, m$*. The optimal value is* $2\sqrt{\frac{\alpha\beta}{m}}$*.*

Proof. Denote the objective function of (9.22) by

$$\phi(\mathbf{t}) \equiv \frac{\alpha + \beta\sum_{k=1}^m t_k^2}{\sum_{k=1}^m t_k}.$$

Note that ϕ is a permutation symmetric function, meaning that $\phi(\mathbf{t}) = \phi(\mathbf{Pt})$ for any permutation matrix $\mathbf{P} \in \Lambda_m$. A consequence of this observation is that if problem (9.22) has an optimal solution, then it necessarily has an optimal solution in which all the variables are the same. To show this, take an arbitrary optimal solution $\mathbf{t}^*$ and a permutation matrix $\mathbf{P} \in \Lambda_m$. Since $\phi(\mathbf{Pt}^*) = \phi(\mathbf{t}^*)$, it follows that $\mathbf{Pt}^*$ is also an optimal solution of (9.22). Therefore, since ϕ is convex over the positive orthant,[51] it follows that

$$\frac{1}{m!} \sum_{\mathbf{P} \in \Lambda_m} \mathbf{P}\mathbf{t}^* = \frac{1}{m} \begin{pmatrix} \mathbf{e}^T\mathbf{t} \\ \vdots \\ \mathbf{e}^T\mathbf{t} \end{pmatrix}$$

is also an optimal solution, showing that there always exists an optimal solution with equal components. Problem (9.22) therefore reduces to (after substituting $t_1 = t_2 = \cdots = t_m = t$)

$$\min_{t>0} \frac{\alpha + \beta m t^2}{mt},$$

whose optimal solution is $t = \sqrt{\frac{\alpha}{\beta m}}$, and thus an optimal solution of problem (9.22) is given by $t_k = \sqrt{\frac{\alpha}{\beta m}}$, $k = 1, 2, \ldots, m$. Substituting this value into ϕ, we obtain that the optimal value is $2\sqrt{\frac{\alpha\beta}{m}}$. □

Using Lemma 9.15 with $\alpha = \Theta(\mathbf{x}^0)$, $\beta = \frac{L_f^2}{2\sigma}$ and $m = N+1$, we conclude that the minimum of the right-hand side of (9.20) is attained at $t_k = \frac{\sqrt{2\Theta(\mathbf{x}^0)\sigma}}{L_f\sqrt{N+1}}$. The $O(1/\sqrt{N})$ rate of convergence follows immediately.

Theorem 9.16 ($O(1/\sqrt{N})$ rate of convergence of mirror descent with fixed amount of iterations). *Suppose that Assumptions* 9.1 *and* 9.3 *hold and that* $\|f'(\mathbf{x})\|_* \le L_f$ *for all* $\mathbf{x} \in C$ *for some* $L_f > 0$. *Assume that* $B_\omega(\mathbf{x}, \mathbf{x}^0)$ *is bounded over* C, *and let* $\Theta(\mathbf{x}^0)$ *satisfy*

$$\Theta(\mathbf{x}^0) \ge \max_{\mathbf{x} \in C} B_\omega(\mathbf{x}, \mathbf{x}^0).$$

Let N *be a positive integer, and let* $\{\mathbf{x}^k\}_{k \ge 0}$ *be the sequence generated by the mirror descent method with*

$$t_k = \frac{\sqrt{2\Theta(\mathbf{x}^0)\sigma}}{L_f\sqrt{N+1}},\ k = 0, 1, \ldots, N. \tag{9.23}$$

Then

$$f_{\text{best}}^N - f_{\text{opt}} \le \frac{\sqrt{2\Theta(\mathbf{x}^0)}L_f}{\sqrt{\sigma}\sqrt{N+1}},$$

where f_{best}^N *is defined in* (9.19).

[51] See, for example, [10, Example 7.18].

Proof. By Lemma 9.14,

$$f_{\text{best}}^N - f_{\text{opt}} \leq \frac{\Theta(\mathbf{x}^0) + \frac{L_f^2}{2\sigma}\sum_{k=0}^N t_k^2}{\sum_{k=0}^N t_k}.$$

Plugging the expression (9.23) for the stepsizes into the above inequality, the result follows. □

Example 9.17 (optimization over the unit simplex). Consider the problem

$$\min\{f(\mathbf{x}) : \mathbf{x} \in \Delta_n\},$$

where $f : \mathbb{R}^n \to (-\infty, \infty]$ is proper closed convex and satisfies $\Delta_n \subseteq \text{int}(\text{dom}(f))$. Consider two possible algorithms.

- **Euclidean setting.** We assume that the underlying norm on $\mathbb{R}^n$ is the l_2-norm and $\omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, which is 1-strongly convex w.r.t. the l_2-norm. In this case, the mirror descent algorithm is the same as the projected subgradient method:

  $$\mathbf{x}^{k+1} = P_{\Delta_n}(\mathbf{x}^k - t_k f'(\mathbf{x}^k)).$$

  We will assume that the method starts with the vector $\mathbf{x}^0 = \frac{1}{n}\mathbf{e}$. For this choice,

  $$\max_{\mathbf{x}\in\Delta_n} B_\omega(\mathbf{x}, \mathbf{x}^0) = \max_{\mathbf{x}\in\Delta_n} \frac{1}{2}\left\|\mathbf{x} - \frac{1}{n}\mathbf{e}\right\|_2^2 = \frac{1}{2}\left(1 - \frac{1}{n}\right),$$

  and we will take $\Theta(\mathbf{x}^0) = 1$. By Theorem 9.16, we have that given a positive integer N, by appropriately choosing the stepsizes, we obtain that

  $$f_{\text{best}}^N - f_{\text{opt}} \leq \underbrace{\frac{\sqrt{2}L_{f,2}}{\sqrt{N+1}}}_{C_{\text{e}}^f}, \tag{9.24}$$

  where $L_{f,2} = \max_{\mathbf{x}\in\Delta_n} \|f'(\mathbf{x})\|_2$.

- **Non-Euclidean setting.** Here we assume that the underlying norm on $\mathbb{R}^n$ is the l_1-norm and that the convex function ω is chosen as the negative entropy function

  $$\omega(\mathbf{x}) = \begin{cases} \sum_{i=1}^n x_i \log(x_i), & \mathbf{x} \in \mathbb{R}^n_{++}, \\ \infty & \text{else.} \end{cases} \tag{9.25}$$

  By Example 5.27, $\omega + \delta_{\Delta_n}$ is 1-strongly convex w.r.t. the l_1-norm. By Example 9.10, the mirror descent method takes the form

  $$x_i^{k+1} = \frac{x_i^k e^{-t_k f_i'(\mathbf{x}^k)}}{\sum_{j=1}^n x_j^k e^{-t_k f_j'(\mathbf{x}^k)}}, \quad i = 1, 2, \ldots, n.$$

As in the Euclidean setting, we will also initialize the method with $\mathbf{x}^0 = \frac{1}{n}\mathbf{e}$. For this choice, using the fact that the Bregman distance coincides with the Kullback–Leibler divergence (see (9.13)), we obtain

$$\begin{aligned}\max_{\mathbf{x}\in\Delta_n} B_\omega\left(\mathbf{x}, \frac{1}{n}\mathbf{e}\right) &= \max_{\mathbf{x}\in\Delta_n} \sum_{i=1}^n x_i \log(nx_i) = \log(n) + \max_{\mathbf{x}\in\Delta_n} \sum_{i=1}^n x_i \log x_i \\ &= \log(n).\end{aligned}$$

We will thus take $\Theta(\mathbf{x}^0) = \log(n)$. By Theorem 9.16, we have that given a positive integer N, by appropriately choosing the stepsizes, we obtain that

$$f_{\text{best}}^N - f_{\text{opt}} \leq \underbrace{\frac{\sqrt{2\log(n)}L_{f,\infty}}{\sqrt{N+1}}}_{C_{\text{ne}}^f}, \tag{9.26}$$

where $L_{f,\infty} = \max_{\mathbf{x}\in\Delta_n} \|f'(\mathbf{x})\|_\infty$.

The ratio of the two upper bounds in (9.24) and (9.26) is given by

$$\rho^f = \frac{C_{\text{ne}}^f}{C_{\text{e}}^f} = \sqrt{\log(n)}\frac{L_{f,\infty}}{L_{f,2}}.$$

Whether or not ρ^f is greater than 1 (superiority of the Euclidean setting) or smaller than 1 (superiority of the non-Euclidean setting) depends on the properties of the function f. In any case, since $\|\mathbf{y}\|_\infty \leq \|\mathbf{y}\|_2 \leq \sqrt{n}\|\mathbf{y}\|_\infty$ for all $\mathbf{y} \in \mathbb{R}^n$, it follows that

$$\frac{1}{\sqrt{n}} \leq \frac{L_{f,\infty}}{L_{f,2}} \leq 1,$$

and hence that

$$\frac{\sqrt{\log(n)}}{\sqrt{n}} \leq \rho^f \leq \sqrt{\log(n)}.$$

Therefore, the ratio between the efficiency estimates ranges between $\frac{\sqrt{\log(n)}}{\sqrt{n}}$ (superiority of the non-Euclidean setting) and $\sqrt{\log(n)}$ (slight superiority of the Euclidean setting). ∎

9.2.3 Dynamic Stepsize Rule

The constant stepsize rule is relatively easy to analyze but has the disadvantage of requiring the a priori knowledge of the total number of iterations employed by the method. In practical situations, the number of iterations is not fixed a priori, and a stopping criteria different than merely fixing the total number of iterations is usually imposed. This is why dynamic (namely, nonconstant) stepsize rules are important. Similarly to the analysis in Chapter 8 for the projected subgradient method, it is possible to use the fundamental inequality for the mirror descent method (Lemma 9.13) to establish convergence results under dynamic stepsize rules.

Theorem 9.18 (convergence of mirror descent with dynamic stepsizes). *Suppose that Assumptions* 9.1 *and* 9.3 *hold and that* $\|f'(\mathbf{x})\|_* \leq L_f$ *for any* $\mathbf{x} \in C$

for some $L_f > 0$. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the mirror descent method with positive stepsizes $\{t_k\}_{k\geq 0}$, and let $\{f^k_{\text{best}}\}_{k\geq 0}$ be the sequence of best achieved values defined in (9.19).

(a) *If $\frac{\sum_{n=0}^k t_n^2}{\sum_{n=0}^k t_n} \to 0$ as $k \to \infty$, then $f^k_{\text{best}} \to f_{\text{opt}}$ as $k \to \infty$.*

(b) *If t_k is chosen as either* **(predefined diminishing stepsize)**

$$t_k = \frac{\sqrt{2\sigma}}{L_f\sqrt{k+1}}$$

or **(adaptive stepsize)**

$$t_k = \begin{cases} \frac{\sqrt{2\sigma}}{\|f'(\mathbf{x}^k)\|_*\sqrt{k+1}}, & f'(\mathbf{x}^k) \neq \mathbf{0}, \\ \frac{\sqrt{2\sigma}}{L_f\sqrt{k+1}}, & f'(\mathbf{x}^k) = \mathbf{0}, \end{cases}$$

then for all $k \geq 1$,

$$f^k_{\text{best}} - f_{\text{opt}} \leq \frac{L_f}{\sqrt{2\sigma}} \frac{B_\omega(\mathbf{x}^*, \mathbf{x}^0) + 1 + \log(k+1)}{\sqrt{k+1}}.$$

Proof. By the fundamental inequality for mirror descent (Lemma 9.13), we have, for all $n \geq 0$,

$$t_n(f(\mathbf{x}^n) - f_{\text{opt}}) \leq B_\omega(\mathbf{x}^*, \mathbf{x}^n) - B_\omega(\mathbf{x}^*, \mathbf{x}^{n+1}) + \frac{t_n^2}{2\sigma}\|f'(\mathbf{x}^n)\|_*^2.$$

Summing the above inequality over $n = 0, 1, \ldots, k$ gives

$$\sum_{n=0}^k t_n(f(\mathbf{x}^n) - f_{\text{opt}}) \leq B_\omega(\mathbf{x}^*, \mathbf{x}^0) - B_\omega(\mathbf{x}^*, \mathbf{x}^{k+1}) + \frac{1}{2\sigma}\sum_{n=0}^k t_n^2\|f'(\mathbf{x}^n)\|_*^2.$$

Using the inequalities $B_\omega(\mathbf{x}^*, \mathbf{x}^{k+1}) \geq 0$ and $f(\mathbf{x}^n) \geq f^k_{\text{best}} (n \leq k)$, we obtain

$$f^k_{\text{best}} - f_{\text{opt}} \leq \frac{B_\omega(\mathbf{x}^*, \mathbf{x}^0) + \frac{1}{2\sigma}\sum_{n=0}^k t_n^2\|f'(\mathbf{x}^n)\|_*^2}{\sum_{n=0}^k t_n}. \tag{9.27}$$

Since $\|f'(\mathbf{x}^n)\|_* \leq L_f$, we can deduce that

$$f^k_{\text{best}} - f_{\text{opt}} \leq \frac{B_\omega(\mathbf{x}^*, \mathbf{x}^0) + \frac{L_f^2}{2\sigma}\sum_{n=0}^k t_n^2}{\sum_{n=0}^k t_n}.$$

Therefore, if $\frac{\sum_{n=0}^k t_n^2}{\sum_{n=0}^k t_n} \to 0$, then $f^k_{\text{best}} \to f_{\text{opt}}$ as $k \to \infty$, proving claim (a).

To show the validity of claim (b), note that for both stepsize rules we have $t_n^2\|f'(\mathbf{x}^n)\|_*^2 \leq \frac{2\sigma}{n+1}$ and $t_n \geq \frac{\sqrt{2\sigma}}{L_f\sqrt{n+1}}$. Hence, by (9.27),

$$f^k_{\text{best}} - f_{\text{opt}} \leq \frac{L_f}{\sqrt{2\sigma}} \frac{B_\omega(\mathbf{x}^*, \mathbf{x}^0) + \sum_{n=0}^k \frac{1}{n+1}}{\sum_{n=0}^k \frac{1}{\sqrt{n+1}}},$$

which, combined with Lemma 8.27(a), yields the desired result. ∎

Example 9.19 (mirror descent vs. projected subgradient—numerical example). Consider the problem

$$\min\{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_1 : \mathbf{x}\in\Delta_n\}, \tag{9.28}$$

where $\mathbf{A}\in\mathbb{R}^{n\times n}$ and $\mathbf{b}\in\mathbb{R}^n$. Following Example 9.17, we consider two methods. The first is the projected subgradient method where $\mathbb{R}^n$ is assumed to be endowed with the Euclidean l_2-norm. The update formula is given by

$$\mathbf{x}^{k+1} = P_{\Delta_n}(\mathbf{x}^k - t_k f'(\mathbf{x}^k)),$$

with $f'(\mathbf{x}^k)$ taken as $\mathbf{A}^T\mathrm{sgn}(\mathbf{A}\mathbf{x}^k-\mathbf{b})$ and the stepsize t_k chosen by the adaptive stepsize rule (in practice, $f'(\mathbf{x}^k)$ is never the zeros vector):

$$t_k = \frac{\sqrt{2}}{\|f'(\mathbf{x}^k)\|_2\sqrt{k+1}}.$$

The second method is mirror descent in which the underlying norm on $\mathbb{R}^n$ is the l_1-norm and ω is chosen to be the negative entropy function given in (9.25). In this case, the method has the form (see Example 9.17)

$$x_i^{k+1} = \frac{x_i^k e^{-t_k f_i'(\mathbf{x}^k)}}{\sum_{j=1}^n x_j^k e^{-t_k f_j'(\mathbf{x}^k)}}, \quad i=1,2,\dots,n,$$

where here we take

$$t_k = \frac{\sqrt{2}}{\|f'(\mathbf{x}^k)\|_\infty\sqrt{k+1}}.$$

Note that the strong convexity parameter is $\sigma=1$ in both settings. We created an instance of problem (9.28) with $n=100$ by generating the components of $\mathbf{A}$ and $\mathbf{b}$ independently via a standard normal distribution. The values of $f(\mathbf{x}^k)-f_{\mathrm{opt}}$ and $f^k_{\mathrm{best}}-f_{\mathrm{opt}}$ for both methods are described in Figure 9.1. Evidently, the non-Euclidean method, referred to as md, is superior to the Euclidean projected subgradient method (ps). ∎

9.3 Mirror Descent for the Composite Model[52]

In this section we will consider a more general model than model (9.1), which was discussed in Sections 9.1 and 9.2. Consider the problem

$$\min_{\mathbf{x}\in\mathbb{E}}\{F(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x})\}, \tag{9.29}$$

where the following set of assumptions is made on f and g.

Assumption 9.20 (properties of f and g).

(A) $f, g : \mathbb{E}\to(-\infty,\infty]$ *are proper closed and convex.*

(B) $\mathrm{dom}(g)\subseteq\mathrm{int}(\mathrm{dom}(f))$.

[52]The analysis of the mirror-C method is based on the work of Duchi, Shalev-Shwartz, Singer, and Tewari [49], where the algorithm is introduced in an online and stochastic setting.

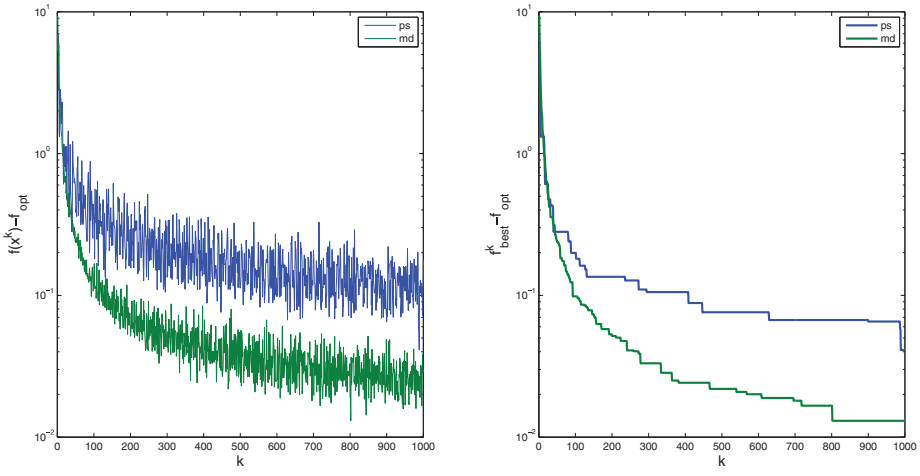

Figure 9.1. *The values* $f(\mathbf{x}^k) - f_{\text{opt}}$ *and* $f^k_{\text{best}} - f_{\text{opt}}$ *generated by the mirror descent and projected subgradient methods.*

(C) $\|f'(\mathbf{x})\|_* \le L_f$ *for any* $\mathbf{x} \in \text{dom}(g)$ $(L_f > 0)$.[53]

(D) *The optimal set of* (9.29) *is nonempty and denoted by* X^*. *The optimal value of the problem is denoted by* F_{opt}.

We will also assume, as usual, that we have at our disposal a convex function ω that satisfies the following properties, which are a slight adjustment of the properties in Assumption 9.3.

Assumption 9.21 (properties of ω).

- ω *is proper closed and convex.*
- ω *is differentiable over* $\text{dom}(\partial\omega)$.
- $\text{dom}(g) \subseteq \text{dom}(\omega)$.
- $\omega + \delta_{\text{dom}(g)}$ *is* σ*-strongly convex* $(\sigma > 0)$.

We can obviously ignore the composite structure of problem (9.29) and just try to employ the mirror descent method on the function $F = f + g$ with $\text{dom}(g)$ taking the role of C:

$$\mathbf{x}^{k+1} = \text{argmin}_{\mathbf{x}\in C} \left\{ \langle f'(\mathbf{x}^k) + g'(\mathbf{x}^k), \mathbf{x}\rangle + \frac{1}{t_k} B_\omega(\mathbf{x}, \mathbf{x}^k) \right\}. \tag{9.30}$$

[53] Recall that we assume that f' represents some rule that takes any $\mathbf{x} \in \text{dom}(\partial f)$ to a vector $f'(\mathbf{x}) \in \partial f(\mathbf{x})$.

However, employing the above scheme might be problematic. First, we did not assume that $C = \mathrm{dom}(g)$ is closed, and thus the argmin in (9.30) might be empty. Second, even if the update step is well defined, we did not assume that g is Lipschitz over C like we did on f in Assumption 9.20(C); this is a key element in the convergence analysis of the mirror descent method. Finally, even if g is Lipschitz over C, it might be that the Lipschitz constant of the sum function $F = f + g$ is much larger than the Lipschitz constant of f, and our objective will be to define a method whose efficiency estimate will depend only on the Lipschitz constant of f over $\mathrm{dom}(g)$.

Instead of linearizing both f and g, as is done in (9.30), we will linearize f and keep g as it is. This leads to the following scheme:

$$\mathbf{x}^{k+1} = \mathrm{argmin}_{\mathbf{x}} \left\{ \langle f'(\mathbf{x}^k), \mathbf{x} \rangle + g(\mathbf{x}) + \frac{1}{t_k} B_\omega(\mathbf{x}, \mathbf{x}^k) \right\}, \tag{9.31}$$

which can also be written as

$$\mathbf{x}^{k+1} = \mathrm{argmin}_{\mathbf{x}} \left\{ \langle t_k f'(\mathbf{x}^k) - \nabla\omega(\mathbf{x}^k), \mathbf{x} \rangle + t_k g(\mathbf{x}) + \omega(\mathbf{x}) \right\}.$$

The algorithm that performs the above update step will be called the *mirror-C* method.

The Mirror-C Method

Initialization: pick $\mathbf{x}^0 \in \mathrm{dom}(g) \cap \mathrm{dom}(\partial\omega)$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick a stepsize $t_k > 0$ and a subgradient $f'(\mathbf{x}^k) \in \partial f(\mathbf{x}^k)$;

(b) set

$$\mathbf{x}^{k+1} = \mathrm{argmin}_{\mathbf{x}} \left\{ \langle t_k f'(\mathbf{x}^k) - \nabla\omega(\mathbf{x}^k), \mathbf{x} \rangle + t_k g(\mathbf{x}) + \omega(\mathbf{x}) \right\}. \tag{9.32}$$

Remark 9.22. *The update formula* (9.32) *can also be rewritten as*

$$\mathbf{x}^{k+1} = \mathrm{argmin}_{\mathbf{x}} \left\{ \langle t_k f'(\mathbf{x}^k), \mathbf{x} \rangle + t_k g(\mathbf{x}) + B_\omega(\mathbf{x}, \mathbf{x}^k) \right\}. \tag{9.33}$$

Remark 9.23 (Euclidean setting—proximal subgradient method). *When the underlying space* $\mathbb{E}$ *is Euclidean and* $\omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$*, then the update formula* (9.33) *reduces to*

$$\mathbf{x}^{k+1} = \mathrm{argmin}_{\mathbf{x}} \left\{ \langle t_k f'(\mathbf{x}^k), \mathbf{x} \rangle + t_k g(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|^2 \right\},$$

which, after some rearrangement of terms and removal of constant terms, takes the form

$$\mathbf{x}^{k+1} = \mathrm{argmin}_{\mathbf{x}} \left\{ t_k g(\mathbf{x}) + \frac{1}{2} \left\| \mathbf{x} - [\mathbf{x}^k - t_k f'(\mathbf{x}^k)] \right\|^2 \right\}.$$

By the definition of the prox operator (see Chapter 6*), the last equation can be rewritten as*

$$\mathbf{x}^{k+1} = \operatorname{prox}_{t_k g}(\mathbf{x}^k - t_k f'(\mathbf{x}^k)).$$

Thus, at each iteration the method takes a step toward minus of the subgradient followed by a prox step. The resulting method is called the proximal subgradient method*. The method will be discussed extensively in Chapter* 10 *in the case where* f *possesses some differentiability properties.*

Of course, the mirror-C method coincides with the mirror descent method when taking $g = \delta_C$ with C being a nonempty closed and convex set. We begin by showing that the mirror-C method is well defined, meaning that the minimum in (9.32) is uniquely attained at $\operatorname{dom}(g) \cap \operatorname{dom}(\partial\omega)$.

Theorem 9.24 (mirror-C is well defined). *Suppose that Assumptions* 9.20 *and* 9.21 *hold. Let* $\mathbf{a} \in \mathbb{E}^*$*. Then the problem*

$$\min_{\mathbf{x}\in\mathbb{E}}\{\langle \mathbf{a}, \mathbf{x}\rangle + g(\mathbf{x}) + \omega(\mathbf{x})\}$$

has a unique minimizer in $\operatorname{dom}(g) \cap \operatorname{dom}(\partial\omega)$*.*

Proof. The proof follows by invoking Lemma 9.7 with $\psi(\mathbf{x}) \equiv \langle \mathbf{a}, \mathbf{x}\rangle + g(\mathbf{x})$. ∎

The analysis of the mirror-C method is based on arguments similar to those used in Section 9.2 to analyze the mirror descent method. We begin by proving a technical lemma establishing an inequality similar to the one derived in Lemma 9.14. Note that in addition to our basic assumptions, we assume that g is a nonnegative function and that the stepsizes are nonincreasing.

Lemma 9.25. *Suppose that Assumptions* 9.20 *and* 9.21 *hold and that* g *is a nonnegative function. Let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by the mirror-C method with positive nonincreasing stepsizes* $\{t_k\}_{k\geq 0}$*. Then for any* $\mathbf{x}^* \in X^*$ *and* $k \geq 0$*,*

$$\min_{n=0,1,\ldots,k} F(\mathbf{x}^n) - F_{\text{opt}} \leq \frac{t_0 g(\mathbf{x}^0) + B_\omega(\mathbf{x}^*, \mathbf{x}^0) + \frac{1}{2\sigma}\sum_{n=0}^k t_n^2 \|f'(\mathbf{x}^n)\|_*^2}{\sum_{n=0}^k t_n}. \quad (9.34)$$

Proof. By the update formula (9.33) and the non-Euclidean second prox theorem (Theorem 9.12) invoked with $\mathbf{b} = \mathbf{x}^n, \mathbf{a} = \mathbf{x}^{n+1}$, and $\psi(\mathbf{x}) \equiv t_n\langle f'(\mathbf{x}^n), \mathbf{x}\rangle + t_n g(\mathbf{x})$, we have

$$\langle \nabla\omega(\mathbf{x}^n) - \nabla\omega(\mathbf{x}^{n+1}), \mathbf{u} - \mathbf{x}^{n+1}\rangle \leq t_n\langle f'(\mathbf{x}^n), \mathbf{u} - \mathbf{x}^{n+1}\rangle + t_n g(\mathbf{u}) - t_n g(\mathbf{x}^{n+1}). \quad (9.35)$$

Invoking the three-points lemma (Lemma 9.11) with $\mathbf{a} = \mathbf{x}^{n+1}, \mathbf{b} = \mathbf{x}^n$, and $\mathbf{c} = \mathbf{u}$ yields

$$\langle \nabla\omega(\mathbf{x}^n) - \nabla\omega(\mathbf{x}^{n+1}), \mathbf{u} - \mathbf{x}^{n+1}\rangle = B_\omega(\mathbf{u}, \mathbf{x}^{n+1}) + B_\omega(\mathbf{x}^{n+1}, \mathbf{x}^n) - B_\omega(\mathbf{u}, \mathbf{x}^n),$$

which, combined with (9.35), gives

$$B_\omega(\mathbf{u}, \mathbf{x}^{n+1}) + B_\omega(\mathbf{x}^{n+1}, \mathbf{x}^n) - B_\omega(\mathbf{u}, \mathbf{x}^n) \leq t_n\langle f'(\mathbf{x}^n), \mathbf{u} - \mathbf{x}^{n+1}\rangle + t_n g(\mathbf{u}) - t_n g(\mathbf{x}^{n+1}).$$

Therefore,

$$\begin{aligned}
&t_n\langle f'(\mathbf{x}^n), \mathbf{x}^n - \mathbf{u}\rangle + t_n g(\mathbf{x}^{n+1}) - t_n g(\mathbf{u}) \\
&\le B_\omega(\mathbf{u}, \mathbf{x}^n) - B_\omega(\mathbf{u}, \mathbf{x}^{n+1}) - B_\omega(\mathbf{x}^{n+1}, \mathbf{x}^n) + t_n\langle f'(\mathbf{x}^n), \mathbf{x}^n - \mathbf{x}^{n+1}\rangle \\
&\le B_\omega(\mathbf{u}, \mathbf{x}^n) - B_\omega(\mathbf{u}, \mathbf{x}^{n+1}) - \frac{\sigma}{2}\|\mathbf{x}^{n+1} - \mathbf{x}^n\|^2 + t_n\langle f'(\mathbf{x}^n), \mathbf{x}^n - \mathbf{x}^{n+1}\rangle \\
&= B_\omega(\mathbf{u}, \mathbf{x}^n) - B_\omega(\mathbf{u}, \mathbf{x}^{n+1}) - \frac{\sigma}{2}\|\mathbf{x}^{n+1} - \mathbf{x}^n\|^2 + \left\langle \frac{t_n}{\sqrt{\sigma}} f'(\mathbf{x}^n), \sqrt{\sigma}(\mathbf{x}^n - \mathbf{x}^{n+1})\right\rangle \\
&\le B_\omega(\mathbf{u}, \mathbf{x}^n) - B_\omega(\mathbf{u}, \mathbf{x}^{n+1}) - \frac{\sigma}{2}\|\mathbf{x}^{n+1} - \mathbf{x}^n\|^2 + \frac{t_n^2}{2\sigma}\|f'(\mathbf{x}^n)\|_*^2 + \frac{\sigma}{2}\|\mathbf{x}^{n+1} - \mathbf{x}^n\|^2 \\
&= B_\omega(\mathbf{u}, \mathbf{x}^n) - B_\omega(\mathbf{u}, \mathbf{x}^{n+1}) + \frac{t_n^2}{2\sigma}\|f'(\mathbf{x}^n)\|_*^2.
\end{aligned}$$

Plugging in $\mathbf{u} = \mathbf{x}^*$ and using the subgradient inequality, we obtain

$$t_n\left[f(\mathbf{x}^n) + g(\mathbf{x}^{n+1}) - F_{\text{opt}}\right] \le B_\omega(\mathbf{x}^*, \mathbf{x}^n) - B_\omega(\mathbf{x}^*, \mathbf{x}^{n+1}) + \frac{t_n^2}{2\sigma}\|f'(\mathbf{x}^n)\|_*^2.$$

Summing the above over $n = 0, 1, \ldots, k$,

$$\sum_{n=0}^{k} t_n\left[f(\mathbf{x}^n) + g(\mathbf{x}^{n+1}) - F_{\text{opt}}\right] \le B_\omega(\mathbf{x}^*, \mathbf{x}^0) - B_\omega(\mathbf{x}^*, \mathbf{x}^{k+1}) + \frac{1}{2\sigma}\sum_{n=0}^{k} t_n^2\|f'(\mathbf{x}^n)\|_*^2.$$

Adding the term $t_0 g(\mathbf{x}^0) - t_k g(\mathbf{x}^{k+1})$ to both sides and using the nonnegativity of the Bregman distance, we get

$$\begin{aligned}
&t_0(F(\mathbf{x}^0) - F_{\text{opt}}) + \sum_{n=1}^{k}\left[t_n f(\mathbf{x}^n) + t_{n-1} g(\mathbf{x}^n) - t_n F_{\text{opt}}\right] \\
&\le t_0 g(\mathbf{x}^0) - t_k g(\mathbf{x}^{k+1}) + B_\omega(\mathbf{x}^*, \mathbf{x}^0) + \frac{1}{2\sigma}\sum_{n=0}^{k} t_n^2\|f'(\mathbf{x}^n)\|_*^2.
\end{aligned}$$

Using the fact that $t_n \le t_{n-1}$ and the nonnegativity of $g(\mathbf{x}^{k+1})$, we conclude that

$$\sum_{n=0}^{k} t_n\left[F(\mathbf{x}^n) - F_{\text{opt}}\right] \le t_0 g(\mathbf{x}^0) + B_\omega(\mathbf{x}^*, \mathbf{x}^0) + \frac{1}{2\sigma}\sum_{n=0}^{k} t_n^2\|f'(\mathbf{x}^n)\|_*^2,$$

which, combined with the fact that

$$\left(\sum_{n=0}^{k} t_n\right)\left(\min_{n=0,1,\ldots,k} F(\mathbf{x}^n) - F_{\text{opt}}\right) \le \sum_{n=0}^{k} t_n\left[F(\mathbf{x}^n) - F_{\text{opt}}\right],$$

implies the inequality (9.34). ∎

Using Lemma 9.25, it is now easy to derive a convergence result under the assumption that the number of iterations is fixed.

Theorem 9.26 ($O(1/\sqrt{N})$ rate of convergence of mirror-C with fixed amount of iterations). *Suppose that Assumptions* 9.20 *and* 9.21 *hold and that*

g is nonnegative. Assume that $B_\omega(\mathbf{x}, \mathbf{x}^0)$ is bounded above over $\mathrm{dom}(g)$*, and let* $\Theta(\mathbf{x}^0)$ *satisfy*

$$\Theta(\mathbf{x}^0) \geq \max_{\mathbf{x} \in \mathrm{dom}(g)} B(\mathbf{x}, \mathbf{x}^0).$$

Suppose that $g(\mathbf{x}^0) = 0$. Let N be a positive integer, and let $\{\mathbf{x}^k\}_{k \geq 0}$ be the sequence generated by the mirror-C method with constant stepsize

$$t_k = \frac{\sqrt{2\Theta(\mathbf{x}^0)\sigma}}{L_f \sqrt{N}}. \tag{9.36}$$

Then

$$\min_{n=0,1,\ldots,N-1} F(\mathbf{x}^n) - F_{\mathrm{opt}} \leq \frac{\sqrt{2\Theta(\mathbf{x}^0)} L_f}{\sqrt{\sigma}\sqrt{N}}.$$

Proof. By Lemma 9.25, using the fact that $g(\mathbf{x}^0) = 0$ and the inequalities $\|f'(\mathbf{x}^n)\|_* \leq L_f$ and $B_\omega(\mathbf{x}^*, \mathbf{x}^0) \leq \Theta(\mathbf{x}^0)$, we have

$$\min_{n=0,1,\ldots,N-1} F(\mathbf{x}^n) - F_{\mathrm{opt}} \leq \frac{\Theta(\mathbf{x}^0) + \frac{L_f^2}{2\sigma} \sum_{n=0}^{N-1} t_n^2}{\sum_{n=0}^{N-1} t_n}.$$

Plugging the expression (9.36) for the stepsizes into the above inequality, the result follows. ∎

We can also establish a rate of convergence of the mirror-C method with a dynamic stepsize rule.

Theorem 9.27 ($O(\log k/\sqrt{k})$ rate of convergence of mirror-C with dynamic stepsizes). *Suppose that Assumptions* 9.20 *and* 9.21 *hold and that g is nonnegative. Let $\{\mathbf{x}^k\}_{k \geq 0}$ be the sequence generated by the mirror-C method with stepsizes $\{t_k\}_{k \geq 0}$ chosen as*

$$t_k = \frac{\sqrt{2\sigma}}{L_f \sqrt{k+1}}.$$

Then for all $k \geq 1$,

$$\min_{n=0,1,\ldots,k} F(\mathbf{x}^n) - F_{\mathrm{opt}} \leq \frac{L_f}{\sqrt{2\sigma}} \frac{B_\omega(\mathbf{x}^*, \mathbf{x}^0) + \frac{\sqrt{2\sigma}}{L_f} g(\mathbf{x}^0) + 1 + \log(k+1)}{\sqrt{k+1}}. \tag{9.37}$$

Proof. By Lemma 9.25, taking into account the fact that $t_0 = \frac{\sqrt{2\sigma}}{L_f}$,

$$\min_{n=0,1,\ldots,k} F(\mathbf{x}^n) - F_{\mathrm{opt}} \leq \frac{B_\omega(\mathbf{x}^*, \mathbf{x}^0) + \frac{\sqrt{2\sigma}}{L_f} g(\mathbf{x}^0) + \frac{1}{2\sigma} \sum_{n=0}^{k} t_n^2 \|f'(\mathbf{x}^n)\|_*^2}{\sum_{n=0}^{k} t_n}, \tag{9.38}$$

which, along with the relations $t_n^2 \|f'(\mathbf{x}^n)\|_*^2 \leq \frac{2\sigma}{n+1}$ and $t_n = \frac{\sqrt{2\sigma}}{L_f\sqrt{n+1}}$, yields the inequality

$$\min_{n=0,1,\ldots,k} F(\mathbf{x}^n) - f_{\mathrm{opt}} \leq \frac{L_f}{\sqrt{2\sigma}} \frac{B_\omega(\mathbf{x}^*, \mathbf{x}^0) + \frac{\sqrt{2\sigma}}{L_f} g(\mathbf{x}^0) + \sum_{n=0}^{k} \frac{1}{n+1}}{\sum_{n=0}^{k} \frac{1}{\sqrt{n+1}}}.$$

The result (9.37) now follows by invoking Lemma 8.27(a). ∎

Example 9.28. Suppose that the underlying space is $\mathbb{R}^n$ endowed with the Euclidean l_2-norm. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function, which is Lipschitz over $\mathbb{R}^n$, implying that there exists $L_f > 0$ for which $\|f'(\mathbf{x})\|_2 \leq L_f$ for all $\mathbf{x} \in \mathbb{R}^n$. Now consider the problem

$$\min_{\mathbf{x}\in\mathbb{R}^n_{++}} \left\{ F(\mathbf{x}) \equiv f(\mathbf{x}) + \sum_{i=1}^{n} \frac{1}{x_i} \right\}$$

with ω chosen as $\omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$. In this case, the mirror descent and mirror-C methods coincide with the projected subgradient and proximal subgradient methods, respectively. It is not possible to employ the projected subgradient method on the problem—it is not even clear what is the feasible set C. If we take it as the open set $\mathbb{R}^n_{++}$, then projections onto C will in general not be in C. In any case, since F is obviously not Lipschitz, no convergence is guaranteed. On the other hand, employing the proximal subgradient method is definitely possible by taking $g(\mathbf{x}) \equiv \sum_{i=1}^n \frac{1}{x_i} + \delta_{\mathbb{R}^n_{++}}$. Both Assumptions 9.20 and 9.21 hold for f, g and $\omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, and in addition g is nonnegative. The resulting method is

$$\mathbf{x}^{k+1} = \mathrm{prox}_{t_k g}\left(\mathbf{x}^k - t_k f'(\mathbf{x}^k)\right).$$

The computation of $\mathrm{prox}_{t_k g}$ amounts to solving n cubic scalar equations. ∎

Example 9.29 (projected subgradient vs. proximal subgradient). Suppose that the underlying space is $\mathbb{R}^n$ endowed with the Euclidean l_2-norm and consider the problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ F(\mathbf{x}) \equiv \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1 + \lambda\|\mathbf{x}\|_1 \right\}, \tag{9.39}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$, and $\lambda > 0$. We will consider two possible methods to solve the problem:

- **projected subgradient** employed on problem (9.39), where here $C = \mathbb{R}^n$. The method takes the form (when making the choice of the subgradient of $\|\mathbf{y}\|_1$ as $\mathrm{sgn}(\mathbf{y})$)

  $$\mathbf{x}^{k+1} = \mathbf{x}^k - t_k(\mathbf{A}^T\mathrm{sgn}(\mathbf{A}\mathbf{x}^k - \mathbf{b}) + \lambda\mathrm{sgn}(\mathbf{x})).$$

  The stepsize is chosen according to Theorem 8.28 as $t_k = \frac{1}{\|F'(\mathbf{x}^k)\|_2\sqrt{k+1}}$.

- **proximal subgradient**, where we take $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_1$ and $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$, so that $F = f + g$. The method then takes the form

  $$\mathbf{x}^{k+1} = \mathrm{prox}_{s_k g}(\mathbf{x}^k - s_k\mathbf{A}^T\mathrm{sgn}(\mathbf{A}\mathbf{x}^k - \mathbf{b})).$$

  Since $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$, it follows that $\mathrm{prox}_{s_k g}$ is a soft thresholding operator. Specifically, by Example 6.8, $\mathrm{prox}_{s_k g} = \mathcal{T}_{\lambda s_k}$, and hence the general update rule becomes

  $$\mathbf{x}^{k+1} = \mathcal{T}_{\lambda s_k}(\mathbf{x}^k - s_k\mathbf{A}^T\mathrm{sgn}(\mathbf{A}\mathbf{x}^k - \mathbf{b})).$$

  The stepsize is chosen as $s_k = \frac{1}{\|f'(\mathbf{x}^k)\|_2\sqrt{k+1}}$.

A priori it seems that the proximal subgradient method should have an advantage over the projected subgradient method since the efficiency estimate bound of the proximal subgradient method depends on L_f, while the corresponding constant for the projected subgradient method depends on the larger constant L_F. This observation is also quite apparent in practice. We created an instance of problem (9.39) with $m = 10, n = 15$ by generating the components of $\mathbf{A}$ and $\mathbf{b}$ independently via a standard normal distribution. The values of $F(\mathbf{x}^k) - F_{\text{opt}}$ for both methods are described in Figure 9.2. Evidently, in this case, the proximal subgradient method is better by orders of magnitude than the projected subgradient method. ∎

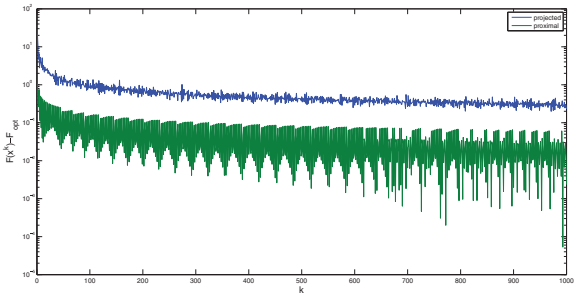

Figure 9.2. *First* 1000 *iterations of the projected and proximal subgradient methods employed on problem* (9.39). *The y-axis describes (in log scale) the quantity* $F(\mathbf{x}^k) - F_{\text{opt}}$.

Chapter 10

The Proximal Gradient Method

Underlying Space: In this chapter, with the exception of Section 10.9, $\mathbb{E}$ is a Euclidean space, meaning a finite dimensional space endowed with an inner product $\langle \cdot, \cdot \rangle$ and the Euclidean norm $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$.

10.1 The Composite Model

In this chapter we will be mostly concerned with the composite model

$$\min_{\mathbf{x} \in \mathbb{E}} \{F(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x})\}, \tag{10.1}$$

where we assume the following.

Assumption 10.1.

(A) $g : \mathbb{E} \to (-\infty, \infty]$ *is proper closed and convex.*

(B) $f : \mathbb{E} \to (-\infty, \infty]$ *is proper and closed,* $\mathrm{dom}(f)$ *is convex,* $\mathrm{dom}(g) \subseteq \mathrm{int}(\mathrm{dom}(f))$, *and* f *is* L_f*-smooth over* $\mathrm{int}(\mathrm{dom}(f))$.

(C) *The optimal set of problem* (10.1) *is nonempty and denoted by* X^*. *The optimal value of the problem is denoted by* F_{opt}.

Three special cases of the general model (10.1) are gathered in the following example.

Example 10.2.

- **Smooth unconstrained minimization**. If $g \equiv 0$ and $\mathrm{dom}(f) = \mathbb{E}$, then (10.1) reduces to the unconstrained smooth minimization problem

$$\min_{\mathbf{x} \in \mathbb{E}} f(\mathbf{x}),$$

where $f : \mathbb{E} \to \mathbb{R}$ is an L_f-smooth function.

- **Convex constrained smooth minimization**. If $g = \delta_C$, where C is a nonempty closed and convex set, then (10.1) amounts to the problem of minimizing a differentiable function over a nonempty closed and convex set:

$$\min_{\mathbf{x}\in C} f(\mathbf{x}),$$

where here f is L_f-smooth over $\mathrm{int}(\mathrm{dom}(f))$ and $C \subseteq \mathrm{int}(\mathrm{dom}(f))$.

- ***l*$_1$-regularized minimization**. Taking $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$ for some $\lambda > 0$, (10.1) amounts to the l_1-regularized problem

$$\min_{\mathbf{x}\in\mathbb{E}}\{f(\mathbf{x}) + \lambda\|\mathbf{x}\|_1\}$$

with f being an L_f-smooth function over the entire space $\mathbb{E}$. ∎

10.2 The Proximal Gradient Method

To understand the idea behind the method for solving (10.1) we are about to study, we begin by revisiting the projected gradient method for solving (10.1) in the case where $g = \delta_C$ with C being a nonempty closed and convex set. In this case, the problem takes the form

$$\min\{f(\mathbf{x}) : \mathbf{x} \in C\}. \tag{10.2}$$

The general update step of the projected gradient method for solving (10.2) takes the form

$$\mathbf{x}^{k+1} = P_C(\mathbf{x}^k - t_k\nabla f(\mathbf{x}^k)),$$

where t_k is the stepsize at iteration k. It is easy to verify that the update step can be also written as (see also Section 9.1 for a similar discussion on the projected subgradient method)

$$\mathbf{x}^{k+1} = \mathrm{argmin}_{\mathbf{x}\in C}\left\{f(\mathbf{x}^k) + \langle\nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k\rangle + \frac{1}{2t_k}\|\mathbf{x} - \mathbf{x}^k\|^2\right\}.$$

That is, the next iterate is the minimizer over C of the sum of the linearization of the smooth part around the current iterate plus a quadratic prox term.

Back to the more general model (10.1), it is natural to generalize the above idea and to define the next iterate as the minimizer of the sum of the linearization of f around $\mathbf{x}^k$, the nonsmooth function g, and a quadratic prox term:

$$\mathbf{x}^{k+1} = \mathrm{argmin}_{\mathbf{x}\in\mathbb{E}}\left\{f(\mathbf{x}^k) + \langle\nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k\rangle + g(\mathbf{x}) + \frac{1}{2t_k}\|\mathbf{x} - \mathbf{x}^k\|^2\right\}. \tag{10.3}$$

After some simple algebraic manipulation and cancellation of constant terms, we obtain that (10.3) can be rewritten as

$$\mathbf{x}^{k+1} = \mathrm{argmin}_{\mathbf{x}\in\mathbb{E}}\left\{t_k g(\mathbf{x}) + \frac{1}{2}\left\|\mathbf{x} - (\mathbf{x}^k - t_k\nabla f(\mathbf{x}^k))\right\|^2\right\},$$

which by the definition of the proximal operator is the same as

$$\mathbf{x}^{k+1} = \mathrm{prox}_{t_k g}(\mathbf{x}^k - t_k\nabla f(\mathbf{x}^k)).$$

The above method is called the *proximal gradient method*, as it consists of a gradient step followed by a proximal mapping. From now on, we will take the stepsizes as $t_k = \frac{1}{L_k}$, leading to the following description of the method.

The Proximal Gradient Method

Initialization: pick $\mathbf{x}^0 \in \mathrm{int}(\mathrm{dom}(f))$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick $L_k > 0$;

(b) set $\mathbf{x}^{k+1} = \mathrm{prox}_{\frac{1}{L_k}g}\left(\mathbf{x}^k - \frac{1}{L_k}\nabla f(\mathbf{x}^k)\right)$.

The general update step of the proximal gradient method can be compactly written as

$$\mathbf{x}^{k+1} = T_{L_k}^{f,g}(\mathbf{x}^k),$$

where $T_L^{f,g} : \mathrm{int}(\mathrm{dom}(f)) \to \mathbb{E}$ $(L > 0)$ is the so-called *prox-grad operator* defined by

$$T_L^{f,g}(\mathbf{x}) \equiv \mathrm{prox}_{\frac{1}{L}g}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right).$$

When the identities of f and g are clear from the context, we will often omit the superscripts f, g and write $T_L(\cdot)$ instead of $T_L^{f,g}(\cdot)$.

Later on, we will consider two stepsize strategies, constant and backtracking, where the meaning of "backtracking" slightly changes under the different settings that will be considered, and hence several backtracking procedures will be defined.

Example 10.3. The table below presents the explicit update step of the proximal gradient method when applied to the three particular models discussed in Example 10.2.[54] The exact assumptions on the models are described in Example 10.2.

Model	Update step	Name of method
$\min_{\mathbf{x}\in\mathbb{E}} f(\mathbf{x})$	$\mathbf{x}^{k+1} = \mathbf{x}^k - t_k\nabla f(\mathbf{x}^k)$	gradient
$\min_{\mathbf{x}\in C} f(\mathbf{x})$	$\mathbf{x}^{k+1} = P_C(\mathbf{x}^k - t_k\nabla f(\mathbf{x}^k))$	projected gradient
$\min_{\mathbf{x}\in\mathbb{E}}\{f(\mathbf{x}) + \lambda\|\mathbf{x}\|_1\}$	$\mathbf{x}^{k+1} = \mathcal{T}_{\lambda t_k}(\mathbf{x}^k - t_k\nabla f(\mathbf{x}^k))$	ISTA

The third method is known as the *iterative shrinkage-thresholding algorithm* (ISTA) in the literature, since at each iteration a soft-thresholding operation (also known as "shrinkage") is performed. ∎

[54] Here we use the facts that $\mathrm{prox}_{t_k g_0} = \mathcal{I}, \mathrm{prox}_{t_k\delta_C} = P_C$ and $\mathrm{prox}_{t_k\lambda\|\cdot\|_1} = \mathcal{T}_{\lambda t_k}$, where $g_0(\mathbf{x}) \equiv 0$.

10.3 Analysis of the Proximal Gradient Method—The Nonconvex Case[55]

10.3.1 Sufficient Decrease

To establish the convergence of the proximal gradient method, we will prove a sufficient decrease lemma for composite functions.

Lemma 10.4 (sufficient decrease lemma). *Suppose that f and g satisfy properties* (A) *and* (B) *of Assumption* 10.1. *Let $F = f + g$ and $T_L \equiv T_L^{f,g}$. Then for any $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$ and $L \in \left(\frac{L_f}{2}, \infty\right)$ the following inequality holds:*

$$F(\mathbf{x}) - F(T_L(\mathbf{x})) \geq \frac{L - \frac{L_f}{2}}{L^2} \left\| G_L^{f,g}(\mathbf{x}) \right\|^2, \tag{10.4}$$

where $G_L^{f,g} : \mathrm{int}(\mathrm{dom}(f)) \to \mathbb{E}$ is the operator defined by $G_L^{f,g}(\mathbf{x}) = L(\mathbf{x} - T_L(\mathbf{x}))$ for all $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$.

Proof. For the sake of simplicity, we use the shorthand notation $\mathbf{x}^+ = T_L(\mathbf{x})$. By the descent lemma (Lemma 5.7), we have that

$$f(\mathbf{x}^+) \leq f(\mathbf{x}) + \left\langle \nabla f(\mathbf{x}), \mathbf{x}^+ - \mathbf{x} \right\rangle + \frac{L_f}{2} \|\mathbf{x} - \mathbf{x}^+\|^2. \tag{10.5}$$

By the second prox theorem (Theorem 6.39), since $\mathbf{x}^+ = \mathrm{prox}_{\frac{1}{L}g}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right)$, we have

$$\left\langle \mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x}) - \mathbf{x}^+, \mathbf{x} - \mathbf{x}^+ \right\rangle \leq \frac{1}{L} g(\mathbf{x}) - \frac{1}{L} g(\mathbf{x}^+),$$

from which it follows that

$$\left\langle \nabla f(\mathbf{x}), \mathbf{x}^+ - \mathbf{x} \right\rangle \leq -L \left\| \mathbf{x}^+ - \mathbf{x} \right\|^2 + g(\mathbf{x}) - g(\mathbf{x}^+),$$

which, combined with (10.5), yields

$$f(\mathbf{x}^+) + g(\mathbf{x}^+) \leq f(\mathbf{x}) + g(\mathbf{x}) + \left(-L + \frac{L_f}{2} \right) \left\| \mathbf{x}^+ - \mathbf{x} \right\|^2.$$

Hence, taking into account the definitions of $\mathbf{x}^+, G_L^{f,g}(\mathbf{x})$ and the identities $F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x}), F(\mathbf{x}^+) = f(\mathbf{x}^+) + g(\mathbf{x}^+)$, the desired result follows. ∎

10.3.2 The Gradient Mapping

The operator $G_L^{f,g}$ that appears in the right-hand side of (10.4) is an important mapping that can be seen as a generalization of the notion of the gradient.

Definition 10.5 (gradient mapping). *Suppose that f and g satisfy properties* (A) *and* (B) *of Assumption* 10.1. *Then the* **gradient mapping** *is the operator*

[55] The analysis of the proximal gradient method in Sections 10.3 and 10.4 mostly follows the presentation of Beck and Teboulle in [18] and [19].

$G_L^{f,g} : \mathrm{int}(\mathrm{dom}(f)) \to \mathbb{E}$ *defined by*

$$G_L^{f,g}(\mathbf{x}) \equiv L\left(\mathbf{x} - T_L^{f,g}(\mathbf{x})\right)$$

for any $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$.

When the identities of f and g will be clear from the context, we will use the notation G_L instead of $G_L^{f,g}$. With the terminology of the gradient mapping, the update step of the proximal gradient method can be rewritten as

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L_k} G_{L_k}(\mathbf{x}^k).$$

In the special case where $L = L_f$, the sufficient decrease inequality (10.4) takes a simpler form.

Corollary 10.6. *Under the setting of Lemma* 10.4*, the following inequality holds for any* $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$*:*

$$F(\mathbf{x}) - F(T_{L_f}(\mathbf{x})) \geq \frac{1}{2L_f} \left\| G_{L_f}(\mathbf{x}) \right\|^2.$$

The next result shows that the gradient mapping is a generalization of the "usual" gradient operator $\mathbf{x} \mapsto \nabla f(\mathbf{x})$ in the sense that they coincide when $g \equiv 0$ and that, for a general g, the points in which the gradient mapping vanishes are the stationary points of the problem of minimizing $f + g$. Recall (see Definition 3.73) that a point $\mathbf{x}^* \in \mathrm{dom}(g)$ is a stationary point of problem (10.1) if and only if $-\nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*)$ and that this condition is a necessary optimality condition for local optimal points (see Theorem 3.72).

Theorem 10.7. *Let* f *and* g *satisfy properties* (A) *and* (B) *of Assumption* 10.1 *and let* $L > 0$*. Then*

(a) $G_L^{f,g_0}(\mathbf{x}) = \nabla f(\mathbf{x})$ *for any* $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$*, where* $g_0(\mathbf{x}) \equiv 0$*;*

(b) *for* $\mathbf{x}^* \in \mathrm{int}(\mathrm{dom}(f))$*, it holds that* $G_L^{f,g}(\mathbf{x}^*) = \mathbf{0}$ *if and only if* $\mathbf{x}^*$ *is a stationary point of problem* (10.1)*.*

Proof. (a) Since $\mathrm{prox}_{\frac{1}{L}g_0}(\mathbf{y}) = \mathbf{y}$ for all $\mathbf{y} \in \mathbb{E}$, it follows that

$$\begin{aligned} G_L^{f,g_0}(\mathbf{x}) &= L(\mathbf{x} - T_L^{f,g_0}(\mathbf{x})) = L\left(\mathbf{x} - \mathrm{prox}_{\frac{1}{L}g_0}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right)\right) \\ &= L\left(\mathbf{x} - \left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right)\right) = \nabla f(\mathbf{x}). \end{aligned}$$

(b) $G_L^{f,g}(\mathbf{x}^*) = \mathbf{0}$ if and only if $\mathbf{x}^* = \mathrm{prox}_{\frac{1}{L}g}\left(\mathbf{x}^* - \frac{1}{L}\nabla f(\mathbf{x}^*)\right)$. By the second prox theorem (Theorem 6.39), the latter relation holds if and only if

$$\mathbf{x}^* - \frac{1}{L}\nabla f(\mathbf{x}^*) - \mathbf{x}^* \in \frac{1}{L}\partial g(\mathbf{x}^*),$$

that is, if and only if

$$-\nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*),$$

which is exactly the condition for stationarity. □

If in addition f is convex, then stationarity is a necessary and sufficient optimality condition (Theorem 3.72(b)), which leads to the following corollary.

Corollary 10.8 (necessary and sufficient optimality condition under convexity). *Let f and g satisfy properties* (A) *and* (B) *of Assumption* 10.1*, and let $L > 0$. Suppose that in addition f is convex. Then for $\mathbf{x}^* \in \mathrm{dom}(g)$, $G_L^{f,g}(\mathbf{x}^*) = \mathbf{0}$ if and only if $\mathbf{x}^*$ is an optimal solution of problem* (10.1)*.*

We can think of the quantity $\|G_L(\mathbf{x})\|$ as an "optimality measure" in the sense that it is always nonnegative, and equal to zero if and only if $\mathbf{x}$ is a stationary point. The next result establishes important monotonicity properties of $\|G_L(\mathbf{x})\|$ w.r.t. the parameter L.

Theorem 10.9 (monotonicity of the gradient mapping). *Suppose that f and g satisfy properties* (A) *and* (B) *of Assumption* 10.1 *and let $G_L \equiv G_L^{f,g}$. Suppose that $L_1 \geq L_2 > 0$. Then*

$$\|G_{L_1}(\mathbf{x})\| \geq \|G_{L_2}(\mathbf{x})\| \tag{10.6}$$

and

$$\frac{\|G_{L_1}(\mathbf{x})\|}{L_1} \leq \frac{\|G_{L_2}(\mathbf{x})\|}{L_2} \tag{10.7}$$

for any $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$.

Proof. Recall that by the second prox theorem (Theorem 6.39), for any $\mathbf{v}, \mathbf{w} \in \mathbb{E}$ and $L > 0$, the following inequality holds:

$$\langle \mathbf{v} - \mathrm{prox}_{\frac{1}{L}g}(\mathbf{v}), \mathrm{prox}_{\frac{1}{L}g}(\mathbf{v}) - \mathbf{w} \rangle \geq \frac{1}{L} g\left(\mathrm{prox}_{\frac{1}{L}g}(\mathbf{v})\right) - \frac{1}{L} g(\mathbf{w}).$$

Plugging $L = L_1, \mathbf{v} = \mathbf{x} - \frac{1}{L_1}\nabla f(\mathbf{x})$, and $\mathbf{w} = \mathrm{prox}_{\frac{1}{L_2}g}\big(\mathbf{x} - \frac{1}{L_2}\nabla f(\mathbf{x})\big) = T_{L_2}(\mathbf{x})$ into the last inequality, it follows that

$$\left\langle \mathbf{x} - \frac{1}{L_1}\nabla f(\mathbf{x}) - T_{L_1}(\mathbf{x}), T_{L_1}(\mathbf{x}) - T_{L_2}(\mathbf{x}) \right\rangle \geq \frac{1}{L_1} g(T_{L_1}(\mathbf{x})) - \frac{1}{L_1} g(T_{L_2}(\mathbf{x}))$$

or

$$\left\langle \frac{1}{L_1} G_{L_1}(\mathbf{x}) - \frac{1}{L_1}\nabla f(\mathbf{x}), \frac{1}{L_2} G_{L_2}(\mathbf{x}) - \frac{1}{L_1} G_{L_1}(\mathbf{x}) \right\rangle \geq \frac{1}{L_1} g(T_{L_1}(\mathbf{x})) - \frac{1}{L_1} g(T_{L_2}(\mathbf{x})).$$

Exchanging the roles of L_1 and L_2 yields the following inequality:

$$\left\langle \frac{1}{L_2} G_{L_2}(\mathbf{x}) - \frac{1}{L_2}\nabla f(\mathbf{x}), \frac{1}{L_1} G_{L_1}(\mathbf{x}) - \frac{1}{L_2} G_{L_2}(\mathbf{x}) \right\rangle \geq \frac{1}{L_2} g(T_{L_2}(\mathbf{x})) - \frac{1}{L_2} g(T_{L_1}(\mathbf{x})).$$

Multiplying the first inequality by L_1 and the second by L_2 and adding them, we obtain

$$\left\langle G_{L_1}(\mathbf{x}) - G_{L_2}(\mathbf{x}), \frac{1}{L_2} G_{L_2}(\mathbf{x}) - \frac{1}{L_1} G_{L_1}(\mathbf{x}) \right\rangle \geq 0,$$

which after some expansion of terms can be seen to be the same as

$$\frac{1}{L_1}\|G_{L_1}(\mathbf{x})\|^2 + \frac{1}{L_2}\|G_{L_2}(\mathbf{x})\|^2 \le \left(\frac{1}{L_1} + \frac{1}{L_2}\right)\langle G_{L_1}(\mathbf{x}), G_{L_2}(\mathbf{x})\rangle.$$

Using the Cauchy–Schwarz inequality, we obtain that

$$\frac{1}{L_1}\|G_{L_1}(\mathbf{x})\|^2 + \frac{1}{L_2}\|G_{L_2}(\mathbf{x})\|^2 \le \left(\frac{1}{L_1} + \frac{1}{L_2}\right)\|G_{L_1}(\mathbf{x})\| \cdot \|G_{L_2}(\mathbf{x})\|. \tag{10.8}$$

Note that if $G_{L_2}(\mathbf{x}) = \mathbf{0}$, then by the last inequality, $G_{L_1}(\mathbf{x}) = \mathbf{0}$, implying that in this case the inequalities (10.6) and (10.7) hold trivially. Assume then that $G_{L_2}(\mathbf{x}) \neq \mathbf{0}$ and define $t = \frac{\|G_{L_1}(\mathbf{x})\|}{\|G_{L_2}(\mathbf{x})\|}$. Then, by (10.8),

$$\frac{1}{L_1}t^2 - \left(\frac{1}{L_1} + \frac{1}{L_2}\right)t + \frac{1}{L_2} \le 0.$$

Since the roots of the quadratic function on the left-hand side of the above inequality are $t = 1, \frac{L_1}{L_2}$, we obtain that

$$1 \le t \le \frac{L_1}{L_2},$$

showing that

$$\|G_{L_2}(\mathbf{x})\| \le \|G_{L_1}(\mathbf{x})\| \le \frac{L_1}{L_2}\|G_{L_2}(\mathbf{x})\|. \qquad \square$$

A straightforward result of the nonexpansivity of the prox operator and the L_f-smoothness of f over $\mathrm{int}(\mathrm{dom}(f))$ is that $G_L(\cdot)$ is Lipschitz continuous with constant $2L + L_f$. Indeed, for any $\mathbf{x}, \mathbf{y} \in \mathrm{int}(\mathrm{dom}(f))$,

$$\begin{aligned}
\|G_L(\mathbf{x}) - G_L(\mathbf{y})\| &= L\left\|\mathbf{x} - \mathrm{prox}_{\frac{1}{L}g}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right) - \mathbf{y} + \mathrm{prox}_{\frac{1}{L}g}\left(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\right)\right\| \\
&\le L\|\mathbf{x} - \mathbf{y}\| + L\left\|\mathrm{prox}_{\frac{1}{L}g}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right) - \mathrm{prox}_{\frac{1}{L}g}\left(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\right)\right\| \\
&\le L\|\mathbf{x} - \mathbf{y}\| + L\left\|\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right) - \left(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\right)\right\| \\
&\le 2L\|\mathbf{x} - \mathbf{y}\| + \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \\
&\le (2L + L_f)\|\mathbf{x} - \mathbf{y}\|.
\end{aligned}$$

In particular, for $L = L_f$, we obtain the inequality

$$\|G_{L_f}(\mathbf{x}) - G_{L_f}(\mathbf{y})\| \le 3L_f\|\mathbf{x} - \mathbf{y}\|.$$

The above discussion is summarized in the following lemma.

Lemma 10.10 (Lipschitz continuity of the gradient mapping). *Let f and g satisfy properties* (A) *and* (B) *of Assumption* 10.1. *Let $G_L = G_L^{f,g}$. Then*

(a) $\|G_L(\mathbf{x}) - G_L(\mathbf{y})\| \le (2L + L_f)\|\mathbf{x} - \mathbf{y}\|$ *for any* $\mathbf{x}, \mathbf{y} \in \mathrm{int}(\mathrm{dom}(f))$*;*

(b) $\|G_{L_f}(\mathbf{x}) - G_{L_f}(\mathbf{y})\| \le 3L_f\|\mathbf{x} - \mathbf{y}\|$ *for any* $\mathbf{x}, \mathbf{y} \in \mathrm{int}(\mathrm{dom}(f))$.

Lemma 10.11 below shows that when f is assumed to be convex and L_f-smooth over the entire space, then the operator $\frac{3}{4L_f}G_{L_f}$ is firmly nonexpansive. A direct consequence is that G_{L_f} is Lipschitz continuous with constant $\frac{4L_f}{3}$.

Lemma 10.11 (firm nonexpansivity of $\frac{3}{4L_f}G_{L_f}$). *Let f be a convex and L_f-smooth function ($L_f > 0$), and let $g : \mathbb{E} \to (-\infty, \infty]$ be a proper closed and convex function. Then*

(a) *the gradient mapping $G_{L_f} \equiv G_{L_f}^{f,g}$ satisfies the relation*

$$\left\langle G_{L_f}(\mathbf{x}) - G_{L_f}(\mathbf{y}), \mathbf{x} - \mathbf{y} \right\rangle \geq \frac{3}{4L_f} \left\| G_{L_f}(\mathbf{x}) - G_{L_f}(\mathbf{y}) \right\|^2 \tag{10.9}$$

for any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$;

(b) $\|G_{L_f}(\mathbf{x}) - G_{L_f}(\mathbf{y})\| \leq \frac{4L_f}{3}\|\mathbf{x} - \mathbf{y}\|$ *for any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$.*

Proof. Part (b) is a direct consequence of (a) and the Cauchy–Schwarz inequality. We will therefore prove (a). To simplify the presentation, we will use the notation $L = L_f$. By the firm nonexpansivity of the prox operator (Theorem 6.42(a)), it follows that for any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$,

$$\left\langle T_L(\mathbf{x}) - T_L(\mathbf{y}), \left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right) - \left(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\right) \right\rangle \geq \|T_L(\mathbf{x}) - T_L(\mathbf{y})\|^2,$$

where $T_L \equiv T_L^{f,g}$ is the prox-grad mapping. Since $T_L = \mathcal{I} - \frac{1}{L}G_L$, we obtain that

$$\begin{aligned}
&\left\langle \left(\mathbf{x} - \frac{1}{L}G_L(\mathbf{x})\right) - \left(\mathbf{y} - \frac{1}{L}G_L(\mathbf{y})\right), \left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right) - \left(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\right) \right\rangle \\
&\geq \left\| \left(\mathbf{x} - \frac{1}{L}G_L(\mathbf{x})\right) - \left(\mathbf{y} - \frac{1}{L}G_L(\mathbf{y})\right) \right\|^2,
\end{aligned}$$

which is the same as

$$\left\langle \left(\mathbf{x} - \frac{1}{L}G_L(\mathbf{x})\right) - \left(\mathbf{y} - \frac{1}{L}G_L(\mathbf{y})\right), (G_L(\mathbf{x}) - \nabla f(\mathbf{x})) - (G_L(\mathbf{y}) - \nabla f(\mathbf{y})) \right\rangle \geq 0.$$

Therefore,

$$\begin{aligned}
\langle G_L(\mathbf{x}) - G_L(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq{}& \frac{1}{L}\|G_L(\mathbf{x}) - G_L(\mathbf{y})\|^2 + \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \\
&- \frac{1}{L}\langle G_L(\mathbf{x}) - G_L(\mathbf{y}), \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) \rangle.
\end{aligned}$$

Since f is L-smooth, it follows from Theorem 5.8 (equivalence between (i) and (iv)) that

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2.$$

Consequently,

$$\begin{aligned}
L\langle G_L(\mathbf{x}) - G_L(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq{}& \|G_L(\mathbf{x}) - G_L(\mathbf{y})\|^2 + \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \\
&- \langle G_L(\mathbf{x}) - G_L(\mathbf{y}), \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) \rangle.
\end{aligned}$$

From the Cauchy–Schwarz inequality we get

$$L\left\langle G_L(\mathbf{x}) - G_L(\mathbf{y}), \mathbf{x} - \mathbf{y}\right\rangle \geq \|G_L(\mathbf{x}) - G_L(\mathbf{y})\|^2 + \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 - \|G_L(\mathbf{x}) - G_L(\mathbf{y})\| \cdot \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|. \quad (10.10)$$

By denoting $\alpha = \|G_L(\mathbf{x}) - G_L(\mathbf{y})\|$ and $\beta = \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|$, the right-hand side of (10.10) reads as $\alpha^2 + \beta^2 - \alpha\beta$ and satisfies

$$\alpha^2 + \beta^2 - \alpha\beta = \frac{3}{4}\alpha^2 + \left(\frac{\alpha}{2} - \beta\right)^2 \geq \frac{3}{4}\alpha^2,$$

which, combined with (10.10), yields the inequality

$$L\left\langle G_L(\mathbf{x}) - G_L(\mathbf{y}), \mathbf{x} - \mathbf{y}\right\rangle \geq \frac{3}{4}\|G_L(\mathbf{x}) - G_L(\mathbf{y})\|^2.$$

Thus, (10.9) holds. ∎

The next result shows a different kind of a monotonicity property of the gradient mapping norm under the setting of Lemma 10.11—the norm of the gradient mapping does not increase if a prox-grad step is employed on its argument.

Lemma 10.12 (monotonicity of the norm of the gradient mapping w.r.t. the prox-grad operator).[56] *Let f be a convex and L_f-smooth function ($L_f > 0$), and let $g : \mathbb{E} \to (-\infty, \infty]$ be a proper closed and convex function. Then for any $\mathbf{x} \in \mathbb{E}$,*

$$\|G_{L_f}(T_{L_f}(\mathbf{x}))\| \leq \|G_{L_f}(\mathbf{x})\|,$$

where $G_{L_f} \equiv G^{f,g}_{L_f}$ and $T_{L_f} \equiv T^{f,g}_{L_f}$.

Proof. Let $\mathbf{x} \in \mathbb{E}$. We will use the shorthand notation $\mathbf{x}^+ = T_{L_f}(\mathbf{x})$. By Theorem 5.8 (equivalence between (i) and (iv)), it follows that

$$\|\nabla f(\mathbf{x}^+) - \nabla f(\mathbf{x})\|^2 \leq L_f\langle \nabla f(\mathbf{x}^+) - \nabla f(\mathbf{x}), \mathbf{x}^+ - \mathbf{x}\rangle. \quad (10.11)$$

Denoting $\mathbf{a} = \nabla f(\mathbf{x}^+) - \nabla f(\mathbf{x})$ and $\mathbf{b} = \mathbf{x}^+ - \mathbf{x}$, inequality (10.11) can be rewritten as $\|\mathbf{a}\|^2 \leq L_f\langle \mathbf{a}, \mathbf{b}\rangle$, which is the same as

$$\left\|\mathbf{a} - \frac{L_f}{2}\mathbf{b}\right\|^2 \leq \frac{L_f^2}{4}\|\mathbf{b}\|^2$$

and as

$$\left\|\frac{1}{L_f}\mathbf{a} - \frac{1}{2}\mathbf{b}\right\| \leq \frac{1}{2}\|\mathbf{b}\|.$$

Using the triangle inequality,

$$\left\|\frac{1}{L_f}\mathbf{a} - \mathbf{b}\right\| \leq \left\|\frac{1}{L_f}\mathbf{a} - \mathbf{b} + \frac{1}{2}\mathbf{b}\right\| + \frac{1}{2}\|\mathbf{b}\| \leq \|\mathbf{b}\|.$$

[56] Lemma 10.12 is a minor variation of Lemma 2.4 from Necoara and Patrascu [88].

Plugging the expressions for $\mathbf{a}$ and $\mathbf{b}$ into the above inequality, we obtain that

$$\left\| \mathbf{x} - \frac{1}{L_f}\nabla f(\mathbf{x}) - \mathbf{x}^+ + \frac{1}{L_f}\nabla f(\mathbf{x}^+) \right\| \leq \|\mathbf{x}^+ - \mathbf{x}\|.$$

Combining the above inequality with the nonexpansivity of the prox operator (Theorem 6.42(b)), we finally obtain

$$\begin{aligned}
\|G_{L_f}(T_{L_f}(\mathbf{x}))\| &= \|G_{L_f}(\mathbf{x}^+)\| = L_f\|\mathbf{x}^+ - T_{L_f}(\mathbf{x}^+)\| = L_f\|T_{L_f}(\mathbf{x}) - T_{L_f}(\mathbf{x}^+)\| \\
&= L_f \left\| \operatorname{prox}_{\frac{1}{L_f}g}\left(\mathbf{x} - \frac{1}{L_f}\nabla f(\mathbf{x})\right) - \operatorname{prox}_{\frac{1}{L_f}g}\left(\mathbf{x}^+ - \frac{1}{L_f}\nabla f(\mathbf{x}^+)\right) \right\| \\
&\leq L_f \left\| \mathbf{x} - \frac{1}{L_f}\nabla f(\mathbf{x}) - \mathbf{x}^+ + \frac{1}{L_f}\nabla f(\mathbf{x}^+) \right\| \\
&\leq L_f\|\mathbf{x}^+ - \mathbf{x}\| = L_f\|T_{L_f}(\mathbf{x}) - \mathbf{x}\| = \|G_{L_f}(\mathbf{x})\|,
\end{aligned}$$

which is the desired result. ∎

10.3.3 Convergence of the Proximal Gradient Method—The Nonconvex Case

We will now analyze the convergence of the proximal gradient method under the validity of Assumption 10.1. Note that we do not assume at this stage that f is convex. The two stepsize strategies that will be considered are constant and backtracking.

- **Constant.** $L_k = \bar{L} \in \left(\frac{L_f}{2}, \infty\right)$ for all k.
- **Backtracking procedure B1.** The procedure requires three parameters (s, γ, η), where $s > 0, \gamma \in (0,1)$, and $\eta > 1$. The choice of L_k is done as follows. First, L_k is set to be equal to the initial guess s. Then, while

$$F(\mathbf{x}^k) - F(T_{L_k}(\mathbf{x}^k)) < \frac{\gamma}{L_k}\|G_{L_k}(\mathbf{x}^k)\|^2,$$

we set $L_k := \eta L_k$. In other words, L_k is chosen as $L_k = s\eta^{i_k}$, where i_k is the smallest nonnegative integer for which the condition

$$F(\mathbf{x}^k) - F(T_{s\eta^{i_k}}(\mathbf{x}^k)) \geq \frac{\gamma}{s\eta^{i_k}}\|G_{s\eta^{i_k}}(\mathbf{x}^k)\|^2$$

is satisfied.

Remark 10.13. *Note that the backtracking procedure is finite under Assumption* 10.1. *Indeed, plugging* $\mathbf{x} = \mathbf{x}^k$ *into* (10.4)*, we obtain*

$$F(\mathbf{x}^k) - F(T_L(\mathbf{x}^k)) \geq \frac{L - \frac{L_f}{2}}{L^2}\left\|G_L(\mathbf{x}^k)\right\|^2. \tag{10.12}$$

If $L \geq \frac{L_f}{2(1-\gamma)}$, *then* $\frac{L-\frac{L_f}{2}}{L} \geq \gamma$, *and hence, by* (10.12), *the inequality*

$$F(\mathbf{x}^k) - F(T_L(\mathbf{x}^k)) \geq \frac{\gamma}{L}\|G_L(\mathbf{x}^k)\|^2$$

holds, implying that the backtracking procedure must end when $L_k \geq \frac{L_f}{2(1-\gamma)}$.

We can also compute an upper bound on L_k*: either* L_k *is equal to* s, *or the backtracking procedure is invoked, meaning that* $\frac{L_k}{\eta}$ *did not satisfy the backtracking condition, which by the above discussion implies that* $\frac{L_k}{\eta} < \frac{L_f}{2(1-\gamma)}$, *so that* $L_k < \frac{\eta L_f}{2(1-\gamma)}$. *To summarize, in the backtracking procedure* B1, *the parameter* L_k *satisfies*

$$L_k \leq \max\left\{s, \frac{\eta L_f}{2(1-\gamma)}\right\}. \tag{10.13}$$

The convergence of the proximal gradient method in the nonconvex case is heavily based on the sufficient decrease lemma (Lemma 10.4). We begin with the following lemma showing that consecutive function values of the sequence generated by the proximal gradient method decrease by at least a constant times the squared norm of the gradient mapping.

Lemma 10.14 (sufficient decrease of the proximal gradient method). *Suppose that Assumption* 10.1 *holds. Let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by the proximal gradient method for solving problem* (10.1) *with either a constant stepsize defined by* $L_k = \bar{L} \in \left(\frac{L_f}{2}, \infty\right)$ *or with a stepsize chosen by the backtracking procedure* B1 *with parameters* (s, γ, η), *where* $s > 0, \gamma \in (0,1), \eta > 1$. *Then for any* $k \geq 0$,

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq M\|G_d(\mathbf{x}^k)\|^2, \tag{10.14}$$

where

$$M = \begin{cases} \frac{\bar{L}-\frac{L_f}{2}}{\left(\bar{L}\right)^2}, & \textit{constant stepsize}, \\ \frac{\gamma}{\max\left\{s, \frac{\eta L_f}{2(1-\gamma)}\right\}}, & \textit{backtracking}, \end{cases} \tag{10.15}$$

and

$$d = \begin{cases} \bar{L}, & \textit{constant stepsize}, \\ s, & \textit{backtracking}. \end{cases} \tag{10.16}$$

Proof. The result for the constant stepsize setting follows by plugging $L = \bar{L}$ and $\mathbf{x} = \mathbf{x}^k$ into (10.4). As for the case where the backtracking procedure is used, by its definition we have

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq \frac{\gamma}{L_k}\|G_{L_k}(\mathbf{x}^k)\|^2 \geq \frac{\gamma}{\max\left\{s, \frac{\eta L_f}{2(1-\gamma)}\right\}}\|G_{L_k}(\mathbf{x}^k)\|^2,$$

where the last inequality follows from the upper bound on L_k given in (10.13). The result for the case where the backtracking procedure is invoked now follows by

the monotonicity property of the gradient mapping (Theorem 10.9) along with the bound $L_k \geq s$, which imply the inequality $\|G_{L_k}(\mathbf{x}^k)\| \geq \|G_s(\mathbf{x}^k)\|$. ∎

We are now ready to prove the convergence of the norm of the gradient mapping to zero and that limit points of the sequence generated by the method are stationary points of problem (10.1).

Theorem 10.15 (convergence of the proximal gradient method—nonconvex case). *Suppose that Assumption* 10.1 *holds and let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by the proximal gradient method for solving problem* (10.1) *either with a constant stepsize defined by* $L_k = \bar{L} \in \left(\frac{L_f}{2}, \infty\right)$ *or with a stepsize chosen by the backtracking procedure* B1 *with parameters* (s, γ, η)*, where* $s > 0, \gamma \in (0,1)$*, and* $\eta > 1$*. Then*

(a) *the sequence* $\{F(\mathbf{x}^k)\}_{k\geq 0}$ *is nonincreasing. In addition,* $F(\mathbf{x}^{k+1}) < F(\mathbf{x}^k)$ *if and only if* $\mathbf{x}^k$ *is not a stationary point of* (10.1)*;*

(b) $G_d(\mathbf{x}^k) \to \mathbf{0}$ *as* $k \to \infty$*, where* d *is given in* (10.16)*;*

(c)
$$\min_{n=0,1,\ldots,k} \|G_d(\mathbf{x}^n)\| \leq \frac{\sqrt{F(\mathbf{x}^0) - F_{\text{opt}}}}{\sqrt{M(k+1)}}, \tag{10.17}$$
where M *is given in* (10.15)*;*

(d) *all limit points of the sequence* $\{\mathbf{x}^k\}_{k\geq 0}$ *are stationary points of problem* (10.1)*.*

Proof. (a) By Lemma 10.14 we have that
$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq M\|G_d(\mathbf{x}^k)\|^2, \tag{10.18}$$
from which it readily follows that $F(\mathbf{x}^k) \geq F(\mathbf{x}^{k+1})$. If $\mathbf{x}^k$ is not a stationary point of problem (10.1), then $G_d(\mathbf{x}^k) \neq \mathbf{0}$, and hence, by (10.18), $F(\mathbf{x}^k) > F(\mathbf{x}^{k+1})$. If $\mathbf{x}^k$ is a stationary point of problem (10.1), then $G_{L_k}(\mathbf{x}^k) = \mathbf{0}$, from which it follows that $\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L_k} G_{L_k}(\mathbf{x}^k) = \mathbf{x}^k$, and consequently $F(\mathbf{x}^k) = F(\mathbf{x}^{k+1})$.

(b) Since the sequence $\{F(\mathbf{x}^k)\}_{k\geq 0}$ is nonincreasing and bounded below, it converges. Thus, in particular, $F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \to 0$ as $k \to \infty$, which, combined with (10.18), implies that $\|G_d(\mathbf{x}^k)\| \to 0$ as $k \to \infty$.

(c) Summing the inequality
$$F(\mathbf{x}^n) - F(\mathbf{x}^{n+1}) \geq M\|G_d(\mathbf{x}^n)\|^2$$
over $n = 0, 1, \ldots, k$, we obtain
$$F(\mathbf{x}^0) - F(\mathbf{x}^{k+1}) \geq M \sum_{n=0}^{k} \|G_d(\mathbf{x}^n)\|^2 \geq M(k+1) \min_{n=0,1,\ldots,k} \|G_d(\mathbf{x}^n)\|^2.$$
Using the fact that $F(\mathbf{x}^{k+1}) \geq F_{\text{opt}}$, the inequality (10.17) follows.

(d) Let $\bar{\mathbf{x}}$ be a limit point of $\{\mathbf{x}^k\}_{k\geq 0}$. Then there exists a subsequence $\{\mathbf{x}^{k_j}\}_{j\geq 0}$ converging to $\bar{\mathbf{x}}$. For any $j \geq 0$,

$$\|G_d(\bar{\mathbf{x}})\| \leq \|G_d(\mathbf{x}^{k_j}) - G_d(\bar{\mathbf{x}})\| + \|G_d(\mathbf{x}^{k_j})\| \leq (2d + L_f)\|\mathbf{x}^{k_j} - \bar{\mathbf{x}}\| + \|G_d(\mathbf{x}^{k_j})\|, \tag{10.19}$$

where Lemma 10.10(a) was used in the second inequality. Since the right-hand side of (10.19) goes to 0 as $j \to \infty$, it follows that $G_d(\bar{\mathbf{x}}) = \mathbf{0}$, which by Theorem 10.7(b) implies that $\bar{\mathbf{x}}$ is a stationary point of problem (10.1). ∎

10.4 Analysis of the Proximal Gradient Method—The Convex Case

10.4.1 The Fundamental Prox-Grad Inequality

The analysis of the proximal gradient method in the case where f is convex is based on the following key inequality (which actually does not assume that f is convex).

Theorem 10.16 (fundamental prox-grad inequality). *Suppose that f and g satisfy properties* (A) *and* (B) *of Assumption* 10.1. *For any $\mathbf{x} \in \mathbb{E}$, $\mathbf{y} \in \operatorname{int}(\operatorname{dom}(f))$ and $L > 0$ satisfying*

$$f(T_L(\mathbf{y})) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), T_L(\mathbf{y}) - \mathbf{y}\rangle + \frac{L}{2}\|T_L(\mathbf{y}) - \mathbf{y}\|^2, \tag{10.20}$$

it holds that

$$F(\mathbf{x}) - F(T_L(\mathbf{y})) \geq \frac{L}{2}\|\mathbf{x} - T_L(\mathbf{y})\|^2 - \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 + \ell_f(\mathbf{x}, \mathbf{y}), \tag{10.21}$$

where

$$\ell_f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle.$$

Proof. Consider the function

$$\varphi(\mathbf{u}) = f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{u} - \mathbf{y}\rangle + g(\mathbf{u}) + \frac{L}{2}\|\mathbf{u} - \mathbf{y}\|^2.$$

Since φ is an L-strongly convex function and $T_L(\mathbf{y}) = \operatorname{argmin}_{\mathbf{u}\in\mathbb{E}}\varphi(\mathbf{u})$, it follows by Theorem 5.25(b) that

$$\varphi(\mathbf{x}) - \varphi(T_L(\mathbf{y})) \geq \frac{L}{2}\|\mathbf{x} - T_L(\mathbf{y})\|^2. \tag{10.22}$$

Note that by (10.20),

$$\begin{aligned}\varphi(T_L(\mathbf{y})) &= f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), T_L(\mathbf{y}) - \mathbf{y}\rangle + \frac{L}{2}\|T_L(\mathbf{y}) - \mathbf{y}\|^2 + g(T_L(\mathbf{y})) \\ &\geq f(T_L(\mathbf{y})) + g(T_L(\mathbf{y})) = F(T_L(\mathbf{y})),\end{aligned}$$

and thus (10.22) implies that for any $\mathbf{x} \in \mathbb{E}$,

$$\varphi(\mathbf{x}) - F(T_L(\mathbf{y})) \geq \frac{L}{2}\|\mathbf{x} - T_L(\mathbf{y})\|^2.$$

Plugging the expression for $\varphi(\mathbf{x})$ into the above inequality, we obtain

$$f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + g(\mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 - F(T_L(\mathbf{y})) \geq \frac{L}{2}\|\mathbf{x} - T_L(\mathbf{y})\|^2,$$

which is the same as the desired result:

$$\begin{aligned} F(\mathbf{x}) - F(T_L(\mathbf{y})) \geq \frac{L}{2}\|\mathbf{x} - T_L(\mathbf{y})\|^2 - \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 \\ + f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle. \end{aligned}$$ ∎

Remark 10.17. *Obviously, by the descent lemma,* (10.20) *is satisfied for* $L = L_f$*, and hence, for any* $\mathbf{x} \in \mathbb{E}$ *and* $\mathbf{y} \in \operatorname{int}(\operatorname{dom}(f))$*, the inequality*

$$F(\mathbf{x}) - F(T_{L_f}(\mathbf{y})) \geq \frac{L_f}{2}\|\mathbf{x} - T_{L_f}(\mathbf{y})\|^2 - \frac{L_f}{2}\|\mathbf{x} - \mathbf{y}\|^2 + \ell_f(\mathbf{x}, \mathbf{y})$$

holds.

A direct consequence of Theorem 10.16 is another version of the sufficient decrease lemma (Lemma 10.4). This is accomplished by substituting $\mathbf{y} = \mathbf{x}$ in the fundamental prox-grad inequality.

Corollary 10.18 (sufficient decrease lemma—second version). *Suppose that* f *and* g *satisfy properties* (A) *and* (B) *of Assumption* 10.1. *For any* $\mathbf{x} \in \operatorname{int}(\operatorname{dom}(f))$ *for which*

$$f(T_L(\mathbf{x})) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), T_L(\mathbf{x}) - \mathbf{x} \rangle + \frac{L}{2}\|T_L(\mathbf{x}) - \mathbf{x}\|^2,$$

it holds that

$$F(\mathbf{x}) - F(T_L(\mathbf{x})) \geq \frac{1}{2L}\|G_L(\mathbf{x})\|^2.$$

10.4.2 Stepsize Strategies in the Convex Case

When f is also convex, we will consider, as in the nonconvex case, both constant and backtracking stepsize strategies. The backtracking procedure, which we will refer to as "backtracking procedure B2," will be slightly different than the one considered in the nonconvex case, and it will aim to find a constant L_k satisfying

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L_k}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2. \tag{10.23}$$

In the special case where $g \equiv 0$, the proximal gradient method reduces to the gradient method $\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L_k}\nabla f(\mathbf{x}^k)$, and condition (10.23) reduces to

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2L_k}\|\nabla f(\mathbf{x}^k)\|^2,$$

which is similar to the sufficient decrease condition described in Lemma 10.4, and this is why condition (10.23) can also be viewed as a "sufficient decrease condition."

- **Constant.** $L_k = L_f$ for all k.
- **Backtracking procedure B2.** The procedure requires two parameters (s, η), where $s > 0$ and $\eta > 1$. Define $L_{-1} = s$. At iteration k $(k \geq 0)$ the choice of L_k is done as follows. First, L_k is set to be equal to L_{k-1}. Then, while

$$f(T_{L_k}(\mathbf{x}^k)) > f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), T_{L_k}(\mathbf{x}^k) - \mathbf{x}^k \rangle + \frac{L_k}{2} \|T_{L_k}(\mathbf{x}^k) - \mathbf{x}^k\|^2,$$

we set $L_k := \eta L_k$. In other words, L_k is chosen as $L_k = L_{k-1}\eta^{i_k}$, where i_k is the smallest nonnegative integer for which the condition

$$f(T_{L_{k-1}\eta^{i_k}}(\mathbf{x}^k)) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), T_{L_{k-1}\eta^{i_k}}(\mathbf{x}^k) - \mathbf{x}^k \rangle + \frac{L_k}{2} \|T_{L_{k-1}\eta^{i_k}}(\mathbf{x}^k) - \mathbf{x}^k\|^2$$

is satisfied.

Remark 10.19 (upper and lower bounds on L_k). *Under Assumption* 10.1 *and by the descent lemma (Lemma* 5.7*), it follows that both stepsize rules ensure that the sufficient decrease condition* (10.23) *is satisfied at each iteration. In addition, the constants L_k that the backtracking procedure* B2 *produces satisfy the following bounds for all $k \geq 0$:*

$$s \leq L_k \leq \max\{\eta L_f, s\}. \tag{10.24}$$

The inequality $s \leq L_k$ is obvious. To understand the inequality $L_k \leq \max\{\eta L_f, s\}$, note that there are two options. Either $L_k = s$ or $L_k > s$, and in the latter case there exists an index $0 \leq k' \leq k$ for which the inequality (10.23) *is not satisfied with $k = k'$ and $\frac{L_k}{\eta}$ replacing L_k. By the descent lemma, this implies in particular that $\frac{L_k}{\eta} < L_f$, and we have thus shown that $L_k \leq \max\{\eta L_f, s\}$. We also note that the bounds on L_k can be rewritten as*

$$\beta L_f \leq L_k \leq \alpha L_f,$$

where

$$\alpha = \begin{cases} 1, & \textit{constant}, \\ \max\left\{\eta, \frac{s}{L_f}\right\}, & \textit{backtracking}, \end{cases} \qquad \beta = \begin{cases} 1, & \textit{constant}, \\ \frac{s}{L_f}, & \textit{backtracking}. \end{cases} \tag{10.25}$$

Remark 10.20 (monotonicity of the proximal gradient method). *Since condition* (10.23) *holds for both stepsize rules, for any $k \geq 0$, we can invoke the fundamental prox-grad inequality* (10.21) *with $\mathbf{y} = \mathbf{x} = \mathbf{x}^k, L = L_k$ and obtain the inequality*

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq \frac{L_k}{2} \|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2,$$

which in particular implies that $F(\mathbf{x}^k) \geq F(\mathbf{x}^{k+1})$, meaning that the method produces a nonincreasing sequence of function values.

10.4.3 Convergence Analysis in the Convex Case

We will assume in addition to Assumption 10.1 that f is convex. We begin by establishing an $O(1/k)$ rate of convergence of the generated sequence of function values to the optimal value. Such rate of convergence is called a *sublinear rate*. This is of course an improvement over the $O(1/\sqrt{k})$ rate that was established for the projected subgradient and mirror descent methods. It is also not particularly surprising that an improved rate of convergence can be established since additional properties are assumed on the objective function.

Theorem 10.21 ($O(1/k)$ rate of convergence of proximal gradient). *Suppose that Assumption* 10.1 *holds and that in addition f is convex. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method for solving problem* (10.1) *with either a constant stepsize rule in which $L_k \equiv L_f$ for all $k \geq 0$ or the backtracking procedure* B2. *Then for any $\mathbf{x}^* \in X^*$ and $k \geq 0$,*

$$F(\mathbf{x}^k) - F_{\text{opt}} \leq \frac{\alpha L_f \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2k}, \tag{10.26}$$

where $\alpha = 1$ in the constant stepsize setting and $\alpha = \max\left\{\eta, \frac{s}{L_f}\right\}$ if the backtracking rule is employed.

Proof. For any $n \geq 0$, substituting $L = L_n$, $\mathbf{x} = \mathbf{x}^*$, and $\mathbf{y} = \mathbf{x}^n$ in the fundamental prox-grad inequality (10.21) and taking into account the fact that in both stepsize rules condition (10.20) is satisfied, we obtain

$$\begin{aligned}\frac{2}{L_n}(F(\mathbf{x}^*) - F(\mathbf{x}^{n+1})) &\geq \|\mathbf{x}^* - \mathbf{x}^{n+1}\|^2 - \|\mathbf{x}^* - \mathbf{x}^n\|^2 + \frac{2}{L_n}\ell_f(\mathbf{x}^*, \mathbf{x}^n) \\ &\geq \|\mathbf{x}^* - \mathbf{x}^{n+1}\|^2 - \|\mathbf{x}^* - \mathbf{x}^n\|^2,\end{aligned}$$

where the convexity of f was used in the last inequality. Summing the above inequality over $n = 0, 1, \ldots, k-1$ and using the bound $L_n \leq \alpha L_f$ for all $n \geq 0$ (see Remark 10.19), we obtain

$$\frac{2}{\alpha L_f}\sum_{n=0}^{k-1}(F(\mathbf{x}^*) - F(\mathbf{x}^{n+1})) \geq \|\mathbf{x}^* - \mathbf{x}^k\|^2 - \|\mathbf{x}^* - \mathbf{x}^0\|^2.$$

Thus,

$$\sum_{n=0}^{k-1}(F(\mathbf{x}^{n+1}) - F_{\text{opt}}) \leq \frac{\alpha L_f}{2}\|\mathbf{x}^* - \mathbf{x}^0\|^2 - \frac{\alpha L_f}{2}\|\mathbf{x}^* - \mathbf{x}^k\|^2 \leq \frac{\alpha L_f}{2}\|\mathbf{x}^* - \mathbf{x}^0\|^2.$$

By the monotonicity of $\{F(\mathbf{x}^n)\}_{n\geq 0}$ (see Remark 10.20), we can conclude that

$$k(F(\mathbf{x}^k) - F_{\text{opt}}) \leq \sum_{n=0}^{k-1}(F(\mathbf{x}^{n+1}) - F_{\text{opt}}) \leq \frac{\alpha L_f}{2}\|\mathbf{x}^* - \mathbf{x}^0\|^2.$$

Consequently,

$$F(\mathbf{x}^k) - F_{\text{opt}} \leq \frac{\alpha L_f \|\mathbf{x}^* - \mathbf{x}^0\|^2}{2k}. \qquad \square$$

Remark 10.22. *Note that we did not utilize in the proof of Theorem* 10.21 *the fact that procedure* B2 *produces a nondecreasing sequence of constants* $\{L_k\}_{k\geq 0}$. *This implies in particular that the monotonicity of this sequence of constants is not essential, and we can actually prove the same convergence rate for any backtracking procedure that guarantees the validity of condition* (10.23) *and the bound* $L_k \leq \alpha L_f$.

We can also prove that the generated sequence is Fejér monotone, from which convergence of the sequence to an optimal solution readily follows.

Theorem 10.23 (Fejér monotonicity of the sequence generated by the proximal gradient method). *Suppose that Assumption* 10.1 *holds and that in addition* f *is convex. Let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by the proximal gradient method for solving problem* (10.1) *with either a constant stepsize rule in which* $L_k \equiv L_f$ *for all* $k \geq 0$ *or the backtracking procedure* B2. *Then for any* $\mathbf{x}^* \in X^*$ *and* $k \geq 0$,

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\| \leq \|\mathbf{x}^k - \mathbf{x}^*\|. \tag{10.27}$$

Proof. We will repeat some of the arguments used in the proof of Theorem 10.21. Substituting $L = L_k$, $\mathbf{x} = \mathbf{x}^*$, and $\mathbf{y} = \mathbf{x}^k$ in the fundamental prox-grad inequality (10.21) and taking into account the fact that in both stepsize rules condition (10.20) is satisfied, we obtain

$$\begin{aligned}\frac{2}{L_k}(F(\mathbf{x}^*) - F(\mathbf{x}^{k+1})) &\geq \|\mathbf{x}^* - \mathbf{x}^{k+1}\|^2 - \|\mathbf{x}^* - \mathbf{x}^k\|^2 + \frac{2}{L_k}\ell_f(\mathbf{x}^*, \mathbf{x}^k)\\ &\geq \|\mathbf{x}^* - \mathbf{x}^{k+1}\|^2 - \|\mathbf{x}^* - \mathbf{x}^k\|^2,\end{aligned}$$

where the convexity of f was used in the last inequality. The result (10.27) now follows by the inequality $F(\mathbf{x}^*) - F(\mathbf{x}^{k+1}) \leq 0$. ∎

Thanks to the Fejér monotonicity property, we can now establish the convergence of the sequence generated by the proximal gradient method.

Theorem 10.24 (convergence of the sequence generated by the proximal gradient method). *Suppose that Assumption* 10.1 *holds and that in addition* f *is convex. Let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by the proximal gradient method for solving problem* (10.1) *with either a constant stepsize rule in which* $L_k \equiv L_f$ *for all* $k \geq 0$ *or the backtracking procedure* B2. *Then the sequence* $\{\mathbf{x}^k\}_{k\geq 0}$ *converges to an optimal solution of problem* (10.1).

Proof. By Theorem 10.23, the sequence is Fejér monotone w.r.t. X^*. Therefore, by Theorem 8.16, to show convergence to a point in X^*, it is enough to show that any limit point of the sequence $\{\mathbf{x}^k\}_{k\geq 0}$ is necessarily in X^*. Let then $\tilde{\mathbf{x}}$ be a limit point of the sequence. Then there exists a subsequence $\{\mathbf{x}^{k_j}\}_{j\geq 0}$ converging to $\tilde{\mathbf{x}}$. By Theorem 10.21,

$$F(\mathbf{x}^{k_j}) \to F_{\text{opt}} \text{ as } j \to \infty. \tag{10.28}$$

Since F is closed, it is also lower semicontinuous, and hence $F(\tilde{\mathbf{x}}) \leq \lim_{j\to\infty} F(\mathbf{x}^{k_j}) = F_{\text{opt}}$, implying that $\tilde{\mathbf{x}} \in X^*$. ∎

To derive a complexity result for the proximal gradient method, we will assume that $\|\mathbf{x}^0 - \mathbf{x}^*\| \leq R$ for some $\mathbf{x}^* \in X^*$ and some constant $R > 0$; for example, if $\mathrm{dom}(g)$ is bounded, then R might be taken as its diameter. By inequality (10.26) it follows that in order to obtain an ε-optimal solution of problem (10.1), it is enough to require that

$$\frac{\alpha L_f R^2}{2k} \leq \varepsilon,$$

which is the same as

$$k \geq \frac{\alpha L_f R^2}{2\varepsilon}.$$

Thus, to obtain an ε-optimal solution, an order of $\frac{1}{\varepsilon}$ iterations is required, which is an improvement of the result for the projected subgradient method in which an order of $\frac{1}{\varepsilon^2}$ iterations is needed (see, for example, Theorem 8.18). We summarize the above observations in the following theorem.

Theorem 10.25 (complexity of the proximal gradient method). *Under the setting of Theorem* 10.21*, for any k satisfying*

$$k \geq \left\lceil \frac{\alpha L_f R^2}{2\varepsilon} \right\rceil,$$

it holds that $F(\mathbf{x}^k) - F_{\mathrm{opt}} \leq \varepsilon$, where R is an upper bound on $\|\mathbf{x}^* - \mathbf{x}^0\|$ for some $\mathbf{x}^* \in X^*$.

In the nonconvex case (meaning when f is not necessarily convex), an $O(1/\sqrt{k})$ rate of convergence of the norm of the gradient mapping was established in Theorem 10.15(c). We will now show that with the additional convexity assumption on f, this rate can be improved to $O(1/k)$.

Theorem 10.26 ($O(1/k)$ rate of convergence of the minimal norm of the gradient mapping). *Suppose that Assumption* 10.1 *holds and that in addition f is convex. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method for solving problem* (10.1) *with either a constant stepsize rule in which $L_k \equiv L_f$ for all $k \geq 0$ or the backtracking procedure* B2*. Then for any $\mathbf{x}^* \in X^*$ and $k \geq 1$,*

$$\min_{n=0,1,\ldots,k} \|G_{\alpha L_f}(\mathbf{x}^n)\| \leq \frac{2\alpha^{1.5} L_f \|\mathbf{x}^0 - \mathbf{x}^*\|}{\sqrt{\beta} k}, \tag{10.29}$$

where $\alpha = \beta = 1$ in the constant stepsize setting and $\alpha = \max\left\{\eta, \frac{s}{L_f}\right\}, \beta = \frac{s}{L_f}$ if the backtracking rule is employed.

Proof. By the sufficient decrease lemma (Corollary 10.18), for any $n \geq 0$,

$$F(\mathbf{x}^n) - F(\mathbf{x}^{n+1}) = F(\mathbf{x}^n) - F(T_{L_n}(\mathbf{x}^n)) \geq \frac{1}{2L_n}\|G_{L_n}(\mathbf{x}^n)\|^2. \tag{10.30}$$

By Theorem 10.9 and the fact that $\beta L_f \leq L_n \leq \alpha L_f$ (see Remark 10.19), it follows that

$$\frac{1}{2L_n}\|G_{L_n}(\mathbf{x}^n)\|^2 = \frac{L_n}{2}\frac{\|G_{L_n}(\mathbf{x}^n)\|^2}{L_n^2} \geq \frac{\beta L_f}{2}\frac{\|G_{\alpha L_f}(\mathbf{x}^n)\|^2}{\alpha^2 L_f^2} = \frac{\beta}{2\alpha^2 L_f}\|G_{\alpha L_f}(\mathbf{x}^n)\|^2. \tag{10.31}$$

Therefore, combining (10.30) and (10.31),

$$F(\mathbf{x}^n) - F_{\text{opt}} \geq F(\mathbf{x}^{n+1}) - F_{\text{opt}} + \frac{\beta}{2\alpha^2 L_f}\|G_{\alpha L_f}(\mathbf{x}^n)\|^2. \tag{10.32}$$

Let p be a positive integer. Summing (10.32) over $n = p, p+1, \ldots, 2p-1$ yields

$$F(\mathbf{x}^p) - F_{\text{opt}} \geq F(\mathbf{x}^{2p}) - F_{\text{opt}} + \frac{\beta}{2\alpha^2 L_f}\sum_{n=p}^{2p-1}\|G_{\alpha L_f}(\mathbf{x}^n)\|^2. \tag{10.33}$$

By Theorem 10.21, $F(\mathbf{x}^p) - F_{\text{opt}} \leq \frac{\alpha L_f \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2p}$, which, combined with the fact that $F(\mathbf{x}^{2p}) - F_{\text{opt}} \geq 0$ and (10.33), implies

$$\frac{\beta p}{2\alpha^2 L_f}\min_{n=0,1,\ldots,2p-1}\|G_{\alpha L_f}(\mathbf{x}^n)\|^2 \leq \frac{\beta}{2\alpha^2 L_f}\sum_{n=p}^{2p-1}\|G_{\alpha L_f}(\mathbf{x}^n)\|^2 \leq \frac{\alpha L_f \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2p}.$$

Thus,

$$\min_{n=0,1,\ldots,2p-1}\|G_{\alpha L_f}(\mathbf{x}^n)\|^2 \leq \frac{\alpha^3 L_f^2 \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{\beta p^2} \tag{10.34}$$

and also

$$\min_{n=0,1,\ldots,2p}\|G_{\alpha L_f}(\mathbf{x}^n)\|^2 \leq \frac{\alpha^3 L_f^2 \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{\beta p^2}. \tag{10.35}$$

We conclude that for any $k \geq 1$,

$$\min_{n=0,1,\ldots,k}\|G_{\alpha L_f}(\mathbf{x}^n)\|^2 \leq \frac{\alpha^3 L_f^2 \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{\beta \min\{(k/2)^2, ((k+1)/2)^2\}} = \frac{4\alpha^3 L_f^2 \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{\beta k^2}. \qquad \square$$

When we assume further that f is L_f-smooth over the entire space $\mathbb{E}$, we can use Lemma 10.12 to obtain an improved result in the case of a constant stepsize.

Theorem 10.27 ($O(1/k)$ rate of convergence of the norm of the gradient mapping under the constant stepsize rule). *Suppose that Assumption* 10.1 *holds and that in addition f is convex and L_f-smooth over $\mathbb{E}$. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method for solving problem* (10.1) *with a constant stepsize rule in which $L_k \equiv L_f$ for all $k \geq 0$. Then for any $\mathbf{x}^* \in X^*$ and $k \geq 0$,*

(a) $\|G_{L_f}(\mathbf{x}^{k+1})\| \leq \|G_{L_f}(\mathbf{x}^k)\|$;

(b) $\|G_{L_f}(\mathbf{x}^k)\| \leq \frac{2L_f\|\mathbf{x}^0 - \mathbf{x}^*\|}{k+1}$.

Proof. Invoking Lemma 10.12 with $\mathbf{x} = \mathbf{x}^k$, we obtain (a). Part (b) now follows by substituting $\alpha = \beta = 1$ in the result of Theorem 10.26 and noting that by part (a), $\|G_{L_f}(\mathbf{x}^k)\| = \min_{n=0,1,\ldots,k}\|G_{L_f}(\mathbf{x}^n)\|$. $\square$

10.5 The Proximal Point Method

Consider the problem

$$\min_{\mathbf{x}\in\mathbb{E}} g(\mathbf{x}), \tag{10.36}$$

where $g : \mathbb{E} \to (-\infty, \infty]$ is a proper closed and convex function. Problem (10.36) is actually a special case of the composite problem (10.1) with $f \equiv 0$. The update step of the proximal gradient method in this case takes the form

$$\mathbf{x}^{k+1} = \mathrm{prox}_{\frac{1}{L_k}g}(\mathbf{x}^k).$$

Taking $L_k = \frac{1}{c}$ for some $c > 0$, we obtain the *proximal point method*.

The Proximal Point Method

Initialization: pick $\mathbf{x}^0 \in \mathbb{E}$ and $c > 0$.
General step ($k \geq 0$):

$$\mathbf{x}^{k+1} = \mathrm{prox}_{cg}(\mathbf{x}^k).$$

The proximal point method is actually not a practical algorithm since the general step asks to minimize the function $g(\mathbf{x}) + \frac{c}{2}\|\mathbf{x} - \mathbf{x}^k\|^2$, which in general is as hard to accomplish as solving the original problem of minimizing g. Since the proximal point method is a special case of the proximal gradient method, we can deduce its main convergence results from the corresponding results on the proximal gradient method. Specifically, since the smooth part $f \equiv 0$ is 0-smooth, we can take any constant stepsize to guarantee convergence and Theorems 10.21 and 10.24 imply the following result.

Theorem 10.28 (convergence of the proximal point method). *Let $g : \mathbb{E} \to (-\infty, \infty]$ be a proper closed and convex function. Assume that problem*

$$\min_{\mathbf{x}\in\mathbb{E}} g(\mathbf{x})$$

has a nonempty optimal set X^*, and let the optimal value be given by g_{opt}. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the proximal point method with parameter $c > 0$. Then

(a) $g(\mathbf{x}^k) - g_{\mathrm{opt}} \leq \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2ck}$ *for any $\mathbf{x}^* \in X^*$ and $k \geq 0$;*

(b) *the sequence $\{\mathbf{x}^k\}_{k\geq 0}$ converges to some point in X^*.*

10.6 Convergence of the Proximal Gradient Method—The Strongly Convex Case

In the case where f is assumed to be σ-strongly convex for some $\sigma > 0$, the sublinear rate of convergence can be improved into a *linear rate* of convergence, meaning a rate of the form $O(q^k)$ for some $q \in (0, 1)$. Throughout the analysis of the strongly convex case we denote the unique optimal solution of problem (10.1) by $\mathbf{x}^*$.

Theorem 10.29 (linear rate of convergence of the proximal gradient method—strongly convex case). *Suppose that Assumption* 10.1 *holds and that in addition f is σ-strongly convex ($\sigma > 0$). Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method for solving problem* (10.1) *with either a constant stepsize rule in which $L_k \equiv L_f$ for all $k \geq 0$ or the backtracking procedure* B2. *Let*

$$\alpha = \begin{cases} 1, & \text{constant stepsize,} \\ \max\left\{\eta, \frac{s}{L_f}\right\}, & \text{backtracking.} \end{cases}$$

Then for any $k \geq 0$,

(a) $\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\sigma}{\alpha L_f}\right)\|\mathbf{x}^k - \mathbf{x}^*\|^2$;

(b) $\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\sigma}{\alpha L_f}\right)^k \|\mathbf{x}^0 - \mathbf{x}^*\|^2$;

(c) $F(\mathbf{x}^{k+1}) - F_{\text{opt}} \leq \frac{\alpha L_f}{2}\left(1 - \frac{\sigma}{\alpha L_f}\right)^{k+1}\|\mathbf{x}^0 - \mathbf{x}^*\|^2$.

Proof. Plugging $L = L_k$, $\mathbf{x} = \mathbf{x}^*$, and $\mathbf{y} = \mathbf{x}^k$ into the fundamental prox-grad inequality (10.21) and taking into account the fact that in both stepsize rules condition (10.20) is satisfied, we obtain

$$F(\mathbf{x}^*) - F(\mathbf{x}^{k+1}) \geq \frac{L_k}{2}\|\mathbf{x}^* - \mathbf{x}^{k+1}\|^2 - \frac{L_k}{2}\|\mathbf{x}^* - \mathbf{x}^k\|^2 + \ell_f(\mathbf{x}^*, \mathbf{x}^k).$$

Since f is σ-strongly convex, it follows by Theorem 5.24(ii) that

$$\ell_f(\mathbf{x}^*, \mathbf{x}^k) = f(\mathbf{x}^*) - f(\mathbf{x}^k) - \langle \nabla f(\mathbf{x}^k), \mathbf{x}^* - \mathbf{x}^k\rangle \geq \frac{\sigma}{2}\|\mathbf{x}^k - \mathbf{x}^*\|^2.$$

Thus,

$$F(\mathbf{x}^*) - F(\mathbf{x}^{k+1}) \geq \frac{L_k}{2}\|\mathbf{x}^* - \mathbf{x}^{k+1}\|^2 - \frac{L_k - \sigma}{2}\|\mathbf{x}^* - \mathbf{x}^k\|^2. \tag{10.37}$$

Since $\mathbf{x}^*$ is a minimizer of F, $F(\mathbf{x}^*) - F(\mathbf{x}^{k+1}) \leq 0$, and hence, by (10.37) and the fact that $L_k \leq \alpha L_f$ (see Remark 10.19),

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\sigma}{L_k}\right)\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\sigma}{\alpha L_f}\right)\|\mathbf{x}^k - \mathbf{x}^*\|^2,$$

establishing part (a). Part (b) follows immediately by (a). To prove (c), note that by (10.37),

$$\begin{aligned} F(\mathbf{x}^{k+1}) - F_{\text{opt}} &\leq \frac{L_k - \sigma}{2}\|\mathbf{x}^k - \mathbf{x}^*\|^2 - \frac{L_k}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 \\ &\leq \frac{\alpha L_f - \sigma}{2}\|\mathbf{x}^k - \mathbf{x}^*\|^2 \\ &= \frac{\alpha L_f}{2}\left(1 - \frac{\sigma}{\alpha L_f}\right)\|\mathbf{x}^k - \mathbf{x}^*\|^2 \\ &\leq \frac{\alpha L_f}{2}\left(1 - \frac{\sigma}{\alpha L_f}\right)^{k+1}\|\mathbf{x}^0 - \mathbf{x}^*\|^2, \end{aligned}$$

where part (b) was used in the last inequality. ∎

Theorem 10.29 immediately implies that in the strongly convex case, the proximal gradient method requires an order of $\log(\frac{1}{\varepsilon})$ iterations to obtain an ε-optimal solution.

Theorem 10.30 (complexity of the proximal gradient method—The strongly convex case). *Under the setting of Theorem* 10.29*, for any* $k \geq 1$ *satisfying*

$$k \geq \alpha\kappa \log\left(\frac{1}{\varepsilon}\right) + \alpha\kappa \log\left(\frac{\alpha L_f R^2}{2}\right),$$

it holds that $F(\mathbf{x}^k) - F_{\text{opt}} \leq \varepsilon$*, where* R *is an upper bound on* $\|\mathbf{x}^0 - \mathbf{x}^*\|$ *and* $\kappa = \frac{L_f}{\sigma}$*.*

Proof. Let $k \geq 1$. By Theorem 10.29 and the definition of κ, a sufficient condition for the inequality $F(\mathbf{x}^k) - F_{\text{opt}} \leq \varepsilon$ to hold is that

$$\frac{\alpha L_f}{2}\left(1 - \frac{1}{\alpha\kappa}\right)^k R^2 \leq \varepsilon,$$

which is the same as

$$k \log\left(1 - \frac{1}{\alpha\kappa}\right) \leq \log\left(\frac{2\varepsilon}{\alpha L_f R^2}\right). \tag{10.38}$$

Since $\log(1-x) \leq -x$ for any[57] $x \leq 1$, it follows that a sufficient condition for (10.38) to hold is that

$$-\frac{1}{\alpha\kappa}k \leq \log\left(\frac{2\varepsilon}{\alpha L_f R^2}\right),$$

namely, that

$$k \geq \alpha\kappa \log\left(\frac{1}{\varepsilon}\right) + \alpha\kappa \log\left(\frac{\alpha L_f R^2}{2}\right). \quad \square$$

10.7 The Fast Proximal Gradient Method—FISTA

10.7.1 The Method

The proximal gradient method achieves an $O(1/k)$ rate of convergence in function values to the optimal value. In this section we will show how to accelerate the method in order to obtain a rate of $O(1/k^2)$ in function values. The method is known as the "fast proximal gradient method," but we will also refer to it as "FISTA," which is an acronym for "fast iterative shrinkage-thresholding algorithm"; see Example 10.37 for further explanations. The method was devised and analyzed by Beck and Teboulle in the paper [18], from which the convergence analysis is taken.

We will assume that f is convex and that it is L_f-smooth, meaning that it is L_f-smooth over the entire space $\mathbb{E}$. We gather all the required properties in the following assumption.

[57] The inequality also holds for $x = 1$ since in that case the left-hand side is $-\infty$.

Assumption 10.31.

(A) $g : \mathbb{E} \to (-\infty, \infty]$ *is proper closed and convex.*

(B) $f : \mathbb{E} \to \mathbb{R}$ *is* L_f*-smooth and convex.*

(C) *The optimal set of problem* (10.1) *is nonempty and denoted by* X^**. The optimal value of the problem is denoted by* $F_{\rm opt}$.

The description of FISTA now follows.

FISTA

Input: $(f, g, \mathbf{x}^0)$, where f and g satisfy properties (A) and (B) in Assumption 10.31 and $\mathbf{x}^0 \in \mathbb{E}$.
Initialization: set $\mathbf{y}^0 = \mathbf{x}^0$ and $t_0 = 1$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick $L_k > 0$;

(b) set $\mathbf{x}^{k+1} = \text{prox}_{\frac{1}{L_k} g}\left(\mathbf{y}^k - \frac{1}{L_k}\nabla f(\mathbf{y}^k)\right)$;

(c) set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;

(d) compute $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)$.

As usual, we will consider two options for the choice of L_k: constant and backtracking. The backtracking procedure for choosing the stepsize is referred to as "backtracking procedure B3" and is identical to procedure B2 with the sole difference that it is invoked on the vector $\mathbf{y}^k$ rather than on $\mathbf{x}^k$.

- **Constant.** $L_k = L_f$ for all k.
- **Backtracking procedure B3.** The procedure requires two parameters (s, η), where $s > 0$ and $\eta > 1$. Define $L_{-1} = s$. At iteration k $(k \geq 0)$ the choice of L_k is done as follows: First, L_k is set to be equal to L_{k-1}. Then, while (recall that $T_L(\mathbf{y}) \equiv T_L^{f,g}(\mathbf{y}) = \text{prox}_{\frac{1}{L}g}(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y}))$),

$$f(T_{L_k}(\mathbf{y}^k)) > f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k), T_{L_k}(\mathbf{y}^k) - \mathbf{y}^k \rangle + \frac{L_k}{2}\|T_{L_k}(\mathbf{y}^k) - \mathbf{y}^k\|^2,$$

we set $L_k := \eta L_k$. In other words, the stepsize is chosen as $L_k = L_{k-1}\eta^{i_k}$, where i_k is the smallest nonnegative integer for which the condition

$$f(T_{L_{k-1}\eta^{i_k}}(\mathbf{y}^k)) \leq f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k), T_{L_{k-1}\eta^{i_k}}(\mathbf{y}^k) - \mathbf{y}^k \rangle + \frac{L_k}{2}\|T_{L_{k-1}\eta^{i_k}}(\mathbf{y}^k) - \mathbf{y}^k\|^2$$

is satisfied.

In both stepsize rules, the following inequality is satisfied for any $k \geq 0$:

$$f(T_{L_k}(\mathbf{y}^k)) \leq f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k), T_{L_k}(\mathbf{y}^k) - \mathbf{y}^k \rangle + \frac{L_k}{2} \|T_{L_k}(\mathbf{y}^k) - \mathbf{y}^k\|^2. \quad (10.39)$$

Remark 10.32. *Since the backtracking procedure* B3 *is identical to the* B2 *procedure (only employed on* $\mathbf{y}^k$*), the arguments of Remark* 10.19 *are still valid, and we have that*

$$\beta L_f \leq L_k \leq \alpha L_f,$$

where α *and* β *are given in* (10.25).

The next lemma shows an important lower bound on the sequence $\{t_k\}_{k\geq 0}$ that will be used in the convergence proof.

Lemma 10.33. *Let* $\{t_k\}_{k\geq 0}$ *be the sequence defined by*

$$t_0 = 1,\ t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \quad k \geq 0.$$

Then $t_k \geq \frac{k+2}{2}$ *for all* $k \geq 0$.

Proof. The proof is by induction on k. Obviously, for $k = 0$, $t_0 = 1 \geq \frac{0+2}{2}$. Suppose that the claim holds for k, meaning $t_k \geq \frac{k+2}{2}$. We will prove that $t_{k+1} \geq \frac{k+3}{2}$. By the recursive relation defining the sequence and the induction assumption,

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2} \geq \frac{1 + \sqrt{1 + (k+2)^2}}{2} \geq \frac{1 + \sqrt{(k+2)^2}}{2} = \frac{k+3}{2}. \quad \square$$

10.7.2 Convergence Analysis of FISTA

Theorem 10.34 ($O(1/k^2)$ rate of convergence of FISTA). *Suppose that Assumption* 10.31 *holds. Let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by FISTA for solving problem* (10.1) *with either a constant stepsize rule in which* $L_k \equiv L_f$ *for all* $k \geq 0$ *or the backtracking procedure* B3*. Then for any* $\mathbf{x}^* \in X^*$ *and* $k \geq 1$,

$$F(\mathbf{x}^k) - F_{\text{opt}} \leq \frac{2\alpha L_f \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{(k+1)^2},$$

where $\alpha = 1$ *in the constant stepsize setting and* $\alpha = \max\left\{\eta, \frac{s}{L_f}\right\}$ *if the backtracking rule is employed.*

Proof. Let $k \geq 1$. Substituting $\mathbf{x} = t_k^{-1}\mathbf{x}^* + (1 - t_k^{-1})\mathbf{x}^k$, $\mathbf{y} = \mathbf{y}^k$, and $L = L_k$ in the fundamental prox-grad inequality (10.21), taking into account that inequality

(10.39) is satisfied and that f is convex, we obtain that

$$\begin{aligned}
&F(t_k^{-1}\mathbf{x}^* + (1-t_k^{-1})\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \\
&\geq \frac{L_k}{2}\|\mathbf{x}^{k+1} - (t_k^{-1}\mathbf{x}^* + (1-t_k^{-1})\mathbf{x}^k)\|^2 - \frac{L_k}{2}\|\mathbf{y}^k - (t_k^{-1}\mathbf{x}^* + (1-t_k^{-1})\mathbf{x}^k)\|^2 \\
&= \frac{L_k}{2t_k^2}\|t_k\mathbf{x}^{k+1} - (\mathbf{x}^* + (t_k-1)\mathbf{x}^k)\|^2 - \frac{L_k}{2t_k^2}\|t_k\mathbf{y}^k - (\mathbf{x}^* + (t_k-1)\mathbf{x}^k)\|^2. \quad (10.40)
\end{aligned}$$

By the convexity of F,

$$F(t_k^{-1}\mathbf{x}^* + (1-t_k^{-1})\mathbf{x}^k) \leq t_k^{-1}F(\mathbf{x}^*) + (1-t_k^{-1})F(\mathbf{x}^k).$$

Therefore, using the notation $v_n \equiv F(\mathbf{x}^n) - F_{\text{opt}}$ for any $n \geq 0$,

$$\begin{aligned}
F(t_k^{-1}\mathbf{x}^* + (1-t_k^{-1})\mathbf{x}^k) - F(\mathbf{x}^{k+1}) &\leq (1-t_k^{-1})(F(\mathbf{x}^k) - F(\mathbf{x}^*)) - (F(\mathbf{x}_{k+1}) - F(\mathbf{x}^*)) \\
&= (1-t_k^{-1})v_k - v_{k+1}. \quad (10.41)
\end{aligned}$$

On the other hand, using the relation $\mathbf{y}^k = \mathbf{x}^k + \left(\frac{t_{k-1}-1}{t_k}\right)(\mathbf{x}^k - \mathbf{x}^{k-1})$,

$$\begin{aligned}
\|t_k\mathbf{y}^k - (\mathbf{x}^* + (t_k-1)\mathbf{x}^k)\|^2 &= \|t_k\mathbf{x}^k + (t_{k-1}-1)(\mathbf{x}^k - \mathbf{x}^{k-1}) - (\mathbf{x}^* + (t_k-1)\mathbf{x}^k)\|^2 \\
&= \|t_{k-1}\mathbf{x}^k - (\mathbf{x}^* + (t_{k-1}-1)\mathbf{x}^{k-1})\|^2. \quad (10.42)
\end{aligned}$$

Combining (10.40), (10.41), and (10.42), we obtain that

$$(t_k^2 - t_k)v_k - t_k^2 v_{k+1} \geq \frac{L_k}{2}\|\mathbf{u}^{k+1}\|^2 - \frac{L_k}{2}\|\mathbf{u}^k\|^2,$$

where we use the notation $\mathbf{u}^n = t_{n-1}\mathbf{x}^n - (\mathbf{x}^* + (t_{n-1}-1)\mathbf{x}^{n-1})$ for any $n \geq 0$. By the update rule of t_{k+1}, we have $t_k^2 - t_k = t_{k-1}^2$, and hence

$$\frac{2}{L_k}t_{k-1}^2 v_k - \frac{2}{L_k}t_k^2 v_{k+1} \geq \|\mathbf{u}^{k+1}\|^2 - \|\mathbf{u}^k\|^2.$$

Since $L_k \geq L_{k-1}$, we can conclude that

$$\frac{2}{L_{k-1}}t_{k-1}^2 v_k - \frac{2}{L_k}t_k^2 v_{k+1} \geq \|\mathbf{u}^{k+1}\|^2 - \|\mathbf{u}^k\|^2.$$

Thus,

$$\|\mathbf{u}^{k+1}\|^2 + \frac{2}{L_k}t_k^2 v_{k+1} \leq \|\mathbf{u}^k\|^2 + \frac{2}{L_{k-1}}t_{k-1}^2 v_k,$$

and hence, for any $k \geq 1$,

$$\|\mathbf{u}^k\|^2 + \frac{2}{L_{k-1}}t_{k-1}^2 v_k \leq \|\mathbf{u}^1\|^2 + \frac{2}{L_0}t_0^2 v_1 = \|\mathbf{x}^1 - \mathbf{x}^*\|^2 + \frac{2}{L_0}(F(\mathbf{x}^1) - F_{\text{opt}}) \quad (10.43)$$

Substituting $\mathbf{x} = \mathbf{x}^*, \mathbf{y} = \mathbf{y}^0$, and $L = L_0$ in the fundamental prox-grad inequality (10.21), taking into account the convexity of f yields

$$\frac{2}{L_0}(F(\mathbf{x}^*) - F(\mathbf{x}^1)) \geq \|\mathbf{x}^1 - \mathbf{x}^*\|^2 - \|\mathbf{y}^0 - \mathbf{x}^*\|^2,$$

which, along with the fact that $\mathbf{y}^0 = \mathbf{x}^0$, implies the bound

$$\|\mathbf{x}^1 - \mathbf{x}^*\|^2 + \frac{2}{L_0}(F(\mathbf{x}^1) - F_{\text{opt}}) \leq \|\mathbf{x}^0 - \mathbf{x}^*\|^2.$$

Combining the last inequality with (10.43), we get

$$\frac{2}{L_{k-1}} t_{k-1}^2 v_k \le \|\mathbf{u}^k\|^2 + \frac{2}{L_{k-1}} t_{k-1}^2 v_k \le \|\mathbf{x}^0 - \mathbf{x}^*\|^2.$$

Thus, using the bound $L_{k-1} \le \alpha L_f$, the definition of v_k, and Lemma 10.33,

$$F(\mathbf{x}^k) - F_{\text{opt}} \le \frac{L_{k-1}\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2t_{k-1}^2} \le \frac{2\alpha L_f \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{(k+1)^2}. \qquad \square$$

Remark 10.35 (alternative choice for t_k). *A close inspection of the proof of Theorem* 10.34 *reveals that the result is correct if $\{t_k\}_{k\ge 0}$ is any sequence satisfying the following two properties for any $k \ge 0$:* (a) $t_k \ge \frac{k+2}{2}$; (b) $t_{k+1}^2 - t_{k+1} \le t_k^2$. *The choice $t_k = \frac{k+2}{2}$ also satisfies these two properties. The validity of* (a) *is obvious; to show* (b)*, note that*

$$\begin{aligned} t_{k+1}^2 - t_{k+1} = t_{k+1}(t_{k+1} - 1) &= \frac{k+3}{2} \cdot \frac{k+1}{2} = \frac{k^2 + 4k + 3}{4} \\ &\le \frac{k^2 + 4k + 4}{4} = \frac{(k+2)^2}{4} = t_k^2. \end{aligned}$$

Remark 10.36. *Note that FISTA has an $O(1/k^2)$ rate of convergence in function values, while the proximal gradient method has an $O(1/k)$ rate of convergence. This improvement was achieved despite the fact that the dominant computational steps at each iteration of both methods are essentially the same: one gradient evaluation and one prox computation.*

10.7.3 Examples

Example 10.37. Consider the following model, which was already discussed in Example 10.2:

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) + \lambda\|\mathbf{x}\|_1,$$

where $\lambda > 0$ and $f : \mathbb{R}^n \to \mathbb{R}$ is assumed to be convex and L_f-smooth. The update formula of the proximal gradient method with constant stepsize $\frac{1}{L_f}$ has the form

$$\mathbf{x}^{k+1} = \mathcal{T}_{\frac{\lambda}{L_f}}\left(\mathbf{x}^k - \frac{1}{L_f}\nabla f(\mathbf{x}^k)\right).$$

As was already noted in Example 10.3, since at each iteration one shrinkage/soft-thresholding operation is performed, this method is also known as the *iterative shrinkage-thresholding algorithm* (ISTA). The general update step of the accelerated proximal gradient method discussed in this section takes the following form:

(a) set $\mathbf{x}^{k+1} = \mathcal{T}_{\frac{\lambda}{L_f}}\left(\mathbf{y}^k - \frac{1}{L_f}\nabla f(\mathbf{y}^k)\right)$;

(b) set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;

(c) compute $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)$.

The above scheme truly deserves to be called "fast iterative shrinkage/thresholding algorithm" (FISTA) since it is an accelerated method that performs at each iteration a thresholding step. In this book we adopt the convention and use the acronym FISTA as the name of the fast proximal gradient method for a general nonsmooth part g. ∎

Example 10.38 (l_1-regularized least squares). As a special instance of Example 10.37, consider the problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1, \tag{10.44}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$, and $\lambda > 0$. The problem fits model (10.1) with $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ and $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$. The function f is L_f-smooth with $L_f = \left\|\mathbf{A}^T\mathbf{A}\right\|_{2,2} = \lambda_{\max}(\mathbf{A}^T\mathbf{A})$ (see Example 5.2). The update step of FISTA has the following form:

(a) set $\mathbf{x}^{k+1} = \mathcal{T}_{\frac{\lambda}{L_k}}\left(\mathbf{y}^k - \frac{1}{L_k}\mathbf{A}^T(\mathbf{A}\mathbf{y}^k - \mathbf{b})\right)$;

(b) set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;

(c) compute $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)$.

The update step of the proximal gradient method, which in this case is the same as ISTA, is

$$\mathbf{x}^{k+1} = \mathcal{T}_{\frac{\lambda}{L_k}}\left(\mathbf{x}^k - \frac{1}{L_k}\mathbf{A}^T(\mathbf{A}\mathbf{x}^k - \mathbf{b})\right).$$

The stepsizes in both methods can be chosen to be the constant $L_k \equiv \lambda_{\max}(\mathbf{A}^T\mathbf{A})$.

To illustrate the difference in the actual performance of ISTA and FISTA, we generated an instance of the problem with $\lambda = 1$ and $\mathbf{A} \in \mathbb{R}^{100\times 110}$. The components of $\mathbf{A}$ were independently generated using a standard normal distribution. The "true" vector is $\mathbf{x}_{\text{true}} = \mathbf{e}_3 - \mathbf{e}_7$, and $\mathbf{b}$ was chosen as $\mathbf{b} = \mathbf{A}\mathbf{x}_{\text{true}}$. We ran 200 iterations of ISTA and FISTA in order to solve problem (10.44) with initial vector $\mathbf{x} = \mathbf{e}$, the vector of all ones. It is well known that the l_1-norm element in the objective function is a regularizer that promotes sparsity, and we thus expect that the optimal solution of (10.44) will be close to the "true" sparse vector $\mathbf{x}_{\text{true}}$. The distances to optimality in terms of function values of the sequences generated by the two methods as a function of the iteration index are plotted in Figure 10.1, where it is apparent that FISTA is far superior to ISTA.

In Figure 10.2 we plot the vectors that were obtained by the two methods. Obviously, the solution produced by 200 iterations of FISTA is much closer to the optimal solution (which is very close to $\mathbf{e}_3 - \mathbf{e}_7$) than the solution obtained after 200 iterations of ISTA. ∎

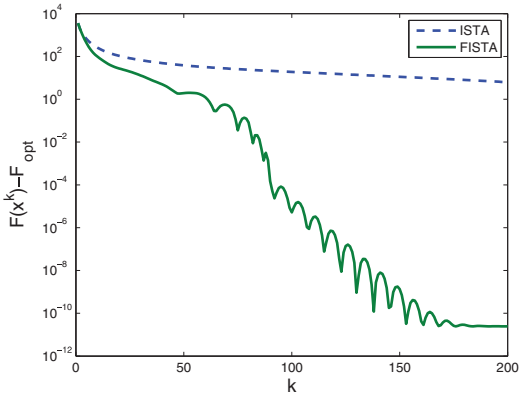

Figure 10.1. *Results of* 200 *iterations of ISTA and FISTA on an* l_1*-regularized least squares problem.*

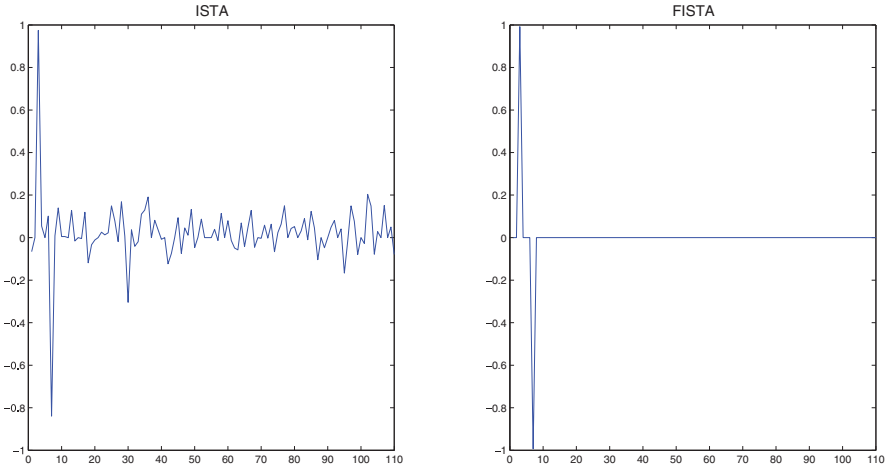

Figure 10.2. *Solutions obtained by ISTA (left) and FISTA (right).*

10.7.4 MFISTA[58]

FISTA is not a monotone method, meaning that the sequence of function values it produces is not necessarily nonincreasing. It is possible to define a monotone version of FISTA, which we call MFISTA, which is a descent method and at the same time preserves the same rate of convergence as FISTA.

[58] MFISTA and its convergence analysis are from the work of Beck and Teboulle [17].

MFISTA

Input: $(f, g, \mathbf{x}^0)$, where f and g satisfy properties (A) and (B) in Assumption 10.31 and $\mathbf{x}^0 \in \mathbb{E}$.
Initialization: set $\mathbf{y}^0 = \mathbf{x}^0$ and $t_0 = 1$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick $L_k > 0$;

(b) set $\mathbf{z}^k = \operatorname{prox}_{\frac{1}{L_k} g}\left(\mathbf{y}^k - \frac{1}{L_k}\nabla f(\mathbf{y}^k)\right)$;

(c) choose $\mathbf{x}^{k+1} \in \mathbb{E}$ such that $F(\mathbf{x}^{k+1}) \leq \min\{F(\mathbf{z}^k), F(\mathbf{x}^k)\}$;

(d) set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;

(e) compute $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \frac{t_k}{t_{k+1}}(\mathbf{z}^k - \mathbf{x}^{k+1}) + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)$.

Remark 10.39. *The choice* $\mathbf{x}^{k+1} \in \operatorname{argmin}\{F(\mathbf{x}) : \mathbf{x} = \mathbf{x}^k, \mathbf{z}^k\}$ *is a very simple rule ensuring the condition* $F(\mathbf{x}^{k+1}) \leq \min\{F(\mathbf{z}^k), F(\mathbf{x}^k)\}$. *We also note that the convergence established in Theorem* 10.40 *only requires the condition* $F(\mathbf{x}^{k+1}) \leq F(\mathbf{z}^k)$.

The convergence result of MFISTA, whose proof is a minor adjustment of the proof of Theorem 10.34, is given below.

Theorem 10.40 ($O(1/k^2)$ rate of convergence of MFISTA). *Suppose that Assumption* 10.31 *holds. Let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by MFISTA for solving problem* (10.1) *with either a constant stepsize rule in which* $L_k \equiv L_f$ *for all* $k \geq 0$ *or the backtracking procedure* B3. *Then for any* $\mathbf{x}^* \in X^*$ *and* $k \geq 1$,

$$F(\mathbf{x}^k) - F_{\text{opt}} \leq \frac{2\alpha L_f \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{(k+1)^2},$$

where $\alpha = 1$ *in the constant stepsize setting and* $\alpha = \max\left\{\eta, \frac{s}{L_f}\right\}$ *if the backtracking rule is employed.*

Proof. Let $k \geq 1$. Substituting $\mathbf{x} = t_k^{-1}\mathbf{x}^* + (1 - t_k^{-1})\mathbf{x}^k$, $\mathbf{y} = \mathbf{y}^k$, and $L = L_k$ in the fundamental prox-grad inequality (10.21), taking into account that inequality (10.39) is satisfied and that f is convex, we obtain that

$$\begin{aligned}
&F(t_k^{-1}\mathbf{x}^* + (1 - t_k^{-1})\mathbf{x}^k) - F(\mathbf{z}^k) \\
&\geq \frac{L_k}{2}\|\mathbf{z}^k - (t_k^{-1}\mathbf{x}^* + (1 - t_k^{-1})\mathbf{x}^k)\|^2 - \frac{L_k}{2}\|\mathbf{y}^k - (t_k^{-1}\mathbf{x}^* + (1 - t_k^{-1})\mathbf{x}^k)\|^2 \\
&= \frac{L_k}{2t_k^2}\|t_k\mathbf{z}^k - (\mathbf{x}^* + (t_k - 1)\mathbf{x}^k)\|^2 - \frac{L_k}{2t_k^2}\|t_k\mathbf{y}^k - (\mathbf{x}^* + (t_k - 1)\mathbf{x}^k)\|^2. \quad (10.45)
\end{aligned}$$

By the convexity of F,

$$F(t_k^{-1}\mathbf{x}^* + (1 - t_k^{-1})\mathbf{x}^k) \leq t_k^{-1}F(\mathbf{x}^*) + (1 - t_k^{-1})F(\mathbf{x}^k).$$

Therefore, using the notation $v_n \equiv F(\mathbf{x}^n) - F_{\text{opt}}$ for any $n \geq 0$ and the fact that $F(\mathbf{x}^{k+1}) \leq F(\mathbf{z}^k)$, it follows that

$$\begin{aligned} F(t_k^{-1}\mathbf{x}^* + (1-t_k^{-1})\mathbf{x}^k) - F(\mathbf{z}^k) &\leq (1-t_k^{-1})(F(\mathbf{x}^k) - F(\mathbf{x}^*)) - (F(\mathbf{x}_{k+1}) - F(\mathbf{x}^*)) \\ &= (1-t_k^{-1})v_k - v_{k+1}. \end{aligned} \tag{10.46}$$

On the other hand, using the relation $\mathbf{y}^k = \mathbf{x}^k + \frac{t_{k-1}}{t_k}(\mathbf{z}^{k-1} - \mathbf{x}^k) + \left(\frac{t_{k-1}-1}{t_k}\right)(\mathbf{x}^k - \mathbf{x}^{k-1})$, we have

$$t_k\mathbf{y}^k - (\mathbf{x}^* + (t_k - 1)\mathbf{x}^k) = t_{k-1}\mathbf{z}^{k-1} - (\mathbf{x}^* + (t_{k-1} - 1)\mathbf{x}^{k-1}). \tag{10.47}$$

Combining (10.45), (10.46), and (10.47), we obtain that

$$(t_k^2 - t_k)v_k - t_k^2 v_{k+1} \geq \frac{L_k}{2}\|\mathbf{u}^{k+1}\|^2 - \frac{L_k}{2}\|\mathbf{u}^k\|^2,$$

where we use the notation $\mathbf{u}^n = t_{n-1}\mathbf{z}^{n-1} - (\mathbf{x}^* + (t_{n-1} - 1)\mathbf{x}^{n-1})$ for any $n \geq 0$. By the update rule of t_{k+1}, we have $t_k^2 - t_k = t_{k-1}^2$, and hence

$$\frac{2}{L_k}t_{k-1}^2 v_k - \frac{2}{L_k}t_k^2 v_{k+1} \geq \|\mathbf{u}^{k+1}\|^2 - \|\mathbf{u}^k\|^2.$$

Since $L_k \geq L_{k-1}$, we can conclude that

$$\frac{2}{L_{k-1}}t_{k-1}^2 v_k - \frac{2}{L_k}t_k^2 v_{k+1} \geq \|\mathbf{u}^{k+1}\|^2 - \|\mathbf{u}^k\|^2.$$

Thus,

$$\|\mathbf{u}^{k+1}\|^2 + \frac{2}{L_k}t_k^2 v_{k+1} \leq \|\mathbf{u}^k\|^2 + \frac{2}{L_{k-1}}t_{k-1}^2 v_k,$$

and hence, for any $k \geq 1$,

$$\|\mathbf{u}^k\|^2 + \frac{2}{L_{k-1}}t_{k-1}^2 v_k \leq \|\mathbf{u}^1\|^2 + \frac{2}{L_0}t_0^2 v_1 = \|\mathbf{z}^0 - \mathbf{x}^*\|^2 + \frac{2}{L_0}(F(\mathbf{x}^1) - F_{\text{opt}}). \tag{10.48}$$

Substituting $\mathbf{x} = \mathbf{x}^*, \mathbf{y} = \mathbf{y}^0$, and $L = L_0$ in the fundamental prox-grad inequality (10.21), taking into account the convexity of f, yields

$$\frac{2}{L_0}(F(\mathbf{x}^*) - F(\mathbf{z}^0)) \geq \|\mathbf{z}^0 - \mathbf{x}^*\|^2 - \|\mathbf{y}^0 - \mathbf{x}^*\|^2,$$

which, along with the facts that $\mathbf{y}^0 = \mathbf{x}^0$ and $F(\mathbf{x}^1) \leq F(\mathbf{z}^0)$, implies the bound

$$\|\mathbf{z}^0 - \mathbf{x}^*\|^2 + \frac{2}{L_0}(F(\mathbf{x}^1) - F_{\text{opt}}) \leq \|\mathbf{x}^0 - \mathbf{x}^*\|^2.$$

Combining the last inequality with (10.48), we get

$$\frac{2}{L_{k-1}}t_{k-1}^2 v_k \leq \|\mathbf{u}^k\|^2 + \frac{2}{L_{k-1}}t_{k-1}^2 v_k \leq \|\mathbf{x}^0 - \mathbf{x}^*\|^2.$$

Thus, using the bound $L_{k-1} \leq \alpha L_f$, the definition of v_k, and Lemma 10.33,

$$F(\mathbf{x}^k) - F_{\text{opt}} \leq \frac{L_{k-1}\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2t_{k-1}^2} \leq \frac{2\alpha L_f\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{(k+1)^2}. \qquad \square$$

10.7.5 Weighted FISTA

Consider the main composite model (10.1) under Assumption 10.31. Suppose that $\mathbb{E} = \mathbb{R}^n$. Recall that a standing assumption in this chapter is that the underlying space is Euclidean, but this does not mean that the endowed inner product is the dot product. Assume that the endowed inner product is the $\mathbf{Q}$-inner product: $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{Q} \mathbf{y}$, where $\mathbf{Q} \in \mathbb{S}^n_{++}$. In this case, as explained in Remark 3.32, the gradient is given by

$$\nabla f(\mathbf{x}) = \mathbf{Q}^{-1} D_f(\mathbf{x}),$$

where

$$D_f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \frac{\partial f}{\partial x_2}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{pmatrix}.$$

We will use a Lipschitz constant of ∇f w.r.t. the $\mathbf{Q}$-norm, which we will denote by $L_f^{\mathbf{Q}}$. The constant is essentially defined by the relation

$$\|\mathbf{Q}^{-1} D_f(\mathbf{x}) - \mathbf{Q}^{-1} D_f(\mathbf{y})\|_{\mathbf{Q}} \leq L_f^{\mathbf{Q}} \|\mathbf{x} - \mathbf{y}\|_{\mathbf{Q}} \text{ for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

The general update rule for FISTA in this case will have the following form:

(a) set $\mathbf{x}^{k+1} = \operatorname{prox}_{\frac{1}{L_f^{\mathbf{Q}}} g}\big(\mathbf{y}^k - \frac{1}{L_f^{\mathbf{Q}}} \mathbf{Q}^{-1} D_f(\mathbf{y}^k)\big)$;

(b) set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;

(c) compute $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)$.

Obviously, the prox operator in step (a) is computed in terms of the $\mathbf{Q}$-norm, meaning that

$$\operatorname{prox}_h(\mathbf{x}) = \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^n} \left\{ h(\mathbf{u}) + \frac{1}{2} \|\mathbf{u} - \mathbf{x}\|_{\mathbf{Q}}^2 \right\}.$$

The convergence result of Theorem 10.34 will also be written in terms of the $\mathbf{Q}$-norm:

$$F(\mathbf{x}^k) - F_{\text{opt}} \leq \frac{2 L_f^{\mathbf{Q}} \|\mathbf{x}^0 - \mathbf{x}^*\|_{\mathbf{Q}}^2}{(k+1)^2}.$$

10.7.6 Restarting FISTA in the Strongly Convex Case

We will now assume that in addition to Assumption 10.31, f is σ-strongly convex for some $\sigma > 0$. Recall that by Theorem 10.30, the proximal gradient method attains an ε-optimal solution after an order of $O(\kappa \log(\frac{1}{\varepsilon}))$ iterations ($\kappa = \frac{L_f}{\sigma}$). The natural question is obviously how the complexity result improves when using FISTA instead of the proximal gradient method. Perhaps surprisingly, one option for obtaining such an improved result is by considering a version of FISTA that incorporates a restarting of the method after a constant amount of iterations.

Restarted FISTA

Initialization: pick $\mathbf{z}^{-1} \in \mathbb{E}$ and a positive integer N. Set $\mathbf{z}^0 = T_{L_f}(\mathbf{z}^{-1})$.
General step ($k \geq 0$):

- run N iterations of FISTA with constant stepsize ($L_k \equiv L_f$) and input $(f, g, \mathbf{z}^k)$ and obtain a sequence $\{\mathbf{x}^n\}_{n=0}^N$;
- set $\mathbf{z}^{k+1} = \mathbf{x}^N$.

The algorithm essentially consists of "outer" iterations, and each one employs N iterations of FISTA. To avoid confusion, the outer iterations will be called *cycles*. Theorem 10.41 below shows that an order of $O(\sqrt{\kappa}\log(\frac{1}{\varepsilon}))$ FISTA iterations are enough to guarantee that an ε-optimal solution is attained.

Theorem 10.41 ($O(\sqrt{\kappa}\log(\frac{1}{\varepsilon}))$ complexity of restarted FISTA). *Suppose that Assumption* 10.31 *holds and that f is σ-strongly convex ($\sigma > 0$). Let $\{\mathbf{z}^k\}_{k\geq 0}$ be the sequence generated by the restarted FISTA method employed with $N = \lceil\sqrt{8\kappa}-1\rceil$, where $\kappa = \frac{L_f}{\sigma}$. Let R be an upper bound on $\|\mathbf{z}^{-1} - \mathbf{x}^*\|$, where $\mathbf{x}^*$ is the unique optimal solution of problem* (10.1). *Then*[59]

(a) *for any $k \geq 0$,*

$$F(\mathbf{z}^k) - F_{\rm opt} \leq \frac{L_f R^2}{2}\left(\frac{1}{2}\right)^k;$$

(b) *after k iterations of FISTA with k satisfying*

$$k \geq \sqrt{8\kappa}\left(\frac{\log(\frac{1}{\varepsilon})}{\log(2)} + \frac{\log(L_f R^2)}{\log(2)}\right),$$

an ε-optimal solution is obtained at the end of the last completed cycle. That is,

$$F(\mathbf{z}^{\lfloor \frac{k}{N}\rfloor}) - F_{\rm opt} \leq \varepsilon.$$

Proof. (a) By Theorem 10.34, for any $n \geq 0$,

$$F(\mathbf{z}^{n+1}) - F_{\rm opt} \leq \frac{2L_f\|\mathbf{z}^n - \mathbf{x}^*\|^2}{(N+1)^2}. \tag{10.49}$$

Since f is σ-strongly convex, it follows by Theorem 5.25(b) that

$$F(\mathbf{z}^n) - F_{\rm opt} \geq \frac{\sigma}{2}\|\mathbf{z}^n - \mathbf{x}^*\|^2,$$

which, combined with (10.49), yields (recalling that $\kappa = L_f/\sigma$)

$$F(\mathbf{z}^{n+1}) - F_{\rm opt} \leq \frac{4\kappa(F(\mathbf{z}^n) - F_{\rm opt})}{(N+1)^2}. \tag{10.50}$$

[59]Note that the index k in part (a) stands for the number of cycles, while in part (b) it is the number of FISTA iterations.

Since $N \geq \sqrt{8\kappa} - 1$, it follows that $\frac{4\kappa}{(N+1)^2} \leq \frac{1}{2}$, and hence by (10.50)

$$F(\mathbf{z}^{n+1}) - F_{\text{opt}} \leq \frac{1}{2}(F(\mathbf{z}^n) - F_{\text{opt}}).$$

Employing the above inequality for $n = 0, 1, \ldots, k-1$, we conclude that

$$F(\mathbf{z}^k) - F_{\text{opt}} \leq \left(\frac{1}{2}\right)^k (F(\mathbf{z}^0) - F_{\text{opt}}). \tag{10.51}$$

Note that $\mathbf{z}^0 = T_{L_f}(\mathbf{z}^{-1})$. Invoking the fundamental prox-grad inequality (10.21) with $\mathbf{x} = \mathbf{x}^*, \mathbf{y} = \mathbf{z}^{-1}$, $L = L_f$, and taking into account the convexity of f, we obtain

$$F(\mathbf{x}^*) - F(\mathbf{z}^0) \geq \frac{L_f}{2}\|\mathbf{x}^* - \mathbf{z}^0\|^2 - \frac{L_f}{2}\|\mathbf{x}^* - \mathbf{z}^{-1}\|^2,$$

and hence

$$F(\mathbf{z}^0) - F_{\text{opt}} \leq \frac{L_f}{2}\|\mathbf{x}^* - \mathbf{z}^{-1}\|^2 \leq \frac{L_f R^2}{2}. \tag{10.52}$$

Combining (10.51) and (10.52), we obtain

$$F(\mathbf{z}^k) - F_{\text{opt}} \leq \frac{L_f R^2}{2}\left(\frac{1}{2}\right)^k.$$

(b) If k iterations of FISTA were employed, then $\lfloor \frac{k}{N} \rfloor$ cycles were completed. By part (a),

$$F(\mathbf{z}^{\lfloor \frac{k}{N} \rfloor}) - F_{\text{opt}} \leq \frac{L_f R^2}{2}\left(\frac{1}{2}\right)^{\lfloor \frac{k}{N} \rfloor} \leq L_f R^2 \left(\frac{1}{2}\right)^{\frac{k}{N}}.$$

Therefore, a sufficient condition for the inequality $F(\mathbf{z}^{\lfloor \frac{k}{N} \rfloor}) - F_{\text{opt}} \leq \varepsilon$ to hold is that

$$L_f R^2 \left(\frac{1}{2}\right)^{\frac{k}{N}} \leq \varepsilon,$$

which is equivalent to the inequality

$$k \geq N\left(\frac{\log(\frac{1}{\varepsilon})}{\log(2)} + \frac{\log(L_f R^2)}{\log(2)}\right).$$

The claim now follows by the fact that $N = \lceil \sqrt{8\kappa} - 1 \rceil \leq \sqrt{8\kappa}$. ∎

10.7.7 The Strongly Convex Case (Once Again)—Variation on FISTA

As in the previous section, we will assume that in addition to Assumption 10.31, f is σ-strongly convex for some $\sigma > 0$. We will define a variant of FISTA, called V-FISTA, that will exhibit the improved linear rate of convergence of the restarted FISTA. This rate is established without any need of restarting of the method.

V-FISTA

Input: $(f, g, \mathbf{x}^0)$, where f and g satisfy properties (A) and (B) in Assumption 10.31, f is σ-strongly convex ($\sigma > 0$), and $\mathbf{x}^0 \in \mathbb{E}$.
Initialization: set $\mathbf{y}^0 = \mathbf{x}^0, t_0 = 1$ and $\kappa = \frac{L_f}{\sigma}$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) set $\mathbf{x}^{k+1} = \text{prox}_{\frac{1}{L_f} g}\left(\mathbf{y}^k - \frac{1}{L_f}\nabla f(\mathbf{y}^k)\right)$;

(b) compute $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)$.

The improved linear rate of convergence is established in the next result, whose proof is a variation on the proof of the rate of convergence of FISTA for the non–strongly convex case (Theorem 10.34).

Theorem 10.42 ($O((1-1/\sqrt{\kappa})^k)$ rate of convergence of V-FISTA).[60] *Suppose that Assumption* 10.31 *holds and that f is σ-strongly convex ($\sigma > 0$). Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by V-FISTA for solving problem* (10.1). *Then for any $\mathbf{x}^* \in X^*$ and $k \geq 0$,*

$$F(\mathbf{x}^k) - F_{\text{opt}} \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \left(F(\mathbf{x}^0) - F_{\text{opt}} + \frac{\sigma}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|^2\right), \tag{10.53}$$

where $\kappa = \frac{L_f}{\sigma}$.

Proof. By the fundamental prox-grad inequality (Theorem 10.16) and the σ-strong convexity of f (invoking Theorem 5.24), it follows that for any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$,

$$\begin{aligned} F(\mathbf{x}) - F(T_{L_f}(\mathbf{y})) &\geq \frac{L_f}{2}\|\mathbf{x} - T_{L_f}\mathbf{y})\|^2 - \frac{L_f}{2}\|\mathbf{x} - \mathbf{y}\|^2 + f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \\ &\geq \frac{L_f}{2}\|\mathbf{x} - T_{L_f}(\mathbf{y})\|^2 - \frac{L_f}{2}\|\mathbf{x} - \mathbf{y}\|^2 + \frac{\sigma}{2}\|\mathbf{x} - \mathbf{y}\|^2. \end{aligned}$$

Therefore,

$$F(\mathbf{x}) - F(T_{L_f}(\mathbf{y})) \geq \frac{L_f}{2}\|\mathbf{x} - T_{L_f}(\mathbf{y})\|^2 - \frac{L_f - \sigma}{2}\|\mathbf{x} - \mathbf{y}\|^2. \tag{10.54}$$

Let $k \geq 0$ and $t = \sqrt{\kappa} = \sqrt{\frac{L_f}{\sigma}}$. Substituting $\mathbf{x} = t^{-1}\mathbf{x}^* + (1 - t^{-1})\mathbf{x}^k$ and $\mathbf{y} = \mathbf{y}^k$ into (10.54), we obtain that

$$\begin{aligned} &F(t^{-1}\mathbf{x}^* + (1 - t^{-1})\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \\ &\geq \frac{L_f}{2}\|\mathbf{x}^{k+1} - (t^{-1}\mathbf{x}^* + (1 - t^{-1})\mathbf{x}^k)\|^2 - \frac{L_f - \sigma}{2}\|\mathbf{y}^k - (t^{-1}\mathbf{x}^* + (1 - t^{-1})\mathbf{x}^k)\|^2 \\ &= \frac{L_f}{2t^2}\|t\mathbf{x}^{k+1} - (\mathbf{x}^* + (t-1)\mathbf{x}^k)\|^2 - \frac{L_f - \sigma}{2t^2}\|t\mathbf{y}^k - (\mathbf{x}^* + (t-1)\mathbf{x}^k)\|^2. \end{aligned} \tag{10.55}$$

[60] The proof of Theorem 10.42 follows the proof of Theorem 4.10 from the review paper of Chambolle and Pock [42].

By the σ-strong convexity of F,

$$F(t^{-1}\mathbf{x}^* + (1-t^{-1})\mathbf{x}^k) \le t^{-1}F(\mathbf{x}^*) + (1-t^{-1})F(\mathbf{x}^k) - \frac{\sigma}{2}t^{-1}(1-t^{-1})\|\mathbf{x}^k - \mathbf{x}^*\|^2.$$

Therefore, using the notation $v_n \equiv F(\mathbf{x}^n) - F_{\text{opt}}$ for any $n \ge 0$,

$$\begin{aligned}
&F(t^{-1}\mathbf{x}^* + (1-t^{-1})\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \\
&\le (1-t^{-1})(F(\mathbf{x}^k) - F(\mathbf{x}^*)) - (F(\mathbf{x}_{k+1}) - F(\mathbf{x}^*)) - \frac{\sigma}{2}t^{-1}(1-t^{-1})\|\mathbf{x}^k - \mathbf{x}^*\|^2 \\
&= (1-t^{-1})v_k - v_{k+1} - \frac{\sigma}{2}t^{-1}(1-t^{-1})\|\mathbf{x}^k - \mathbf{x}^*\|^2,
\end{aligned}$$

which, combined with (10.55), yields the inequality

$$\begin{aligned}
&t(t-1)v_k + \frac{L_f - \sigma}{2}\|t\mathbf{y}^k - (\mathbf{x}^* + (t-1)\mathbf{x}^k)\|^2 - \frac{\sigma(t-1)}{2}\|\mathbf{x}^k - \mathbf{x}^*\|^2 \\
&\ge t^2 v_{k+1} + \frac{L_f}{2}\|t\mathbf{x}^{k+1} - (\mathbf{x}^* + (t-1)\mathbf{x}^k)\|^2. \qquad (10.56)
\end{aligned}$$

We will use the following identity that holds for any $\mathbf{a}, \mathbf{b} \in \mathbb{E}$ and $\beta \in [0,1)$:

$$\|\mathbf{a}+\mathbf{b}\|^2 - \beta\|\mathbf{a}\|^2 = (1-\beta)\left\|\mathbf{a} + \frac{1}{1-\beta}\mathbf{b}\right\|^2 - \frac{\beta}{1-\beta}\|\mathbf{b}\|^2.$$

Plugging $\mathbf{a} = \mathbf{x}^k - \mathbf{x}^*, \mathbf{b} = t(\mathbf{y}^k - \mathbf{x}^k)$, and $\beta = \frac{\sigma(t-1)}{L_f - \sigma}$ into the above inequality, we obtain

$$\begin{aligned}
&\frac{L_f - \sigma}{2}\|t(\mathbf{y}^k - \mathbf{x}^k) + \mathbf{x}^k - \mathbf{x}^*\|^2 - \frac{\sigma(t-1)}{2}\|\mathbf{x}^k - \mathbf{x}^*\|^2 \\
&= \frac{L_f - \sigma}{2}\left[\|t(\mathbf{y}^k - \mathbf{x}^k) + \mathbf{x}^k - \mathbf{x}^*\|^2 - \frac{\sigma(t-1)}{L_f - \sigma}\|\mathbf{x}^k - \mathbf{x}^*\|^2\right] \\
&= \frac{L_f - \sigma}{2}\left[\frac{L_f - \sigma t}{L_f - \sigma}\left\|\mathbf{x}^k - \mathbf{x}^* + \frac{L_f - \sigma}{L_f - \sigma t}t(\mathbf{y}^k - \mathbf{x}^k)\right\|^2 - \frac{\sigma(t-1)}{L_f - \sigma t}\|\mathbf{x}^k - \mathbf{x}^*\|^2\right] \\
&\le \frac{L_f - \sigma t}{2}\left\|\mathbf{x}^k - \mathbf{x}^* + \frac{L_f - \sigma}{L_f - \sigma t}t(\mathbf{y}^k - \mathbf{x}^k)\right\|^2.
\end{aligned}$$

We can therefore conclude from the above inequality and (10.56) that

$$\begin{aligned}
&t(t-1)v_k + \frac{L_f - \sigma t}{2}\left\|\mathbf{x}^k - \mathbf{x}^* + \frac{L_f - \sigma}{L_f - \sigma t}t(\mathbf{y}^k - \mathbf{x}^k)\right\|^2 \\
&\ge t^2 v_{k+1} + \frac{L_f}{2}\|t\mathbf{x}^{k+1} - (\mathbf{x}^* + (t-1)\mathbf{x}^k)\|^2. \qquad (10.57)
\end{aligned}$$

If $k \ge 1$, then using the relations $\mathbf{y}^k = \mathbf{x}^k + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(\mathbf{x}^k - \mathbf{x}^{k-1})$ and $t = \sqrt{\kappa} = \sqrt{\frac{L_f}{\sigma}}$, we obtain

$$\begin{aligned}
\mathbf{x}^k - \mathbf{x}^* + \frac{L_f - \sigma}{L_f - \sigma t}t(\mathbf{y}^k - \mathbf{x}^k) &= \mathbf{x}^k - \mathbf{x}^* + \frac{L_f - \sigma}{L_f - \sigma t}\frac{t(t-1)}{t+1}(\mathbf{x}^k - \mathbf{x}^{k-1}) \\
&= \mathbf{x}^k - \mathbf{x}^* + \frac{\kappa - 1}{\kappa - \sqrt{\kappa}}\frac{\sqrt{\kappa}(\sqrt{\kappa}-1)}{\sqrt{\kappa}+1}(\mathbf{x}^k - \mathbf{x}^{k-1}) \\
&= \mathbf{x}^k - \mathbf{x}^* + (\sqrt{\kappa}-1)(\mathbf{x}^k - \mathbf{x}^{k-1}) \\
&= t\mathbf{x}^k - (\mathbf{x}^* + (t-1)\mathbf{x}^{k-1}),
\end{aligned}$$

and obviously, for the case $k = 0$ (recalling that $\mathbf{y}^0 = \mathbf{x}^0$),

$$\mathbf{x}^0 - \mathbf{x}^* + \frac{L_f - \sigma}{L_f - \sigma t} t(\mathbf{y}^0 - \mathbf{x}^0) = \mathbf{x}^0 - \mathbf{x}^*.$$

We can thus deduce that (10.57) can be rewritten as (after division by t^2 and using again the definition of t as $t = \sqrt{\frac{L_f}{\sigma}}$)

$$v_{k+1} + \frac{\sigma}{2}\|t\mathbf{x}^{k+1} - (\mathbf{x}^* + (t-1)\mathbf{x}^k)\|^2$$
$$\leq \begin{cases} \left(1 - \frac{1}{t}\right)\left[v_k + \frac{\sigma}{2}\|t\mathbf{x}^k - (\mathbf{x}^* + (t-1)\mathbf{x}^{k-1})\|^2\right], & k \geq 1, \\ \left(1 - \frac{1}{t}\right)\left[v_0 + \frac{\sigma}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|^2\right], & k = 0. \end{cases}$$

We can thus conclude that for any $k \geq 0$,

$$v_k \leq \left(1 - \frac{1}{t}\right)^k \left(v_0 + \frac{\sigma}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|^2\right),$$

which is the desired result (10.53). □

10.8 Smoothing[61]

10.8.1 Motivation

In Chapters 8 and 9 we considered methods for solving nonsmooth convex optimization problems with complexity $O(1/\varepsilon^2)$, meaning that an order of $1/\varepsilon^2$ iterations were required in order to obtain an ε-optimal solution. On the other hand, FISTA requires $O(1/\sqrt{\varepsilon})$ iterations in order to find an ε-optimal solution of the composite model

$$\min_{\mathbf{x}\in\mathbb{E}} f(\mathbf{x}) + g(\mathbf{x}), \tag{10.58}$$

where f is L_f-smooth and convex and g is a proper closed and convex function. In this section we will show how FISTA can be used to devise a method for more general nonsmooth convex problems in an improved complexity of $O(1/\varepsilon)$. In particular, the model that will be considered includes an additional third term to (10.58):

$$\min\{f(\mathbf{x}) + h(\mathbf{x}) + g(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}. \tag{10.59}$$

The function h will be assumed to be real-valued and convex; we will not assume that it is easy to compute its prox operator (as is implicitly assumed on g), and hence solving it directly using FISTA with smooth and nonsmooth parts taken as $(f, g + h)$ is not a practical solution approach. The idea will be to find a smooth approximation of h, say $\tilde{h}$, and solve the problem via FISTA with smooth and nonsmooth parts taken as $(f + \tilde{h}, g)$. This simple idea will be the basis for the improved $O(1/\varepsilon)$ complexity. To be able to describe the method, we will need to study in more detail the notions of *smooth approximations* and *smoothability*.

[61] The idea of producing an $O(1/\varepsilon)$ complexity result for nonsmooth problems by employing an accelerated gradient method was first presented and developed by Nesterov in [95]. The extension presented in Section 10.8 to the three-part composite model and to the setting of more general smooth approximations was developed by Beck and Teboulle in [20], where additional results and extensions can also be found.

10.8.2 Smoothable Functions and Smooth Approximations

Definition 10.43 (smoothable functions). *A convex function $h : \mathbb{E} \to \mathbb{R}$ is called* **(α, β)-smoothable** *$(\alpha, \beta > 0)$ if for any $\mu > 0$ there exists a convex differentiable function $h_\mu : \mathbb{E} \to \mathbb{R}$ such that the following holds:*

(a) *$h_\mu(\mathbf{x}) \le h(\mathbf{x}) \le h_\mu(\mathbf{x}) + \beta\mu$ for all $\mathbf{x} \in \mathbb{E}$.*

(b) *h_μ is $\frac{\alpha}{\mu}$-smooth.*

The function h_μ is called a **$\frac{1}{\mu}$-smooth approximation** *of h with parameters (α, β).*

Example 10.44 (smooth approximation of $\|\mathbf{x}\|_2$). Consider the function $h : \mathbb{R}^n \to \mathbb{R}$ given by $h(\mathbf{x}) = \|\mathbf{x}\|_2$. For any $\mu > 0$, define $h_\mu(\mathbf{x}) \equiv \sqrt{\|\mathbf{x}\|_2^2 + \mu^2} - \mu$. Then for any $\mathbf{x} \in \mathbb{R}^n$,

$$h_\mu(\mathbf{x}) = \sqrt{\|\mathbf{x}\|_2^2 + \mu^2} - \mu \le \|\mathbf{x}\|_2 + \mu - \mu = \|\mathbf{x}\|_2 = h(\mathbf{x}),$$
$$h(\mathbf{x}) = \|\mathbf{x}\|_2 \le \sqrt{\|\mathbf{x}\|_2^2 + \mu^2} = h_\mu(\mathbf{x}) + \mu,$$

showing that property (a) in the definition of smoothable functions holds with $\beta = 1$. To show that property (b) holds with $\alpha = 1$, note that by Example 5.14, the function $\varphi(\mathbf{x}) \equiv \sqrt{\|\mathbf{x}\|_2^2 + 1}$ is 1-smooth, and hence $h_\mu(\mathbf{x}) = \mu\varphi(\mathbf{x}/\mu) - \mu$ is $\frac{1}{\mu}$-smooth. We conclude that h_μ is a $\frac{1}{\mu}$-smooth approximation of h with parameters $(1, 1)$. In the terminology described in Definition 10.43, we showed that h is $(1, 1)$-smoothable. ∎

Example 10.45 (smooth approximation of $\max_i\{x_i\}$). Consider the function $h : \mathbb{R}^n \to \mathbb{R}$ given by $h(\mathbf{x}) = \max\{x_1, x_2, \ldots, x_n\}$. For any $\mu > 0$, define the function

$$h_\mu(\mathbf{x}) = \mu \log\left(\textstyle\sum_{i=1}^n e^{x_i/\mu}\right) - \mu \log n.$$

Then for any $\mathbf{x} \in \mathbb{R}^n$,

$$\begin{aligned} h_\mu(\mathbf{x}) &= \mu \log\left(\sum_{i=1}^n e^{x_i/\mu}\right) - \mu \log n \\ &\le \mu \log\left(n e^{\max_i\{x_i\}/\mu}\right) - \mu \log n = h(\mathbf{x}), \qquad (10.60) \\ h(\mathbf{x}) &= \max_i\{x_i\} \le \mu \log\left(\sum_{i=1}^n e^{x_i/\mu}\right) = h_\mu(\mathbf{x}) + \mu \log n. \qquad (10.61) \end{aligned}$$

By Example 5.15, the function $\varphi(\mathbf{x}) = \log(\sum_{i=1}^n e^{x_i})$ is 1-smooth, and hence the function $h_\mu(\mathbf{x}) = \mu\varphi(\mathbf{x}/\mu) - \mu \log n$ is $\frac{1}{\mu}$-smooth. Combining this with (10.60) and (10.61), it follows that h_μ is a $\frac{1}{\mu}$-smooth approximation of h with parameters $(1, \log n)$. We conclude in particular that h is $(1, \log n)$-smoothable. ∎

The following result describes two important calculus rules of smooth approximations.

Theorem 10.46 (calculus of smooth approximations).

(a) *Let $h^1, h^2 : \mathbb{E} \to \mathbb{R}$ be convex functions, and let γ_1, γ_2 be nonnegative numbers. Suppose that for a given $\mu > 0$, h^i_μ is a $\frac{1}{\mu}$-smooth approximation of h^i with parameters (α_i, β_i) for $i = 1, 2$. Then $\gamma_1 h^1_\mu + \gamma_2 h^2_\mu$ is a $\frac{1}{\mu}$-smooth approximation of $\gamma_1 h^1 + \gamma_2 h^2$ with parameters $(\gamma_1\alpha_1 + \gamma_2\alpha_2, \gamma_1\beta_1 + \gamma_2\beta_2)$.*

(b) *Let $\mathcal{A} : \mathbb{E} \to \mathbb{V}$ be a linear transformation between the Euclidean spaces $\mathbb{E}$ and $\mathbb{V}$. Let $h : \mathbb{V} \to \mathbb{R}$ be a convex function and define*

$$q(\mathbf{x}) \equiv h(\mathcal{A}(\mathbf{x}) + \mathbf{b}),$$

where $\mathbf{b} \in \mathbb{V}$. Suppose that for a given $\mu > 0$, h_μ is a $\frac{1}{\mu}$-smooth approximation of h with parameters (α, β). Then the function $q_\mu(\mathbf{x}) \equiv h_\mu(\mathcal{A}(\mathbf{x}) + \mathbf{b})$ is a $\frac{1}{\mu}$-smooth approximation of q with parameters$(\alpha\|\mathcal{A}\|^2, \beta)$.

Proof. (a) By its definition, h^i_μ $(i = 1, 2)$ is convex, $\frac{\alpha_i}{\mu}$-smooth and satisfies $h^i_\mu(\mathbf{x}) \le h^i(\mathbf{x}) \le h^i_\mu(\mathbf{x}) + \beta_i\mu$ for any $\mathbf{x} \in \mathbb{E}$. We can thus conclude that $\gamma_1 h^1_\mu + \gamma_2 h^2_\mu$ is convex and that for any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$,

$$\gamma_1 h^1_\mu(\mathbf{x}) + \gamma_2 h^2_\mu(\mathbf{x}) \le \gamma_1 h^1(\mathbf{x}) + \gamma_2 h^2(\mathbf{x}) \le \gamma_1 h^1_\mu(\mathbf{x}) + \gamma_2 h^2_\mu(\mathbf{x}) + (\gamma_1\beta_1 + \gamma_2\beta_2)\mu,$$

as well as

$$\begin{aligned}\|\nabla(\gamma_1 h^1_\mu + \gamma_2 h^2_\mu)(\mathbf{x}) - \nabla(\gamma_1 h^1_\mu + \gamma_2 h^2_\mu)(\mathbf{y})\| &\le \gamma_1\|\nabla h^1_\mu(\mathbf{x}) - \nabla h^1_\mu(\mathbf{y})\| \\ &\quad + \gamma_2\|\nabla h^2_\mu(\mathbf{x}) - \nabla h^2_\mu(\mathbf{y})\| \\ &\le \gamma_1\frac{\alpha_1}{\mu}\|\mathbf{x} - \mathbf{y}\| + \gamma_2\frac{\alpha_2}{\mu}\|\mathbf{x} - \mathbf{y}\| \\ &= \frac{\gamma_1\alpha_1 + \gamma_2\alpha_2}{\mu}\|\mathbf{x} - \mathbf{y}\|,\end{aligned}$$

establishing the fact that $\gamma_1 h^1_\mu + \gamma_2 h^2_\mu$ is a $\frac{1}{\mu}$-smooth approximation of $\gamma_1 h^1 + \gamma_2 h^2$ with parameters $(\gamma_1\alpha_1 + \gamma_2\alpha_2, \gamma_1\beta_1 + \gamma_2\beta_2)$.

(b) Since h_μ is a $\frac{1}{\mu}$-smooth approximation of h with parameters (α, β), it follows that h_μ is convex, $\frac{\alpha}{\mu}$-smooth and for any $\mathbf{y} \in \mathbb{V}$,

$$h_\mu(\mathbf{y}) \le h(\mathbf{y}) \le h_\mu(\mathbf{y}) + \beta\mu. \tag{10.62}$$

Let $\mathbf{x} \in \mathbb{E}$. Plugging $\mathbf{y} = \mathcal{A}(\mathbf{x}) + \mathbf{b}$ into (10.62), we obtain that

$$q_\mu(\mathbf{x}) \le q(\mathbf{x}) \le q_\mu(\mathbf{x}) + \beta\mu. \tag{10.63}$$

In addition, by the $\frac{\alpha}{\mu}$-smoothness of h_μ, we have for any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$,

$$\begin{aligned}\|\nabla q_\mu(\mathbf{x}) - \nabla q_\mu(\mathbf{y})\| &= \|\mathcal{A}^T\nabla h_\mu(\mathcal{A}(\mathbf{x}) + \mathbf{b}) - \mathcal{A}^T\nabla h_\mu(\mathcal{A}(\mathbf{y}) + \mathbf{b})\| \\ &\le \|\mathcal{A}^T\| \cdot \|\nabla h_\mu(\mathcal{A}(\mathbf{x}) + \mathbf{b}) - \nabla h_\mu(\mathcal{A}(\mathbf{y}) + \mathbf{b})\| \\ &\le \frac{\alpha}{\mu}\|\mathcal{A}^T\| \cdot \|\mathcal{A}(\mathbf{x}) + \mathbf{b} - \mathcal{A}(\mathbf{y}) - \mathbf{b}\| \\ &\le \frac{\alpha}{\mu}\|\mathcal{A}^T\| \cdot \|\mathcal{A}\| \cdot \|\mathbf{x} - \mathbf{y}\| \\ &= \frac{\alpha\|\mathcal{A}\|^2}{\mu}\|\mathbf{x} - \mathbf{y}\|,\end{aligned}$$

where the last equality follows by the fact that $\|\mathcal{A}\| = \|\mathcal{A}^T\|$ (see Section 1.14). We have thus shown that the convex function h_μ is $\frac{\alpha\|\mathcal{A}\|^2}{\mu}$-smooth and satisfies (10.63) for any $\mathbf{x} \in \mathbb{E}$, establishing the desired result. □

A direct result of Theorem 10.46 is the following corollary stating the preservation of smoothability under nonnegative linear combinations and affine transformations of variables.

Corollary 10.47 (operations preserving smoothability).

(a) *Let $h^1, h^2 : \mathbb{E} \to \mathbb{R}$ be convex functions which are (α_1, β_1)- and (α_2, β_2)-smoothable, respectively, and let γ_1, γ_2 be nonnegative numbers. Then $\gamma_1 h^1 + \gamma_2 h^2$ is a $(\gamma_1\alpha_1 + \gamma_2\alpha_2, \gamma_1\beta_1 + \gamma_2\beta_2)$-smoothable function.*

(b) *Let $\mathcal{A} : \mathbb{E} \to \mathbb{V}$ be a linear transformation between the Euclidean spaces $\mathbb{E}$ and $\mathbb{V}$. Let $h : \mathbb{V} \to \mathbb{R}$ be a convex (α, β)-smoothable function and define*

$$q(\mathbf{x}) \equiv h(\mathcal{A}(\mathbf{x}) + \mathbf{b}),$$

where $\mathbf{b} \in \mathbb{V}$. Then q is $(\alpha\|\mathcal{A}\|^2, \beta)$-smoothable.

Example 10.48 (smooth approximation of $\|\mathbf{Ax} + \mathbf{b}\|_2$). Let $q : \mathbb{R}^n \to \mathbb{R}$ be given by $q(\mathbf{x}) = \|\mathbf{Ax}+\mathbf{b}\|_2$, where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$. Then $q(\mathbf{x}) = g(\mathbf{Ax}+\mathbf{b})$, where $g : \mathbb{R}^m \to \mathbb{R}$ is given by $g(\mathbf{y}) = \|\mathbf{y}\|_2$. Let $\mu > 0$. By Example 10.44, $g_\mu(\mathbf{y}) = \sqrt{\|\mathbf{y}\|_2^2 + \mu^2} - \mu$ is a $\frac{1}{\mu}$-smooth approximation of g with parameters $(1, 1)$, and hence, by Theorem 10.46(b),

$$q_\mu(\mathbf{x}) \equiv g_\mu(\mathbf{Ax} + \mathbf{b}) = \sqrt{\|\mathbf{Ax} + \mathbf{b}\|_2^2 + \mu^2} - \mu$$

is a $\frac{1}{\mu}$-smooth approximation of q with parameters $(\|\mathbf{A}\|_{2,2}^2, 1)$. ■

Example 10.49 (smooth approximation of piecewise affine functions). Let $q : \mathbb{R}^n \to \mathbb{R}$ be given by $q(\mathbf{x}) = \max_{i=1,\dots,m}\{\mathbf{a}_i^T\mathbf{x} + b_i\}$, where $\mathbf{a}_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$ for any $i = 1, 2, \dots, m$. Then $q(\mathbf{x}) = g(\mathbf{Ax}+\mathbf{b})$, where $g(\mathbf{y}) = \max\{y_1, y_2, \dots, y_m\}$, $\mathbf{A}$ is the matrix whose rows are $\mathbf{a}_1^T, \mathbf{a}_2^T, \dots, \mathbf{a}_m^T$, and $\mathbf{b} = (b_1, b_2, \dots, b_m)^T$. Let $\mu > 0$. By Example 10.45, $g_\mu(\mathbf{y}) = \mu \log\left(\sum_{i=1}^m e^{y_i/\mu}\right) - \mu \log m$ is a $\frac{1}{\mu}$-smooth approximation of g with parameters $(1, \log m)$. Therefore, by Theorem 10.46(b), the function

$$q_\mu(\mathbf{x}) \equiv g_\mu(\mathbf{Ax} + \mathbf{b}) = \mu \log\left(\sum_{i=1}^m e^{(\mathbf{a}_i^T\mathbf{x}+b_i)/\mu}\right) - \mu \log m$$

is a $\frac{1}{\mu}$-smooth approximation of q with parameters $(\|\mathbf{A}\|_{2,2}^2, \log m)$. ■

Example 10.50 (tightness of the smoothing parameters). Consider the absolute value function $q : \mathbb{R} \to \mathbb{R}$ given by $q(x) = |x|$. By Example 10.44, for any $\mu > 0$ the function $\sqrt{x^2 + \mu^2} - \mu$ is a $\frac{1}{\mu}$-smooth approximation of q with parameters $(1, 1)$. Let us consider an alternative way to construct a smooth approximation of

q using Theorem 10.46. Note that $q(x) = \max\{x, -x\}$. Thus, by Example 10.49 the function $q_\mu(x) = \mu \log(e^{x/\mu} + e^{-x/\mu}) - \mu \log 2$ is a $\frac{1}{\mu}$-smooth approximation of q with parameters $(\|\mathbf{A}\|_{2,2}^2, \log 2)$, where $\mathbf{A} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Since $\|\mathbf{A}\|_{2,2}^2 = 2$, we conclude that q_μ is a $\frac{1}{\mu}$-smooth approximation of q with parameters $(2, \log 2)$. The question that arises is whether these parameters are tight, meaning whether they are the smallest ones possible. The β-parameter is indeed tight (since $\lim_{x\to\infty} q(x) - q_\mu(x) = \mu \log(2)$); however, the α-parameter is not tight. To see this, note that for any $x \in \mathbb{R}$,

$$q_1''(x) = \frac{4}{(e^x + e^{-x})^2}.$$

Therefore, for any $x \in \mathbb{R}$, it holds that $|q_1''(x)| \le 1$, and hence, by Theorem 5.12, q_1 is 1-smooth. Consequently, q_μ, which can also be written as $q_\mu(\mathbf{x}) = \mu q_1(\mathbf{x}/\mu)$, is $\frac{1}{\mu}$-smooth. We conclude that q_μ, is a $\frac{1}{\mu}$-smooth approximation of q with parameters $(1, \log 2)$. ∎

10.8.3 The Moreau Envelope Revisited

A natural $\frac{1}{\mu}$-smooth approximation of a given real-valued convex function $h : \mathbb{E} \to \mathbb{R}$ is its Moreau envelope M_h^μ, which was discussed in detail in Section 6.7. Recall that the Moreau envelope of h is given by

$$M_h^\mu(\mathbf{x}) = \min_{\mathbf{u}\in\mathbb{E}} \left\{ h(\mathbf{u}) + \frac{1}{2\mu}\|\mathbf{x} - \mathbf{u}\|^2 \right\}.$$

We will now show that whenever h is in addition Lipschitz, the Moreau envelope is indeed a $\frac{1}{\mu}$-smooth approximation.

Theorem 10.51 (smoothability of real-valued Lipschitz convex functions). *Let $h : \mathbb{E} \to \mathbb{R}$ be a convex function satisfying*

$$|h(\mathbf{x}) - h(\mathbf{y})| \le \ell_h \|\mathbf{x} - \mathbf{y}\| \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{E}.$$

Then for any $\mu > 0$, M_h^μ is a $\frac{1}{\mu}$-smooth approximation of h with parameters $(1, \frac{\ell_h^2}{2})$.

Proof. By Theorem 6.60, M_h^μ is $\frac{1}{\mu}$-smooth. For any $\mathbf{x} \in \mathbb{E}$,

$$M_h^\mu(\mathbf{x}) = \min_{\mathbf{u}\in\mathbb{E}} \left\{ h(\mathbf{u}) + \frac{1}{2\mu}\|\mathbf{u} - \mathbf{x}\|^2 \right\} \le h(\mathbf{x}) + \frac{1}{2\mu}\|\mathbf{x} - \mathbf{x}\|^2 = h(\mathbf{x}).$$

Let $\mathbf{g}_\mathbf{x} \in \partial h(\mathbf{x})$. Since h is Lipschitz with constant ℓ_h, it follows by Theorem 3.61 that $\|\mathbf{g}_\mathbf{x}\| \le \ell_h$, and hence

$$\begin{aligned} M_h^\mu(\mathbf{x}) - h(\mathbf{x}) &= \min_{\mathbf{u}\in\mathbb{E}} \left\{ h(\mathbf{u}) - h(\mathbf{x}) + \frac{1}{2\mu}\|\mathbf{u} - \mathbf{x}\|^2 \right\} \\ &\ge \min_{\mathbf{u}\in\mathbb{E}} \left\{ \langle \mathbf{g}_\mathbf{x}, \mathbf{u} - \mathbf{x}\rangle + \frac{1}{2\mu}\|\mathbf{u} - \mathbf{x}\|^2 \right\} \\ &= -\frac{\mu}{2}\|\mathbf{g}_\mathbf{x}\|^2 \\ &\ge -\frac{\ell_h^2}{2}\mu, \end{aligned}$$

where the subgradient inequality was used in the first inequality. To summarize, we obtained that the convex function M_h^μ is $\frac{1}{\mu}$-smooth and satisfies

$$M_h^\mu(\mathbf{x}) \le h(\mathbf{x}) \le M_h^\mu(\mathbf{x}) + \frac{\ell_h^2}{2}\mu,$$

showing that M_h^μ is a $\frac{1}{\mu}$-smooth approximation of h with parameters $(1, \frac{\ell_h^2}{2})$. □

Corollary 10.52. *Let $h : \mathbb{E} \to \mathbb{R}$ be convex and Lipschitz with constant ℓ_h. Then h is $(1, \frac{\ell_h^2}{2})$-smoothable.*

Example 10.53 (smooth approximation of the l_2-norm). Consider the function $h : \mathbb{R}^n \to \mathbb{R}$ given by $h(\mathbf{x}) = \|\mathbf{x}\|_2$. Then h is convex and Lipschitz with constant $\ell_h = 1$. Hence, by Theorem 10.51, for any $\mu > 0$, the function (see Example 6.54)

$$M_h^\mu(\mathbf{x}) = H_\mu(\mathbf{x}) = \begin{cases} \frac{1}{2\mu}\|\mathbf{x}\|_2^2, & \|\mathbf{x}\|_2 \le \mu, \\ \|\mathbf{x}\|_2 - \frac{\mu}{2}, & \|\mathbf{x}\|_2 > \mu, \end{cases}$$

is a $\frac{1}{\mu}$-smooth approximation of h with parameters $(1, \frac{1}{2})$. ■

Example 10.54 (smooth approximation of the l_1-norm). Consider the function $h : \mathbb{R}^n \to \mathbb{R}$ given by $h(\mathbf{x}) = \|\mathbf{x}\|_1$. Then h is convex and Lipschitz with constant $\ell_h = \sqrt{n}$. Hence, by Theorem 10.51, for any $\mu > 0$, the Moreau envelope of h given by

$$M_h^\mu(\mathbf{x}) = \sum_{i=1}^n H_\mu(x_i)$$

is a $\frac{1}{\mu}$-smooth approximation of h with parameters $(1, \frac{n}{2})$. ■

Example 10.55 (smooth approximations of the absolute value function). Let us consider again the absolute value function $h(x) = |x|$. In our discussions we actually considered three possible $\frac{1}{\mu}$-smooth approximations of h, which are detailed below along with their parameters:

- **(Example 10.44)** $h_\mu^1(x) = \sqrt{x^2 + \mu^2} - \mu$, $(\alpha, \beta) = (1, 1)$.
- **(Example 10.50)** $h_\mu^2(x) = \mu \log(e^{x/\mu} + e^{-x/\mu}) - \mu \log 2$, $(\alpha, \beta) = (1, \log 2)$.
- **(Example 10.53)** $h_\mu^3(x) = H_\mu(x)$, $(\alpha, \beta) = (1, \frac{1}{2})$.

Obviously, the Huber function is the best $\frac{1}{\mu}$-smooth approximation out of the three functions since all the functions have the same α-parameter, but h_μ^3 has the smallest β-parameter. This phenomenon is illustrated in Figure 10.3, where the three functions are plotted (for the case $\mu = 0.2$). ■

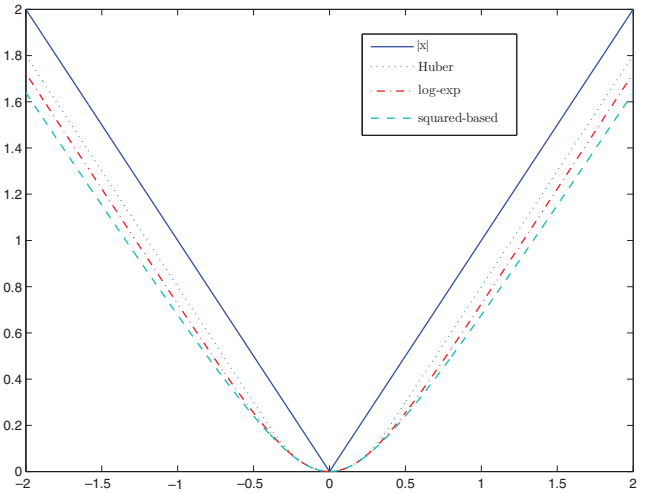

Figure 10.3. *The absolute value function along with its three 5-smooth approximations ($\mu = 0.2$). "squared-based" is the function $h^1_\mu(x) = \sqrt{x^2+\mu^2} - \mu$, "log-exp" is $h^2_\mu(x) = \mu \log(e^{x/\mu} + e^{-x/\mu}) - \mu \log 2$, and "Huber" is $h^3_\mu(x) = H_\mu(x)$.*

10.8.4 The S-FISTA Method

The optimization model that we consider is

$$\min_{\mathbf{x}\in\mathbb{E}} \{H(\mathbf{x}) \equiv f(\mathbf{x}) + h(\mathbf{x}) + g(\mathbf{x})\}, \tag{10.64}$$

where the following assumptions are made.

Assumption 10.56.

(A) *$f : \mathbb{E} \to \mathbb{R}$ is L_f-smooth ($L_f \geq 0$).*

(B) *$h : \mathbb{E} \to \mathbb{R}$ is (α, β)-smoothable ($\alpha, \beta > 0$). For any $\mu > 0$, h_μ denotes a $\frac{1}{\mu}$-smooth approximation of h with parameters (α, β).*

(C) *$g : \mathbb{E} \to (-\infty, \infty]$ is proper closed and convex.*

(D) *H has bounded level sets. Specifically, for any $\delta > 0$, there exists $R_\delta > 0$ such that*

$$\|\mathbf{x}\| \leq R_\delta \text{ for any } \mathbf{x} \text{ satisfying } H(\mathbf{x}) \leq \delta.$$

(E) *The optimal set of problem (10.64) is nonempty and denoted by X^*. The optimal value of the problem is denoted by H_{opt}.*

Assumption (E) is actually a consequence of assumptions (A)–(D). The idea is to consider the smoothed version of (10.64),

$$\min_{\mathbf{x}\in\mathbb{E}}\{H_\mu(\mathbf{x}) \equiv \underbrace{f(\mathbf{x}) + h_\mu(\mathbf{x})}_{F_\mu(\mathbf{x})} + g(\mathbf{x})\}, \tag{10.65}$$

for some *smoothing parameter* $\mu > 0$, and solve it using an accelerated method with convergence rate of $O(1/k^2)$ in function values. Actually, *any* accelerated method can be employed, but we will describe the version in which FISTA with constant stepsize is employed on (10.65) with the smooth and nonsmooth parts taken as F_μ and g, respectively. The method is described in detail below. Note that a Lipschitz constant of the gradient of F_μ is $L_f + \frac{\alpha}{\mu}$, and thus the stepsize is taken as $\frac{1}{L_f+\frac{\alpha}{\mu}}$.

S-FISTA

Input: $\mathbf{x}^0 \in \mathrm{dom}(g)$, $\mu > 0$.
Initialization: set $\mathbf{y}^0 = \mathbf{x}^0, t_0 = 1$; construct h_μ—a $\frac{1}{\mu}$-smooth approximation of h with parameters (α, β); set $F_\mu = f + h_\mu$, $\tilde{L} = L_f + \frac{\alpha}{\mu}$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) set $\mathbf{x}^{k+1} = \mathrm{prox}_{\frac{1}{\tilde{L}}g}\left(\mathbf{y}^k - \frac{1}{\tilde{L}}\nabla F_\mu(\mathbf{y}^k)\right)$;

(b) set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;

(c) compute $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{t_k-1}{t_{k+1}}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)$.

The next result shows how, given an accuracy level $\varepsilon > 0$, the parameter μ can be chosen to ensure that an ε-optimal solution of the original problem (10.64) is reached in $O(1/\varepsilon)$ iterations.

Theorem 10.57 ($O(1/\varepsilon)$ complexity of S-FISTA). *Suppose that Assumption* 10.56 *holds. Let $\varepsilon \in (0, \bar{\varepsilon})$ for some fixed $\bar{\varepsilon} > 0$. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by S-FISTA with smoothing parameter*

$$\mu = \sqrt{\frac{\alpha}{\beta}}\frac{\varepsilon}{\sqrt{\alpha\beta} + \sqrt{\alpha\beta + L_f\varepsilon}}.$$

Then for any k satisfying

$$k \geq 2\sqrt{2\alpha\beta\Gamma}\frac{1}{\varepsilon} + \sqrt{2L_f\Gamma}\frac{1}{\sqrt{\varepsilon}},$$

where $\Gamma = (R_{H(\mathbf{x}^0)+\frac{\bar{\varepsilon}}{2}} + \|\mathbf{x}^0\|)^2$, it holds that $H(\mathbf{x}^k) - H_{\mathrm{opt}} \leq \varepsilon$.

Proof. By definition of S-FISTA, $\{\mathbf{x}^k\}_{k\geq 0}$ is the sequence generated by FISTA employed on problem (10.65) with input $(F_\mu, g, \mathbf{x}^0)$. Note that

$$\mathrm{argmin}_{\mathbf{x}\in\mathbb{E}} H_\mu(\mathbf{x}) = \mathrm{argmin}_{\mathbf{x}\in\mathbb{E}}\left\{H_\mu(\mathbf{x}) : H_\mu(\mathbf{x}) \leq H_\mu(\mathbf{x}^0)\right\}. \tag{10.66}$$

Since H_μ is closed, the feasible set $C \equiv \{\mathbf{x} \in \mathbb{E} : H_\mu(\mathbf{x}) \leq H_\mu(\mathbf{x}^0)\}$ of the right-hand side problem in (10.66) is closed. We will show that it is also bounded. Indeed, since h_μ is a $\frac{1}{\mu}$-smooth approximation of h with parameters (α, β), it follows in particular that $h(\mathbf{x}) \leq h_\mu(\mathbf{x}) + \beta\mu$ for all $\mathbf{x} \in \mathbb{E}$, and consequently $H(\mathbf{x}) \leq H_\mu(\mathbf{x}) + \beta\mu$ for all $\mathbf{x} \in \mathbb{E}$. Thus,

$$C \subseteq \{\mathbf{x} \in \mathbb{E} : H(\mathbf{x}) \leq H_\mu(\mathbf{x}^0) + \beta\mu\},$$

which by Assumption 10.56(D) implies that C is bounded and hence, by its closedness, also compact. We can therefore conclude by Weierstrass theorem for closed functions (Theorem 2.12) that an optimal solution of problem (10.65) is attained at some point $\mathbf{x}^*_\mu$ with an optimal value $H_{\mu,\text{opt}}$. By Theorem 10.34, since F_μ is $(L_f + \frac{\alpha}{\mu})$-smooth,

$$H_\mu(\mathbf{x}^k) - H_{\mu,\text{opt}} \leq 2\left(L_f + \frac{\alpha}{\mu}\right)\frac{\|\mathbf{x}^0 - \mathbf{x}^*_\mu\|^2}{(k+1)^2} = 2\left(L_f + \frac{\alpha}{\mu}\right)\frac{\Lambda}{(k+1)^2}, \quad (10.67)$$

where $\Lambda = \|\mathbf{x}^0 - \mathbf{x}^*_\mu\|^2$. We use again the fact that h_μ is a $\frac{1}{\mu}$-smooth approximation of h with parameters (α, β), from which it follows that for any $\mathbf{x} \in \mathbb{E}$,

$$H_\mu(\mathbf{x}) \leq H(\mathbf{x}) \leq H_\mu(\mathbf{x}) + \beta\mu. \quad (10.68)$$

In particular, the following two inequalities hold:

$$H_{\text{opt}} \geq H_{\mu,\text{opt}} \quad \text{and} \quad H(\mathbf{x}^k) \leq H_\mu(\mathbf{x}^k) + \beta\mu,\ k = 0, 1, \ldots, \quad (10.69)$$

which, combined with (10.67), yields

$$\begin{aligned} H(\mathbf{x}^k) - H_{\text{opt}} &\leq H_\mu(\mathbf{x}^k) + \beta\mu - H_{\mu,\text{opt}} \leq 2L_f\frac{\Lambda}{(k+1)^2} + \left(\frac{2\alpha\Lambda}{(k+1)^2}\right)\frac{1}{\mu} + \beta\mu \\ &\leq 2L_f\frac{\Lambda}{k^2} + \left(\frac{2\alpha\Lambda}{k^2}\right)\frac{1}{\mu} + \beta\mu. \end{aligned}$$

Therefore, for a given $K > 0$, it holds that for any $k \geq K$,

$$H(\mathbf{x}^k) - H_{\text{opt}} \leq 2L_f\frac{\Lambda}{K^2} + \left(\frac{2\alpha\Lambda}{K^2}\right)\frac{1}{\mu} + \beta\mu. \quad (10.70)$$

Minimizing the right-hand side w.r.t. μ, we obtain

$$\mu = \sqrt{\frac{2\alpha\Lambda}{\beta}}\frac{1}{K}. \quad (10.71)$$

Plugging the above expression into (10.70), we conclude that for any $k \geq K$,

$$H(\mathbf{x}^k) - H_{\text{opt}} \leq 2L_f\frac{\Lambda}{K^2} + 2\sqrt{2\alpha\beta\Lambda}\frac{1}{K}.$$

Thus, to guarantee that $\mathbf{x}^k$ is an ε-optimal solution for any $k \geq K$, it is enough that K will satisfy

$$2L_f\frac{\Lambda}{K^2} + 2\sqrt{2\alpha\beta\Lambda}\frac{1}{K} \leq \varepsilon.$$

Denoting $t = \frac{\sqrt{2\Lambda}}{K}$, the above inequality reduces to

$$L_f t^2 + 2\sqrt{\alpha\beta}t - \varepsilon \le 0,$$

which, by the fact that $t > 0$, is equivalent to

$$\frac{\sqrt{2\Lambda}}{K} = t \le \frac{-\sqrt{\alpha\beta} + \sqrt{\alpha\beta + L_f\varepsilon}}{L_f} = \frac{\varepsilon}{\sqrt{\alpha\beta} + \sqrt{\alpha\beta + L_f\varepsilon}}.$$

We conclude that K should satisfy

$$K \ge \frac{\sqrt{2\Lambda\alpha\beta} + \sqrt{2\Lambda\alpha\beta + 2\Lambda L_f\varepsilon}}{\varepsilon}.$$

In particular, if we choose

$$K = K_1 \equiv \frac{\sqrt{2\Lambda\alpha\beta} + \sqrt{2\Lambda\alpha\beta + 2\Lambda L_f\varepsilon}}{\varepsilon}$$

and μ according to (10.71), meaning that

$$\mu = \sqrt{\frac{2\alpha\Lambda}{\beta}}\frac{1}{K_1} = \sqrt{\frac{\alpha}{\beta}}\frac{\varepsilon}{\sqrt{\alpha\beta} + \sqrt{\alpha\beta + L_f\varepsilon}},$$

then for any $k \ge K_1$ it holds that $H(\mathbf{x}^k) - H_{\text{opt}} \le \varepsilon$. By (10.68) and (10.69),

$$H(\mathbf{x}_\mu^*) - \beta\mu \le H_\mu(\mathbf{x}_\mu^*) = H_{\mu,\text{opt}} \le H_{\text{opt}} \le H(\mathbf{x}^0),$$

which along with the inequality

$$\mu = \sqrt{\frac{\alpha}{\beta}}\frac{\varepsilon}{\sqrt{\alpha\beta} + \sqrt{\alpha\beta + L_f\varepsilon}} \le \sqrt{\frac{\alpha}{\beta}}\frac{\varepsilon}{\sqrt{\alpha\beta} + \sqrt{\alpha\beta}} \le \frac{\bar{\varepsilon}}{2\beta}$$

implies that $H(\mathbf{x}_\mu^*) \le H(\mathbf{x}^0) + \frac{\bar{\varepsilon}}{2}$, and hence, by Assumption 10.56(D), it follows that $\|\mathbf{x}_\mu^*\| \le R_\delta$, where $\delta = H(\mathbf{x}^0) + \frac{\bar{\varepsilon}}{2}$. Therefore, $\Lambda = \|\mathbf{x}_\mu^* - \mathbf{x}^0\|^2 \le (R_\delta + \|\mathbf{x}^0\|)^2 = \Gamma$. Consequently,

$$\begin{aligned} K_1 &= \frac{\sqrt{2\Lambda\alpha\beta} + \sqrt{2\Lambda\alpha\beta + 2\Lambda L_f\varepsilon}}{\varepsilon} \\ &\overset{\sqrt{\gamma+\delta}\le\sqrt{\gamma}+\sqrt{\delta}\ \forall\gamma,\delta\ge 0}{\le} \frac{2\sqrt{2\Lambda\alpha\beta} + \sqrt{2\Lambda L_f\varepsilon}}{\varepsilon} \\ &\le \frac{2\sqrt{2\Gamma\alpha\beta} + \sqrt{2\Gamma L_f\varepsilon}}{\varepsilon} \\ &\equiv K_2, \end{aligned}$$

and hence for any $k \ge K_2$, we have that $H(\mathbf{x}^k) - H_{\text{opt}} \le \varepsilon$, establishing the desired result. ∎

Remark 10.58. *Note that the smoothing parameter chosen in Theorem* 10.57 *does not depend on* Γ*, although the number of iterations required to obtain an* ε*-optimal solution does depend on* Γ*.*

Example 10.59. Consider the problem

$$\min_{\mathbf{x}\in\mathbb{E}}\{h(\mathbf{x}) : \mathbf{x}\in C\}, \tag{10.72}$$

where C is a nonempty closed and convex set and $h:\mathbb{E}\to\mathbb{R}$ is convex function, which is Lipschitz with constant ℓ_h. Problem (10.72) fits model (10.64) with $f\equiv 0$ and $g=\delta_C$. By Theorem 10.51, for any $\mu>0$ the Moreau envelope M_h^μ is a $\frac{1}{\mu}$-smooth approximation of h with parameters $(\alpha,\beta)=(1,\frac{\ell_h^2}{2})$. In addition, by Theorem 6.60, $\nabla M_h^\mu(\mathbf{x})=\frac{1}{\mu}(\mathbf{x}-\mathrm{prox}_{\mu h}(\mathbf{x}))$. We will pick $h_\mu=M_h^\mu$, and therefore $F_\mu=f+M_h^\mu=M_h^\mu$. By Theorem 10.57, after employing $O(1/\varepsilon)$ iterations of the S-FISTA method with (recalling that $L_f=0$)

$$\mu=\sqrt{\frac{\alpha}{\beta}}\frac{\varepsilon}{\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}}=\sqrt{\frac{\alpha}{\beta}}\frac{\varepsilon}{\sqrt{\alpha\beta}+\sqrt{\alpha\beta}}=\frac{\varepsilon}{2\beta}=\frac{\varepsilon}{\ell_h^2},$$

an ε-optimal solution will be achieved. The stepsize is $\frac{1}{\tilde{L}}$, where $\tilde{L}=\frac{\alpha}{\mu}=\frac{1}{\mu}$. The main update step of S-FISTA has the following form:

$$\begin{aligned}\mathbf{x}^{k+1}&=\mathrm{prox}_{\frac{1}{\tilde{L}}g}\left(\mathbf{y}^k-\frac{1}{\tilde{L}}\nabla F_\mu(\mathbf{y}^k)\right)=P_C\left(\mathbf{y}^k-\frac{1}{\tilde{L}\mu}(\mathbf{y}^k-\mathrm{prox}_{\mu h}(\mathbf{y}^k))\right)\\&=P_C(\mathrm{prox}_{\mu h}(\mathbf{y}^k)).\end{aligned}$$

The S-FISTA method for solving (10.72) is described below.

S-FISTA for solving (10.72)

Initialization: set $\mathbf{y}^0=\mathbf{x}^0\in C, t_0=1, \mu=\frac{\varepsilon}{\ell_h^2}$, and $\tilde{L}=\frac{\ell_h^2}{\varepsilon}$.
General step: for any $k=0,1,2,\ldots$ execute the following steps:

(a) $\mathbf{x}^{k+1}=P_C(\mathrm{prox}_{\mu h}(\mathbf{y}^k))$;

(b) $t_{k+1}=\frac{1+\sqrt{1+4t_k^2}}{2}$;

(c) $\mathbf{y}^{k+1}=\mathbf{x}^{k+1}+\left(\frac{t_k-1}{t_{k+1}}\right)(\mathbf{x}^{k+1}-\mathbf{x}^k)$.

■

Example 10.60. Consider the problem

$$\text{(P)}\quad \min_{\mathbf{x}\in\mathbb{R}^n}\left\{\frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2+\|\mathbf{D}\mathbf{x}\|_1+\lambda\|\mathbf{x}\|_1\right\},$$

where $\mathbf{A}\in\mathbb{R}^{m\times n}, \mathbf{b}\in\mathbb{R}^m, \mathbf{D}\in\mathbb{R}^{p\times n}$, and $\lambda>0$. Problem (P) fits model (10.64) with $f(\mathbf{x})=\frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$, $h(\mathbf{x})=\|\mathbf{D}\mathbf{x}\|_1$, and $g(\mathbf{x})=\lambda\|\mathbf{x}\|_1$. Assumption 10.56 holds: f is convex and L_f-smooth with $L_f=\|\mathbf{A}^T\mathbf{A}\|_{2,2}=\|\mathbf{A}\|_{2,2}^2$, g is proper closed and convex, h is real-valued and convex, and the level sets of the objective function are bounded. To show that h is smoothable, and to find its parameters, note that $h(\mathbf{x})=q(\mathbf{D}\mathbf{x})$, where $q:\mathbb{R}^p\to\mathbb{R}$ is given by $q(\mathbf{y})=\|\mathbf{y}\|_1$. By Example 10.54, for

any $\mu > 0$, $q_\mu(\mathbf{y}) = M_q^\mu(\mathbf{y}) = \sum_{i=1}^p H_\mu(y_i)$ is a $\frac{1}{\mu}$-smooth approximation of q with parameters $(1, \frac{p}{2})$. By Theorem 10.46(b), $q_\mu(\mathbf{Dx})$ is a $\frac{1}{\mu}$-smooth approximation of h with parameters $(\alpha, \beta) = (\|\mathbf{D}\|_{2,2}^2, \frac{p}{2})$, and we will set $h_\mu(\mathbf{x}) = M_q^\mu(\mathbf{Dx})$ and $F_\mu(\mathbf{x}) = f(\mathbf{x}) + h_\mu(\mathbf{x})$. Therefore, invoking Theorem 10.57, to obtain an ε-optimal solution of problem (P), we need to employ the S-FISTA method with

$$\begin{aligned}\mu &= \sqrt{\frac{\alpha}{\beta}}\frac{\varepsilon}{\sqrt{\alpha\beta} + \sqrt{\alpha\beta + L_f\varepsilon}} \\ &= \frac{2\|\mathbf{D}\|_{2,2}}{\sqrt{p}} \cdot \frac{\varepsilon}{\sqrt{\|\mathbf{D}\|_{2,2}^2 p} + \sqrt{\|\mathbf{D}\|_{2,2}^2 p + 2\|\mathbf{A}^T\mathbf{A}\|_{2,2}\varepsilon}}. \end{aligned} \tag{10.73}$$

Since $F_\mu(\mathbf{x}) = f(\mathbf{x}) + M_q^\mu(\mathbf{Dx})$, it follows that

$$\begin{aligned}\nabla F_\mu(\mathbf{x}) &= \nabla f(\mathbf{x}) + \mathbf{D}^T \nabla M_q^\mu(\mathbf{Dx}) \\ &= \nabla f(\mathbf{x}) + \tfrac{1}{\mu}\mathbf{D}^T(\mathbf{Dx} - \mathrm{prox}_{\mu q}(\mathbf{Dx})) \quad \text{[Theorem 6.60]} \\ &= \nabla f(\mathbf{x}) + \tfrac{1}{\mu}\mathbf{D}^T(\mathbf{Dx} - \mathcal{T}_\mu(\mathbf{Dx})). \quad \text{[Example 6.8]}\end{aligned}$$

Below we write the S-FISTA method for solving problem (P) for a given tolerance parameter $\varepsilon > 0$.

S-FISTA for solving (P)

Initialization: set $\mathbf{y}^0 = \mathbf{x}^0 \in \mathbb{R}^n, t_0 = 1$; set μ as in (10.73) and $\tilde{L} = \|\mathbf{A}\|_{2,2}^2 + \frac{\|\mathbf{D}\|_{2,2}^2}{\mu}$.

General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) $\mathbf{x}^{k+1} = \mathcal{T}_{\lambda/\tilde{L}}\left(\mathbf{y}^k - \frac{1}{\tilde{L}}(\mathbf{A}^T(\mathbf{A}\mathbf{y}^k - \mathbf{b}) + \frac{1}{\mu}\mathbf{D}^T(\mathbf{D}\mathbf{y}^k - \mathcal{T}_\mu(\mathbf{D}\mathbf{y}^k)))\right)$;

(b) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;

(c) $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)$.

It is interesting to note that in the case of problem (P) we can actually compute the constant Γ that appears in Theorem 10.57. Indeed, if $H(\mathbf{x}) \le \alpha$, then

$$\lambda\|\mathbf{x}\|_1 \le \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \|\mathbf{D}\mathbf{x}\|_1 + \lambda\|\mathbf{x}\|_1 \le \alpha,$$

and since $\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1$, it follows that R_α can be chosen as $\frac{\alpha}{\lambda}$, from which Γ can be computed. ∎

10.9 Non-Euclidean Proximal Gradient Methods

In this section, and in this section only, the underlying space will *not* be assumed to be Euclidean. We will consider two different approaches for handling this situation.

The first tackles unconstrained smooth problems through a variation of the gradient method, and the second, which is aimed at solving the composite model, is based on replacing the Euclidean prox operator by a mapping based on the Bregman distance.

10.9.1 The Non-Euclidean Gradient Method

Consider the unconstrained problem

$$\min\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}, \tag{10.74}$$

where we assume that f is L_f-smooth w.r.t. the underlying norm. Recall that the gradient method (see Section 10.2) has the form

$$\mathbf{x}^{k+1} = \mathbf{x}^k - t_k \nabla f(\mathbf{x}^k). \tag{10.75}$$

As was already discussed in Section 9.1 (in the context of the mirror descent method), this scheme has a "philosophical" flaw since $\mathbf{x}^k \in \mathbb{E}$ while $\nabla f(\mathbf{x}^k) \in \mathbb{E}^*$. Obviously, as the only difference between $\mathbb{E}$ and $\mathbb{E}^*$ in this book is their underlying norm, there is no practical problem to invoke the scheme (10.75). Nonetheless, we will change the scheme (10.75) and replace $\nabla f(\mathbf{x}^k) \in \mathbb{E}^*$ with a "primal counterpart" in $\mathbb{E}$. For any vector $\mathbf{a} \in \mathbb{E}^*$, we define the *set of primal counterparts of* $\mathbf{a}$ as

$$\Lambda_{\mathbf{a}} = \operatorname{argmax}_{\mathbf{v}\in\mathbb{E}}\{\langle \mathbf{a}, \mathbf{v}\rangle : \|\mathbf{v}\| \le 1\}. \tag{10.76}$$

The lemma below presents some elementary properties of $\Lambda_{\mathbf{a}}$ that follow immediately by its definition and the definition of the dual norm.

Lemma 10.61 (basic properties of the set of primal counterparts). *Let* $\mathbf{a} \in \mathbb{E}^*$.

(a) *If* $\mathbf{a} \neq \mathbf{0}$, *then* $\|\mathbf{a}^\dagger\| = 1$ *for any* $\mathbf{a}^\dagger \in \Lambda_{\mathbf{a}}$.

(b) *If* $\mathbf{a} = \mathbf{0}$, *then* $\Lambda_{\mathbf{a}} = B_{\|\cdot\|}[\mathbf{0}, 1]$.

(c) $\langle \mathbf{a}, \mathbf{a}^\dagger\rangle = \|\mathbf{a}\|_*$ *for any* $\mathbf{a}^\dagger \in \Lambda_{\mathbf{a}}$.

We also note that by the conjugate subgradient theorem (Corollary 4.21),

$$\Lambda_{\mathbf{a}} = \partial h(\mathbf{a}), \text{ where } h(\cdot) = \|\cdot\|_*.$$

Example 10.62. Suppose that $\mathbb{E} = \mathbb{R}^n$ endowed with the Euclidean l_2-norm. In this case, for any $\mathbf{a} \neq \mathbf{0}$,

$$\Lambda_{\mathbf{a}} = \left\{ \frac{\mathbf{a}}{\|\mathbf{a}\|_2} \right\}. \quad \blacksquare$$

Example 10.63. Suppose that $\mathbb{E} = \mathbb{R}^n$ endowed with the l_1-norm. In this case, for any $\mathbf{a} \neq \mathbf{0}$, by Example 3.52,

$$\Lambda_{\mathbf{a}} = \partial\|\cdot\|_\infty(\mathbf{a}) = \left\{ \sum_{i\in I(\mathbf{a})} \lambda_i \operatorname{sgn}(a_i)\mathbf{e}_i : \sum_{i\in I(\mathbf{a})} \lambda_i = 1, \lambda_j \ge 0, j \in I(\mathbf{a}) \right\},$$

where $I(\mathbf{a}) = \operatorname{argmax}_{i=1,2,\ldots,n}|a_i|$. $\blacksquare$

Example 10.64. Suppose that $\mathbb{E} = \mathbb{R}^n$ endowed with the l_∞-norm. For any $\mathbf{a} \neq \mathbf{0}$, $\Lambda_{\mathbf{a}} = \partial h(\mathbf{a})$, where $h(\cdot) = \|\cdot\|_1$. Then, by Example 3.41,

$$\Lambda_{\mathbf{a}} = \{\mathbf{z} \in \mathbb{R}^n : z_i = \operatorname{sgn}(a_i), i \in I_{\neq}(\mathbf{a}), |z_j| \leq 1, j \in I_0(\mathbf{a})\},$$

where

$$I_{\neq}(\mathbf{a}) = \{i \in \{1, 2, \ldots, n\} : a_i \neq 0\}, I_0(\mathbf{a}) = \{i \in \{1, 2, \ldots, n\} : a_i = 0\}. \quad \blacksquare$$

We are now ready to present the non-Euclidean gradient method in which the gradient $\nabla f(\mathbf{x}^k)$ is replaced by a primal counterpart $\nabla f(\mathbf{x}^k)^\dagger \in \Lambda_{\nabla f(\mathbf{x}^k)}$.

The Non-Euclidean Gradient Method

Initialization: pick $\mathbf{x}^0 \in \mathbb{E}$ arbitrarily.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick $\nabla f(\mathbf{x}^k)^\dagger \in \Lambda_{\nabla f(\mathbf{x}^k)}$ and $L_k > 0$;

(b) set $\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{\|\nabla f(\mathbf{x}^k)\|_*}{L_k} \nabla f(\mathbf{x}^k)^\dagger$.

We begin by establishing a sufficient decrease property. The proof is almost identical to the proof of Lemma 10.4.

Lemma 10.65 (sufficient decrease for the non-Euclidean gradient method). *Let $f : \mathbb{E} \to \mathbb{R}$ be an L_f-smooth function, and let $\{\mathbf{x}^k\}_{k \geq 0}$ be the sequence generated by the non-Euclidean gradient method. Then for any $k \geq 0$,*

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{L_k - \frac{L_f}{2}}{L_k^2} \|\nabla f(\mathbf{x}^k)\|_*^2. \tag{10.77}$$

Proof. By the descent lemma (Lemma 5.7) we have

$$\begin{aligned} f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L_f}{2} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \\ &= f(\mathbf{x}^k) - \frac{\|\nabla f(\mathbf{x}^k)\|_*}{L_k} \langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^k)^\dagger \rangle + \frac{L_f \|\nabla f(\mathbf{x}^k)\|_*^2}{2L_k^2} \\ &\overset{(*)}{=} f(\mathbf{x}^k) - \frac{\|\nabla f(\mathbf{x}^k)\|_*^2}{L_k} + \frac{L_f \|\nabla f(\mathbf{x}^k)\|_*^2}{2L_k^2} \\ &= f(\mathbf{x}^k) - \frac{L_k - \frac{L_f}{2}}{L_k^2} \|\nabla f(\mathbf{x}^k)\|_*^2, \end{aligned}$$

where $(*)$ follows by Lemma 10.61(c). $\square$

Similarly to Section 10.3.3, we will consider both constant and backtracking stepsize strategies. In addition, we will also consider an exact line search procedure.

- **Constant.** $L_k = \bar{L} \in \left(\frac{L_f}{2}, \infty\right)$ for all k.
- **Backtracking procedure B4.** The procedure requires three parameters (s, γ, η), where $s > 0, \gamma \in (0,1)$, and $\eta > 1$. The choice of L_k is done as follows: First, L_k is set to be equal to the initial guess s. Then, while

$$f(\mathbf{x}^k) - f\left(\mathbf{x}^k - \frac{\|\nabla f(\mathbf{x}^k)\|_*}{L_k}\nabla f(\mathbf{x}^k)^\dagger\right) < \frac{\gamma}{L_k}\|\nabla f(\mathbf{x}^k)\|_*^2,$$

we set $L_k := \eta L_k$. In other words, L_k is chosen as $L_k = s\eta^{i_k}$, where i_k is the smallest nonnegative integer for which the condition

$$f(\mathbf{x}^k) - f\left(\mathbf{x}^k - \frac{\|\nabla f(\mathbf{x}^k)\|_*}{s\eta^{i_k}}\nabla f(\mathbf{x}^k)^\dagger\right) \geq \frac{\gamma}{s\eta^{i_k}}\|\nabla f(\mathbf{x}^k)\|_*^2$$

is satisfied.

- **Exact line search.** L_k is chosen as

$$L_k \in \operatorname{argmin}_{L>0} f\left(\mathbf{x}^k - \frac{\|\nabla f(\mathbf{x}^k)\|_*}{L}\nabla f(\mathbf{x}^k)^\dagger\right).$$

By the same arguments given in Remark 10.13, it follows that if the backtracking procedure B4 is used, then

$$L_k \leq \max\left\{s, \frac{\eta L_f}{2(1-\gamma)}\right\}. \tag{10.78}$$

Convergence Analysis in the Nonconvex Case

The statements and proofs of the next two results (Lemma 10.66 and Theorem 10.67) are similar those of Lemma 10.14 and Theorem 10.15.

Lemma 10.66 (sufficient decrease of the non-Euclidean gradient method). *Let f be an L_f-smooth function. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the non-Euclidean gradient method for solving problem* (10.74) *with either a constant stepsize corresponding to $L_k = \bar{L} \in \left(\frac{L_f}{2}, \infty\right)$; a stepsize chosen by the backtracking procedure* B4 *with parameters (s, γ, η) satisfying $s > 0, \gamma \in (0,1), \eta > 1$; or an exact line search for computing the stepsize. Then for any $k \geq 0$,*

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq M\|\nabla f(\mathbf{x}^k)\|_*^2, \tag{10.79}$$

where

$$M = \begin{cases} \frac{\bar{L} - \frac{L_f}{2}}{\left(\bar{L}\right)^2}, & \text{constant stepsize,} \\ \frac{\gamma}{\max\left\{s, \frac{\eta L_f}{2(1-\gamma)}\right\}}, & \text{backtracking,} \\ \frac{1}{2L_f}, & \text{exact line search.} \end{cases} \tag{10.80}$$

Proof. The result for the constant stepsize setting follows by plugging $L_k = \bar{L}$ in (10.77). If L_k is chosen by the exact line search procedure, then, in particular, $f(\mathbf{x}^{k+1}) \leq f(\tilde{\mathbf{x}}^k)$, where $\tilde{\mathbf{x}}^k = \mathbf{x}^k - \frac{\|\nabla f(\mathbf{x}^k)\|_*}{L_f} \nabla f(\mathbf{x}^k)^\dagger$, and hence

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq f(\mathbf{x}^k) - f(\tilde{\mathbf{x}}^k) \geq \frac{1}{2L_f} \|\nabla f(\mathbf{x}^k)\|_*^2,$$

where we used the result already established for the constant stepsize in the second inequality. As for the backtracking procedure, by its definition and the upper bound (10.78) on L_k we have

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{\gamma}{L_k} \|\nabla f(\mathbf{x}^k)\|_*^2 \geq \frac{\gamma}{\max\left\{s, \frac{\eta L_f}{2(1-\gamma)}\right\}} \|\nabla f(\mathbf{x}^k)\|_*^2. \qquad \square$$

Theorem 10.67 (convergence of the non-Euclidean gradient method—nonconvex case). *Suppose that f is an L_f-smooth function. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the non-Euclidean gradient method for solving the problem*

$$\min_{\mathbf{x}\in\mathbb{E}} f(\mathbf{x}) \tag{10.81}$$

with either a constant stepsize corresponding to $L_k = \bar{L} \in \left(\frac{L_f}{2}, \infty\right)$; a stepsize chosen by the backtracking procedure B4 *with parameters (s, γ, η) satisfying $s > 0, \gamma \in (0,1), \eta > 1$; or an exact line search for computing the stepsize. Then*

(a) *the sequence $\{f(\mathbf{x}^k)\}_{k\geq 0}$ is nonincreasing; in addition, $f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k)$ if and only if $\nabla f(\mathbf{x}^k) \neq \mathbf{0}$;*

(b) *if the sequence $\{f(\mathbf{x}^k)\}_{k\geq 0}$ is bounded below, then $\nabla f(\mathbf{x}^k) \to \mathbf{0}$ as $k \to \infty$;*

(c) *if the optimal value of* (10.81) *is finite and equal to $f_{\rm opt}$, then*

$$\min_{n=0,1,\ldots,k} \|\nabla f(\mathbf{x}^k)\|_* \leq \frac{\sqrt{f(\mathbf{x}^0) - f_{\rm opt}}}{\sqrt{M(k+1)}}, \tag{10.82}$$

where M is given in (10.80)*;*

(d) *all limit points of the sequence $\{\mathbf{x}^k\}_{k\geq 0}$ are stationary points of problem* (10.81).

Proof. (a) By Lemma 10.66,

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq M \|\nabla f(\mathbf{x}^k)\|_*^2, \tag{10.83}$$

where $M > 0$ is given in (10.80). The inequality (10.83) readily implies that $f(\mathbf{x}^k) \geq f(\mathbf{x}^{k+1})$ and that if $\nabla f(\mathbf{x}^k) \neq \mathbf{0}$, then $f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k)$. Finally, if $\nabla f(\mathbf{x}^k) = \mathbf{0}$, then $\mathbf{x}^k = \mathbf{x}^{k+1}$, and hence $f(\mathbf{x}^k) = f(\mathbf{x}^{k+1})$.

(b) Since the sequence $\{f(\mathbf{x}^k)\}_{k\geq 0}$ is nonincreasing and bounded below, it converges. Thus, in particular $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \to 0$ as $k \to \infty$, which, combined with (10.83), implies that $\nabla f(\mathbf{x}^k) \to \mathbf{0}$ as $k \to \infty$.

(c) By Lemma 10.66, for any $n \geq 0$,

$$f(\mathbf{x}^n) - f(\mathbf{x}^{n+1}) \geq M\|\nabla f(\mathbf{x}^n)\|_*^2.$$

Summing the above over $n = 0, 1, \ldots, k$, we obtain

$$f(\mathbf{x}^0) - f(\mathbf{x}^{k+1}) \geq M \sum_{n=0}^{k} \|\nabla f(\mathbf{x}^n)\|_*^2 \geq (k+1)M \min_{n=0,1,\ldots,k} \|\nabla f(\mathbf{x}^n)\|_*^2.$$

Using the fact that $f(\mathbf{x}^{k+1}) \geq f_{\text{opt}}$, the inequality (10.82) follows.

(d) Let $\bar{\mathbf{x}}$ be a limit point of $\{\mathbf{x}^k\}_{k\geq 0}$. Then there exists a subsequence $\{\mathbf{x}^{k_j}\}_{j\geq 0}$ converging to $\bar{\mathbf{x}}$. For any $j \geq 0$,

$$\|\nabla f(\bar{\mathbf{x}})\|_* \leq \|\nabla f(\mathbf{x}^{k_j}) - \nabla f(\bar{\mathbf{x}})\|_* + \|\nabla f(\mathbf{x}^{k_j})\|_* \leq L_f\|\mathbf{x}^{k_j} - \bar{\mathbf{x}}\| + \|\nabla f(\mathbf{x}^{k_j})\|_*. \tag{10.84}$$

Since the right-hand side of (10.84) goes to 0 as $j \to \infty$, it follows that $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$. ∎

Convergence Analysis in the Convex Case

To establish a rate of convergence in the case where f is convex, we will require an additional boundedness-type assumption. We gather all the required assumptions in the following.

Assumption 10.68.

(A) *$f : \mathbb{E} \to \mathbb{R}$ is L_f-smooth and convex.*

(B) *The optimal set of the problem*

$$\min_{\mathbf{x}\in\mathbb{E}} f(\mathbf{x})$$

is nonempty and denoted by X^*. The optimal value is denoted by f_{opt}.

(C) *For any $\alpha > 0$, there exists $R_\alpha > 0$ such that*

$$\max_{\mathbf{x},\mathbf{x}^*}\{\|\mathbf{x}^* - \mathbf{x}\| : f(\mathbf{x}) \leq \alpha, \mathbf{x}^* \in X^*\} \leq R_\alpha.$$

The proof of the convergence rate is based on the following very simple lemma.

Lemma 10.69. *Suppose that Assumption* 10.68 *holds. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the non-Euclidean gradient method for solving the problem of minimizing f over $\mathbb{E}$ with either a constant stepsize corresponding to $L_k = \bar{L} \in \left(\frac{L_f}{2}, \infty\right)$; a stepsize chosen by the backtracking procedure* B4 *with parameters (s, γ, η) satisfying $s > 0, \gamma \in (0, 1), \eta > 1$; or an exact line search for computing the stepsize. Then*

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{C}(f(\mathbf{x}^k) - f_{\text{opt}})^2, \tag{10.85}$$

where

$$C = \begin{cases} \frac{R_\alpha^2 \bar{L}^2}{\bar{L} - \frac{L_f}{2}}, & \textit{constant stepsize}, \\ \frac{R_\alpha^2}{\gamma} \max\left\{s, \frac{\eta L_f}{2(1-\gamma)}\right\}, & \textit{backtracking}, \\ 2R_\alpha^2 L_f, & \textit{exact line search}, \end{cases} \tag{10.86}$$

with $\alpha = f(\mathbf{x}^0)$.

Proof. Note that, by Theorem 10.67(a), $\{f(\mathbf{x}^k)\}_{k\geq 0}$ is nonincreasing, and in particular for any $k \geq 0$ it holds that $f(\mathbf{x}^k) \leq f(\mathbf{x}^0)$. Therefore, for any $\mathbf{x}^* \in X^*$ and $k \geq 0$,

$$\|\mathbf{x}^k - \mathbf{x}^*\| \leq R_\alpha,$$

where $\alpha = f(\mathbf{x}^0)$. To prove (10.85), we note that on the one hand, by Lemma 10.66,

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq M\|\nabla f(\mathbf{x}^k)\|_*^2, \tag{10.87}$$

where M is given in (10.80). On the other hand, by the gradient inequality along with the generalized Cauchy–Schwarz inequality (Lemma 1.4), for any $\mathbf{x}^* \in X^*$,

$$\begin{aligned} f(\mathbf{x}^k) - f_{\text{opt}} &= f(\mathbf{x}^k) - f(\mathbf{x}^*) \\ &\leq \langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^* \rangle \\ &\leq \|\nabla f(\mathbf{x}^k)\|_* \|\mathbf{x}^k - \mathbf{x}^*\| \\ &\leq R_\alpha \|\nabla f(\mathbf{x}^k)\|_*. \end{aligned} \tag{10.88}$$

Combining (10.87) and (10.88), we obtain that

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq M\|\nabla f(\mathbf{x}^k)\|_*^2 \geq \frac{M}{R_\alpha^2}(f(\mathbf{x}^k) - f_{\text{opt}})^2.$$

Plugging the expression for M given in (10.80) into the above inequality, the result (10.85) is established. ∎

To derive the rate of convergence in function values, we will use the following lemma on convergence of nonnegative scalar sequences.

Lemma 10.70. *Let* $\{a_k\}_{k\geq 0}$ *be a sequence of nonnegative real numbers satisfying for any* $k \geq 0$

$$a_k - a_{k+1} \geq \frac{1}{\gamma} a_k^2$$

for some $\gamma > 0$. *Then for any* $k \geq 1$,

$$a_k \leq \frac{\gamma}{k}. \tag{10.89}$$

Proof. Let k be a positive integer. If $a_k = 0$, then obviously (10.89) holds. Suppose that $a_k > 0$. Then by the monotonicity of $\{a_n\}_{n\geq 0}$, we have that $a_0, a_1, \ldots, a_k > 0$. For any $n = 1, 2, \ldots, k$,

$$\frac{1}{a_n} - \frac{1}{a_{n-1}} = \frac{a_{n-1} - a_n}{a_{n-1}a_n} \geq \frac{1}{\gamma} \frac{a_{n-1}^2}{a_{n-1}a_n} = \frac{1}{\gamma} \frac{a_{n-1}}{a_n} \geq \frac{1}{\gamma}, \tag{10.90}$$

where the last inequality follows from the monotonicity of the sequence. Summing (10.90) over $n = 1, 2, \ldots, k$, we obtain

$$\frac{1}{a_k} \geq \frac{1}{a_0} + \frac{k}{\gamma} \geq \frac{k}{\gamma},$$

proving (10.89). □

Combining Lemmas 10.69 and 10.70, we can establish an $O(1/k)$ rate of convergence in function values of the sequence generated by the non-Euclidean gradient method.

Theorem 10.71 ($O(1/k)$ rate of convergence of the non-Euclidean gradient method). *Under the setting of Lemma* 10.69*, for any* $k \geq 1$*,*

$$f(\mathbf{x}^k) - f_{\text{opt}} \leq \frac{C}{k}, \tag{10.91}$$

where C *is given in* (10.86)*.*

Proof. By Lemma 10.69,

$$a_k - a_{k+1} \geq \frac{1}{C} a_k^2,$$

where $a_k = f(\mathbf{x}^k) - f_{\text{opt}}$. Invoking Lemma 10.70 with $\gamma = C$, the inequality $a_k \leq \frac{C}{k}$, which is the same as (10.91), follows. □

Remark 10.72. *When a constant stepsize* $\frac{1}{L_f}$ *is used (meaning that* $L_k \equiv \bar{L} \equiv L_f$*),* (10.91) *has the form*

$$f(\mathbf{x}^k) - f_{\text{opt}} \leq \frac{2R_\alpha^2 L_f}{k},$$

which is similar in form to the result in the Euclidean setting in which the following bound was derived (see Theorem 10.21*):*

$$f(\mathbf{x}^k) - f_{\text{opt}} \leq \frac{L_f \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2k}.$$

The Non-Euclidean Gradient Method in $\mathbb{R}^n$ Endowed with the l_1-Norm

Example 10.73. Suppose that the underlying space is $\mathbb{R}^n$ endowed with the l_1-norm, and let f be an L_f-smooth function w.r.t. the l_1-norm. Recall (see Example 10.63) that the set of primal counterparts in this case is given for any $\mathbf{a} \neq \mathbf{0}$ by

$$\Lambda_{\mathbf{a}} = \left\{ \sum_{i \in I(\mathbf{a})} \lambda_i \text{sgn}(a_i) \mathbf{e}_i : \sum_{i \in I(\mathbf{a})} \lambda_i = 1, \lambda_j \geq 0, j \in I(\mathbf{a}) \right\},$$

where $I(\mathbf{a}) = \text{argmax}_{i=1,2,\ldots,n} |a_i|$. When employing the method, we can always choose $\mathbf{a}^\dagger = \text{sgn}(a_i)\mathbf{e}_i$ for some arbitrary $i \in I(\mathbf{a})$. The method thus takes the following form:

Non-Euclidean Gradient under the l_1-Norm

- **Initialization:** pick $\mathbf{x}^0 \in \mathbb{R}^n$.
- **General step:** for any $k = 0, 1, 2, \ldots$ execute the following steps:
  - pick $i_k \in \operatorname{argmax}_i \left| \frac{\partial f(\mathbf{x}^k)}{\partial x_i} \right|$;
  - set $\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{\|\nabla f(\mathbf{x}^k)\|_\infty}{L_k} \operatorname{sgn}\left(\frac{\partial f(\mathbf{x}^k)}{\partial x_{i_k}} \right) \mathbf{e}_{i_k}$.

The constants L_k can be chosen by either one of the three options: a constant stepsize rule $L_k \equiv \bar{L} \in \left(\frac{L_f}{2}, \infty\right)$, the backtracking procedure B4, or an exact line search. Note that at each iteration only one coordinate is altered. This is a variant of a coordinate descent method that actually has an interpretation as a non-Euclidean gradient method. ∎

Example 10.74. Consider the problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} \right\},$$

where $\mathbf{A} \in \mathbb{S}^n_{++}$ and $\mathbf{b} \in \mathbb{R}^n$. The underlying space is $\mathbb{E} = \mathbb{R}^n$ endowed with the l_p-norm ($p \in [1, \infty]$). By Example 5.2, f is $L_f^{(p)}$-smooth with

$$L_f^{(p)} = \|\mathbf{A}\|_{p,q} = \max_{\mathbf{x}} \{ \|\mathbf{A}\mathbf{x}\|_q : \|\mathbf{x}\|_p \leq 1 \}$$

with $q \in [1, \infty]$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Two examples of smoothness parameters are the following:

- $p = 2$. In this case, since $\mathbf{A}$ is positive definite, $L_f^{(2)} = \|\mathbf{A}\|_{2,2} = \lambda_{\max}(\mathbf{A})$.
- $p = 1$. Here $L_f^{(1)} = \|\mathbf{A}\|_{1,\infty} = \max_{i,j} |A_{i,j}|$.

The non-Euclidean gradient method for $p = 2$ is actually the Euclidean gradient method; taking a constant stepsize corresponding to $L_k = L_f^{(2)} = \lambda_{\max}(\mathbf{A})$, the method takes the following form:

Algorithm G2

- **Initialization:** pick $\mathbf{x}^0 \in \mathbb{R}^n$.
- **General step ($k \geq 0$):** $\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L_f^{(2)}} (\mathbf{A}\mathbf{x}^k + \mathbf{b})$.

In the case $p = 1$ the method is a coordinate descent-type method, and with a constant stepsize corresponding to $L_k = L_f^{(1)} = \max_{i,j} |A_{i,j}|$ it takes the following form:

Algorithm G1

- **Initialization:** pick $\mathbf{x}^0 \in \mathbb{R}^n$.
- **General step ($k \geq 0$):**
  - pick $i_k \in \operatorname{argmax}_{i=1,2,\ldots,n} |\mathbf{A}_i \mathbf{x}^k + b_i|$, where $\mathbf{A}_i$ denotes the ith row of $\mathbf{A}$.
  - update $\mathbf{x}_j^{k+1} = \begin{cases} \mathbf{x}_j^k, & j \neq i_k, \\ \mathbf{x}_{i_k}^k - \frac{1}{L_f^{(1)}}(\mathbf{A}_{i_k}\mathbf{x}^k + b_{i_k}), & j = i_k. \end{cases}$

By Theorem 10.71,[62]

$$f(\mathbf{x}^k) - f_{\text{opt}} \leq \frac{2L_f^{(p)} R_{f(\mathbf{x}^0)}^2}{k}.$$

Therefore, the ratio $\frac{L_f^{(2)}}{L_f^{(1)}}$ might indicate which of the methods should have an advantage over the other. ∎

Remark 10.75. *Note that Algorithm* G2 *(from Example* 10.74*) requires $O(n^2)$ operations at each iteration since the matrix/vector multiplication $\mathbf{A}\mathbf{x}^k$ is computed. On the other hand, a careful implementation of Algorithm* G1 *will only require $O(n)$ operations at each iteration; this can be accomplished by updating the gradient $\mathbf{g}^k \equiv \mathbf{A}\mathbf{x}^k + \mathbf{b}$ using the relation $\mathbf{g}^{k+1} = \mathbf{g}^k - \frac{\mathbf{A}_{i_k}\mathbf{x}^k + b_{i_k}}{L_f^{(1)}} \mathbf{A}\mathbf{e}_{i_k}$ ($\mathbf{A}\mathbf{e}_{i_k}$ is obviously the i_kth column of $\mathbf{A}$). Therefore, a fair comparison between Algorithms* G1 *and* G2 *will count each n iterations of algorithm* G1 *as "one iteration." We will call such an iteration a "meta-iteration."*

Example 10.76. Continuing Example 10.74, consider, for example, the matrix $\mathbf{A} = \mathbf{A}^{(d)} \equiv \mathbf{J} + d\mathbf{I}$, where the matrix $\mathbf{J}$ is the matrix of all ones. Then for any $d > 0$, $\mathbf{A}^{(d)}$ is positive definite and $\lambda_{\max}(\mathbf{A}^{(d)}) = d+n$, $\max_{i,j} |A_{i,j}^{(d)}| = d+1$. Therefore, as the ratio $\rho_f \equiv \frac{L_f^{(2)}}{L_f^{(1)}} = \frac{d+n}{d+1}$ gets larger, the Euclidean gradient method (Algorithm G2) should become more inferior to the non-Euclidean version (Algorithm G1).

We ran the two algorithms for the choice $\mathbf{A} = \mathbf{A}^{(2)}$ and $\mathbf{b} = 10\mathbf{e}_1$ with initial point $\mathbf{x}^0 = \mathbf{e}_n$. The values $f(\mathbf{x}^k) - f_{\text{opt}}$ as a function of the iteration index k are plotted in Figures 10.4 and 10.5 for $n = 10$ and $n = 100$, respectively. As can be seen in the left images of both figures, when meta-iterations of algorithm G1 are compared with iterations of algorithm G2, the superiority of algorithm G1 is significant. We also made the comparison when each iteration of algorithm G1 is just an update of one coordinate, meaning that we do not consider meta-iterations. For $n = 10$, the methods behave similarly, and there does not seem to be any preference to G1 or G2. However, when $n = 100$, there is still a substantial advantage of algorithm G1 compared to G2, despite the fact that it is a much cheaper method w.r.t. the number of operations performed per iteration. A possible reason for this

[62] Note that also $R_{f(\mathbf{x}^0)}$ might depend on the choice of norm.

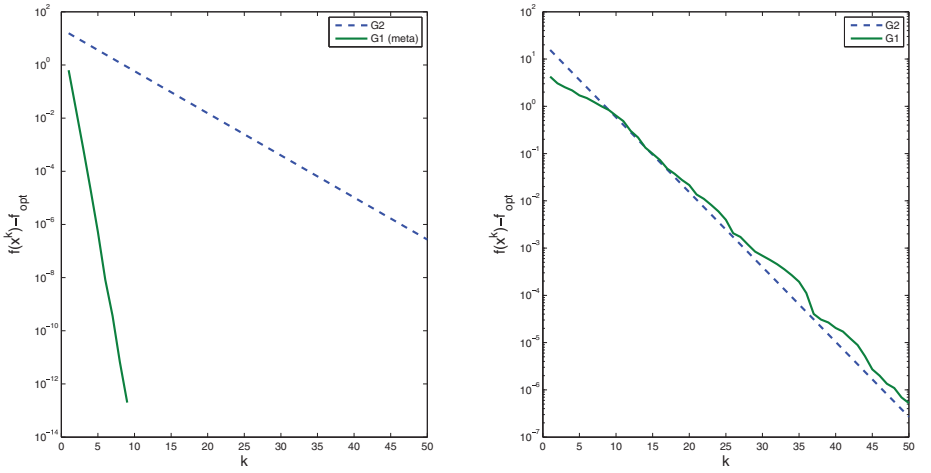

Figure 10.4. *Comparison of the Euclidean gradient method* (G2) *with the non-Euclidean gradient method* (G1) *applied on the problem from Example* 10.76 *with* $n = 10$. *The left image considers "meta-iterations" of* G1, *meaning that* 10 *iterations of* G1 *are counted as one iteration, while the right image counts each coordinate update as one iteration.*

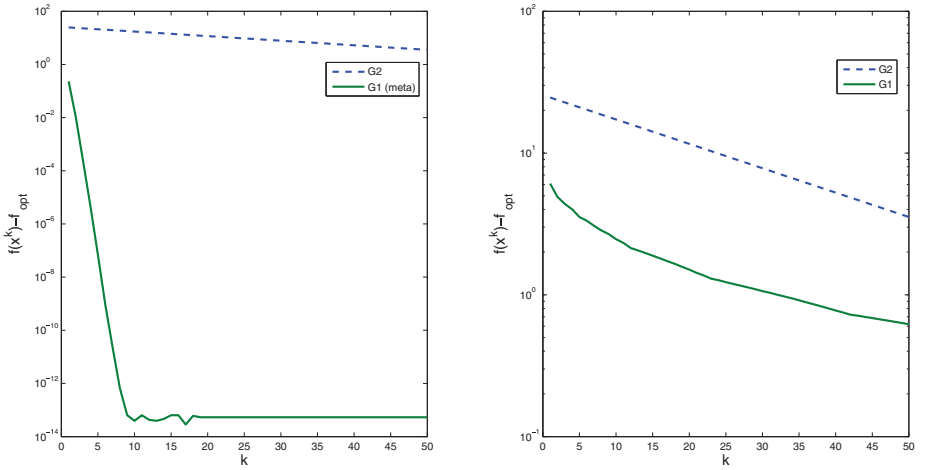

Figure 10.5. *Comparison of the Euclidean gradient method* (G2) *with the non-Euclidean gradient method* (G1) *applied on the problem from Example* 10.76 *with* $n = 100$. *The left image considers "meta-iterations" of* G1, *meaning that* 100 *iterations of* G1 *are counted as one iteration, while the right image counts each coordinate update as one iteration.*

is the fact that for $n = 10$, $\rho_f = \frac{2+10}{2+1} = 4$, while for $n = 100$, $\frac{2+100}{2+1} = 34$, and hence it is expected that the advantage of algorithm G1 over algorithm G2 will be more substantial when $n = 100$. ∎

10.9.2 The Non-Euclidean Proximal Gradient Method[63]

In this section we return to the composite model

$$\min_{\mathbf{x}\in\mathbb{E}}\{F(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x})\}, \tag{10.92}$$

where the endowed norm on $\mathbb{E}$ is not assume to be Euclidean. Our main objective will be to develop a non-Euclidean version of the proximal gradient method. We note that when $g \equiv 0$, the method will *not* coincide with the non-Euclidean gradient method discussed in Section 10.9.1, meaning that the approach described here, which is similar to the generalization of projected subgradient to mirror descent (see Chapter 9), is fundamentally different than the approach considered in the non-Euclidean gradient method. We will make the following assumption.

Assumption 10.77.

(A) $g : \mathbb{E} \to (-\infty, \infty]$ *is proper closed and convex.*

(B) $f : \mathbb{E} \to (-\infty, \infty]$ *is proper closed and convex;* $\mathrm{dom}(g) \subseteq \mathrm{int}(\mathrm{dom}(f))$ *and* f *is* L_f*-smooth over* $\mathrm{int}(\mathrm{dom}(f))$.

(C) *The optimal solution of problem* (10.1) *is nonempty and denoted by* X^*. *The optimal value of the problem is denoted by* F_{opt}.

In the Euclidean setting, the general update rule of the proximal gradient method (see the discussion in Section 10.2) can be written in the following form:

$$\mathbf{x}^{k+1} = \mathrm{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + g(\mathbf{x}) + \frac{L_k}{2}\|\mathbf{x} - \mathbf{x}^k\|^2 \right\}.$$

We will use the same idea as in the mirror descent method and replace the half-squared Euclidean distance with a Bregman distance, leading to the following update rule:

$$\mathbf{x}^{k+1} = \mathrm{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + g(\mathbf{x}) + L_k B_\omega(\mathbf{x}, \mathbf{x}^k) \right\},$$

where B_ω is the Bregman distance associated with ω (see Definition 9.2). The function ω will satisfy the following properties.

Assumption 10.78 (properties of ω).

- ω *is proper closed and convex.*
- ω *is differentiable over* $\mathrm{dom}(\partial\omega)$.
- $\mathrm{dom}(g) \subseteq \mathrm{dom}(\omega)$.
- $\omega + \delta_{\mathrm{dom}(g)}$ *is* 1*-strongly convex.*

[63] The non-Euclidean proximal gradient method presented in Section 10.9.2 was analyzed in the work of Tseng [121].

The proximal gradient method is defined below.

The Non-Euclidean Proximal Gradient Method

Initialization: pick $\mathbf{x}^0 \in \mathrm{dom}(g) \cap \mathrm{dom}(\partial\omega)$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick $L_k > 0$;

(b) compute

$$\mathbf{x}^{k+1} = \mathrm{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ \left\langle \frac{1}{L_k}\nabla f(\mathbf{x}^k) - \nabla\omega(\mathbf{x}^k), \mathbf{x} \right\rangle + \frac{1}{L_k} g(\mathbf{x}) + \omega(\mathbf{x}) \right\}. \tag{10.93}$$

Our first observation is that under Assumptions 10.77 and 10.78, the non-Euclidean proximal gradient method is well defined, meaning that if $\mathbf{x}^k \in \mathrm{dom}(g) \cap \mathrm{dom}(\partial\omega)$, then the minimization problem in (10.93) has a unique optimal solution in $\mathrm{dom}(g) \cap \mathrm{dom}(\partial\omega)$. This is a direct result of Lemma 9.7 invoked with $\psi(\mathbf{x}) = \left\langle \frac{1}{L_k}\nabla f(\mathbf{x}^k) - \nabla\omega(\mathbf{x}^k), \mathbf{x} \right\rangle + \frac{1}{L_k} g(\mathbf{x})$. The two stepsize rules that will be analyzed are detailed below. We use the notation

$$V_L(\bar{\mathbf{x}}) \equiv \mathrm{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ \left\langle \frac{1}{L}\nabla f(\bar{\mathbf{x}}) - \nabla\omega(\bar{\mathbf{x}}), \mathbf{x} \right\rangle + \frac{1}{L} g(\mathbf{x}) + \omega(\mathbf{x}) \right\}.$$

- **Constant.** $L_k = \bar{L} = L_f$ for all k.
- **Backtracking procedure B5.** The procedure requires two parameters (s, η), where $s > 0$ and $\eta > 1$. Define $L_{-1} = s$. At iteration k ($k \geq 0$) the choice of L_k is done as follows: First, L_k is set to be equal to L_{k-1}. Then, while

$$f(V_{L_k}(\mathbf{x}^k)) > f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), V_{L_k}(\mathbf{x}^k) - \mathbf{x}^k \rangle + \frac{L_k}{2}\|V_{L_k}(\mathbf{x}^k) - \mathbf{x}^k\|^2,$$

set $L_k := \eta L_k$. In other words, the stepsize is chosen as $L_k = L_{k-1}\eta^{i_k}$, where i_k is the smallest nonnegative integer for which the condition

$$f(V_{L_{k-1}\eta^{i_k}}(\mathbf{x}^k)) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), V_{L_{k-1}\eta^{i_k}}(\mathbf{x}^k) - \mathbf{x}^k \rangle + \frac{L_k}{2}\|V_{L_{k-1}\eta^{i_k}}(\mathbf{x}^k) - \mathbf{x}^k\|^2$$

is satisfied.

Remark 10.79. *In both stepsize rules the following inequality holds:*

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L_k}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2.$$

Remark 10.80. *By the same arguments as in Remark* 10.19 *we have that* $L_k \leq \alpha L_f$*, where* $\alpha = 1$ *for the constant stepsize case and* $\alpha = \max\left\{\eta, \frac{s}{L_f}\right\}$ *in the setting of the backtracking procedure* B5*.*

The rate of convergence result will now be stated and proved.

Theorem 10.81 ($O(1/k)$ rate of convergence of the non-Euclidean proximal gradient method). *Suppose that Assumptions* 10.77 *and* 10.78 *hold. Let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by the non-Euclidean proximal gradient method for solving problem* (10.92) *with either a constant stepsize rule in which* $L_k \equiv L_f$ *for all* $k \geq 0$ *or the backtracking procedure* B5*. Then*

(a) *the sequence* $\{F(\mathbf{x}^k)\}_{k\geq 0}$ *is nonincreasing;*

(b) *for any* $k \geq 1$ *and* $\mathbf{x}^* \in X^*$*,*

$$F(\mathbf{x}^k) - F_{\text{opt}} \leq \frac{\alpha L_f B_\omega(\mathbf{x}^*, \mathbf{x}^0)}{k},$$

where $\alpha = 1$ *in the constant stepsize setting and* $\alpha = \max\left\{\eta, \frac{s}{L_f}\right\}$ *if the backtracking rule is employed.*

Proof. (a) We will use the notation $m(\mathbf{x}, \mathbf{y}) \equiv f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle$. For both stepsize rules we have, for any $n \geq 0$ (see Remark 10.79),

$$f(\mathbf{x}^{n+1}) \leq m(\mathbf{x}^{n+1}, \mathbf{x}^n) + \frac{L_n}{2}\|\mathbf{x}^{n+1} - \mathbf{x}^n\|^2.$$

Therefore,

$$\begin{aligned} F(\mathbf{x}^{n+1}) &= f(\mathbf{x}^{n+1}) + g(\mathbf{x}^{n+1}) \\ &\leq m(\mathbf{x}^{n+1}, \mathbf{x}^n) + g(\mathbf{x}^{n+1}) + \frac{L_n}{2}\|\mathbf{x}^{n+1} - \mathbf{x}^n\|^2 \\ &\leq m(\mathbf{x}^{n+1}, \mathbf{x}^n) + g(\mathbf{x}^{n+1}) + L_n B_\omega(\mathbf{x}^{n+1}, \mathbf{x}^n), \end{aligned} \tag{10.94}$$

where the 1-strong convexity of $\omega + \delta_{\text{dom}(g)}$ was used in the last inequality. Note that

$$\mathbf{x}^{n+1} = \text{argmin}_{\mathbf{x}\in\mathbb{E}}\{m(\mathbf{x}, \mathbf{x}^n) + g(\mathbf{x}) + L_n B_\omega(\mathbf{x}, \mathbf{x}^n)\}. \tag{10.95}$$

Therefore, in particular,

$$\begin{aligned} m(\mathbf{x}^{n+1}, \mathbf{x}^n) + g(\mathbf{x}^{n+1}) + L_n B_\omega(\mathbf{x}^{n+1}, \mathbf{x}^n) &\leq m(\mathbf{x}^n, \mathbf{x}^n) + g(\mathbf{x}^n) + L_n B_\omega(\mathbf{x}^n, \mathbf{x}^n) \\ &= f(\mathbf{x}^n) + g(\mathbf{x}^n) \\ &= F(\mathbf{x}^n), \end{aligned}$$

which, combined with (10.94), implies that $F(\mathbf{x}^{n+1}) \leq F(\mathbf{x}^n)$, meaning that the sequence of function values $\{F(\mathbf{x}^n)\}_{n\geq 0}$ is nonincreasing.

(b) Let $k \geq 1$ and $\mathbf{x}^* \in X^*$. Using the relation (10.95) and invoking the non-Euclidean second prox theorem (Theorem 9.12) with $\psi(\mathbf{x}) = \frac{m(\mathbf{x},\mathbf{x}^n)+g(\mathbf{x})}{L_n}$, $\mathbf{b} = \mathbf{x}^n$, and $\mathbf{a} = \mathbf{x}^{n+1}$, it follows that for all $\mathbf{x} \in \text{dom}(g)$,

$$\langle \nabla\omega(\mathbf{x}^n) - \nabla\omega(\mathbf{x}^{n+1}), \mathbf{x} - \mathbf{x}^{n+1}\rangle \leq \frac{m(\mathbf{x}, \mathbf{x}^n) - m(\mathbf{x}^{n+1}, \mathbf{x}^n) + g(\mathbf{x}) - g(\mathbf{x}^{n+1})}{L_n},$$

which, combined with the three-points lemma (Lemma 9.11) with $\mathbf{a} = \mathbf{x}^{n+1}$, $\mathbf{b} = \mathbf{x}^n$, and $\mathbf{c} = \mathbf{x}$, yields the inequality

$$B_\omega(\mathbf{x}, \mathbf{x}^{n+1}) + B_\omega(\mathbf{x}^{n+1}, \mathbf{x}^n) - B_\omega(\mathbf{x}, \mathbf{x}^n) \leq \frac{m(\mathbf{x}, \mathbf{x}^n) - m(\mathbf{x}^{n+1}, \mathbf{x}^n) + g(\mathbf{x}) - g(\mathbf{x}^{n+1})}{L_n}.$$

Rearranging terms, we obtain that

$$\begin{aligned} m(\mathbf{x}^{n+1}, \mathbf{x}^n) + g(\mathbf{x}^{n+1}) + L_n B_\omega(\mathbf{x}^{n+1}, \mathbf{x}^n) \leq m(\mathbf{x}, \mathbf{x}^n) + g(\mathbf{x}) + L_n B_\omega(\mathbf{x}, \mathbf{x}^n) \\ - L_n B_\omega(\mathbf{x}, \mathbf{x}^{n+1}), \end{aligned}$$

which, combined with (10.94), yields the inequality

$$F(\mathbf{x}^{n+1}) \leq m(\mathbf{x}, \mathbf{x}^n) + g(\mathbf{x}) + L_n B_\omega(\mathbf{x}, \mathbf{x}^n) - L_n B_\omega(\mathbf{x}, \mathbf{x}^{n+1}).$$

Since f is convex, $m(\mathbf{x}, \mathbf{x}^n) \leq f(\mathbf{x})$, and hence

$$F(\mathbf{x}^{n+1}) - F(\mathbf{x}) \leq L_n B_\omega(\mathbf{x}, \mathbf{x}^n) - L_n B_\omega(\mathbf{x}, \mathbf{x}^{n+1}).$$

Plugging in $\mathbf{x} = \mathbf{x}^*$ and dividing by L_n, we obtain

$$\frac{F(\mathbf{x}^{n+1}) - F(\mathbf{x}^*)}{L_n} \leq B_\omega(\mathbf{x}^*, \mathbf{x}^n) - B_\omega(\mathbf{x}^*, \mathbf{x}^{n+1}).$$

Using the bound $L_n \leq \alpha L_f$ (see Remark 10.80),

$$\frac{F(\mathbf{x}^{n+1}) - F(\mathbf{x}^*)}{\alpha L_f} \leq B_\omega(\mathbf{x}^*, \mathbf{x}^n) - B_\omega(\mathbf{x}^*, \mathbf{x}^{n+1}),$$

and hence

$$F(\mathbf{x}^{n+1}) - F_{\text{opt}} \leq \alpha L_f B_\omega(\mathbf{x}^*, \mathbf{x}^n) - \alpha L_f B_\omega(\mathbf{x}^*, \mathbf{x}^{n+1}).$$

Summing the above inequality for $n = 0, 1, \ldots, k-1$, we obtain that

$$\sum_{n=0}^{k-1} (F(\mathbf{x}^{n+1}) - F_{\text{opt}}) \leq \alpha L_f B_\omega(\mathbf{x}^*, \mathbf{x}^0) - \alpha L_f B_\omega(\mathbf{x}^*, \mathbf{x}^k) \leq \alpha L_f B_\omega(\mathbf{x}^*, \mathbf{x}^0).$$

Using the monotonicity of the sequence of function values, we conclude that

$$k(F(\mathbf{x}^k) - F_{\text{opt}}) \leq \alpha L_f B_\omega(\mathbf{x}^*, \mathbf{x}^0),$$

thus obtaining the result

$$F(\mathbf{x}^k) - F_{\text{opt}} \leq \frac{\alpha L_f B_\omega(\mathbf{x}^*, \mathbf{x}^0)}{k}. \qquad \square$$

Chapter 11

The Block Proximal Gradient Method

Underlying Spaces: In this chapter, all the underlying spaces are Euclidean (see the details in Section 11.2).

11.1 Decomposition Methods

Many of the methods discussed in this book are *decomposition methods*, which, loosely speaking, are methods that utilize at each step only a certain portion of the problem's data or resort to solving a smaller-dimension problem at each step. One class of decomposition methods is the class of *functional decomposition methods*, in which the data of the problem comprise several functions, and at each iteration only a few of them (perhaps only one) are processed. Examples of functional decomposition methods were studied in the context of the model

$$\min_{\mathbf{x}} \left\{ \sum_{i=1}^{m} f_i(\mathbf{x}) : \mathbf{x} \in C \right\}.$$

In Example 8.36 it was shown that an implementation of the stochastic projected subgradient method amounts to a method of the form

$$\mathbf{x}^{k+1} = P_C(\mathbf{x}^k - t_k f'_{i_k}(\mathbf{x}^k)),$$

where the index i_k is picked randomly by a uniform distribution. A deterministic version of this method is the incremental projected subgradient method, which was studied in Section 8.4, in which i_k is picked by a cyclic order. In both methods, each step exploits only one of the m functions that constitute the data of the problem. The proximal gradient method is actually another example of a functional decomposition method, where the relevant model (see Chapter 10) is

$$\min_{\mathbf{x}\in\mathbb{E}} f(\mathbf{x}) + g(\mathbf{x}).$$

The general step of the proximal gradient method is of the form

$$\mathbf{x}^{k+1} = \mathrm{prox}_{t_k g}(\mathbf{x}^k - t_k \nabla f(\mathbf{x}^k)).$$

The functions f and g are treated separately in the above update formula. First, a gradient step w.r.t. f is taken, and then a prox operator w.r.t. g is computed.

Another class of decomposition methods is the class of *variables decomposition methods*, in which at each iteration only a subset of the decision variables is altered while all the other variables remain fixed. One example for such a method was given in Example 10.73, where the problem of minimizing a differentiable function over $\mathbb{R}^n$ was considered. The method described in Example 10.73 (non-Euclidean gradient method under the l_1-norm) picks one variable at each iteration by a certain greedy rule and performs a gradient step w.r.t. the chosen variable while keeping all the other variables fixed.

In this chapter we will consider additional variables decomposition methods; these methods pick at each iteration one block of variables and perform a proximal gradient step w.r.t. the chosen block.

11.2 Model and Assumptions

In this chapter we will consider methods for solving the composite model $f + g$ in the case where g has a block separable structure. More specifically, the main model of this chapter is

$$\min_{\mathbf{x}_1 \in \mathbb{E}_1, \mathbf{x}_2 \in \mathbb{E}_2, \ldots, \mathbf{x}_p \in \mathbb{E}_p} \left\{ F(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) = f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) + \sum_{j=1}^{p} g_j(\mathbf{x}_j) \right\}, \quad (11.1)$$

where $\mathbb{E}_1, \mathbb{E}_2, \ldots, \mathbb{E}_p$ are Euclidean spaces. We will denote the product space by $\mathbb{E} = \mathbb{E}_1 \times \mathbb{E}_2 \times \cdots \times \mathbb{E}_p$ and use our convention (see Section 1.9) that the product space is also Euclidean with endowed norm

$$\|(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p)\|_{\mathbb{E}} = \sqrt{\sum_{i=1}^{p} \|\mathbf{u}_i\|_{\mathbb{E}_i}^2}.$$

In most cases we will omit the subscript of the norm indicating the underlying vector space (whose identity will be clear from the context). The function $g : \mathbb{E} \to (-\infty, \infty]$ is defined by

$$g(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) \equiv \sum_{i=1}^{p} g_i(\mathbf{x}_i).$$

The gradient w.r.t. the ith block ($i \in \{1, 2, \ldots, p\}$) is denoted by $\nabla_i f$, and whenever the function is differentiable it holds that

$$\nabla f(\mathbf{x}) = (\nabla_1 f(\mathbf{x}), \nabla_2 f(\mathbf{x}), \ldots, \nabla_p f(\mathbf{x})).$$

For any $i \in \{1, 2, \ldots, p\}$ we define $\mathcal{U}_i : \mathbb{E}_i \to \mathbb{E}$ to be the linear transformation given by

$$\mathcal{U}_i(\mathbf{d}) = (\mathbf{0}, \ldots, \mathbf{0}, \underbrace{\mathbf{d}}_{i\text{th block}}, \mathbf{0}, \ldots, \mathbf{0}), \quad \mathbf{d} \in \mathbb{E}_i.$$

We also use throughout this chapter the notation that a vector $\mathbf{x} \in \mathbb{E}$ can be written as

$$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p),$$

and this relation will also be written as $\mathbf{x} = (\mathbf{x}_i)_{i=1}^p$. Thus, in our notation, the main model (11.1) can be simply written as

$$\min_{\mathbf{x}\in\mathbb{E}}\{F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})\}.$$

The basic assumptions on the model are summarized below.

Assumption 11.1.

(A) $g_i : \mathbb{E}_i \to (-\infty, \infty]$ *is proper closed and convex for any* $i \in \{1, 2, \ldots, p\}$.

(B) $f : \mathbb{E} \to (-\infty, \infty]$ *is proper and closed, and* $\mathrm{dom}(f)$ *is convex;* $\mathrm{dom}(g) \subseteq \mathrm{int}(\mathrm{dom}(f))$, *and* f *is differentiable over* $\mathrm{int}(\mathrm{dom}(f))$.

(C) f *is* L_f*-smooth over* $\mathrm{int}(\mathrm{dom}(f))$ *(*$L_f > 0$*).*

(D) *There exist* $L_1, L_2, \ldots, L_p > 0$ *such that for any* $i \in \{1, 2, \ldots, p\}$ *it holds that*

$$\|\nabla_i f(\mathbf{x}) - \nabla_i f(\mathbf{x} + \mathcal{U}_i(\mathbf{d}))\| \le L_i\|\mathbf{d}\| \tag{11.2}$$

for all $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$ *and* $\mathbf{d} \in \mathbb{E}_i$ *for which* $\mathbf{x} + \mathcal{U}_i(\mathbf{d}) \in \mathrm{int}(\mathrm{dom}(f))$.

(E) *The optimal set of problem* (11.1) *is nonempty and denoted by* X^*. *The optimal value is denoted by* F_{opt}.

Remark 11.2 (block/global Lipschitz constants). *The constant* L_f *will be called the "global Lipschitz constant," while the constants* $L_1, L_2, \ldots, L_p$ *are the "block Lipschitz constants." Obviously, we can choose* $L_i = L_f$ *for all* i *since by the definition of* L_f, (11.2) *holds for* $L_i = L_f$. *However, the block Lipschitz constants can be significantly smaller than the global Lipschitz constant—a fact that might have significant influence on the performance of the derived algorithms, as well as their convergence rate.*

11.3 The Toolbox

11.3.1 The Partial Gradient Mapping

Recall that the gradient mapping associated with the functions f, g and a constant $L > 0$, as defined in Section 10.3.2, is a mapping from $\mathrm{int}(\mathrm{dom}(f))$ to $\mathbb{E}$ given by

$$G_L^{f,g}(\mathbf{x}) = L\left(\mathbf{x} - T_L^{f,g}(\mathbf{x})\right),$$

where $T_L^{f,g} : \mathrm{int}(\mathrm{dom}(f)) \to \mathbb{E}$ is the prox-grad mapping given by

$$T_L^{f,g}(\mathbf{x}) = \mathrm{prox}_{\frac{1}{L}g}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right).$$

From now on we will always omit the superscripts and write T_L and G_L instead of $T_L^{f,g}$ and $G_L^{f,g}$. In the context of block variables decomposition methods, it is also important to consider the notions of *partial prox-grad mappings* and *partial gradient mappings*.

Definition 11.3 (partial prox-grad mapping). *Suppose that f and $g_1, g_2, \ldots, g_p$ satisfy properties* (A) *and* (B) *of Assumption* 11.1*, $L > 0$, and let $i \in \{1, 2, \ldots, p\}$. Then the* ***i*th partial prox-grad mapping** *is the operator $T_L^i : \mathrm{int}(\mathrm{dom}(f)) \to \mathbb{E}_i$ defined by*

$$T_L^i(\mathbf{x}) = \mathrm{prox}_{\frac{1}{L}g_i}\left(\mathbf{x}_i - \frac{1}{L}\nabla_i f(\mathbf{x})\right).$$

Definition 11.4 (partial gradient mapping). *Suppose that f and $g_1, g_2, \ldots, g_p$ satisfy properties* (A) *and* (B) *of Assumption* 11.1*, $L > 0$, and let $i \in \{1, 2, \ldots, p\}$. Then the* ***i*th partial gradient mapping** *is the operator $G_L^i : \mathrm{int}(\mathrm{dom}(f)) \to \mathbb{E}_i$ defined by*

$$G_L^i(\mathbf{x}) = L\left(\mathbf{x}_i - T_L^i(\mathbf{x})\right).$$

The ith partial prox-grad and gradient mappings depend on f and g_i, but this dependence is not indicated in our notation. If $g_i \equiv 0$ for some $i \in \{1, 2, \ldots, p\}$, then $G_L^i(\mathbf{x}) = \nabla_i f(\mathbf{x})$; that is, in this case the partial gradient mapping coincides with the mapping $\mathbf{x} \mapsto \nabla_i f(\mathbf{x})$. Some basic properties of the partial prox-grad and gradient mappings are summarized in the following lemma.

Lemma 11.5. *Suppose that f and $g_1, g_2, \ldots, g_p$ satisfy properties* (A) *and* (B) *of Assumption* 11.1*, $L > 0$, and let $i \in \{1, 2, \ldots, p\}$. Then for any $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$,*

$$\begin{aligned} T_L(\mathbf{x}) &= (T_L^1(\mathbf{x}), T_L^2(\mathbf{x}), \ldots, T_L^p(\mathbf{x})), \\ G_L(\mathbf{x}) &= (G_L^1(\mathbf{x}), G_L^2(\mathbf{x}), \ldots, G_L^p(\mathbf{x})). \end{aligned} \tag{11.3}$$

Proof. By Theorem 6.6, we have that for any $\mathbf{y} \in \mathrm{dom}(f)$,

$$\mathrm{prox}_{\frac{1}{L}g}(\mathbf{y}) = (\mathrm{prox}_{\frac{1}{L}g_i}(\mathbf{y}_i))_{i=1}^p.$$

Thus, for any $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$,

$$\begin{aligned} T_L(\mathbf{x}) &= \mathrm{prox}_{\frac{1}{L}g}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right) = \left(\mathrm{prox}_{\frac{1}{L}g_i}\left(\left[\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right]_i\right)\right)_{i=1}^p \\ &= \left(\mathrm{prox}_{\frac{1}{L}g_i}\left(\mathbf{x}_i - \frac{1}{L}\nabla_i f(\mathbf{x})\right)\right)_{i=1}^p \\ &= (T_L^i(\mathbf{x}))_{i=1}^p. \end{aligned}$$

The second identity follows immediately:

$$\begin{aligned} G_L(\mathbf{x}) &= L(\mathbf{x} - T_L(\mathbf{x})) = L\left((\mathbf{x}_i)_{i=1}^p - (T_L^i(\mathbf{x}))_{i=1}^p\right) \\ &= \left(L(\mathbf{x}_i - T_L^i(\mathbf{x}))\right)_{i=1}^p \\ &= \left(G_L^i(\mathbf{x})\right)_{i=1}^p. \end{aligned}$$ ∎

A point $\mathbf{x}^* \in \mathrm{dom}(g)$ is a stationary point of problem (11.1) if $-\nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*)$ (see Definition 3.73). The following simple theorem shows that the stationarity condition for problem (11.1) can be decomposed into p conditions expressed in terms of the partial gradient mappings.

Theorem 11.6. *Suppose that f and $g_1, g_2, \ldots, g_p$ satisfy properties* (A) *and* (B) *of Assumption* 11.1. *Then*

(a) $\mathbf{x}^* \in \mathrm{dom}(g)$ *is a stationary point of problem* (11.1) *if and only if*

$$-\nabla_i f(\mathbf{x}^*) \in \partial g_i(\mathbf{x}_i^*),\ i = 1, 2, \ldots, p; \tag{11.4}$$

(b) *for any p positive numbers $M_1, M_2, \ldots, M_p > 0$, $\mathbf{x}^* \in \mathrm{dom}(g)$ is a stationary point of problem* (11.1) *if and only if*

$$G^i_{M_i}(\mathbf{x}^*) = \mathbf{0},\ i = 1, 2, \ldots, p.$$

Proof. (a) By definition, $\mathbf{x}^* \in \mathrm{dom}(g)$ is a stationary point of problem (11.1) if and only if

$$-\nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*). \tag{11.5}$$

By the block separable structure of g, it is easy to show that

$$\partial g(\mathbf{x}^*) = \partial g_1(\mathbf{x}_1^*) \times \partial g_2(\mathbf{x}_2^*) \times \cdots \times \partial g_p(\mathbf{x}_p^*),$$

which, combined with the fact that $\nabla f(\mathbf{x}^*) = (\nabla_1 f(\mathbf{x}^*), \nabla_2 f(\mathbf{x}^*), \ldots, \nabla_p f(\mathbf{x}^*))$, implies that the relation (11.5) is equivalent to

$$-(\nabla_1 f(\mathbf{x}^*), \nabla_2 f(\mathbf{x}^*), \ldots, \nabla_p f(\mathbf{x}^*)) \in \partial g_1(\mathbf{x}_1^*) \times \partial g_2(\mathbf{x}_2^*) \times \cdots \times \partial g_p(\mathbf{x}_p^*),$$

that is, to (11.4).

(b) By the definition of the partial gradient mapping, $G^i_{M_i}(\mathbf{x}^*) = \mathbf{0}$ if and only if $\mathbf{x}_i^* = \mathrm{prox}_{\frac{1}{M_i} g_i}\big(\mathbf{x}_i^* - \frac{1}{M_i}\nabla_i f(\mathbf{x}^*)\big)$, which, by the second prox theorem (Theorem 6.39), is equivalent to

$$\left(\mathbf{x}_i^* - \frac{1}{M_i}\nabla_i f(\mathbf{x}^*)\right) - \mathbf{x}_i^* \in \frac{1}{M_i}\partial g_i(\mathbf{x}_i^*),$$

that is, to

$$-\nabla_i f(\mathbf{x}^*) \in \partial g_i(\mathbf{x}_i^*).$$

To summarize, $G^i_{M_i}(\mathbf{x}^*) = \mathbf{0}$ for all i if and only if $-\nabla_i f(\mathbf{x}^*) \in \partial g_i(\mathbf{x}_i^*)$ for all i, which, by part (a), is equivalent to saying that $\mathbf{x}^*$ is a stationary point of problem (11.1). ∎

The next results shows some monotonicity properties of the partial gradient mapping w.r.t. its parameter. The result is presented without its proof, which is an almost verbatim repetition of the arguments in Theorem 10.9.

Theorem 11.7 (monotonicity of the partial gradient mapping). *Suppose that f and $g_1, g_2, \ldots, g_p$ satisfy properties* (A) *and* (B) *of Assumption* 11.1, *and let $i \in \{1, 2, \ldots, p\}$. Suppose that $L_1 \geq L_2 > 0$. Then*

$$\|G^i_{L_1}(\mathbf{x})\| \geq \|G^i_{L_2}(\mathbf{x})\|$$

and

$$\frac{\|G^i_{L_1}(\mathbf{x})\|}{L_1} \leq \frac{\|G^i_{L_2}(\mathbf{x})\|}{L_2}$$

for any $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$.

11.3.2 The Block Descent Lemma

The block descent lemma is a variant of the descent lemma (Lemma 5.7), and its proof is almost identical.

Lemma 11.8 (block descent lemma). *Let $f : \mathbb{E}_1 \times \mathbb{E}_2 \times \cdots \times \mathbb{E}_p \to (-\infty, \infty]$ be a proper function whose domain $\mathrm{dom}(f)$ is convex. Assume that f is differentiable over $\mathrm{int}(\mathrm{dom}(f))$. Let $i \in \{1, 2, \ldots, p\}$. Suppose that there exists $L_i > 0$ for which*

$$\|\nabla_i f(\mathbf{y}) - \nabla_i f(\mathbf{y} + \mathcal{U}_i(\mathbf{d}))\| \leq L_i \|\mathbf{d}\|$$

for any $\mathbf{y} \in \mathrm{int}(\mathrm{dom}(f))$ and $\mathbf{d} \in \mathbb{E}_i$ for which $\mathbf{y} + \mathcal{U}_i(\mathbf{d}) \in \mathrm{int}(\mathrm{dom}(f))$. Then

$$f(\mathbf{x} + \mathcal{U}_i(\mathbf{d})) \leq f(\mathbf{x}) + \langle \nabla_i f(\mathbf{x}), \mathbf{d} \rangle + \frac{L_i}{2} \|\mathbf{d}\|^2$$

for any $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$ and $\mathbf{d} \in \mathbb{E}_i$ for which $\mathbf{x} + \mathcal{U}_i(\mathbf{d}) \in \mathrm{int}(\mathrm{dom}(f))$.

Proof. Let $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$ and $\mathbf{d} \in \mathbb{E}_i$ such that $\mathbf{x} + \mathcal{U}_i(\mathbf{d}) \in \mathrm{int}(\mathrm{dom}(f))$. Denote $\mathbf{x}^{(t)} = \mathbf{x} + t\mathcal{U}_i(\mathbf{d})$ and define $g(t) = f(\mathbf{x}^{(t)})$. By the fundamental theorem of calculus,

$$\begin{aligned} f(\mathbf{x}^{(1)}) - f(\mathbf{x}) = g(1) - g(0) &= \int_0^1 g'(t) dt \\ &= \int_0^1 \langle \nabla f(\mathbf{x}^{(t)}), \mathcal{U}_i(\mathbf{d}) \rangle dt = \int_0^1 \langle \nabla_i f(\mathbf{x}^{(t)}), \mathbf{d} \rangle dt \\ &= \langle \nabla_i f(\mathbf{x}), \mathbf{d} \rangle + \int_0^1 \langle \nabla_i f(\mathbf{x}^{(t)}) - \nabla_i f(\mathbf{x}), \mathbf{d} \rangle dt. \end{aligned}$$

Thus,

$$\begin{aligned} |f(\mathbf{x}^{(1)}) - f(\mathbf{x}) - \langle \nabla_i f(\mathbf{x}), \mathbf{d} \rangle| &= \left| \int_0^1 \langle \nabla_i f(\mathbf{x}^{(t)}) - \nabla_i f(\mathbf{x}), \mathbf{d} \rangle dt \right| \\ &\leq \int_0^1 |\langle \nabla_i f(\mathbf{x}^{(t)}) - \nabla_i f(\mathbf{x}), \mathbf{d} \rangle| dt \\ &\overset{(*)}{\leq} \int_0^1 \|\nabla_i f(\mathbf{x}^{(t)}) - \nabla_i f(\mathbf{x})\| \cdot \|\mathbf{d}\| dt \\ &\leq \int_0^1 t L_i \|\mathbf{d}\|^2 dt \\ &= \frac{L_i}{2} \|\mathbf{d}\|^2, \end{aligned}$$

where the Cauchy–Schwarz inequality was used in $(*)$. ∎

11.3.3 Sufficient Decrease

The basic step that will be employed by all the methods discussed in this chapter is a proximal gradient step w.r.t. a given block. Specifically, for a given $\mathbf{x} \in \mathbb{E}$ and

$i \in \{1, 2, \ldots, p\}$, the next updated vector $\mathbf{x}^+$ will have the form

$$\mathbf{x}_j^+ = \begin{cases} \mathbf{x}_j, & j \neq i, \\ T_{L_i}^i(\mathbf{x}), & j = i. \end{cases}$$

The above update formula can be compactly written as

$$\mathbf{x}^+ = \mathbf{x} + \mathcal{U}_i(T_{L_i}^i(\mathbf{x}) - \mathbf{x}_i).$$

We will now prove a variant of the sufficient decrease lemma (Lemma 10.4), in which only Lipschitz continuity w.r.t. a certain block of the gradient of the function is assumed.

Lemma 11.9 (block sufficient decrease lemma). *Suppose that f and $g_1, g_2, \ldots, g_p$ satisfy properties* (A) *and* (B) *of Assumption* 11.1. *Let $i \in \{1, 2, \ldots, p\}$. Suppose that there exists $L_i > 0$ for which*

$$\|\nabla_i f(\mathbf{y}) - \nabla_i f(\mathbf{y} + \mathcal{U}_i(\mathbf{d}))\| \leq L_i \|\mathbf{d}\|$$

for any $\mathbf{y} \in \operatorname{int}(\operatorname{dom}(f))$ and $\mathbf{d} \in \mathbb{E}_i$ for which $\mathbf{y} + \mathcal{U}_i(\mathbf{d}) \in \operatorname{int}(\operatorname{dom}(f))$. Then

$$F(\mathbf{x}) - F(\mathbf{x} + \mathcal{U}_i(T_{L_i}^i(\mathbf{x}) - \mathbf{x}_i)) \geq \frac{1}{2L_i} \|G_{L_i}^i(\mathbf{x})\|^2 \tag{11.6}$$

for all $\mathbf{x} \in \operatorname{int}(\operatorname{dom}(f))$.

Proof. For the sake of simplicity, we use the shorthand notation $\mathbf{x}^+ = \mathbf{x} + \mathcal{U}_i(T_{L_i}^i(\mathbf{x}) - \mathbf{x}_i)$. By the block descent lemma (Lemma 11.8), we have that

$$f(\mathbf{x}^+) \leq f(\mathbf{x}) + \left\langle \nabla_i f(\mathbf{x}), T_{L_i}^i(\mathbf{x}) - \mathbf{x}_i \right\rangle + \frac{L_i}{2} \|T_{L_i}^i(\mathbf{x}) - \mathbf{x}_i\|^2. \tag{11.7}$$

By the second prox theorem (Theorem 6.39), since $T_{L_i}^i(\mathbf{x}) = \operatorname{prox}_{\frac{1}{L_i} g_i}\big(\mathbf{x}_i - \frac{1}{L_i}\nabla_i f(\mathbf{x})\big)$, we have

$$\left\langle \mathbf{x}_i - \frac{1}{L_i}\nabla_i f(\mathbf{x}) - T_{L_i}^i(\mathbf{x}), \mathbf{x}_i - T_{L_i}^i(\mathbf{x}) \right\rangle \leq \frac{1}{L_i} g_i(\mathbf{x}_i) - \frac{1}{L_i} g_i(T_{L_i}^i(\mathbf{x})),$$

and hence

$$\left\langle \nabla_i f(\mathbf{x}), T_{L_i}^i(\mathbf{x}) - \mathbf{x}_i \right\rangle \leq -L_i \left\| T_{L_i}^i(\mathbf{x}) - \mathbf{x}_i \right\|^2 + g_i(\mathbf{x}_i) - g_i(\mathbf{x}_i^+),$$

which, combined with (11.7), yields

$$f(\mathbf{x}^+) + g_i(\mathbf{x}_i^+) \leq f(\mathbf{x}) + g_i(\mathbf{x}_i) - \frac{L_i}{2} \left\| T_{L_i}^i(\mathbf{x}) - \mathbf{x}_i \right\|^2.$$

Adding the identity $\sum_{j \neq i} g_j(\mathbf{x}_j^+) = \sum_{j \neq i} g_j(\mathbf{x}_j)$ to the last inequality yields

$$F(\mathbf{x}^+) \leq F(\mathbf{x}) - \frac{L_i}{2} \left\| T_{L_i}^i(\mathbf{x}) - \mathbf{x}_i \right\|^2,$$

which, by the definition of the partial gradient mapping, is equivalent to the desired result (11.6). □

Remark 11.10. *Under the setting of Lemma* 11.9*, if we denote* $\mathbf{x}^+ = \mathbf{x} + \mathcal{U}_i(T^i_{L_i}(\mathbf{x}) - \mathbf{x}_i)$*, then the sufficient decrease condition* (11.6) *can be written in the following form:*

$$F(\mathbf{x}) - F(\mathbf{x}^+) \geq \frac{L_i}{2}\|\mathbf{x} - \mathbf{x}^+\|^2.$$

11.4 The Cyclic Block Proximal Gradient Method

In the cyclic block proximal gradient (CBPG) method we successively pick a block in a cyclic manner and perform a prox-grad step w.r.t. the chosen block. The kth iterate is denoted by $\mathbf{x}^k = (\mathbf{x}^k_1, \mathbf{x}^k_2, \ldots, \mathbf{x}^k_p)$. Each iteration of the CBPG method involves p "subiterations," and the by-products of these subiterations will be denoted by the following auxiliary subsequences:

$$\begin{aligned}
\mathbf{x}^{k,0} &= \mathbf{x}^k = (\mathbf{x}^k_1, \mathbf{x}^k_2, \ldots, \mathbf{x}^k_p),\\
\mathbf{x}^{k,1} &= (\mathbf{x}^{k+1}_1, \mathbf{x}^k_2, \ldots, \mathbf{x}^k_p),\\
\mathbf{x}^{k,2} &= (\mathbf{x}^{k+1}_1, \mathbf{x}^{k+1}_2, \mathbf{x}^k_3, \ldots, \mathbf{x}^k_p),\\
&\vdots\\
\mathbf{x}^{k,p} &= \mathbf{x}^{k+1} = (\mathbf{x}^{k+1}_1, \mathbf{x}^{k+1}_2, \ldots, \mathbf{x}^{k+1}_p).
\end{aligned}$$

We can also write the following formula for the kth member of the ith auxiliary sequence:

$$\mathbf{x}^{k,i} = \sum_{j=1}^{i} \mathcal{U}_j(\mathbf{x}^{k+1}_j) + \sum_{j=i+1}^{p} \mathcal{U}_j(\mathbf{x}^k_j). \tag{11.8}$$

We are now ready to present the method.

The Cyclic Block Proximal Gradient (CBPG) Method

Initialization: pick $\mathbf{x}^0 = (\mathbf{x}^0_1, \mathbf{x}^0_2, \ldots, \mathbf{x}^0_p) \in \operatorname{int}(\operatorname{dom}(f))$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

- set $\mathbf{x}^{k,0} = \mathbf{x}^k$;
- for $i = 1, 2, \ldots, p$, compute

$$\mathbf{x}^{k,i} = \mathbf{x}^{k,i-1} + \mathcal{U}_i(T^i_{L_i}(\mathbf{x}^{k,i-1}) - \mathbf{x}^{k,i-1}_i);$$

- set $\mathbf{x}^{k+1} = \mathbf{x}^{k,p}$.

11.4.1 Convergence Analysis of the CBPG Method—The Nonconvex Case

The convergence analysis of the CBPG method relies on the following technical lemma, which is a direct consequence of the sufficient decrease property of Lemma 11.9.

Lemma 11.11 (sufficient decrease of the CBPG method—version I). *Suppose that Assumption* 11.1 *holds, and let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by the CBPG method for solving problem* (11.1) *with the auxiliary sequences defined in* (11.8). *Then*

(a) *for all* $k \geq 0$ *and* $j \in \{0, 1, \ldots, p-1\}$ *it holds that*

$$F(\mathbf{x}^{k,j}) - F(\mathbf{x}^{k,j+1}) \geq \frac{1}{2L_{j+1}} \|G^{j+1}_{L_{j+1}}(\mathbf{x}^{k,j})\|^2, \tag{11.9}$$

or equivalently,

$$F(\mathbf{x}^{k,j}) - F(\mathbf{x}^{k,j+1}) \geq \frac{L_{j+1}}{2} \|\mathbf{x}^{k,j} - \mathbf{x}^{k,j+1}\|^2; \tag{11.10}$$

(b) *for all* $k \geq 0$,

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq \frac{L_{\min}}{2} \|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2, \tag{11.11}$$

where $L_{\min} = \min_{i=1,2,\ldots,p} L_i$.

Proof. (a) Inequality (11.9) follows by invoking Lemma 11.9 with $\mathbf{x} = \mathbf{x}^{k,j}$ and $i = j + 1$. The result (11.10) now follows by the identity $\|\mathbf{x}^{k,j} - \mathbf{x}^{k,j+1}\|^2 = \|T^{j+1}_{L_{j+1}}(\mathbf{x}^{k,j}) - \mathbf{x}^k_{j+1}\|^2 = \frac{1}{L^2_{j+1}} \|G^{j+1}_{L_{j+1}}(\mathbf{x}^{k,j})\|^2$.

(b) Summing the inequality (11.10) over $j = 0, 1, \ldots, p-1$, we obtain

$$\begin{aligned} F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) &= \sum_{j=0}^{p-1} (F(\mathbf{x}^{k,j}) - F(\mathbf{x}^{k,j+1})) \geq \sum_{j=0}^{p-1} \frac{L_{j+1}}{2} \|\mathbf{x}^{k,j} - \mathbf{x}^{k,j+1}\|^2 \\ &= \sum_{j=0}^{p-1} \frac{L_{j+1}}{2} \|\mathbf{x}^k_{j+1} - \mathbf{x}^{k+1}_{j+1}\|^2 \geq \frac{L_{\min}}{2} \sum_{j=0}^{p-1} \|\mathbf{x}^k_{j+1} - \mathbf{x}^{k+1}_{j+1}\|^2 \\ &= \frac{L_{\min}}{2} \|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2. \quad \square \end{aligned}$$

A direct result of the last lemma is the monotonicity in function values of the sequence generated by the CBPG method.

Corollary 11.12 (monotonicity of the sequence generated by the CBPG method). *Under the setting of Lemma* 11.11, *for any* $k \geq 0$, $F(\mathbf{x}^{k+1}) \leq F(\mathbf{x}^k)$, *and equality holds if and only if* $\mathbf{x}^k = \mathbf{x}^{k+1}$.

We can now prove a sufficient decrease property of the CBPG method in terms of the (nonpartial) gradient mapping.

Lemma 11.13 (sufficient decrease of the CBPG method—version II). *Suppose that Assumption* 11.1 *holds, and let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by the CBPG method for solving problem* (11.1). *Then for any* $k \geq 0$,

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq \frac{C}{p}\|G_{L_{\min}}(\mathbf{x}^k)\|^2, \tag{11.12}$$

where

$$C = \frac{L_{\min}}{2(L_f + 2L_{\max} + \sqrt{L_{\min}L_{\max}})^2} \tag{11.13}$$

and

$$L_{\min} = \min_{i=1,2,\ldots,p} L_i,\ L_{\max} = \max_{i=1,2,\ldots,p} L_i.$$

Proof. Let $i \in \{0,1,\ldots,p-1\}$. By (11.9),

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq F(\mathbf{x}^{k,i}) - F(\mathbf{x}^{k,i+1}) \geq \frac{1}{2L_{i+1}}\|G^{i+1}_{L_{i+1}}(\mathbf{x}^{k,i})\|^2. \tag{11.14}$$

We can bound $\|G^{i+1}_{L_{i+1}}(\mathbf{x}^k)\|$ as follows:

$$\begin{aligned}
\|G^{i+1}_{L_{i+1}}(\mathbf{x}^k)\| &\leq \|G^{i+1}_{L_{i+1}}(\mathbf{x}^k) - G^{i+1}_{L_{i+1}}(\mathbf{x}^{k,i})\| + \|G^{i+1}_{L_{i+1}}(\mathbf{x}^{k,i})\| && \text{[triangle inequality]}\\
&\leq \|G_{L_{i+1}}(\mathbf{x}^k) - G_{L_{i+1}}(\mathbf{x}^{k,i})\| + \|G^{i+1}_{L_{i+1}}(\mathbf{x}^{k,i})\| && \text{[(11.3)]}\\
&\leq (2L_{i+1} + L_f)\|\mathbf{x}^k - \mathbf{x}^{k,i}\| + \|G^{i+1}_{L_{i+1}}(\mathbf{x}^{k,i})\| && \text{[Lemma 10.10(a)]}\\
&\leq (2L_{i+1} + L_f)\|\mathbf{x}^k - \mathbf{x}^{k+1}\| + \|G^{i+1}_{L_{i+1}}(\mathbf{x}^{k,i})\|,
\end{aligned}$$

where the last inequality follows by the following argument:

$$\|\mathbf{x}^k - \mathbf{x}^{k,i}\| = \sqrt{\sum_{j=1}^{i}\|\mathbf{x}_j^k - \mathbf{x}_j^{k+1}\|^2} \leq \sqrt{\sum_{j=1}^{p}\|\mathbf{x}_j^k - \mathbf{x}_j^{k+1}\|^2} = \|\mathbf{x}^k - \mathbf{x}^{k+1}\|.$$

Using the inequalities (11.11) and (11.14), it follows that we can continue to bound $\|G^{i+1}_{L_{i+1}}(\mathbf{x}^k)\|$ as follows:

$$\begin{aligned}
\|G^{i+1}_{L_{i+1}}(\mathbf{x}^k)\| &\leq (2L_{i+1} + L_f)\|\mathbf{x}^k - \mathbf{x}^{k+1}\| + \|G^{i+1}_{L_{i+1}}(\mathbf{x}^{k,i})\|\\
&\leq \left[\frac{\sqrt{2}(2L_{i+1} + L_f)}{\sqrt{L_{\min}}} + \sqrt{2L_{i+1}}\right]\sqrt{F(\mathbf{x}^k) - F(\mathbf{x}^{k+1})}\\
&\overset{L_{i+1}\leq L_{\max}}{\leq} \sqrt{\frac{2}{L_{\min}}}(L_f + 2L_{\max} + \sqrt{L_{\min}L_{\max}})\sqrt{F(\mathbf{x}^k) - F(\mathbf{x}^{k+1})}.
\end{aligned}$$

By the monotonicity of the partial gradient mapping (Theorem 11.7), it follows that $\|G^{i+1}_{L_{\min}}(\mathbf{x}^k)\| \leq \|G^{i+1}_{L_{i+1}}(\mathbf{x}^k)\|$, and hence, for any $i \in \{0,1,\ldots,p-1\}$,

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq C\|G^{i+1}_{L_{\min}}(\mathbf{x}^k)\|^2,$$

where C is given in (11.13). We can thus conclude that

$$\|G_{L_{\min}}(\mathbf{x}^k)\|^2 = \sum_{i=0}^{p-1} \|G^{i+1}_{L_{\min}}(\mathbf{x}^k)\|^2 \le \sum_{i=0}^{p-1} \frac{F(\mathbf{x}^k) - F(\mathbf{x}^{k+1})}{C} = \frac{p}{C}(F(\mathbf{x}^k) - F(\mathbf{x}^{k+1})),$$

which is the same as (11.12). ∎

Equipped with Lemma 11.13, it is easy to show some standard convergence properties of the CBPG method.

Theorem 11.14 (convergence of the CBPG method—nonconvex case). *Suppose that Assumption* 11.1 *holds, and let* $\{\mathbf{x}^k\}_{k\ge 0}$ *be the sequence generated by the CBPG method for solving problem* (11.1). *Denote*

$$L_{\min} = \min_{i=1,2,\dots,p} L_i,\ L_{\max} = \max_{i=1,2,\dots,p} L_i,$$

and let C *be given in* (11.13). *Then*

(a) $G_{L_{\min}}(\mathbf{x}^k) \to \mathbf{0}$ *as* $k \to \infty$;

(b) $\min_{n=0,1,\dots,k} \|G_{L_{\min}}(\mathbf{x}^n)\| \le \frac{\sqrt{p(F(\mathbf{x}^0) - F_{\text{opt}})}}{\sqrt{C(k+1)}}$;

(c) *all limit points of the sequence* $\{\mathbf{x}^k\}_{k\ge 0}$ *are stationary points of problem* (11.1).

Proof. (a) Since the sequence $\{F(\mathbf{x}^k)\}_{k\ge 0}$ is nonincreasing (Corollary 11.12) and bounded below (by Assumption 11.1(E)), it converges. Thus, in particular $F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \to 0$ as $k \to \infty$, which, combined with (11.12), implies that $\|G_{L_{\min}}(\mathbf{x}^k)\| \to 0$ as $k \to \infty$.

(b) By Lemma 11.13, for any $n \ge 0$,

$$F(\mathbf{x}^n) - F(\mathbf{x}^{n+1}) \ge \frac{C}{p}\|G_{L_{\min}}(\mathbf{x}^n)\|^2. \tag{11.15}$$

Summing the above inequality over $n = 0, 1, \dots, k$, we obtain

$$F(\mathbf{x}^0) - F(\mathbf{x}^{k+1}) \ge \frac{C}{p}\sum_{n=0}^{k} \|G_{L_{\min}}(\mathbf{x}^n)\|^2 \ge \frac{C(k+1)}{p} \min_{n=0,1,\dots,k} \|G_{L_{\min}}(\mathbf{x}^n)\|^2.$$

Using the fact that $F(\mathbf{x}^{k+1}) \ge F_{\text{opt}}$, the result follows.

(c) Let $\bar{\mathbf{x}}$ be a limit point of $\{\mathbf{x}^k\}_{k\ge 0}$. Then there exists a subsequence $\{\mathbf{x}^{k_j}\}_{j\ge 0}$ converging to $\bar{\mathbf{x}}$. For any $j \ge 0$,

$$\begin{aligned}\|G_{L_{\min}}(\bar{\mathbf{x}})\| &\le \|G_{L_{\min}}(\mathbf{x}^{k_j}) - G_{L_{\min}}(\bar{\mathbf{x}})\| + \|G_{L_{\min}}(\mathbf{x}^{k_j})\| \\ &\le (2L_{\min} + L_f)\|\mathbf{x}^{k_j} - \bar{\mathbf{x}}\| + \|G_{L_{\min}}(\mathbf{x}^{k_j})\|, \end{aligned} \tag{11.16}$$

where Lemma 10.10(a) was used in the last inequality. Since the expression in (11.16) goes to 0 as $j \to \infty$, it follows that $G_{L_{\min}}(\bar{\mathbf{x}}) = \mathbf{0}$, which, by Theorem 10.7(b), implies that $\bar{\mathbf{x}}$ is a stationary point of problem (11.1). ∎

11.4.2 Convergence Analysis of the CBPG Method—The Convex Case[64]

We will now show a rate of convergence in function values of the CBPG method in the case where f is assumed to be convex and a certain boundedness property of the level sets of F holds.

Assumption 11.15.

(A) *f is convex.*

(B) *For any $\alpha > 0$, there exists $R_\alpha > 0$ such that*

$$\max_{\mathbf{x},\mathbf{x}^* \in \mathbb{E}} \{\|\mathbf{x} - \mathbf{x}^*\| : F(\mathbf{x}) \le \alpha, \mathbf{x}^* \in X^*\} \le R_\alpha.$$

The analysis in the convex case is based on the following key lemma describing a recursive inequality relation of the sequence of function values.

Lemma 11.16. *Suppose that Assumptions* 11.1 *and* 11.15 *hold. Let $\{\mathbf{x}^k\}_{k\ge 0}$ be the sequence generated by the CBPG method for solving problem* (11.1)*. Then for any $k \ge 0$,*

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \ge \frac{L_{\min}}{2p(L_f + L_{\max})^2 R^2}(F(\mathbf{x}^{k+1}) - F_{\text{opt}})^2,$$

where $R = R_{F(\mathbf{x}^0)}$, $L_{\max} = \max_{j=1,2,\ldots,p} L_j$, and $L_{\min} = \min_{j=1,2,\ldots,p} L_j$.

Proof. Let $\mathbf{x}^* \in X^*$. By the definition of the CBPG method, for any $k \ge 0$ and $j \in \{1, 2, \ldots, p\}$,

$$\mathbf{x}_j^{k,j} = \text{prox}_{\frac{1}{L_j} g_j}\left(\mathbf{x}_j^{k,j-1} - \frac{1}{L_j}\nabla_j f(\mathbf{x}^{k,j-1})\right).$$

Thus, invoking the second prox theorem (Theorem 6.39), for any $\mathbf{y} \in \mathbb{E}_j$,

$$g_j(\mathbf{y}) \ge g_j(\mathbf{x}_j^{k,j}) + L_j\left\langle \mathbf{x}_j^{k,j-1} - \frac{1}{L_j}\nabla_j f(\mathbf{x}^{k,j-1}) - \mathbf{x}_j^{k,j}, \mathbf{y} - \mathbf{x}_j^{k,j}\right\rangle.$$

By the definition of the auxiliary sequences given in (11.8), $\mathbf{x}_j^{k,j-1} = \mathbf{x}_j^k$, $\mathbf{x}_j^{k,j} = \mathbf{x}_j^{k+1}$, and therefore

$$g_j(\mathbf{y}) \ge g_j(\mathbf{x}_j^{k+1}) + L_j\left\langle \mathbf{x}_j^k - \frac{1}{L_j}\nabla_j f(\mathbf{x}^{k,j-1}) - \mathbf{x}_j^{k+1}, \mathbf{y} - \mathbf{x}_j^{k+1}\right\rangle.$$

Thus, in particular, if we substitute $\mathbf{y} = \mathbf{x}_j^*$,

$$g_j(\mathbf{x}_j^*) \ge g_j(\mathbf{x}_j^{k+1}) + L_j\left\langle \mathbf{x}_j^k - \frac{1}{L_j}\nabla_j f(\mathbf{x}^{k,j-1}) - \mathbf{x}_j^{k+1}, \mathbf{x}_j^* - \mathbf{x}_j^{k+1}\right\rangle.$$

[64] The type of analysis in Section 11.4.2 originates from Beck and Tetruashvili [22], who studied the case in which the nonsmooth functions are indicators. The extension to the general composite model can be found in Shefi and Teboulle [115] and Hong, Wang, Razaviyayn, and Luo [69].

Summing the above inequality over $j = 1, 2, \ldots, p$ yields the inequality

$$g(\mathbf{x}^*) \geq g(\mathbf{x}^{k+1}) + \sum_{j=1}^{p} L_j \left\langle \mathbf{x}_j^k - \frac{1}{L_j}\nabla_j f(\mathbf{x}^{k,j-1}) - \mathbf{x}_j^{k+1}, \mathbf{x}_j^* - \mathbf{x}_j^{k+1} \right\rangle. \quad (11.17)$$

We can now utilize the convexity of f and write

$$\begin{aligned} F(\mathbf{x}^{k+1}) - F(\mathbf{x}^*) &= f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*) + g(\mathbf{x}^{k+1}) - g(\mathbf{x}^*) \\ &\leq \langle \nabla f(\mathbf{x}^{k+1}), \mathbf{x}^{k+1} - \mathbf{x}^* \rangle + g(\mathbf{x}^{k+1}) - g(\mathbf{x}^*) \\ &= \sum_{j=1}^{p} \left\langle \nabla_j f(\mathbf{x}^{k+1}), \mathbf{x}_j^{k+1} - \mathbf{x}_j^* \right\rangle + g(\mathbf{x}^{k+1}) - g(\mathbf{x}^*), \end{aligned}$$

which, combined with (11.17), implies

$$\begin{aligned} F(\mathbf{x}^{k+1}) - F(\mathbf{x}^*) &\leq \sum_{j=1}^{p} \left\langle \nabla_j f(\mathbf{x}^{k+1}), \mathbf{x}_j^{k+1} - \mathbf{x}_j^* \right\rangle \\ &\quad + \sum_{j=1}^{p} L_j \left\langle \mathbf{x}_j^k - \frac{1}{L_j}\nabla_j f(\mathbf{x}^{k,j-1}) - \mathbf{x}_j^{k+1}, \mathbf{x}_j^{k+1} - \mathbf{x}_j^* \right\rangle \\ &= \sum_{j=1}^{p} \left\langle \nabla_j f(\mathbf{x}^{k+1}) - \nabla_j f(\mathbf{x}^{k,j-1}) + L_j(\mathbf{x}_j^k - \mathbf{x}_j^{k+1}), \mathbf{x}_j^{k+1} - \mathbf{x}_j^* \right\rangle. \end{aligned}$$

Using the Cauchy–Schwarz and triangle inequalities, we can conclude that

$$\begin{aligned} F(\mathbf{x}^{k+1}) - F(\mathbf{x}^*) &\leq \sum_{j=1}^{p} \left(\|\nabla_j f(\mathbf{x}^{k+1}) - \nabla_j f(\mathbf{x}^{k,j-1})\| + L_j \|\mathbf{x}_j^k - \mathbf{x}_j^{k+1}\| \right) \|\mathbf{x}_j^{k+1} - \mathbf{x}_j^*\| \\ &\leq \sum_{j=1}^{p} \left(\|\nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^{k,j-1})\| + L_j \|\mathbf{x}_j^k - \mathbf{x}_j^{k+1}\| \right) \|\mathbf{x}_j^{k+1} - \mathbf{x}_j^*\| \\ &\leq \sum_{j=1}^{p} \left(L_f \|\mathbf{x}^{k+1} - \mathbf{x}^{k,j-1}\| + L_{\max} \|\mathbf{x}^k - \mathbf{x}^{k+1}\| \right) \|\mathbf{x}_j^{k+1} - \mathbf{x}_j^*\| \\ &\leq (L_f + L_{\max}) \|\mathbf{x}^{k+1} - \mathbf{x}^k\| \sum_{j=1}^{p} \|\mathbf{x}_j^{k+1} - \mathbf{x}_j^*\|. \end{aligned}$$

Hence,

$$\begin{aligned} (F(\mathbf{x}_{k+1}) - F(\mathbf{x}^*))^2 &\leq (L_f + L_{\max})^2 \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \left(\sum_{j=1}^{p} \|\mathbf{x}_j^{k+1} - \mathbf{x}_j^*\| \right)^2 \\ &\leq p(L_f + L_{\max})^2 \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \sum_{j=1}^{p} \|\mathbf{x}_j^{k+1} - \mathbf{x}_j^*\|^2 \\ &= p(L_f + L_{\max})^2 \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \cdot \|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 \\ &\leq p(L_f + L_{\max})^2 R_{F(\mathbf{x}^0)}^2 \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2, \end{aligned} \quad (11.18)$$

where the last inequality follows by the monotonicity of the sequence of function values (Corollary 11.12) and Assumption 11.15(B). Combining (11.18) with (11.11), we obtain that

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq \frac{L_{\min}}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \geq \frac{L_{\min}}{2p(L_f + L_{\max})^2 R^2}(F(\mathbf{x}^{k+1}) - F(\mathbf{x}^*))^2,$$

where $R = R_{F(\mathbf{x}^0)}$. □

To derive the rate of convergence in function values, we will use the following lemma on the convergence of nonnegative scalar sequences satisfying a certain recursive inequality relation. The result resembles the one derived in Lemma 10.70, but the recursive inequality is different.

Lemma 11.17. *Let $\{a_k\}_{k\geq 0}$ be a nonnegative sequence of real numbers satisfying*

$$a_k - a_{k+1} \geq \frac{1}{\gamma}a_{k+1}^2, \quad k = 0, 1, \ldots, \tag{11.19}$$

for some $\gamma > 0$. Then for any $n \geq 2$,

$$a_n \leq \max\left\{\left(\frac{1}{2}\right)^{(n-1)/2} a_0, \frac{4\gamma}{n-1}\right\}. \tag{11.20}$$

In addition, for any $\varepsilon > 0$, if $n \geq 2$ satisfies

$$n \geq \max\left\{\frac{2}{\log(2)}(\log(a_0) + \log(1/\varepsilon)), \frac{4\gamma}{\varepsilon}\right\} + 1,$$

then $a_n \leq \varepsilon$.

Proof. Let $n \geq 2$. If $a_n = 0$, then (11.20) is trivial. We can thus assume that $a_n > 0$, from which it follows that $a_1, a_2, \ldots, a_{n-1} > 0$. For any $k \in \{0, 1, \ldots, n-1\}$,

$$\frac{1}{a_{k+1}} - \frac{1}{a_k} = \frac{a_k - a_{k+1}}{a_k a_{k+1}} \geq \frac{1}{\gamma}\frac{a_{k+1}}{a_k}. \tag{11.21}$$

For each k, there are two options:

(i) $\frac{a_{k+1}}{a_k} \leq \frac{1}{2}$.

(ii) $\frac{a_{k+1}}{a_k} > \frac{1}{2}$.

By (11.21), under option (ii) we have

$$\frac{1}{a_{k+1}} - \frac{1}{a_k} \geq \frac{1}{2\gamma}.$$

Suppose that n is a positive even integer. If there are at least $\frac{n}{2}$ indices (out of $k = 0, 1, \ldots, n-1$) for which option (ii) occurs, then

$$\frac{1}{a_n} \geq \frac{n}{4\gamma},$$

and hence

$$a_n \leq \frac{4\gamma}{n}.$$

On the other hand, if this is not the case, then there are at least $\frac{n}{2}$ indices for which option (i) occurs, and consequently

$$a_n \leq \left(\frac{1}{2}\right)^{n/2} a_0.$$

We therefore obtain that in any case, for an even n,

$$a_n \leq \max\left\{\left(\frac{1}{2}\right)^{n/2} a_0, \frac{4\gamma}{n}\right\}. \tag{11.22}$$

If $n \geq 3$ is a positive odd integer, then

$$a_n \leq a_{n-1} \leq \max\left\{\left(\frac{1}{2}\right)^{(n-1)/2} a_0, \frac{4\gamma}{n-1}\right\}. \tag{11.23}$$

Since the right-hand side of (11.23) is larger than the right-hand side of (11.22), the result (11.20) follows. Let $n \geq 2$. To guarantee that the inequality $a_n \leq \varepsilon$ holds, it is sufficient that the inequality

$$\max\left\{\left(\frac{1}{2}\right)^{(n-1)/2} a_0, \frac{4\gamma}{n-1}\right\} \leq \varepsilon$$

will hold, meaning that the following two inequalities will be satisfied:

$$\left(\frac{1}{2}\right)^{(n-1)/2} a_0 \leq \varepsilon, \quad \frac{4\gamma}{n-1} \leq \varepsilon.$$

These inequalities are obviously equivalent to

$$n \geq \frac{2}{\log(2)}(\log(a_0) + \log(1/\varepsilon)) + 1, \quad n \geq \frac{4\gamma}{\varepsilon} + 1.$$

Therefore, if

$$n \geq \max\left\{\frac{2}{\log(2)}(\log(a_0) + \log(1/\varepsilon)), \frac{4\gamma}{\varepsilon}\right\} + 1,$$

then the inequality $a_n \leq \varepsilon$ is guaranteed. ∎

Combining Lemmas 11.16 and 11.17, we can establish an $O(1/k)$ rate of convergence in function values of the sequence generated by the CBPG method, as well as a complexity result.

Theorem 11.18 ($O(1/k)$ rate of convergence of CBPG). *Suppose that Assumptions* 11.1 *and* 11.15 *hold. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the CBPG method for solving problem* (11.1). *For any $k \geq 2$,*

$$F(\mathbf{x}^k) - F_{\text{opt}} \leq \max\left\{\left(\frac{1}{2}\right)^{(k-1)/2} (F(\mathbf{x}^0) - F_{\text{opt}}), \frac{8p(L_f + L_{\max})^2 R^2}{L_{\min}(k-1)}\right\}, \tag{11.24}$$

where $L_{\min} = \min_{i=1,2,\ldots,p} L_i$, $L_{\max} = \max_{i=1,2,\ldots,p} L_i$, *and* $R = R_{F(\mathbf{x}^0)}$. *In addition, if* $n \geq 2$ *satisfies*

$$n \geq \max\left\{\frac{2}{\log(2)}(\log(F(\mathbf{x}^0) - F_{\text{opt}}) + \log(1/\varepsilon)), \frac{8p(L_f + L_{\max})^2 R^2}{L_{\min}\varepsilon}\right\} + 1,$$

then $F(\mathbf{x}^n) - F_{\text{opt}} \leq \varepsilon$.

Proof. Denote $a_k = F(\mathbf{x}^k) - F_{\text{opt}}$. Then by Lemma 11.16,

$$a_k - a_{k+1} \geq \frac{1}{D} a_{k+1}^2,$$

where $D = \frac{2p(L_f + L_{\max})^2 R^2}{L_{\min}}$. The result now follows by invoking Lemma 11.17 with $\gamma = D$. ∎

Remark 11.19 (index order). *The analysis of the CBPG method was done under the assumption that the index selection strategy is cyclic. However, it is easy to see that the same analysis, and consequently the main results (Theorems* 11.14 *and* 11.18*), hold for any index selection strategy in which each block is updated exactly once between consecutive iterations. One example of such an index selection strategy is the "cyclic shuffle" order in which the order of blocks is picked at the beginning of each iteration by a random permutation; in a sense, this is a "quasi-randomized" strategy. In the next section we will study a fully randomized approach.*

We end this section by showing that for convex differentiable functions (over the entire space) block Lipschitz continuity (Assumption 11.1(D)) implies that the function is L-smooth (Assumption 11.1(C)) with L being the sum of the block Lipschitz constants. This means that in this situation we can actually drop Assumption 11.1(C).

Theorem 11.20.[65] *Let* $\phi : \mathbb{E} \to \mathbb{R}$ $(\mathbb{E} = \mathbb{E}_1 \times \mathbb{E}_2 \times \cdots \times \mathbb{E}_p)$ *be a convex function satisfying the following assumptions:*

(A) ϕ *is differentiable over* $\mathbb{E}$*;*

(B) *there exist* $L_1, L_2, \ldots, L_p > 0$ *such that for any* $i \in \{1, 2, \ldots, p\}$ *it holds that*

$$\|\nabla_i \phi(\mathbf{x}) - \nabla_i \phi(\mathbf{x} + \mathcal{U}_i(\mathbf{d}))\| \leq L_i \|\mathbf{d}\|$$

for all $\mathbf{x} \in \mathbb{E}$ *and* $\mathbf{d} \in \mathbb{E}_i$.

Then ϕ *is* L*-smooth with* $L = L_1 + L_2 + \cdots + L_p$.

Proof. Let $\mathbf{y} \in \mathbb{E}$. Define the function

$$f(\mathbf{x}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \nabla \phi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle. \tag{11.25}$$

Then it is immediate to show that f also satisfies properties (A) and (B). In addition, the convexity of ϕ implies the convexity of f as well as the fact that f is nonnegative.

[65] Theorem 11.20 is a specialization of Lemma 2 from Nesterov [96].

Invoking Lemma 11.9 with $g_1 = g_2 = \cdots = g_p \equiv 0$, we obtain that for all $i \in \{1, 2, \ldots, p\}$ and $\mathbf{x} \in \mathbb{E}$,

$$f(\mathbf{x}) - f\left(\mathbf{x} - \frac{1}{L_i}\mathcal{U}_i(\nabla_i f(\mathbf{x}))\right) \geq \frac{1}{2L_i}\|\nabla_i f(\mathbf{x})\|^2,$$

which, along with the nonnegativity of f, implies that

$$f(\mathbf{x}) \geq \frac{1}{2L_i}\|\nabla_i f(\mathbf{x})\|^2.$$

Since the last inequality holds for any $i \in \{1, 2, \ldots, p\}$, it follows that

$$\begin{aligned} f(\mathbf{x}) &\geq \max_{i=1,2,\ldots,p}\left\{\frac{1}{2L_i}\|\nabla_i f(\mathbf{x})\|^2\right\} \geq \sum_{i=1}^{p}\frac{L_i}{\sum_{j=1}^{p} L_j}\frac{1}{2L_i}\|\nabla_i f(\mathbf{x})\|^2 \\ &= \frac{1}{2(\sum_{j=1}^{p} L_j)}\|\nabla f(\mathbf{x})\|^2. \end{aligned}$$

Plugging the expression (11.25) for f into the above inequality, we obtain

$$\phi(\mathbf{x}) \geq \phi(\mathbf{y}) + \langle \nabla\phi(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + \frac{1}{2(\sum_{j=1}^{p} L_j)}\|\nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{y})\|^2.$$

Since the above inequality holds for any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$, it follows by Theorem 5.8 (equivalence between (i) and (iii)) that ϕ is $(L_1 + L_2 + \cdots + L_p)$-smooth. □

11.5 The Randomized Block Proximal Gradient Method[66]

In this section we will analyze a version of the block proximal gradient method in which at each iteration a prox-grad step is performed at a randomly chosen block. The analysis is made under Assumption 11.21 given below. Note that at this point we do not assume that f is convex, but the main convergence result, Theorem 11.25, will require the convexity of f.

Assumption 11.21.

(A) $g_i : \mathbb{E}_i \to (-\infty, \infty]$ *is proper closed and convex for any* $i \in \{1, 2, \ldots, p\}$.

(B) $f : \mathbb{E} \to (-\infty, \infty]$ *is proper closed and convex,* $\mathrm{dom}(g) \subseteq \mathrm{int}(\mathrm{dom}(f))$*, and* f *is differentiable over* $\mathrm{int}(\mathrm{dom}(f))$.

(C) *There exist* $L_1, L_2, \ldots, L_p > 0$ *such that for any* $i \in \{1, 2, \ldots, p\}$ *it holds that*

$$\|\nabla_i f(\mathbf{x}) - \nabla_i f(\mathbf{x} + \mathcal{U}_i(\mathbf{d}))\| \leq L_i\|\mathbf{d}\|$$

for all $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$ *and* $\mathbf{d} \in \mathbb{E}_i$ *for which* $\mathbf{x} + \mathcal{U}_i(\mathbf{d}) \in \mathrm{int}(\mathrm{dom}(f))$.

[66] The derivation of the randomized complexity result in Section 11.5 mostly follows the presentation in the work of Lin, Lu, and Xiao [82].

(D) *The optimal set of problem* (11.1) *is nonempty and denoted by* X^*. *The optimal value is denoted by* $F_{\rm opt}$.

The Randomized Block Proximal Gradient (RBPG) Method

Initialization: pick $\mathbf{x}^0 = (\mathbf{x}_1^0, \mathbf{x}_2^0, \ldots, \mathbf{x}_p^0) \in \mathrm{int}(\mathrm{dom}(f))$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick $i_k \in \{1, 2, \ldots, p\}$ randomly via a uniform distribution;

(b) $\mathbf{x}^{k+1} = \mathbf{x}^k + \mathcal{U}_{i_k}(T^{i_k}_{L_{i_k}}(\mathbf{x}^k) - \mathbf{x}^k_{i_k})$.

Remark 11.22. *Step* (b) *of the algorithm can also be written as*

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L_{i_k}} \mathcal{U}_{i_k}(G^{i_k}_{L_{i_k}}(\mathbf{x}^k)).$$

From the point of view of computational complexity, loosely speaking, each p iterations of the RBPG method are comparable to one iteration of the CBPG method.

Using the block sufficient decrease lemma (Lemma 11.9), it is easy to show a sufficient decrease property of the RBPG method.

Theorem 11.23 (sufficient decrease of the RBPG method). *Suppose that Assumption* 11.21 *holds, and let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by the RBPG method. Then for any* $k \geq 0$,

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq \frac{1}{2L_{i_k}} \|G^{i_k}_{L_{i_k}}(\mathbf{x}^k)\|^2.$$

Proof. Invoke Lemma 11.9 with $\mathbf{x} = \mathbf{x}^k$ and $i = i_k$. ∎

Remark 11.24. *A direct consequence of Theorem* 11.23 *is that the sequence of function values* $\{F(\mathbf{x}^k)\}_{k\geq 0}$ *generated by the RBPG method is nonincreasing. As a result, it is also correct that the sequence of expected function values*

$$\{\mathsf{E}_{i_0,\ldots,i_{k-1}}(F(\mathbf{x}^k))\}_{k\geq 0}$$

is nonincreasing.

In our analysis the following notation is used:

- $\xi_{k-1} \equiv \{i_0, i_1, \ldots, i_{k-1}\}$ is a multivariate random variable.

- In addition to the underlying Euclidean norm of the space $\mathbb{E}$, we define the following weighted norm:

$$\|\mathbf{x}\|_L \equiv \sqrt{\sum_{i=1}^{p} L_i \|\mathbf{x}_i\|^2}$$

and its dual norm

$$\|\mathbf{x}\|_{L,*} = \sqrt{\sum_{i=1}^{p} \frac{1}{L_i} \|\mathbf{x}_i\|^2}.$$

- We will consider the following variation of the gradient mapping:

$$\widetilde{G}(\mathbf{x}) = (G^1_{L_1}(\mathbf{x}^k), G^2_{L_2}(\mathbf{x}^k), \ldots, G^p_{L_p}(\mathbf{x}^k)). \tag{11.26}$$

Obviously, if $L_1 = L_2 = \cdots = L_p = L$, then $\widetilde{G}(\mathbf{x}) = G_L(\mathbf{x})$.

The main convergence result will now be stated and proved.

Theorem 11.25 ($O(1/k)$ rate of convergence of the RBPG method). *Suppose that Assumption* 11.21 *holds and that f is convex. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the RBPG method for solving problem* (11.1). *Let $\mathbf{x}^* \in X^*$. Then for any $k \geq 0$,*

$$\mathsf{E}_{\xi_k}(F(\mathbf{x}^{k+1})) - F_{\rm opt} \leq \frac{p}{p+k+1}\left(\frac{1}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|_L^2 + F(\mathbf{x}^0) - F_{\rm opt}\right). \tag{11.27}$$

Proof. Let $\mathbf{x}^* \in X^*$. We denote for any $n \geq 0$, $r_n \equiv \|\mathbf{x}^n - \mathbf{x}^*\|_L$. Then for any $k \geq 0$,

$$\begin{aligned}
r_{k+1}^2 &= \|\mathbf{x}^{k+1} - \mathbf{x}^*\|_L^2 \\
&= \left\|\mathbf{x}^k - \frac{1}{L_{i_k}}\mathcal{U}_{i_k}\left(G^{i_k}_{L_{i_k}}(\mathbf{x}^k)\right) - \mathbf{x}^*\right\|_L^2 \\
&= \|\mathbf{x}^k - \mathbf{x}^*\|_L^2 - \frac{2}{L_{i_k}} L_{i_k}\langle G^{i_k}_{L_{i_k}}(\mathbf{x}^k), \mathbf{x}^k_{i_k} - \mathbf{x}^*_{i_k}\rangle + \frac{L_{i_k}}{L_{i_k}^2}\|G^{i_k}_{L_{i_k}}(\mathbf{x}^k)\|^2 \\
&= \|\mathbf{x}^k - \mathbf{x}^*\|_L^2 - 2\langle G^{i_k}_{L_{i_k}}(\mathbf{x}^k), \mathbf{x}^k_{i_k} - \mathbf{x}^*_{i_k}\rangle + \frac{1}{L_{i_k}}\|G^{i_k}_{L_{i_k}}(\mathbf{x}^k)\|^2 \\
&= r_k^2 - 2\langle G^{i_k}_{L_{i_k}}(\mathbf{x}^k), \mathbf{x}^k_{i_k} - \mathbf{x}^*_{i_k}\rangle + \frac{1}{L_{i_k}}\|G^{i_k}_{L_{i_k}}(\mathbf{x}^k)\|^2.
\end{aligned}$$

Taking expectation w.r.t. i_k, we obtain (using the notation (11.26))

$$\begin{aligned}
\mathsf{E}_{i_k}\left(\frac{1}{2}r_{k+1}^2\right) &= \frac{1}{2}r_k^2 - \frac{1}{p}\sum_{i=1}^{p}\langle G^i_{L_i}(\mathbf{x}_k), \mathbf{x}^k_i - \mathbf{x}^*_i\rangle + \frac{1}{2p}\sum_{i=1}^{p}\frac{1}{L_i}\|G^i_{L_i}(\mathbf{x}^k)\|^2 \\
&= \frac{1}{2}r_k^2 - \frac{1}{p}\langle \widetilde{G}(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^*\rangle + \frac{1}{2p}\|\widetilde{G}(\mathbf{x}^k)\|_{L,*}^2.
\end{aligned} \tag{11.28}$$

By the block descent lemma (Lemma 11.8),

$$\begin{aligned} f(\mathbf{x}^{k+1}) &= f\left(\mathbf{x}^k - \frac{1}{L_{i_k}}\mathcal{U}_{i_k}(G^{i_k}_{L_{i_k}}(\mathbf{x}^k))\right) \\ &\leq f(\mathbf{x}^k) - \frac{1}{L_{i_k}}\langle \nabla_{i_k} f(\mathbf{x}^k), G^{i_k}_{L_{i_k}}(\mathbf{x}^k)\rangle + \frac{1}{2L_{i_k}}\|G^{i_k}_{L_{i_k}}(\mathbf{x}^k)\|^2. \end{aligned}$$

Hence,

$$F(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \frac{1}{L_{i_k}}\langle \nabla_{i_k} f(\mathbf{x}^k), G^{i_k}_{L_{i_k}}(\mathbf{x}^k)\rangle + \frac{1}{2L_{i_k}}\|G^{i_k}_{L_{i_k}}(\mathbf{x}^k)\|^2 + g(\mathbf{x}^{k+1}).$$

Taking expectation of both sides of the last inequality w.r.t. i_k, we obtain

$$\mathsf{E}_{i_k}(F(\mathbf{x}^{k+1})) \leq f(\mathbf{x}^k) - \frac{1}{p}\sum_{i=1}^{p}\frac{1}{L_i}\langle \nabla_i f(\mathbf{x}^k), G^i_{L_i}(\mathbf{x}^k)\rangle + \frac{1}{2p}\|\widetilde{G}(\mathbf{x}^k)\|^2_{L,*} + \mathsf{E}_{i_k}(g(\mathbf{x}^{k+1})). \tag{11.29}$$

Since $\mathbf{x}^{k+1}_{i_k} = \mathbf{x}^k_{i_k} - \frac{1}{L_{i_k}}G^{i_k}_{L_{i_k}}(\mathbf{x}^k) = \operatorname{prox}_{\frac{1}{L_{i_k}}g_{i_k}}\left(\mathbf{x}^k_{i_k} - \frac{1}{L_{i_k}}\nabla_{i_k} f(\mathbf{x}^k)\right)$, it follows by the second prox theorem (Theorem 6.39) that

$$\begin{aligned} g_{i_k}(\mathbf{x}^*_{i_k}) &\geq g_{i_k}\left(\mathbf{x}^k_{i_k} - \frac{1}{L_{i_k}}G^{i_k}_{L_{i_k}}(\mathbf{x}^k)\right) \\ &\quad + L_{i_k}\left\langle \mathbf{x}^k_{i_k} - \frac{1}{L_{i_k}}\nabla_{i_k} f(\mathbf{x}^k) - \mathbf{x}^k_{i_k} + \frac{1}{L_{i_k}}G^{i_k}_{L_{i_k}}(\mathbf{x}^k), \mathbf{x}^*_{i_k} - \mathbf{x}^k_{i_k} + \frac{1}{L_{i_k}}G^{i_k}_{L_{i_k}}(\mathbf{x}^k)\right\rangle. \end{aligned}$$

That is,

$$\begin{aligned} g_{i_k}(\mathbf{x}^*_{i_k}) &\geq g_{i_k}\left(\mathbf{x}^k_{i_k} - \frac{1}{L_{i_k}}G^{i_k}_{L_{i_k}}(\mathbf{x}^k)\right) \\ &\quad + \left\langle -\nabla_{i_k} f(\mathbf{x}^k) + G^{i_k}_{L_{i_k}}(\mathbf{x}^k), \mathbf{x}^*_{i_k} - \mathbf{x}^k_{i_k} + \frac{1}{L_{i_k}}G^{i_k}_{L_{i_k}}(\mathbf{x}^k)\right\rangle. \end{aligned} \tag{11.30}$$

Note that

$$\mathsf{E}_{i_k}(g_{i_k}(\mathbf{x}^*_{i_k})) = \frac{1}{p}g(\mathbf{x}^*), \tag{11.31}$$

$$\mathsf{E}_{i_k}(g(\mathbf{x}^{k+1})) = \frac{p-1}{p}g(\mathbf{x}^k) + \frac{1}{p}\sum_{i=1}^{p} g_i\left(\mathbf{x}^k_i - \frac{1}{L_i}G^i_{L_i}(\mathbf{x}^k)\right). \tag{11.32}$$

Taking expectation w.r.t. i_k in (11.30) and plugging in the relations (11.31) and (11.32) leads to the following inequality:

$$\begin{aligned} \frac{1}{p}g(\mathbf{x}^*) &\geq \mathsf{E}_{i_k}(g(\mathbf{x}^{k+1})) - \frac{p-1}{p}g(\mathbf{x}^k) + \frac{1}{p}\langle -\nabla f(\mathbf{x}^k) + \widetilde{G}(\mathbf{x}^k), \mathbf{x}^* - \mathbf{x}^k\rangle \\ &\quad - \frac{1}{p}\sum_{i=1}^{p}\frac{1}{L_i}\langle \nabla_i f(\mathbf{x}^k), G^i_{L_i}(\mathbf{x}^k)\rangle + \frac{1}{p}\|\widetilde{G}(\mathbf{x}^k)\|^2_{L,*}. \end{aligned}$$

The last inequality can be equivalently written as

$$\begin{aligned} &\mathsf{E}_{i_k}(g(\mathbf{x}^{k+1})) - \frac{1}{p}\sum_{i=1}^{p}\frac{1}{L_i}\langle \nabla_i f(\mathbf{x}^k), G^i_{L_i}(\mathbf{x}^k)\rangle \\ &\leq \frac{1}{p}g(\mathbf{x}^*) + \frac{p-1}{p}g(\mathbf{x}^k) + \frac{1}{p}\langle \nabla f(\mathbf{x}^k) - \widetilde{G}(\mathbf{x}^k), \mathbf{x}^* - \mathbf{x}^k\rangle - \frac{1}{p}\|\widetilde{G}(\mathbf{x}^k)\|^2_{L,*}. \end{aligned}$$

Plugging the last inequality into (11.29) we obtain that

$$\mathsf{E}_{i_k}(F(\mathbf{x}^{k+1})) \leq f(\mathbf{x}^k) - \frac{1}{2p}\|\tilde{G}(\mathbf{x}^k)\|^2_{L,*} + \frac{1}{p}g(\mathbf{x}^*) + \frac{1}{p}\langle \nabla f(\mathbf{x}^k) - \widetilde{G}(\mathbf{x}^k), \mathbf{x}^* - \mathbf{x}^k\rangle + \frac{p-1}{p}g(\mathbf{x}^k),$$

which, along with the gradient inequality $\langle \nabla f(\mathbf{x}^k), \mathbf{x}^* - \mathbf{x}^k\rangle \leq f(\mathbf{x}^*) - f(\mathbf{x}^k)$, implies

$$\mathsf{E}_{i_k}(F(\mathbf{x}^{k+1})) \leq \frac{p-1}{p}F(\mathbf{x}^k) + \frac{1}{p}F(\mathbf{x}^*) - \frac{1}{2p}\|\widetilde{G}(\mathbf{x}^k)\|^2_{L,*} - \frac{1}{p}\langle \widetilde{G}(\mathbf{x}^k), \mathbf{x}^* - \mathbf{x}^k\rangle.$$

The last inequality, combined with (11.28), yields the relation

$$\mathsf{E}_{i_k}\left(\frac{1}{2}r^2_{k+1}\right) \leq \frac{1}{2}r^2_k + \frac{p-1}{p}F(\mathbf{x}^k) + \frac{1}{p}F(\mathbf{x}^*) - \mathsf{E}_{i_k}(F(\mathbf{x}^{k+1})),$$

which can be rearranged as

$$\mathsf{E}_{i_k}\left(\frac{1}{2}r^2_{k+1} + F(\mathbf{x}^{k+1}) - F_{\text{opt}}\right) \leq \left(\frac{1}{2}r^2_k + F(\mathbf{x}^k) - F_{\text{opt}}\right) - \frac{1}{p}(F(\mathbf{x}^k) - F_{\text{opt}}).$$

Taking expectation over ξ_{k-1} of both sides we obtain (where we make the convention that the expression $\mathbb{E}_{\xi_{-1}}(F(\mathbf{x}^0))$ means $F(\mathbf{x}^0)$)

$$\mathsf{E}_{\xi_k}\left(\frac{1}{2}r^2_{k+1} + F(\mathbf{x}^{k+1}) - F_{\text{opt}}\right) \leq \mathsf{E}_{\xi_{k-1}}\left(\frac{1}{2}r^2_k + F(\mathbf{x}^k) - F_{\text{opt}}\right) - \frac{1}{p}(\mathsf{E}_{\xi_{k-1}}(F(\mathbf{x}^k)) - F_{\text{opt}}).$$

We can thus conclude that

$$\begin{aligned}\mathsf{E}_{\xi_k}(F(\mathbf{x}^{k+1})) - F_{\text{opt}} &\leq \mathsf{E}_{\xi_k}\left(\frac{1}{2}r^2_{k+1} + F(\mathbf{x}^{k+1}) - F_{\text{opt}}\right)\\ &\leq \frac{1}{2}r^2_0 + F(\mathbf{x}^0) - F_{\text{opt}} - \frac{1}{p}\sum_{j=0}^{k}\left(\mathsf{E}_{\xi_{j-1}}(F(\mathbf{x}^j)) - F_{\text{opt}}\right),\end{aligned}$$

which, together with the monotonicity of the sequence of expected values $\{\mathsf{E}_{\xi_{k-1}}(F(\mathbf{x}^k))\}_{k\geq 0}$ (see Remark 11.24), implies that

$$\mathsf{E}_{\xi_k}(F(\mathbf{x}^{k+1})) - F_{\text{opt}} \leq \frac{1}{2}r^2_0 + F(\mathbf{x}^0) - F_{\text{opt}} - \frac{k+1}{p}\left(\mathsf{E}_{\xi_k}(F(\mathbf{x}^{k+1})) - F_{\text{opt}}\right). \quad (11.33)$$

The desired result (11.27) follows immediately from (11.33). ∎

Chapter 12

Dual-Based Proximal Gradient Methods

Underlying Spaces: In this chapter, all the underlying spaces are Euclidean.

12.1 The Primal and Dual Models

The main model discussed in this chapter is

$$f_{\text{opt}} = \min_{\mathbf{x}\in\mathbb{E}} \{f(\mathbf{x}) + g(\mathcal{A}(\mathbf{x}))\}, \tag{12.1}$$

where the following assumptions are made.

Assumption 12.1.

(A) $f : \mathbb{E} \to (-\infty, +\infty]$ *is proper closed and* σ*-strongly convex* ($\sigma > 0$).

(B) $g : \mathbb{V} \to (-\infty, +\infty]$ *is proper closed and convex.*

(C) $\mathcal{A} : \mathbb{E} \to \mathbb{V}$ *is a linear transformation.*

(D) *There exists* $\hat{\mathbf{x}} \in \text{ri}(\text{dom}(f))$ *and* $\hat{\mathbf{z}} \in \text{ri}(\text{dom}(g))$ *such that* $\mathcal{A}(\hat{\mathbf{x}}) = \hat{\mathbf{z}}$.

Under Assumption 12.1 the function $\mathbf{x} \mapsto f(\mathbf{x}) + g(\mathcal{A}(\mathbf{x}))$ is proper closed and σ-strongly convex, and hence, by Theorem 5.25(a), problem (12.1) has a unique optimal solution, which we denote throughout this chapter by $\mathbf{x}^*$.

To construct a dual problem to (12.1), we first rewrite it in the form

$$\begin{array}{ll} \min_{\mathbf{x},\mathbf{z}} & f(\mathbf{x}) + g(\mathbf{z}) \\ \text{s.t.} & \mathcal{A}(\mathbf{x}) - \mathbf{z} = \mathbf{0}. \end{array} \tag{12.2}$$

Associating a Lagrange dual vector $\mathbf{y} \in \mathbb{V}$ to the equality constraints in (12.2), the Lagrangian can be written as

$$L(\mathbf{x}, \mathbf{z}; \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{z}) - \langle \mathbf{y}, \mathcal{A}(\mathbf{x}) - \mathbf{z} \rangle = f(\mathbf{x}) + g(\mathbf{z}) - \langle \mathcal{A}^T(\mathbf{y}), \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle. \tag{12.3}$$

Minimizing the Lagrangian w.r.t. $\mathbf{x}$ and $\mathbf{z}$, the obtained dual problem is

$$q_{\text{opt}} = \max_{\mathbf{y}\in\mathbb{V}} \left\{ q(\mathbf{y}) \equiv -f^*(\mathcal{A}^T(\mathbf{y})) - g^*(-\mathbf{y}) \right\}. \tag{12.4}$$

By the strong duality theorem for convex problems (see Theorem A.1), it follows that strong duality holds for the pair of problems (12.1) and (12.4).

Theorem 12.2 (strong duality for the pair of problems (12.1) and (12.4)). *Suppose that Assumption* 12.1 *holds, and let* $f_{\text{opt}}, q_{\text{opt}}$ *be the optimal values of the primal and dual problems* (12.1) *and* (12.4)*, respectively. Then* $f_{\text{opt}} = q_{\text{opt}}$*, and the dual problem* (12.4) *possesses an optimal solution.*

We will consider the dual problem in its minimization form:

$$\min_{\mathbf{y}\in\mathbb{V}} \{F(\mathbf{y}) + G(\mathbf{y})\}, \tag{12.5}$$

where

$$F(\mathbf{y}) \equiv f^*(\mathcal{A}^T(\mathbf{y})), \tag{12.6}$$

$$G(\mathbf{y}) \equiv g^*(-\mathbf{y}). \tag{12.7}$$

The basic properties of F and G are gathered in the following lemma.

Lemma 12.3 (properties of F and G). *Suppose that Assumption* 12.1 *holds, and let* F *and* G *be defined by* (12.6) *and* (12.7)*, respectively. Then*

(a) $F : \mathbb{V} \to \mathbb{R}$ *is convex and* L_F*-smooth with* $L_F = \frac{\|\mathcal{A}\|^2}{\sigma}$*;*

(b) $G : \mathbb{V} \to (-\infty, \infty]$ *is proper closed and convex.*

Proof. (a) Since f is proper closed and σ-strongly convex, then by the conjugate correspondence theorem (Theorem 5.26(b)), f^* is $\frac{1}{\sigma}$-smooth. Therefore, for any $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{V}$,

$$\begin{aligned}
\|\nabla F(\mathbf{y}_1) - \nabla F(\mathbf{y}_2)\| &= \|\mathcal{A}(\nabla f^*(\mathcal{A}^T(\mathbf{y}_1))) - \mathcal{A}(\nabla f^*(\mathcal{A}^T(\mathbf{y}_2)))\| \\
&\le \|\mathcal{A}\| \cdot \|\nabla f^*(\mathcal{A}^T(\mathbf{y}_1)) - \nabla f^*(\mathcal{A}^T(\mathbf{y}_2))\| \\
&\le \frac{1}{\sigma}\|\mathcal{A}\| \cdot \|\mathcal{A}^T(\mathbf{y}_1) - \mathcal{A}^T(\mathbf{y}_2)\| \\
&\le \frac{\|\mathcal{A}\| \cdot \|\mathcal{A}^T\|}{\sigma}\|\mathbf{y}_1 - \mathbf{y}_2\| = \frac{\|\mathcal{A}\|^2}{\sigma}\|\mathbf{y}_1 - \mathbf{y}_2\|,
\end{aligned}$$

where we used in the last equality the fact that $\|\mathcal{A}\| = \|\mathcal{A}^T\|$ (see Section 1.14). To show the convexity of F, note that f^* is convex as a conjugate function (Theorem 4.3), and hence, by Theorem 2.16, F, as a composition of a convex function and a linear mapping, is convex.

(b) Since g is proper closed and convex, so is g^* (Theorems 4.3 and 4.5). Thus, $G(\mathbf{y}) \equiv g^*(-\mathbf{y})$ is also proper closed and convex. ∎

12.2 The Dual Proximal Gradient Method[67]

Problem (12.5) consists of minimizing the sum of a convex L-smooth function and a proper closed and convex function. It is therefore possible to employ in this setting the proximal gradient method on problem (12.5), which is equivalent to the dual problem of (12.1). Naturally we will refer to this algorithm as the "dual proximal gradient" (DPG) method. The dual representation of the method is given below.

Dual Proximal Gradient—dual representation

- **Initialization:** pick $\mathbf{y}^0 \in \mathbb{V}$ and $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}$.
- **General step ($k \geq 0$):**

$$\mathbf{y}^{k+1} = \text{prox}_{\frac{1}{L}G}\left(\mathbf{y}^k - \frac{1}{L}\nabla F(\mathbf{y}^k)\right). \tag{12.8}$$

Since F is convex and L_F-smooth and G is proper closed and convex, we can invoke Theorem 10.21 to obtain an $O(1/k)$ rate of convergence in terms of the dual objective function values.

Theorem 12.4. *Suppose that Assumption* 12.1 *holds, and let* $\{\mathbf{y}^k\}_{k\geq 0}$ *be the sequence generated by the DPG method with* $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}$. *Then for any optimal solution* $\mathbf{y}^*$ *of the dual problem* (12.4) *and* $k \geq 1$,

$$q_{\text{opt}} - q(\mathbf{y}^k) \leq \frac{L\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{2k}.$$

Our goal now will be to find a primal representation of the method, which will be written in a more explicit way in terms of the data of the problem, meaning $(f, g, \mathcal{A})$. To achieve this goal, we will require the following technical lemma.

Lemma 12.5. *Let* $F(\mathbf{y}) = f^*(\mathcal{A}^T(\mathbf{y}) + \mathbf{b}), G(\mathbf{y}) = g^*(-\mathbf{y})$, *where* f, g, *and* $\mathcal{A}$ *satisfy properties* (A), (B), *and* (C) *of Assumption* 12.1 *and* $\mathbf{b} \in \mathbb{E}$. *Then for any* $\mathbf{y}, \mathbf{v} \in \mathbb{V}$ *and* $L > 0$ *the relation*

$$\mathbf{y} = \text{prox}_{\frac{1}{L}G}\left(\mathbf{v} - \frac{1}{L}\nabla F(\mathbf{v})\right) \tag{12.9}$$

holds if and only if

$$\mathbf{y} = \mathbf{v} - \frac{1}{L}\mathcal{A}(\tilde{\mathbf{x}}) + \frac{1}{L}\text{prox}_{Lg}(\mathcal{A}(\tilde{\mathbf{x}}) - L\mathbf{v}),$$

where

$$\tilde{\mathbf{x}} = \text{argmax}_{\mathbf{x}}\left\{\langle \mathbf{x}, \mathcal{A}^T(\mathbf{v}) + \mathbf{b}\rangle - f(\mathbf{x})\right\}.$$

Proof. By the conjugate subgradient theorem (Corollary 4.21), since f is proper closed and convex,

$$\nabla f^*(\mathcal{A}^T(\mathbf{v}) + \mathbf{b}) = \tilde{\mathbf{x}} \in \mathbb{E} \equiv \text{argmax}_{\mathbf{x}}\left\{\langle \mathbf{x}, \mathcal{A}^T(\mathbf{v}) + \mathbf{b}\rangle - f(\mathbf{x})\right\}.$$

[67] Sections 12.2 and 12.3 follow the work of Beck and Teboulle [21].

Therefore, since $\nabla F(\mathbf{v}) = \mathcal{A}(\nabla f^*(\mathcal{A}^T(\mathbf{v}) + \mathbf{b})) = \mathcal{A}(\tilde{\mathbf{x}})$,

$$\mathbf{y} = \text{prox}_{\frac{1}{L}G}\left(\mathbf{v} - \frac{1}{L}\mathcal{A}(\tilde{\mathbf{x}})\right). \tag{12.10}$$

Invoking Theorem 6.15 with $g \leftarrow \frac{1}{L}g^*$, $\mathcal{A} = -\mathcal{I}$, $\mathbf{b} = \mathbf{0}$, we obtain that for any $\mathbf{z} \in \mathbb{V}$,

$$\text{prox}_{\frac{1}{L}G}(\mathbf{z}) = -\text{prox}_{\frac{1}{L}g^*}(-\mathbf{z}). \tag{12.11}$$

Combining (12.10) and (12.11) and using the extended Moreau decomposition formula (Theorem 6.45), we finally obtain that

$$\begin{aligned}\mathbf{y} &= \text{prox}_{\frac{1}{L}G}\left(\mathbf{v} - \frac{1}{L}\mathcal{A}(\tilde{\mathbf{x}})\right) = -\text{prox}_{\frac{1}{L}g^*}\left(\frac{1}{L}\mathcal{A}(\tilde{\mathbf{x}}) - \mathbf{v}\right)\\ &= -\left[\frac{1}{L}\mathcal{A}(\tilde{\mathbf{x}}) - \mathbf{v} - \frac{1}{L}\text{prox}_{Lg}(\mathcal{A}(\tilde{\mathbf{x}}) - L\mathbf{v})\right]\\ &= \mathbf{v} - \frac{1}{L}\mathcal{A}(\tilde{\mathbf{x}}) + \frac{1}{L}\text{prox}_{Lg}(\mathcal{A}(\tilde{\mathbf{x}}) - L\mathbf{v}). \quad \square\end{aligned}$$

Equipped with Lemma 12.5, we can write a primal representation of the DPG method.

The Dual Proximal Gradient (DPG) Method—primal representation

Initialization: pick $\mathbf{y}^0 \in \mathbb{V}$, and $L \geq \frac{\|\mathcal{A}\|^2}{\sigma}$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) set $\mathbf{x}^k = \text{argmax}_{\mathbf{x}}\left\{\langle \mathbf{x}, \mathcal{A}^T(\mathbf{y}^k)\rangle - f(\mathbf{x})\right\}$;

(b) set $\mathbf{y}^{k+1} = \mathbf{y}^k - \frac{1}{L}\mathcal{A}(\mathbf{x}^k) + \frac{1}{L}\text{prox}_{Lg}(\mathcal{A}(\mathbf{x}^k) - L\mathbf{y}^k)$.

Remark 12.6 (the primal sequence). *The sequence $\{\mathbf{x}^k\}_{k\geq 0}$ generated by the method will be called "the primal sequence." The elements of the sequence are actually not necessarily feasible w.r.t. the primal problem* (12.1) *since they are not guaranteed to belong to* $\text{dom}(g)$*; nevertheless, we will show that the primal sequence does converge to the optimal solution* $\mathbf{x}^*$.

To prove a convergence result in terms of the primal sequence, we will require the following fundamental primal-dual relation.

Lemma 12.7 (primal-dual relation). *Suppose that Assumption* 12.1 *holds. Let* $\bar{\mathbf{y}} \in \text{dom}(G)$*, where* G *is given in* (12.7)*, and let*

$$\bar{\mathbf{x}} = \text{argmax}_{\mathbf{x}\in\mathbb{E}}\left\{\langle \mathbf{x}, \mathcal{A}^T(\bar{\mathbf{y}})\rangle - f(\mathbf{x})\right\}. \tag{12.12}$$

Then

$$\|\bar{\mathbf{x}} - \mathbf{x}^*\|^2 \leq \frac{2}{\sigma}(q_{\text{opt}} - q(\bar{\mathbf{y}})). \tag{12.13}$$

Proof. Recall that the primal problem (12.1) can be equivalently written as the problem

$$\min_{\mathbf{x}\in\mathbb{E},\mathbf{z}\in\mathbb{V}}\{f(\mathbf{x})+g(\mathbf{z}):\mathcal{A}(\mathbf{x})-\mathbf{z}=\mathbf{0}\},$$

whose Lagrangian is (see also (12.3))

$$L(\mathbf{x},\mathbf{z};\mathbf{y})=f(\mathbf{x})-\langle\mathcal{A}^T(\mathbf{y}),\mathbf{x}\rangle+g(\mathbf{z})+\langle\mathbf{y},\mathbf{z}\rangle.$$

In particular,

$$L(\mathbf{x},\mathbf{z};\bar{\mathbf{y}})=h(\mathbf{x})+s(\mathbf{z}), \tag{12.14}$$

where

$$\begin{aligned}h(\mathbf{x})&=f(\mathbf{x})-\langle\mathcal{A}^T(\bar{\mathbf{y}}),\mathbf{x}\rangle,\\ s(\mathbf{z})&=g(\mathbf{z})+\langle\bar{\mathbf{y}},\mathbf{z}\rangle.\end{aligned}$$

Since h is σ-strongly convex and $\bar{\mathbf{x}}$ is its minimizer (see relation (12.12)), it follows by Theorem 5.25(b) that

$$h(\mathbf{x})-h(\bar{\mathbf{x}})\geq\frac{\sigma}{2}\|\mathbf{x}-\bar{\mathbf{x}}\|^2. \tag{12.15}$$

Since the relation $\bar{\mathbf{y}}\in\mathrm{dom}(G)$ is equivalent to $-\bar{\mathbf{y}}\in\mathrm{dom}(g^*)$, it follows that

$$\min_{\mathbf{z}\in\mathbb{V}}\{g(\mathbf{z})+\langle\bar{\mathbf{y}},\mathbf{z}\rangle\}=\min_{\mathbf{z}\in\mathbb{V}}s(\mathbf{z})>-\infty.$$

Let $\varepsilon>0$. Then there exists $\bar{\mathbf{z}}_\varepsilon$ for which

$$s(\bar{\mathbf{z}}_\varepsilon)\leq\min_{\mathbf{z}\in\mathbb{V}}s(\mathbf{z})+\varepsilon. \tag{12.16}$$

Combining (12.14), (12.15), and (12.16), we obtain that for all $\mathbf{x}\in\mathrm{dom}(f)$ and $\mathbf{z}\in\mathrm{dom}(g)$,

$$L(\mathbf{x},\mathbf{z};\bar{\mathbf{y}})-L(\bar{\mathbf{x}},\bar{\mathbf{z}}_\varepsilon;\bar{\mathbf{y}})=h(\mathbf{x})-h(\bar{\mathbf{x}})+s(\mathbf{z})-s(\bar{\mathbf{z}}_\varepsilon)\geq\frac{\sigma}{2}\|\mathbf{x}-\bar{\mathbf{x}}\|^2-\varepsilon.$$

In particular, substituting $\mathbf{x}=\mathbf{x}^*,\mathbf{z}=\mathbf{z}^*\equiv\mathcal{A}(\mathbf{x}^*)$, then $L(\mathbf{x}^*,\mathbf{z}^*;\bar{\mathbf{y}})=f(\mathbf{x}^*)+g(\mathcal{A}(\mathbf{x}^*))=f_{\mathrm{opt}}=q_{\mathrm{opt}}$ (by Theorem 12.2), and we obtain

$$q_{\mathrm{opt}}-L(\bar{\mathbf{x}},\bar{\mathbf{z}}_\varepsilon;\bar{\mathbf{y}})\geq\frac{\sigma}{2}\|\mathbf{x}^*-\bar{\mathbf{x}}\|^2-\varepsilon. \tag{12.17}$$

In addition, by the definition of the dual objective function value,

$$L(\bar{\mathbf{x}},\bar{\mathbf{z}}_\varepsilon;\bar{\mathbf{y}})\geq\min_{\mathbf{x}\in\mathbb{E},\mathbf{z}\in\mathbb{V}}L(\mathbf{x},\mathbf{z};\bar{\mathbf{y}})=q(\bar{\mathbf{y}}),$$

which, combined with (12.17), results in the inequality

$$\|\bar{\mathbf{x}}-\mathbf{x}^*\|^2\leq\frac{2}{\sigma}(q_{\mathrm{opt}}-q(\bar{\mathbf{y}}))+\frac{2}{\sigma}\varepsilon.$$

Since the above inequality holds for any $\varepsilon>0$, the desired result (inequality (12.13)) follows. ∎

Combining the primal-dual relation of Lemma 12.7 with the rate of convergence of the sequence of dual objective function values stated in Theorem 12.4, we can deduce a rate of convergence result for the primal sequence to the unique optimal solution.

Theorem 12.8 ($O(1/k)$ rate of convergence of the primal sequence of the DPG method). *Suppose that Assumption* 12.1 *holds, and let* $\{\mathbf{x}^k\}_{k\geq 0}$ *and* $\{\mathbf{y}^k\}_{k\geq 0}$ *be the primal and dual sequences generated by the DPG method with* $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}$. *Then for any optimal solution* $\mathbf{y}^*$ *of the dual problem* (12.4) *and* $k \geq 1$,

$$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{L\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{\sigma k}. \tag{12.18}$$

Proof. Invoking Lemma 12.7 with $\bar{\mathbf{y}} = \mathbf{y}^k$, we obtain by the definition of $\bar{\mathbf{x}}$ (equation (12.12)) that $\bar{\mathbf{x}} = \mathbf{x}^k$, and hence (12.13) reads as

$$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{2}{\sigma}(q_{\text{opt}} - q(\mathbf{y}^k)),$$

which, combined with Theorem 12.4, yields the desired result. ∎

12.3 Fast Dual Proximal Gradient

The DPG method employs the proximal gradient method on the dual problem. Alternatively, we can also employ FISTA (see Section 10.7) on the dual problem (12.4). The dual representation of the method is given below.

The Fast Dual Proximal Gradient (FDPG) Method—dual representation

- **Initialization:** $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}, \mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{E}, t_0 = 1.$
- **General step ($k \geq 0$):**
  - (a) $\mathbf{y}^{k+1} = \text{prox}_{\frac{1}{L}G}\left(\mathbf{w}^k - \frac{1}{L}\nabla F(\mathbf{w}^k)\right)$;
  - (b) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
  - (c) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.

Since this is exactly FISTA employed on the dual problem, we can invoke Theorem 10.34 and obtain a convergence result in terms of dual objective function values.

Theorem 12.9. *Suppose that Assumption* 12.1 *holds and that* $\{\mathbf{y}^k\}_{k\geq 0}$ *is the sequence generated by the FDPG method with* $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}$. *Then for any optimal solution* $\mathbf{y}^*$ *of problem* (12.4) *and* $k \geq 1$,

$$q_{\text{opt}} - q(\mathbf{y}^k) \leq \frac{2L\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{(k+1)^2}.$$

Using Lemma 12.5 with $\mathbf{v} = \mathbf{w}^k$, $\mathbf{y} = \mathbf{y}^{k+1}$, and $\mathbf{b} = \mathbf{0}$, we obtain that step (a) of the FDPG method, namely,

$$\mathbf{y}^{k+1} = \mathrm{prox}_{\frac{1}{L}G}\left(\mathbf{w}^k - \frac{1}{L}\nabla F(\mathbf{w}^k)\right),$$

can be equivalently written as

$$\begin{aligned}\mathbf{u}^k &= \mathrm{argmax}_{\mathbf{u}}\left\{\langle \mathbf{u}, \mathcal{A}^T(\mathbf{w}^k)\rangle - f(\mathbf{u})\right\},\\ \mathbf{y}^{k+1} &= \mathbf{w}^k - \frac{1}{L}\mathcal{A}(\mathbf{u}^k) + \frac{1}{L}\mathrm{prox}_{Lg}(\mathcal{A}(\mathbf{u}^k) - L\mathbf{w}^k).\end{aligned}$$

We can thus formulate a primal representation of the method.

The Fast Dual Proximal Gradient (FDPG) Method—primal representation

Initialization: $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}$, $\mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{V}, t_0 = 1$.
General step ($k \geq 0$):

(a) $\mathbf{u}^k = \mathrm{argmax}_{\mathbf{u}}\left\{\langle \mathbf{u}, \mathcal{A}^T(\mathbf{w}^k)\rangle - f(\mathbf{u})\right\}$;

(b) $\mathbf{y}^{k+1} = \mathbf{w}^k - \frac{1}{L}\mathcal{A}(\mathbf{u}^k) + \frac{1}{L}\mathrm{prox}_{Lg}(\mathcal{A}(\mathbf{u}^k) - L\mathbf{w}^k)$;

(c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;

(d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.

The primal sequence that we will be interested in is actually not computed during the steps of the FDPG method. The definition of the primal sequence on which a convergence result will be proved is

$$\mathbf{x}^k = \mathrm{argmax}_{\mathbf{x}\in\mathbb{E}}\left\{\langle \mathbf{x}, \mathcal{A}^T(\mathbf{y}^k)\rangle - f(\mathbf{x})\right\}. \tag{12.19}$$

The convergence result on the primal sequence is given below, and its proof is almost a verbatim repetition of the proof of Theorem 12.8.

Theorem 12.10 ($O(1/k^2)$ convergence of the primal sequence of the FDPG method). *Suppose that Assumption* 12.1 *holds, and let* $\{\mathbf{y}^k\}_{k\geq 0}$ *be the sequence generated by the FDPG method with* $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}$. *Let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence defined by* (12.19). *Then for any optimal solution* $\mathbf{y}^*$ *of the dual problem* (12.4) *and* $k \geq 1$,

$$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{4L\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{\sigma(k+1)^2}.$$

Proof. Invoking Lemma 12.7 with $\bar{\mathbf{y}} = \mathbf{y}^k$, we obtain by the definition of $\bar{\mathbf{x}}$ (equation (12.12)) that $\bar{\mathbf{x}} = \mathbf{x}^k$, and hence the result (12.13) reads as

$$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{2}{\sigma}(q_{\text{opt}} - q(\mathbf{y}^k)),$$

which, combined with Theorem 12.9, yields the desired result. ∎

12.4 Examples I

12.4.1 Orthogonal Projection onto a Polyhedral Set

Let

$$S = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{b}\},$$

where $\mathbf{A} \in \mathbb{R}^{p\times n}$, $\mathbf{b} \in \mathbb{R}^p$. We assume that S is nonempty. Let $\mathbf{d} \in \mathbb{R}^n$. The orthogonal projection of $\mathbf{d}$ onto S is the unique optimal solution of the problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{2}\|\mathbf{x} - \mathbf{d}\|^2 : \mathbf{A}\mathbf{x} \leq \mathbf{b} \right\}. \tag{12.20}$$

Problem (12.20) fits model (12.1) with $\mathbb{E} = \mathbb{R}^n, \mathbb{V} = \mathbb{R}^p$, $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{d}\|^2$,

$$g(\mathbf{z}) = \delta_{\text{Box}[-\infty\mathbf{e},\mathbf{b}]}(\mathbf{z}) = \begin{cases} \mathbf{0}, & \mathbf{z} \leq \mathbf{b}, \\ \infty & \text{else}, \end{cases}$$

and $\mathcal{A}(\mathbf{x}) \equiv \mathbf{A}\mathbf{x}$. We have

- $\operatorname{argmax}_{\mathbf{x}}\{\langle \mathbf{v}, \mathbf{x}\rangle - f(\mathbf{x})\} = \mathbf{v} + \mathbf{d}$ for any $\mathbf{v} \in \mathbb{R}^n$;
- $\|\mathcal{A}\| = \|\mathbf{A}\|_{2,2}$;
- $\sigma = 1$;
- $\mathcal{A}^T(\mathbf{y}) = \mathbf{A}^T\mathbf{y}$ for any $\mathbf{y} \in \mathbb{R}^p$;
- $\operatorname{prox}_{Lg}(\mathbf{z}) = P_{\text{Box}[-\infty\mathbf{e},\mathbf{b}]}(\mathbf{z}) = \min\{\mathbf{z}, \mathbf{b}\}$, where $\min\{\mathbf{z}, \mathbf{b}\}$ is the vector $(\min\{z_i, b_i\})_{i=1}^p$.

Using these facts, the DPG and FDPG methods for solving problem (12.20) can be explicitly written.

Algorithm 1 [DPG for solving (12.20)]

- **Initialization:** $L \geq \|\mathbf{A}\|_{2,2}^2, \mathbf{y}^0 \in \mathbb{R}^p$.
- **General step ($k \geq 0$):**
  - (a) $\mathbf{x}^k = \mathbf{A}^T\mathbf{y}^k + \mathbf{d}$;
  - (b) $\mathbf{y}^{k+1} = \mathbf{y}^k - \frac{1}{L}\mathbf{A}\mathbf{x}^k + \frac{1}{L}\min\{\mathbf{A}\mathbf{x}^k - L\mathbf{y}^k, \mathbf{b}\}$.

Algorithm 2 [FDPG for solving (12.20)]

- **Initialization:** $L \geq \|\mathbf{A}\|_{2,2}^2, \mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{R}^p, t_0 = 1.$
- **General step ($k \geq 0$):**
  - (a) $\mathbf{u}^k = \mathbf{A}^T\mathbf{w}^k + \mathbf{d}$;
  - (b) $\mathbf{y}^{k+1} = \mathbf{w}^k - \frac{1}{L}\mathbf{A}\mathbf{u}^k + \frac{1}{L}\min\{\mathbf{A}\mathbf{u}^k - L\mathbf{w}^k, \mathbf{b}\}$;
  - (c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
  - (d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.

The primal sequence for the FDPG method is given by $\mathbf{x}^k = \mathbf{A}^T\mathbf{y}^k + \mathbf{d}$.

12.4.2 Orthogonal Projection onto the Intersection of Closed Convex Sets

Given p closed and convex sets $C_1, C_2, \ldots, C_p \subseteq \mathbb{E}$ and a point $\mathbf{d} \in \mathbb{E}$, the orthogonal projection of $\mathbf{d}$ onto the intersection $\cap_{i=1}^p C_i$ is the optimal solution of the problem

$$\min_{\mathbf{x}\in\mathbb{E}} \left\{ \frac{1}{2}\|\mathbf{x} - \mathbf{d}\|^2 : \mathbf{x} \in \cap_{i=1}^p C_i \right\}. \tag{12.21}$$

We will assume that the intersection $\cap_{i=1}^p C_i$ is nonempty and that projecting onto each set C_i is an easy task. Our purpose will be to devise a method for solving problem (12.21) that only requires computing at each iteration—in addition to elementary linear algebra operations—orthogonal projections onto the sets C_i. Problem (12.21) fits model (12.1) with $\mathbb{V} = \mathbb{E}^p, f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{d}\|^2, g(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) = \sum_{i=1}^p \delta_{C_i}(\mathbf{x}_i)$, and $\mathcal{A} : \mathbb{E} \to \mathbb{V}$ given by

$$\mathcal{A}(\mathbf{z}) = (\underbrace{\mathbf{z}, \mathbf{z}, \ldots, \mathbf{z}}_{p \text{ times}}) \text{ for any } \mathbf{z} \in \mathbb{E}.$$

We have

- $\operatorname{argmax}_{\mathbf{x}}\{\langle \mathbf{v}, \mathbf{x}\rangle - f(\mathbf{x})\} = \mathbf{v} + \mathbf{d}$ for any $\mathbf{v} \in \mathbb{E}$;
- $\|\mathcal{A}\|^2 = p$;
- $\sigma = 1$;
- $\mathcal{A}^T(\mathbf{y}) = \sum_{i=1}^p y_i$ for any $\mathbf{y} \in \mathbb{E}^p$;
- $\operatorname{prox}_{Lg}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p) = (P_{C_1}(\mathbf{v}_1), P_{C_2}(\mathbf{v}_2), \ldots, P_{C_p}(\mathbf{v}_p))$ for any $\mathbf{v} \in \mathbb{E}^p$.

Using these facts, the DPG and FDPG methods for solving problem (12.21) can be explicitly written.

Algorithm 3 [DPG for solving (12.21)]

- **Initialization:** $L \geq p, \mathbf{y}^0 \in \mathbb{E}^p$.
- **General step ($k \geq 0$):**
  (a) $\mathbf{x}^k = \sum_{i=1}^p \mathbf{y}_i^k + \mathbf{d}$;
  (b) $\mathbf{y}_i^{k+1} = \mathbf{y}_i^k - \frac{1}{L}\mathbf{x}^k + \frac{1}{L}P_{C_i}(\mathbf{x}^k - L\mathbf{y}_i^k),\ i = 1, 2, \ldots, p.$

Algorithm 4 [FDPG for solving (12.21)]

- **Initialization:** $L \geq p, \mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{E}^p, t_0 = 1.$
- **General step ($k \geq 0$):**
  (a) $\mathbf{u}^k = \sum_{i=1}^p \mathbf{w}_i^k + \mathbf{d}$;
  (b) $\mathbf{y}_i^{k+1} = \mathbf{w}_i^k - \frac{1}{L}\mathbf{u}^k + \frac{1}{L}P_{C_i}(\mathbf{u}^k - L\mathbf{w}_i^k),\ i = 1, 2, \ldots, p$;
  (c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
  (d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.

To actually guarantee convergence of the method, Assumption 12.1 needs to be satisfied, meaning that we assume that $\cap_{i=1}^p \mathrm{ri}(C_i) \neq \emptyset$.

The primal sequence for the FDPG method is given by $\mathbf{x}^k = \sum_{i=1}^p \mathbf{y}_i^k + \mathbf{d}$.

Example 12.11 (orthogonal projection onto a polyhedral set revisited). Note that Algorithm 4 can also be used to find an orthogonal projection of a point $\mathbf{d} \in \mathbb{R}^n$ onto the polyhedral set $C = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$, where $\mathbf{A} \in \mathbb{R}^{p \times n}, \mathbf{b} \in \mathbb{R}^p$. Indeed, C can be written as the following intersection of half-spaces:

$$C = \cap_{i=1}^p C_i,$$

where

$$C_i = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}_i^T \mathbf{x} \leq b_i\}, \tag{12.22}$$

with $\mathbf{a}_1^T, \mathbf{a}_2^T, \ldots, \mathbf{a}_p^T$ being the rows of $\mathbf{A}$. The projections on the half-spaces are simple and given by (see Lemma 6.26) $P_{C_i}(\mathbf{x}) = \mathbf{x} - \frac{[\mathbf{a}_i^T \mathbf{x} - b_i]_+}{\|\mathbf{a}_i\|^2}\mathbf{a}_i$. To summarize, the problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \frac{1}{2}\|\mathbf{x} - \mathbf{d}\|_2^2 : \mathbf{A}\mathbf{x} \leq \mathbf{b} \right\}$$

can be solved by two different FDPG methods. The first one is Algorithm 2, and the second one is the following algorithm, which is Algorithm 4 specified to the case where C_i is given by (12.22) for any $i \in \{1, 2, \ldots, p\}$.

Algorithm 5 [second version of FDPG for solving (12.20)]

- **Initialization:** $L \geq p, \mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{E}^p, t_0 = 1.$
- **General step ($k \geq 0$):**
  - (a) $\mathbf{u}^k = \sum_{i=1}^p \mathbf{w}_i^k + \mathbf{d}$;
  - (b) $\mathbf{y}_i^{k+1} = -\frac{1}{L\|\mathbf{a}_i\|^2}[\mathbf{a}_i^T(\mathbf{u}^k - L\mathbf{w}_i^k) - b_i]_+\mathbf{a}_i,\ i = 1, 2, \ldots, p$;
  - (c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
  - (d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.

Example 12.12 (comparison between DPG and FDPG). The $O(1/k^2)$ rate of convergence obtained for the FDPG method (Theorem 12.10) is better than the $O(1/k)$ result obtained for the DPG method (Theorem 12.8). To illustrate that this theoretical advantage is also reflected in practice, we consider the problem of projecting the point $(0.5, 1.9)^T$ onto a dodecagon—a regular polygon with 12 edges, which is represented as the intersection of 12 half-spaces. The first 10 iterations of the DPG and FDPG methods with $L = p = 12$ can be seen in Figure 12.1, where the DPG and FDPG methods that were used are those described by Algorithms 3 and 4 for the intersection of closed convex sets (which are taken as the 12 half-spaces in this example) and not Algorithms 1 and 2. Evidently, the FDPG method was able to find a good approximation of the projection after 10 iterations, while the DPG method was rather far from the required solution.

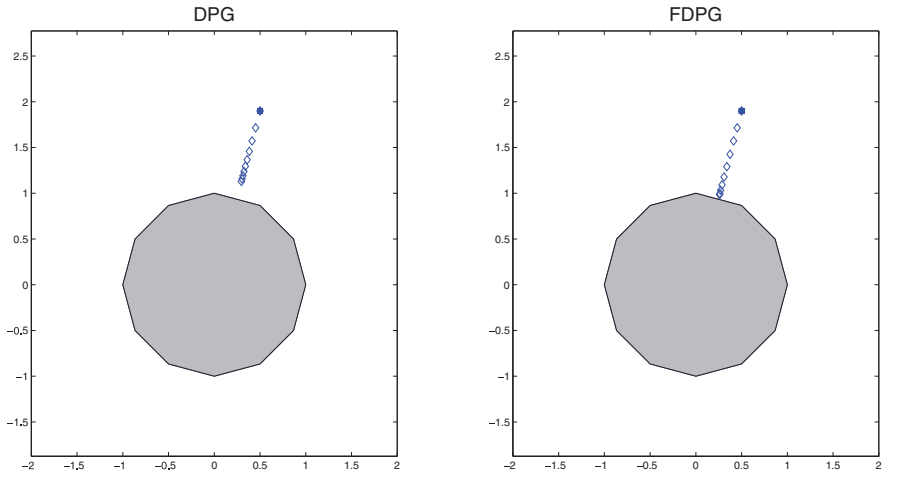

Figure 12.1. *First* 10 *iterations of the DPG method (Algorithm* 3*) and the FDPG method (Algorithm* 4/5*). The initial value of the dual vector* $\mathbf{y}$ *was the zeros vector in both methods.*

12.4.3 One-Dimensional Total Variation Denoising

In the denoising problem we are given a signal $\mathbf{d} \in \mathbb{E}$, which is contaminated by noise, and we seek to find another vector $\mathbf{x} \in \mathbb{E}$, which, on the one hand, is close to $\mathbf{d}$ in the sense that the norm $\|\mathbf{x}-\mathbf{d}\|$ is small and, on the other hand, yields a small regularization term $R(\mathcal{A}(\mathbf{x}))$, where here $\mathcal{A} : \mathbb{E} \to \mathbb{V}$ is a linear transformation that in many applications accounts for the smoothness of the signal and $R : \mathbb{V} \to \mathbb{R}_+$ is a given convex function that measures the magnitude of its argument in some sense. The denoising problem is then defined to be

$$\min_{\mathbf{x}\in\mathbb{E}} \left\{ \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|^2 + R(\mathcal{A}(\mathbf{x})) \right\}. \tag{12.23}$$

In the one-dimensional total variation denoising problem, we are interested in the case where $\mathbb{E} = \mathbb{R}^n, \mathbb{V} = \mathbb{R}^{n-1}, \mathcal{A}(\mathbf{x}) = \mathbf{D}\mathbf{x}$, and $R(\mathbf{z}) = \lambda\|\mathbf{z}\|_1$ with $\lambda > 0$ being a "regularization parameter" and $\mathbf{D}$ being the matrix satisfying $\mathbf{D}\mathbf{x} = (x_1 - x_2, x_2 - x_3, \ldots, x_{n-1} - x_n)^T$ for all $\mathbf{x} \in \mathbb{R}^n$. Thus, problem (12.23) takes the form[68]

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|_2^2 + \lambda\|\mathbf{D}\mathbf{x}\|_1 \right\} \tag{12.24}$$

or, more explicitly,

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|_2^2 + \lambda \sum_{i=1}^{n-1} |x_i - x_{i+1}| \right\}.$$

The function $\mathbf{x} \mapsto \|\mathbf{D}\mathbf{x}\|_1$ is known as a one-dimensional *total variation* function and is actually only one instance of many variants of total variation functions. Problem (12.24) fits model (12.1) with $\mathbb{E} = \mathbb{R}^n, \mathbb{V} = \mathbb{R}^{n-1}, f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|_2^2, g(\mathbf{y}) = \lambda\|\mathbf{y}\|_1$, and $\mathcal{A}(\mathbf{x}) \equiv \mathbf{D}\mathbf{x}$. In order to explicitly write the DPG and FDPG methods, we note that

- $\operatorname{argmax}_{\mathbf{x}}\{\langle \mathbf{v}, \mathbf{x}\rangle - f(\mathbf{x})\} = \mathbf{v} + \mathbf{d}$ for any $\mathbf{v} \in \mathbb{E}$;
- $\|\mathcal{A}\|^2 = \|\mathbf{D}\|_{2,2}^2 \le 4$;
- $\sigma = 1$;
- $\mathcal{A}^T(\mathbf{y}) = \mathbf{D}^T\mathbf{y}$ for any $\mathbf{y} \in \mathbb{R}^{n-1}$;
- $\operatorname{prox}_{Lg}(\mathbf{y}) = \mathcal{T}_{\lambda L}(\mathbf{y})$.

The bound on $\|\mathbf{D}\|_{2,2}$ was achieved by the following argument:

$$\|\mathbf{D}\mathbf{x}\|_2^2 = \sum_{i=1}^{n-1}(x_i - x_{i+1})^2 \le 2\sum_{i=1}^{n-1}(x_i^2 + x_{i+1}^2) \le 4\|\mathbf{x}\|^2.$$

The DPG and FDPG methods with $L = 4$ are explicitly written below.

[68] Since in this chapter all underlying spaces are Euclidean and since the standing assumption is that (unless otherwise stated) $\mathbb{R}^n$ is embedded with the dot product, it follows that $\mathbb{R}^n$ is endowed with the l_2-norm.

Algorithm 6 [DPG for solving (12.24)]

- **Initialization:** $\mathbf{y}^0 \in \mathbb{R}^{n-1}$.
- **General step ($k \geq 0$):**
  (a) $\mathbf{x}^k = \mathbf{D}^T\mathbf{y}^k + \mathbf{d}$;
  (b) $\mathbf{y}^{k+1} = \mathbf{y}^k - \frac{1}{4}\mathbf{D}\mathbf{x}^k + \frac{1}{4}\mathcal{T}_{4\lambda}(\mathbf{D}\mathbf{x}^k - 4\mathbf{y}^k)$.

Algorithm 7 [FDPG for solving (12.24)]

- **Initialization:** $\mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{R}^{n-1}, t_0 = 1$.
- **General step ($k \geq 0$):**
  (a) $\mathbf{u}^k = \mathbf{D}^T\mathbf{w}^k + \mathbf{d}$;
  (b) $\mathbf{y}^{k+1} = \mathbf{w}^k - \frac{1}{4}\mathbf{D}\mathbf{u}^k + \frac{1}{4}\mathcal{T}_{4\lambda}(\mathbf{D}\mathbf{u}^k - 4\mathbf{w}^k)$;
  (c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
  (d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.

Example 12.13. Consider the case where $n = 1000$ and the "clean" (actually unknown) signal is the vector $\mathbf{d}^{\text{true}}$, which is a discretization of a step function:

$$d_i^{\text{true}} = \begin{cases} 1, & 1 \leq i \leq 250, \\ 3, & 251 \leq i \leq 500, \\ 0, & 501 \leq i \leq 750, \\ 2, & 751 \leq i \leq 1000. \end{cases}$$

The observed vector $\mathbf{d}$ was constructed by adding independently to each component of $\mathbf{d}^{\text{true}}$ a normally distributed noise with zero mean and standard deviation 0.05. The true and noisy signals can be seen in Figure 12.2. We ran 100 iterations of Algorithms 6 (DPG) and 7 (FDPG) initialized with $\mathbf{y}^0 = \mathbf{0}$, and the resulting signals can be seen in Figure 12.3. Clearly, the FDPG method produces a much better quality reconstruction of the original step function than the DPG method. This is reflected in the objective function values of the vectors produced by each of the methods. The objective function values of the vectors generated by the DPG and FDPG methods after 100 iterations are 9.1667 and 8.4621, respectively, where the optimal value is 8.3031. ∎

12.4.4 Two-Dimensional Total Variation Denoising

In the two-dimensional total variation denoising problem, we are given an observed noisy matrix $\mathbf{d} \in \mathbb{R}^{m \times n}$, and we seek to solve the problem

$$\min_{\mathbf{x} \in \mathbb{R}^{m \times n}} \left\{ \frac{1}{2}\|\mathbf{x} - \mathbf{d}\|_F^2 + \lambda \text{TV}(\mathbf{x}) \right\}. \tag{12.25}$$

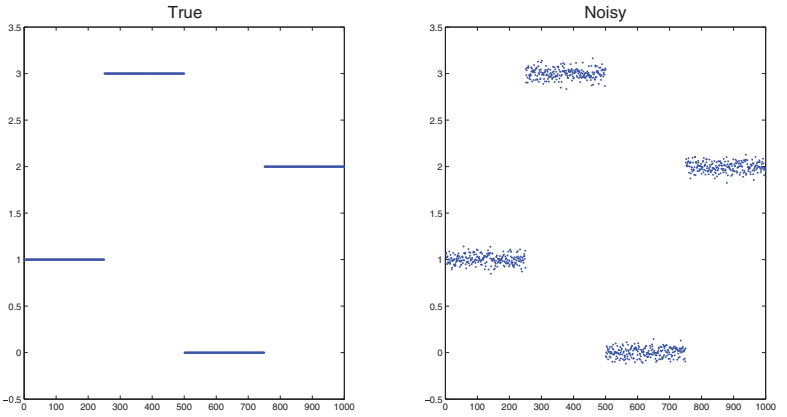

Figure 12.2. *True signal (left) and noisy signal (right).*

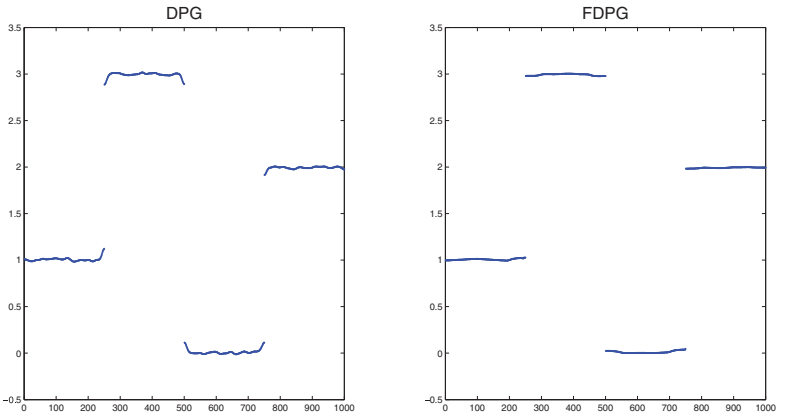

Figure 12.3. *Results of the DPG and FDPG methods.*

There are many possible choices for the two-dimensional total variation function $\mathrm{TV}(\cdot)$. Two popular choices are the isotropic TV defined for any $\mathbf{x} \in \mathbb{R}^{m\times n}$ by

$$\mathrm{TV}_I(\mathbf{x}) = \sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\sqrt{(x_{i,j}-x_{i,j+1})^2+(x_{i,j}-x_{i+1,j})^2} + \sum_{j=1}^{n-1}|x_{m,j}-x_{m,j+1}| + \sum_{i=1}^{m-1}|x_{i,n}-x_{i+1,n}| \tag{12.26}$$

and the l_1-based, anisotropic TV defined by

$$\mathbf{x}\in\mathbb{R}^{m\times n}, \quad \mathrm{TV}_{l_1}(\mathbf{x}) = \sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\{|x_{i,j}-x_{i,j+1}|+|x_{i,j}-x_{i+1,j}|\} + \sum_{j=1}^{n-1}|x_{m,j}-x_{m,j+1}| + \sum_{i=1}^{m-1}|x_{i,n}-x_{i+1,n}|.$$

Problem (12.25) fits the main model (12.1) with $\mathbb{E} = \mathbb{R}^{m\times n}, \mathbb{V} = \mathbb{R}^{m\times(n-1)} \times \mathbb{R}^{(m-1)\times n}, f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{d}\|_F^2$, and $\mathcal{A}(\mathbf{x}) = (\mathbf{p}^{\mathbf{x}}, \mathbf{q}^{\mathbf{x}})$, where $\mathbf{p}^{\mathbf{x}} \in \mathbb{R}^{m\times(n-1)}$ and $\mathbf{q}^{\mathbf{x}} \in \mathbb{R}^{(m-1)\times n}$ are given by

$$\begin{aligned} p^{\mathbf{x}}_{i,j} &= x_{i,j} - x_{i,j+1}, \quad i = 1, 2, \ldots, m, j = 1, 2, \ldots, n-1, \\ q^{\mathbf{x}}_{i,j} &= x_{i,j} - x_{i+1,j}, \quad i = 1, 2, \ldots, m-1, j = 1, 2, \ldots, n. \end{aligned}$$

The function $g : \mathbb{V} \to \mathbb{R}$ is given in the isotropic case by

$$g(\mathbf{p}, \mathbf{q}) = g_{\mathrm{I}}(\mathbf{p}, \mathbf{q}) \equiv \sum_{i=1}^{m-1}\sum_{j=1}^{n-1} \sqrt{p_{i,j}^2 + q_{i,j}^2} + \sum_{j=1}^{n-1} |p_{m,j}| + \sum_{i=1}^{m-1} |q_{i,n}|$$

and in the anisotropic case by

$$g(\mathbf{p}, \mathbf{q}) = g_{l_1}(\mathbf{p}, \mathbf{q}) \equiv \sum_{i=1}^{m}\sum_{j=1}^{n-1} |p_{i,j}| + \sum_{i=1}^{m-1}\sum_{j=1}^{n} |q_{i,j}|.$$

Since g_{I} and g_{l_1} are a separable sum of either absolute values or l_2 norms, it is easy to compute their prox mappings using Theorem 6.6 (prox of separable functions), Example 6.8 (prox of the l_1-norm), and Example 6.19 (prox of Euclidean norms) and obtain that for any $\mathbf{p} \in \mathbb{R}^{m\times(n-1)}$ and $\mathbf{q} \in \mathbb{R}^{(m-1)\times n}$,

$$\mathrm{prox}_{\lambda g_{\mathrm{I}}}(\mathbf{p}, \mathbf{q}) = (\bar{\mathbf{p}}, \bar{\mathbf{q}}),$$

where

$$\begin{aligned} \bar{p}_{i,j} &= \left(1 - \lambda / \max\left\{\sqrt{p_{i,j}^2 + q_{i,j}^2}, \lambda\right\}\right) p_{i,j}, \quad i = 1, 2, \ldots, m-1, j = 1, 2, \ldots n-1, \\ \bar{p}_{m,j} &= \mathcal{T}_\lambda(p_{m,j}), \quad j = 1, 2, \ldots, n-1, \\ \bar{q}_{i,j} &= \left(1 - \lambda / \max\left\{\sqrt{p_{i,j}^2 + q_{i,j}^2}, \lambda\right\}\right) q_{i,j}, \quad i = 1, 2, \ldots, m-1, j = 1, 2, \ldots n-1, \\ \bar{q}_{i,n} &= \mathcal{T}_\lambda(q_{i,n}), \quad i = 1, 2, \ldots, m-1, \end{aligned}$$

and

$$\mathrm{prox}_{\lambda g_{l_1}}(\mathbf{p}, \mathbf{q}) = (\tilde{\mathbf{p}}, \tilde{\mathbf{q}}),$$

where

$$\begin{aligned} \tilde{p}_{i,j} &= \mathcal{T}_\lambda(p_{i,j}), \quad i = 1, 2, \ldots, m, j = 1, 2, \ldots n-1, \\ \tilde{q}_{i,j} &= \mathcal{T}_\lambda(q_{i,j}), \quad i = 1, 2, \ldots, m-1, j = 1, 2, \ldots, n. \end{aligned}$$

The last detail that is missing in order to explicitly write the DPG or FDPG methods for solving problem (12.25) is the computation of $\mathcal{A}^T : \mathbb{V} \to \mathbb{E}$ at points in $\mathbb{V}$. For that, note that for any $\mathbf{x} \in \mathbb{E}$ and $(\mathbf{p}, \mathbf{q}) \in \mathbb{V}$,

$$\begin{aligned} \langle \mathcal{A}(\mathbf{x}), (\mathbf{p}, \mathbf{q}) \rangle &= \sum_{i=1}^{m}\sum_{j=1}^{n-1} (x_{i,j} - x_{i,j+1}) p_{i,j} + \sum_{i=1}^{m-1}\sum_{j=1}^{n} (x_{i,j} - x_{i+1,j}) q_{i,j} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} x_{i,j}(p_{i,j} + q_{i,j} - p_{i,j-1} - q_{i-1,j}) \\ &= \langle \mathbf{x}, \mathcal{A}^T(\mathbf{p}, \mathbf{q}) \rangle, \end{aligned}$$

where we use a convention that

$$p_{i,0} = p_{i,n} = q_{0,j} = q_{m,j} = 0 \quad \text{for any } i = 1, 2, \ldots, m, j = 1, 2, \ldots, n.$$

Therefore, with the above convention in mind, for any $(\mathbf{p}, \mathbf{q}) \in \mathbb{V}$,

$$\mathcal{A}^T(\mathbf{p}, \mathbf{q})_{i,j} = p_{i,j} + q_{i,j} - p_{i,j-1} - q_{i-1,j}, \quad i = 1, 2, \ldots, m, j = 1, 2, \ldots, n.$$

We also want to compute an upper bound on $\|\mathcal{A}\|^2$. This can be done using the same technique as in the one-dimensional case; note that for any $\mathbf{x} \in \mathbb{R}^{m \times n}$,

$$\begin{aligned}
\|\mathcal{A}(\mathbf{x})\|^2 &= \sum_{i=1}^{m} \sum_{j=1}^{n-1} (x_{i,j} - x_{i,j+1})^2 + \sum_{i=1}^{m-1} \sum_{j=1}^{n} (x_{i,j} - x_{i+1,j})^2 \\
&\leq 2 \sum_{i=1}^{m} \sum_{j=1}^{n-1} (x_{i,j}^2 + x_{i,j+1}^2) + 2 \sum_{i=1}^{m-1} \sum_{j=1}^{n} (x_{i,j}^2 + x_{i+1,j}^2) \\
&\leq 8 \sum_{i=1}^{n} \sum_{j=1}^{m} x_{i,j}^2.
\end{aligned}$$

Therefore, $\|\mathcal{A}\|^2 \leq 8$. We will now explicitly write the FDPG method for solving the two-dimensional anisotropic total variation problem, meaning problem (12.25) with $g = g_{l_1}$. For the stepsize, we use $L = 8$.

Algorithm 8 [FDPG for solving (12.25) with $g = \lambda \mathrm{TV}_{l_1}$]

- **Initialization:** $\tilde{\mathbf{p}}^0 = \mathbf{p}^0 \in \mathbb{R}^{m \times (n-1)}, \tilde{\mathbf{q}}^0 = \mathbf{q}^0 \in \mathbb{R}^{(m-1) \times n}, t_0 = 1$.
- **General step ($k \geq 0$):**

  (a) compute $\mathbf{u}^k \in \mathbb{R}^{m \times n}$ by setting for $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$,

  $$u_{i,j}^k = \tilde{p}_{i,j}^k + \tilde{q}_{i,j}^k - \tilde{p}_{i,j-1}^k - \tilde{q}_{i-1,j}^k + d_{i,j};$$

  (b) set $(\mathbf{p}^{k+1}, \mathbf{q}^{k+1})$ as

  $$\begin{aligned}
  p_{i,j}^{k+1} &= \tilde{p}_{i,j}^k - \frac{1}{8}(u_{i,j}^k - u_{i,j+1}^k) + \frac{1}{8}\mathcal{T}_{8\lambda}(u_{i,j}^k - u_{i,j+1}^k - 8\tilde{p}_{i,j}^k), \\
  q_{i,j}^{k+1} &= \tilde{q}_{i,j}^k - \frac{1}{8}(u_{i,j}^k - u_{i+1,j}^k) + \frac{1}{8}\mathcal{T}_{8\lambda}(u_{i,j}^k - u_{i+1,j}^k - 8\tilde{q}_{i,j}^k);
  \end{aligned}$$

  (c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;

  (d) $(\tilde{\mathbf{p}}^{k+1}, \tilde{\mathbf{q}}^{k+1}) = (\mathbf{p}^{k+1}, \mathbf{q}^{k+1}) + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{p}^{k+1} - \mathbf{p}^k, \mathbf{q}^{k+1} - \mathbf{q}^k)$.

12.5 The Dual Block Proximal Gradient Method

12.5.1 Preliminaries

In this section we will consider the problem

$$\min_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}) + \sum_{i=1}^{p} g_i(\mathbf{x}) \right\}, \tag{12.27}$$

where the following assumptions are made.

Assumption 12.14.

(A) $f : \mathbb{E} \to (-\infty, +\infty]$ *is proper closed and* σ*-strongly convex* ($\sigma > 0$).

(B) $g_i : \mathbb{E} \to (-\infty, +\infty]$ *is proper closed and convex for any* $i \in \{1, 2, \ldots, p\}$.

(C) $\mathrm{ri}(\mathrm{dom}(f)) \cap (\cap_{i=1}^{p} \mathrm{ri}(\mathrm{dom}(g_i))) \neq \emptyset$.

Problem (12.27) is actually a generalization of the projection problem discussed in Section 12.4.2, and we can use a similar observation to the one made there and note that problem (12.27) fits model (12.1) with $\mathbb{V} = \mathbb{E}^p$, $g(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) = \sum_{i=1}^{p} g_i(\mathbf{x}_i)$, and $\mathcal{A} : \mathbb{E} \to \mathbb{V}$ given by

$$\mathcal{A}(\mathbf{z}) = (\underbrace{\mathbf{z}, \mathbf{z}, \ldots, \mathbf{z}}_{p \text{ times}}) \text{ for any } \mathbf{z} \in \mathbb{E}.$$

Noting that

- $\|\mathcal{A}\|^2 = p$;
- $\mathcal{A}^T(\mathbf{y}) = \sum_{i=1}^{p} y_i$ for any $\mathbf{y} \in \mathbb{E}^p$;
- $\mathrm{prox}_{Lg}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p) = (\mathrm{prox}_{Lg_1}(\mathbf{v}_1), \mathrm{prox}_{Lg_2}(\mathbf{v}_2), \ldots, \mathrm{prox}_{Lg_p}(\mathbf{v}_p))$ for any $\mathbf{v}_i \in \mathbb{E}$, $i = 1, 2, \ldots, p$,

we can explicitly write the FDPG method with $L = \frac{\|\mathcal{A}\|^2}{\sigma} = \frac{p}{\sigma}$.

Algorithm 9 [FDPG for solving (12.27)]

- **Initialization:** $\mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{E}^p, t_0 = 1$.
- **General step ($k \geq 0$):**
  - (a) $\mathbf{u}^k = \mathrm{argmax}_{\mathbf{u}\in\mathbb{E}} \left\{ \langle \mathbf{u}, \sum_{i=1}^{p} \mathbf{w}_i^k \rangle - f(\mathbf{u}) \right\}$;
  - (b) $\mathbf{y}_i^{k+1} = \mathbf{w}_i^k - \frac{\sigma}{p}\mathbf{u}^k + \frac{\sigma}{p}\mathrm{prox}_{\frac{p}{\sigma}g_i}(\mathbf{u}^k - \frac{p}{\sigma}\mathbf{w}_i^k)$, $i = 1, 2, \ldots, p$;
  - (c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
  - (d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.

The primal sequence is given by

$$\mathbf{x}^k = \mathrm{argmax}_{\mathbf{x}\in\mathbb{E}} \left\{ \left\langle \mathbf{x}, \sum_{i=1}^{p} \mathbf{y}_i^k \right\rangle - f(\mathbf{x}) \right\}.$$

12.5.2 The Dual Block Proximal Gradient Method

Note that the stepsize taken at each iteration of Algorithm 9 is $\frac{\sigma}{p}$, which might be extremely small when the number of blocks (p) is large. The natural question is therefore whether it is possible to define a dual-based method whose stepsize is independent of the dimension. For that, let us consider the dual of problem (12.27), meaning problem (12.4). Keeping in mind that $\mathcal{A}^T(\mathbf{y}) = \sum_{i=1}^p \mathbf{y}_i$ and the fact that $g^*(\mathbf{y}) = \sum_{i=1}^p g_i^*(\mathbf{y}_i)$ (see Theorem 4.12), we obtain the following form of the dual problem:

$$q_{\text{opt}} = \max_{\mathbf{y}\in\mathbb{E}^p} \left\{ -f^*\left(\sum_{i=1}^p \mathbf{y}_i\right) - \sum_{i=1}^p \underbrace{g_i^*(-\mathbf{y}_i)}_{G_i(\mathbf{y}_i)} \right\}. \tag{12.28}$$

Since the nonsmooth part in (12.28) is block separable, we can employ a block proximal gradient method (see Chapter 11) on the dual problem (in its minimization form). Suppose that the current point is $\mathbf{y}^k = (\mathbf{y}_1^k, \mathbf{y}_2^k, \ldots, \mathbf{y}_p^k)$. At each iteration of a block proximal gradient method we pick an index i according to some rule and perform a proximal gradient step only on the ith block which is thus updated by the formula

$$\mathbf{y}_i^{k+1} = \text{prox}_{\sigma G_i}\left(\mathbf{y}_i^k - \sigma\nabla f^*\left(\sum_{j=1}^p \mathbf{y}_j^k\right)\right).$$

The stepsize was chosen to be σ since f is proper closed and σ-strongly convex, and thus, by the conjugate correspondence theorem (Theorem 5.26), f^* is $\frac{1}{\sigma}$-smooth, from which it follows that the block Lipschitz constants of the function $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p) \mapsto f^*(\sum_{i=1}^p \mathbf{y}_i)$ are $\frac{1}{\sigma}$. Thus, the constant stepsize can be taken as σ. We can now write a dual representation of the dual block proximal gradient (DBPG) method.

The Dual Block Proximal Gradient (DBPG) Method—dual representation

- **Initialization:** pick $\mathbf{y}^0 = (\mathbf{y}_1^0, \mathbf{y}_2^0, \ldots, \mathbf{y}_p^0) \in \mathbb{E}^p$.
- **General step ($k \geq 0$):**
  - pick an index $i_k \in \{1, 2, \ldots, p\}$;
  - compute $\mathbf{y}_j^{k+1} = \begin{cases} \text{prox}_{\sigma G_{i_k}}\left(\mathbf{y}_{i_k}^k - \sigma\nabla f^*(\sum_{j=1}^p \mathbf{y}_j^k)\right), & j = i_k, \\ \mathbf{y}_j^k, & j \neq i_k. \end{cases}$

We can utilize Lemma 12.5 to obtain a primal representation of the general step of the DBPG method.

Lemma 12.15. *Let f and $g_1, g_2, \ldots, g_p$ satisfy properties* (A) *and* (B) *of Assumption* 12.14. *Let $i \in \{1, 2, \ldots, p\}$ and $G_i(\mathbf{y}_i) \equiv g_i^*(-\mathbf{y}_i)$. Let $L > 0$. Then $\mathbf{y}_i \in \mathbb{E}$ and $\mathbf{v} \in \mathbb{E}^p$ satisfy the relation*

$$\mathbf{y}_i = \mathrm{prox}_{\frac{1}{L}G_i}\left(\mathbf{v}_i - \frac{1}{L}\nabla f^*\left(\sum_{j=1}^p \mathbf{v}_j\right)\right)$$

if and only if

$$\mathbf{y}_i = \mathbf{v}_i - \frac{1}{L}\tilde{\mathbf{x}} + \frac{1}{L}\mathrm{prox}_{Lg_i}\left(\tilde{\mathbf{x}} - L\mathbf{v}_i\right),$$

where

$$\tilde{\mathbf{x}} = \mathrm{argmax}_{\mathbf{x}\in\mathbb{E}}\left\{\left\langle \mathbf{x}, \sum_{j=1}^p \mathbf{v}_j\right\rangle - f(\mathbf{x})\right\}.$$

Proof. Follows by invoking Lemma 12.5 with $\mathbb{V} = \mathbb{E}$, $\mathcal{A} = \mathcal{I}$, $\mathbf{b} = \sum_{j\neq i}\mathbf{v}_j$, $g = g_i$, $\mathbf{y} = \mathbf{y}_i$, and $\mathbf{v} = \mathbf{v}_i$. ∎

Using Lemma 12.15, we can now write a primal representation of the DBPG method.

The Dual Block Proximal Gradient (DBPG) Method—primal representation

Initialization: pick $\mathbf{y}^0 = (\mathbf{y}_1^0, \mathbf{y}_2^0, \ldots, \mathbf{y}_p^0) \in \mathbb{E}$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick $i_k \in \{1, 2, \ldots, p\}$;

(b) set $\mathbf{x}^k = \mathrm{argmax}_{\mathbf{x}\in\mathbb{E}}\left\{\langle \mathbf{x}, \sum_{j=1}^p \mathbf{y}_j^k\rangle - f(\mathbf{x})\right\}$;

(c) set $\mathbf{y}_j^{k+1} = \begin{cases} \mathbf{y}_{i_k}^k - \sigma\mathbf{x}^k + \sigma\mathrm{prox}_{g_i/\sigma}\left(\mathbf{x}^k - \mathbf{y}_{i_k}^k/\sigma\right), & j = i_k, \\ \mathbf{y}_j^k, & j \neq i_k. \end{cases}$

Note that the derived DBPG method is a functional decomposition method, as it utilizes only one of the functions $g_1, g_2, \ldots, g_p$ at each iteration, and in addition the computation involving the function f (step (b)) does not involve any other function. Thus, we obtained that in this case a variables decomposition method in the dual space gives rise to a functional decomposition method in the primal space.

What is missing from the above description of the DBPG method is the index selection strategy, meaning the rule for choosing i_k at each iteration. We will consider two variations.

- **Cyclic.** $i_k = (k \mod p) + 1$.
- **Randomized.** i_k is randomly picked from $\{1, 2, \ldots, p\}$ by a uniform distribution.

12.5.3 Convergence Analysis

The rate of convergence of the DBPG method is a simple consequence of the rates of convergence already established for the block proximal gradient method in Chapter 11 combined with the primal-dual relation presented in Lemma 12.7.

Cyclic Block Order

Recall that since the model (12.27) is a special case of the general model (12.1) (with $\mathbb{V} = \mathbb{E}^p, \mathcal{A} : \mathbf{z} \mapsto (\mathbf{z}, \mathbf{z}, \ldots, \mathbf{z}), g(\mathbf{x}) = \sum_{i=1}^p g_i(\mathbf{x}_i)$), then under Assumption 12.14 the strong duality theorem (Theorem 12.2) holds, and thus the dual problem (12.28) has a nonempty optimal set. We will denote the set of dual optimal solutions by Λ^*. The following assumption is required to present a convergence result for the DBPG method with a cyclic index selection strategy.

Assumption 12.16. *For any $\alpha > 0$, there exists $R_\alpha > 0$ such that*

$$\max_{\mathbf{y}, \mathbf{y}^* \in \mathbb{E}^p} \{\|\mathbf{y} - \mathbf{y}^*\| : q(\mathbf{y}) \geq \alpha, \mathbf{y}^* \in \Lambda^*\} \leq R_\alpha,$$

where $q(\mathbf{y}) \equiv -f^*(\sum_{i=1}^p \mathbf{y}_i) - \sum_{i=1}^p g_i^*(-\mathbf{y}_i)$.

Theorem 12.17 ($O(1/k)$ rate of convergence of DBPG with cyclic order). *Suppose that Assumptions* 12.14 *and* 12.16 *hold. Let $\{\mathbf{x}^k\}_{k\geq 0}$ and $\{\mathbf{y}^k\}_{k\geq 0}$ be the primal and dual sequences generated by the DBPG method with cyclic index selection strategy for solving problem* (12.27)*. Then for any $k \geq 2$,*

(a) $q_{\rm opt} - q(\mathbf{y}^{pk}) \leq \max\left\{\left(\frac{1}{2}\right)^{(k-1)/2}(q_{\rm opt} - q(\mathbf{y}^0)), \frac{8p(p+1)^2R^2}{\sigma(k-1)}\right\}$;

(b) $\|\mathbf{x}^{pk} - \mathbf{x}^*\|^2 \leq \frac{2}{\sigma}\max\left\{\left(\frac{1}{2}\right)^{(k-1)/2}(q_{\rm opt} - q(\mathbf{y}^0)), \frac{8p(p+1)^2R^2}{\sigma(k-1)}\right\}$.

In the above two formulas $R = R_{q(\mathbf{y}^0)}$.

Proof. (a) The proof follows by invoking Theorem 11.18 while taking into account that in this case the constants in (11.24) are given by $L_{\max} = L_{\min} = \frac{1}{\sigma}, L_f = \frac{p}{\sigma}$.

(b) By the primal-dual relation, Lemma 12.7, $\|\mathbf{x}^{pk} - \mathbf{x}^*\|^2 \leq \frac{2}{\sigma}(q_{\rm opt} - q(\mathbf{y}^{pk}))$, which, combined with part (a), yields the inequality of part (b). ∎

Randomized Block Order

A direct result of the $O(1/k)$ rate of convergence of the RBPG method presented in Theorem 11.25 along with the primal-dual relation (Lemma 12.7) yields the following result on the convergence of the DBPG method with random index selection strategy. As in Section 11.5, we will use the notation of the random variable

$\xi_k \equiv \{i_0, i_1, \ldots, i_k\}$. Note that in the randomized setting we do not require Assumption 12.16 to hold.

Theorem 12.18 ($O(1/k)$ rate of convergence of DBPG with randomized order). *Suppose that Assumption* 12.14 *holds. Let* $\{\mathbf{x}^k\}_{k\geq 0}$ *and* $\{\mathbf{y}^k\}_{k\geq 0}$ *be primal and dual sequences generated by the DBPG method with randomized index selection strategy. Then for any* $k \geq 0$,

(a) $q_{\text{opt}} - \mathsf{E}_{\xi_k}(q(\mathbf{y}^{k+1})) \leq \frac{p}{p+k+1}\left(\frac{1}{2\sigma}\|\mathbf{y}^0 - \mathbf{y}^*\|^2 + q_{\text{opt}} - q(\mathbf{y}^0)\right)$;

(b) $\mathsf{E}_{\xi_k}\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 \leq \frac{2p}{\sigma(p+k+1)}\left(\frac{1}{2\sigma}\|\mathbf{y}^0 - \mathbf{y}^*\|^2 + q_{\text{opt}} - q(\mathbf{y}^0)\right)$.

12.5.4 Acceleration in the Two-Block Case[69]

Both the deterministic and the randomized DBPG methods are not accelerated methods, and consequently it was only possible to show that they exhibit an $O(1/k)$ rate of convergence. In the case where $p = 2$, we will show that it is actually possible to derive an accelerated dual block proximal gradient method by using a simple trick. For that, note that when $p = 2$, the model amounts to

$$f_{\text{opt}} = \min_{\mathbf{x}\in\mathbb{E}}\{F(\mathbf{x}) \equiv f(\mathbf{x}) + g_1(\mathbf{x}) + g_2(\mathbf{x})\}. \tag{12.29}$$

We can rewrite the problem as

$$\min_{\mathbf{x}\in\mathbb{E}}\{\tilde{f}(\mathbf{x}) + g_2(\mathbf{x})\}, \tag{12.30}$$

where $\tilde{f} = f + g_1$. If Assumption 12.14 holds with $p = 2$, then $\tilde{f}$ is proper closed and σ-strongly convex, g_2 is proper closed and convex, and the regularity condition $\text{ri}(\text{dom}(\tilde{f})) \cap \text{ri}(\text{dom}(g_2)) \neq \emptyset$ is satisfied. This means that Assumption 12.1 holds for $f = \tilde{f}$, $g = g_2$, and $\mathcal{A} = \mathcal{I}$. We can now define the *accelerated dual block proximal gradient* (ADBPG), which is the FDPG method with stepsize σ employed on the model (12.1) with $f = \tilde{f}$, $g = g_2$, and $\mathcal{A} = \mathcal{I}$.

The ADBPG Method

Initialization: $\mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{E}, t_0 = 1$.
General step ($k \geq 0$):

(a) $\mathbf{u}^k = \text{argmax}_{\mathbf{u}}\left\{\langle \mathbf{u}, \mathbf{w}^k\rangle - f(\mathbf{u}) - g_1(\mathbf{u})\right\}$;

(b) $\mathbf{y}^{k+1} = \mathbf{w}^k - \sigma\mathbf{u}^k + \sigma\text{prox}_{g_2/\sigma}(\mathbf{u}^k - \mathbf{w}^k/\sigma)$;

(c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;

(d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.

[69] The accelerated method ADBPG is a different representation of the accelerated method proposed by Chambolle and Pock in [41].

A direct consequence of Theorem 12.10 is the following result on the rate of convergence of the ADBPG method.

Theorem 12.19 ($O(1/k^2)$ rate of convergence of ADBPG). *Suppose that Assumption* 12.14 *holds with $p = 2$, and let $\{\mathbf{y}^k\}_{k\geq 0}$ be the sequence generated by the ADBPG method. Then for any optimal solution $\mathbf{y}^*$ of the dual problem*

$$\min_{\mathbf{y}\in\mathbb{E}}\{(\tilde{f})^*(\mathbf{y}) + g_2^*(-\mathbf{y})\}$$

and $k \geq 1$, it holds that

$$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{4\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{\sigma^2(k+1)^2},$$

where $\mathbf{x}^k = \operatorname{argmax}_{\mathbf{x}} \left\{\langle \mathbf{x}, \mathbf{y}^k\rangle - f(\mathbf{x}) - g_1(\mathbf{x})\right\}$.

Remark 12.20. *When $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{d}\|^2$ for some $\mathbf{d} \in \mathbb{E}$, step* (a) *of the ADBPG can be written as a prox computation:*

$$\mathbf{u}^k = \operatorname{prox}_{g_1}(\mathbf{d} + \mathbf{w}^k).$$

Remark 12.21. *Note that the ADBPG is not a full functional decomposition method since step* (a) *is a computation involving both f and g_1, but it still separates between g_1 and g_2. The method has two main features. First, it is an accelerated method. Second, the stepsize taken in the method is σ, in contrast to the stepsize of $\frac{\sigma}{2}$ that is used in Algorithm* 9*, which is another type of an FDPG method.*

12.6 Examples II

Example 12.22 (one-dimensional total variation denoising). In this example we will compare the performance of the ADBPG method and Algorithm 9 (with $p = 2$)—both are FDPG methods, although quite different. We will consider the one-dimensional total variation problem (see also Section 12.4.3)

$$f_{\text{opt}} = \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ F(\mathbf{x}) \equiv \frac{1}{2}\|\mathbf{x} - \mathbf{d}\|_2^2 + \lambda \sum_{i=1}^{n-1} |x_{i-1} - x_i| \right\}, \tag{12.31}$$

where $\mathbf{d} \in \mathbb{R}^n$ and $\lambda > 0$. The above problem can be written as

$$\min_{\mathbf{x}\in\mathbb{R}^n}\{f(\mathbf{x}) + g_1(\mathbf{x}) + g_2(\mathbf{x})\},$$

where

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{d}\|_2^2,$$

$$g_1(\mathbf{x}) = \lambda \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} |x_{2i-1} - x_{2i}|,$$

$$g_2(\mathbf{x}) = \lambda \sum_{i=1}^{\lfloor \frac{n-1}{2} \rfloor} |x_{2i} - x_{2i+1}|.$$

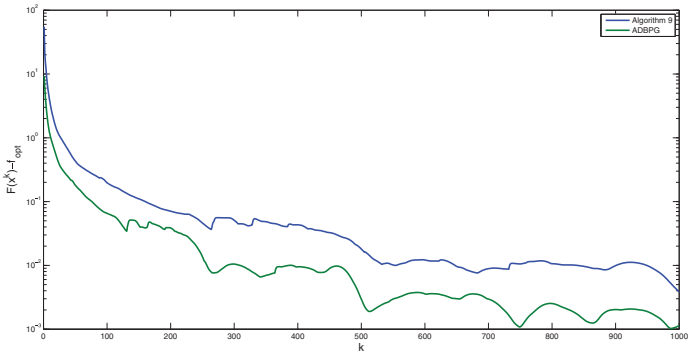

Figure 12.4. *Comparison of the ADBPG method and Algorithm* 9 *employed on the one-dimensional total variation denoising problem.*

By Example 6.17 we have that the prox of λ times the two-dimensional function $h(y,z) = |y-z|$ is given by

$$\begin{aligned}\mathrm{prox}_{\lambda h}(y,z) &= (y,z) + \frac{1}{2\lambda^2}(\mathcal{T}_{2\lambda^2}(\lambda y - \lambda z) - \lambda y + \lambda z)(\lambda, -\lambda) \\ &= (y,z) + \frac{1}{2}([|y-z| - 2\lambda]_+ \mathrm{sgn}(y-z) - y + z)(1,-1).\end{aligned}$$

Therefore, using the separability of g_1 w.r.t. the pairs of variables $\{x_1, x_2\}$, $\{x_3, x_4\}, \ldots$, it follows that

$$\mathrm{prox}_{g_1}(\mathbf{x}) = \mathbf{x} + \frac{1}{2}\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor}([|x_{2i-1} - x_{2i}| - 2\lambda]_+ \mathrm{sgn}(x_{2i-1} - x_{2i}) - x_{2i-1} + x_{2i})(\mathbf{e}_{2i-1} - \mathbf{e}_{2i}),$$

and similarly

$$\mathrm{prox}_{g_2}(\mathbf{x}) = \mathbf{x} + \frac{1}{2}\sum_{i=1}^{\lfloor \frac{n-1}{2} \rfloor}([|x_{2i} - x_{2i+1}| - 2\lambda]_+ \mathrm{sgn}(x_{2i} - x_{2i+1}) - x_{2i} + x_{2i+1})(\mathbf{e}_{2i} - \mathbf{e}_{2i+1}).$$

Equipped with the above expressions for prox_{g_1} and prox_{g_2} (recalling that step (a) only requires a single computation of prox_{g_1}; see Remark 12.20), we can employ the ADBPG method and Algorithm 9 on problem (12.31). The computational effort per iteration in both methods is almost identical and is dominated by single evaluations of the prox mappings of g_1 and g_2. We ran 1000 iterations of both algorithms starting with a dual vector which is all zeros. In Figure 12.4 we plot the distance in function values[70] $F(\mathbf{x}^k) - f_{\mathrm{opt}}$ as a function of the iteration index k. Evidently, the ADBPG method exhibits the superior performance. Most likely, the reason is the fact that the ADBPG method uses a larger stepsize (σ) than the one used by Algorithm 9 ($\frac{\sigma}{2}$). ∎

[70] Since the specific example is unconstrained, the distance in function values is indeed in some sense an "optimality measure."

Example 12.23 (two-dimensional total variation denoising). Consider the isotropic two-dimensional total variation problem

$$\min_{\mathbf{x}\in\mathbb{R}^{m\times n}} \left\{ \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|_F^2 + \lambda \mathrm{TV}_I(\mathbf{x}) \right\},$$

where $\mathbf{d} \in \mathbb{R}^{m\times n}$, $\lambda > 0$, and TV_I is given in (12.26). It does not seem possible to decompose TV_I into two functions whose prox can be directly computed as in the one-dimensional case. However, a decomposition into three separable functions (w.r.t. triplets of variables) is possible. To describe the decomposition, we introduce the following notation. Let D_k denote the set of indices that correspond to the elements of the kth diagonal of an $m \times n$ matrix, where D_0 represents the indices set of the main diagonal, and D_k for $k > 0$ and $k < 0$ stand for the diagonals above and below the main diagonal, respectively. In addition, consider the partition of the diagonal indices set, $\{-(m-1), \ldots, n-1\}$, into three sets

$$K_i \equiv \big\{k \in \{-(m-1), \ldots, n-1\} : (k+1-i) \mod 3 = 0\big\}, \qquad i = 1, 2, 3.$$

With the above notation, we are now ready to write the function TV_I as

$$\begin{aligned}
\mathrm{TV}_I(\mathbf{x}) &= \sum_{i=1}^{m}\sum_{j=1}^{n} \sqrt{(x_{i,j}-x_{i,j+1})^2 + (x_{i,j}-x_{i+1,j})^2} \\
&= \sum_{k\in K_1} \sum_{(i,j)\in D_k} \sqrt{(x_{i,j}-x_{i,j+1})^2 + (x_{i,j}-x_{i+1,j})^2} \\
&\quad + \sum_{k\in K_2} \sum_{(i,j)\in D_k} \sqrt{(x_{i,j}-x_{i,j+1})^2 + (x_{i,j}-x_{i+1,j})^2} \\
&\quad + \sum_{k\in K_3} \sum_{(i,j)\in D_k} \sqrt{(x_{i,j}-x_{i,j+1})^2 + (x_{i,j}-x_{i+1,j})^2} \\
&= \psi_1(\mathbf{x}) + \psi_2(\mathbf{x}) + \psi_3(\mathbf{x}),
\end{aligned}$$

where we assume in the above expressions that $x_{i,n+1} = x_{i,n}$ and $x_{m+1,j} = x_{m,j}$. The fact that each of the functions ψ_i is separable w.r.t. triplets of variables $\{x_{i,j}, x_{i+1,j}, x_{i,j+1}\}$ is evident from the illustration in Figure 12.5.

The denoising problem can thus be rewritten as

$$\min_{\mathbf{x}\in\mathbb{R}^{m\times n}} \left\{ \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|_F^2 + \lambda\psi_1(\mathbf{x}) + \lambda\psi_2(\mathbf{x}) + \lambda\psi_3(\mathbf{x}) \right\}.$$

It is not possible to employ the ADBPG method since the nonsmooth part is decomposed into three functions. However, it is possible to employ the DBPG method, which has no restriction on the number of functions. The algorithm requires evaluating a prox mapping of one of the functions $\lambda\psi_i$ at each iteration. By the separability of these functions, it follows that each prox computation involves several prox computations of three-dimensional functions of the form λh, where

$$h(x, y, z) = \sqrt{(x-y)^2 + (x-z)^2}.$$

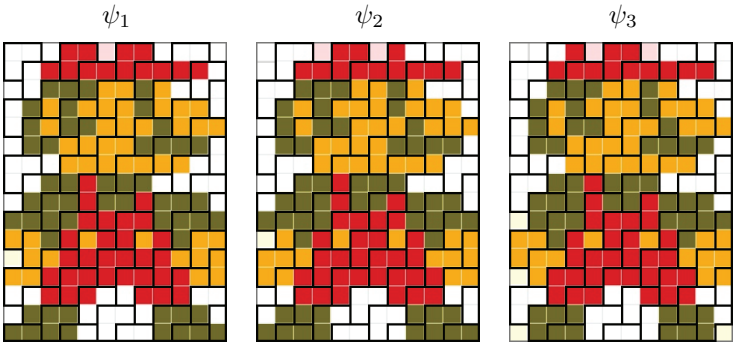

Figure 12.5. *The decomposition of a* 16×12 *pixels Mario image according to the isotropic TV into three separable functions. The images are partitioned into blocks of three pixels positioned in an r-shaped structure. Each block encompasses the three pixels that form the term* $\sqrt{(x_{i,j} - x_{i+1,j})^2 + (x_{i,j} - x_{i,j+1})^2}$. *Summing over all the terms represented by the blocks of any of the above images yields the appropriate separable function. Reprinted with permission from Elsevier.* [23]

The prox of λh can be computed using Lemma 6.68 and is given by

$$\text{prox}_{\lambda h}(\mathbf{x}) = \begin{cases} \mathbf{x} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{x}, & \|(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{x}\|_2 \le \lambda, \\ \mathbf{x} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \alpha^*\mathbf{I})^{-1}\mathbf{A}\mathbf{x}, & \|(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{x}\|_2 > \lambda, \end{cases}$$

where α^* is the unique root of the decreasing function

$$g(\alpha) = \|(\mathbf{A}\mathbf{A}^T + \alpha^*\mathbf{I})^{-1}\mathbf{A}\mathbf{x}\|_2^2 - \lambda^2$$

and $\mathbf{A}$ is the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}. \quad \blacksquare$$

Chapter 13
The Generalized Conditional Gradient Method

Underlying Spaces: In this chapter, all the underlying spaces are Euclidean.

13.1 The Frank–Wolfe/Conditional Gradient Method

Consider the problem

$$\min\{f(\mathbf{x}) : \mathbf{x} \in C\}, \tag{13.1}$$

where $C \subseteq \mathbb{E}$ is a nonempty convex and compact set and $f : \mathbb{E} \to (-\infty, \infty]$ is a convex function satisfying $C \subseteq \operatorname{dom}(f)$. We further assume that $\operatorname{dom}(f)$ is open and that f is differentiable over $\operatorname{dom}(f)$. One method that can be employed in order to solve the problem is the projected gradient method (see Section 10.2) whose update step is

$$\mathbf{x}^{k+1} = P_C(\mathbf{x}^k - t_k \nabla f(\mathbf{x}^k)),$$

with t_k being an appropriately chosen stepsize. In this chapter we will consider an alternative approach that does not require the evaluation of the orthogonal projection operator at each iteration. Instead, the approach, known as the *conditional gradient* method or *Frank–Wolfe* algorithm, computes the next step as a convex combination of the current iterate and a minimizer of a linearized version of the objective function over C.

The Conditional Gradient Method

Initialization: pick $\mathbf{x}^0 \in C$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) compute $\mathbf{p}^k \in \operatorname{argmin}_{\mathbf{p} \in C} \langle \mathbf{p}, \nabla f(\mathbf{x}^k) \rangle$;

(b) choose $t_k \in [0, 1]$ and set $\mathbf{x}^{k+1} = \mathbf{x}^k + t_k(\mathbf{p}^k - \mathbf{x}^k)$.

The conditional gradient approach is potentially beneficial in cases where computation of a linear oracle over the feasible set (that is, computation of a minimizer of

a linear function over C) is a simpler task than evaluating the orthogonal projection onto C. We will actually analyze an extension of the method that tackles the problem of minimizing a composite function $f + g$, where the case $g = \delta_C$ brings us back to the model (13.1).

13.2 The Generalized Conditional Gradient Method

13.2.1 Model and Method

Consider the composite problem

$$\min \{F(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x})\}, \tag{13.2}$$

where we assume the following set of properties.

Assumption 13.1.

(A) $g : \mathbb{E} \to (-\infty, \infty]$ *is proper closed and convex and* $\mathrm{dom}(g)$ *is compact.*

(B) $f : \mathbb{E} \to (-\infty, \infty]$ *is* L_f*-smooth over* $\mathrm{dom}(f)$ $(L_f > 0)$*, which is assumed to be an open and convex set satisfying* $\mathrm{dom}(g) \subseteq \mathrm{dom}(f)$*.*

(C) *The optimal set of problem* (13.2) *is nonempty and denoted by* X^**. The optimal value of the problem is denoted by* F_{opt}*.*

It is not difficult to deduce that property (C) is implied by properties (A) and (B). The *generalized conditional gradient method* for solving the composite model (13.2) is similar to the conditional gradient method, but instead of linearizing the entire objective function, the algorithm computes a minimizer of the sum of the linearized smooth part f around the current iterate and leaves g unchanged.

The Generalized Conditional Gradient Method

Initialization: pick $\mathbf{x}^0 \in \mathrm{dom}(g)$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) compute $\mathbf{p}^k \in \mathrm{argmin}_{\mathbf{p} \in \mathbb{E}} \left\{ \langle \mathbf{p}, \nabla f(\mathbf{x}^k) \rangle + g(\mathbf{p}) \right\}$;

(b) choose $t_k \in [0, 1]$ and set $\mathbf{x}^{k+1} = \mathbf{x}^k + t_k(\mathbf{p}^k - \mathbf{x}^k)$.

13.2.2 The Conditional Gradient Norm

Throughout this chapter we will use the following notation:

$$\mathbf{p}(\mathbf{x}) \in \mathrm{argmin}_{\mathbf{p}} \{ \langle \mathbf{p}, \nabla f(\mathbf{x}) \rangle + g(\mathbf{p}) \}. \tag{13.3}$$

Of course, $\mathbf{p}(\mathbf{x})$ is not uniquely defined in the sense that the above minimization problem might have multiple optimal solutions. We assume that there exists some rule for choosing an optimal solution whenever the optimal set of (13.3) is not a

singleton and that the vector $\mathbf{p}^k$ computed by the generalized conditional gradient method is chosen by the same rule, meaning that $\mathbf{p}^k = \mathbf{p}(\mathbf{x}^k)$. We can write the update step of the generalized conditional gradient method as

$$\mathbf{x}^{k+1} = \mathbf{x}^k + t_k(\mathbf{p}(\mathbf{x}^k) - \mathbf{x}^k).$$

A natural optimality measure in the context of proximal gradient methods is the gradient mapping (see Section 10.3.2). However, the analysis of the conditional gradient method relies on a different optimality measure, which we will refer to as the *conditional gradient norm*.

Definition 13.2 (conditional gradient norm). *Suppose that f and g satisfy properties* (A) *and* (B) *of Assumption* 13.1. *Then the* **conditional gradient norm** *is the function $S : \mathrm{dom}(f) \to \mathbb{R}$ defined by*

$$S(\mathbf{x}) \equiv \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{p}(\mathbf{x}) \rangle + g(\mathbf{x}) - g(\mathbf{p}(\mathbf{x})).$$

Remark 13.3. *The conditional gradient norm obviously depends on f and g, so a more precise notation would be $S^{f,g}(\mathbf{x})$. However, since the identities of f and g will be clear from the context, we will keep the notation $S(\mathbf{x})$.*

Remark 13.4. *By the definition of $\mathbf{p}(\mathbf{x})$ (equation* (13.3)*), we can also write $S(\mathbf{x})$ as*

$$S(\mathbf{x}) = \max_{\mathbf{p} \in \mathbb{E}} \{ \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{p} \rangle + g(\mathbf{x}) - g(\mathbf{p}) \}. \tag{13.4}$$

The following lemma shows how to write the conditional gradient norm in terms of the conjugate of g.

Lemma 13.5. *Suppose that f and g satisfy properties* (A) *and* (B) *of Assumption* 13.1. *Then for any $\mathbf{x} \in \mathrm{dom}(f)$,*

$$S(\mathbf{x}) = \langle \nabla f(\mathbf{x}), \mathbf{x} \rangle + g(\mathbf{x}) + g^*(-\nabla f(\mathbf{x})). \tag{13.5}$$

Proof. Follows by the definition of the conjugate function:

$$\begin{aligned} S(\mathbf{x}) &= \max_{\mathbf{p} \in \mathbb{E}} \{ \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{p} \rangle + g(\mathbf{x}) - g(\mathbf{p}) \} \\ &= \langle \nabla f(\mathbf{x}), \mathbf{x} \rangle + g(\mathbf{x}) + \max_{\mathbf{p} \in \mathbb{E}} \{ \langle -\nabla f(\mathbf{x}), \mathbf{p} \rangle - g(\mathbf{p}) \} \\ &= \langle \nabla f(\mathbf{x}), \mathbf{x} \rangle + g(\mathbf{x}) + g^*(-\nabla f(\mathbf{x})). \end{aligned}$$

∎

A direct consequence of Lemma 13.5 is that $S(\cdot)$ is an optimality measure in the sense that it is always nonnegative and is equal to zero only at stationary points of problem (13.2).

Theorem 13.6 (conditional gradient norm as an optimality measure). *Suppose that f and g satisfy properties* (A) *and* (B) *of Assumption* 13.1. *Then*

(a) $S(\mathbf{x}) \geq 0$ *for any* $\mathbf{x} \in \mathrm{dom}(f)$*;*

(b) $S(\mathbf{x}^*) = 0$ *if and only if* $-\nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*)$*, that is, if and only if* $\mathbf{x}^*$ *is a stationary point of problem* (13.2).

Proof. (a) Follows by the expression (13.5) for the conditional gradient norm and Fenchel's inequality (Theorem 4.6).

(b) By part (a), it follows that $S(\mathbf{x}^*) = 0$ if and only if $S(\mathbf{x}^*) \leq 0$, which is the same as the relation (using the expression (13.4) for $S(\mathbf{x}^*)$)

$$\langle \nabla f(\mathbf{x}^*), \mathbf{x}^* - \mathbf{p} \rangle + g(\mathbf{x}^*) - g(\mathbf{p}) \leq 0 \text{ for all } \mathbf{p} \in \mathbb{E}.$$

After some rearrangement of terms, the above can be rewritten as

$$g(\mathbf{p}) \geq g(\mathbf{x}^*) + \langle -\nabla f(\mathbf{x}^*), \mathbf{p} - \mathbf{x}^* \rangle,$$

which is equivalent to the relation $-\nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*)$, namely, to stationarity (see Definition 3.73). ∎

The basic inequality that will be used in the analysis of the generalized conditional gradient method is the following recursive inequality.

Lemma 13.7 (fundamental inequality for generalized conditional gradient). *Suppose that* f *and* g *satisfy properties of* (A) *and* (B) *of Assumption* 13.1. *Let* $\mathbf{x} \in \mathrm{dom}(g)$ *and* $t \in [0, 1]$. *Then*

$$F(\mathbf{x} + t(\mathbf{p}(\mathbf{x}) - \mathbf{x})) \leq F(\mathbf{x}) - tS(\mathbf{x}) + \frac{t^2 L_f}{2} \|\mathbf{p}(\mathbf{x}) - \mathbf{x}\|^2. \tag{13.6}$$

Proof. Using the descent lemma (Lemma 5.7), the convexity of g, and the notation $\mathbf{p}^+ = \mathbf{p}(\mathbf{x})$, we can write the following:

$$\begin{aligned}
F(\mathbf{x} + t(\mathbf{p}^+ - \mathbf{x})) &= f(\mathbf{x} + t(\mathbf{p}^+ - \mathbf{x})) + g(\mathbf{x} + t(\mathbf{p}^+ - \mathbf{x})) \\
&\leq f(\mathbf{x}) - t\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{p}^+ \rangle + \frac{t^2 L_f}{2} \|\mathbf{p}^+ - \mathbf{x}\|^2 + g((1-t)\mathbf{x} + t\mathbf{p}^+) \\
&\leq f(\mathbf{x}) - t\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{p}^+ \rangle + \frac{t^2 L_f}{2} \|\mathbf{p}^+ - \mathbf{x}\|^2 + (1-t)g(\mathbf{x}) + tg(\mathbf{p}^+) \\
&= F(\mathbf{x}) - t(\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{p}^+ \rangle + g(\mathbf{x}) - g(\mathbf{p}^+)) + \frac{t^2 L_f}{2} \|\mathbf{p}^+ - \mathbf{x}\|^2 \\
&= F(\mathbf{x}) - tS(\mathbf{x}) + \frac{t^2 L_f}{2} \|\mathbf{p}^+ - \mathbf{x}\|^2. \quad \blacksquare
\end{aligned}$$

13.2.3 Convergence Analysis in the Nonconvex Case

Note that we do not assume at this point that f is convex, and therefore convergence (if any) will be proven to stationary points. Before we delve into the convergence analysis, we mention the different options of stepsize strategies that will be considered.

- **Predefined diminishing stepsize.** $t_k = \frac{2}{k+2}$.
- **Adaptive stepsize.** $t_k = \min\left\{1, \frac{S(\mathbf{x}^k)}{L_f\|\mathbf{p}^k - \mathbf{x}^k\|^2}\right\}$.
- **Exact line search.** $t_k \in \operatorname{argmin}_{t\in[0,1]} F(\mathbf{x}^k + t(\mathbf{p}^k - \mathbf{x}^k))$.

The motivation for considering the adaptive stepsize comes from the fundamental inequality (13.6)—it is easy to verify that $t_k = \min\left\{1, \frac{S(\mathbf{x}^k)}{L_f\|\mathbf{p}^k - \mathbf{x}^k\|^2}\right\}$ is the minimizer of the right-hand side of (13.6) w.r.t. $t \in [0,1]$ when $\mathbf{x} = \mathbf{x}^k$. Much like the analysis of the proximal gradient method, the convergence of the generalized conditional gradient method is based on a sufficient decrease property.

Lemma 13.8 (sufficient decrease for the generalized conditional gradient method). *Suppose that f and g satisfy properties* (A) *and* (B) *of Assumption* 13.1*, and let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the generalized conditional gradient method for solving problem* (13.2) *with stepsizes chosen by either the adaptive or exact line search strategies. Then for any $k \geq 0$,*

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq \frac{1}{2}\min\left\{S(\mathbf{x}^k), \frac{S^2(\mathbf{x}^k)}{L_f\Omega^2}\right\}, \tag{13.7}$$

where Ω be an upper bound on the diameter of $\operatorname{dom}(g)$*:*

$$\Omega \geq \max_{\mathbf{x},\mathbf{y}\in\operatorname{dom}(g)} \|\mathbf{x} - \mathbf{y}\|.$$

Proof. Let $k \geq 0$ and let $\tilde{\mathbf{x}}^k = \mathbf{x}^k + s_k(\mathbf{p}^k - \mathbf{x}^k)$, where

$$s_k = \min\left\{1, \frac{S(\mathbf{x}^k)}{L_f\|\mathbf{x}^k - \mathbf{p}^k\|^2}\right\}.$$

By the fundamental inequality (13.6) invoked with $\mathbf{x} = \mathbf{x}^k$ and $t = s_k$, we have

$$F(\mathbf{x}^k) - F(\tilde{\mathbf{x}}^k) \geq s_k S(\mathbf{x}^k) - \frac{s_k^2 L_f}{2}\|\mathbf{p}^k - \mathbf{x}^k\|^2. \tag{13.8}$$

There are two options: Either $\frac{S(\mathbf{x}^k)}{L_f\|\mathbf{x}^k - \mathbf{p}^k\|^2} \leq 1$, and in this case $s_k = \frac{S(\mathbf{x}^k)}{L_f\|\mathbf{x}^k - \mathbf{p}^k\|^2}$, and hence, by (13.8),

$$F(\mathbf{x}^k) - F(\tilde{\mathbf{x}}^k) \geq \frac{S^2(\mathbf{x}^k)}{2L_f\|\mathbf{p}^k - \mathbf{x}^k\|^2} \geq \frac{S^2(\mathbf{x}^k)}{2L_f\Omega^2}.$$

Or, on the other hand, if

$$\frac{S(\mathbf{x}^k)}{L_f\|\mathbf{x}^k - \mathbf{p}^k\|^2} \geq 1, \tag{13.9}$$

then $s_k = 1$, and by (13.8),

$$F(\mathbf{x}^k) - F(\tilde{\mathbf{x}}^k) \geq S(\mathbf{x}^k) - \frac{L_f}{2}\|\mathbf{p}^k - \mathbf{x}^k\|^2 \overset{(13.9)}{\geq} \frac{1}{2}S(\mathbf{x}^k).$$

Combining the two cases, we obtain

$$F(\mathbf{x}^k) - F(\tilde{\mathbf{x}}^k) \geq \frac{1}{2} \min \left\{ S(\mathbf{x}^k), \frac{S^2(\mathbf{x}^k)}{L_f \Omega^2} \right\}. \tag{13.10}$$

If the adaptive stepsize strategy is used, then $\tilde{\mathbf{x}}^k = \mathbf{x}^{k+1}$ and (13.10) is the same as (13.7). If an exact line search strategy is employed, then

$$F(\mathbf{x}^{k+1}) = \min_{t \in [0,1]} F(\mathbf{x}^k + t(\mathbf{p}^k - \mathbf{x}^k)) \leq F(\mathbf{x}^k + s_k(\mathbf{p}^k - \mathbf{x}^k)) = F(\tilde{\mathbf{x}}^k),$$

which, combined with (13.10), implies that also in this case (13.7) holds. ∎

Using Lemma 13.8 we can establish the main convergence result for the generalized conditional gradient method with stepsizes chosen by either the adaptive or exact line search strategies.

Theorem 13.9 (convergence of the generalized conditional gradient). *Suppose that Assumption* 13.1 *holds, and let* $\{\mathbf{x}^k\}_{k \geq 0}$ *be the sequence generated by the generalized conditional gradient method for solving problem* (13.2) *with stepsizes chosen by either the adaptive or exact line search strategies. Then*

(a) *for any* $k \geq 0$, $F(\mathbf{x}^k) \geq F(\mathbf{x}^{k+1})$ *and* $F(\mathbf{x}^k) > F(\mathbf{x}^{k+1})$ *if* $\mathbf{x}^k$ *is not a stationary point of problem* (13.2)*;*

(b) $S(\mathbf{x}^k) \to 0$ *as* $k \to \infty$*;*

(c) *for any* $k \geq 0$,

$$\min_{n=0,1,\ldots,k} S(\mathbf{x}^n) \leq \max \left\{ \frac{2(F(\mathbf{x}^0) - F_{\text{opt}})}{k+1}, \frac{\sqrt{2L_f \Omega^2 (F(\mathbf{x}^0) - F_{\text{opt}})}}{\sqrt{k+1}} \right\}, \tag{13.11}$$

where Ω *is an upper bound on the diameter of* $\text{dom}(g)$*;*

(d) *all limit points of the sequence* $\{\mathbf{x}^k\}_{k \geq 0}$ *are stationary points of problem* (13.2).

Proof. (a) The monotonicity of $\{F(\mathbf{x}^k)\}_{k \geq 0}$ is a direct result of the sufficient decrease inequality (13.7) and the nonnegativity of $S(\mathbf{x}^k)$ (Theorem 13.6(a)). As for the second claim, if $\mathbf{x}^k$ is not a stationary point of problem (13.2), then $S(\mathbf{x}^k) > 0$ (see Theorem 13.6(b)), and hence, by the sufficient decrease inequality, $F(\mathbf{x}^k) > F(\mathbf{x}^{k+1})$.

(b) Since $\{F(\mathbf{x}^k)\}_{k \geq 0}$ is nonincreasing and bounded below (by F_{opt}), it follows that it is convergent, and in particular, $F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \to 0$ as $k \to \infty$. Therefore, by the sufficient decrease inequality (13.7), it follows that $\min \left\{ S(\mathbf{x}^k), \frac{S^2(\mathbf{x}^k)}{L_f \Omega^2} \right\} \to 0$ as $k \to \infty$, implying that $S(\mathbf{x}^k) \to 0$ as $k \to \infty$.

(c) By the sufficient decrease inequality (13.7), for all $n \geq 0$,

$$F(\mathbf{x}^n) - F(\mathbf{x}^{n+1}) \geq \frac{1}{2} \min \left\{ S(\mathbf{x}^n), \frac{S^2(\mathbf{x}^n)}{L_f \Omega^2} \right\}. \tag{13.12}$$

Summing the above inequality over $n = 0, 1, \ldots, k$,

$$F(\mathbf{x}^0) - F(\mathbf{x}^{k+1}) \geq \frac{1}{2} \sum_{n=0}^{k} \min \left\{ S(\mathbf{x}^n), \frac{S^2(\mathbf{x}^n)}{L_f \Omega^2} \right\}. \tag{13.13}$$

Using the facts that $F(\mathbf{x}^{k+1}) \geq F_{\text{opt}}$ and

$$\sum_{n=0}^{k} \min \left\{ S(\mathbf{x}^n), \frac{S^2(\mathbf{x}^n)}{L_f \Omega^2} \right\} \geq (k+1) \min_{n=0,1,\ldots,k} \left[\min \left\{ S(\mathbf{x}^n), \frac{S^2(\mathbf{x}^n)}{L_f \Omega^2} \right\} \right],$$

we obtain that

$$\min_{n=0,1,\ldots,k} \left[\min \left\{ S(\mathbf{x}^n), \frac{S^2(\mathbf{x}^n)}{L_f \Omega^2} \right\} \right] \leq \frac{2(F(\mathbf{x}^0) - F_{\text{opt}})}{k+1},$$

which implies in particular that there exists an $n \in \{0, 1, \ldots, k\}$ for which

$$\min \left\{ S(\mathbf{x}^n), \frac{S^2(\mathbf{x}^n)}{L_f \Omega^2} \right\} \leq \frac{2(F(\mathbf{x}^0) - F_{\text{opt}})}{k+1},$$

that is,

$$S(\mathbf{x}^n) \leq \max \left\{ \frac{2(F(\mathbf{x}^0) - F_{\text{opt}})}{k+1}, \frac{\sqrt{2L_f \Omega^2 (F(\mathbf{x}^0) - F_{\text{opt}})}}{\sqrt{k+1}} \right\}.$$

Since there exists $n \in \{0, 1, \ldots, k\}$ for which the above inequality holds, the result (13.11) immediately follows.

(d) Suppose that $\bar{\mathbf{x}}$ is a limit point of $\{\mathbf{x}^k\}_{k \geq 0}$. Then there exists a subsequence $\{\mathbf{x}^{k_j}\}_{j \geq 0}$ that converges to $\bar{\mathbf{x}}$. By the definition of the conditional gradient norm $S(\cdot)$, it follows that for any $\mathbf{v} \in \mathbb{E}$,

$$S(\mathbf{x}^{k_j}) \geq \langle \nabla f(\mathbf{x}^{k_j}), \mathbf{x}^{k_j} - \mathbf{v} \rangle + g(\mathbf{x}^{k_j}) - g(\mathbf{v}).$$

Passing to the limit $j \to \infty$ and using the fact that $S(\mathbf{x}^{k_j}) \to 0$ as $j \to \infty$, as well as the continuity of ∇f and the lower semicontinuity of g, we obtain that

$$0 \geq \langle \nabla f(\bar{\mathbf{x}}), \bar{\mathbf{x}} - \mathbf{v} \rangle + g(\bar{\mathbf{x}}) - g(\mathbf{v}) \text{ for any } \mathbf{v} \in \mathbb{E},$$

which is the same as the relation $-\nabla f(\bar{\mathbf{x}}) \in \partial g(\bar{\mathbf{x}})$, that is, the same as stationarity. ∎

Example 13.10 (optimization over the unit ball). Consider the problem

$$\min\{f(\mathbf{x}) : \|\mathbf{x}\| \leq 1\}, \tag{13.14}$$

where $f : \mathbb{E} \to \mathbb{R}$ is L_f-smooth. Problem (13.14) fits the general model (13.2) with $g = \delta_{B_{\|\cdot\|}[\mathbf{0},1]}$. Obviously, in this case the generalized conditional gradient method amounts to the conditional gradient method with feasible set $C = B_{\|\cdot\|}[\mathbf{0}, 1]$. Take

$\mathbf{x} \in B_{\|\cdot\|}[\mathbf{0}, 1]$. In order to find an expression for the conditional gradient norm $S(\mathbf{x})$, we first note that

$$\mathbf{p}(\mathbf{x}) \in \operatorname{argmin}_{\mathbf{p}:\|\mathbf{p}\|\leq 1}\langle \mathbf{p}, \nabla f(\mathbf{x})\rangle$$

is given by $\mathbf{p}(\mathbf{x}) = -\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}$ if $\nabla f(\mathbf{x}) \neq 0$ and can be chosen as $\mathbf{p}(\mathbf{x}) = \mathbf{0}$ if $\nabla f(\mathbf{x}) = \mathbf{0}$. Thus, in both cases, we obtain that for any $\mathbf{x} \in B_{\|\cdot\|}[\mathbf{0}, 1]$,

$$S(\mathbf{x}) = \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{p}(\mathbf{x})\rangle = \langle \nabla f(\mathbf{x}), \mathbf{x}\rangle + \|\nabla f(\mathbf{x})\|. \tag{13.15}$$

By its definition, $S(\mathbf{x}) = \infty$ for any $\mathbf{x} \notin B_{\|\cdot\|}[\mathbf{0}, 1]$. By Theorem 13.6 the above expression (13.15) is nonnegative and is equal to zero if and only if $\mathbf{x}$ is a stationary point of (13.14), which in this case means that either $\nabla f(\mathbf{x}) = \mathbf{0}$ or $\nabla f(\mathbf{x}) = \lambda \mathbf{x}$ for some $\lambda \leq 0$ (see [10, Example 9.6]).

Assuming that $S(\mathbf{x}^k) \neq 0$, the general update formula of the conditional gradient method for solving (13.14) is

$$\mathbf{x}^{k+1} = (1 - t_k)\mathbf{x}^k - t_k \frac{\nabla f(\mathbf{x}^k)}{\|\nabla f(\mathbf{x}^k)\|},$$

where $t_k \in [0, 1]$ is an appropriately chosen stepsize. By Theorem 13.9 if the stepsize is chosen by either an adaptive or exact line search strategies, convergence of $S(\mathbf{x}^k)$ to zero is guaranteed. ∎

Example 13.11 (the power method).[71] Continuing Example 13.10, let us consider the problem

$$\max_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} : \|\mathbf{x}\|_2 \leq 1 \right\}, \tag{13.16}$$

where $\mathbf{A} \in \mathbb{S}^n_+$. Problem (13.16) fits the model (13.14) with $f : \mathbb{R}^n \to \mathbb{R}$ given by $f(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}$. Consider the conditional gradient method for solving (13.16) and assume that $\mathbf{x}^k$ is not a stationary point of problem (13.2). Then

$$\mathbf{x}^{k+1} = (1 - t_k)\mathbf{x}^k + t_k \underbrace{\frac{\mathbf{A}\mathbf{x}^k}{\|\mathbf{A}\mathbf{x}^k\|_2}}_{\mathbf{p}^k}. \tag{13.17}$$

If the stepsizes are chosen by an exact line search strategy, then

$$t_k \in \operatorname{argmin}_{t\in[0,1]} f(\mathbf{x}^k + t(\mathbf{p}^k - \mathbf{x}^k)). \tag{13.18}$$

Since f is concave, it follows that either 0 or 1 is an optimal solution of (13.18), and by the fact that $\mathbf{x}^k$ is not a stationary point of problem (13.2), we can conclude by Theorem 13.9(a) that $t_k \neq 0$. We can thus choose $t_k = 1$, and the method (13.17) becomes

$$\mathbf{x}^{k+1} = \frac{\mathbf{A}\mathbf{x}^k}{\|\mathbf{A}\mathbf{x}^k\|_2},$$

which is the well-known *power method* for finding the eigenvector of $\mathbf{A}$ corresponding to the maximal eigenvalue. Theorem 13.9 guarantees that limit points of the method are stationary points of problem (13.16), meaning eigenvectors $\mathbf{A}$ corresponding to nonnegative eigenvalues. ∎

[71] The interpretation of the power method as the conditional gradient method was described in the work of Luss and Teboulle [85].

13.2.4 Convergence Analysis in the Convex Case

We will now further assume that f is convex. In this case, obviously all stationary points of problem (13.2) are also optimal points (Theorem 3.72(b)), so that Theorem 13.9 guarantees that all limit points of the sequence generated by the generalized conditional gradient method with either adaptive or exact line search stepsize strategies are optimal points. We also showed in Theorem 13.9 an $O(1/\sqrt{k})$ rate of convergence of the conditional gradient norm. Our objectives will be to show an $O(1/k)$ rate of convergence of function values to the optimal value, as well as of the conditional gradient norm to zero.

We begin by showing that when f is convex, the conditional gradient norm is lower bounded by the distance to optimality in terms of function values.

Lemma 13.12. *Suppose that Assumption* 13.1 *holds and that f is convex. Then for any $\mathbf{x} \in \mathrm{dom}(g)$,*

$$S(\mathbf{x}) \geq F(\mathbf{x}) - F_{\mathrm{opt}}.$$

Proof. Let $\mathbf{x}^* \in X^*$. Then for any $\mathbf{x} \in \mathrm{dom}(g)$,

$$\begin{aligned}
S(\mathbf{x}) &= \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{p}(\mathbf{x}) \rangle + g(\mathbf{x}) - g(\mathbf{p}(\mathbf{x})) && \text{[definition of } S\text{]} \\
&= \langle \nabla f(\mathbf{x}), \mathbf{x} \rangle + g(\mathbf{x}) - (\langle \nabla f(\mathbf{x}), \mathbf{p}(\mathbf{x}) \rangle + g(\mathbf{p}(\mathbf{x}))) \\
&\geq \langle \nabla f(\mathbf{x}), \mathbf{x} \rangle + g(\mathbf{x}) - (\langle \nabla f(\mathbf{x}), \mathbf{x}^* \rangle + g(\mathbf{x}^*)) && \text{[definition of } \mathbf{p}(\cdot) \text{ (13.3)]} \\
&= \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle + g(\mathbf{x}) - g(\mathbf{x}^*) \\
&\geq f(\mathbf{x}) - f(\mathbf{x}^*) + g(\mathbf{x}) - g(\mathbf{x}^*) && \text{[convexity of } f\text{]} \\
&= F(\mathbf{x}) - F_{\mathrm{opt}}. \quad \square
\end{aligned}$$

The convergence analysis relies on the following technical lemma on sequences of scalars.

Lemma 13.13.[72] *Let p be a positive integer, and let $\{a_k\}_{k\geq 0}$ and $\{b_k\}_{k\geq 0}$ be nonnegative sequences satisfying for any $k \geq 0$*

$$a_{k+1} \leq a_k - \gamma_k b_k + \frac{A}{2}\gamma_k^2, \tag{13.19}$$

where $\gamma_k = \frac{2}{k+2p}$ and A is a positive number. Suppose that $a_k \leq b_k$ for all k. Then

(a) $a_k \leq \frac{2\max\{A,(p-1)a_0\}}{k+2p-2}$ *for any $k \geq 1$;*

(b) *for any $k \geq 3$,*

$$\min_{n=\lfloor k/2 \rfloor + 2,\ldots,k} b_n \leq \frac{8\max\{A,(p-1)a_0\}}{k-2}.$$

[72] Lemma 13.13 is an extension of Lemma 4.4 from Bach [4].

Proof. (a) By (13.19) and the fact that $a_k \le b_k$, it follows that

$$a_{k+1} \le (1-\gamma_k)a_k + \frac{A}{2}\gamma_k^2.$$

Therefore,

$$\begin{aligned}
a_1 &\le (1-\gamma_0)a_0 + \frac{A}{2}\gamma_0^2,\\
a_2 &\le (1-\gamma_1)a_1 + \frac{A}{2}\gamma_1^2 = (1-\gamma_1)(1-\gamma_0)a_0 + \frac{A}{2}(1-\gamma_1)\gamma_0^2 + \frac{A}{2}\gamma_1^2,\\
a_3 &\le (1-\gamma_2)a_2 + \frac{A}{2}\gamma_2^2 = (1-\gamma_2)(1-\gamma_1)(1-\gamma_0)a_0\\
&\quad + \frac{A}{2}\left[(1-\gamma_2)(1-\gamma_1)\gamma_0^2 + (1-\gamma_2)\gamma_1^2 + \gamma_2^2\right].
\end{aligned}$$

In general,[73]

$$a_k \le a_0\prod_{s=0}^{k-1}(1-\gamma_s) + \frac{A}{2}\sum_{u=0}^{k-1}\left[\prod_{s=u+1}^{k-1}(1-\gamma_s)\right]\gamma_u^2. \tag{13.20}$$

Since $\gamma_k = \frac{2}{k+2p}$, it follows that

$$\begin{aligned}
\frac{A}{2}\sum_{u=0}^{k-1}\left[\prod_{s=u+1}^{k-1}(1-\gamma_s)\gamma_u^2\right] &= \frac{A}{2}\sum_{u=0}^{k-1}\left[\prod_{s=u+1}^{k-1}\frac{s+2p-2}{s+2p}\gamma_u^2\right]\\
&= \frac{A}{2}\sum_{u=0}^{k-1}\frac{(u+2p-1)(u+2p)}{(k+2p-2)(k+2p-1)}\cdot\frac{4}{(u+2p)^2}\\
&= \frac{A}{2}\sum_{u=0}^{k-1}\frac{u+2p-1}{(k+2p-2)(k+2p-1)}\cdot\frac{4}{u+2p}\\
&\le \frac{2Ak}{(k+2p-2)(k+2p-1)}.
\end{aligned} \tag{13.21}$$

In addition,

$$a_0\prod_{s=0}^{k-1}(1-\gamma_s) = a_0\prod_{s=0}^{k-1}\frac{s+2p-2}{s+2p} = a_0\frac{(2p-2)(2p-1)}{(k+2p-2)(k+2p-1)}. \tag{13.22}$$

Therefore, combining (13.20), (13.21), and (13.22),

$$\begin{aligned}
a_k &\le \frac{2Ak}{(k+2p-2)(k+2p-1)} + \frac{a_0(2p-2)(2p-1)}{(k+2p-2)(k+2p-1)}\\
&\le \frac{2\max\{A,(p-1)a_0\}(k+2p-1)}{(k+2p-2)(k+2p-1)}\\
&= \frac{2\max\{A,(p-1)a_0\}}{k+2p-2}.
\end{aligned}$$

[73] We use the convention that $\Pi_{k=\ell}^{u}c_k = 1$ whenever $\ell > u$.

(b) Replacing the index k with n in (13.19), we have

$$a_{n+1} \le a_n - \gamma_n b_n + \frac{A}{2}\gamma_n^2.$$

Summing the above inequality over $n = j, j+1, \dots, k$, we obtain that

$$a_{k+1} \le a_j - \sum_{n=j}^{k} \gamma_n b_n + \frac{A}{2}\sum_{n=j}^{k} \gamma_n^2.$$

Thus, using the result of part (a) (assuming that $j \ge 1$),

$$\begin{aligned}
\left(\sum_{n=j}^{k} \gamma_n\right) \min_{n=j,\dots,k} b_n &\le a_j + \frac{A}{2}\sum_{n=j}^{k} \gamma_n^2 \\
&\le \frac{2\max\{A,(p-1)a_0\}}{j+2p-2} + 2A\sum_{n=j}^{k} \frac{1}{(n+2p)^2} \\
&\le \frac{2\max\{A,(p-1)a_0\}}{j+2p-2} + 2A\sum_{n=j}^{k} \frac{1}{(n+2p-1)(n+2p)} \\
&= \frac{2\max\{A,(p-1)a_0\}}{j+2p-2} + 2A\sum_{n=j}^{k} \left[\frac{1}{n+2p-1} - \frac{1}{n+2p}\right] \\
&= \frac{2\max\{A,(p-1)a_0\}}{j+2p-2} + 2A\left[\frac{1}{j+2p-1} - \frac{1}{k+2p}\right] \\
&\le \frac{4\max\{A,(p-1)a_0\}}{j+2p-2}. \qquad (13.23)
\end{aligned}$$

On the other hand,

$$\sum_{n=j}^{k} \gamma_n = 2\sum_{n=j}^{k} \frac{1}{n+2p} \ge 2\frac{k-j+1}{k+2p},$$

which, combined with (13.23), yields

$$\min_{n=j,\dots,k} b_n \le \frac{2\max\{A,(p-1)a_0\}(k+2p)}{(j+2p-2)(k-j+1)}.$$

Taking $j = \lfloor k/2 \rfloor + 2$, we conclude that for any $k \ge 3$,

$$\min_{n=\lfloor k/2 \rfloor+2,\dots,k} b_n \le \frac{2\max\{A,(p-1)a_0\}(k+2p)}{(\lfloor k/2 \rfloor+2p)(k-\lfloor k/2 \rfloor-1)}. \qquad (13.24)$$

Now,

$$\begin{aligned}\frac{k+2p}{(\lfloor k/2\rfloor+2p)(k-\lfloor k/2\rfloor-1)} &\le \frac{k+2p}{(k/2+2p-0.5)(k-\lfloor k/2\rfloor-1)}\\ &= 2\frac{k+2p}{k+4p-1}\cdot\frac{1}{k-\lfloor k/2\rfloor-1}\\ &\le \frac{2}{k-\lfloor k/2\rfloor-1}\\ &\le \frac{2}{k/2-1},\end{aligned}$$

which, combined with (13.24), yields

$$\min_{n=\lfloor k/2\rfloor+2,\ldots,k} b_n \le \frac{8\max\{A,(p-1)a_0\}}{k-2}. \qquad \square$$

Equipped with Lemma 13.13, we will now establish a sublinear rate of convergence of the generalized conditional gradient method under the three stepsize strategies described at the beginning of Section 13.2.3: predefined, adaptive, and exact line search.

Theorem 13.14. *Suppose that Assumption* 13.1 *holds and that f is convex. Let $\{\mathbf{x}^k\}_{k\ge 0}$ be the sequence generated by the generalized conditional gradient method for solving problem* (13.2) *with either a predefined stepsize $t_k = \alpha_k \equiv \frac{2}{k+2}$, adaptive stepsize, or exact line search. Let Ω be an upper bound on the diameter of* $\mathrm{dom}(g)$*:*

$$\Omega \ge \max_{\mathbf{x},\mathbf{y}\in\mathrm{dom}(g)} \|\mathbf{x}-\mathbf{y}\|.$$

Then

(a) $F(\mathbf{x}^k) - F_{\mathrm{opt}} \le \frac{2L_f\Omega^2}{k}$ *for any* $k \ge 1$*;*

(b) $\min_{n=\lfloor k/2\rfloor+2,\ldots,k} S(\mathbf{x}^n) \le \frac{8L_f\Omega^2}{k-2}$ *for any* $k \ge 3$.

Proof. By the fundamental inequality (13.6) invoked with $\mathbf{x} = \mathbf{x}^k$ and $t = t_k$, it follows that for any $k \ge 0$,

$$F(\mathbf{x}^k + t_k(\mathbf{p}^k - \mathbf{x}^k)) - F_{\mathrm{opt}} \le F(\mathbf{x}^k) - F_{\mathrm{opt}} - t_k S(\mathbf{x}^k) + \frac{t_k^2 L_f}{2}\|\mathbf{p}^k - \mathbf{x}^k\|^2, \quad (13.25)$$

where $\mathbf{p}^k = \mathbf{p}(\mathbf{x}^k)$. Specifically, if a predefined stepsize is used, meaning that $t_k = \alpha_k \equiv \frac{2}{k+2}$, then

$$F(\mathbf{x}^k + \alpha_k(\mathbf{p}^k - \mathbf{x}^k)) - F_{\mathrm{opt}} \le F(\mathbf{x}^k) - F_{\mathrm{opt}} - \alpha_k S(\mathbf{x}^k) + \frac{\alpha_k^2 L_f}{2}\|\mathbf{p}^k - \mathbf{x}^k\|^2. \quad (13.26)$$

If an exact line search is used, meaning that $t_k = u_k \in \mathrm{argmin}_{t\in[0,1]} F(\mathbf{x}^k + t(\mathbf{p}^k - \mathbf{x}^k))$, then

$$\begin{aligned}F(\mathbf{x}^k + u_k(\mathbf{p}^k - \mathbf{x}^k)) - F_{\mathrm{opt}} &\le F(\mathbf{x}^k + \alpha_k(\mathbf{p}^k - \mathbf{x}^k)) - F_{\mathrm{opt}} \qquad (13.27)\\ &\le F(\mathbf{x}^k) - F_{\mathrm{opt}} - \alpha_k S(\mathbf{x}^k) + \frac{\alpha_k^2 L_f}{2}\|\mathbf{p}^k - \mathbf{x}^k\|^2,\end{aligned}$$

where the first inequality follows by the definition of u_k and the second is the inequality (13.26). Finally, in the adaptive stepsize strategy, $t_k = v_k \equiv \min\left\{1, \frac{S(\mathbf{x}^k)}{L_f\|\mathbf{p}^k - \mathbf{x}^k\|^2}\right\}$. Note that v_k satisfies

$$v_k = \operatorname{argmin}_{t\in[0,1]} \left\{-tS(\mathbf{x}^k) + \frac{t^2 L_f}{2}\|\mathbf{p}^k - \mathbf{x}^k\|^2\right\}. \tag{13.28}$$

Thus,

$$\begin{aligned} F(\mathbf{x}^k + v_k(\mathbf{p}^k - \mathbf{x}^k)) - F_{\text{opt}} &\le F(\mathbf{x}^k) - F_{\text{opt}} - v_k S(\mathbf{x}^k) + \frac{v_k^2 L_f}{2}\|\mathbf{p}^k - \mathbf{x}^k\|^2 \\ &\le F(\mathbf{x}^k) - F_{\text{opt}} - \alpha_k S(\mathbf{x}^k) + \frac{\alpha_k^2 L_f}{2}\|\mathbf{p}^k - \mathbf{x}^k\|^2, \end{aligned}$$

where the first inequality is the inequality (13.25) with $t_k = v_k$ and the second is due to (13.28). Combining the last inequality with (13.26) and (13.27), we conclude that for the three stepsize strategies, the following inequality holds:

$$F(\mathbf{x}^{k+1}) - F_{\text{opt}} \le F(\mathbf{x}^k) - F_{\text{opt}} - \alpha_k S(\mathbf{x}^k) + \frac{\alpha_k^2 L_f}{2}\|\mathbf{p}^k - \mathbf{x}^k\|^2,$$

which, combined with the inequality $\|\mathbf{p}^k - \mathbf{x}^k\| \le \Omega$, implies that

$$F(\mathbf{x}^{k+1}) - F_{\text{opt}} \le F(\mathbf{x}^k) - F_{\text{opt}} - \alpha_k S(\mathbf{x}^k) + \frac{\alpha_k^2 L_f \Omega^2}{2}.$$

Invoking Lemma 13.13 with $a_k = F(\mathbf{x}^k) - F_{\text{opt}}, b_k = S(\mathbf{x}^k)$, $A = L_f\Omega^2$, and $p = 1$ and noting that $a_k \le b_k$ by Lemma 13.12, both parts (a) and (b) follow. ∎

13.3 The Strongly Convex Case

We will focus on the case where the nonsmooth part is an indicator of a compact and convex set C, meaning that $g = \delta_C$, so that problem (13.2) becomes

$$\min\{f(\mathbf{x}) : \mathbf{x} \in C\},$$

and the method under consideration is the conditional gradient method. In Section 10.6 we showed that the proximal gradient method enjoys an improved linear convergence when the smooth part (in the composite model) is strongly convex. Unfortunately, as we will see in Section 13.3.1, in general, the conditional gradient method does not converge in a linear rate even if an additional strong convexity assumption is made on the objective function. Later on, in Section 13.3.2 we will show how, under a strong convexity assumption on the *feasible set* (and not on the objective function), linear rate can be established.

13.3.1 The Negative Result of Canon and Cullum

The arguments go back to Canon and Cullum [37], and we follow them. We begin with some technical lemmas.

Lemma 13.15. *Let $\{a_n\}_{n\geq 0}$ be a sequence of real numbers such that $\sum_{n=0}^{\infty}|a_n|$ diverges. Then for every $\varepsilon>0$, $\sum_{n=k}^{\infty}a_n^2\geq\frac{1}{k^{1+\varepsilon}}$ for infinitely many k's.*

Proof. Suppose by contradiction that there is $\varepsilon>0$ and a positive integer K such that for all $k\geq K$

$$\sum_{n=k}^{\infty}a_n^2<\frac{1}{k^{1+2\varepsilon}}. \tag{13.29}$$

We will show that $\sum_{n=1}^{\infty}|a_n|$ converges. Note that by the Cauchy–Schwarz inequality,

$$\sum_{n=1}^{\infty}|a_n|=\sum_{n=1}^{\infty}|a_n|n^{(1+\varepsilon)/2}n^{-(1+\varepsilon)/2}\leq\sqrt{\sum_{n=1}^{\infty}n^{1+\varepsilon}a_n^2}\sqrt{\sum_{n=1}^{\infty}n^{-(1+\varepsilon)}}. \tag{13.30}$$

Since $\sum_{n=1}^{\infty}n^{-(1+\varepsilon)}$ converges, it is enough to show that $\sum_{n=1}^{\infty}n^{1+\varepsilon}a_n^2$ converges. For that, note that by (13.29), for any $m\geq K$,

$$\sum_{k=K}^{m}\left[k^{\varepsilon}\sum_{n=k}^{m}a_n^2\right]\leq\sum_{k=K}^{m}\left[k^{\varepsilon}\sum_{n=k}^{\infty}a_n^2\right]\leq\sum_{k=K}^{m}\frac{1}{k^{1+\varepsilon}}. \tag{13.31}$$

On the other hand,

$$\sum_{k=K}^{m}\left[k^{\varepsilon}\sum_{n=k}^{m}a_n^2\right]=\sum_{n=K}^{m}\left[a_n^2\sum_{k=K}^{n}k^{\varepsilon}\right],$$

which, combined with the inequality

$$\sum_{k=K}^{n}k^{\varepsilon}\geq\int_K^n x^{\varepsilon}dx=\frac{1}{1+\varepsilon}(n^{1+\varepsilon}-K^{1+\varepsilon})$$

and (13.31), implies that (taking $m\to\infty$)

$$\frac{1}{1+\varepsilon}\sum_{n=K}^{\infty}(n^{1+\varepsilon}-K^{1+\varepsilon})a_n^2\leq\sum_{k=K}^{\infty}\frac{1}{k^{1+\varepsilon}}.$$

Since both $\sum_{k=K}^{\infty}\frac{1}{k^{1+\varepsilon}}$ and $\sum_{n=K}^{\infty}a_n^2$ converge, it follows that $\sum_{n=K}^{\infty}n^{1+\varepsilon}a_n^2$ converges and hence, by (13.30), that $\sum_{n=1}^{\infty}|a_n|$ converges, which is a contradiction to our underlying assumptions. □

We will also use the following well-known lemma.

Lemma 13.16 (see [75, Chapter VII, Theorem 4]). *Let $\{b_n\}_{n\geq 0}$ be a sequence satisfying $0\leq b_n<1$ for any n. Then $\prod_{n=0}^{m}(1-b_n)\to 0$ as $m\to\infty$ if and only if $\sum_{n=0}^{\infty}b_n$ diverges.*

Our main goal will be to describe an example of a minimization problem of a strongly convex function over a nonempty compact convex set for which the

conditional gradient method does not exhibit a linear rate of convergence. For that, let us consider the following quadratic problem over $\mathbb{R}^n$:

$$f_{\text{opt}} \equiv \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ f_q(\mathbf{x}) \equiv \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{b}^T\mathbf{x} : \mathbf{x} \in \Omega \right\}, \tag{13.32}$$

where $\mathbf{Q} \in \mathbb{S}^n_{++}$, $\mathbf{b} \in \mathbb{R}^n$, and $\Omega = \text{conv}\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_l\}$, where $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_l \in \mathbb{R}^n$. We will make the following assumption on problem (13.32).

Assumption 13.17. $\text{int}(\Omega) \neq \emptyset$ *and the optimal solution of problem* (13.32)*, denoted by* $\mathbf{x}^*$*, is on the boundary of* Ω *and is not an extreme point of* Ω*.*

Denoting $\mathbf{A} \in \mathbb{R}^{n\times l}$ as the matrix whose columns are $\mathbf{a}_1, \ldots, \mathbf{a}_l$, we can also write problem (13.32) as

$$\min_{\mathbf{x}\in\mathbb{R}^n, \mathbf{v}\in\mathbb{R}^l} \left\{ \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{b}^T\mathbf{x} : \mathbf{x} = \mathbf{A}\mathbf{v}, \mathbf{v} \in \Delta_l \right\}.$$

The conditional gradient method with exact line search strategy for solving (13.32) reads as follows. Given the kth iterate $\mathbf{x}^k$, the next point $\mathbf{x}^{k+1}$ is computed as follows:

- Choose
$$i_k \in \text{argmin}_{i=1,2,\ldots,l} \langle \mathbf{a}_i, \nabla f_q(\mathbf{x}^k)\rangle.$$
- Define
$$\mathbf{d}^k = \mathbf{a}_{i_k} - \mathbf{x}^k. \tag{13.33}$$
If $\langle \mathbf{d}^k, \nabla f_q(\mathbf{x}^k)\rangle \geq 0$, then $\mathbf{x}^k$ is the optimal solution of problem (13.32). Otherwise, set
$$\mathbf{x}^{k+1} = \mathbf{x}^k + t_k\mathbf{d}^k,$$
where
$$t_k = \text{argmin}_{t\in[0,1]} f_q(\mathbf{x}^k + t\mathbf{d}^k) = \min\{\lambda_k, 1\},$$
with λ_k defined as
$$\lambda_k = -\frac{\langle \mathbf{d}^k, \nabla f_q(\mathbf{x}^k)\rangle}{(\mathbf{d}^k)^T\mathbf{Q}\mathbf{d}^k}. \tag{13.34}$$

We will make the following assumption on the starting point of the conditional gradient method.

Assumption 13.18. $f_q(\mathbf{x}^0) < \min_{i=1,2,\ldots,l} f_q(\mathbf{a}_i)$ *and* $\mathbf{x}^0 = \mathbf{A}\mathbf{v}^0 \in \Omega$*, where* $\mathbf{v}^0 \in \Delta_l \cap \mathbb{R}^n_{++}$*. In particular,* $\mathbf{x}^0 \in \text{int}(\Omega)$*.*

A vector $\mathbf{x}^0$ satisfying Assumption 13.18 can be easily obtained by the following procedure.

- Pick $p \in \text{argmin}_{i=1,2,\ldots,l} f_q(\mathbf{a}_i)$.
- Employ one step of the conditional gradient method starting from $\mathbf{a}_p$ and obtain a point $\tilde{\mathbf{x}}^0 \in \Omega$ for which $f_q(\tilde{\mathbf{x}}^0) < f_q(\mathbf{a}^p)$ (the latter is satisfied since $\mathbf{a}^p$ is not an optimal solution—see Theorem 13.9(a)).

- Find $\tilde{\mathbf{v}}^0 \in \Delta_l$ for which $\tilde{\mathbf{x}}^0 = \mathbf{A}\tilde{\mathbf{v}}^0$.
- If $\tilde{\mathbf{v}}^0 \in \mathbb{R}^l_{++}$, define $\mathbf{v}^0 = \tilde{\mathbf{v}}^0$ and $\mathbf{x}^0 = \tilde{\mathbf{x}}^0$. If $\tilde{\mathbf{v}}^0 \notin \mathbb{R}^l_{++}$, then take a point $\mathbf{v}^0 \in \Delta_l \cap \mathbb{R}^l_{++}$ close enough to $\tilde{\mathbf{v}}^0$ such that $\mathbf{x}^0 \equiv \mathbf{A}\mathbf{v}^0$ will satisfy $f_q(\mathbf{x}^0) < f_q(\mathbf{a}^p)$.

The following lemma gathers several technical results that will be key to establishing the slow rate of the conditional gradient method.

Lemma 13.19. *Suppose that Assumption* 13.17 *holds and that* $\{\mathbf{x}^k\}$ *is the sequence generated by the conditional gradient method with exact line search employed on problem* (13.32) *with a starting point* $\mathbf{x}^0$ *satisfying Assumption* 13.18. *Let* $\mathbf{d}^k$ *and* λ_k *be given by* (13.33) *and* (13.34)*, respectively. Then*

(a) $\mathbf{x}^k \in \operatorname{int}(\Omega)$ *and* $t_k = \lambda_k < 1$ *for any* $k \geq 0$*;*

(b) $f_q(\mathbf{x}^{k+1}) = f_q(\mathbf{x}^k) - \frac{1}{2}((\mathbf{d}^k)^T\mathbf{Q}\mathbf{d}^k)\lambda_k^2$ *for any* $k \geq 0$*;*

(c) $\sum_{k=0}^{\infty} \lambda_k = \infty$*;*

(d) *there exists* $\beta > 0$ *such that* $(\mathbf{d}^k)^T\mathbf{Q}\mathbf{d}^k \geq \beta$ *for all* $k \geq 0$.

Proof. (a) The stepsizes must satisfy $t_k = \lambda_k < 1$, since otherwise, if $t_k = 1$ for some k, then this means that $\mathbf{x}^{k+1} = \mathbf{a}_{i_k}$. But $f_q(\mathbf{x}^{k+1}) = f_q(\mathbf{a}_{i_k}) > f_q(\mathbf{x}^0)$, which is a contradiction to the monotonicity of the sequence of function values generated by the conditional gradient method (Theorem 13.9(a)). The proof that $\mathbf{x}^k \in \operatorname{int}(\Omega)$ is by induction on k. For $k = 0$, by Assumption 13.18, $\mathbf{x}^0 \in \operatorname{int}(\Omega)$. Now suppose that $\mathbf{x}^k \in \operatorname{int}(\Omega)$. To prove that the same holds for $k+1$, note that since $t_k < 1$, it follows by the line segment principle (Lemma 5.23) that $\mathbf{x}^{k+1} = (1 - t_k)\mathbf{x}^k + t_k\mathbf{a}_{i_k}$ is also in $\operatorname{int}(\Omega)$.

(b) Since $t_k = \lambda_k$, it follows that

$$\begin{aligned} f_q(\mathbf{x}^{k+1}) &= f_q\left(\mathbf{x}^k + \lambda_k\mathbf{d}^k\right) \\ &= \frac{1}{2}(\mathbf{x}^k + \lambda_k\mathbf{d}^k)^T\mathbf{Q}(\mathbf{x}^k + \lambda_k\mathbf{d}^k) + \mathbf{b}^T(\mathbf{x}^k + \lambda_k\mathbf{d}^k) \\ &= f_q(\mathbf{x}^k) + \lambda_k(\mathbf{d}^k)^T(\mathbf{Q}\mathbf{x}^k + \mathbf{b}) + \frac{\lambda_k^2}{2}(\mathbf{d}^k)^T\mathbf{Q}\mathbf{d}^k \\ &= f_q(\mathbf{x}^k) + ((\mathbf{d}^k)^T\mathbf{Q}\mathbf{d}^k)\left(-\lambda_k^2 + \frac{\lambda_k^2}{2}\right) \\ &= f_q(\mathbf{x}^k) - \frac{1}{2}((\mathbf{d}^k)^T\mathbf{Q}\mathbf{d}^k)\lambda_k^2. \end{aligned}$$

(c) Suppose by contradiction that $\sum_{k=0}^{\infty} \lambda_k < \infty$, then by Lemma 13.16, it follows that $\prod_{k=0}^{\infty}(1 - \lambda_k) = \delta$ for some $\delta > 0$. Note that by the definition of the method, for any $k \geq 0$, $\mathbf{x}^k = \mathbf{A}\mathbf{v}^k$, where $\{\mathbf{v}^k\}_{k\geq 0}$ satisfies

$$\mathbf{v}^{k+1} = (1 - \lambda_k)\mathbf{v}^k + \lambda_k\mathbf{e}_{i_k}.$$

Hence,

$$\mathbf{v}^{k+1} \geq (1 - \lambda_k)\mathbf{v}^k,$$

implying that

$$\mathbf{v}^k \geq \delta \mathbf{v}^0. \tag{13.35}$$

By Theorem 13.9(d), the limit points of $\{\mathbf{x}^k\}_{k\geq 0}$ are stationary points of problem (13.32). Let $\mathbf{x}^*$ be the unique optimal solution of problem (13.32). Since $\mathbf{x}^*$ is the only stationary point of problem (13.32), we can conclude that $\mathbf{x}^k \to \mathbf{x}^*$. The sequence $\{\mathbf{v}^k\}_{k\geq 0}$ is bounded and hence has a convergent subsequence $\{\mathbf{v}^{k_j}\}_{j\geq 0}$. Denoting the limit of the subsequence by $\mathbf{v}^* \in \Delta_l$, we note that by (13.35) it follows that $\mathbf{v}^* \geq \delta \mathbf{v}^0$, and hence $\mathbf{v}^* \in \Delta_l \cap \mathbb{R}^l_{++}$. Taking j to ∞ in the identity $\mathbf{x}^{k_j} = \mathbf{A}\mathbf{v}^{k_j}$, we obtain that $\mathbf{x}^* = \mathbf{A}\mathbf{v}^*$, where $\mathbf{v}^* \in \Delta_l \cap \mathbb{R}^l_{++}$, implying that the $\mathbf{x}^* \in \text{int}(\Omega)$, in contradiction to Assumption 13.17.

(d) Since

$$(\mathbf{d}^k)^T \mathbf{Q}\mathbf{d}^k \geq \gamma \|\mathbf{d}^k\|_2^2 \tag{13.36}$$

with $\gamma = \lambda_{\min}(\mathbf{Q}) > 0$, it follows that we need to show that $\|\mathbf{d}^k\|_2$ is bounded below by a positive number. Note that by Assumption 13.17, $\mathbf{x}^* \notin \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_l\}$, and therefore there exists a positive integer K and $\beta_1 > 0$ such that $\|\mathbf{a}_i - \mathbf{x}^k\| \geq \beta_1$ for all $k > K$ and $i \in \{1, 2, \ldots, l\}$. Since $\mathbf{x}^k \in \text{int}(\Omega)$ for all k, it follows that for β_2 defined as

$$\beta_2 \equiv \min\{\beta_1, \|\mathbf{a}_{i_0} - \mathbf{x}^0\|_2, \|\mathbf{a}_{i_1} - \mathbf{x}^1\|_2, \ldots, \|\mathbf{a}_{i_K} - \mathbf{x}^K\|_2\} > 0,$$

it holds that $\|\mathbf{d}^k\|_2 = \|\mathbf{a}_{i_k} - \mathbf{x}^k\| \geq \beta_2$ for all $k \geq 0$, and we can finally conclude by (13.36) that for $\beta = \gamma\beta_2^2$, $(\mathbf{d}^k)^T \mathbf{Q}\mathbf{d}^k \geq \beta$ for all $k \geq 0$. ∎

The main negative result showing that the rate of convergence of the method cannot be linear is stated in Theorem 13.20 below.

Theorem 13.20 (Canon and Cullum's negative result). *Suppose that Assumption* 13.17 *holds and that* $\{\mathbf{x}^k\}$ *is the sequence generated by the conditional gradient method with exact line search for solving problem* (13.32) *with a starting point* $\mathbf{x}^0$ *satisfying Assumption* 13.18. *Then for every* $\varepsilon > 0$ *we have that* $f_q(\mathbf{x}^k) - f_{\text{opt}} \geq \frac{1}{k^{1+\varepsilon}}$ *for infinitely many* k*'s.*

Proof. Let $\mathbf{d}^k$ and λ_k be given by (13.33) and (13.34), respectively. By Lemma 13.19(b), we have for any two positive integers satisfying $K \geq k$,

$$f_q(\mathbf{x}^K) - f_{\text{opt}} = f_q(\mathbf{x}^k) - f_{\text{opt}} - \frac{1}{2}\sum_{n=k}^{K-1}((\mathbf{d}^n)^T \mathbf{Q}\mathbf{d}^n)\lambda_n^2.$$

Taking $K \to \infty$ and using the fact that $f_q(\mathbf{x}^K) \to f_{\text{opt}}$ and Lemma 13.19(d), we obtain that

$$f_q(\mathbf{x}^k) - f_{\text{opt}} = \frac{1}{2}\sum_{n=k}^{\infty}((\mathbf{d}^n)\mathbf{Q}(\mathbf{d}^n))\lambda_n^2 \geq \frac{\beta}{2}\sum_{n=k}^{\infty}\lambda_n^2. \tag{13.37}$$

By Lemma 13.19(c), $\sum_{k=0}^{\infty}\lambda_k = \infty$, and hence by Lemma 13.15 and (13.37), we conclude that $f_q(\mathbf{x}^k) - f_{\text{opt}} \geq \frac{1}{k^{1+\varepsilon}}$ for infinitely many k's. ∎

Example 13.21. Consider the problem

$$\min\{f_q(x_1, x_2) \equiv x_1^2 + x_2^2 : (x_1, x_2) \in \operatorname{conv}\{(-1, 0), (1, 0), (0, 1)\}\}. \tag{13.38}$$

Assumption 13.17 is satisfied since the feasible set of problem (13.38) has a nonempty interior and the optimal solution, $(x_1^*, x_2^*) = (0, 0)$, is on the boundary of the feasible set but is not an extreme point. The starting point $\mathbf{x}^0 = (0, \frac{1}{2})$ satisfies Assumption 13.18 since

$$f_q(\mathbf{x}^0) = \frac{1}{4} < 1 = \min\{f_q(-1, 0), f_q(1, 0), f_q(0, 1)\}$$

and $\mathbf{x}^0 = \frac{1}{4}(-1, 0) + \frac{1}{4}(1, 0) + \frac{1}{2}(0, 1)$. The first 100 iterations produced by the conditional gradient method are plotted in Figure 13.1. ∎

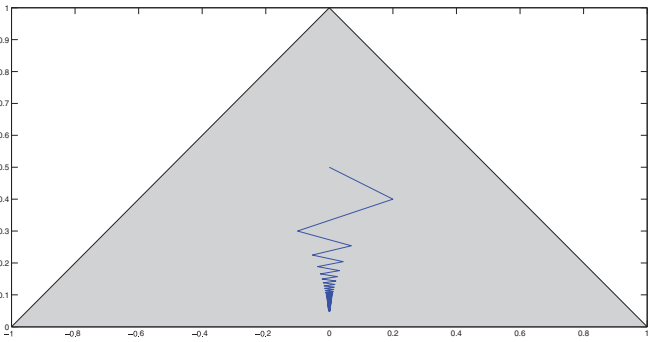

Figure 13.1. *First* 100 *iterations of the conditional gradient method employed on the problem from Example* 13.21.

13.3.2 Linear Rate under Strong Convexity of the Feasible Set

Canon and Cullum's negative result shows that different assumptions than strong convexity of the objective are required in order to establish a linear rate of convergence of the conditional gradient method. One example of such an assumption is strong convexity of the feasible set.

Definition 13.22 (strongly convex set). *A nonempty set $C \subseteq \mathbb{E}$ is called* **σ-strongly convex** *($\sigma > 0$) if for any $\mathbf{x}, \mathbf{y} \in C$ and $\lambda \in [0, 1]$ the inclusion*

$$B\left[\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}, \frac{\sigma}{2}\lambda(1 - \lambda)\|\mathbf{x} - \mathbf{y}\|^2\right] \subseteq C$$

holds.

A set is called strongly convex if it is σ-strongly convex for some $\sigma > 0$. Obviously, any strongly convex set is also convex. The next result states that level sets of nonnegative strongly convex and smooth functions are strongly convex sets.

Theorem 13.23 (strong convexity of level sets of strongly convex and smooth functions).[74] *Suppose that $g : \mathbb{E} \to \mathbb{R}_+$ is nonnegative, L_g-smooth, and σ_g-strongly convex. Let $\alpha > 0$. Then the set*

$$C_\alpha = \{\mathbf{x} \in \mathbb{E} : g(\mathbf{x}) \le \alpha\}$$

is $\frac{\sigma_g}{\sqrt{2\alpha L_g}}$-strongly convex.

Proof. Let $\mathbf{x}, \mathbf{y} \in C_\alpha$ and $\lambda \in [0,1]$. Define $\mathbf{x}_\lambda = \lambda\mathbf{x} + (1-\lambda)\mathbf{y}$. By the nonnegativity of g and the sufficient decrease lemma (Lemma 10.4), we have

$$g(\mathbf{x}_\lambda) \ge g(\mathbf{x}_\lambda) - g\left(\mathbf{x}_\lambda - \frac{1}{L_g}\nabla g(\mathbf{x}_\lambda)\right) \ge \frac{1}{2L_g}\|\nabla g(\mathbf{x}_\lambda)\|^2.$$

Thus,

$$\|\nabla g(\mathbf{x}_\lambda)\| \le \sqrt{2L_g g(\mathbf{x}_\lambda)}. \tag{13.39}$$

By the σ_g-strong convexity of g and the inequalities $g(\mathbf{x}), g(\mathbf{y}) \le \alpha$,

$$g(\mathbf{x}_\lambda) \le \lambda g(\mathbf{x}) + (1-\lambda)g(\mathbf{y}) - \frac{\sigma_g}{2}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2 \le \alpha - \beta, \tag{13.40}$$

where $\beta \equiv \frac{\sigma_g}{2}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2$.

Denote $\tilde{\sigma} = \frac{\sigma_g}{\sqrt{2\alpha L_g}}$. In order to show that C_α is $\tilde{\sigma}$-strongly convex, we will take $\mathbf{u} \in B[\mathbf{0}, 1]$ and show that $\mathbf{x}_\lambda + \gamma\mathbf{u} \in C_\alpha$, where $\gamma = \frac{\tilde{\sigma}}{2}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2$. Indeed,

$$\begin{aligned}
g(\mathbf{x}_\lambda + \gamma\mathbf{u}) &\le g(\mathbf{x}_\lambda) + \gamma\langle\nabla g(\mathbf{x}_\lambda), \mathbf{u}\rangle + \frac{\gamma^2 L_g}{2}\|\mathbf{u}\|^2 && \text{[descent lemma]} \\
&\le g(\mathbf{x}_\lambda) + \gamma\|\nabla g(\mathbf{x}_\lambda)\| \cdot \|\mathbf{u}\| + \frac{\gamma^2 L_g}{2}\|\mathbf{u}\|^2 && \text{[Cauchy–Schwarz]} \\
&\le g(\mathbf{x}_\lambda) + \gamma\sqrt{2L_g g(\mathbf{x}_\lambda)}\|\mathbf{u}\| + \frac{\gamma^2 L_g}{2}\|\mathbf{u}\|^2 && \text{[(13.39)]} \\
&= \left(\sqrt{g(\mathbf{x}_\lambda)} + \gamma\sqrt{\frac{L_g}{2}}\|\mathbf{u}\|\right)^2,
\end{aligned}$$

which, combined with (13.40) and the fact that $\|\mathbf{u}\| \le 1$, implies that

$$g(\mathbf{x}_\lambda + \gamma\mathbf{u}) \le \left(\sqrt{\alpha - \beta} + \gamma\sqrt{\frac{L_g}{2}}\right)^2. \tag{13.41}$$

[74] Theorem 13.23 is from Journée, Nesterov, Richtárik, and Sepulchre, [74, Theorem 12].

By the concavity of the square root function $\varphi(t) = \sqrt{t}$, we have

$$\begin{aligned}\sqrt{\alpha - \beta} &= \varphi(\alpha - \beta) \le \varphi(\alpha) - \varphi'(\alpha)\beta = \sqrt{\alpha} - \frac{\beta}{2\sqrt{\alpha}} \\ &= \sqrt{\alpha} - \frac{\sigma_g \lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2}{4\sqrt{\alpha}} \\ &= \sqrt{\alpha} - \frac{\sqrt{2\alpha L_g}\tilde{\sigma}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2}{4\sqrt{\alpha}} \\ &= \sqrt{\alpha} - \sqrt{\frac{L_g}{2}}\frac{\tilde{\sigma}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2}{2} \\ &= \sqrt{\alpha} - \gamma\sqrt{\frac{L_g}{2}},\end{aligned}$$

which, along with (13.41), leads to the inequality $g(\mathbf{x}_\lambda + \gamma\mathbf{u}) \le \alpha$. □

Example 13.24 (strong convexity of Euclidean balls). Consider the set[75] $C = B[\mathbf{c}, r] \subseteq \mathbb{E}$, where $\mathbf{c} \in \mathbb{E}$ and $r > 0$. Note that $C = \mathrm{Lev}(g, r^2)$, where $g(\mathbf{x}) = \|\mathbf{x} - \mathbf{c}\|^2$. Since here $L_g = \sigma_g = 2, \alpha = r^2$, it follows that the strong convexity parameter of the set is $\frac{2}{\sqrt{2\cdot 2\cdot r^2}} = \frac{1}{r}$. ■

We will consider the problem

$$\min_{\mathbf{x}\in C} f(\mathbf{x}), \tag{13.42}$$

where we assume the following set of properties.

Assumption 13.25.

(A) *C is nonempty, compact, and σ-strongly convex.*

(B) *$f : \mathbb{E} \to (-\infty, \infty]$ is convex L_f-smooth over $\mathrm{dom}(f)$, which is assumed to be an open and convex set satisfying $C \subseteq \mathrm{dom}(f)$.*

(C) *There exists $\delta > 0$ such that $\|\nabla f(\mathbf{x})\| \ge \delta$ for any $\mathbf{x} \in C$.*

(D) *The optimal set of problem (13.42) is nonempty and denoted by X^*. The optimal value of the problem is denoted by f_{opt}.*

As usual, for any $\mathbf{x} \in C$, we use the notation

$$\mathbf{p}(\mathbf{x}) \in \mathrm{argmin}_{\mathbf{p}\in C}\langle \nabla f(\mathbf{x}), \mathbf{p}\rangle,\ S(\mathbf{x}) = \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{p}(\mathbf{x})\rangle.$$

We begin by establishing the following result connecting $S(\mathbf{x})$ and the distance between $\mathbf{x}$ and $\mathbf{p}(\mathbf{x})$.

[75] Recall that in this chapter the underlying norm is assumed to be Euclidean.

Lemma 13.26. *Suppose that Assumption* 13.25 *holds. Then for any* $\mathbf{x} \in C$,

$$S(\mathbf{x}) \geq \frac{\sigma\delta}{4}\|\mathbf{x} - \mathbf{p}(\mathbf{x})\|^2. \tag{13.43}$$

Proof. Let $\mathbf{x} \in C$. Define

$$\mathbf{z} = \frac{\mathbf{x} + \mathbf{p}(\mathbf{x})}{2} - \frac{\sigma}{8}\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}\|\mathbf{x} - \mathbf{p}(\mathbf{x})\|^2.$$

Then obviously $\mathbf{z} \in B\left[\frac{\mathbf{x}+\mathbf{p}(\mathbf{x})}{2}, \frac{\sigma}{8}\|\mathbf{x} - \mathbf{p}(\mathbf{x})\|^2\right]$, and hence, by the σ-strong convexity of C, $\mathbf{z} \in C$. In particular,

$$\langle \nabla f(\mathbf{x}), \mathbf{z}\rangle \geq \langle \nabla f(\mathbf{x}), \mathbf{p}(\mathbf{x})\rangle. \tag{13.44}$$

The result (13.43) follows by the following arguments:

$$\begin{aligned}
\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{p}(\mathbf{x})\rangle &= 2\left\langle \nabla f(\mathbf{x}), \frac{\mathbf{x} + \mathbf{p}(\mathbf{x})}{2} - \mathbf{p}(\mathbf{x})\right\rangle \\
&= 2\langle \nabla f(\mathbf{x}), \mathbf{z} - \mathbf{p}(\mathbf{x})\rangle + 2\left\langle \nabla f(\mathbf{x}), \frac{\sigma}{8}\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}\|\mathbf{x} - \mathbf{p}(\mathbf{x})\|^2\right\rangle \\
&\overset{(13.44)}{\geq} 2\left\langle \nabla f(\mathbf{x}), \frac{\sigma}{8}\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}\|\mathbf{x} - \mathbf{p}(\mathbf{x})\|^2\right\rangle \\
&= \frac{\sigma}{4}\|\nabla f(\mathbf{x})\| \cdot \|\mathbf{x} - \mathbf{p}(\mathbf{x})\|^2 \\
&\geq \frac{\sigma\delta}{4}\|\mathbf{x} - \mathbf{p}(\mathbf{x})\|^2. \quad \square
\end{aligned}$$

We will now establish the main result of this section stating that under Assumption 13.25, the conditional gradient method with either an adaptive or exact line search stepsize strategies enjoys a linear rate of convergence in function values.

Theorem 13.27. *Suppose that Assumption* 13.25 *holds, and let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by the conditional gradient method for solving problem* (13.42) *with stepsizes chosen by either the adaptive or exact line search strategies. Then for any* $k \geq 0$,

(a) $f(\mathbf{x}^{k+1}) - f_{\text{opt}} \leq (1 - \lambda)(f(\mathbf{x}^k) - f_{\text{opt}})$, *where*

$$\lambda = \min\left\{\frac{\sigma\delta}{8L_f}, \frac{1}{2}\right\}; \tag{13.45}$$

(b) $f(\mathbf{x}^k) - f_{\text{opt}} \leq (1 - \lambda)^k(f(\mathbf{x}^0) - f_{\text{opt}})$.

Proof. Let $k \geq 0$ and let $\tilde{\mathbf{x}}^k = \mathbf{x}^k + s_k(\mathbf{p}^k - \mathbf{x}^k)$, where $\mathbf{p}^k = \mathbf{p}(\mathbf{x}^k)$ and s_k is the stepsize chosen by the adaptive strategy:

$$s_k = \min\left\{1, \frac{S(\mathbf{x}^k)}{L_f\|\mathbf{x}^k - \mathbf{p}^k\|^2}\right\}.$$

By Lemma 13.7 (invoked with $\mathbf{x} = \mathbf{x}^k$ and $t = s_k$),

$$f(\mathbf{x}^k) - f(\tilde{\mathbf{x}}^k) \geq s_k S(\mathbf{x}^k) - \frac{s_k^2 L_f}{2}\|\mathbf{p}^k - \mathbf{x}^k\|^2. \tag{13.46}$$

There are two options: Either $s_k = 1$, and in this case $S(\mathbf{x}^k) \geq L_f\|\mathbf{x}^k - \mathbf{p}^k\|^2$, and thus

$$f(\mathbf{x}^k) - f(\tilde{\mathbf{x}}^k) \geq S(\mathbf{x}^k) - \frac{L_f}{2}\|\mathbf{p}^k - \mathbf{x}^k\|^2 \geq \frac{1}{2}S(\mathbf{x}^k), \tag{13.47}$$

or, on the other hand, $s_k = \frac{S(\mathbf{x}^k)}{L_f\|\mathbf{x}^k - \mathbf{p}^k\|^2}$, and then (13.46) amounts to

$$f(\mathbf{x}^k) - f(\tilde{\mathbf{x}}^k) \geq \frac{S^2(\mathbf{x}^k)}{2L_f\|\mathbf{x}^k - \mathbf{p}^k\|^2},$$

which, combined with (13.43) (with $\mathbf{x} = \mathbf{x}^k$), implies the inequality

$$f(\mathbf{x}^k) - f(\tilde{\mathbf{x}}^k) \geq \frac{\sigma\delta}{8L_f}S(\mathbf{x}^k). \tag{13.48}$$

Combining the inequalities (13.47) and (13.48) arising from the two possible cases, we obtain that

$$f(\mathbf{x}^k) - f(\tilde{\mathbf{x}}^k) \geq \lambda S(\mathbf{x}^k),$$

where λ is given in (13.45). If the method is employed with an adaptive stepsize strategy, then $\tilde{\mathbf{x}}^k = \mathbf{x}^{k+1}$, and hence $f(\tilde{\mathbf{x}}^k) = f(\mathbf{x}^{k+1})$. If the method is employed with an exact line search strategy, then $f(\mathbf{x}^{k+1}) \leq f(\tilde{\mathbf{x}}^k)$. Therefore, in both stepsize regimes, we get

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq f(\mathbf{x}^k) - f(\tilde{\mathbf{x}}^k) \geq \lambda S(\mathbf{x}^k). \tag{13.49}$$

On the other hand, by Lemma 13.12,

$$f(\mathbf{x}^k) - f_{\text{opt}} \leq S(\mathbf{x}^k). \tag{13.50}$$

Combining (13.49) and (13.50), we obtain that

$$\lambda(f(\mathbf{x}^k) - f_{\text{opt}}) \leq (f(\mathbf{x}^k) - f_{\text{opt}}) - (f(\mathbf{x}^{k+1}) - f_{\text{opt}}),$$

from which it readily follows that

$$f(\mathbf{x}^{k+1}) - f_{\text{opt}} \leq (1 - \lambda)(f(\mathbf{x}^k) - f_{\text{opt}}).$$

Part (b) is an immediate consequence of (a). □

13.4 The Randomized Generalized Block Conditional Gradient Method[76]

In this section we will consider a block version of the generalized conditional gradient method. The model and underlying assumptions are similar to those made w.r.t.

[76] The randomized generalized block conditional gradient method presented in Section 13.4 is a simple generalization of the randomized block conditional gradient method introduced and analyzed by Lacoste-Julien, Jaggi, Schmidt, and Pletscher in [76].

the block proximal gradient method in Section 11.2. We will consider the problem

$$\min_{\mathbf{x}_1\in\mathbb{E}_1,\mathbf{x}_2\in\mathbb{E}_2,\ldots,\mathbf{x}_p\in\mathbb{E}_p}\left\{F(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_p)\equiv f(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_p)+\sum_{j=1}^{p}g_j(\mathbf{x}_j)\right\},\tag{13.51}$$

where $\mathbb{E}_1,\mathbb{E}_2,\ldots,\mathbb{E}_p$ are Euclidean spaces. We will denote the product space by $\mathbb{E}=\mathbb{E}_1\times\mathbb{E}_2\times\cdots\times\mathbb{E}_p$ and use our convention (see Section 1.9) that the product space is also Euclidean with endowed norm

$$\|(\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_p)\|_{\mathbb{E}}=\sqrt{\sum_{i=1}^{p}\|\mathbf{u}_i\|_{\mathbb{E}_i}^2}.$$

We will omit the subscripts of the norms indicating the underlying vector space (whose identity will be clear from the context). The function $g:\mathbb{E}\to(-\infty,\infty]$ is defined by

$$g(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_p)\equiv\sum_{i=1}^{p}g_i(\mathbf{x}_i),$$

and in particular $\mathrm{dom}(g)=\mathrm{dom}(g_1)\times\mathrm{dom}(g_2)\times\cdots\times\mathrm{dom}(g_p)$. The gradient w.r.t. the ith block ($i\in\{1,2,\ldots,p\}$) is denoted by $\nabla_i f$ and is actually a mapping from $\mathrm{dom}(f)$ to $\mathbb{E}_i$. The following is satisfied:

$$\nabla f(\mathbf{x})=(\nabla_1 f(\mathbf{x}),\nabla_2 f(\mathbf{x}),\ldots,\nabla_p f(\mathbf{x})).$$

For any $i\in\{1,2,\ldots,p\}$ we define $\mathcal{U}_i:\mathbb{E}_i\to\mathbb{E}$ to be the linear transformation given by

$$\mathcal{U}_i(\mathbf{d})=(\mathbf{0},\ldots,\mathbf{0},\underbrace{\mathbf{d}}_{i\text{th block}},\mathbf{0},\ldots,\mathbf{0}),\quad \mathbf{d}\in\mathbb{E}_i.$$

We also use throughout this chapter the notation that a vector $\mathbf{x}\in\mathbb{E}$ can be written as

$$\mathbf{x}=(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_p),$$

and this relation will also be written as $\mathbf{x}=(\mathbf{x}_i)_{i=1}^p$. Thus, in our notation, the main model (13.51) can be simply written as

$$\min_{\mathbf{x}\in\mathbb{E}}\{F(\mathbf{x})=f(\mathbf{x})+g(\mathbf{x})\}.$$

The basic assumptions on the model are summarized below.

Assumption 13.28.

(A) $g_i:\mathbb{E}_i\to(-\infty,\infty]$ *is proper closed and convex with compact* $\mathrm{dom}(g_i)$ *for any* $i\in\{1,2,\ldots,p\}$.

(B) $f:\mathbb{E}\to(-\infty,\infty]$ *is convex and differentiable over* $\mathrm{dom}(f)$*, which is assumed to be an open and convex set satisfying* $\mathrm{dom}(g)\subseteq\mathrm{dom}(f)$.

(C) *There exist* $L_1, L_2, \ldots, L_p > 0$ *such that for any* $i \in \{1, 2, \ldots, p\}$ *it holds that*

$$\|\nabla_i f(\mathbf{x}) - \nabla_i f(\mathbf{x} + \mathcal{U}_i(\mathbf{d}))\| \leq L_i \|\mathbf{d}\|$$

for all $\mathbf{x} \in \mathrm{dom}(f)$ *and* $\mathbf{d} \in \mathbb{E}_i$ *for which* $\mathbf{x} + \mathcal{U}_i(\mathbf{d}) \in \mathrm{dom}(f)$.

(D) *The optimal set of problem* (13.51) *is nonempty and denoted by* X^*. *The optimal value is denoted by* F_{opt}.

For any $i \in \{1, 2, \ldots, p\}$, we denote

$$\mathbf{p}_i(\mathbf{x}) \in \mathrm{argmin}_{\mathbf{v} \in \mathbb{E}_i} \{\langle \mathbf{v}, \nabla_i f(\mathbf{x}) \rangle + g_i(\mathbf{v})\} \tag{13.52}$$

and define the *ith partial conditional gradient norm* as

$$S_i(\mathbf{x}) = \max_{\mathbf{v} \in \mathbb{E}_i} \{\langle \nabla_i f(\mathbf{x}), \mathbf{x}_i - \mathbf{v} \rangle + g_i(\mathbf{x}_i) - g_i(\mathbf{v})\} = \langle \nabla_i f(\mathbf{x}), \mathbf{x}_i - \mathbf{p}_i(\mathbf{x}) \rangle + g_i(\mathbf{x}_i) - g_i(\mathbf{p}_i(\mathbf{x})).$$

Obviously, we have

$$S(\mathbf{x}) = \sum_{i=1}^{p} S_i(\mathbf{x}).$$

There might be multiple optimal solutions for problem (13.52) and also for problem (13.3) defining $\mathbf{p}(\mathbf{x})$. Our only assumption is that $\mathbf{p}(\mathbf{x})$ is chosen as

$$\mathbf{p}(\mathbf{x}) = (\mathbf{p}_1(\mathbf{x}), \mathbf{p}_2(\mathbf{x}), \ldots, \mathbf{p}_p(\mathbf{x})). \tag{13.53}$$

The latter is not a restricting assumption since the vector in the right-hand side of (13.53) is indeed a minimizer of problem (13.3). The randomized generalized block conditional gradient method is described below.

The Randomized Generalized Block Conditional Gradient (RGBCG) Method

Initialization: pick $\mathbf{x}^0 = (\mathbf{x}_1^0, \mathbf{x}_2^0, \ldots, \mathbf{x}_p^0) \in \mathrm{dom}(g)$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick $i_k \in \{1, 2, \ldots, p\}$ randomly via a uniform distribution and $t_k \in [0, 1]$;

(b) set $\mathbf{x}^{k+1} = \mathbf{x}^k + t_k \mathcal{U}_{i_k}(\mathbf{p}_{i_k}(\mathbf{x}^k) - \mathbf{x}_{i_k}^k)$.

In our analysis the following notation is used:

- $\xi_{k-1} \equiv \{i_0, i_1, \ldots, i_{k-1}\}$ is a multivariate random variable.
- We will consider, in addition to the underlying Euclidean norm of the space $\mathbb{E}$, the following weighted norm:

$$\|\mathbf{x}\|_L \equiv \sqrt{\sum_{i=1}^{p} L_i \|\mathbf{x}_i\|^2}.$$

The rate of convergence of the RGBCG method with a specific choice of diminishing stepsizes is established in the following result.

Theorem 13.29. *Suppose that Assumption* 13.28 *holds, and let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by the RGBCG method for solving problem* (13.51) *with stepsizes* $t_k = \frac{2p}{k+2p}$. *Let* Ω *satisfy*

$$\Omega \geq \max_{\mathbf{x},\mathbf{y}\in \mathrm{dom}(g)} \|\mathbf{x}-\mathbf{y}\|_L. \tag{13.54}$$

Then

(a) *for any* $k \geq 1$,

$$\mathsf{E}_{\xi_{k-1}}(F(\mathbf{x}^k)) - F_{\mathrm{opt}} \leq \frac{2\max\{(p-1)(F(\mathbf{x}^0)-F_{\mathrm{opt}}), p\Omega^2\}}{k+2p-2}; \tag{13.55}$$

(b) *for any* $k \geq 3$,

$$\min_{n=\lfloor k/2\rfloor+2,\ldots,k} \mathsf{E}_{\xi_{n-1}}(S(\mathbf{x}^n)) \leq \frac{8\max\{(p-1)(F(\mathbf{x}^0)-F_{\mathrm{opt}}), p\Omega^2\}}{k-2}. \tag{13.56}$$

Proof. We will use the shorthand notation $\mathbf{p}^k = \mathbf{p}(\mathbf{x}^k)$, and by the relation (13.53) it follows that $\mathbf{p}_i^k = \mathbf{p}_i(\mathbf{x}^k)$. Using the block descent lemma (Lemma 11.8) and the convexity of g_{i_k}, we can write the following:

$$\begin{aligned}
F(\mathbf{x}^{k+1}) &= f(\mathbf{x}^{k+1}) + g(\mathbf{x}^{k+1}) \\
&= f(\mathbf{x}^k + t_k\mathcal{U}_{i_k}(\mathbf{p}_{i_k}^k - \mathbf{x}_{i_k}^k)) + g(\mathbf{x}^k + t_k\mathcal{U}_{i_k}(\mathbf{p}_{i_k}^k - \mathbf{x}_{i_k}^k)) \\
&\leq f(\mathbf{x}^k) - t_k\langle \nabla_{i_k} f(\mathbf{x}^k), \mathbf{x}_{i_k}^k - \mathbf{p}_{i_k}^k\rangle + \frac{t_k^2 L_{i_k}}{2}\|\mathbf{p}_{i_k}^k - \mathbf{x}_{i_k}^k\|^2 + \sum_{j=1, j\neq i_k}^{p} g_j(\mathbf{x}^k) \\
&\quad + g_{i_k}((1-t_k)\mathbf{x}_{i_k}^k + t_k\mathbf{p}_{i_k}^k) \\
&= f(\mathbf{x}^k) - t_k\langle \nabla_{i_k} f(\mathbf{x}^k), \mathbf{x}_{i_k}^k - \mathbf{p}_{i_k}^k\rangle + \frac{t_k^2 L_{i_k}}{2}\|\mathbf{p}_{i_k}^k - \mathbf{x}_{i_k}^k\|^2 + g(\mathbf{x}^k) \\
&\quad - g_{i_k}(\mathbf{x}_{i_k}^k) + g_{i_k}((1-t_k)\mathbf{x}_{i_k}^k + t_k\mathbf{p}_{i_k}^k) \\
&\leq f(\mathbf{x}^k) - t_k\langle \nabla_{i_k} f(\mathbf{x}^k), \mathbf{x}_{i_k}^k - \mathbf{p}_{i_k}^k\rangle + \frac{t_k^2 L_{i_k}}{2}\|\mathbf{p}_{i_k}^k - \mathbf{x}_{i_k}^k\|^2 + g(\mathbf{x}^k) \\
&\quad - g_{i_k}(\mathbf{x}_{i_k}^k) + (1-t_k)g_{i_k}(\mathbf{x}_{i_k}^k) + t_k g_{i_k}(\mathbf{p}_{i_k}^k) \\
&= F(\mathbf{x}^k) - t_k S_{i_k}(\mathbf{x}^k) + \frac{t_k^2 L_{i_k}}{2}\|\mathbf{p}_{i_k}^k - \mathbf{x}_{i_k}^k\|^2.
\end{aligned}$$

Taking expectation w.r.t. the random variable i_k, we obtain

$$\begin{aligned}
\mathsf{E}_{i_k}(F(\mathbf{x}^{k+1})) &\leq F(\mathbf{x}^k) - \frac{t_k}{p}\sum_{i=1}^{p} S_i(\mathbf{x}^k) + \frac{t_k^2}{2p}\sum_{i=1}^{p} L_i\|\mathbf{p}_i^k - \mathbf{x}_i^k\|^2 \\
&= F(\mathbf{x}^k) - \frac{t_k}{p}S(\mathbf{x}^k) + \frac{t_k^2}{2p}\|\mathbf{p}^k - \mathbf{x}^k\|_L^2.
\end{aligned}$$

Taking expectation w.r.t. ξ_{k-1} and using the bound (13.54) results with the following inequality:

$$\mathsf{E}_{\xi_k}(F(\mathbf{x}^{k+1})) \leq \mathsf{E}_{\xi_{k-1}}(F(\mathbf{x}^k)) - \frac{t_k}{p}\mathsf{E}_{\xi_{k-1}}(S(\mathbf{x}^k)) + \frac{t_k^2}{2p}\Omega^2.$$

Defining $\alpha_k = \frac{t_k}{p} = \frac{2}{k+2p}$ and subtracting F_{opt} from both sides, we obtain

$$\mathsf{E}_{\xi_k}(F(\mathbf{x}^{k+1})) - F_{\text{opt}} \leq \mathsf{E}_{\xi_{k-1}}(F(\mathbf{x}^k)) - F_{\text{opt}} - \alpha_k\mathsf{E}_{\xi_{k-1}}(S(\mathbf{x}^k)) + \frac{p\alpha_k^2}{2}\Omega^2.$$

Invoking Lemma 13.13 with $a_k = \mathsf{E}_{\xi_{k-1}}(F(\mathbf{x}^k)) - F_{\text{opt}}, b_k = \mathsf{E}_{\xi_{k-1}}(S(\mathbf{x}^k))$, and $A = p\Omega^2$, noting that by Lemma 13.12 $a_k \leq b_k$, the inequalities (13.55) and (13.56) follow. □

Chapter 14
Alternating Minimization

Underlying Spaces: In this chapter, all the underlying spaces are Euclidean.

14.1 The Method

Consider the problem

$$\min_{\mathbf{x}_1\in\mathbb{E}_1,\mathbf{x}_2\in\mathbb{E}_2,\ldots,\mathbf{x}_p\in\mathbb{E}_p} F(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_p), \tag{14.1}$$

where $\mathbb{E}_1,\mathbb{E}_2,\ldots,\mathbb{E}_p$ are Euclidean spaces whose product space is denoted by $\mathbb{E} = \mathbb{E}_1\times\mathbb{E}_2\times\cdots\times\mathbb{E}_p$. We use our convention (see Section 1.9) that the product space is also Euclidean with endowed norm

$$\|(\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_p)\|_{\mathbb{E}} = \sqrt{\sum_{i=1}^{p}\|\mathbf{u}_i\|_{\mathbb{E}_i}^2}.$$

We will omit the subscripts of the norms indicating the underlying vector space whose identity will be clear from the context. At this point we only assume that $F:\mathbb{E}\to(-\infty,\infty]$ is proper, but obviously to assure some kind of convergence, additional assumptions will be imposed.

For any $i\in\{1,2,\ldots,p\}$ we define $\mathcal{U}_i:\mathbb{E}_i\to\mathbb{E}$ to be the linear transformation given by

$$\mathcal{U}_i(\mathbf{d}) = (\mathbf{0},\ldots,\mathbf{0},\underbrace{\mathbf{d}}_{i\text{th block}},\mathbf{0},\ldots,\mathbf{0}),\quad \mathbf{d}\in\mathbb{E}_i.$$

We also use throughout this chapter the notation that a vector $\mathbf{x}\in\mathbb{E}$ can be written as

$$\mathbf{x} = (\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_p),$$

and this relation will also be written as $\mathbf{x} = (\mathbf{x}_i)_{i=1}^p$.

In this chapter we consider the *alternating minimization method* in which we successively pick a block in a cyclic manner and set the new value of the chosen block to be a minimizer of the objective w.r.t. the chosen block. The kth iterate is denoted by $\mathbf{x}^k = (\mathbf{x}_1^k,\mathbf{x}_2^k,\ldots,\mathbf{x}_p^k)$. Each iteration of the alternating minimization

method involves p "subiterations" and the by-products of these sub-iterations will be denoted by the following auxiliary subsequences:

$$\begin{aligned}
\mathbf{x}^{k,0} &= \mathbf{x}^k = (\mathbf{x}_1^k, \mathbf{x}_2^k, \ldots, \mathbf{x}_p^k),\\
\mathbf{x}^{k,1} &= (\mathbf{x}_1^{k+1}, \mathbf{x}_2^k, \ldots, \mathbf{x}_p^k),\\
\mathbf{x}^{k,2} &= (\mathbf{x}_1^{k+1}, \mathbf{x}_2^{k+1}, \mathbf{x}_3^k, \ldots, \mathbf{x}_p^k),\\
&\vdots\\
\mathbf{x}^{k,p} &= \mathbf{x}^{k+1} = (\mathbf{x}_1^{k+1}, \mathbf{x}_2^{k+1}, \ldots, \mathbf{x}_p^{k+1}).
\end{aligned} \tag{14.2}$$

The alternating minimization method for minimizing F is described below.

The Alternating Minimization Method

Initialization: pick $\mathbf{x}^0 = (\mathbf{x}_1^0, \mathbf{x}_2^0, \ldots, \mathbf{x}_p^0) \in \mathrm{dom}(F)$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following step:

- for $i = 1, 2, \ldots, p$, compute

$$\mathbf{x}_i^{k+1} \in \mathrm{argmin}_{\mathbf{x}_i \in \mathbb{E}_i} F(\mathbf{x}_1^{k+1}, \ldots, \mathbf{x}_{i-1}^{k+1}, \mathbf{x}_i, \mathbf{x}_{i+1}^k, \ldots, \mathbf{x}_p^k). \tag{14.3}$$

In our notation, we can alternatively rewrite the general step of the alternating minimization method as follows:

- set $\mathbf{x}^{k,0} = \mathbf{x}^k$;
- for $i = 1, 2, \ldots, p$, compute $\mathbf{x}^{k,i} = \mathbf{x}^{k,i-1} + \mathcal{U}_i(\tilde{\mathbf{y}} - \mathbf{x}_i^k)$, where

$$\tilde{\mathbf{y}} \in \mathrm{argmin}_{\mathbf{y} \in \mathbb{E}_i} F(\mathbf{x}^{k,i-1} + \mathcal{U}_i(\mathbf{y} - \mathbf{x}_i^k)); \tag{14.4}$$

- set $\mathbf{x}^{k+1} = \mathbf{x}^{k,p}$.

The following simple lemma states that if F is proper and closed and has bounded level sets, then problem (14.1) has a minimizer and the alternating minimization method is well defined in the sense that the minimization problems (14.3) (or in their alternative form (14.4)) possess minimizers. In the sequel, we will impose additional assumptions on the structure of F that will enable us to establish convergence results.

Lemma 14.1 (alternating minimization is well defined). *Suppose that $F : \mathbb{E} \to (-\infty, \infty]$ ($\mathbb{E} = \mathbb{E}_1 \times \mathbb{E}_2 \times \cdots \times \mathbb{E}_p$) is a proper and closed function. Assume further that F has bounded level sets; that is, $\mathrm{Lev}(F, \alpha) = \{\mathbf{x} \in \mathbb{E} : F(\mathbf{x}) \le \alpha\}$ is bounded for any $\alpha \in \mathbb{R}$. Then the function F has at least one minimizer, and for any $\bar{\mathbf{x}} \in \mathrm{dom}(F)$ and $i \in \{1, 2, \ldots, p\}$ the problem*

$$\min_{\mathbf{y} \in \mathbb{E}_i} F(\bar{\mathbf{x}} + \mathcal{U}_i(\mathbf{y} - \bar{\mathbf{x}}_i)) \tag{14.5}$$

possesses a minimizer.

Proof. Take $\tilde{\mathbf{x}} \in \mathrm{dom}(F)$. Then

$$\mathrm{argmin}_{\mathbf{x}\in\mathbb{E}} F(\mathbf{x}) = \mathrm{argmin}_{\mathbf{x}\in\mathbb{E}}\{F(\mathbf{x}) : \mathbf{x} \in \mathrm{Lev}(F, F(\tilde{\mathbf{x}}))\}.$$

Since F is closed with bounded level sets, it follows that $\mathrm{Lev}(F, F(\tilde{\mathbf{x}}))$ is compact. Hence, by the Weierstrass theorem for closed functions (Theorem 2.12), it follows that the problem of minimizing the proper and closed function F over $\mathrm{Lev}(F, F(\tilde{\mathbf{x}}))$, and hence also the problem of minimizing F over the entire space, possesses a minimizer. Since the function $\mathbf{y} \mapsto F(\bar{\mathbf{x}} + \mathcal{U}_i(\mathbf{y} - \bar{\mathbf{x}}_i))$ is proper and closed with bounded level sets, the same argument shows that problem (14.5) also possesses a minimizer. ∎

14.2 Coordinate-wise Minima

By the definition of the method, it is clear that convergence will most likely be proved (if at all possible) to *coordinate-wise minimum points*.

Definition 14.2. *A vector* $\mathbf{x}^* \in \mathbb{E}$ *is a* **coordinate-wise minimum** *of a function* $F : \mathbb{E}_1 \times \mathbb{E}_2 \times \cdots \times \mathbb{E}_p \to (-\infty, \infty]$ *if* $\mathbf{x}^* \in \mathrm{dom}(F)$ *and*

$$F(\mathbf{x}^*) \le F(\mathbf{x}^* + \mathcal{U}_i(\mathbf{y})) \text{ for all } i = 1, 2, \ldots, p, \mathbf{y} \in \mathbb{E}_i.$$

The next theorem is a rather standard result showing that under properness and closedness of the objective function, as well as an assumption on the uniqueness of the minimizers of the class of subproblems solved at each iteration, the limit points of the sequence generated by the alternating minimization method are coordinate-wise minima.

Theorem 14.3 (convergence of alternating minimization to coordinate-wise minima).[77] *Suppose that* $F : \mathbb{E} \to (-\infty, \infty]$ *(*$\mathbb{E} = \mathbb{E}_1 \times \mathbb{E}_2 \times \cdots \times \mathbb{E}_p$*) is a proper closed function that is continuous over its domain. Assume that*

(A) *for each* $\bar{\mathbf{x}} \in \mathrm{dom}(F)$ *and* $i \in \{1, 2, \ldots, p\}$ *the problem* $\min_{\mathbf{y}\in\mathbb{E}_i} F(\bar{\mathbf{x}} + \mathcal{U}_i(\mathbf{y} - \bar{\mathbf{x}}_i))$ *has a unique minimizer;*

(B) *the level sets of* F *are bounded, meaning that for any* $\alpha \in \mathbb{R}$*, the set* $\mathrm{Lev}(F, \alpha) = \{\mathbf{x} \in \mathbb{E} : F(\mathbf{x}) \le \alpha\}$ *is bounded.*

Let $\{\mathbf{x}^k\}_{k\ge0}$ *be the sequence generated by the alternating minimization method for minimizing* F*. Then* $\{\mathbf{x}^k\}_{k\ge0}$ *is bounded, and any limit point of the sequence is a coordinate-wise minimum.*

Proof. To prove that the sequence is bounded, note that by the definition of the method, the sequence of function values $\{F(\mathbf{x}^k)\}_{k\ge0}$ is nonincreasing, which in particular implies that $\{\mathbf{x}^k\}_{k\ge0} \subseteq \mathrm{Lev}(F, F(\mathbf{x}^0))$; therefore, by condition (B), it follows that the sequence $\{\mathbf{x}^k\}_{k\ge0}$ is bounded, which along with the closedness of F

[77] Theorem 14.3 and its proof originate from Bertsekas [28, Proposition 2.7.1].

implies that $\{F(\mathbf{x}^k)\}_{k\geq 0}$ is bounded below. We can thus conclude that $\{F(\mathbf{x}^k)\}_{k\geq 0}$ converges to some real number $\bar{F}$. Since $F(\mathbf{x}^k) \geq F(\mathbf{x}^{k,1}) \geq F(\mathbf{x}^{k+1})$, it follows that $\{F(\mathbf{x}^{k,1})\}_{k\geq 0}$ also converges to $\bar{F}$, meaning that the sequences $\{F(\mathbf{x}^k)\}_{k\geq 0}$ and $\{F(\mathbf{x}^{k,1})\}_{k\geq 0}$ converge to the same value.

Now, suppose that $\bar{\mathbf{x}}$ is a limit point of $\{\mathbf{x}^k\}_{k\geq 0}$. Then there exists a subsequence $\{\mathbf{x}^{k_j}\}_{j\geq 0}$ converging to $\bar{\mathbf{x}}$. Since the sequence $\{\mathbf{x}^{k_j,1}\}_{j\geq 0}$ is bounded (follows directly from the boundedness of $\{\mathbf{x}^k\}_{k\geq 0}$), by potentially passing to a subsequence, we can assume that $\{\mathbf{x}^{k_j,1}\}_{j\geq 0}$ converges to some vector $(\mathbf{v}, \bar{\mathbf{x}}_2, \ldots, \bar{\mathbf{x}}_p)$ $(\mathbf{v} \in \mathbb{E}_1)$. By definition of the method,

$$F(\mathbf{x}_1^{k_j+1}, \mathbf{x}_2^{k_j}, \ldots, \mathbf{x}_p^{k_j}) \leq F(\mathbf{x}_1, \mathbf{x}_2^{k_j}, \ldots, \mathbf{x}_p^{k_j}) \text{ for any } \mathbf{x}_1 \in \mathbb{E}_1.$$

Taking the limit $j \to \infty$ and using the closedness of F, as well as the continuity of F over its domain, we obtain that

$$F(\mathbf{v}, \bar{\mathbf{x}}_2, \ldots, \bar{\mathbf{x}}_p) \leq F(\mathbf{x}_1, \bar{\mathbf{x}}_2, \ldots, \bar{\mathbf{x}}_p) \text{ for any } \mathbf{x}_1 \in \mathbb{E}_1.$$

Since $\{F(\mathbf{x}^k)\}_{k\geq 0}$ and $\{F(\mathbf{x}^{k,1})\}_{k\geq 0}$ converge to the same value, we have

$$F(\mathbf{v}, \bar{\mathbf{x}}_2, \ldots, \bar{\mathbf{x}}_p) = F(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \ldots, \bar{\mathbf{x}}_p),$$

which by the uniqueness of the minimizer w.r.t. the first block (condition (A)) implies that $\mathbf{v} = \bar{\mathbf{x}}_1$. Therefore,

$$F(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \ldots, \bar{\mathbf{x}}_p) \leq F(\mathbf{x}_1, \bar{\mathbf{x}}_2, \ldots, \bar{\mathbf{x}}_p) \text{ for any } \mathbf{x}_1 \in \mathbb{E}_1,$$

which is the first condition for coordinate-wise minimality. We have shown that $\mathbf{x}^{k_j,1} \to \bar{\mathbf{x}}$ as $j \to \infty$. This means that we can repeat the arguments when $\mathbf{x}^{k_j,1}$ replaces $\mathbf{x}^{k_j}$ and concentrate on the second coordinate to obtain that

$$F(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \ldots, \bar{\mathbf{x}}_p) \leq F(\bar{\mathbf{x}}_1, \mathbf{x}_2, \bar{\mathbf{x}}_3, \ldots, \bar{\mathbf{x}}_p) \text{ for any } \mathbf{x}_2 \in \mathbb{E}_2,$$

which is the second condition for coordinate-wise minimality. The above argument can be repeated until we show that $\bar{\mathbf{x}}$ satisfies all the conditions for coordinate-wise minimality. ∎

The following famous example of Powell describes a situation in which the alternating minimization method produces a sequence whose limit points are *not* coordinate-wise minima points.

Example 14.4 (Powell's example—failure of alternating minimization I).[78] Let

$$\varphi(x, y, z) = -xy - yz - zx + [x-1]_+^2 + [-x-1]_+^2 + [y-1]_+^2 + [-y-1]_+^2 + [z-1]_+^2 + [-z-1]_+^2.$$

Note that φ is differentiable. Fixing y and z, it is easy to show that

$$\operatorname{argmin}_x \varphi(x, y, z) = \begin{cases} \operatorname{sgn}(y+z)(1 + \frac{1}{2}|y+z|), & y + z \neq 0, \\ [-1, 1], & y + z = 0, \end{cases} \tag{14.6}$$

[78] Powell's example is from [106].

and similarly (by the symmetry of φ),

$$\operatorname{argmin}_y \varphi(x,y,z) = \begin{cases} \operatorname{sgn}(x+z)(1+\frac{1}{2}|x+z|), & x+z \neq 0, \\ [-1,1], & x+z=0, \end{cases} \tag{14.7}$$

$$\operatorname{argmin}_z \varphi(x,y,z) = \begin{cases} \operatorname{sgn}(x+y)(1+\frac{1}{2}|x+y|), & x+y \neq 0, \\ [-1,1], & x+y=0. \end{cases} \tag{14.8}$$

Suppose that $\varepsilon > 0$ and that we initialize the alternating minimization method with the point $\left(-1-\varepsilon, 1+\frac{1}{2}\varepsilon, -1-\frac{1}{4}\varepsilon\right)$. Then the first six iterations are

$$\begin{gathered} \left(1+\frac{1}{8}\varepsilon, 1+\frac{1}{2}\varepsilon, -1-\frac{1}{4}\varepsilon\right), \\ \left(1+\frac{1}{8}\varepsilon, -1-\frac{1}{16}\varepsilon, -1-\frac{1}{4}\varepsilon\right), \\ \left(1+\frac{1}{8}\varepsilon, -1-\frac{1}{16}\varepsilon, 1+\frac{1}{32}\varepsilon\right), \\ \left(-1-\frac{1}{64}\varepsilon, -1-\frac{1}{16}\varepsilon, 1+\frac{1}{32}\varepsilon\right), \\ \left(-1-\frac{1}{64}\varepsilon, 1+\frac{1}{128}\varepsilon, 1+\frac{1}{32}\varepsilon\right), \\ \left(-1-\frac{1}{64}\varepsilon, 1+\frac{1}{128}\varepsilon, -1-\frac{1}{256}\varepsilon\right). \end{gathered}$$

We are essentially back to the first point, but with $\frac{1}{64}\varepsilon$ replacing ε. The process continues by cycling around the six points

$$(1,1,-1),\ (1,-1,-1),\ (1,-1,1),\ (-1,-1,1),\ (-1,1,1),\ (-1,1,-1).$$

None of these points is a stationary point of φ. Indeed,

$$\nabla\varphi(1,1,-1) = (0,0,-2), \quad \nabla\varphi(-1,1,1) = (-2,0,0), \quad \nabla\varphi(1,-1,1) = (0,-2,0),$$
$$\nabla\varphi(-1,-1,1) = (0,0,2), \quad \nabla\varphi(1,-1,-1) = (2,0,0), \quad \nabla\varphi(-1,1,-1) = (0,2,0).$$

Since the limit points are not stationary points of φ, they are also not coordinate-wise minima[79] points. The fact that the limit points of the sequence generated by the alternating minimization method are not coordinate-wise minima is not a contradiction to Theorem 14.3 since two assumptions are not met: the subproblems solved at each iteration do not necessarily possess unique minimizers, and the level sets of φ are not bounded since for any $x > 1$

$$\varphi(x,x,x) = -3x^2 + 3(x-1)^2 = -6x + 3$$

[79] For example, to show that $(1,1,-1)$ is not a coordinate-wise minimum, note that since $\nabla\varphi(1,1,-1) = (0,0,-2)$, then $(1,1,-1+\delta)$ for small enough $\delta > 0$ will have a smaller function value than $(1,1,-1)$.

goes to $-\infty$ as $x \to \infty$. A close inspection of the proof of Theorem 14.3 reveals that the assumption on the boundedness of the level sets in Theorem 14.3 is only required in order to assure the boundedness of the sequence generated by the method. Since the sequence in this example is in any case bounded, it follows that the failure to converge to a coordinate-wise minimum is actually due to the nonuniqueness of the optimal solutions of the subproblems (14.6),(14.7) and (14.8). ∎

Note that if the alternating minimization method reaches a coordinate-wise minimum, then it might get stuck there since the point is optimal w.r.t. each block.[80] The natural question is of course whether coordinate-wise minima are necessarily stationary points of the problem, meaning that they satisfy the most basic optimality condition of the problem. The answer is unfortunately *no* even when the objective function is convex, as the following example illustrates.

Example 14.5 (failure of alternating minimization II). Consider the convex function

$$F(x_1, x_2) = |3x_1 + 4x_2| + |x_1 - 2x_2|.$$

The function satisfies all the assumptions of Theorem 14.3: it is proper, closed, and continuous with bounded level sets and has a unique minimizer w.r.t. each variable. Therefore, Theorem 14.3 guarantees that the limit point points of the alternating minimization method are coordinate-wise minima points. We will see that for the specific problem under consideration, this result is of very little importance.

The unique minimizer of the function is $(x_1, x_2) = (0, 0)$. However, for any $\alpha \in \mathbb{R}$ the point $(-4\alpha, 3\alpha)$ is a coordinate-wise minimum of f. To show this, assume first that $\alpha > 0$. Note that

$$F(-4\alpha, t) = |4t - 12\alpha| + |2t + 4\alpha| = \begin{cases} -6t + 8\alpha, & t < -2\alpha, \\ -2t + 16\alpha, & -2\alpha \le t \le 3\alpha, \\ 6t - 8\alpha, & t > 3\alpha, \end{cases}$$

and obviously $t = 3\alpha$ is the minimizer of $F(-4\alpha, t)$. Similarly, the optimal solution of

$$F(t, 3\alpha) = |3t + 12\alpha| + |t - 6\alpha| = \begin{cases} -4t - 6\alpha, & t < -4\alpha, \\ 2t + 18\alpha, & -4\alpha \le t \le 6\alpha, \\ 4t + 6\alpha, & t > 6\alpha, \end{cases}$$

is $t = -4\alpha$. A similar argument also shows that $(-4\alpha, 3\alpha)$ is a coordinate-wise minimum also for $\alpha < 0$. We conclude that $(-4\alpha, 3\alpha)$ is a coordinate-wise minimum for any $\alpha \in \mathbb{R}$ where only the value $\alpha = 0$ corresponds to the actual minimum of F; all other values correspond to nonoptimal/nonstationary[81] points of F. The severity of the situation is made clear when noting that after only one iteration

[80] Actually, the only situation in which the method might move away from a coordinate-wise minimum is if there are multiple optimal solutions to some of the subproblems solved at each subiteration of the method.

[81] In the sense that $\mathbf{0} \notin \partial F(\mathbf{x})$.

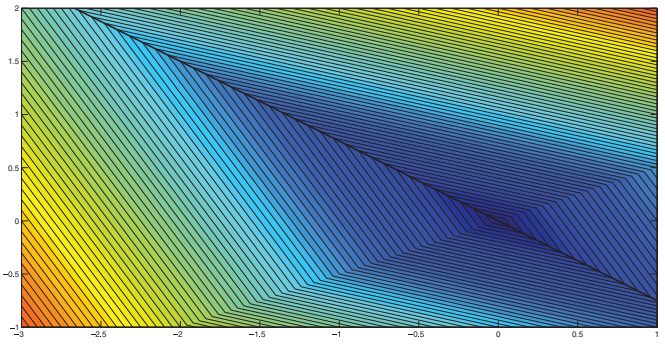

Figure 14.1. *Contour lines of the function* $f(x_1, x_2) = |3x_1 + 4x_2| + |x_1 - 2x_2|$. *All the points on the emphasized line* $\{(-4\alpha, 3\alpha) : \alpha \in \mathbb{R}\}$ *are coordinate-wise minima, and only* $(0, 0)$ *is a global minimum.*

of alternating minimization, the method gets stuck at a coordinate-wise minimum, which, unless the initial vector contains at least one zero element, is a nonoptimal point (easy to show). The contour lines of F, as well as the line comprising the continuum of coordinate-wise minima points is described in Figure 14.1. ∎

Example 14.5 shows that even if convexity is assumed, coordinate-wise minima points are not necessarily stationary points of the objective function; in particular, this means that the alternating minimization method will not be guaranteed to converge to stationary points (which are global minima points in the convex case). One possible reason for this phenomena is that the stationarity condition $\mathbf{0} \in \partial F(\mathbf{x})$ does not decompose into separate conditions on each block. This is why, in the next section, we present a specific model for the function F for which we will be able to prove that coordinate-wise minima points are necessarily stationary points.

14.3 The Composite Model

The model that we will analyze from now on is the composite model, which was discussed in Sections 11.2 and 13.4 in the contexts of the block proximal gradient and block conditional gradient methods. Thus, our main model is

$$\min_{\mathbf{x}_1 \in \mathbb{E}_1, \mathbf{x}_2 \in \mathbb{E}_2, \ldots, \mathbf{x}_p \in \mathbb{E}_p} \left\{ F(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) = f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) + \sum_{j=1}^{p} g_j(\mathbf{x}_j) \right\}. \quad (14.9)$$

The function $g : \mathbb{E} \to (-\infty, \infty]$ is defined by

$$g(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) \equiv \sum_{i=1}^{p} g_i(\mathbf{x}_i).$$

The gradient w.r.t. the ith block ($i \in \{1, 2, \ldots, p\}$) is denoted by $\nabla_i f$, and the following is satisfied:

$$\nabla f(\mathbf{x}) = (\nabla_1 f(\mathbf{x}), \nabla_2 f(\mathbf{x}), \ldots, \nabla_p f(\mathbf{x})).$$

Note that in our notation the main model (14.9) can be simply written as

$$\min_{\mathbf{x} \in \mathbb{E}} \{F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})\}.$$

The basic assumptions on the model are summarized below.

Assumption 14.6.

(A) $g_i : \mathbb{E}_i \to (-\infty, \infty]$ *is proper closed and convex for any* $i \in \{1, 2, \ldots, p\}$. *In addition,* g_i *is continuous over its domain.*

(B) $f : \mathbb{E} \to (-\infty, \infty]$ *is a closed function;* $\mathrm{dom}(f)$ *is convex;* f *is differentiable over* $\mathrm{int}(\mathrm{dom}(f))$ *and* $\mathrm{dom}(g) \subseteq \mathrm{int}(\mathrm{dom}(f))$.

Under the above structure of the function F, the general step of the alternating minimization method (14.3) can be compactly written as

$$\mathbf{x}_i^{k+1} \in \mathrm{argmin}_{\mathbf{x}_i \in \mathbb{E}_i} \{f(\mathbf{x}_1^{k+1}, \ldots, \mathbf{x}_{i-1}^{k+1}, \mathbf{x}_i, \mathbf{x}_{i+1}^{k}, \ldots, \mathbf{x}_p^{k}) + g_i(\mathbf{x}_i)\},$$

where we omitted from the above the constant terms related to the functions g_j, $j \neq i$.

Recall that a point $\mathbf{x}^* \in \mathrm{dom}(g)$ is a stationary point of problem (14.9) if it satisfies $-\nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*)$ (Definition 3.73) and that by Theorem 11.6(a), this condition can be written equivalently as $-\nabla_i f(\mathbf{x}^*) \in \partial g_i(\mathbf{x}^*)$, $i = 1, 2, \ldots, p$. The latter fact will enable us to show that coordinate-wise minima points of F are stationary points of problem (14.9).

Lemma 14.7 (coordinate-wise minimality ⇒ stationarity). *Suppose that Assumption* 14.6 *holds and that* $\mathbf{x}^* \in \mathrm{dom}(g)$ *is a coordinate-wise minimum of* $F = f + g$. *Then* $\mathbf{x}^*$ *is a stationary point of problem* (14.9).

Proof. Since $\mathbf{x}^*$ is a coordinate-wise minimum of F, it follows that for all $i \in \{1, 2, \ldots, p\}$,

$$\mathbf{x}_i^* \in \mathrm{argmin}_{\mathbf{y} \in \mathbb{E}_i} \{\tilde{f}_i(\mathbf{y}) + g_i(\mathbf{y})\},$$

where

$$\tilde{f}_i(\mathbf{y}) \equiv f(\mathbf{x}^* + \mathcal{U}_i(\mathbf{y} - \mathbf{x}_i^*)) = f(\mathbf{x}_1^*, \ldots, \mathbf{x}_{i-1}^*, \mathbf{y}, \mathbf{x}_{i+1}^*, \ldots, \mathbf{x}_p^*).$$

Therefore, by Theorem 3.72(a), $-\nabla \tilde{f}_i(\mathbf{x}_i^*) \in \partial g_i(\mathbf{x}^*)$. Since $\nabla \tilde{f}_i(\mathbf{x}_i^*) = \nabla_i f(\mathbf{x}^*)$, we conclude that for any i, $-\nabla_i f(\mathbf{x}^*) \in \partial g_i(\mathbf{x}^*)$. Thus, invoking Theorem 11.6(a), we obtain that $-\nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*)$, namely, that $\mathbf{x}^*$ is a stationary point of problem (14.9). ∎

Recall that Theorem 14.3 showed under appropriate assumptions that limit points of the sequence generated by the alternating minimization method are coordinate-wise minima points. Combining this result with Lemma 14.7 we obtain the following corollary.

Corollary 14.8. *Suppose that Assumption* 14.6 *holds, and assume further that* $F = f + g$ *satisfies the following:*

- *for each* $\bar{\mathbf{x}} \in \operatorname{dom}(F)$ *and* $i \in \{1, 2, \ldots, p\}$ *the problem* $\min_{\mathbf{y} \in \mathbb{E}_i} F(\bar{\mathbf{x}} + \mathcal{U}_i(\mathbf{y} - \bar{\mathbf{x}}_i))$ *has a unique minimizer;*
- *the level sets of* F *are bounded, meaning that for any* $\alpha \in \mathbb{R}$*, the set* $\{\mathbf{x} \in \mathbb{E} : F(\mathbf{x}) \leq \alpha\}$ *is bounded.*

Let $\{\mathbf{x}^k\}_{k \geq 0}$ *be the sequence generated by the alternating minimization method for solving* (14.9)*. Then* $\{\mathbf{x}^k\}_{k \geq 0}$ *is bounded, and any limit point of the sequence is a stationary point of problem* (14.9)*.*

14.4 Convergence in the Convex Case

The convergence results previously established require a rather strong assumption on the uniqueness of the optimal solution to the class of subproblems that are solved at each sub-iteration of the alternating minimization method. We will show how this assumption can be removed if we assume convexity of the objective function.

Theorem 14.9.[82] *Suppose that Assumption* 14.6 *holds and that in addition*

- f *is convex;*
- f *is continuously differentiable*[83] *over* $\operatorname{int}(\operatorname{dom}(f))$*;*
- *the function* $F = f + g$ *satisfies that the level sets of* F *are bounded, meaning that for any* $\alpha \in \mathbb{R}$*, the set* $\operatorname{Lev}(F, \alpha) = \{\mathbf{x} \in \mathbb{E} : F(\mathbf{x}) \leq \alpha\}$ *is bounded.*

Then the sequence generated by the alternating minimization method for solving problem (14.9) *is bounded, and any limit point of the sequence is an optimal solution of the problem.*

Proof. Let $\{\mathbf{x}^k\}_{k \geq 0}$ be the sequence generated by the alternating minimization method, and let $\{\mathbf{x}^{k,i}\}_{k \geq 0}$ $(i = 0, 1, \ldots, p)$ be the auxiliary sequences given in (14.2). We begin by showing that $\{\mathbf{x}^k\}_{k \geq 0}$ is bounded. Indeed, by the definition of the method, the sequence of function values is nonincreasing, and hence $\{\mathbf{x}^k\}_{k \geq 0} \subseteq \operatorname{Lev}(F, F(\mathbf{x}^0))$. Since $\operatorname{Lev}(F, F(\mathbf{x}^0))$ is bounded by the premise of the theorem, it follows that $\{\mathbf{x}^k\}_{k \geq 0}$ is bounded.

Let $\bar{\mathbf{x}} \in \operatorname{dom}(g)$ be a limit point of $\{\mathbf{x}^k\}_{k \geq 0}$. We will show that $\bar{\mathbf{x}}$ is an optimal solution of problem (14.9). Since $\bar{\mathbf{x}}$ is a limit point of the sequence, there exists a subsequence $\{\mathbf{x}^{k_j}\}_{j \geq 0}$ for which $\mathbf{x}^{k_j} \to \bar{\mathbf{x}}$. By potentially passing to a subsequence, the sequences $\{\mathbf{x}^{k_j, i}\}_{j \geq 0}$ $(i = 1, 2, \ldots, p)$ can also be assumed to be convergent and $\mathbf{x}^{k_j, i} \to \bar{\mathbf{x}}^i \in \operatorname{dom}(g)$ as $j \to \infty$ for all $i \in \{0, 1, 2, \ldots, p\}$. Obviously, the following three properties hold:

- **[P1]** $\bar{\mathbf{x}} = \bar{\mathbf{x}}^0$.

[82] Theorem 14.9 is an extension of Proposition 6 from Grippo and Sciandrone [61] to the composite model.

[83] "Continuously differentiable" means that the gradient is a continuous mapping.

- **[P2]** for any i, $\bar{\mathbf{x}}^i$ is different from $\bar{\mathbf{x}}^{i-1}$ only at the ith block (if at all different).
- **[P3]** $F(\bar{\mathbf{x}}) = F(\bar{\mathbf{x}}^i)$ for all $i \in \{0, 1, 2, \ldots, p\}$ (easily shown by taking the limit $j \to \infty$ in the inequality $F(\mathbf{x}^{k_j}) \geq F(\mathbf{x}^{k_j,i}) \geq F(\mathbf{x}^{k_j+1})$ and using the closedness F, as well as the continuity of F over its domain).

By the definition of the sequence we have for all $j \geq 0$ and $i \in \{1, 2, \ldots, p\}$,

$$\mathbf{x}_i^{k_j,i} \in \operatorname{argmin}_{\mathbf{x}_i \in \mathbb{E}_i} F(\mathbf{x}_1^{k_j+1}, \ldots, \mathbf{x}_{i-1}^{k_j+1}, \mathbf{x}_i, \mathbf{x}_{i+1}^{k_j}, \ldots, \mathbf{x}_p^{k_j}).$$

Therefore, since $\mathbf{x}_i^{k_j,i}$ is a stationary point of the above minimization problem (see Theorem 3.72(a)),

$$-\nabla_i f(\mathbf{x}^{k_j,i}) \in \partial g_i(\mathbf{x}_i^{k_j,i}).$$

Taking the limit[84] $j \to \infty$ and using the continuity of ∇f, we obtain that

$$-\nabla_i f(\bar{\mathbf{x}}^i) \in \partial g_i(\bar{\mathbf{x}}_i^i). \tag{14.10}$$

Note that for any $\mathbf{x}_{i+1} \in \operatorname{dom}(g_{i+1})$,

$$F(\mathbf{x}^{k_j,i+1}) \leq F(\mathbf{x}_1^{k_j+1}, \ldots, \mathbf{x}_i^{k_j+1}, \mathbf{x}_{i+1}, \mathbf{x}_{i+2}^{k_j}, \ldots, \mathbf{x}_p^{k_j}).$$

Taking the limit $j \to \infty$ and using [P3], we conclude that for any $\mathbf{x}_{i+1} \in \operatorname{dom}(g_{i+1})$,

$$F(\bar{\mathbf{x}}^i) = F(\bar{\mathbf{x}}^{i+1}) \leq F(\bar{\mathbf{x}}_1^i, \ldots, \bar{\mathbf{x}}_i^i, \mathbf{x}_{i+1}, \bar{\mathbf{x}}_{i+2}^i, \ldots, \bar{\mathbf{x}}_p^i),$$

from which we obtain, using Theorem 3.72(a) again, that for any $i \in \{0, 1, \ldots, p-1\}$,

$$-\nabla_{i+1} f(\bar{\mathbf{x}}^i) \in \partial g_{i+1}(\bar{\mathbf{x}}_{i+1}^i). \tag{14.11}$$

We need to show that the following implication holds for any $i \in \{2, 3, \ldots, p\}, l \in \{1, 2, \ldots, p-1\}$ such that $l < i$:

$$-\nabla_i f(\bar{\mathbf{x}}^l) \in \partial g_i(\bar{\mathbf{x}}_i^l) \Rightarrow -\nabla_i f(\bar{\mathbf{x}}^{l-1}) \in \partial g_i(\bar{\mathbf{x}}_i^{l-1}). \tag{14.12}$$

To prove the above implication, assume that $-\nabla_i f(\bar{\mathbf{x}}^l) \in \partial g_i(\bar{\mathbf{x}}_i^l)$ and let $\boldsymbol{\eta} \in \mathbb{E}_i$. Then

$$\begin{aligned}
\langle \nabla f(\bar{\mathbf{x}}^l), \bar{\mathbf{x}}^{l-1} + \mathcal{U}_i(\boldsymbol{\eta}) - \bar{\mathbf{x}}^l \rangle &\overset{(*)}{=} \langle \nabla_l f(\bar{\mathbf{x}}^l), \bar{\mathbf{x}}_l^{l-1} - \bar{\mathbf{x}}_l^l \rangle + \langle \nabla_i f(\bar{\mathbf{x}}^l), \boldsymbol{\eta} \rangle \\
&\overset{(**)}{\geq} g_l(\bar{\mathbf{x}}_l^l) - g_l(\bar{\mathbf{x}}_l^{l-1}) + \langle \nabla_i f(\bar{\mathbf{x}}^l), \boldsymbol{\eta} \rangle \\
&\overset{(***)}{=} g_l(\bar{\mathbf{x}}_l^l) - g_l(\bar{\mathbf{x}}_l^{l-1}) + \langle \nabla_i f(\bar{\mathbf{x}}^l), (\bar{\mathbf{x}}_i^{l-1} + \boldsymbol{\eta}) - \bar{\mathbf{x}}_i^l \rangle \\
&\overset{(****)}{\geq} g_l(\bar{\mathbf{x}}_l^l) - g_l(\bar{\mathbf{x}}_l^{l-1}) + g_i(\bar{\mathbf{x}}_i^l) - g_i(\bar{\mathbf{x}}_i^{l-1} + \boldsymbol{\eta}) \\
&= g(\bar{\mathbf{x}}^l) - g(\bar{\mathbf{x}}^{l-1} + \mathcal{U}_i(\boldsymbol{\eta})),
\end{aligned} \tag{14.13}$$

[84] We use here the following simple result: if $h : \mathbb{V} \to (-\infty, \infty]$ is proper closed and convex and $\mathbf{a}^k \in \partial h(\mathbf{b}^k)$ for all k and $\mathbf{a}^k \to \bar{\mathbf{a}}, \mathbf{b}^k \to \bar{\mathbf{b}}$, then $\bar{\mathbf{a}} \in \partial h(\bar{\mathbf{b}})$. To prove the result, take an arbitrary $\mathbf{z} \in \mathbb{V}$. Then since $\mathbf{a}^k \in \partial h(\mathbf{b}^k)$, it follows that $h(\mathbf{z}) \geq h(\mathbf{b}^k) + \langle \mathbf{a}^k, \mathbf{z} - \mathbf{b}^k \rangle$. Taking the liminf of both sides and using the closedness (hence lower semicontinuity) of h, we obtain that $h(\mathbf{z}) \geq h(\bar{\mathbf{b}}) + \langle \bar{\mathbf{a}}, \mathbf{z} - \bar{\mathbf{b}} \rangle$, showing that $\bar{\mathbf{a}} \in \partial h(\bar{\mathbf{b}})$.

where $(*)$ follows by [P2], $(**)$ is a consequence of the relation (14.10) with $i = l$, $(***)$ follows by the fact that for any $l < i$, $\bar{\mathbf{x}}_i^l = \bar{\mathbf{x}}_i^{l-1}$, and $(****)$ is due to our underlying assumption that $-\nabla_i f(\bar{\mathbf{x}}^l) \in \partial g_i(\bar{\mathbf{x}}_i^l)$. Using inequality (14.13) and the gradient inequality on the function f (utilizing its convexity), we obtain

$$\begin{aligned} F(\bar{\mathbf{x}}^{l-1} + \mathcal{U}_i(\boldsymbol{\eta})) &= f(\bar{\mathbf{x}}^{l-1} + \mathcal{U}_i(\boldsymbol{\eta})) + g(\bar{\mathbf{x}}^{l-1} + \mathcal{U}_i(\boldsymbol{\eta})) \\ &\geq f(\bar{\mathbf{x}}^l) + \langle \nabla f(\bar{\mathbf{x}}^l), \bar{\mathbf{x}}^{l-1} + \mathcal{U}_i(\boldsymbol{\eta}) - \bar{\mathbf{x}}^l \rangle + g(\bar{\mathbf{x}}^{l-1} + \mathcal{U}_i(\boldsymbol{\eta})) \\ &\geq F(\bar{\mathbf{x}}^l) \\ &\overset{[P3]}{=} F(\bar{\mathbf{x}}^{l-1}). \end{aligned}$$

We thus obtain that

$$\bar{\mathbf{x}}_i^{l-1} \in \operatorname{argmin}_{\mathbf{x}_i \in \mathbb{E}_i} F(\bar{\mathbf{x}}_1^{l-1}, \ldots, \bar{\mathbf{x}}_{i-1}^{l-1}, \mathbf{x}_i, \bar{\mathbf{x}}_{i+1}^{l-1}, \ldots, \bar{\mathbf{x}}_p^{l-1}),$$

which implies that $-\nabla_i f(\bar{\mathbf{x}}^{l-1}) \in \partial g_i(\bar{\mathbf{x}}_i^{l-1})$, establishing the implication (14.12). We are now ready to prove that $\bar{\mathbf{x}} = \bar{\mathbf{x}}^0$ is an optimal solution of problem (14.9). For that, we will show that for any $m \in \{1, 2, \ldots, p\}$ it holds that

$$-\nabla_m f(\bar{\mathbf{x}}) \in \partial g_m(\bar{\mathbf{x}}_m). \tag{14.14}$$

By Theorem 11.6 these relations are equivalent to stationarity of $\bar{\mathbf{x}}$, and using Theorem 3.72(b) and the convexity of f, we can deduce that $\bar{\mathbf{x}}$ is an optimal solution of problem (14.9). For $m = 1$ the relation (14.14) follows by substituting $i = 0$ in (14.11) and using the fact that $\bar{\mathbf{x}} = \bar{\mathbf{x}}^0$ (property [P1]). Let $m > 1$. Then by (14.11) we have that $-\nabla_m f(\bar{\mathbf{x}}^{m-1}) \in \partial g_m(\bar{\mathbf{x}}_m^{m-1})$. We can now utilize the implication (14.12) several times and obtain

$$\begin{array}{c} -\nabla_m f(\bar{\mathbf{x}}^{m-1}) \in \partial g_m(\bar{\mathbf{x}}_m^{m-1}) \\ \Downarrow \\ -\nabla_m f(\bar{\mathbf{x}}^{m-2}) \in \partial g_m(\bar{\mathbf{x}}_m^{m-2}) \\ \Downarrow \\ \vdots \\ \Downarrow \\ -\nabla_m f(\bar{\mathbf{x}}^0) \in \partial g_m(\bar{\mathbf{x}}_m^0), \end{array}$$

and thus, since $\bar{\mathbf{x}} = \bar{\mathbf{x}}^0$ (property [P1]), we conclude that $-\nabla_m f(\bar{\mathbf{x}}) \in \partial g_m(\bar{\mathbf{x}}_m)$ for any m, implying that $\bar{\mathbf{x}}$ is an optimal solution of problem (14.9). □

14.5 Rate of Convergence in the Convex Case

In this section we will prove some rates of convergence results of the alternating minimization method in the convex setting. We begin by showing a general result that holds for any number of blocks, and we will then establish an improved result for the case $p = 2$.

14.5.1 General p

We will consider the model (14.9) that was studied in the previous two sections. The basic assumptions on the model are gathered in the following.

Assumption 14.10.

(A) $g_i : \mathbb{E}_i \to (-\infty, \infty]$ *is proper closed and convex for any* $i \in \{1, 2, \ldots, p\}$.

(B) $f : \mathbb{E} \to \mathbb{R}$ *is convex and* L_f*-smooth.*

(C) *For any* $\alpha > 0$*, there exists* $R_\alpha > 0$ *such that*

$$\max_{\mathbf{x}, \mathbf{x}^* \in \mathbb{E}} \{\|\mathbf{x} - \mathbf{x}^*\| : F(\mathbf{x}) \le \alpha, \mathbf{x}^* \in X^*\} \le R_\alpha.$$

(D) *The optimal set of problem* (14.9) *is nonempty and denoted by* X^**. The optimal value is denoted by* $F_{\rm opt}$.[85]

Theorem 14.11 ($O(1/k)$ rate of convergence of alternating minimization).[86] *Suppose that Assumption* 14.10 *holds, and let* $\{\mathbf{x}^k\}_{k\ge 0}$ *be the sequence generated by the alternating minimization method for solving problem* (14.9)*. Then for all* $k \ge 2$*,*

$$F(\mathbf{x}^k) - F_{\rm opt} \le \max\left\{\left(\frac{1}{2}\right)^{(k-1)/2}(F(\mathbf{x}^0) - F_{\rm opt}), \frac{8L_f p^2 R^2}{k-1}\right\}, \tag{14.15}$$

where $R = R_{F(\mathbf{x}^0)}$.

Proof. Let $\mathbf{x}^* \in X^*$. Since the sequence of function values $\{F(\mathbf{x}^k)\}_{k\ge 0}$ generated by the method is nonincreasing, it follows that $\{\mathbf{x}^k\}_{k\ge 0} \subseteq \mathrm{Lev}(F, F(\mathbf{x}_0))$, and hence, by Assumption 14.10(C),

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\| \le R, \tag{14.16}$$

where $R = R_{F(\mathbf{x}^0)}$. Let $\{\mathbf{x}^{k,j}\}_{k\ge 0}$ $(j = 0, 1, \ldots, p)$ be the auxiliary sequences given in (14.2). Then for any $k \ge 0$ and $j \in \{0, 1, 2, \ldots, p-1\}$,

$$\begin{aligned}
&F(\mathbf{x}^{k,j}) - F(\mathbf{x}^{k,j+1})\\
&= f(\mathbf{x}^{k,j}) - f(\mathbf{x}^{k,j+1}) + g(\mathbf{x}^{k,j}) - g(\mathbf{x}^{k,j+1})\\
&\ge \langle \nabla f(\mathbf{x}^{k,j+1}), \mathbf{x}^{k,j} - \mathbf{x}^{k,j+1}\rangle + \frac{1}{2L_f}\|\nabla f(\mathbf{x}^{k,j}) - \nabla f(\mathbf{x}^{k,j+1})\|^2 + g(\mathbf{x}^{k,j}) - g(\mathbf{x}^{k,j+1})\\
&= \langle \nabla_{j+1} f(\mathbf{x}^{k,j+1}), \mathbf{x}^k_{j+1} - \mathbf{x}^{k+1}_{j+1}\rangle + \frac{1}{2L_f}\|\nabla f(\mathbf{x}^{k,j}) - \nabla f(\mathbf{x}^{k,j+1})\|^2\\
&\quad + g_{j+1}(\mathbf{x}^k_{j+1}) - g_{j+1}(\mathbf{x}^{k+1}_{j+1}),
\end{aligned} \tag{14.17}$$

where the inequality follows by the convexity and L_f-smoothness of f along with Theorem 5.8 (equivalence between (i) and (iii)). Since

$$\mathbf{x}^{k+1}_{j+1} \in \mathrm{argmin}_{\mathbf{x}_{j+1}} F(\mathbf{x}^{k+1}_1, \ldots, \mathbf{x}^{k+1}_j, \mathbf{x}_{j+1}, \mathbf{x}^k_{j+2}, \ldots, \mathbf{x}^k_p),$$

[85] Property (D) actually follows from properties (A), (B), and (C); see Lemma 14.1.

[86] The proof of Theorem 14.11 follows the proof of Theorem 3.1 from the work of Hong, Wang, Razaviyayn, and Luo [69].

it follows that

$$-\nabla_{j+1} f(\mathbf{x}^{k,j+1}) \in \partial g_{j+1}(\mathbf{x}_{j+1}^{k+1}), \tag{14.18}$$

and hence, by the subgradient inequality,

$$g_{j+1}(\mathbf{x}_{j+1}^{k}) \geq g_{j+1}(\mathbf{x}_{j+1}^{k+1}) - \langle \nabla_{j+1} f(\mathbf{x}^{k,j+1}), \mathbf{x}_{j+1}^{k} - \mathbf{x}_{j+1}^{k+1} \rangle,$$

which, combined with (14.17), yields

$$F(\mathbf{x}^{k,j}) - F(\mathbf{x}^{k,j+1}) \geq \frac{1}{2L_f} \|\nabla f(\mathbf{x}^{k,j}) - \nabla f(\mathbf{x}^{k,j+1})\|^2.$$

Summing the above inequality over $j = 0, 1, \ldots, p-1$, we obtain that

$$F(\mathbf{x}^{k}) - F(\mathbf{x}^{k+1}) \geq \frac{1}{2L_f} \sum_{j=0}^{p-1} \|\nabla f(\mathbf{x}^{k,j}) - \nabla f(\mathbf{x}^{k,j+1})\|^2. \tag{14.19}$$

On the other hand, for any $k \geq 0$,

$$\begin{aligned}
F(\mathbf{x}^{k+1}) - F(\mathbf{x}^{*}) &= f(\mathbf{x}^{k+1}) - f(\mathbf{x}^{*}) + g(\mathbf{x}^{k+1}) - g(\mathbf{x}^{*}) \\
&\leq \langle \nabla f(\mathbf{x}^{k+1}), \mathbf{x}^{k+1} - \mathbf{x}^{*} \rangle + g(\mathbf{x}^{k+1}) - g(\mathbf{x}^{*}) \\
&= \sum_{j=0}^{p-1} \left[\langle \nabla_{j+1} f(\mathbf{x}^{k+1}), \mathbf{x}_{j+1}^{k+1} - \mathbf{x}_{j+1}^{*} \rangle + (g_{j+1}(\mathbf{x}_{j+1}^{k+1}) - g_{j+1}(\mathbf{x}_{j+1}^{*})) \right] \\
&= \sum_{j=0}^{p-1} \left[\langle \nabla_{j+1} f(\mathbf{x}^{k,j+1}), \mathbf{x}_{j+1}^{k+1} - \mathbf{x}_{j+1}^{*} \rangle + (g_{j+1}(\mathbf{x}_{j+1}^{k+1}) - g_{j+1}(\mathbf{x}_{j+1}^{*})) \right] \\
&\quad + \sum_{j=0}^{p-1} \langle \nabla_{j+1} f(\mathbf{x}^{k+1}) - \nabla_{j+1} f(\mathbf{x}^{k,j+1}), \mathbf{x}_{j+1}^{k+1} - \mathbf{x}_{j+1}^{*} \rangle \\
&\leq \sum_{j=0}^{p-1} \langle \nabla_{j+1} f(\mathbf{x}^{k+1}) - \nabla_{j+1} f(\mathbf{x}^{k,j+1}), \mathbf{x}_{j+1}^{k+1} - \mathbf{x}_{j+1}^{*} \rangle,
\end{aligned} \tag{14.20}$$

where the first inequality follows by the gradient inequality employed on the function f, and the second inequality follows by the relation (14.18). Using the Cauchy–Schwarz and triangle inequalities, we can continue (14.20) and obtain that

$$F(\mathbf{x}^{k+1}) - F(\mathbf{x}^{*}) \leq \sum_{j=0}^{p-1} \|\nabla_{j+1} f(\mathbf{x}^{k+1}) - \nabla_{j+1} f(\mathbf{x}^{k,j+1})\| \cdot \|\mathbf{x}_{j+1}^{k+1} - \mathbf{x}_{j+1}^{*}\|. \tag{14.21}$$

Note that

$$\begin{aligned}
\|\nabla_{j+1} f(\mathbf{x}^{k+1}) - \nabla_{j+1} f(\mathbf{x}^{k,j+1})\| &\leq \|\nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^{k,j+1})\| \\
&\leq \sum_{t=j+1}^{p-1} \|\nabla f(\mathbf{x}^{k,t}) - \nabla f(\mathbf{x}^{k,t+1})\| \\
&\leq \sum_{t=0}^{p-1} \|\nabla f(\mathbf{x}^{k,t}) - \nabla f(\mathbf{x}^{k,t+1})\|,
\end{aligned}$$

which, combined with (14.21), yields the inequality

$$F(\mathbf{x}^{k+1}) - F(\mathbf{x}^*) \leq \left(\sum_{t=0}^{p-1} \|\nabla f(\mathbf{x}^{k,t}) - \nabla f(\mathbf{x}^{k,t+1})\| \right) \left(\sum_{j=0}^{p-1} \|\mathbf{x}_{j+1}^{k+1} - \mathbf{x}_{j+1}^*\| \right).$$

Taking the square of both sides and using (14.16), we obtain

$$\begin{aligned}
(F(\mathbf{x}^{k+1}) - F(\mathbf{x}^*))^2 &\leq \left(\sum_{t=0}^{p-1} \|\nabla f(\mathbf{x}^{k,t}) - \nabla f(\mathbf{x}^{k,t+1})\| \right)^2 \left(\sum_{j=0}^{p-1} \|\mathbf{x}_{j+1}^{k+1} - \mathbf{x}_{j+1}^*\| \right)^2 \\
&\leq p^2 \left(\sum_{t=0}^{p-1} \|\nabla f(\mathbf{x}^{k,t}) - \nabla f(\mathbf{x}^{k,t+1})\|^2 \right) \left(\sum_{j=0}^{p-1} \|\mathbf{x}_{j+1}^{k+1} - \mathbf{x}_{j+1}^*\|^2 \right) \\
&= p^2 \left(\sum_{t=0}^{p-1} \|\nabla f(\mathbf{x}^{k,t}) - \nabla f(\mathbf{x}^{k,t+1})\|^2 \right) \|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 \\
&\leq p^2 R^2 \sum_{t=0}^{p-1} \|\nabla f(\mathbf{x}^{k,t}) - \nabla f(\mathbf{x}^{k,t+1})\|^2. \qquad (14.22)
\end{aligned}$$

We can thus conclude by (14.19) and (14.22) that for any $k \geq 0$,

$$(F(\mathbf{x}^{k+1}) - F_{\text{opt}})^2 \leq 2L_f p^2 R^2 (F(\mathbf{x}^k) - F(\mathbf{x}^{k+1})).$$

Denoting $a_k = F(\mathbf{x}^k) - F_{\text{opt}}$, the last inequality can be rewritten as

$$a_k - a_{k+1} \geq \frac{1}{\gamma} a_{k+1}^2,$$

where $\gamma = 2L_f p^2 R^2$. Invoking Lemma 11.17, we obtain that for all $k \geq 2$,

$$a_k \leq \max \left\{ \left(\frac{1}{2} \right)^{(k-1)/2} a_0, \frac{8L_f p^2 R^2}{k-1} \right\},$$

which is the desired result (14.15). □

14.5.2 $p = 2$

The dependency of the efficiency estimate (14.15) on the global Lipschitz constant L_f is problematic since it might be a very large number. We will now develop a different line of analysis in the case where there are only two blocks ($p = 2$). The new analysis will produce an improved efficiency estimate that depends on the smallest block Lipschitz constant rather than on L_f. The general model (14.9) in the case $p = 2$ amounts to

$$\min_{\mathbf{x}_1 \in \mathbb{E}_1, \mathbf{x}_2 \in \mathbb{E}_2} \{F(\mathbf{x}_1, \mathbf{x}_2) \equiv f(\mathbf{x}_1, \mathbf{x}_2) + g_1(\mathbf{x}_1) + g_2(\mathbf{x}_2)\}. \qquad (14.23)$$

As usual, we use the notation $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ and $g(\mathbf{x}) = g_1(\mathbf{x}_1) + g_2(\mathbf{x}_2)$. We gather below the required assumptions.

Assumption 14.12.

(A) *For $i \in \{1,2\}$, the function $g_i : \mathbb{E}_i \to (-\infty,\infty]$ is proper closed and convex.*

(B) *$f : \mathbb{E} \to \mathbb{R}$ is convex. In addition, f is differentiable over an open set containing* $\mathrm{dom}(g)$.

(C) *For any $i \in \{1,2\}$ the gradient of f is Lipschitz continuous w.r.t. $\mathbf{x}_i$ over* $\mathrm{dom}(g_i)$ *with constant $L_i \in (0,\infty)$, meaning that*

$$\|\nabla_1 f(\mathbf{x}_1+\mathbf{d}_1,\mathbf{x}_2)-\nabla_1 f(\mathbf{x}_1,\mathbf{x}_2)\| \le L_1\|\mathbf{d}_1\|,$$
$$\|\nabla_2 f(\mathbf{x}_1,\mathbf{x}_2+\mathbf{d}_2)-\nabla_2 f(\mathbf{x}_1,\mathbf{x}_2)\| \le L_2\|\mathbf{d}_2\|$$

for any $\mathbf{x}_1 \in \mathrm{dom}(g_1), \mathbf{x}_2 \in \mathrm{dom}(g_2)$, and $\mathbf{d}_1 \in \mathbb{E}_1, \mathbf{d}_2 \in \mathbb{E}_2$ such that $\mathbf{x}+\mathbf{d}_1 \in \mathrm{dom}(g_1), \mathbf{x}_2+\mathbf{d}_2 \in \mathrm{dom}(g_2)$.

(D) *The optimal set of* (14.23)*, denoted by X^*, is nonempty, and the corresponding optimal value is denoted by F_{opt}.*

(E) *For any $\alpha > 0$, there exists $R_\alpha > 0$ such that*

$$\max_{\mathbf{x},\mathbf{x}^* \in \mathbb{E}} \{\|\mathbf{x}-\mathbf{x}^*\| : F(\mathbf{x}) \le \alpha, \mathbf{x}^* \in X^*\} \le R_\alpha.$$

The alternating minimization method for solving problem (14.23) is described below.

The Alternating Minimization Method

Initialization: $\mathbf{x}_1^0 \in \mathrm{dom}(g_1), \mathbf{x}_2^0 \in \mathrm{dom}(g_2)$ such that

$$\mathbf{x}_2^0 \in \mathrm{argmin}_{\mathbf{x}_2 \in \mathbb{E}_2} f(\mathbf{x}_1^0,\mathbf{x}_2)+g_2(\mathbf{x}_2).$$

General step ($k = 0,1,\ldots$):

$$\mathbf{x}_1^{k+1} \in \mathrm{argmin}_{\mathbf{x}_1 \in \mathbb{E}_1} f(\mathbf{x}_1,\mathbf{x}_2^k)+g_1(\mathbf{x}_1), \tag{14.24}$$
$$\mathbf{x}_2^{k+1} \in \mathrm{argmin}_{\mathbf{x}_2 \in \mathbb{E}_2} f(\mathbf{x}_1^{k+1},\mathbf{x}_2)+g_2(\mathbf{x}_2). \tag{14.25}$$

Note that, as opposed to the description of the method so far, we assume that "half" an iteration was performed prior to the first iteration (that is, $\mathbf{x}_2^0 \in \mathrm{argmin}_{\mathbf{x}_2 \in \mathbb{E}_2} f(\mathbf{x}_1^0,\mathbf{x}_2)+g_2(\mathbf{x}_2)$). We will also utilize the auxiliary sequence $\{\mathbf{x}^{k,1}\}_{k\ge 0}$ as defined in (14.2) but use the following simpler notation:

$$\mathbf{x}^{k+\frac{1}{2}} = (\mathbf{x}_1^{k+1},\mathbf{x}_2^k).$$

We will adopt the notation used in Section 11.3.1 and consider for any $M > 0$ the partial prox-grad mappings

$$T_M^i(\mathbf{x}) = \mathrm{prox}_{\frac{1}{M}g_i}\left(\mathbf{x}_i - \frac{1}{M}\nabla_i f(\mathbf{x})\right), \quad i = 1,2,$$

as well as the partial gradient mappings

$$G_M^i(\mathbf{x}) = M\left(\mathbf{x}_i - T_M^i(\mathbf{x})\right), \quad i = 1, 2.$$

Obviously, for any $M > 0$,

$$T_M(\mathbf{x}) = (T_M^1(\mathbf{x}), T_M^2(\mathbf{x})),\ G_M(\mathbf{x}) = (G_M^1(\mathbf{x}), G_M^2(\mathbf{x})),$$

and from the definition of the alternating minimization method we have for all $k \geq 0$,

$$G_M^1(\mathbf{x}^{k+\frac{1}{2}}) = \mathbf{0}, G_M^2(\mathbf{x}^k) = \mathbf{0}. \tag{14.26}$$

We begin by proving the following sufficient decrease-type result.

Lemma 14.13. *Suppose that Assumption* 14.12 *holds. Let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by the alternating minimization method for solving problem* (14.23). *Then for any* $k \geq 0$ *the following inequalities hold:*

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+\frac{1}{2}}) \geq \frac{1}{2L_1}\|G_{L_1}^1(\mathbf{x}^k)\|^2, \tag{14.27}$$

$$F(\mathbf{x}^{k+\frac{1}{2}}) - F(\mathbf{x}^{k+1}) \geq \frac{1}{2L_2}\|G_{L_2}^2(\mathbf{x}^{k+\frac{1}{2}})\|^2. \tag{14.28}$$

Proof. Invoking the block sufficient decrease lemma (Lemma 11.9) with $\mathbf{x} = \mathbf{x}^k$ and $i = 1$, we obtain

$$F(\mathbf{x}_1^k, \mathbf{x}_2^k) - F(T_{L_1}^1(\mathbf{x}^k), \mathbf{x}_2^k) \geq \frac{1}{2L_1}\|G_{L_1}^1(\mathbf{x}_1^k, \mathbf{x}_2^k)\|^2.$$

The inequality (14.27) now follows from the inequality $F(\mathbf{x}^{k+\frac{1}{2}}) \leq F(T_{L_1}^1(\mathbf{x}^k)), \mathbf{x}_2^k)$. The inequality (14.28) follows by invoking the block sufficient decrease lemma with $\mathbf{x} = \mathbf{x}^{k+\frac{1}{2}}$, $i = 2$, and using the inequality $F(\mathbf{x}^{k+1}) \leq F(\mathbf{x}_1^{k+1}, T_{L_2}^2(\mathbf{x}^{k+\frac{1}{2}}))$. ∎

The next lemma establishes an upper bound on the distance in function values of the iterates of the method.

Lemma 14.14. *Let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by the alternating minimization method for solving problem* (14.23). *Then for any* $\mathbf{x}^* \in X^*$ *and* $k \geq 0$,

$$F(\mathbf{x}^{k+\frac{1}{2}}) - F(\mathbf{x}^*) \leq \|G_{L_1}^1(\mathbf{x}^k)\| \cdot \|\mathbf{x}^k - \mathbf{x}^*\|, \tag{14.29}$$

$$F(\mathbf{x}^{k+1}) - F(\mathbf{x}^*) \leq \|G_{L_2}^2(\mathbf{x}^{k+\frac{1}{2}})\| \cdot \|\mathbf{x}^{k+\frac{1}{2}} - \mathbf{x}^*\|. \tag{14.30}$$

Proof. Note that

$$T_{L_1}(\mathbf{x}^k) = (T_{L_1}^1(\mathbf{x}^k), T_{L_1}^2(\mathbf{x}^k)) = \left(T_{L_1}^1(\mathbf{x}^k), \mathbf{x}_2^k - \frac{1}{L_1}G_{L_1}^2(\mathbf{x}^k)\right) = (T_{L_1}^1(\mathbf{x}^k), \mathbf{x}_2^k),$$

where in the last equality we used (14.26). Combining this with the block descent lemma (Lemma 11.8), we obtain that

$$\begin{aligned} f(T_{L_1}(\mathbf{x}^k)) - f(\mathbf{x}^*) &\leq f(\mathbf{x}^k) + \langle \nabla_1 f(\mathbf{x}^k), T^1_{L_1}(\mathbf{x}^k) - \mathbf{x}^k_1 \rangle \\ &\quad + \frac{L_1}{2}\|T^1_{L_1}(\mathbf{x}^k) - \mathbf{x}^k_1\|^2 - f(\mathbf{x}^*) \\ &= f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), T_{L_1}(\mathbf{x}^k) - \mathbf{x}^k \rangle \\ &\quad + \frac{L_1}{2}\|T^1_{L_1}(\mathbf{x}^k) - \mathbf{x}^k_1\|^2 - f(\mathbf{x}^*). \end{aligned} \tag{14.31}$$

Since f is convex, it follows that $f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^* \rangle$, which, combined with (14.31), yields

$$f(T_{L_1}(\mathbf{x}^k)) - f(\mathbf{x}^*) \leq \langle \nabla f(\mathbf{x}^k), T_{L_1}(\mathbf{x}^k) - \mathbf{x}^* \rangle + \frac{L_1}{2}\|T^1_{L_1}(\mathbf{x}^k) - \mathbf{x}^k_1\|^2. \tag{14.32}$$

Since $T_{L_1}(\mathbf{x}^k) = \operatorname{prox}_{\frac{1}{L_1}g}\big(\mathbf{x}^k - \frac{1}{L_1}\nabla f(\mathbf{x}^k)\big)$, then by invoking the second prox theorem (Theorem 6.39) with $f = \frac{1}{L_1}g, \mathbf{x} = \mathbf{x}^k - \frac{1}{L_1}\nabla f(\mathbf{x}^k)$, and $\mathbf{y} = \mathbf{x}^*$, we have

$$g(T_{L_1}(\mathbf{x}^k)) - g(\mathbf{x}^*) \leq L_1 \left\langle \mathbf{x}^k - \frac{1}{L_1}\nabla f(\mathbf{x}^k) - T_{L_1}(\mathbf{x}^k), T_{L_1}(\mathbf{x}^k) - \mathbf{x}^* \right\rangle. \tag{14.33}$$

Combining inequalities (14.32) and (14.33), along with the fact that $F(\mathbf{x}^{k+\frac{1}{2}}) \leq F(T^1_{L_1}(\mathbf{x}^k), \mathbf{x}^k_2) = F(T_{L_1}(\mathbf{x}^k))$, we finally have

$$\begin{aligned} F(\mathbf{x}^{k+\frac{1}{2}}) - F(\mathbf{x}^*) &\leq F(T_{L_1}(\mathbf{x}^k)) - F(\mathbf{x}^*) \\ &= f(T_{L_1}(\mathbf{x}^k)) + g(T_{L_1}(\mathbf{x}^k)) - f(\mathbf{x}^*) - g(\mathbf{x}^*) \\ &\leq L_1\langle \mathbf{x}^k - T_{L_1}(\mathbf{x}^k), T_{L_1}(\mathbf{x}^k) - \mathbf{x}^* \rangle + \frac{L_1}{2}\|T^1_{L_1}(\mathbf{x}^k) - \mathbf{x}^k_1\|^2 \\ &= \langle G_{L_1}(\mathbf{x}^k), T_{L_1}(\mathbf{x}^k) - \mathbf{x}^* \rangle + \frac{1}{2L_1}\|G_{L_1}(\mathbf{x}^k)\|^2 \\ &= \langle G_{L_1}(\mathbf{x}^k), T_{L_1}(\mathbf{x}^k) - \mathbf{x}^k \rangle + \langle G_{L_1}(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^* \rangle + \frac{1}{2L_1}\|G_{L_1}(\mathbf{x}^k)\|^2 \\ &= -\frac{1}{L_1}\|G_{L_1}(\mathbf{x}^k)\|^2 + \langle G_{L_1}(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^* \rangle + \frac{1}{2L_1}\|G_{L_1}(\mathbf{x}^k)\|^2 \\ &\leq \langle G_{L_1}(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^* \rangle \\ &\leq \|G_{L_1}(\mathbf{x}^k)\| \cdot \|\mathbf{x}^k - \mathbf{x}^*\| \\ &= \|G^1_{L_1}(\mathbf{x}^k)\| \cdot \|\mathbf{x}^k - \mathbf{x}^*\|, \end{aligned}$$

establishing (14.29). The inequality (14.30) follows by using the same argument but on the sequence generated by the alternating minimization method with starting point $(\mathbf{x}^1_1, \mathbf{x}^0_2)$ and assuming that the first index to be updated is $i = 2$. ∎

With the help of Lemmas 14.13 and 14.14, we can prove a sublinear rate of convergence of the alternating minimization method with an improved constant.

Theorem 14.15 ($O(1/k)$ rate of alternating minimization—improved result). *Suppose that Assumption* 14.12 *holds, and let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by the alternating minimization method for solving problem* (14.23). *Then*

for all $k \geq 2$,

$$F(\mathbf{x}^k) - F_{\text{opt}} \leq \max\left\{\left(\frac{1}{2}\right)^{(k-1)/2}(F(\mathbf{x}^0) - F_{\text{opt}}), \frac{8\min\{L_1, L_2\}R^2}{k-1}\right\}, \quad (14.34)$$

where $R = R_{F(\mathbf{x}^0)}$.

Proof. By Lemma 14.14 and Assumption 14.12(E),

$$F(\mathbf{x}^{k+\frac{1}{2}}) - F_{\text{opt}} \leq \|G^1_{L_1}(\mathbf{x}^k)\|R,$$

where $R = R_{F(\mathbf{x}^0)}$. Now, by Lemma 14.13,

$$\begin{aligned} F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq F(\mathbf{x}^k) - F(\mathbf{x}^{k+\frac{1}{2}}) &\geq \frac{1}{2L_1}\|G^1_{L_1}(\mathbf{x}^k)\|^2 \\ &\geq \frac{(F(\mathbf{x}^{k+\frac{1}{2}}) - F_{\text{opt}})^2}{2L_1R^2} \\ &\geq \frac{1}{2L_1R^2}(F(\mathbf{x}^{k+1}) - F_{\text{opt}})^2. \end{aligned} \quad (14.35)$$

Similarly, by Lemma 14.14 and Assumption 14.12(E),

$$F(\mathbf{x}^{k+1}) - F_{\text{opt}} \leq \|G^2_{L_2}(\mathbf{x}^{k+\frac{1}{2}})\|R.$$

Thus, utilizing Lemma 14.13 we obtain

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq F(\mathbf{x}^{k+\frac{1}{2}}) - F(\mathbf{x}^{k+1}) \geq \frac{1}{2L_2}\|G^2_{L_2}(\mathbf{x}^{k+\frac{1}{2}})\|^2 \geq \frac{(F(\mathbf{x}^{k+1}) - F_{\text{opt}})^2}{2L_2R^2},$$

which, combined with (14.35), yields the inequality

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq \frac{1}{2\min\{L_1, L_2\}R^2}(F(\mathbf{x}^{k+1}) - F_{\text{opt}})^2. \quad (14.36)$$

Denoting $a_k = F(\mathbf{x}^k) - F_{\text{opt}}$ and $\gamma = 2\min\{L_1, L_2\}R^2$, we obtain that for all $k \geq 0$,

$$a_k - a_{k+1} \geq \frac{1}{\gamma}a^2_{k+1},$$

and thus, by Lemma 11.17, it holds that for all $k \geq 2$

$$a_k \leq \max\left\{\left(\frac{1}{2}\right)^{(k-1)/2} a_0, \frac{8\min\{L_1, L_2\}R^2}{k-1}\right\},$$

which is the desired result (14.34). ∎

Remark 14.16. *Note that the constant in the efficiency estimate* (14.34) *depends on* $\min\{L_1, L_2\}$. *This means that the rate of convergence of the alternating minimization method in the case of two blocks is dictated by the smallest block Lipschitz constant, meaning by the smoother part of the function. This is not the case for the efficiency estimate obtained in Theorem* 14.11 *for the convergence of alternating minimization with an arbitrary number of blocks, which depends on the global Lipschitz constant* L_f *and is thus dictated by the "worst" block w.r.t. the level of smoothness.*

Chapter 15
ADMM

Underlying Spaces: In this chapter all the underlying spaces are Euclidean $\mathbb{R}^n$ spaces endowed with the dot product and the l_2-norm.

15.1 The Augmented Lagrangian Method

Consider the problem

$$H_{\text{opt}} = \min\{H(\mathbf{x},\mathbf{z}) \equiv h_1(\mathbf{x}) + h_2(\mathbf{z}) : \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c}\}, \tag{15.1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{m\times p}$, and $\mathbf{c} \in \mathbb{R}^m$. For now, we will assume that h_1 and h_2 are proper closed and convex functions. Later on, we will specify exact conditions on the data $(h_1, h_2, \mathbf{A}, \mathbf{B}, \mathbf{c})$ that will guarantee the validity of some convergence results. To find a dual problem of (15.1), we begin by constructing a Lagrangian:

$$L(\mathbf{x},\mathbf{z};\mathbf{y}) = h_1(\mathbf{x}) + h_2(\mathbf{z}) + \langle \mathbf{y}, \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}\rangle.$$

The dual objective function is therefore given by

$$\begin{aligned} q(\mathbf{y}) &= \min_{\mathbf{x}\in\mathbb{R}^n, \mathbf{z}\in\mathbb{R}^p} \{h_1(\mathbf{x}) + h_2(\mathbf{z}) + \langle \mathbf{y}, \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}\rangle\} \\ &= -h_1^*(-\mathbf{A}^T\mathbf{y}) - h_2^*(-\mathbf{B}^T\mathbf{y}) - \langle \mathbf{c}, \mathbf{y}\rangle, \end{aligned}$$

and the dual problem is given by

$$q_{\text{opt}} = \max_{\mathbf{y}\in\mathbb{R}^m} \{-h_1^*(-\mathbf{A}^T\mathbf{y}) - h_2^*(-\mathbf{B}^T\mathbf{y}) - \langle \mathbf{c}, \mathbf{y}\rangle\} \tag{15.2}$$

or, in minimization form, by

$$\min_{\mathbf{y}\in\mathbb{R}^m} \left\{h_1^*(-\mathbf{A}^T\mathbf{y}) + h_2^*(-\mathbf{B}^T\mathbf{y}) + \langle \mathbf{c}, \mathbf{y}\rangle\right\}. \tag{15.3}$$

The proximal point method was discussed in Section 10.5, where its convergence was established. The general update step of the proximal point method employed on problem (15.3) takes the form ($\rho > 0$ being a given constant)

$$\mathbf{y}^{k+1} = \text{argmin}_{\mathbf{y}\in\mathbb{R}^m} \left\{h_1^*(-\mathbf{A}^T\mathbf{y}) + h_2^*(-\mathbf{B}^T\mathbf{y}) + \langle \mathbf{c}, \mathbf{y}\rangle + \frac{1}{2\rho}\|\mathbf{y} - \mathbf{y}^k\|^2\right\}. \tag{15.4}$$

Assuming that the sum and affine rules of subdifferential calculus (Theorems 3.40 and 3.43) hold for the relevant functions, we can conclude by Fermat's optimality condition (Theorem 3.63) that (15.4) holds if and only if

$$\mathbf{0} \in -\mathbf{A}\partial h_1^*(-\mathbf{A}^T\mathbf{y}^{k+1}) - \mathbf{B}\partial h_2^*(-\mathbf{B}^T\mathbf{y}^{k+1}) + \mathbf{c} + \frac{1}{\rho}(\mathbf{y}^{k+1} - \mathbf{y}^k). \tag{15.5}$$

Using the conjugate subgradient theorem (Corollary 4.21), we obtain that $\mathbf{y}^{k+1}$ satisfies (15.5) if and only if $\mathbf{y}^{k+1} = \mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^{k+1} - \mathbf{c})$, where $\mathbf{x}^{k+1}$ and $\mathbf{z}^{k+1}$ satisfy

$$\begin{aligned}\mathbf{x}^{k+1} &\in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^n}\{\langle \mathbf{A}^T\mathbf{y}^{k+1}, \mathbf{x}\rangle + h_1(\mathbf{x})\},\\ \mathbf{z}^{k+1} &\in \operatorname{argmin}_{\mathbf{z}\in\mathbb{R}^p}\{\langle \mathbf{B}^T\mathbf{y}^{k+1}, \mathbf{z}\rangle + h_2(\mathbf{z})\}.\end{aligned}$$

Plugging the update equation for $\mathbf{y}^{k+1}$ into the above, we conclude that $\mathbf{y}^{k+1}$ satisfies (15.5) if and only if

$$\begin{aligned}\mathbf{y}^{k+1} &= \mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^{k+1} - \mathbf{c}),\\ \mathbf{x}^{k+1} &\in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^n}\{\langle \mathbf{A}^T(\mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^{k+1} - \mathbf{c})), \mathbf{x}\rangle + h_1(\mathbf{x})\},\\ \mathbf{z}^{k+1} &\in \operatorname{argmin}_{\mathbf{z}\in\mathbb{R}^p}\{\langle \mathbf{B}^T(\mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^{k+1} - \mathbf{c})), \mathbf{z}\rangle + h_2(\mathbf{z})\},\end{aligned}$$

meaning if and only if (using the properness and convexity of h_1 and h_2, as well as Fermat's optimality condition)

$$\mathbf{y}^{k+1} = \mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^{k+1} - \mathbf{c}), \tag{15.6}$$

$$\mathbf{0} \in \mathbf{A}^T(\mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^{k+1} - \mathbf{c})) + \partial h_1(\mathbf{x}^{k+1}), \tag{15.7}$$

$$\mathbf{0} \in \mathbf{B}^T(\mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^{k+1} - \mathbf{c})) + \partial h_2(\mathbf{z}^{k+1}). \tag{15.8}$$

Conditions (15.7) and (15.8) are satisfied if and only if $(\mathbf{x}^{k+1}, \mathbf{z}^{k+1})$ is a coordinate-wise minimum (see Definition 14.2) of the function

$$\tilde{H}(\mathbf{x}, \mathbf{z}) \equiv h_1(\mathbf{x}) + h_2(\mathbf{z}) + \frac{\rho}{2}\left\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k\right\|^2.$$

By Lemma 14.7, coordinate-wise minima points of $\tilde{H}$ are exactly the minimizers of $\tilde{H}$, and therefore the system (15.6), (15.7), (15.8) leads us to the following primal representation of the dual proximal point method, known as the *augmented Lagrangian method*.

The Augmented Lagrangian Method

Initialization: $\mathbf{y}^0 \in \mathbb{R}^m$, $\rho > 0$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

$$(\mathbf{x}^{k+1}, \mathbf{z}^{k+1}) \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n, \mathbf{z}\in\mathbb{R}^p}\left\{h_1(\mathbf{x}) + h_2(\mathbf{z}) + \frac{\rho}{2}\left\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k\right\|^2\right\} \tag{15.9}$$

$$\mathbf{y}^{k+1} = \mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^{k+1} - \mathbf{c}). \tag{15.10}$$

Naturally, step (15.9) is called the *primal update step*, while (15.10) is the *dual update step.*

Remark 15.1 (augmented Lagrangian). *The augmented Lagrangian associated with the main problem* (15.1) *is defined to be*

$$L_\rho(\mathbf{x}, \mathbf{z}; \mathbf{y}) = h_1(\mathbf{x}) + h_2(\mathbf{z}) + \langle \mathbf{y}, \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c} \rangle + \frac{\rho}{2} \|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}\|^2.$$

Obviously, $L_0 = L$ is the Lagrangian function, and L_ρ for $\rho > 0$ can be considered as a penalized version of the Lagrangian. The primal update step (15.9) *can be equivalently written as*

$$(\mathbf{x}^{k+1}, \mathbf{z}^{k+1}) \in \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^n, \mathbf{z} \in \mathbb{R}^p} L_\rho(\mathbf{x}, \mathbf{z}; \mathbf{y}^k).$$

The above representation of the primal update step as the outcome of the minimization of the augmented Lagrangian function is the reason for the name of the method.

15.2 Alternating Direction Method of Multipliers (ADMM)

The augmented Lagrangian method is in general not an implementable method since the primal update step (15.9) can be as hard to solve as the original problem. One source of difficulty is the coupling term between the $\mathbf{x}$ and the $\mathbf{z}$ variables, which is of the form $\rho(\mathbf{x}^T \mathbf{A}^T \mathbf{B}\mathbf{z})$. The approach used in the *alternating direction method of multipliers* (ADMM) to tackle this difficulty is to replace the exact minimization in the primal update step (15.9) by one iteration of the alternating minimization method; that is, the objective function of (15.9) is first minimized w.r.t. $\mathbf{x}$, and then w.r.t. $\mathbf{z}$.

ADMM

Initialization: $\mathbf{x}^0 \in \mathbb{R}^n$, $\mathbf{z}^0 \in \mathbb{R}^p$, $\mathbf{y}^0 \in \mathbb{R}^m$, $\rho > 0$.
General step: for any $k = 0, 1, \ldots$ execute the following:

(a) $\mathbf{x}^{k+1} \in \operatorname{argmin}_{\mathbf{x}} \left\{ h_1(\mathbf{x}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}^k - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k \right\|^2 \right\}$;

(b) $\mathbf{z}^{k+1} \in \operatorname{argmin}_{\mathbf{z}} \left\{ h_2(\mathbf{z}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z} - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k \right\|^2 \right\}$;

(c) $\mathbf{y}^{k+1} = \mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^{k+1} - \mathbf{c})$.

15.2.1 Alternating Direction Proximal Method of Multipliers (AD-PMM)

We will actually analyze a more general method than ADMM in which a quadratic proximity term is added to the objective in the minimization problems of steps

(a) and (b). We will assume that we are given two positive semidefinite matrices $\mathbf{G} \in \mathbb{S}^n_+, \mathbf{Q} \in \mathbb{S}^p_+$, and recall that $\|\mathbf{x}\|^2_{\mathbf{G}} = \mathbf{x}^T\mathbf{G}\mathbf{x}, \|\mathbf{x}\|^2_{\mathbf{Q}} = \mathbf{x}^T\mathbf{Q}\mathbf{x}$.

AD-PMM

Initialization: $\mathbf{x}^0 \in \mathbb{R}^n$, $\mathbf{z}^0 \in \mathbb{R}^p$, $\mathbf{y}^0 \in \mathbb{R}^m$, $\rho > 0$.
General step: for any $k = 0, 1, \ldots$ execute the following:

(a) $\mathbf{x}^{k+1} \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^n} \left\{ h_1(\mathbf{x}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}^k - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k \right\|^2 + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathbf{G}} \right\}$;

(b) $\mathbf{z}^{k+1} \in \operatorname{argmin}_{\mathbf{z}\in\mathbb{R}^p} \left\{ h_2(\mathbf{z}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z} - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k \right\| + \frac{1}{2}\|\mathbf{z} - \mathbf{z}^k\|^2_{\mathbf{Q}} \right\}$;

(c) $\mathbf{y}^{k+1} = \mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^{k+1} - \mathbf{c})$.

One important motivation for considering AD-PMM is that by using the proximity terms, the minimization problems in steps (a) and (b) of ADMM can be simplified considerably by choosing $\mathbf{G} = \alpha\mathbf{I} - \rho\mathbf{A}^T\mathbf{A}$ with $\alpha \geq \rho\lambda_{\max}(\mathbf{A}^T\mathbf{A})$ and $\mathbf{Q} = \beta\mathbf{I} - \rho\mathbf{B}^T\mathbf{B}$ with $\beta \geq \rho\lambda_{\max}(\mathbf{B}^T\mathbf{B})$. Then obviously $\mathbf{G}, \mathbf{Q} \in \mathbb{S}^n_+$, and the function that needs to be minimized in the $\mathbf{x}$-step can be simplified as follows:

$$
\begin{aligned}
& h_1(\mathbf{x}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}^k - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k \right\|^2 + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathbf{G}} \\
&= h_1(\mathbf{x}) + \frac{\rho}{2} \left\| \mathbf{A}(\mathbf{x} - \mathbf{x}^k) + \mathbf{A}\mathbf{x}^k + \mathbf{B}\mathbf{z}^k - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k \right\|^2 + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathbf{G}} \\
&= h_1(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{A}(\mathbf{x} - \mathbf{x}^k)\|^2 + \left\langle \rho\mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{x}^k + \mathbf{B}\mathbf{z}^k - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k \right\rangle \\
&\quad + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{x}^k\|^2 - \frac{\rho}{2}\|\mathbf{A}(\mathbf{x} - \mathbf{x}^k)\|^2 + \text{constant} \\
&= h_1(\mathbf{x}) + \rho \left\langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{x}^k + \mathbf{B}\mathbf{z}^k - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k \right\rangle + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{x}^k\|^2 + \text{constant},
\end{aligned}
$$

where by "constant" we mean a term that does not depend on $\mathbf{x}$. We can therefore conclude that step (a) of AD-PMM amounts to

$$
\mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^n} \left\{ h_1(\mathbf{x}) + \rho \left\langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{x}^k + \mathbf{B}\mathbf{z}^k - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k \right\rangle + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{x}^k\|^2 \right\}, \tag{15.11}
$$

and, similarly, step (b) of AD-PMM is the same as

$$
\mathbf{z}^{k+1} = \operatorname{argmin}_{\mathbf{z}\in\mathbb{R}^p} \left\{ h_2(\mathbf{z}) + \rho \left\langle \mathbf{B}\mathbf{z}, \mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^k - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k \right\rangle + \frac{\beta}{2}\|\mathbf{z} - \mathbf{z}^k\|^2 \right\}. \tag{15.12}
$$

The functions minimized in the update formulas (15.11) and (15.12) are actually constructed from the functions minimized in steps (a) and (b) of ADMM by linearizing the quadratic term and adding a proximity term. This is the reason why the resulting method will be called the *alternating direction linearized proximal method of multipliers* (AD-LPMM). We can also write the update formulas (15.11) and

(15.12) in terms of proximal operators. Indeed, (15.11) can be rewritten equivalently as

$$\mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{x}} \left\{ \frac{1}{\alpha} h_1(\mathbf{x}) + \frac{1}{2} \left\| \mathbf{x} - \left(\mathbf{x}^k - \frac{\rho}{\alpha} \mathbf{A}^T \left(\mathbf{A}\mathbf{x}^k + \mathbf{B}\mathbf{z}^k - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k \right) \right) \right\|^2 \right\}.$$

That is,

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\frac{1}{\alpha}h_1} \left[\mathbf{x}^k - \frac{\rho}{\alpha} \mathbf{A}^T \left(\mathbf{A}\mathbf{x}^k + \mathbf{B}\mathbf{z}^k - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k \right) \right].$$

Similarly, the $\mathbf{z}$-step can be rewritten as

$$\mathbf{z}^{k+1} = \operatorname{prox}_{\frac{1}{\beta}h_2} \left[\mathbf{z}^k - \frac{\rho}{\beta} \mathbf{B}^T \left(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^k - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k \right) \right].$$

We can now summarize and write explicitly the AD-LPMM method.

AD-LPMM

Initialization: $\mathbf{x}^0 \in \mathbb{R}^n$, $\mathbf{z}^0 \in \mathbb{R}^p$, $\mathbf{y}^0 \in \mathbb{R}^m$, $\rho > 0$, $\alpha \geq \rho\lambda_{\max}(\mathbf{A}^T\mathbf{A})$, $\beta \geq \rho\lambda_{\max}(\mathbf{B}^T\mathbf{B})$.
General step: for any $k = 0, 1, \ldots$ execute the following:

(a) $\mathbf{x}^{k+1} = \operatorname{prox}_{\frac{1}{\alpha}h_1} \left[\mathbf{x}^k - \frac{\rho}{\alpha} \mathbf{A}^T \left(\mathbf{A}\mathbf{x}^k + \mathbf{B}\mathbf{z}^k - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k \right) \right]$;

(b) $\mathbf{z}^{k+1} = \operatorname{prox}_{\frac{1}{\beta}h_2} \left[\mathbf{z}^k - \frac{\rho}{\beta} \mathbf{B}^T \left(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^k - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k \right) \right]$;

(c) $\mathbf{y}^{k+1} = \mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^{k+1} - \mathbf{c})$.

15.3 Convergence Analysis of AD-PMM

In this section we will develop a rate of convergence analysis of AD-PMM employed on problem (15.1). Note that both ADMM and AD-LPMM are special cases of AD-PMM. The following set of assumptions will be made.

Assumption 15.2.

(A) $h_1 : \mathbb{R}^n \to (-\infty, \infty]$ *and* $h_2 : \mathbb{R}^p \to (-\infty, \infty]$ *are proper closed convex functions.*

(B) $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{m\times p}$, $\mathbf{c} \in \mathbb{R}^m$, $\rho > 0$.

(C) $\mathbf{G} \in \mathbb{S}^n_+$, $\mathbf{Q} \in \mathbb{S}^p_+$.

(D) *For any* $\mathbf{a} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^p$ *the optimal sets of the problems*

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ h_1(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{x}\|^2 + \frac{1}{2}\|\mathbf{x}\|^2_{\mathbf{G}} + \langle \mathbf{a}, \mathbf{x} \rangle \right\}$$

and

$$\min_{\mathbf{z}\in\mathbb{R}^p}\left\{h_2(\mathbf{z})+\frac{\rho}{2}\|\mathbf{B}\mathbf{z}\|^2+\frac{1}{2}\|\mathbf{z}\|_{\mathbf{Q}}^2+\langle\mathbf{b},\mathbf{z}\rangle\right\}$$

are nonempty.

(E) *There exists* $\hat{\mathbf{x}}\in\mathrm{ri}(\mathrm{dom}(h_1))$ *and* $\hat{\mathbf{z}}\in\mathrm{ri}(\mathrm{dom}(h_2))$ *for which* $\mathbf{A}\hat{\mathbf{x}}+\mathbf{B}\hat{\mathbf{z}}=\mathbf{c}$.

(F) *Problem* (15.1) *has a nonempty optimal set, denoted by* X^**, and the corresponding optimal value is* H_{opt}.

Property (D) guarantees that the AD-PMM method is actually a well-defined method.

By the strong duality theorem for convex problems (see Theorem A.1), under Assumption 15.2, it follows that strong duality holds for the pair of problems (15.1) and (15.2).

Theorem 15.3 (strong duality for the pair of problems (15.1) and (15.2)). *Suppose that Assumption* 15.2 *holds, and let* $H_{\mathrm{opt}}, q_{\mathrm{opt}}$ *be the optimal values of the primal and dual problems* (15.1) *and* (15.2)*, respectively. Then* $H_{\mathrm{opt}}=q_{\mathrm{opt}}$*, and the dual problem* (15.2) *possesses an optimal solution.*

We will now prove an $O(1/k)$ rate of convergence result of the sequence generated by AD-PMM.

Theorem 15.4 ($O(1/k)$ rate of convergence of AD-PMM).[87] *Suppose that Assumption* 15.2 *holds. Let* $\{(\mathbf{x}^k,\mathbf{z}^k)\}_{k\geq 0}$ *be the sequence generated by AD-PMM for solving problem* (15.1)*. Let* $(\mathbf{x}^*,\mathbf{z}^*)$ *be an optimal solution of problem* (15.1) *and* $\mathbf{y}^*$ *be an optimal solution of the dual problem* (15.2)*. Suppose that* $\gamma>0$ *is any constant satisfying* $\gamma\geq 2\|\mathbf{y}^*\|$*. Then for all* $n\geq 0$*,*

$$H(\mathbf{x}^{(n)},\mathbf{z}^{(n)})-H_{\mathrm{opt}}\leq\frac{\|\mathbf{x}^*-\mathbf{x}^0\|_{\mathbf{G}}^2+\|\mathbf{z}^*-\mathbf{z}^0\|_{\mathbf{C}}^2+\frac{1}{\rho}(\gamma+\|\mathbf{y}^0\|)^2}{2(n+1)},\quad(15.13)$$

$$\|\mathbf{A}\mathbf{x}^{(n)}+\mathbf{B}\mathbf{z}^{(n)}-\mathbf{c}\|\leq\frac{\|\mathbf{x}^*-\mathbf{x}^0\|_{\mathbf{G}}^2+\|\mathbf{z}^*-\mathbf{z}^0\|_{\mathbf{C}}^2+\frac{1}{\rho}(\gamma+\|\mathbf{y}^0\|)^2}{\gamma(n+1)},\quad(15.14)$$

where $\mathbf{C}=\rho\mathbf{B}^T\mathbf{B}+\mathbf{Q}$ *and*

$$\mathbf{x}^{(n)}=\frac{1}{n+1}\sum_{k=0}^{n}\mathbf{x}^{k+1},\mathbf{z}^{(n)}=\frac{1}{n+1}\sum_{k=0}^{n}\mathbf{z}^{k+1}.$$

Proof. By Fermat's optimality condition (Theorem 3.63) and the update steps (a) and (b) of AD-PMM, it follows that $\mathbf{x}^{k+1}$ and $\mathbf{z}^{k+1}$ satisfy

$$-\rho\mathbf{A}^T\left(\mathbf{A}\mathbf{x}^{k+1}+\mathbf{B}\mathbf{z}^k-\mathbf{c}+\frac{1}{\rho}\mathbf{y}^k\right)-\mathbf{G}(\mathbf{x}^{k+1}-\mathbf{x}^k)\in\partial h_1(\mathbf{x}^{k+1}),\quad(15.15)$$

$$-\rho\mathbf{B}^T\left(\mathbf{A}\mathbf{x}^{k+1}+\mathbf{B}\mathbf{z}^{k+1}-\mathbf{c}+\frac{1}{\rho}\mathbf{y}^k\right)-\mathbf{Q}(\mathbf{z}^{k+1}-\mathbf{z}^k)\in\partial h_2(\mathbf{z}^{k+1}).\quad(15.16)$$

[87] The proof of Theorem 15.4 on the rate of convergence of AD-PMM is based on a combination of the proof techniques of He and Yuan [65] and Gao and Zhang [58].

We will use the following notation:

$$\begin{aligned}\tilde{\mathbf{x}}^k &= \mathbf{x}^{k+1},\\ \tilde{\mathbf{z}}^k &= \mathbf{z}^{k+1},\\ \tilde{\mathbf{y}}^k &= \mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^k - \mathbf{c}).\end{aligned}$$

Using (15.15), (15.16), the subgradient inequality, and the above notation, we obtain that for any $\mathbf{x} \in \mathrm{dom}(h_1)$ and $\mathbf{z} \in \mathrm{dom}(h_2)$,

$$\begin{aligned}h_1(\mathbf{x}) - h_1(\tilde{\mathbf{x}}^k) + \left\langle \rho\mathbf{A}^T\left(\mathbf{A}\tilde{\mathbf{x}}^k + \mathbf{B}\mathbf{z}^k - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k\right) + \mathbf{G}(\tilde{\mathbf{x}}^k - \mathbf{x}^k), \mathbf{x} - \tilde{\mathbf{x}}^k \right\rangle \geq 0,\\ h_2(\mathbf{z}) - h_2(\tilde{\mathbf{z}}^k) + \left\langle \rho\mathbf{B}^T\left(\mathbf{A}\tilde{\mathbf{x}}^k + \mathbf{B}\tilde{\mathbf{z}}^k - \mathbf{c} + \frac{1}{\rho}\mathbf{y}^k\right) + \mathbf{Q}(\tilde{\mathbf{z}}^k - \mathbf{z}^k), \mathbf{z} - \tilde{\mathbf{z}}^k \right\rangle \geq 0.\end{aligned}$$

Using the definition of $\tilde{\mathbf{y}}^k$, the above two inequalities can be rewritten as

$$\begin{aligned}h_1(\mathbf{x}) - h_1(\tilde{\mathbf{x}}^k) + \left\langle \mathbf{A}^T\tilde{\mathbf{y}}^k + \mathbf{G}(\tilde{\mathbf{x}}^k - \mathbf{x}^k), \mathbf{x} - \tilde{\mathbf{x}}^k \right\rangle \geq 0,\\ h_2(\mathbf{z}) - h_2(\tilde{\mathbf{z}}^k) + \left\langle \mathbf{B}^T\tilde{\mathbf{y}}^k + (\rho\mathbf{B}^T\mathbf{B} + \mathbf{Q})(\tilde{\mathbf{z}}^k - \mathbf{z}^k), \mathbf{z} - \tilde{\mathbf{z}}^k \right\rangle \geq 0.\end{aligned}$$

Adding the above two inequalities and using the identity

$$\mathbf{y}^{k+1} - \mathbf{y}^k = \rho(\mathbf{A}\tilde{\mathbf{x}}^k + \mathbf{B}\tilde{\mathbf{z}}^k - \mathbf{c}),$$

we can conclude that for any $\mathbf{x} \in \mathrm{dom}(h_1)$, $\mathbf{z} \in \mathrm{dom}(h_2)$, and $\mathbf{y} \in \mathbb{R}^m$,

$$H(\mathbf{x},\mathbf{z}) - H(\tilde{\mathbf{x}}^k,\tilde{\mathbf{z}}^k) + \left\langle \begin{pmatrix}\mathbf{x} - \tilde{\mathbf{x}}^k\\ \mathbf{z} - \tilde{\mathbf{z}}^k\\ \mathbf{y} - \tilde{\mathbf{y}}^k\end{pmatrix}, \begin{pmatrix}\mathbf{A}^T\tilde{\mathbf{y}}^k\\ \mathbf{B}^T\tilde{\mathbf{y}}^k\\ -\mathbf{A}\tilde{\mathbf{x}}^k - \mathbf{B}\tilde{\mathbf{z}}^k + \mathbf{c}\end{pmatrix} - \begin{pmatrix}\mathbf{G}(\mathbf{x}^k - \tilde{\mathbf{x}}^k)\\ \mathbf{C}(\mathbf{z}^k - \tilde{\mathbf{z}}^k)\\ \frac{1}{\rho}(\mathbf{y}^k - \mathbf{y}^{k+1})\end{pmatrix} \right\rangle \geq 0, \tag{15.17}$$

where $\mathbf{C} = \rho\mathbf{B}^T\mathbf{B} + \mathbf{Q}$. We will use the following identity that holds for any positive semidefinite matrix $\mathbf{P}$:

$$(\mathbf{a} - \mathbf{b})^T\mathbf{P}(\mathbf{c} - \mathbf{d}) = \frac{1}{2}\left(\|\mathbf{a} - \mathbf{d}\|_{\mathbf{P}}^2 - \|\mathbf{a} - \mathbf{c}\|_{\mathbf{P}}^2 + \|\mathbf{b} - \mathbf{c}\|_{\mathbf{P}}^2 - \|\mathbf{b} - \mathbf{d}\|_{\mathbf{P}}^2\right).$$

Using the above identity, we can conclude that

$$\begin{aligned}(\mathbf{x} - \tilde{\mathbf{x}}^k)^T\mathbf{G}(\mathbf{x}^k - \tilde{\mathbf{x}}^k) &= \frac{1}{2}\left(\|\mathbf{x} - \tilde{\mathbf{x}}^k\|_{\mathbf{G}}^2 - \|\mathbf{x} - \mathbf{x}^k\|_{\mathbf{G}}^2 + \|\tilde{\mathbf{x}}^k - \mathbf{x}^k\|_{\mathbf{G}}^2\right)\\ &\geq \frac{1}{2}\|\mathbf{x} - \tilde{\mathbf{x}}^k\|_{\mathbf{G}}^2 - \frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|_{\mathbf{G}}^2,\end{aligned} \tag{15.18}$$

as well as

$$(\mathbf{z} - \tilde{\mathbf{z}}^k)^T\mathbf{C}(\mathbf{z}^k - \tilde{\mathbf{z}}^k) = \frac{1}{2}\|\mathbf{z} - \tilde{\mathbf{z}}^k\|_{\mathbf{C}}^2 - \frac{1}{2}\|\mathbf{z} - \mathbf{z}^k\|_{\mathbf{C}}^2 + \frac{1}{2}\|\mathbf{z}^k - \tilde{\mathbf{z}}^k\|_{\mathbf{C}}^2 \tag{15.19}$$

and

$$\begin{aligned}&2(\mathbf{y} - \tilde{\mathbf{y}}^k)^T(\mathbf{y}^k - \mathbf{y}^{k+1})\\ &= \|\mathbf{y} - \mathbf{y}^{k+1}\|^2 - \|\mathbf{y} - \mathbf{y}^k\|^2 + \|\tilde{\mathbf{y}}^k - \mathbf{y}^k\|^2 - \|\tilde{\mathbf{y}}^k - \mathbf{y}^{k+1}\|^2\\ &= \|\mathbf{y} - \mathbf{y}^{k+1}\|^2 - \|\mathbf{y} - \mathbf{y}^k\|^2 + \rho^2\|\mathbf{A}\tilde{\mathbf{x}}^k + \mathbf{B}\mathbf{z}^k - \mathbf{c}\|^2\\ &-\|\mathbf{y}^k + \rho(\mathbf{A}\tilde{\mathbf{x}}^k + \mathbf{B}\mathbf{z}^k - \mathbf{c}) - \mathbf{y}^k - \rho(\mathbf{A}\tilde{\mathbf{x}}^k + \mathbf{B}\tilde{\mathbf{z}}^k - \mathbf{c})\|^2\\ &= \|\mathbf{y} - \mathbf{y}^{k+1}\|^2 - \|\mathbf{y} - \mathbf{y}^k\|^2 + \rho^2\|\mathbf{A}\tilde{\mathbf{x}}^k + \mathbf{B}\mathbf{z}^k - \mathbf{c}\|^2 - \rho^2\|\mathbf{B}(\mathbf{z}^k - \tilde{\mathbf{z}}^k)\|^2.\end{aligned}$$

Therefore,

$$\frac{1}{\rho}(\mathbf{y}-\tilde{\mathbf{y}}^k)^T(\mathbf{y}^k-\mathbf{y}^{k+1}) \geq \frac{1}{2\rho}\left(\|\mathbf{y}-\mathbf{y}^{k+1}\|^2 - \|\mathbf{y}-\mathbf{y}^k\|^2\right) - \frac{\rho}{2}\|\mathbf{B}(\mathbf{z}^k-\tilde{\mathbf{z}}^k)\|^2. \quad (15.20)$$

Denoting

$$\mathbf{H} = \begin{pmatrix} \mathbf{G} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{\rho}\mathbf{I} \end{pmatrix},$$

as well as

$$\mathbf{w} = \begin{pmatrix} \mathbf{x} \\ \mathbf{z} \\ \mathbf{y} \end{pmatrix}, \quad \mathbf{w}^k = \begin{pmatrix} \mathbf{x}^k \\ \mathbf{z}^k \\ \mathbf{y}^k \end{pmatrix}, \quad \tilde{\mathbf{w}}^k = \begin{pmatrix} \tilde{\mathbf{x}}^k \\ \tilde{\mathbf{z}}^k \\ \tilde{\mathbf{y}}^k \end{pmatrix},$$

we obtain by combining (15.18), (15.19), and (15.20) that

$$\begin{aligned} \left\langle \begin{pmatrix} \mathbf{x}-\tilde{\mathbf{x}}^k \\ \mathbf{z}-\tilde{\mathbf{z}}^k \\ \mathbf{y}-\tilde{\mathbf{y}}^k \end{pmatrix}, \begin{pmatrix} \mathbf{G}(\mathbf{x}^k-\tilde{\mathbf{x}}^k) \\ \mathbf{C}(\mathbf{z}^k-\tilde{\mathbf{z}}^k) \\ \frac{1}{\rho}(\mathbf{y}^k-\mathbf{y}^{k+1}) \end{pmatrix} \right\rangle &\geq \frac{1}{2}\|\mathbf{w}-\mathbf{w}^{k+1}\|_{\mathbf{H}}^2 - \frac{1}{2}\|\mathbf{w}-\mathbf{w}^k\|_{\mathbf{H}}^2 \\ &\quad + \frac{1}{2}\|\mathbf{z}^k-\tilde{\mathbf{z}}^k\|_{\mathbf{C}}^2 - \frac{\rho}{2}\|\mathbf{B}(\mathbf{z}^k-\tilde{\mathbf{z}}^k)\|^2 \\ &\geq \frac{1}{2}\|\mathbf{w}-\mathbf{w}^{k+1}\|_{\mathbf{H}}^2 - \frac{1}{2}\|\mathbf{w}-\mathbf{w}^k\|_{\mathbf{H}}^2. \end{aligned}$$

Combining the last inequality with (15.17), we obtain that for any $\mathbf{x} \in \mathrm{dom}(h_1)$, $\mathbf{z} \in \mathrm{dom}(h_2)$, and $\mathbf{y} \in \mathbb{R}^m$,

$$H(\mathbf{x},\mathbf{z}) - H(\tilde{\mathbf{x}}^k,\tilde{\mathbf{z}}^k) + \langle \mathbf{w}-\tilde{\mathbf{w}}^k, \mathbf{F}\tilde{\mathbf{w}}^k + \tilde{\mathbf{c}}\rangle \geq \frac{1}{2}\|\mathbf{w}-\mathbf{w}^{k+1}\|_{\mathbf{H}}^2 - \frac{1}{2}\|\mathbf{w}-\mathbf{w}^k\|_{\mathbf{H}}^2, \quad (15.21)$$

where

$$\mathbf{F} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{A}^T \\ \mathbf{0} & \mathbf{0} & \mathbf{B}^T \\ -\mathbf{A} & -\mathbf{B} & \mathbf{0} \end{pmatrix}, \qquad \tilde{\mathbf{c}} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{c} \end{pmatrix}.$$

Note that

$$\begin{aligned} \langle \mathbf{w}-\tilde{\mathbf{w}}^k, \mathbf{F}\tilde{\mathbf{w}}^k + \tilde{\mathbf{c}}\rangle &= \langle \mathbf{w}-\tilde{\mathbf{w}}^k, \mathbf{F}(\tilde{\mathbf{w}}^k - \mathbf{w}) + \mathbf{F}\mathbf{w} + \tilde{\mathbf{c}}\rangle \\ &= \langle \mathbf{w}-\tilde{\mathbf{w}}^k, \mathbf{F}\mathbf{w} + \tilde{\mathbf{c}}\rangle, \end{aligned}$$

where the second equality follows from the fact that $\mathbf{F}$ is skew symmetric (meaning $\mathbf{F}^T = -\mathbf{F}$). We can thus conclude that (15.21) can be rewritten as

$$H(\mathbf{x},\mathbf{z}) - H(\tilde{\mathbf{x}}^k,\tilde{\mathbf{z}}^k) + \langle \mathbf{w}-\tilde{\mathbf{w}}^k, \mathbf{F}\mathbf{w} + \tilde{\mathbf{c}}\rangle \geq \frac{1}{2}\|\mathbf{w}-\mathbf{w}^{k+1}\|_{\mathbf{H}}^2 - \frac{1}{2}\|\mathbf{w}-\mathbf{w}^k\|_{\mathbf{H}}^2.$$

Summing the above inequality over $k = 0, 1, \ldots, n$ yields the inequality

$$(n+1)H(\mathbf{x},\mathbf{z}) - \sum_{k=0}^{n} H(\tilde{\mathbf{x}}^k, \tilde{\mathbf{z}}^k) + \left\langle (n+1)\mathbf{w} - \sum_{k=0}^{n} \tilde{\mathbf{w}}^k, \mathbf{F}\mathbf{w} + \tilde{\mathbf{c}} \right\rangle \geq -\frac{1}{2}\|\mathbf{w} - \mathbf{w}^0\|_{\mathbf{H}}^2.$$

Defining

$$\mathbf{w}^{(n)} = \frac{1}{n+1}\sum_{k=0}^{n} \tilde{\mathbf{w}}^k, \mathbf{x}^{(n)} = \frac{1}{n+1}\sum_{k=0}^{n} \mathbf{x}^{k+1}, \mathbf{z}^{(n)} = \frac{1}{n+1}\sum_{k=0}^{n} \mathbf{z}^{k+1}$$

and using the convexity of H, we obtain that

$$H(\mathbf{x},\mathbf{z}) - H(\mathbf{x}^{(n)},\mathbf{z}^{(n)}) + \langle \mathbf{w} - \mathbf{w}^{(n)}, \mathbf{F}\mathbf{w} + \tilde{\mathbf{c}} \rangle + \frac{1}{2(n+1)}\|\mathbf{w} - \mathbf{w}^0\|_{\mathbf{H}}^2 \geq 0.$$

Using (again) the skew-symmetry of $\mathbf{F}$, we can conclude that the above inequality is the same as

$$H(\mathbf{x},\mathbf{z}) - H(\mathbf{x}^{(n)},\mathbf{z}^{(n)}) + \langle \mathbf{w} - \mathbf{w}^{(n)}, \mathbf{F}\mathbf{w}^{(n)} + \tilde{\mathbf{c}} \rangle + \frac{1}{2(n+1)}\|\mathbf{w} - \mathbf{w}^0\|_{\mathbf{H}}^2 \geq 0.$$

In other words, for any $\mathbf{x} \in \mathrm{dom}(h_1)$ and $\mathbf{z} \in \mathrm{dom}(h_1)$,

$$H(\mathbf{x}^{(n)},\mathbf{z}^{(n)}) - H(\mathbf{x},\mathbf{z}) + \langle \mathbf{w}^{(n)} - \mathbf{w}, \mathbf{F}\mathbf{w}^{(n)} + \tilde{\mathbf{c}} \rangle \leq \frac{1}{2(n+1)}\|\mathbf{w} - \mathbf{w}^0\|_{\mathbf{H}}^2. \quad (15.22)$$

Let $(\mathbf{x}^*, \mathbf{z}^*)$ be an optimal solution of problem (15.1). Then $H(\mathbf{x}^*, \mathbf{z}^*) = H_{\mathrm{opt}}$ and $\mathbf{A}\mathbf{x}^* + \mathbf{B}\mathbf{z}^* = \mathbf{c}$. Plugging $\mathbf{x} = \mathbf{x}^*, \mathbf{z} = \mathbf{z}^*$, and the expressions for $\mathbf{w}^{(n)}, \mathbf{w}, \mathbf{w}^0, \mathbf{F}, \mathbf{H}, \tilde{\mathbf{c}}$ into (15.22), we obtain (denoting $\mathbf{y}^{(n)} = \frac{1}{n+1}\sum_{k=0}^{n} \tilde{\mathbf{y}}^k$)

$$\begin{aligned} &H(\mathbf{x}^{(n)},\mathbf{z}^{(n)}) - H_{\mathrm{opt}} + \langle \mathbf{x}^{(n)} - \mathbf{x}^*, \mathbf{A}^T\mathbf{y}^{(n)} \rangle + \langle \mathbf{z}^{(n)} - \mathbf{z}^*, \mathbf{B}^T\mathbf{y}^{(n)} \rangle \\ &+ \langle \mathbf{y}^{(n)} - \mathbf{y}, -\mathbf{A}\mathbf{x}^{(n)} - \mathbf{B}\mathbf{z}^{(n)} + \mathbf{c} \rangle \\ &\leq \frac{1}{2(n+1)} \left\{ \|\mathbf{x}^* - \mathbf{x}^0\|_{\mathbf{G}}^2 + \|\mathbf{z}^* - \mathbf{z}^0\|_{\mathbf{C}}^2 + \frac{1}{\rho}\|\mathbf{y} - \mathbf{y}^0\|^2 \right\}. \end{aligned}$$

Cancelling terms and using the fact that $\mathbf{A}\mathbf{x}^* + \mathbf{B}\mathbf{z}^* = \mathbf{c}$, we obtain that the last inequality is the same as

$$H(\mathbf{x}^{(n)},\mathbf{z}^{(n)}) - H_{\mathrm{opt}} + \langle \mathbf{y}, \mathbf{A}\mathbf{x}^{(n)} + \mathbf{B}\mathbf{z}^{(n)} - \mathbf{c} \rangle \leq \frac{\|\mathbf{x}^* - \mathbf{x}^0\|_{\mathbf{G}}^2 + \|\mathbf{z}^* - \mathbf{z}^0\|_{\mathbf{C}}^2 + \frac{1}{\rho}\|\mathbf{y} - \mathbf{y}^0\|^2}{2(n+1)}.$$

Since the above inequality holds for any $\mathbf{y} \in \mathbb{R}^m$, we can take the maximum of both sides over all $\mathbf{y} \in B[\mathbf{0}, \gamma]$ and obtain the inequality

$$H(\mathbf{x}^{(n)},\mathbf{z}^{(n)}) - H_{\mathrm{opt}} + \gamma\|\mathbf{A}\mathbf{x}^{(n)} + \mathbf{B}\mathbf{z}^{(n)} - \mathbf{c}\| \leq \frac{\|\mathbf{x}^* - \mathbf{x}^0\|_{\mathbf{G}}^2 + \|\mathbf{z}^* - \mathbf{z}^0\|_{\mathbf{C}}^2 + \frac{1}{\rho}(\gamma + \|\mathbf{y}^0\|)^2}{2(n+1)}.$$

Since $\gamma \geq 2\|\mathbf{y}^*\|$ for some optimal dual solution $\mathbf{y}^*$ and strong duality holds (Theorem 15.3), it follows by Theorem 3.60 that the two inequalities (15.13) and (15.14) hold. ∎

15.4 Minimizing $f_1(\mathbf{x}) + f_2(\mathbf{A}\mathbf{x})$

In this section we consider the model

$$\min_{\mathbf{x}\in\mathbb{R}^n} \{f_1(\mathbf{x}) + f_2(\mathbf{A}\mathbf{x})\}, \tag{15.23}$$

where f_1, f_2 are proper closed convex functions and $\mathbf{A} \in \mathbb{R}^{m\times n}$. As usual, $\rho > 0$ is a given constant. An implicit assumption will be that f_1 and f_2 are "proximable," which loosely speaking means that the prox operator of λf_1 and λf_2 can be efficiently computed for any $\lambda > 0$. This is obviously a "virtual" assumption, and its importance is only in the fact that it dictates the development of algorithms that rely on prox computations of λf_1 and λf_2.

Problem (15.23) can rewritten as

$$\min_{\mathbf{x}\in\mathbb{R}^n, \mathbf{z}\in\mathbb{R}^m} \{f_1(\mathbf{x}) + f_2(\mathbf{z}) : \mathbf{A}\mathbf{x} - \mathbf{z} = \mathbf{0}\}. \tag{15.24}$$

This fits the general model (15.1) with $h_1 = f_1, h_2 = f_2$, $\mathbf{B} = -\mathbf{I}$, and $\mathbf{c} = \mathbf{0}$. A direct implementation of ADMM leads to the following scheme ($\rho > 0$ is a given constant):

$$\begin{aligned}
\mathbf{x}^{k+1} &\in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^n} \left[f_1(\mathbf{x}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x} - \mathbf{z}^k + \frac{1}{\rho}\mathbf{y}^k \right\|^2 \right], \\
\mathbf{z}^{k+1} &= \operatorname{argmin}_{\mathbf{z}\in\mathbb{R}^m} \left[f_2(\mathbf{z}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x}^{k+1} - \mathbf{z} + \frac{1}{\rho}\mathbf{y}^k \right\|^2 \right], \\
\mathbf{y}^{k+1} &= \mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{z}^{k+1}).
\end{aligned} \tag{15.25}$$

The $\mathbf{z}$-step can be rewritten as a prox step, thus resulting in the following algorithm for solving problem (15.23).

Algorithm 1 [ADMM for solving (15.23)—version 1]

- **Initialization:** $\mathbf{x}^0 \in \mathbb{R}^n, \mathbf{z}^0, \mathbf{y}^0 \in \mathbb{R}^m$, $\rho > 0$.
- **General step ($k \geq 0$):**
  - (a) $\mathbf{x}^{k+1} \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^n} \left[f_1(\mathbf{x}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x} - \mathbf{z}^k + \frac{1}{\rho}\mathbf{y}^k \right\|^2 \right]$;
  - (b) $\mathbf{z}^{k+1} = \operatorname{prox}_{\frac{1}{\rho}f_2} \left(\mathbf{A}\mathbf{x}^{k+1} + \frac{1}{\rho}\mathbf{y}^k \right)$;
  - (c) $\mathbf{y}^{k+1} = \mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{z}^{k+1})$.

Step (a) of Algorithm 1 might be difficult to compute since the minimization in step (a) is more involved than a prox computation due to the quadratic term $\frac{\rho}{2}\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}$. We can actually employ ADMM in a different way that will refrain

from the type of computation made in step (a). For that, we will rewrite problem (15.23) as

$$\min_{\mathbf{x},\mathbf{w}\in\mathbb{R}^n,\mathbf{z}\in\mathbb{R}^m} \{f_1(\mathbf{w}) + f_2(\mathbf{z}) : \mathbf{A}\mathbf{x} - \mathbf{z} = \mathbf{0}, \mathbf{x} - \mathbf{w} = \mathbf{0}\}.$$

The above problem fits model (15.1) with $h_1 \equiv 0, h_2(\mathbf{z},\mathbf{w}) = f_1(\mathbf{z}) + f_2(\mathbf{w})$, $\mathbf{B} = -\mathbf{I}$, and $\binom{\mathbf{A}}{\mathbf{I}}$ taking the place of $\mathbf{A}$. The dual vector $\mathbf{y} \in \mathbb{R}^{m+n}$ is of the form $\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T)^T$, where $\mathbf{y}_1 \in \mathbb{R}^m$ and $\mathbf{y}_2 \in \mathbb{R}^n$. In the above reformulation we have two blocks of vectors: $\mathbf{x}$ and $(\mathbf{z},\mathbf{w})$. The $\mathbf{x}$-step is given by

$$\begin{aligned}\mathbf{x}^{k+1} &= \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^n} \left[\left\|\mathbf{A}\mathbf{x} - \mathbf{z}^k + \frac{1}{\rho}\mathbf{y}_1^k\right\|^2 + \left\|\mathbf{x} - \mathbf{w}^k + \frac{1}{\rho}\mathbf{y}_2^k\right\|^2\right] \\ &= (\mathbf{I} + \mathbf{A}^T\mathbf{A})^{-1}\left(\mathbf{A}^T\left[\mathbf{z}^k - \frac{1}{\rho}\mathbf{y}_1^k\right] + \mathbf{w}^k - \frac{1}{\rho}\mathbf{y}_2^k\right).\end{aligned}$$

The $(\mathbf{z},\mathbf{w})$-step is

$$\begin{aligned}\mathbf{z}^{k+1} &= \operatorname{prox}_{\frac{1}{\rho}f_2}\left(\mathbf{A}\mathbf{x}^{k+1} + \frac{1}{\rho}\mathbf{y}_1^k\right), \\ \mathbf{w}^{k+1} &= \operatorname{prox}_{\frac{1}{\rho}f_1}\left(\mathbf{x}^{k+1} + \frac{1}{\rho}\mathbf{y}_2^k\right).\end{aligned}$$

The method is summarized in the following.

Algorithm 2 [ADMM for solving (15.23)—version 2]

- **Initialization:** $\mathbf{x}^0, \mathbf{w}^0, \mathbf{y}_2^0 \in \mathbb{R}^n, \mathbf{z}^0, \mathbf{y}_1^0 \in \mathbb{R}^m, \rho > 0$.
- **General step ($k \geq 0$):**

$$\begin{aligned}\mathbf{x}^{k+1} &= (\mathbf{I} + \mathbf{A}^T\mathbf{A})^{-1}\left(\mathbf{A}^T\left[\mathbf{z}^k - \frac{1}{\rho}\mathbf{y}_1^k\right] + \mathbf{w}^k - \frac{1}{\rho}\mathbf{y}_2^k\right), \\ \mathbf{z}^{k+1} &= \operatorname{prox}_{\frac{1}{\rho}f_2}\left(\mathbf{A}\mathbf{x}^{k+1} + \frac{1}{\rho}\mathbf{y}_1^k\right), \\ \mathbf{w}^{k+1} &= \operatorname{prox}_{\frac{1}{\rho}f_1}\left(\mathbf{x}^{k+1} + \frac{1}{\rho}\mathbf{y}_2^k\right), \\ \mathbf{y}_1^{k+1} &= \mathbf{y}_1^k + \rho(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{z}^{k+1}), \\ \mathbf{y}_2^{k+1} &= \mathbf{y}_2^k + \rho(\mathbf{x}^{k+1} - \mathbf{w}^{k+1}).\end{aligned}$$

Algorithm 2 might still be too computationally demanding since it involves the evaluation of the inverse of $\mathbf{I} + \mathbf{A}^T\mathbf{A}$ (or at least the evaluation of $\mathbf{A}^T\mathbf{A}$ and a solution of a linear system at each iteration), which might be a difficult task in large-scale problems. We can alternatively employ AD-LPMM on problem (15.24) and obtain the following scheme that does not involve any matrix inverse calculations.

Algorithm 3 [AD-LPMM for solving (15.23)]

- **Initialization:** $\mathbf{x}^0 \in \mathbb{R}^n, \mathbf{z}^0, \mathbf{y}^0 \in \mathbb{R}^m$, $\rho > 0$, $\alpha \geq \rho\lambda_{\max}(\mathbf{A}^T\mathbf{A}), \beta \geq \rho$.
- **General step ($k \geq 0$):**

$$\begin{aligned}
\mathbf{x}^{k+1} &= \text{prox}_{\frac{1}{\alpha}f_1}\left[\mathbf{x}^k - \frac{\rho}{\alpha}\mathbf{A}^T\left(\mathbf{A}\mathbf{x}^k - \mathbf{z}^k + \frac{1}{\rho}\mathbf{y}^k\right)\right], \\
\mathbf{z}^{k+1} &= \text{prox}_{\frac{1}{\beta}f_2}\left[\mathbf{z}^k + \frac{\rho}{\beta}\left(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{z}^k + \frac{1}{\rho}\mathbf{y}^k\right)\right], \\
\mathbf{y}^{k+1} &= \mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{z}^{k+1}).
\end{aligned}$$

The above scheme has the advantage that it only requires simple linear algebra operations (no more than matrix/vector multiplications) and prox evaluations of λf_1 and λf_2 for different values of $\lambda > 0$.

Example 15.5 (l_1-regularized least squares). Consider the problem

$$\min_{\mathbf{x}\in\mathbb{R}^n}\left\{\frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1\right\}, \tag{15.26}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$ and $\lambda > 0$. Problem (15.26) fits the composite model (15.23) with $f_1(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$ and $f_2(\mathbf{y}) \equiv \frac{1}{2}\|\mathbf{y} - \mathbf{b}\|_2^2$. For any $\gamma > 0$, $\text{prox}_{\gamma f_1} = \mathcal{T}_{\gamma\lambda}$ (by Example 6.8) and $\text{prox}_{\gamma f_2}(\mathbf{y}) = \frac{\mathbf{y}+\gamma\mathbf{b}}{\gamma+1}$ (by Section 6.2.3). Step (a) of Algorithm 1 (first version of ADMM) has the form

$$\mathbf{x}^{k+1} \in \text{argmin}_{\mathbf{x}\in\mathbb{R}^n}\left[\lambda\|\mathbf{x}\|_1 + \frac{\rho}{2}\left\|\mathbf{A}\mathbf{x} - \mathbf{z}^k + \frac{1}{\rho}\mathbf{y}^k\right\|^2\right],$$

which actually means that this version of ADMM is completely useless since it suggests to solve an l_1-regularized least squares problem by a sequence of l_1-regularized least squares problems.

Algorithm 2 (second version of ADMM) has the following form.

ADMM, version 2 (Algorithm 2):

$$\begin{aligned}
\mathbf{x}^{k+1} &= (\mathbf{I} + \mathbf{A}^T\mathbf{A})^{-1}\left(\mathbf{A}^T\left[\mathbf{z}^k - \frac{1}{\rho}\mathbf{y}_1^k\right] + \mathbf{w}^k - \frac{1}{\rho}\mathbf{y}_2^k\right), \\
\mathbf{z}^{k+1} &= \frac{\rho\mathbf{A}\mathbf{x}^{k+1} + \mathbf{y}_1^k + \mathbf{b}}{\rho+1}, \\
\mathbf{w}^{k+1} &= \mathcal{T}_{\frac{\lambda}{\rho}}\left(\mathbf{x}^{k+1} + \frac{1}{\rho}\mathbf{y}_2^k\right), \\
\mathbf{y}_1^{k+1} &= \mathbf{y}_1^k + \rho(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{z}^{k+1}), \\
\mathbf{y}_2^{k+1} &= \mathbf{y}_2^k + \rho(\mathbf{x}^{k+1} - \mathbf{w}^{k+1}).
\end{aligned}$$

An implementation of the above ADMM variant will require to compute the matrix $\mathbf{A}^T\mathbf{A}$ in a preprocess and to solve at each iteration an $n \times n$ linear system

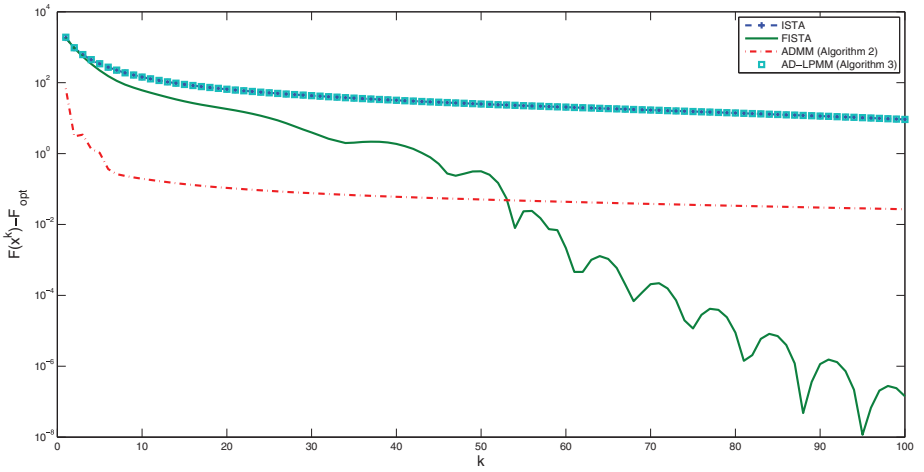

Figure 15.1. *Results of* 100 *iterations of ISTA, FISTA, ADMM (Algorithm* 2*) and AD-LPMM (Algorithm* 3*) on an* l_1*-regularized least squares problem.*

(or, alternatively, compute the inverse of $\mathbf{I} + \mathbf{A}^T\mathbf{A}$ in a preprocess). These operations might be difficult to execute in large-scale problems.

The general step of Algorithm 3 (which is essentially AD-LPMM) with $\alpha = \lambda_{\max}(\mathbf{A}^T\mathbf{A})\rho$ and $\beta = \rho$ takes the following form (denoting $L = \lambda_{\max}(\mathbf{A}^T\mathbf{A})$).

AD-LPMM (Algorithm 3):

$$\begin{aligned}
\mathbf{x}^{k+1} &= \mathcal{T}_{\frac{\lambda}{L\rho}}\left[\mathbf{x}^k - \frac{1}{L}\mathbf{A}^T\left(\mathbf{A}\mathbf{x}^k - \mathbf{z}^k + \frac{1}{\rho}\mathbf{y}^k\right)\right],\\
\mathbf{z}^{k+1} &= \frac{\rho\mathbf{A}\mathbf{x}^{k+1} + \mathbf{y}^k + \mathbf{b}}{\rho + 1},\\
\mathbf{y}^{k+1} &= \mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{z}^{k+1}).
\end{aligned}$$

The dominant computations in AD-LPMM are matrix/vector multiplications.

To illustrate the performance of the above two methods, we repeat the experiment described in Example 10.38 on the l_1-regularized least squares problem. We ran ADMM and AD-LPMM on the exact same instance, and the decay of the function values as a function of the iteration index k for the first 100 iterations is described in Figure 15.1. Clearly, ISTA and AD-LPMM exhibit the same performance, while ADMM seems to outperform both of them. This is actually not surprising since the computations carried out at each iteration of ADMM (solution of linear systems) are much heavier than the computations per iteration of AD-LPMM and ISTA (matrix/vector multiplications). In that respect, the comparison is in fact not fair and biased in favor of ADMM. What is definitely interesting is that FISTA significantly outperforms ADMM starting from approximately 50 iterations despite the fact that it is a simpler algorithm that requires substantially less computational effort per iteration. One possible reason is that FISTA is a method with a provably

$O(1/k^2)$ rate of convergence in function values, while ADMM is only guaranteed to converge at a rate of $O(1/k)$. ∎

Example 15.6 (robust regression). Consider the problem

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1, \tag{15.27}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$. Problem (15.27) fits the composite model (15.23) with $f_1 \equiv 0$ and $f_2(\mathbf{y}) = \|\mathbf{y} - \mathbf{b}\|_1$. Let $\rho > 0$. For any $\gamma > 0$, $\text{prox}_{\gamma f_1}(\mathbf{y}) = \mathbf{y}$ and $\text{prox}_{\gamma f_2}(\mathbf{y}) = \mathcal{T}_\gamma(\mathbf{y} - \mathbf{b}) + \mathbf{b}$ (by Example 6.8 and Theorem 6.11). Therefore, the general step of Algorithm 1 (first version of ADMM) takes the following form.

ADMM, version 1 (Algorithm 1):

$$\begin{aligned}
\mathbf{x}^{k+1} &= \text{argmin}_{\mathbf{x}\in\mathbb{R}^m} \left\| \mathbf{A}\mathbf{x} - \mathbf{z}^k + \frac{1}{\rho}\mathbf{y}^k \right\|^2, \\
\mathbf{z}^{k+1} &= \mathcal{T}_{\frac{1}{\rho}} \left(\mathbf{A}\mathbf{x}^{k+1} + \frac{1}{\rho}\mathbf{y}^k - \mathbf{b} \right) + \mathbf{b}, \\
\mathbf{y}^{k+1} &= \mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{z}^{k+1}).
\end{aligned}$$

The general step of Algorithm 2 (second version of ADMM) reads as follows:

$$\begin{aligned}
\mathbf{x}^{k+1} &= (\mathbf{I} + \mathbf{A}^T\mathbf{A})^{-1} \left(\mathbf{A}^T \left[\mathbf{z}^k - \frac{1}{\rho}\mathbf{y}_1^k \right] + \mathbf{w}^k - \frac{1}{\rho}\mathbf{y}_2^k \right), \\
\mathbf{z}^{k+1} &= \mathcal{T}_{\frac{1}{\rho}} \left(\mathbf{A}\mathbf{x}^{k+1} + \frac{1}{\rho}\mathbf{y}_1^k - \mathbf{b} \right) + \mathbf{b}, \\
\mathbf{w}^{k+1} &= \mathbf{x}^{k+1} + \frac{1}{\rho}\mathbf{y}_2^k, \\
\mathbf{y}_1^{k+1} &= \mathbf{y}_1^k + \rho(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{z}^{k+1}), \\
\mathbf{y}_2^{k+1} &= \mathbf{y}_2^k + \rho(\mathbf{x}^{k+1} - \mathbf{w}^{k+1}).
\end{aligned}$$

Plugging the expression for $\mathbf{w}^{k+1}$ into the update formula of $\mathbf{y}_2^{k+1}$, we obtain that $\mathbf{y}_2^{k+1} = \mathbf{0}$. Thus, if we start with $\mathbf{y}_2^0 = 0$, then $\mathbf{y}_2^k = \mathbf{0}$ for all $k \geq 0$, and consequently $\mathbf{w}^k = \mathbf{x}^k$ for all k. The algorithm thus reduces to the following.

ADMM, version 2 (Algorithm 2):

$$\begin{aligned}
\mathbf{x}^{k+1} &= (\mathbf{I} + \mathbf{A}^T\mathbf{A})^{-1} \left(\mathbf{A}^T \left[\mathbf{z}^k - \frac{1}{\rho}\mathbf{y}_1^k \right] + \mathbf{x}^k \right), \\
\mathbf{z}^{k+1} &= \mathcal{T}_{\frac{1}{\rho}} \left(\mathbf{A}\mathbf{x}^{k+1} + \frac{1}{\rho}\mathbf{y}_1^k - \mathbf{b} \right) + \mathbf{b}, \\
\mathbf{y}_1^{k+1} &= \mathbf{y}_1^k + \rho(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{z}^{k+1}).
\end{aligned}$$

Algorithm 3 (which is essentially AD-LPMM) with $\alpha = \lambda_{\max}(\mathbf{A}^T\mathbf{A})\rho$ and $\beta = \rho$ takes the following form (denoting $L = \lambda_{\max}(\mathbf{A}^T\mathbf{A})$).

AD-LPMM (Algorithm 3):

$$\begin{aligned}
\mathbf{x}^{k+1} &= \mathbf{x}^k - \frac{1}{L}\mathbf{A}^T\left(\mathbf{A}\mathbf{x}^k - \mathbf{z}^k + \frac{1}{\rho}\mathbf{y}^k\right),\\
\mathbf{z}^{k+1} &= \mathcal{T}_{\frac{1}{\rho}}\left[\left(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b} + \frac{1}{\rho}\mathbf{y}^k\right)\right] + \mathbf{b},\\
\mathbf{y}^{k+1} &= \mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{z}^{k+1}). \quad \blacksquare
\end{aligned}$$

Example 15.7 (basis pursuit). Consider the problem

$$\begin{aligned}
&\min \quad \|\mathbf{x}\|_1\\
&\text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b},
\end{aligned} \tag{15.28}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$. Problem (15.28) fits the composite model (15.23) with $f_1(\mathbf{x}) = \|\mathbf{x}\|_1$ and $f_2 = \delta_{\{\mathbf{b}\}}$. Let $\rho > 0$. For any $\gamma > 0$, $\mathrm{prox}_{\gamma f_1} = \mathcal{T}_\gamma$ (by Example 6.8) and $\mathrm{prox}_{\gamma_2 f_2} \equiv \mathbf{b}$. Algorithm 1 is actually not particularly implementable since its first update step is

$$\mathbf{x}^{k+1} \in \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^n}\left\{\|\mathbf{x}\|_1 + \frac{\rho}{2}\left\|\mathbf{A}\mathbf{x} - \mathbf{z}^k + \frac{1}{\rho}\mathbf{y}^k\right\|^2\right\},$$

which does not seem to be simpler to solve than the original problem (15.28).

Algorithm 2 takes the following form (assuming that $\mathbf{z}^0 = \mathbf{b}$).

ADMM, version 2 (Algorithm 2):

$$\begin{aligned}
\mathbf{x}^{k+1} &= (\mathbf{I} + \mathbf{A}^T\mathbf{A})^{-1}\left(\mathbf{A}^T\left[\mathbf{b} - \frac{1}{\rho}\mathbf{y}_1^k\right] + \mathbf{w}^k - \frac{1}{\rho}\mathbf{y}_2^k\right),\\
\mathbf{w}^{k+1} &= \mathcal{T}_{\frac{1}{\rho}}\left(\mathbf{x}^{k+1} + \frac{1}{\rho}\mathbf{y}_2^k\right),\\
\mathbf{y}_1^{k+1} &= \mathbf{y}_1^k + \rho(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}),\\
\mathbf{y}_2^{k+1} &= \mathbf{y}_2^k + \rho(\mathbf{x}^{k+1} - \mathbf{w}^{k+1}).
\end{aligned}$$

Finally, assuming that $\mathbf{z}^0 = \mathbf{b}$, Algorithm 3 with $\alpha = \lambda_{\max}(\mathbf{A}^T\mathbf{A})\rho$ and $\beta = \rho$ reduces to the following simple update steps (denoting $L = \lambda_{\max}(\mathbf{A}^T\mathbf{A})$).

AD-LPMM (Algorithm 3):

$$\begin{aligned}
\mathbf{x}^{k+1} &= \mathcal{T}_{\frac{1}{\rho L}}\left[\mathbf{x}^k - \frac{1}{L}\mathbf{A}^T\left(\mathbf{A}\mathbf{x}^k - \mathbf{b} + \frac{1}{\rho}\mathbf{y}^k\right)\right],\\
\mathbf{y}^{k+1} &= \mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}). \quad \blacksquare
\end{aligned}$$

Example 15.8 (minimizing $\sum_{i=1}^{p} g_i(\mathbf{A}_i\mathbf{x})$). Consider now the problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \sum_{i=1}^{p} g_i(\mathbf{A}_i\mathbf{x}), \tag{15.29}$$

where $g_1, g_2, \ldots, g_p$ are proper closed and convex functions and $\mathbf{A}_i \in \mathbb{R}^{m_i \times n}$ for all $i = 1, 2, \ldots, p$. Problem (15.29) fits the composite model (15.23) with

- $f_1 \equiv 0$;
- $f_2(\mathbf{y}) = \sum_{i=1}^{p} g_i(\mathbf{y}_i)$, where we assume that $\mathbf{y} \in \mathbb{R}^{m_1+m_2+\cdots+m_p}$ is of the form $\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_p^T)^T$, where $\mathbf{y}_i \in \mathbb{R}^{m_i}$;
- the matrix $\mathbf{A} \in \mathbb{R}^{(m_1+m_2+\ldots+m_p)\times n}$ given by $\mathbf{A} = (\mathbf{A}_1^T, \mathbf{A}_2^T, \ldots, \mathbf{A}_p^T)^T$.

For any $\gamma > 0$, $\mathrm{prox}_{\gamma f_1}(\mathbf{x}) = \mathbf{x}$ and $\mathrm{prox}_{\gamma f_2}(\mathbf{y})_i = \mathrm{prox}_{\gamma g_i}(\mathbf{y}_i)$, $i = 1, 2, \ldots, p$ (by Theorem 6.6). The general update step of the first version of ADMM (Algorithm 1) has the form

$$\begin{aligned}
\mathbf{x}^{k+1} &\in \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^n} \sum_{i=1}^{p} \left\| \mathbf{A}_i\mathbf{x} - \mathbf{z}_i^k + \frac{1}{\rho}\mathbf{y}_i^k \right\|^2, \qquad (15.30)\\
\mathbf{z}_i^{k+1} &= \mathrm{prox}_{\frac{1}{\rho}g_i}\left(\mathbf{A}_i\mathbf{x}^{k+1} + \frac{1}{\rho}\mathbf{y}_i^k \right), \quad i = 1, 2, \ldots, p,\\
\mathbf{y}_i^{k+1} &= \mathbf{y}_i^k + \rho(\mathbf{A}_i\mathbf{x}^{k+1} - \mathbf{z}_i^{k+1}), \quad i = 1, 2, \ldots, p.
\end{aligned}$$

In the case where $\mathbf{A}$ has full column rank, step (15.30) can be written more explicitly, leading to the following representation.

ADMM, version 1 (Algorithm 1):

$$\begin{aligned}
\mathbf{x}^{k+1} &= \left(\sum_{i=1}^{p} \mathbf{A}_i^T\mathbf{A}_i \right)^{-1} \sum_{i=1}^{p} \mathbf{A}_i^T \left(\mathbf{z}_i^k - \frac{1}{\rho}\mathbf{y}_i^k \right),\\
\mathbf{z}_i^{k+1} &= \mathrm{prox}_{\frac{1}{\rho}g_i}\left(\mathbf{A}_i\mathbf{x}^{k+1} + \frac{1}{\rho}\mathbf{y}_i^k \right), \quad i = 1, 2, \ldots, p,\\
\mathbf{y}_i^{k+1} &= \mathbf{y}_i^k + \rho(\mathbf{A}_i\mathbf{x}^{k+1} - \mathbf{z}_i^{k+1}), \quad i = 1, 2, \ldots, p.
\end{aligned}$$

The second version of ADMM (Algorithm 2) is not simpler than the first version, and we will therefore not write it explicitly. AD-LPMM (Algorithm 3) invoked with the constants $\alpha = \lambda_{\max}(\sum_{i=1}^{p} \mathbf{A}_i^T\mathbf{A}_i)\rho$ and $\beta = \rho$ reads as follows (denoting $L = \lambda_{\max}(\sum_{i=1}^{p} \mathbf{A}_i^T\mathbf{A}_i)$).

AD-LPMM (Algorithm 3):

$$\begin{aligned}
\mathbf{x}^{k+1} &= \mathbf{x}^k - \frac{1}{L}\sum_{i=1}^{p} \mathbf{A}_i^T \left(\mathbf{A}_i\mathbf{x}^k - \mathbf{z}_i^k + \frac{1}{\rho}\mathbf{y}_i^k \right),\\
\mathbf{z}_i^{k+1} &= \mathrm{prox}_{\frac{1}{\rho}g_i}\left(\mathbf{A}_i\mathbf{x}^{k+1} + \frac{1}{\rho}\mathbf{y}_i^k \right), \quad i = 1, 2, \ldots, p,\\
\mathbf{y}_i^{k+1} &= \mathbf{y}_i^k + \rho(\mathbf{A}_i\mathbf{x}^{k+1} - \mathbf{z}_i^{k+1}), \quad i = 1, 2, \ldots, p. \qquad \blacksquare
\end{aligned}$$

Appendix A

Strong Duality and Optimality Conditions

The following strong duality theorem is taken from [29, Proposition 6.4.4].

Theorem A.1 (strong duality theorem). *Consider the optimization problem*

$$\begin{aligned} f_{\text{opt}} = \quad & \min \quad f(\mathbf{x}) \\ & s.t. \quad g_i(\mathbf{x}) \le 0, \quad i = 1, 2, \dots, m, \\ & \qquad h_j(\mathbf{x}) \le 0, \quad j = 1, 2, \dots, p, \\ & \qquad s_k(\mathbf{x}) = 0, \quad k = 1, 2, \dots, q, \\ & \qquad \mathbf{x} \in X, \end{aligned} \tag{A.1}$$

where $X = P \cap C$ with $P \subseteq \mathbb{E}$ being a convex polyhedral set and $C \subseteq \mathbb{E}$ convex. The functions $f, g_i, i = 1, 2, \dots, m : \mathbb{E} \to (-\infty, \infty]$ are convex, and their domains satisfy $X \subseteq \text{dom}(f), X \subseteq \text{dom}(g_i)$, $i = 1, 2, \dots, m$. The functions $h_j, s_k, j = 1, 2, \dots, p, k = 1, 2, \dots, q$, are affine functions. Suppose there exist

(i) *a feasible solution $\bar{\mathbf{x}}$ satisfying $g_i(\bar{\mathbf{x}}) < 0$ for all $i = 1, 2, \dots, m$;*

(ii) *a vector satisfying all the affine constraints $h_j(\mathbf{x}) \le 0, j = 1, 2, \dots, p$, $s_k(\mathbf{x}) = 0, k = 1, 2, \dots, q$, and that is in $P \cap \text{ri}(C)$.*

Then if problem (A.1) has a finite optimal value, then the optimal value of the dual problem

$$q_{\text{opt}} = \max\{q(\boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\mu}) : (\boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\mu}) \in \text{dom}(-q)\},$$

where $q : \mathbb{R}^m_+ \times \mathbb{R}^p_+ \times \mathbb{R}^q \to \mathbb{R} \cup \{-\infty\}$ is given by

$$\begin{aligned} q(\boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\mu}) &= \min_{\mathbf{x} \in X} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\mu}) \\ &= \min_{\mathbf{x} \in X} \left[f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^{p} \eta_j h_j(\mathbf{x}) + \sum_{k=1}^{q} \mu_k s_k(\mathbf{x}) \right], \end{aligned}$$

is attained, and the optimal values of the primal and dual problems are the same:

$$f_{\text{opt}} = q_{\text{opt}}.$$

We also recall some well-known optimality conditions expressed in terms of the Lagrangian function in cases where strong duality holds.

Theorem A.2 (optimality conditions under strong duality). *Consider the problem*

$$\text{(P)} \qquad \begin{array}{ll} \min & f(\mathbf{x}) \\ s.t. & g_i(\mathbf{x}) \le 0, i = 1, 2, \dots, m, \\ & h_j(\mathbf{x}) = 0, j = 1, 2, \dots, p, \\ & \mathbf{x} \in X, \end{array}$$

where $f, g_1, g_2, \dots, g_m, h_1, h_2, \dots, h_p : \mathbb{E} \to (-\infty, \infty]$, *and* $X \subseteq \mathbb{E}$. *Assume that* $X \subseteq \text{dom}(f)$, $X \subseteq \text{dom}(g_i)$, *and* $X \subseteq \text{dom}(h_j)$ *for all* $i = 1, 2, \dots, m$, $j = 1, 2, \dots, p$. *Let* (D) *be the following dual problem:*

$$\text{(D)} \qquad \begin{array}{ll} \max & q(\boldsymbol{\lambda}, \boldsymbol{\mu}) \\ s.t. & (\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \text{dom}(-q), \end{array}$$

where

$$q(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \min_{\mathbf{x} \in X} \left\{ L(\mathbf{x}; \boldsymbol{\lambda}, \boldsymbol{\mu}) \equiv f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{x}) \right\},$$
$$\text{dom}(-q) = \{(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathbb{R}^m_+ \times \mathbb{R}^p : q(\boldsymbol{\lambda}, \boldsymbol{\mu}) > -\infty\}.$$

Suppose that the optimal value of problem (P) *is finite and equal to the optimal value of problem* (D). *Then* $\mathbf{x}^*, (\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ *are optimal solutions of problems* (P) *and* (D), *respectively, if and only if*

(i) $\mathbf{x}^*$ *is a feasible solution of (P);*

(ii) $\boldsymbol{\lambda}^* \ge \mathbf{0}$;

(iii) $\lambda_i^* g_i(\mathbf{x}^*) = 0, i = 1, 2, \dots, m$;

(iv) $\mathbf{x}^* \in \text{argmin}_{\mathbf{x} \in X} L(\mathbf{x}; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$.

Proof. Denote the optimal values of problem (P) and (D) by f_{opt} and q_{opt}, respectively. An underlying assumption of the theorem is that $f_{\text{opt}} = q_{\text{opt}}$. If $\mathbf{x}^*$ and $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ are the optimal solutions of problems (P) and (D), then obviously (i) and

(ii) are satisfied. In addition,

$$\begin{aligned} f_{\text{opt}} &= q_{\text{opt}} = q(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \\ &= \min_{\mathbf{x} \in X} L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \\ &\le L(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \\ &= f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* g_i(\mathbf{x}^*) + \sum_{j=1}^{p} \mu_j^* h_j(\mathbf{x}^*) \\ &\le f(\mathbf{x}^*), \end{aligned}$$

where the last inequality follows by the facts that $h_j(\mathbf{x}^*) = 0$, $\lambda_i^* \ge 0$, and $g_i(\mathbf{x}^*) \le 0$. Since $f_{\text{opt}} = f(\mathbf{x}^*)$, all of the inequalities in the above chain of equalities and inequalities are actually equalities. This implies in particular that $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x} \in X} L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$, meaning property (iv), and that $\sum_{i=1}^{m} \lambda_i^* g_i(\mathbf{x}^*) = 0$, which by the fact that $\lambda_i^* g_i(\mathbf{x}^*) \le 0$ for any i, implies that $\lambda_i^* g_i(\mathbf{x}^*) = 0$ for any i, showing the validity of property (iii).

To prove the reverse direction, assume that properties (i)–(iv) are satisfied. Then

$$\begin{aligned} q(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) &= \min_{\mathbf{x} \in X} L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) && \text{[definition of } q\text{]} \\ &= L(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) && \text{[property (iv)]} \\ &= f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* g_i(\mathbf{x}^*) + \sum_{j=1}^{p} \mu_j^* h_j(\mathbf{x}^*) \\ &= f(\mathbf{x}^*). && \text{[property (iii)]} \end{aligned}$$

By the weak duality theorem, since $\mathbf{x}^*$ and $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ are primal and dual feasible solutions with equal primal and dual objective functions, it follows that they are the optimal solutions of their corresponding problems. □

Appendix B
Tables

Support Functions

C	$\sigma_C(\mathbf{y})$	Assumptions	Reference
$\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$	$\max_{i=1,2,\ldots,n}\langle \mathbf{b}_i, \mathbf{y}\rangle$	$\mathbf{b}_i \in \mathbb{E}$	Example 2.25
K	$\delta_{K^\circ}(\mathbf{y})$	K – cone	Example 2.26
$\mathbb{R}^n_+$	$\delta_{\mathbb{R}^n_-}(\mathbf{y})$	$\mathbb{E} = \mathbb{R}^n$	Example 2.27
Δ_n	$\max\{y_1, y_2, \ldots, y_n\}$	$\mathbb{E} = \mathbb{R}^n$	Example 2.36
$\{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{0}\}$	$\delta_{\{\mathbf{A}^T\boldsymbol{\lambda}:\boldsymbol{\lambda}\in\mathbb{R}^m_+\}}(\mathbf{y})$	$\mathbb{E} = \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{m\times n}$	Example 2.29
$\{\mathbf{x} \in \mathbb{R}^n : \mathbf{B}\mathbf{x} = \mathbf{b}\}$	$\langle \mathbf{y}, \mathbf{x}_0\rangle + \delta_{\mathrm{Range}(\mathbf{B}^T)}(\mathbf{y})$	$\mathbb{E} = \mathbb{R}^n$, $\mathbf{B} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{B}\mathbf{x}_0 = \mathbf{b}$	Example 2.30
$B_{\|\cdot\|}[\mathbf{0}, 1]$	$\|\mathbf{y}\|_*$	-	Example 2.31

Weak Subdifferential Results

Function	Weak result	Setting	Reference
$-q$ = negative dual function	$-\mathbf{g}(\mathbf{x}_0) \in \partial(-q)(\boldsymbol{\lambda}_0)$	$q(\boldsymbol{\lambda}) = \min_{\mathbf{x}\in X} f(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{g}(\mathbf{x})$, $f : \mathbb{E} \to \mathbb{R}$, $\mathbf{g} : \mathbb{E} \to \mathbb{R}^m$, $\mathbf{x}_0$ = a minimizer of $f(\mathbf{x}) + \boldsymbol{\lambda}_0^T\mathbf{g}(\mathbf{x})$ over X	Example 3.7
$f(\mathbf{X}) = \lambda_{\max}(\mathbf{X})$	$\mathbf{v}\mathbf{v}^T \in \partial f(\mathbf{X})$	$f : \mathbb{S}^n \to \mathbb{R}$, $\mathbf{v}$ = normalized maximum eigenvector of $X \in \mathbb{S}^n$	Example 3.8
$f(\mathbf{x}) = \|\mathbf{x}\|_1$	$\mathrm{sgn}(\mathbf{x}) \in \partial f(\mathbf{x})$	$\mathbb{E} = \mathbb{R}^n$	Example 3.42
$f(\mathbf{x}) = \lambda_{\max}(\mathbf{A}_0 + \sum_{i=1}^m x_i\mathbf{A}_i)$	$(\tilde{\mathbf{y}}^T\mathbf{A}_i\tilde{\mathbf{y}})_{i=1}^m \in \partial f(\mathbf{x})$	$\tilde{\mathbf{y}}$ = normalized maximum eigenvector of $\mathbf{A}_0 + \sum_{i=1}^m x_i\mathbf{A}_i$	Example 3.56

Strong Subdifferential Results

$f(\mathbf{x})$	$\partial f(\mathbf{x})$	Assumptions	Reference
$\|\mathbf{x}\|$	$B_{\|\cdot\|_*}[\mathbf{0},1]$	$\mathbf{x}=\mathbf{0}$	Example 3.3
$\|\mathbf{x}\|_1$	$\left\{\sum_{i\in I_{\neq}(\mathbf{x})}\operatorname{sgn}(x_i)\mathbf{e}_i+\sum_{i\in I_0(\mathbf{x})}[-\mathbf{e}_i,\mathbf{e}_i]\right\}$	$\mathbb{E}=\mathbb{R}^n$, $I_{\neq}(\mathbf{x})=\{i:x_i\neq 0\}$, $I_0(\mathbf{x})=\{i:x_i=0\}$.	Example 3.41
$\|\mathbf{x}\|_2$	$\begin{cases}\left\{\frac{\mathbf{x}}{\|\mathbf{x}\|_2}\right\}, & \mathbf{x}\neq\mathbf{0},\\ B_{\|\cdot\|_2}[\mathbf{0},1], & \mathbf{x}=\mathbf{0}.\end{cases}$	$\mathbb{E}=\mathbb{R}^n$	Example 3.34
$\|\mathbf{x}\|_\infty$	$\left\{\sum_{i\in I(\mathbf{x})}\lambda_i\operatorname{sgn}(x_i)\mathbf{e}_i:\ \begin{array}{l}\sum_{i\in I(\mathbf{x})}\lambda_i=1\\ \lambda_i\geq 0\end{array}\right\}$	$\mathbb{E}=\mathbb{R}^n$, $I(\mathbf{x})=\{i:\|\mathbf{x}\|_\infty=\|x_i\|\}$, $\mathbf{x}\neq\mathbf{0}$	Example 3.52
$\max(\mathbf{x})$	$\left\{\sum_{i\in I(\mathbf{x})}\lambda_i\mathbf{e}_i:\sum_{i\in I(\mathbf{x})}\lambda_i=1,\lambda_i\geq 0\right\}$	$\mathbb{E}=\mathbb{R}^n$, $I(\mathbf{x})=\{i:\max(\mathbf{x})=x_i\}$	Example 3.51
$\max(\mathbf{x})$	Δ_n	$\mathbb{E}=\mathbb{R}^n$, $\mathbf{x}=\alpha\mathbf{e}$ for some $\alpha\in\mathbb{R}$	Example 3.51
$\delta_S(\mathbf{x})$	$N_S(\mathbf{x})$	$\emptyset\neq S\subseteq\mathbb{E}$	Example 3.5
$\delta_{B[\mathbf{0},1]}(\mathbf{x})$	$\begin{cases}\{\mathbf{y}\in\mathbb{E}^*:\|\mathbf{y}\|_*\leq\langle\mathbf{y},\mathbf{x}\rangle\}, & \|\mathbf{x}\|\leq 1,\\ \emptyset, & \|\mathbf{x}\|>1.\end{cases}$		Example 3.6
$\|\mathbf{A}\mathbf{x}+\mathbf{b}\|_1$	$\sum_{i\in I_{\neq}(\mathbf{x})}\operatorname{sgn}(\mathbf{a}_i^T\mathbf{x}+b_i)\mathbf{a}_i+\sum_{i\in I_0(\mathbf{x})}[-\mathbf{a}_i,\mathbf{a}_i]$	$\mathbb{E}=\mathbb{R}^n$, $\mathbf{A}\in\mathbb{R}^{m\times n}$, $\mathbf{b}\in\mathbb{R}^m$, $I_{\neq}(\mathbf{x})=\{i:\mathbf{a}_i^T\mathbf{x}+b_i\neq 0\}$, $I_0(\mathbf{x})=\{i:\mathbf{a}_i^T\mathbf{x}+b_i=0\}$	Example 3.44
$\|\mathbf{A}\mathbf{x}+\mathbf{b}\|_2$	$\begin{cases}\frac{\mathbf{A}^T(\mathbf{A}\mathbf{x}+\mathbf{b})}{\|\mathbf{A}\mathbf{x}+\mathbf{b}\|_2}, & \mathbf{A}\mathbf{x}+\mathbf{b}\neq\mathbf{0},\\ \mathbf{A}^TB_{\|\cdot\|_2}[\mathbf{0},1], & \mathbf{A}\mathbf{x}+\mathbf{b}=\mathbf{0}.\end{cases}$	$\mathbb{E}=\mathbb{R}^n$, $\mathbf{A}\in\mathbb{R}^{m\times n}$, $\mathbf{b}\in\mathbb{R}^m$	Example 3.45
$\|\mathbf{A}\mathbf{x}+\mathbf{b}\|_\infty$	$\left\{\sum_{i\in I_{\mathbf{x}}}\lambda_i\operatorname{sgn}(\mathbf{a}_i^T\mathbf{x}+b_i)\mathbf{a}_i:\ \begin{array}{l}\sum_{i\in I_{\mathbf{x}}}\lambda_i=1\\ \lambda_i\geq 0\end{array}\right\}$	$\mathbb{E}=\mathbb{R}^n$, $\mathbf{A}\in\mathbb{R}^{m\times n}$, $\mathbf{b}\in\mathbb{R}^m$, $I_{\mathbf{x}}=\{i:\|\mathbf{A}\mathbf{x}+\mathbf{b}\|_\infty=\|\mathbf{a}_i^T\mathbf{x}+b_i\|\}$, $\mathbf{A}\mathbf{x}+\mathbf{b}\neq\mathbf{0}$	Example 3.54
$\|\mathbf{A}\mathbf{x}+\mathbf{b}\|_\infty$	$\mathbf{A}^TB_{\|\cdot\|_1}[\mathbf{0},1]$	same as above but with $\mathbf{A}\mathbf{x}+\mathbf{b}=\mathbf{0}$	Example 3.54
$\max_i\{\mathbf{a}_i^T\mathbf{x}+\mathbf{b}\}$	$\left\{\sum_{i\in I(\mathbf{x})}\lambda_i\mathbf{a}_i:\sum_{i\in I(\mathbf{x})}\lambda_i=1,\lambda_i\geq 0\right\}$	$\mathbb{E}=\mathbb{R}^n$,$\mathbf{a}_i\in\mathbb{R}^n$, $b_i\in\mathbb{R}$, $I(\mathbf{x})=\{i:f(\mathbf{x})=\mathbf{a}_i^T\mathbf{x}+b_i\}$	Example 3.53
$\frac{1}{2}d_C(\mathbf{x})^2$	$\{\mathbf{x}-P_C(\mathbf{x})\}$	C = nonempty closed and convex, $\mathbb{E}$ = Euclidean	Example 3.31
$d_C(\mathbf{x})$	$\begin{cases}\left\{\frac{\mathbf{x}-P_C(\mathbf{x})}{d_C(\mathbf{x})}\right\}, & \mathbf{x}\notin C,\\ N_C(\mathbf{x})\cap B[\mathbf{0},1] & \mathbf{x}\in C.\end{cases}$	C = nonempty closed and convex, $\mathbb{E}$ = Euclidean	Example 3.49

Conjugate Calculus Rules

$g(\mathbf{x})$	$g^*(\mathbf{y})$	Reference
$\sum_{i=1}^m f_i(\mathbf{x}_i)$	$\sum_{i=1}^m f_i^*(\mathbf{y}_i)$	Theorem 4.12
$\alpha f(\mathbf{x})$ $(\alpha > 0)$	$\alpha f^*(\mathbf{y}/\alpha)$	Theorem 4.14
$\alpha f(\mathbf{x}/\alpha)$ $(\alpha > 0)$	$\alpha f^*(\mathbf{y})$	Theorem 4.14
$f(\mathcal{A}(\mathbf{x}-\mathbf{a}))+\langle \mathbf{b},\mathbf{x}\rangle + c$	$f^*\left((\mathcal{A}^T)^{-1}(\mathbf{y}-\mathbf{b})\right)+\langle \mathbf{a},\mathbf{y}\rangle - c - \langle \mathbf{a},\mathbf{b}\rangle$	Theorem 4.13

Conjugate Functions

f	$\mathrm{dom}(f)$	f^*	Assumptions	Reference
e^x	$\mathbb{R}$	$y\log y - y$ $(\mathrm{dom}(f^*) = \mathbb{R}_+)$	–	Section 4.4.1
$-\log x$	$\mathbb{R}_{++}$	$-1-\log(-y)$ $(\mathrm{dom}(f^*) = \mathbb{R}_{--})$	–	Section 4.4.2
$\max\{1-x,0\}$	$\mathbb{R}$	$y + \delta_{[-1,0]}(y)$	–	Section 4.4.3
$\frac{1}{p}\lvert x\rvert^p$	$\mathbb{R}$	$\frac{1}{q}\lvert y\rvert^q$	$p>1$, $\frac{1}{p}+\frac{1}{q}=1$	Section 4.4.4
$-\frac{x^p}{p}$	$\mathbb{R}_+$	$-\frac{(-y)^q}{q}$ $(\mathrm{dom}(f^*) = \mathbb{R}_{--})$	$0<p<1$, $\frac{1}{p}+\frac{1}{q}=1$	Section 4.4.5
$\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}+\mathbf{b}^T\mathbf{x}+c$	$\mathbb{R}^n$	$\frac{1}{2}(\mathbf{y}-\mathbf{b})^T\mathbf{A}^{-1}(\mathbf{y}-\mathbf{b})-c$	$\mathbf{A}\in\mathbb{S}^n_{++}$, $\mathbf{b}\in\mathbb{R}^n, c\in\mathbb{R}$	Section 4.4.6
$\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}+\mathbf{b}^T\mathbf{x}+c$	$\mathbb{R}^n$	$\frac{1}{2}(\mathbf{y}-\mathbf{b})^T\mathbf{A}^\dagger(\mathbf{y}-\mathbf{b})-c$ $(\mathrm{dom}(f^*) = \mathbf{b}+\mathrm{Range}(\mathbf{A}))$	$\mathbf{A}\in\mathbb{S}^n_{+}$, $\mathbf{b}\in\mathbb{R}^n, c\in\mathbb{R}$	Section 4.4.7
$\sum_{i=1}^n x_i\log x_i$	$\mathbb{R}^n_+$	$\sum_{i=1}^n e^{y_i-1}$	–	Section 4.4.8
$\sum_{i=1}^n x_i\log x_i$	Δ_n	$\log\left(\sum_{i=1}^n e^{y_i}\right)$	–	Section 4.4.10
$-\sum_{i=1}^n \log x_i$	$\mathbb{R}^n_{++}$	$-n-\sum_{i=1}^n\log(-y_i)$ $(\mathrm{dom}(f^*) = \mathbb{R}^n_{--})$	–	Section 4.4.9
$\log\left(\sum_{i=1}^n e^{x_i}\right)$	$\mathbb{R}^n$	$\sum_{i=1}^n y_i\log y_i$ $(\mathrm{dom}(f^*) = \Delta_n)$	–	Section 4.4.11
$\max_i\{x_i\}$	$\mathbb{R}^n$	$\delta_{\Delta_n}(\mathbf{y})$	–	Example 4.10
$\delta_C(\mathbf{x})$	C	$\sigma_C(\mathbf{y})$	$\emptyset\neq C\subseteq\mathbb{E}$	Example 4.2
$\sigma_C(\mathbf{x})$	$\mathrm{dom}(\sigma_C)$	$\delta_{\mathrm{cl}(\mathrm{conv}(C))}(\mathbf{y})$	$\emptyset\neq C\subseteq\mathbb{E}$	Example 4.9
$\lVert\mathbf{x}\rVert$	$\mathbb{E}$	$\delta_{B_{\lVert\cdot\rVert_*}[\mathbf{0},1]}(\mathbf{y})$	–	Section 4.4.12
$-\sqrt{\alpha^2-\lVert\mathbf{x}\rVert^2}$	$B[\mathbf{0},\alpha]$	$\alpha\sqrt{\lVert\mathbf{y}\rVert_*^2+1}$	$\alpha>0$	Section 4.4.13
$\sqrt{\alpha^2+\lVert\mathbf{x}\rVert^2}$	$\mathbb{E}$	$-\alpha\sqrt{1-\lVert\mathbf{y}\rVert_*^2}$ $(\mathrm{dom} f^* = B_{\lVert\cdot\rVert_*}[\mathbf{0},1])$	$\alpha>0$	Section 4.4.14
$\frac{1}{2}\lVert\mathbf{x}\rVert^2$	$\mathbb{E}$	$\frac{1}{2}\lVert\mathbf{y}\rVert_*^2$	–	Section 4.4.15
$\frac{1}{2}\lVert\mathbf{x}\rVert^2+\delta_C(\mathbf{x})$	C	$\frac{1}{2}\lVert\mathbf{y}\rVert^2-\frac{1}{2}d_C^2(\mathbf{y})$	$\emptyset\neq C\subseteq\mathbb{E}$, $\mathbb{E}$ Euclidean	Example 4.4
$\frac{1}{2}\lVert\mathbf{x}\rVert^2-\frac{1}{2}d_C^2(\mathbf{x})$	$\mathbb{E}$	$\frac{1}{2}\lVert\mathbf{y}\rVert^2+\delta_C(\mathbf{y})$	$\emptyset\neq C\subseteq\mathbb{E}$ closed convex. $\mathbb{E}$ Euclidean	Example 4.11

Conjugates of Symmetric Spectral Functions over $\mathbb{S}^n$ (from Example 7.16)

$g(\mathbf{X})$	$\mathrm{dom}(g)$	$g^*(\mathbf{Y})$	$\mathrm{dom}(g^*)$
$\lambda_{\max}(\mathbf{X})$	$\mathbb{S}^n$	$\delta_{\Upsilon_n}(\mathbf{Y})$	Υ_n
$\alpha\|\mathbf{X}\|_F \ (\alpha > 0)$	$\mathbb{S}^n$	$\delta_{B_{\|\cdot\|_F}[\mathbf{0},\alpha]}(\mathbf{Y})$	$B_{\|\cdot\|_F}[\mathbf{0},\alpha]$
$\alpha\|\mathbf{X}\|_F^2 \ (\alpha > 0)$	$\mathbb{S}^n$	$\frac{1}{4\alpha}\|\mathbf{Y}\|_F^2$	$\mathbb{S}^n$
$\alpha\|\mathbf{X}\|_{2,2} \ (\alpha > 0)$	$\mathbb{S}^n$	$\delta_{B_{\|\cdot\|_{S_1}}[\mathbf{0},\alpha]}(\mathbf{Y})$	$B_{\|\cdot\|_{S_1}}[\mathbf{0},\alpha]$
$\alpha\|\mathbf{X}\|_{S_1} \ (\alpha > 0)$	$\mathbb{S}^n$	$\delta_{B_{\|\cdot\|_{2,2}}[\mathbf{0},\alpha]}(\mathbf{Y})$	$B_{\|\cdot\|_{2,2}}[\mathbf{0},\alpha]$
$-\log\det(\mathbf{X})$	$\mathbb{S}^n_{++}$	$-n - \log\det(-\mathbf{Y})$	$\mathbb{S}^n_{--}$
$\sum_{i=1}^n \lambda_i(\mathbf{X})\log(\lambda_i(\mathbf{X}))$	$\mathbb{S}^n_{+}$	$\sum_{i=1}^n e^{\lambda_i(\mathbf{Y})-1}$	$\mathbb{S}^n$
$\sum_{i=1}^n \lambda_i(\mathbf{X})\log(\lambda_i(\mathbf{X}))$	Υ_n	$\log\left(\sum_{i=1}^n e^{\lambda_i(\mathbf{Y})}\right)$	$\mathbb{S}^n$

Conjugates of Symmetric Spectral Functions over $\mathbb{R}^{m\times n}$ (from Example 7.27)

$g(\mathbf{X})$	$\mathrm{dom}(g)$	$g^*(\mathbf{Y})$	$\mathrm{dom}(g^*)$
$\alpha\sigma_1(\mathbf{X}) \ (\alpha > 0)$	$\mathbb{R}^{m\times n}$	$\delta_{B_{\|\cdot\|_{S_1}}[\mathbf{0},\alpha]}(\mathbf{Y})$	$B_{\|\cdot\|_{S_1}}[\mathbf{0},\alpha]$
$\alpha\|\mathbf{X}\|_F \ (\alpha > 0)$	$\mathbb{R}^{m\times n}$	$\delta_{B_{\|\cdot\|_F}[\mathbf{0},\alpha]}(\mathbf{Y})$	$B_{\|\cdot\|_F}[\mathbf{0},\alpha]$
$\alpha\|\mathbf{X}\|_F^2 \ (\alpha > 0)$	$\mathbb{R}^{m\times n}$	$\frac{1}{4\alpha}\|\mathbf{Y}\|_F^2$	$\mathbb{R}^{m\times n}$
$\alpha\|\mathbf{X}\|_{S_1} \ (\alpha > 0)$	$\mathbb{R}^{m\times n}$	$\delta_{B_{\|\cdot\|_{S_\infty}}[\mathbf{0},\alpha]}(\mathbf{Y})$	$B_{\|\cdot\|_{S_\infty}}[\mathbf{0},\alpha]$

Smooth Functions

$f(\mathbf{x})$	$\mathrm{dom}(f)$	Parameter	Norm	Reference
$\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$ $(\mathbf{A}\in\mathbb{S}^n, \mathbf{b}\in\mathbb{R}^n, c\in\mathbb{R})$	$\mathbb{R}^n$	$\|\mathbf{A}\|_{p,q}$	l_p	Example 5.2
$\langle\mathbf{b},\mathbf{x}\rangle + c$ $(\mathbf{b}\in\mathbb{E}^*, c\in\mathbb{R})$	$\mathbb{E}$	0	any norm	Example 5.3
$\frac{1}{2}\|\mathbf{x}\|_p^2, \ p\in[2,\infty)$	$\mathbb{R}^n$	$p-1$	l_p	Example 5.11
$\sqrt{1+\|\mathbf{x}\|_2^2}$	$\mathbb{R}^n$	1	l_2	Example 5.14
$\log(\sum_{i=1}^n e^{xi})$	$\mathbb{R}^n$	1	l_2, l_∞	Example 5.15
$\frac{1}{2}d_C^2(\mathbf{x})$ $(\emptyset\neq C\subseteq\mathbb{E}$ closed convex)	$\mathbb{E}$	1	Euclidean	Example 5.5
$\frac{1}{2}\|\mathbf{x}\|^2 - \frac{1}{2}d_C^2(\mathbf{x})$ $(\emptyset\neq C\subseteq\mathbb{E}$ closed convex)	$\mathbb{E}$	1	Euclidean	Example 5.6
$H_\mu(\mathbf{x}) \ (\mu > 0)$	$\mathbb{E}$	$\frac{1}{\mu}$	Euclidean	Example 6.62

Strongly Convex Functions

$f(\mathbf{x})$	$\mathrm{dom}(f)$	Strongly convex parameter	Norm	Reference
$\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}+2\mathbf{b}^T\mathbf{x}+c$ $(\mathbf{A}\in\mathbb{S}^n_{++},\mathbf{b}\in\mathbb{R}^n,c\in\mathbb{R})$	$\mathbb{R}^n$	$\lambda_{\min}(\mathbf{A})$	l_2	Example 5.19
$\frac{1}{2}\|\mathbf{x}\|^2+\delta_C(\mathbf{x})$ $(\emptyset\neq C\subseteq\mathbb{E}$ convex$)$	C	1	Euclidean	Example 5.21
$-\sqrt{1-\|\mathbf{x}\|_2^2}$	$B_{\|\cdot\|_2}[\mathbf{0},1]$	1	l_2	Example 5.29
$\frac{1}{2}\|\mathbf{x}\|_p^2$ $(p\in(1,2])$	$\mathbb{R}^n$	$p-1$	l_p	Example 5.28
$\sum_{i=1}^n x_i\log x_i$	Δ_n	1	l_2 or l_1	Example 5.27

Orthogonal Projections

Set (C)	$P_C(\mathbf{x})$	Assumptions	Reference
$\mathbb{R}^n_+$	$[\mathbf{x}]_+$	–	Lemma 6.26
$\mathrm{Box}[\boldsymbol{\ell},\mathbf{u}]$	$P_C(\mathbf{x})_i=\min\{\max\{x_i,\ell_i\},u_i\}$	$\ell_i\le u_i$	Lemma 6.26
$B_{\|\cdot\|_2}[\mathbf{c},r]$	$\mathbf{c}+\frac{r}{\max\{\|\mathbf{x}-\mathbf{c}\|_2,r\}}(\mathbf{x}-\mathbf{c})$	$\mathbf{c}\in\mathbb{R}^n,r>0$	Lemma 6.26
$\{\mathbf{x}:\mathbf{A}\mathbf{x}=\mathbf{b}\}$	$\mathbf{x}-\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{x}-\mathbf{b})$	$\mathbf{A}\in\mathbb{R}^{m\times n}$, $\mathbf{b}\in\mathbb{R}^m$, $\mathbf{A}$ full row rank	Lemma 6.26
$\{\mathbf{x}:\mathbf{a}^T\mathbf{x}\le b\}$	$\mathbf{x}-\frac{[\mathbf{a}^T\mathbf{x}-b]_+}{\|\mathbf{a}\|^2}\mathbf{a}$	$\mathbf{0}\neq\mathbf{a}\in\mathbb{R}^n,b\in\mathbb{R}$	Lemma 6.26
Δ_n	$[\mathbf{x}-\mu^*\mathbf{e}]_+$ where $\mu^*\in\mathbb{R}$ satisfies $\mathbf{e}^T[\mathbf{x}-\mu^*\mathbf{e}]_+=1$		Corollary 6.29
$H_{\mathbf{a},b}\cap\mathrm{Box}[\boldsymbol{\ell},\mathbf{u}]$	$P_{\mathrm{Box}[\boldsymbol{\ell},\mathbf{u}]}(\mathbf{x}-\mu^*\mathbf{a})$ where $\mu^*\in\mathbb{R}$ satisfies $\mathbf{a}^TP_{\mathrm{Box}[\boldsymbol{\ell},\mathbf{u}]}(\mathbf{x}-\mu^*\mathbf{a})=b$	$\mathbf{a}\in\mathbb{R}^n\backslash\{\mathbf{0}\},b\in\mathbb{R}$	Theorem 6.27
$H^-_{\mathbf{a},b}\cap\mathrm{Box}[\boldsymbol{\ell},\mathbf{u}]$	$\begin{cases}P_{\mathrm{Box}[\boldsymbol{\ell},\mathbf{u}]}(\mathbf{x}), & \mathbf{a}^T\mathbf{v}_{\mathbf{x}}\le b,\\ P_{\mathrm{Box}[\boldsymbol{\ell},\mathbf{u}]}(\mathbf{x}-\lambda^*\mathbf{a}), & \mathbf{a}^T\mathbf{v}_{\mathbf{x}}>b,\end{cases}$ $\mathbf{v}_{\mathbf{x}}=P_{\mathrm{Box}[\boldsymbol{\ell},\mathbf{u}]}(\mathbf{x})$, $\mathbf{a}^TP_{\mathrm{Box}[\boldsymbol{\ell},\mathbf{u}]}(\mathbf{x}-\lambda^*\mathbf{a})=b,\lambda^*>0$	$\mathbf{a}\in\mathbb{R}^n\backslash\{\mathbf{0}\},b\in\mathbb{R}$	Example 6.32
$B_{\|\cdot\|_1}[\mathbf{0},\alpha]$	$\begin{cases}\mathbf{x}, & \|\mathbf{x}\|_1\le\alpha,\\ \mathcal{T}_{\lambda^*}(\mathbf{x}), & \|\mathbf{x}\|_1>\alpha,\end{cases}$ $\|\mathcal{T}_{\lambda^*}(\mathbf{x})\|_1=\alpha,\ \lambda^*>0$	$\alpha>0$	Example 6.33
$\{\mathbf{x}:\boldsymbol{\omega}^T\|\mathbf{x}\|\le\beta,$ $-\boldsymbol{\alpha}\le\mathbf{x}\le\boldsymbol{\alpha}\}$	$\begin{cases}\mathbf{v}_{\mathbf{x}}, & \boldsymbol{\omega}^T\|\mathbf{v}_{\mathbf{x}}\|\le\beta,\\ \mathcal{S}_{\lambda^*\boldsymbol{\omega},\boldsymbol{\alpha}}(\mathbf{x}), & \boldsymbol{\omega}^T\|\mathbf{v}_{\mathbf{x}}\|>\beta,\end{cases}$ $\mathbf{v}_{\mathbf{x}}=P_{\mathrm{Box}[-\boldsymbol{\alpha},\boldsymbol{\alpha}]}(\mathbf{x})$, $\boldsymbol{\omega}^T\|\mathcal{S}_{\lambda^*\boldsymbol{\omega},\boldsymbol{\alpha}}(\mathbf{x})\|=\beta,\lambda^*>0$	$\boldsymbol{\omega}\in\mathbb{R}^n_+$, $\boldsymbol{\alpha}\in[0,\infty]^n$, $\beta\in\mathbb{R}_{++}$	Example 6.34
$\{\mathbf{x}>\mathbf{0}:\Pi x_i\ge\alpha\}$	$\begin{cases}\mathbf{x}, & \mathbf{x}\in C,\\ \left(\frac{x_j+\sqrt{x_j^2+4\lambda^*}}{2}\right)_{j=1}^n, & \mathbf{x}\notin C,\end{cases}$ $\Pi_{j=1}^n\left((x_j+\sqrt{x_j^2+4\lambda^*})/2\right)=\alpha,\lambda^*>0$	$\alpha>0$	Example 6.35
$\{(\mathbf{x},s):\|\mathbf{x}\|_2\le s\}$	$\left(\frac{\|\mathbf{x}\|_2+s}{2\|\mathbf{x}\|_2}\mathbf{x},\frac{\|\mathbf{x}\|_2+s}{2}\right)$ if $\|\mathbf{x}\|_2\ge\|s\|$ $(\mathbf{0},0)$ if $s<\|\mathbf{x}\|_2<-s$, $(\mathbf{x},s)$ if $\|\mathbf{x}\|_2\le s$.	–	Example 6.37
$\{(\mathbf{x},s):\|\mathbf{x}\|_1\le s\}$	$\begin{cases}(\mathbf{x},s), & \|\mathbf{x}\|_1\le s,\\ (\mathcal{T}_{\lambda^*}(\mathbf{x}),s+\lambda^*), & \|\mathbf{x}\|_1>s,\end{cases}$ $\|\mathcal{T}_{\lambda^*}(\mathbf{x})\|_1-\lambda^*-s=0,\ \lambda^*>0$	–	Example 6.38

Orthogonal Projections onto Symmetric Spectral Sets in $\mathbb{S}^n$

set (T)	$P_T(\mathbf{X})$	Assumptions
$\mathbb{S}^n_+$	$\mathbf{U}\text{diag}([\boldsymbol{\lambda}(\mathbf{X})]_+)\mathbf{U}^T$	–
$\{\mathbf{X} : \ell\mathbf{I} \preceq \mathbf{X} \preceq u\mathbf{I}\}$	$\mathbf{U}\text{diag}(\mathbf{v})\mathbf{U}^T$, $v_i = \min\{\max\{\lambda_i(\mathbf{X}), \ell\}, u\}$	$\ell \leq u$
$B_{\|\cdot\|_F}[\mathbf{0}, r]$	$\frac{r}{\max\{\|\mathbf{X}\|_F, r\}}\mathbf{X}$	$r > 0$
$\{\mathbf{X} : \text{Tr}(\mathbf{X}) \leq b\}$	$\mathbf{U}\text{diag}(\mathbf{v})\mathbf{U}^T$, $\mathbf{v} = \boldsymbol{\lambda}(\mathbf{X}) - \frac{[\mathbf{e}^T\boldsymbol{\lambda}(\mathbf{X})-b]_+}{n}\mathbf{e}$	$b \in \mathbb{R}$
Υ_n	$\mathbf{U}\text{diag}(\mathbf{v})\mathbf{U}^T$, $\mathbf{v} = [\boldsymbol{\lambda}(\mathbf{X}) - \mu^*\mathbf{e}]_+$ where $\mu^* \in \mathbb{R}$ satisfies $\mathbf{e}^T[\boldsymbol{\lambda}(\mathbf{X}) - \mu^*\mathbf{e}]_+ = 1$	–
$B_{\|\cdot\|_{S_1}}[\mathbf{0}, \alpha]$	$\begin{cases} \mathbf{X}, & \|\mathbf{X}\|_{S_1} \leq \alpha, \\ \mathbf{U}\text{diag}(\mathcal{T}_{\beta^*}(\boldsymbol{\lambda}(\mathbf{X})))\mathbf{U}^T, & \|\mathbf{X}\|_{S_1} > \alpha, \end{cases}$ $\|\mathcal{T}_{\beta^*}(\boldsymbol{\lambda}(\mathbf{X}))\|_1 = \alpha$, $\beta^* > 0$	$\alpha > 0$

Orthogonal Projections onto Symmetric Spectral Sets in $\mathbb{R}^{m\times n}$ (from Example 7.31)

set (T)	$P_T(\mathbf{X})$	Assumptions
$B_{\|\cdot\|_{S_\infty}}[\mathbf{0}, \alpha]$	$\mathbf{U}\text{diag}(\mathbf{v})\mathbf{V}^T$, $v_i = \min\{\sigma_i(\mathbf{X}), \alpha\}$	$\alpha > 0$
$B_{\|\cdot\|_F}[\mathbf{0}, r]$	$\frac{r}{\max\{\|\mathbf{X}\|_F, r\}}\mathbf{X}$	$r > 0$
$B_{\|\cdot\|_{S_1}}[\mathbf{0}, \alpha]$	$\begin{cases} \mathbf{X}, & \|\mathbf{X}\|_{S_1} \leq \alpha, \\ \mathbf{U}\text{diag}(\mathcal{T}_{\beta^*}(\sigma(\mathbf{X})))\mathbf{V}^T, & \|\mathbf{X}\|_{S_1} > \alpha, \end{cases}$ $\|\mathcal{T}_{\beta^*}(\boldsymbol{\sigma}(\mathbf{X}))\|_1 = \alpha$, $\beta^* > 0$	$\alpha > 0$

Prox Calculus Rules

$f(\mathbf{x})$	$\text{prox}_f(\mathbf{x})$	Assumptions	Reference
$\sum_{i=1}^m f_i(\mathbf{x}_i)$	$\text{prox}_{f_1}(\mathbf{x}_1) \times \cdots \times \text{prox}_{f_m}(\mathbf{x}_m)$	–	Theorem 6.6
$g(\lambda\mathbf{x} + \mathbf{a})$	$\frac{1}{\lambda}\left[\text{prox}_{\lambda^2 g}(\lambda\mathbf{x} + \mathbf{a}) - \mathbf{a}\right]$	$\lambda \neq 0, \mathbf{a} \in \mathbb{E}$, g proper	Theorem 6.11
$\lambda g(\mathbf{x}/\lambda)$	$\lambda\text{prox}_{g/\lambda}(\mathbf{x}/\lambda)$	$\lambda \neq 0$, g proper	Theorem 6.12
$g(\mathbf{x}) + \frac{c}{2}\|\mathbf{x}\|^2 + \langle\mathbf{a}, \mathbf{x}\rangle + \gamma$	$\text{prox}_{\frac{1}{c+1}g}(\frac{\mathbf{x}-\mathbf{a}}{c+1})$	$\mathbf{a} \in \mathbb{E}, c > 0, \gamma \in \mathbb{R}$, g proper	Theorem 6.13
$g(\mathcal{A}(\mathbf{x}) + \mathbf{b})$	$\mathbf{x} + \frac{1}{\alpha}\mathcal{A}^T(\text{prox}_{\alpha g}(\mathcal{A}(\mathbf{x}) + \mathbf{b}) - \mathcal{A}(\mathbf{x}) - \mathbf{b})$	$\mathbf{b} \in \mathbb{R}^m$, $\mathcal{A} : \mathbb{V} \to \mathbb{R}^m$, g proper closed convex, $\mathcal{A} \circ \mathcal{A}^T = \alpha I$, $\alpha > 0$	Theorem 6.15
$g(\|\mathbf{x}\|)$	$\text{prox}_g(\|\mathbf{x}\|)\frac{\mathbf{x}}{\|\mathbf{x}\|}$, $\mathbf{x} \neq \mathbf{0}$ $\{\mathbf{u} : \|\mathbf{u}\| = \text{prox}_g(0)\}$, $\mathbf{x} = \mathbf{0}$	g proper closed convex, $\text{dom}(g) \subseteq [0, \infty)$	Theorem 6.18

Prox Computations

$f(\mathbf{x})$	$\mathrm{dom}(f)$	$\mathrm{prox}_f(\mathbf{x})$	Assumptions	Reference
$\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$	$\mathbb{R}^n$	$(\mathbf{A}+\mathbf{I})^{-1}(\mathbf{x}-\mathbf{b})$	$\mathbf{A} \in \mathbb{S}^n_+$, $\mathbf{b} \in \mathbb{R}^n$, $c \in \mathbb{R}$	Section 6.2.3
λx^3	$\mathbb{R}_+$	$\frac{-1+\sqrt{1+12\lambda[x]_+}}{6\lambda}$	$\lambda > 0$	Lemma 6.5
μx	$[0,\alpha] \cap \mathbb{R}$	$\min\{\max\{x-\mu,0\},\alpha\}$	$\mu \in \mathbb{R}$, $\alpha \in [0,\infty]$	Example 6.14
$\lambda\|\mathbf{x}\|$	$\mathbb{E}$	$\left(1-\frac{\lambda}{\max\{\|\mathbf{x}\|,\lambda\}}\right)\mathbf{x}$	$\|\cdot\|$—Euclidean norm, $\lambda > 0$	Example 6.19
$-\lambda\|\mathbf{x}\|$	$\mathbb{E}$	$\left(1+\frac{\lambda}{\|\mathbf{x}\|}\right)\mathbf{x}$, $\mathbf{x} \neq \mathbf{0}$; $\{\mathbf{u} : \|\mathbf{u}\| = \lambda\}$, $\mathbf{x} = \mathbf{0}$.	$\|\cdot\|$—Euclidean norm, $\lambda > 0$	Example 6.21
$\lambda\|\mathbf{x}\|_1$	$\mathbb{R}^n$	$\mathcal{T}_\lambda(\mathbf{x}) = [\|\mathbf{x}\| - \lambda\mathbf{e}]_+ \odot \mathrm{sgn}(\mathbf{x})$	$\lambda > 0$	Example 6.8
$\|\boldsymbol{\omega} \odot \mathbf{x}\|_1$	$\mathrm{Box}[-\boldsymbol{\alpha},\boldsymbol{\alpha}]$	$\mathcal{S}_{\boldsymbol{\omega},\boldsymbol{\alpha}}(\mathbf{x})$	$\boldsymbol{\alpha} \in [0,\infty]^n, \boldsymbol{\omega} \in \mathbb{R}^n_+$	Example 6.23
$\lambda\|\mathbf{x}\|_\infty$	$\mathbb{R}^n$	$\mathbf{x} - \lambda P_{B_{\|\cdot\|_1}[\mathbf{0},1]}(\mathbf{x}/\lambda)$	$\lambda > 0$	Example 6.48
$\lambda\|\mathbf{x}\|_a$	$\mathbb{E}$	$\mathbf{x} - \lambda P_{B_{\|\cdot\|_{a,*}}[\mathbf{0},1]}(\mathbf{x}/\lambda)$	$\|\mathbf{x}\|_a$—norm, $\lambda > 0$	Example 6.47
$\lambda\|\mathbf{x}\|_0$	$\mathbb{R}^n$	$\mathcal{H}_{\sqrt{2\lambda}}(x_1) \times \cdots \times \mathcal{H}_{\sqrt{2\lambda}}(x_n)$	$\lambda > 0$	Example 6.10
$\lambda\|\mathbf{x}\|^3$	$\mathbb{E}$	$\frac{2}{1+\sqrt{1+12\lambda\|\mathbf{x}\|}}\mathbf{x}$	$\|\cdot\|$—Euclidean norm, $\lambda > 0$,	Example 6.20
$-\lambda\sum_{j=1}^n \log x_j$	$\mathbb{R}^n_{++}$	$\left(\frac{x_j+\sqrt{x_j^2+4\lambda}}{2}\right)_{j=1}^n$	$\lambda > 0$	Example 6.9
$\delta_C(\mathbf{x})$	$\mathbb{E}$	$P_C(\mathbf{x})$	$\emptyset \neq C \subseteq \mathbb{E}$	Theorem 6.24
$\lambda\sigma_C(\mathbf{x})$	$\mathbb{E}$	$\mathbf{x} - \lambda P_C(\mathbf{x}/\lambda)$	$\lambda > 0$, $C \neq \emptyset$ closed convex	Theorem 6.46
$\lambda\max\{x_i\}$	$\mathbb{R}^n$	$\mathbf{x} - P_{\Delta_n}(\mathbf{x}/\lambda)$	$\lambda > 0$	Example 6.49
$\lambda\sum_{i=1}^k x_{[i]}$	$\mathbb{R}^n$	$\mathbf{x} - \lambda P_C(\mathbf{x}/\lambda)$, $C = H_{\mathbf{e},k} \cap \mathrm{Box}[\mathbf{0},\mathbf{e}]$	$\lambda > 0$	Example 6.50
$\lambda\sum_{i=1}^k \|x_{\langle i\rangle}\|$	$\mathbb{R}^n$	$\mathbf{x} - \lambda P_C(\mathbf{x}/\lambda)$, $C = B_{\|\cdot\|_1}[\mathbf{0},k] \cap \mathrm{Box}[-\mathbf{e},\mathbf{e}]$	$\lambda > 0$	Example 6.51
$\lambda M_f^\mu(\mathbf{x})$	$\mathbb{E}$	$\mathbf{x} + \frac{\lambda}{\mu+\lambda}\left(\mathrm{prox}_{(\mu+\lambda)f}(\mathbf{x}) - \mathbf{x}\right)$	$\lambda, \mu > 0$, f proper closed convex	Corollary 6.64
$\lambda d_C(\mathbf{x})$	$\mathbb{E}$	$\mathbf{x} + \min\left\{\frac{\lambda}{d_C(\mathbf{x})}, 1\right\}(P_C(\mathbf{x}) - \mathbf{x})$	$\emptyset \neq C$ closed convex, $\lambda > 0$	Lemma 6.43
$\frac{\lambda}{2}d_C^2(\mathbf{x})$	$\mathbb{E}$	$\frac{\lambda}{\lambda+1}P_C(\mathbf{x}) + \frac{1}{\lambda+1}\mathbf{x}$	$\emptyset \neq C$ closed convex, $\lambda > 0$	Example 6.65
$\lambda H_\mu(\mathbf{x})$	$\mathbb{E}$	$\left(1-\frac{\lambda}{\max\{\|\mathbf{x}\|,\mu+\lambda\}}\right)\mathbf{x}$	$\lambda, \mu > 0$	Example 6.66
$\rho\|\mathbf{x}\|_1^2$	$\mathbb{R}^n$	$\left(\frac{v_i x_i}{v_i+2\rho}\right)_{i=1}^n$, $\mathbf{v} = \left[\sqrt{\frac{\rho}{\mu}}\|\mathbf{x}\| - 2\rho\right]_+$, $\mathbf{e}^T\mathbf{v} = 1$ ($\mathbf{0}$ when $\mathbf{x} = \mathbf{0}$)	$\rho > 0$	Lemma 6.70
$\lambda\|\mathbf{A}\mathbf{x}\|_2$	$\mathbb{R}^n$	$\mathbf{x} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \alpha^*\mathbf{I})^{-1}\mathbf{A}\mathbf{x}$, $\alpha^* = 0$ if $\|\mathbf{v}_0\|_2 \leq \lambda$; otherwise, $\|\mathbf{v}_{\alpha^*}\|_2 = \lambda$; $\mathbf{v}_\alpha \equiv (\mathbf{A}\mathbf{A}^T + \alpha\mathbf{I})^{-1}\mathbf{A}\mathbf{x}$	$\mathbf{A} \in \mathbb{R}^{m\times n}$ with full row rank, $\lambda > 0$	Lemma 6.68

Prox of Symmetric Spectral Functions over $\mathbb{S}^n$ (from Example 7.19)

$F(\mathbf{X})$	$\mathrm{dom}(F)$	$\mathrm{prox}_F(\mathbf{X})$	Reference
$\alpha\|\mathbf{X}\|_F^2$	$\mathbb{S}^n$	$\frac{1}{1+2\alpha}\mathbf{X}$	Section 6.2.3
$\alpha\|\mathbf{X}\|_F$	$\mathbb{S}^n$	$\left(1-\frac{\alpha}{\max\{\|\mathbf{X}\|_F,\alpha\}}\right)\mathbf{X}$	Example 6.19
$\alpha\|\mathbf{X}\|_{S_1}$	$\mathbb{S}^n$	$\mathbf{U}\mathrm{diag}(\mathcal{T}_\alpha(\boldsymbol{\lambda}(\mathbf{X})))\mathbf{U}^T$	Example 6.8
$\alpha\|\mathbf{X}\|_{2,2}$	$\mathbb{S}^n$	$\mathbf{U}\mathrm{diag}(\boldsymbol{\lambda}(\mathbf{X})-\alpha P_{B_{\|\cdot\|_1}[\mathbf{0},1]}(\boldsymbol{\lambda}(\mathbf{X})/\alpha))\mathbf{U}^T$	Example 6.48
$-\alpha\log\det(\mathbf{X})$	$\mathbb{S}^n_{++}$	$\mathbf{U}\mathrm{diag}\left(\frac{\lambda_j(\mathbf{X})+\sqrt{\lambda_j(\mathbf{X})^2+4\alpha}}{2}\right)\mathbf{U}^T$	Example 6.9
$\alpha\lambda_1(\mathbf{X})$	$\mathbb{S}^n$	$\mathbf{U}\mathrm{diag}(\boldsymbol{\lambda}(\mathbf{X})-\alpha P_{\Delta_n}(\boldsymbol{\lambda}(\mathbf{X})/\alpha))\mathbf{U}^T$	Example 6.49
$\alpha\sum_{i=1}^k\lambda_i(\mathbf{X})$	$\mathbb{S}^n$	$\mathbf{X}-\alpha\mathbf{U}\mathrm{diag}(P_C(\boldsymbol{\lambda}(\mathbf{X})/\alpha))\mathbf{U}^T$, $C=H_{\mathbf{e},k}\cap\mathrm{Box}[\mathbf{0},\mathbf{e}]$	Example 6.50

Prox of Symmetric Spectral Functions over $\mathbb{R}^{m\times n}$ (from Example 7.30)

$F(\mathbf{X})$	$\mathrm{prox}_F(\mathbf{X})$
$\alpha\|\mathbf{X}\|_F^2$	$\frac{1}{1+2\alpha}\mathbf{X}$
$\alpha\|\mathbf{X}\|_F$	$\left(1-\frac{\alpha}{\max\{\|\mathbf{X}\|_F,\alpha\}}\right)\mathbf{X}$
$\alpha\|\mathbf{X}\|_{S_1}$	$\mathbf{U}\mathrm{dg}(\mathcal{T}_\alpha(\boldsymbol{\sigma}(\mathbf{X})))\mathbf{V}^T$
$\alpha\|\mathbf{X}\|_{S_\infty}$	$\mathbf{X}-\alpha\mathbf{U}\mathrm{dg}(P_{B_{\|\cdot\|_1}[\mathbf{0},1]}(\boldsymbol{\sigma}(\mathbf{X})/\alpha))\mathbf{V}^T$
$\alpha\|\mathbf{X}\|_{\langle k\rangle}$	$\mathbf{X}-\alpha\mathbf{U}\mathrm{dg}(P_C(\boldsymbol{\sigma}(\mathbf{X})/\alpha))\mathbf{V}^T$, $C=B_{\|\cdot\|_1}[\mathbf{0},k]\cap B_{\|\cdot\|_\infty}[\mathbf{0},1]$

Appendix C

Symbols and Notation

Vector Spaces

$\mathbb{E}, \mathbb{V}$		underlying vector spaces
$\mathbb{E}^*$	p. 9	dual space of $\mathbb{E}$
$\|\cdot\|_*$	p. 9	dual norm
$\dim(V)$	p. 2	dimension of a vector space V
$\operatorname{aff}(S)$	p. 3	affine hull of a set S
$\|\cdot\|$	p. 2	norm
$\|\cdot\|_{\mathbb{E}}$	p. 2	norm of a vector space $\mathbb{E}$
$\langle \mathbf{x}, \mathbf{y} \rangle$	p. 2	inner product of $\mathbf{x}$ and $\mathbf{y}$
$\mathbb{R}^n$	p. 4	space of n-dimensional real column vectors
$[\mathbf{x}, \mathbf{y}]$	p. 3	closed line segment between $\mathbf{x}$ and $\mathbf{y}$
$(\mathbf{x}, \mathbf{y})$	p. 3	open line segment between $\mathbf{x}$ and $\mathbf{y}$
$B(\mathbf{c}, r), B_{\|\cdot\|}(\mathbf{c}, r)$	p. 2	open ball with center $\mathbf{c}$ and radius r
$B[\mathbf{c}, r], B_{\|\cdot\|}[\mathbf{c}, r]$	p. 2	closed ball with center $\mathbf{c}$ and radius r
$\mathbb{R}^{m \times n}$	p. 6	space of $m \times n$ real-valued matrices
$\mathcal{A}^T$	p. 11	adjoint of the linear transformation $\mathcal{A}$
$\mathcal{I}$	p. 8	identity transformation

The Space $\mathbb{R}^n$

$\mathbf{e}_i$	p. 4	ith vector in the standard basis of $\mathbb{R}^n$
$\mathbf{e}$	p. 4	vector of all ones
$\mathbf{0}$	p. 4	vector of all zeros
$\|\cdot\|_p$	p. 5	l_p-norm
Δ_n	p. 5	unit simplex
$\mathrm{Box}[\boldsymbol{\ell}, \mathbf{u}]$	pp. 5, 147	box with lower bounds $\boldsymbol{\ell}$ and upper bounds $\mathbf{u}$
$\mathbb{R}^n_+$	p. 5	nonnegative orthant
$\mathbb{R}^n_{++}$	p. 5	positive orthant
$H_{\mathbf{a},b}$	p. 3	the hyperplane $\{\mathbf{x} : \langle \mathbf{a}, \mathbf{x} \rangle = b\}$
$H^-_{\mathbf{a},b}$	p. 3	the half-space $\{\mathbf{x} : \langle \mathbf{a}, \mathbf{x} \rangle \leq b\}$
$[\mathbf{x}]_+$	p. 5	nonnegative part of $\mathbf{x}$
$\lvert\mathbf{x}\rvert$	p. 5	absolute values vector of $\mathbf{x}$
$\mathrm{sgn}(\mathbf{x})$	p. 5	sign vector of $\mathbf{x}$
$\mathbf{a} \odot \mathbf{b}$	p. 5	Hadamard product
$\mathbf{x}^{\downarrow}$	p. 180	$\mathbf{x}$ reordered nonincreasingly

The Space $\mathbb{R}^{m\times n}$

$\mathbb{S}^n$	p. 6	set of all $n \times n$ symmetric matrices
$\mathbb{S}^n_+$	p. 6	set of all $n \times n$ positive semidefinite matrices
$\mathbb{S}^n_{++}$	p. 6	set of all $n \times n$ positive definite matrices
$\mathbb{S}^n_-$	p. 6	set of all $n \times n$ negative semidefinite matrices
$\mathbb{S}^n_{--}$	p. 6	set of all $n \times n$ negative definite matrices
$\mathbb{O}_n$	p. 6	set of all $n \times n$ orthogonal matrices
Υ_n	p. 183	spectahedron
$\|\mathbf{A}\|_F$	p. 6	Frobenius norm of $\mathbf{A}$
$\|\mathbf{A}\|_{S_p}$	p. 189	Schatten p-norm of $\mathbf{A}$
$\|\mathbf{A}\|_{\langle k\rangle}$	p. 190	Ky Fan k-norm of $\mathbf{A}$
$\|\mathbf{A}\|_{ab}$	p. 7	induced norm of $\mathbf{A} \in \mathbb{R}^{m\times n}$ when $\mathbb{R}^n$ and $\mathbb{R}^m$ the norms $\|\cdot\|_a$ and $\|\cdot\|_b$ respectively
$\|\mathbf{A}\|_2$	p. 7	spectral norm of $\mathbf{A}$
$\lambda_{\max}(\mathbf{A})$		maximum eigenvalue of a symmetric matrix $\mathbf{A}$
$\lambda_{\min}(\mathbf{A})$		maximum eigenvalue of a symmetric matrix $\mathbf{A}$

Sets

$\operatorname{int}(S)$		interior of S
$\operatorname{cl}(S)$		closure of S
$\operatorname{conv}(S)$		convex hull of S
$A + B$	p. 26	Minkowski sum of A and B
K°	p. 27	polar cone of K
$N_S(\mathbf{x})$	p. 36	normal cone of S at $\mathbf{x}$
$\operatorname{ri}(S)$	p. 43	relative interior of S
$\#A$		number of elements in A
Λ_n	p. 180	$n \times n$ permutation matrices
Λ_n^G	p. 180	$n \times n$ generalized permutation matrices

Functions and Operators

$\log x$		natural logarithm of x
$\mathrm{dom}(f)$	p. 14	(effective) domain of f
δ_C	p. 14	indicator function of the set C
$\mathrm{epi}(f)$	p. 14	epigraph of f
$\mathrm{Lev}(f, \alpha)$	p. 15	α-level set of f
d_C	p. 22	distance function to C
σ_C	p. 26	support function of C
$h_1 \square h_2$	p. 24	infimal convolution of h_1 and h_2
$\partial f(\mathbf{x})$	p. 35	subdifferential set of f at $\mathbf{x}$
$f'(\mathbf{x})$	p. 202	subgradient of f at $\mathbf{x}$
$\mathrm{dom}(\partial f)$	p. 40	set of points of differentiability
$f'(\mathbf{x}; \mathbf{d})$	p. 44	directional derivative of f at $\mathbf{x}$ in the direction $\mathbf{d}$
$\nabla f(\mathbf{x})$	p. 48	gradient of f at $\mathbf{x}$
$\nabla^2 f(\mathbf{x})$		Hessian of a function over $\mathbb{R}^n$ at $\mathbf{x}$
P_C	p. 49	orthogonal projection on C
$f \circ g$		f composed with g
f^*	p. 87	conjugate of f
$C_L^{1,1}(D)$	p. 107	class of L-smooth functions over D
$\mathrm{prox}_f(\mathbf{x})$	p. 129	proximal mapping of f evaluated at $\mathbf{x}$
$\mathcal{T}_\lambda(\mathbf{x})$	p. 136	soft thresholding with level λ evaluated at $\mathbf{x}$
$\mathcal{S}_{\mathbf{a},\mathbf{b}}$	p. 151	two-sided soft thresholding
H_μ	p. 163	Huber function with smoothing parameter μ
$T_L^{f,g}(\mathbf{x}), T_L(\mathbf{x})$	p. 271	prox-grad mapping evaluated at $\mathbf{x}$
$G_L^{f,g}(\mathbf{x}), G_L(\mathbf{x})$	p. 273	gradient mapping evaluated at $\mathbf{x}$

Matrices

$\mathbf{A}^\dagger$		Moore–Penrose pseudoinverse
$\lambda_{\max}(\mathbf{A})$		maximum eigenvalue of $\mathbf{A}$
$\lambda_{\min}(\mathbf{A})$		minimum eigenvalue of $\mathbf{A}$
$\sigma_{\max}(\mathbf{A})$		maximum singular of $\mathbf{A}$
Range($\mathbf{A}$)		range of $\mathbf{A}$—all linear combinations of the columns of $\mathbf{A}$
Null($\mathbf{A}$)		null space/kernel of $\mathbf{A}$
diag($\mathbf{x}$)		diagonal matrix with diagonal $\mathbf{x}$
dg($\mathbf{x}$)	p. 188	generalized diagonal matrix with diagonal $\mathbf{x}$

Appendix D

Bibliographic Notes

Chapter 1. For a comprehensive treatment of finite-dimensional vector spaces and advanced linear algebra topics, the reader can refer to the classical book of Halmos [64], as well as to the textbooks of Meyer [86] and Strang [117].

Chapters 2, 3, 4. Most of the material in these chapters is classical. Additional materials and extensions can be found, for example, in Bauschke and Combettes [8], Bertsekas [29], Borwein and Lewis [32], Hiriart-Urruty and Lemaréchal [67], Nesterov [94] and Rockafellar [108]. Example 2.17 is taken from the book of Hiriart-Urruty and Lemaréchal [67, Example 2.1.4]. Example 2.32 is from Rockafellar [108, p. 83]. The proof in Example 3.31 follows Beck and Teboulle [20, Theorem 4.1]. Section 3.5, excluding Theorem 3.60, follows Hiriart-Urruty and Lemaréchal [67, Section VII.3.3]. Theorem 3.60 is a slight extension of Lemma 6 from Lan [78]. The optimality conditions derived in Example 3.66 are rather old and can be traced back to Sturm, who proved them in his work from 1884 [118]. Actually, (re)proving these conditions was the main motivation for Weiszfeld to devise the (now-called) Weiszfeld's method in 1937 [124]. For more information on the Fermat–Weber problem and Weiszfeld's method, see the review paper of Beck and Sabach [14] and references therein.

Chapter 5. The proof of the descent lemma can be found in Bertsekas [28]. The proof of Theorem 5.8 follows the proof of Nesterov in [94, Theorem 2.1.5]. The equivalence between claims (i) and (iv) in Theorem 5.8 is also known as the Baillon-Haddad theorem [5]. The analysis in Example 5.11 of the smoothness parameter of the squared l_p-norm follows the derivation in the work of Ben-Tal, Margalit, and Nemirovski [24, Appendix 1]. The conjugate correspondence theorem can be deduced from the work of Zalinescu [128, Theorem 2.2] and can also be found in the paper of Azé and Penot [3] as well as Zalinescu's book [129, Corollary 3.5.11]. In its Euclidean form, the result can be found in the book of Rockafellar and Wets [111, Proposition 12.60]. Further characterizations appear in the paper of Bauschke and Combettes [7]. The proof of Theorem 5.30 follows Beck and Teboulle [20, Theorem 4.1].

Chapter 6. The seminal 1965 paper of Moreau [87] already contains much of the properties of the proximal mapping discussed in the chapter. Excellent references for the subject are the book of Bauschke and Combettes [8], the paper of Combettes and Wajs [44], and the review paper of Parikh and Boyd [102]. The computation of the prox of the squared l_1-norm in Section 6.8.2 is due to Evgeniou, Pontil, Spinellis, and Nassuphis [54].

Chapter 7. The notion of symmetry w.r.t. a given set of orthogonal matrices was studied by Rockafellar [108, Chapter 12]. A variant of the symmetric conjugate theorem (Theorem 7.9) can be found in Rockafellar [108, Corollary 12.3.1]. Fan's inequality can be found in Theobald [119]. Von Neumann's trace inequality [123], as well as Fan's inequality, are often formulated over the complex field, but the adaptation to the real field is straightforward. Sections 7.2 and 7.3, excluding the spectral proximal theorem, are based on the seminal papers of Lewis [80, 81] on unitarily invariant functions. See also Borwein and Lewis [32, Section 1.2], as well as Borwein and Vanderwerff [33, Section 3.2]. The equivalence between the convexity of spectral functions and their associated functions was first established by Davis in [47]. The spectral proximal formulas can be found in Parikh and Boyd [102].

Chapter 8. Example 8.3 is taken from Vandenberghe's lecture notes [122]. Wolfe's example with $\gamma = \frac{16}{9}$ originates from his work [125]. The version with general $\gamma > 1$, along with the support form of the function, can be found in the set of exercises [35]. Studies of subgradient methods and extensions can be found in many books; to name a few, the books of Nemirovsky and Yudin [92], Shor [116] and Polyak [104] are classical; modern accounts of the subject can be found, for example, in Bertsekas [28, 29, 30], Nesterov [94], and Ruszczyński [113]. The analysis of the stochastic and deterministic projected subgradient method in the strongly convex case is based on the work of Lacoste-Julien, Schmidt, and Bach [77]. The fundamental inequality for the incremental projected subgradient is taken from Nedić and Bertsekas [89], where many additional results on incremental methods are derived. Theorem 8.42 and Lemma 8.47 are Lemmas 1 and 3 from the work of from Nedić and Ozdaglar [90]. The latter work also contains additional results on the dual projected subgradient method with constant stepsize. The presentation of the network utility maximization problem, as well as the distributed subgradient method for solving it, originates from Nedić and Ozdaglar [91].

Chapter 9. The mirror descent method was introduced by Nemirovsky and Yudin in [92]. The interpretation of the method as a non-Euclidean projected subgradient method was presented by Beck and Teboulle in [15]. The rate of convergence analysis of the mirror descent method is based on [15]. The three-points lemma was proven by Chen and Teboulle in [43]. The analysis of the mirror-C method is based on the work of Duchi, Shalev-Shwartz, Singer, and Tewari [49], where the algorithm is introduced in an online and stochastic setting.

Chapter 10. The proximal gradient method can be traced back to the forward-backward algorithm introduced by Bruck [36], Pasty [103], and Lions and Mercier [83]. More modern accounts of the topic can be found, for example, in Bauschke and Combettes [8, Chapter 27], Combettes and Wajs [44], and Facchinei and Pang

[55, Chapter 12]. The proximal gradient method is a generalization of the gradient method, which goes back to Cauchy [38] and was extensively studied and generalized by many authors; see, for example, the books of Bertsekas [28], Nesterov [94], Polyak [104], and Nocedal and Wright [99], as well as the many references therein. ISTA and its variations was studied in the literature in several contexts; see, for example, the works of Daubechies, Defrise, and De Mol [46]; Hale, Yin, and Zhang [63]; Wright, Nowak, and Figueiredo [127]; and Elad [52]. The analysis of the proximal gradient method in Sections 10.3 and 10.4 mostly follows the presentation of Beck and Teboulle in [18] and [19]. Lemma 10.11 was stated and proved for the case where g is an indicator of a nonempty closed and convex set in [9]; see also [13, Lemma 2.3]. Theorem 10.9 on the monotonicity of the gradient mapping is a simple generalization of [10, Lemma 9.12]. The first part of the monotonicity result was shown in the case where g is an indicator of a nonempty closed and convex set in Bertsekas [28, Lemma 2.3.1]. Lemma 10.12 is a minor variation of Lemma 2.4 from Necoara and Patrascu [88]. Theorem 10.26 is an extension of a result of Nesterov from [97] on the convergence of the gradient method for convex functions. The proximal point method was studied by Rockafellar in [110], as well as by many other authors; see, for example, the book of Bauschke and Combettes [8] and its extensive list of references. FISTA was developed by Beck and Teboulle in [18]; see also the book chapter [19]; the convergence analysis presented in Section 10.7 is taken from these sources. When the nonsmooth part is an indicator function of a closed and convex set, the method reduces to the optimal gradient method of Nesterov from 1983 [93]. Other accelerated proximal gradient methods can be found in the works of Nesterov [98] and Tseng [121]—the latter also describes a generalization to the non-Euclidean setting, which is an extension of the work of Auslender and Teboulle [2]. MFISTA and its convergence analysis are from the work of Beck and Teboulle [17]. The idea of using restarting in order to gain an improved rate of convergence in the strongly convex case can be found in Nesterov's work [98] in the context of a different accelerated proximal gradient method, but the idea works for any method that gains an $O(1/k^2)$ rate in the (not necessarily strongly) convex case. The proof of Theorem 10.42 follows the proof of Theorem 4.10 from the review paper of Chambolle and Pock [42]. The idea of solving nonsmooth problems through a smooth approximation was studied by many authors; see, for example, the works of Ben-Tal and Teboulle [25], Bertsekas [26], Moreau [87], and the more recent book of Auslender and Teboulle [1] and references therein. Lemma 10.70 can be found in Levitin and Polyak [79]. The idea of producing an $O(1/\varepsilon)$ complexity result for nonsmooth problems by employing an accelerated gradient method was first presented and developed by Nesterov in [95]. The extension to the three-part composite model and to the setting of more general smooth approximations was studied by Beck and Teboulle [20], where additional results and extensions can also be found. The non-Euclidean gradient method was proposed by Nutini, Schmidt, Laradji, Friendlander, and Koepke [100], where its rate of convergence in the strongly convex case was analyzed; the work [100] also contains a comparison between two coordinate selection strategies: Gauss–Southwell (which is the one considered in the chapter) and randomized selection. The non-Euclidean proximal gradient method was presented in the work of Tseng [121], where an accelerated non-Euclidean version was also analyzed.

Chapter 11. The version of the block proximal gradient method in which the nonsmooth functions g_i are indicators was studied by Luo and Tseng in [84], where some error bounds on the model were assumed. It was shown that under the model assumptions, the CBPG method with each block consisting of a single variable has a linear rate of convergence. Nesterov studied in [96] a randomized version of the method (again, in the setting where the nonsmooth functions are indicators) in which the selection of the block on which a gradient projection step is performed at each iteration is done randomly via a pre-described distribution. For the first time, Nesterov was able to establish global nonasymptotic rates of convergence in the convex case without any strict convexity, strong convexity, uniqueness, or error bound assumptions. Specifically, it was shown that the rate of convergence to the optimal value of the expectation sequence of the function values of the sequence generated by the randomized method is sublinear under the assumption of Lipschitz continuity of the gradient and linear under a strong convexity assumption. In addition, an accelerated $O(1/k^2)$ was devised in the unconstrained setting. Probabilistic results on the convergence of the function values were also provided. In [107] Richtarik and Takac generalized Nesterovs results to the composite model. The derivation of the randomized complexity result in Section 11.5 mostly follows the presentation in the work of Lin, Lu, and Xiao [82]. The type of analysis in the deterministic convex case (Section 11.4.2) originates from Beck and Tetruashvili [22], who studied the case in which the nonsmooth functions are indicators. The extension to the general composite model can be found in Shefi and Teboulle [115] as well as in Hong, Wang, Razaviyayn, and Luo [69]. Lemma 11.17 is Lemma 3.8 from [11]. Theorem 11.20 is a specialization of Lemma 2 from Nesterov [96]. Additional related methods and discussions can be found in the extensive survey of Wright [126].

Chapter 12. The idea of using a proximal gradient method on the dual of the main model (12.1) was originally developed by Tseng in [120], where the algorithm was named "alternating minimization." The primal representations of the DPG and FDPG methods, convergence analysis, as well as the primal-dual relation are from Beck and Teboulle [21]. The DPG method for solving the total variation problem was initially devised by Chambolle in [39], and the accelerated version was considered by Beck and Teboulle [17]. The one-dimensional total variation denoising problem is presented as an illustration for the DPG and FDPG methods; however, more direct and efficient methods exist for tackling the problem; see Hochbaum [68], Condat [45], Johnson [73], and Barbero and Sra [6]. The dual block proximal gradient method was discussed in Beck, Tetruashvili, Vaisbourd, and Shemtov [23], from which the specific decomposition of the isotropic two-dimensional total variation function is taken. The accelerated method ADBPG is a different representation of the accelerated method proposed by Chambolle and Pock in [41]. The latter work also discusses dual block proximal gradient methods and contains many other suggestions for decompositions of total variation functions.

Chapter 13. The conditional gradient algorithm was presented by Frank and Wolfe [56] in 1956 for minimizing a convex quadratic function over a compact polyhedral set. The original paper of Frank and Wolfe also contained a proof of an $O(1/k)$ rate of convergence in function values. Levitin and Polyak [79] showed that this $O(1/k)$ rate can also be extended to the case where the feasible set is a general compact con-

vex set and the objective function is L-smooth and convex. Dunn and Harshbarger [50] were probably the first to suggest a diminishing stepsize rule for the conditional gradient method and to establish a sublinear rate under such a strategy. The generalized conditional gradient method was introduced and analyzed by Bach in [4], where it was shown that under a certain setting, it can be viewed as a dual mirror descent method. Lemma 13.7 (fundamental inequality for generalized conditional gradient) can be found in the setting of the conditional gradient method in Levitin and Polyak [79]. The interpretation of the power method as the conditional gradient method was described in the work of Luss and Teboulle [85], where many other connections between the conditional gradient method and the sparse PCA problem are explored. Lemma 13.13 is an extension of Lemma 4.4 from Bach's work [4], and the proof is almost identical. Similar results on sequences of nonnegative numbers can be found in the book of Polyak [104, p. 45]. Section 13.3.1 originates from the work of Canon and Cullum [37]. Polyak in [104, p. 214, Exercise 10] seems to be the first to mention the linear rate of convergence of the conditional gradient method under a strong convexity assumption on the feasible set. Theorem 13.23 is from Journée, Nesterov, Richtárik, and Sepulchre [74, Theorem 12]. Lemma 13.26 and Theorem 13.27 are from Levitin and Polyak [79], and the exact form of the proof is due to Edouard Pauwels. Another situation, which was not discussed in the chapter, in which linear rate of converge can be established, is when the objective function is strongly convex and the optimal solution resides in the interior of the feasible set (Guélat and Marcotte [62]). Epelman and Freund [53], as well as Beck and Teboulle [16], showed a linear rate of convergence of the conditional gradient method with a special stepsize choice in the context of finding a point in the intersection of an affine space and a closed and convex set under a Slater-type assumption. The randomized generalized block conditional gradient method presented in Section 13.4 is a simple generalization of the randomized block conditional gradient method introduced and analyzed by Lacoste-Julien, Jaggi, Schmidt, and Pletscher in [76]. A deterministic version was analyzed by Beck, Pauwels, and Sabach in [12]. An excellent overview of the conditional gradient method, including many more theoretical results and applications, can be found in the thesis of Jaggi [72].

Chapter 14. The alternating minimization method is a rather old and fundamental algorithm. It appears in the literature under various names such as the block-nonlinear Gauss-Seidel method or the block coordinate descent method. Powell's example appears in [106]. Theorem 14.3 and its proof originate from Bertsekas [28, Proposition 2.7.1]. Theorem 14.9 and its proof are an extension of Proposition 6 from Grippo and Sciandrone [61] to the composite model. The proof of Theorem 14.11 follows the proof of Theorem 3.1 from the work of Hong, Wang, Razaviyayn, and Luo [69], where more general schemes than alternating minimization are also considered. Section 14.5.2 follows [11].

Chapter 15 The augmented Lagrangian method can be traced back to Hestenes [66] and Powell [105]. The method and its many variants was studied extensively in the literature, see, for example, the books of Bertsekas [27] and Bertsekas and Tsitsiklis [31] and references therein. Rockafellar [109] was first to establish the duality between the proximal point and the augmented Lagrangian methods; see also additional discussions in the work of Iusem [71]. ADMM is equivalent to an

operator splitting method called Douglas–Rachford splitting, which was introduced in the 1950s for the numerical solution of partial differential equations [48]. ADMM, as presented in the chapter, was first introduced by Gabay and Mercier [57] and Glowinski and Marrocco [59]. An extremely extensive survey on ADMM method can be found in the work of Boyd, Parikh, Chu, Peleato, and Eckstein [34]. AD-PMM was suggested by Eckstein [51]. The proof of Theorem 15.4 on the rate of convergence of AD-PMM is based on a combination of the proof techniques of He and Yuan [65] and Gao and Zhang [58]. Shefi and Teboulle provided in [114] a unified analysis for general classes of algorithm that include AD-PMM as a special instance. Shefi and Teboulle also showed the relation between AD-LPMM and the Chambolle–Pock algorithm [40].

Bibliography

[1] A. Auslender and M. Teboulle, *Asymptotic cones and functions in optimization and variational inequalities*, Springer Monographs in Mathematics, Springer-Verlag, New York, 2003. (Cited on p. 459)

[2] ———, *Interior gradient and proximal methods for convex and conic optimization*, SIAM J. Optim., 16 (2006), pp. 697–725, https://doi.org/10.1137/S1052623403427823. (Cited on p. 459)

[3] D. Azé and J. Penot, *Uniformly convex and uniformly smooth convex functions*, Ann. Fac. Sci. Toulouse Math. (6), 4 (1995), pp. 705–730. (Cited on p. 457)

[4] F. Bach, *Duality between subgradient and conditional gradient methods*, SIAM J. Optim., 25 (2015), pp. 115–129, https://doi.org/10.1137/130941961. (Cited on pp. 387, 461)

[5] J. B. Baillon and G. Haddad, *Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones*, Israel J. Math., 26 (1977), pp. 137–150. (Cited on p. 457)

[6] A. Barbero and S. Sra, *Modular proximal optimization for multidimensional total-variation regularization*. Available at https://arxiv.org/abs/1411.0589. (Cited on p. 460)

[7] H. H. Bauschke and P. L. Combettes, *The Baillon-Haddad theorem revisited*, J. Convex Anal., 17 (2010), pp. 781–787. (Cited on p. 457)

[8] ———, *Convex analysis and monotone operator theory in Hilbert spaces*, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, Springer, New York, 2011. With a foreword by Hédy Attouch. (Cited on pp. 457, 458, 459)

[9] A. Beck, *Convergence Rate Analysis of Gradient Based Algorithms*, Ph.D. thesis, School of Mathematical Sciences, Tel-Aviv University, 2003. (Cited on p. 459)

[10] ———, *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*, MOS-SIAM Series on Optimization, SIAM, Philadelphia, PA, 2014, https://doi.org/10.1137/1.9781611973655. (Cited on pp. 24, 28, 31, 45, 49, 63, 108, 112, 147, 195, 256, 386, 459)

[11] ——, *On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes*, SIAM J. Optim., 25 (2015), pp. 185–209, https://doi.org/10.1137/13094829X. (Cited on pp. 460, 461)

[12] A. Beck, E. Pauwels, and S. Sabach, *The cyclic block conditional gradient method for convex optimization problems*, SIAM J. Optim., 25 (2015), pp. 2024–2049, https://doi.org/10.1137/15M1008397. (Cited on p. 461)

[13] A. Beck and S. Sabach, *A first order method for finding minimal norm-like solutions of convex optimization problems*, Math. Program., 147 (2014), pp. 25–46. (Cited on p. 459)

[14] ——, *Weiszfeld's method: Old and new results*, J. Optim. Theory Appl., 164 (2015), pp. 1–40. (Cited on p. 457)

[15] A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Oper. Res. Lett., 31 (2003), pp. 167–175. (Cited on p. 458)

[16] ——, *A conditional gradient method with linear rate of convergence for solving convex linear systems*, Math. Methods Oper. Res., 59 (2004), pp. 235–247. (Cited on p. 461)

[17] ——, *Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems*, IEEE Trans. Image Process., 18 (2009), pp. 2419–2434. (Cited on pp. 296, 459, 460)

[18] ——, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202, https://doi.org/10.1137/080716542. (Cited on pp. 272, 290, 459)

[19] ——, *Gradient-based algorithms with applications to signal-recovery problems*, in Convex Optimization in Signal Processing and Communications, Cambridge University Press, Cambridge, 2010, pp. 42–88. (Cited on pp. 272, 459)

[20] ——, *Smoothing and first order methods: A unified framework*, SIAM J. Optim., 22 (2012), pp. 557–580, https://doi.org/10.1137/100818327. (Cited on pp. 49, 304, 457, 459)

[21] ——, *A fast dual proximal gradient algorithm for convex minimization and applications*, Oper. Res. Lett., 42 (2014), pp. 1–6. (Cited on pp. 355, 460)

[22] A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM J. Optim., 23 (2013), pp. 2037–2060, https://doi.org/10.1137/120887679. (Cited on pp. 342, 460)

[23] A. Beck, L. Tetruashvili, Y. Vaisbourd, and A. Shemtov, *Rate of convergence analysis of dual-based variables decomposition methods for strongly convex problems*, Oper. Res. Lett., 44 (2016), pp. 61–66. (Cited on pp. 377, 460)

[24] A. Ben-Tal, T. Margalit, and A. Nemirovski, *The ordered subsets mirror descent optimization method with applications to tomography*, SIAM J. Optim., 12 (2001), pp. 79–108, https://doi.org/10.1137/S1052623499354564. (Cited on pp. 112, 457)

[25] A. Ben-Tal and M. Teboulle, *A smoothing technique for nondifferentiable optimization problems*, in Optimization (Varetz, 1988), vol. 1405 of Lecture Notes in Math., Springer, Berlin, 1989, pp. 1–11. (Cited on p. 459)

[26] D. P. Bertsekas, *Nondifferentiable optimization via approximation: Nondifferentiable optimization*, Math. Programming Stud., (1975), pp. 1–25. (Cited on p. 459)

[27] ——, *Constrained optimization and Lagrange multiplier methods*, Computer Science and Applied Mathematics, Academic Press. [Harcourt Brace Jovanovich], New York, London, 1982. (Cited on p. 461)

[28] ——, *Nonlinear Programming*, Athena Scientific, Belmont, MA, second ed., 1999. (Cited on pp. 407, 457, 458, 459, 461)

[29] ——, *Convex Analysis and Optimization*, Athena Scientific, Belmont, MA, 2003. With Angelia Nedić and Asuman E. Ozdaglar. (Cited on pp. 41, 439, 457, 458)

[30] ——, *Convex Optimization Algorithms*, Athena Scientific, Belmont, MA, 2015. (Cited on p. 458)

[31] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Upper Saddle River, NJ, 1989. (Cited on p. 461)

[32] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, 3, Springer, New York, second ed., 2006. (Cited on pp. 183, 457, 458)

[33] J. M. Borwein and J. D. Vanderwerff, *Convex Functions: Constructions, Characterizations and Counterexamples*, vol. 109 of Encyclopedia of Mathematics and Its Applications, Cambridge University Press, Cambridge, 2010. (Cited on p. 458)

[34] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Mach. Learn., 3 (2011), pp. 1–122. (Cited on p. 462)

[35] S. Boyd and L. Vandenberghe, *Additional exercises for convex optimization*. Available at http://www.stanford.edu/~boyd/cvxbook/bv_cvxbook. (Cited on p. 458)

[36] R. E. Bruck, *On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space*, J. Math. Anal. Appl., 61 (1977), pp. 159–164. (Cited on p. 458)

[37] M. D. Canon and C. D. Cullum, *A tight upper bound on the rate of convergence of the Frank-Wolfe algorithm*, SIAM J. Control, 6 (1968), pp. 509–516, https://doi.org/10.1137/0306032. (Cited on pp. 391, 461)

[38] A. L. Cauchy, *Méthode generales pour la résolution des systèmes d'equations simultanées*, Comptes Rendues Acad. Sci. Paris, 25 (1847), pp. 536–538. (Cited on p. 459)

[39] A. Chambolle, *An algorithm for total variation minimization and applications*, J. Math. Imaging Vision, 20 (2004), pp. 89–97. Special issue on mathematics and image analysis. (Cited on p. 460)

[40] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vision, 40 (2011), pp. 120–145. (Cited on p. 462)

[41] ———, *A remark on accelerated block coordinate descent for computing the proximity operators of a sum of convex functions*, SMAI J. Comput. Math., 1 (2015), pp. 29–54. (Cited on pp. 373, 460)

[42] ———, *An introduction to continuous optimization for imaging*, Acta Numerica, 25 (2016), 161–319. (Cited on pp. 302, 459)

[43] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM J. Optim., 3 (1993), pp. 538–543, https://doi.org/10.1137/0803026. (Cited on pp. 252, 458)

[44] P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200, https://doi.org/10.1137/050626090. (Cited on p. 458)

[45] L. Condat, *A direct algorithm for* 1*-d total variation denoising*, IEEE Signal Process. Lett., 20 (2013), pp. 1054–1057. (Cited on p. 460)

[46] I. Daubechies, M. Defrise, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Comm. Pure Appl. Math., 57 (2004), pp. 1413–1457. (Cited on p. 459)

[47] C. Davis, *All convex invariant functions of hermitian matrices*, Arch. Math., 8 (1957), pp. 276–278. (Cited on p. 458)

[48] J. Douglas and H. H. Rachford, *On the numerical solution of heat conduction problems in two and three space variables*, Trans. Amer. Math. Soc., 82 (1956), pp. 421–439. (Cited on p. 462)

[49] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari, *Composite objective mirror descent*, in COLT 2010—The 23rd Conference on Learning Theory, 2010, pp. 14–26. (Cited on pp. 260, 458)

[50] J. C. Dunn and S. Harshbarger, *Conditional gradient algorithms with open loop step size rules*, J. Math. Anal. Appl., 62 (1978), pp. 432–444. (Cited on p. 461)

[51] J. Eckstein, *Some saddle-function splitting methods for convex programming*, Optim. Methods Softw., 4 (1994), pp. 75–83. (Cited on p. 462)

[52] M. Elad, *Why simple shrinkage is still relevant for redundant representations?*, IEEE Trans. Inform. Theory, 52 (2006), pp. 5559–5569. (Cited on p. 459)

[53] M. Epelman and R. M. Freund, *Condition number complexity of an elementary algorithm for computing a reliable solution of a conic linear system*, Math. Program., 88 (2000), pp. 451–485. (Cited on p. 461)

[54] T. Evgeniou, M. Pontil, D. Spinellis, and N. Nassuphis, *Regularized robust portfolio estimation*, in Regularization, Optimization, Kernels, and Support Vector Machines, CRC Press, Boca Raton, FL, 2015. (Cited on pp. 173, 458)

[55] F. Facchinei and J. S. Pang, *Finite-dimensional variational inequalities and complementarity problems. Vol.* II, Springer Series in Operations Research, Springer-Verlag, New York, 2003. (Cited on p. 459)

[56] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110. (Cited on p. 460)

[57] D. Gabay and B. Mercier, *A dual algorithm for the solution of nonlinear variational problems via finite element approximations*, Comp. Math. Appl., 2 (1976), pp. 17–40. (Cited on p. 462)

[58] X. Gao and S.-Z. Zhang, *First-order algorithms for convex optimization with nonseparable objective and coupled constraints*, J. Oper. Res. Soc. China, 5 (2017), pp. 131–159. (Cited on pp. 428, 462)

[59] R. Glowinski and A. Marroco, *Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires*, ESAIM: Mathematical Modelling and Numerical Analysis—Modélisation Mathématique et Analyse Numérique, 9 (1975), pp. 41–76. (Cited on p. 462)

[60] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, third ed., 1996. (Cited on p. 188)

[61] L. Grippo and M. Sciandrone, *On the convergence of the block nonlinear Gauss-Seidel method under convex constraints*, Oper. Res. Lett., 26 (2000), pp. 127–136. (Cited on pp. 413, 461)

[62] J. Guélat and P. Marcotte, *Some comments on Wolfe's "away step,"* Math. Program., 35 (1986), pp. 110–119. (Cited on p. 461)

[63] E. T. Hale, W. Yin, and Y. Zhang, *Fixed-point continuation for ℓ_1-minimization: Methodology and convergence*, SIAM J. Optim., 19 (2008), pp. 1107–1130, https://doi.org/10.1137/070698920. (Cited on p. 459)

[64] P. R. Halmos, *Finite-Dimensional Vector Spaces*, Undergraduate Texts in Mathematics, Springer-Verlag, New York, Heidelberg, second ed., 1974. (Cited on p. 457)

[65] B. He and X. Yuan, *On the $O(1/n)$ convergence rate of the Douglas–Rachford alternating direction method*, SIAM J. Numer. Anal., 50 (2012), pp. 700–709, https://doi.org/10.1137/110836936. (Cited on pp. 428, 462)

[66] M. R. Hestenes, *Multiplier and gradient methods*, J. Optimization Theory Appl., 4 (1969), pp. 303–320. (Cited on p. 461)

[67] J. B. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and minimization algorithms.* I, vol. 305 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer-Verlag, Berlin, 1996. Second Printing. (Cited on pp. 22, 67, 119, 457)

[68] D. S. Hochbaum, *An efficient algorithm for image segmentation, Markov random fields and related problems*, J. ACM, 48 (2001), pp. 686–701. (Cited on p. 460)

[69] M. Hong, X. Wang, M. Razaviyayn, and Z. Q. Luo, *Iteration complexity analysis of block coordinate descent methods.* Available at http://arxiv.org/abs/1310.6957. (Cited on pp. 342, 416, 460, 461)

[70] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, second ed., 2013. (Cited on p. 189)

[71] A. N. Iusem, *Augmented Lagrangian methods and proximal point methods for convex optimization*, Investigacion Operativa, 8 (1999), pp. 11–49. (Cited on p. 461)

[72] M. Jaggi, *Sparse Convex Optimization Methods for Machine Learning*, Ph.D. thesis, ETH Zurich, 2011. (Cited on p. 461)

[73] N. A. Johnson, *A dynamic programming algorithm for the fused lasso and L_0-segmentation*, J. Comput. Graph. Statist., 22 (2013), pp. 246–260. (Cited on p. 460)

[74] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, *Generalized power method for sparse principal component analysis*, J. Mach. Learn. Res., 11 (2010), pp. 517–553. (Cited on pp. 397, 461)

[75] K. Knopp, *Theory and Application of Infinite Series*, Blackie & Son Limited, 1951. (Cited on p. 392)

[76] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher, *Block-coordinate Frank-Wolfe optimization for structural SVMs*, in Proceedings of the 30th International Conference on Machine Learning (ICML-13), vol. 28, 2013, pp. 53–61. (Cited on pp. 400, 461)

[77] S. Lacoste-Julien, M. Schmidt, and F. Bach, *A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method*, ArXiv e-prints, 2012. (Cited on pp. 219, 458)

[78] G. Lan, *An optimal method for stochastic composite optimization*, Math. Program., 133 (2011), pp. 365–397. (Cited on pp. 70, 457)

[79] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, U.S.S.R. Comput. Math. Math. Phys., 6 (1966), pp. 787–823. (Cited on pp. 459, 460, 461)

[80] A. S. Lewis, *The convex analysis of unitarily invariant matrix functions*, J. Convex Anal., 2 (1995), pp. 173–183. (Cited on pp. 182, 458)

[81] ———, *Convex analysis on the Hermitian matrices*, SIAM J. Optim., 6 (1996), pp. 164–177 https://doi.org/10.1137/0806009. (Cited on pp. 182, 458)

[82] Q. Lin, Z. Lu, and L. Xiao, *An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization*, SIAM J. Optim., 25 (2015), pp. 2244–2273, https://doi.org/10.1137/141000270. (Cited on pp. 347, 460)

[83] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979, https://doi.org/10.1137/0716071. (Cited on p. 458)

[84] Z. Q. Luo and P. Tseng, *On the convergence of the coordinate descent method for convex differentiable minimization*, J. Optim. Theory Appl., 72 (1992), pp. 7–35. (Cited on p. 460)

[85] R. Luss and M. Teboulle, *Conditional gradient algorithms for rank-one matrix approximations with a sparsity constraint*, SIAM Rev., 55 (2013), pp. 65–98, https://doi.org/10.1137/110839072. (Cited on pp. 386, 461)

[86] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, PA, 2000. (Cited on p. 457)

[87] J. J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299. (Cited on pp. 458, 459)

[88] I. Necoara and A. Patrascu, *Iteration complexity analysis of dual first-order methods for conic convex programming*, Optim. Methods Softw., 31 (2016), pp. 645–678. (Cited on pp. 277, 459)

[89] A. Nedić and D. Bertsekas, *Convergence rate of incremental subgradient algorithms*, in Stochastic Optimization: Algorithms and Applications, Springer, Boston, MA, 2001, pp. 223–264. (Cited on pp. 230, 458)

[90] A. Nedić and A. Ozdaglar, *Approximate primal solutions and rate analysis for dual subgradient methods*, SIAM J. Optim., 19 (2009), pp. 1757–1780, https://doi.org/10.1137/070708111. (Cited on pp. 233, 238, 458)

[91] A. Nedić and A. Ozdaglar, *Distributed multi-agent optimization*, in Convex Optimization in Signal Processing and Communications, D. Palomar and Y. Eldar, eds., Cambridge University Press, Cambridge, 2009, pp. 340–386. (Cited on p. 458)

[92] A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, A Wiley-Interscience Publication, New York, 1983. (Cited on p. 458)

[93] Y. Nesterov, *A method for solving the convex programming problem with convergence rate $O(1/k^2)$*, Dokl. Akad. Nauk SSSR, 269 (1983), pp. 543–547. (Cited on p. 459)

[94] ———, *Introductory Lectures on Convex Optimization: A Basic Course*, vol. 87 of Applied Optimization, Kluwer Academic Publishers, Boston, MA, 2004. (Cited on pp. 457, 458, 459)

[95] ———, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152. (Cited on pp. 304, 459)

[96] ———, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM J. Optim., 22 (2012), pp. 341–362, https://doi.org/10.1137/100802001. (Cited on pp. 346, 460)

[97] ———, *How to make the gradients small*, Optima, 88 (2012), pp. 10–11. (Cited on p. 459)

[98] ———, *Gradient methods for minimizing composite functions*, Math. Program., 140 (2013), pp. 125–161. (Cited on p. 459)

[99] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, Springer, New York, second ed., 2006. (Cited on p. 459)

[100] J. Nutini, M. Schmidt, I. H. Laradji, M. Friendlander, and H. Koepke, *Coordinate descent converges faster with the Gauss-Southwell rule than random selection*, in Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015. (Cited on p. 459)

[101] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, vol. 30 of Classics in Applied Mathematics, SIAM, Philadelphia, PA, 2000. Reprint of the 1970 original, https://doi.org/10.1137/1.9780898719468. (Cited on p. 112)

[102] N. Parikh and S. Boyd, *Proximal algorithms*, Found. Trends Optim., 1 (2014), pp. 123–231. (Cited on pp. 182, 458)

[103] G. B. Passty, *Ergodic convergence to a zero of the sum of monotone operators in Hilbert space*, J. Math. Anal. Appl., 72 (1979), pp. 383–390. (Cited on p. 458)

[104] B. T. Polyak, *Introduction to Optimization*, Translations Series in Mathematics and Engineering, Optimization Software Inc., New York, 1987. (Cited on pp. 204, 458, 459, 461)

[105] M. J. D. Powell, *A method for nonlinear constraints in minimization problems*, in Optimization (Sympos., Univ. Keele, Keele, 1968), Academic Press, London, 1969, pp. 283–298. (Cited on p. 461)

[106] ——, *On search directions for minimization algorithms*, Math. Program., 4 (1973), pp. 193–201. (Cited on pp. 408, 461)

[107] P. Richtárik and M. Takáč, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, Math. Program., 144 (2014), pp. 1–38. (Cited on p. 460)

[108] R. T. Rockafellar, *Convex Analysis*, vol. 28 of Princeton Mathematical Series, Princeton University Press, Princeton, NJ, 1970. (Cited on pp. 30, 43, 44, 45, 56, 102, 119, 181, 457, 458)

[109] ——, *A dual approach to solving nonlinear programming problems by unconstrained optimization*, Math. Program., 5 (1973), pp. 354–373. (Cited on p. 461)

[110] ——, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898, https://doi.org/10.1137/0314056. (Cited on p. 459)

[111] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, vol. 317 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer-Verlag, Berlin, 1998. (Cited on p. 457)

[112] W. Rudin, *Principles of Mathematical Analysis*, International Series in Pure and Applied Mathematics, McGraw-Hill, New York, Auckland, Düsseldorf, third ed., 1976. (Cited on pp. 59, 113)

[113] A. Ruszczyński, *Nonlinear Optimization*, Princeton University Press, Princeton, NJ, 2006. (Cited on p. 458)

[114] R. Shefi and M. Teboulle, *Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization*, SIAM J. Optim., 24 (2014), pp. 269–297, https://doi.org/10.1137/130910774. (Cited on p. 462)

[115] ——, *On the rate of convergence of the proximal alternating linearized minimization algorithm for convex problems*, EURO J. Comput. Optim., 4 (2016), pp. 27–46. (Cited on pp. 342, 460)

[116] N. Z. Shor, *Minimization Methods for Nondifferentiable Functions*, vol. 3 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 1985. Translated from the Russian by K. C. Kiwiel and A. Ruszczyński. (Cited on p. 458)

[117] G. Strang, *Introduction to Linear Algebra*, Wellesley-Cambridge Press, fourth ed., 2009. (Cited on p. 457)

[118] R. Sturm, *Ueber den Punkt kleinster Entfernungssumme von gegebenen Punkten*, J. Reine Angew. Math., 97 (1884), pp. 49–61. (Cited on p. 457)

[119] C. M. Theobald, *An inequality for the trace of the product of two symmetric matrices*, Math. Proc. Cambridge Philos. Soc., 77 (1975), pp. 265–267. (Cited on pp. 183, 458)

[120] P. Tseng, *Applications of a splitting algorithm to decomposition in convex programming and variational inequalities*, SIAM J. Control Optim., 29 (1991), pp. 119–138, https://doi.org/10.1137/0329006. (Cited on p. 460)

[121] ———, *Approximation accuracy, gradient methods, and error bound for structured convex optimization*, Math. Program., 125 (2010), pp. 263–295. (Cited on pp. 326, 459)

[122] L. Vandenberghe. *Optimization Methods for Large-Scale Systems*, EE236C lecture notes, UCLA, 2016. (Cited on pp. 196, 458)

[123] J. Von Neumann, *Some matrix inequalities and metrization of matric space*, Tomsk. Univ. Rev., 1 (1937), pp. 286–300. (Cited on pp. 190, 458)

[124] E. V. Weiszfeld, *Sur le point pour lequel la somme des distances de n points donnés est minimum*, Tôhoku Math. J., 43 (1937), pp. 355–386. (Cited on p. 457)

[125] P. Wolfe, *Note on a method of conjugate subgradients for minimizing nondifferentiable functions*, Math. Program., 7 (1974), pp. 380–383. (Cited on p. 458)

[126] S. J. Wright, *Coordinate descent algorithms*, Math. Program., 151 (2015), pp. 3–34. (Cited on p. 460)

[127] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Trans. Signal Process., 57 (2009), pp. 2479–2493. (Cited on p. 459)

[128] C. Zalinescu, *On uniformly convex functions*, J. Math. Anal. Appl., 95 (1983), pp. 344 – 374. (Cited on p. 457)

[129] C. Zalinescu, *Convex Analysis in General Vector Spaces*, World Scientific, River Edge, NJ, 2002. (Cited on p. 457)

Index

The primary goal of this book is to provide a self-contained, comprehensive study of the main first-order methods that are frequently used in solving large-scale problems. First-order methods exploit information on values and gradients/subgradients (but not Hessians) of the functions composing the model under consideration. With the increase in the number of applications that can be modeled as large or even huge-scale optimization problems, there has been a revived interest in using simple methods that require low iteration cost as well as low memory storage.

The author has gathered, reorganized, and synthesized (in a unified manner) many results that are currently scattered throughout the literature, many of which cannot be typically found in optimization books.

First-Order Methods in Optimization

- offers comprehensive study of first-order methods with the theoretical foundations;
- provides plentiful examples and illustrations;
- emphasizes rates of convergence and complexity analysis of the main first-order methods used to solve large-scale problems; and
- covers both variables and functional decomposition methods.

This book is intended primarily for researchers and graduate students in mathematics, computer science, and electrical and other engineering departments. Readers with a background in advanced calculus and linear algebra, as well as prior knowledge in the fundamentals of optimization (some convex analysis, optimality conditions, and duality), will be best prepared for the material.

Amir Beck is a Professor at the School of Mathematical Sciences, Tel-Aviv University. His research interests are in continuous optimization, including theory, algorithmic analysis, and its applications. He has published numerous papers and has given invited lectures at international conferences. He serves on the editorial board of several journals. His research has been supported by various funding agencies, including the Israel Science Foundation, the German-Israeli Foundation, the United States–Israel Binational Science Foundation, the Israeli Science and Energy ministries, and the European Community.

For more information about MOS and SIAM books, journals, conferences, memberships, or activities, contact:

Society for Industrial
and Applied Mathematics
3600 Market Street, 6th Floor
Philadelphia, PA 19104-2688 USA
+1-215-382-9800 • Fax +1-215-386-7999
siam@siam.org • *www.siam.org*

Mathematical Optimization Society
3600 Market Street, 6th Floor
Philadelphia, PA 19104-2688 USA
+1-215-382-9800 x319
Fax +1-215-386-7999
service@mathopt.org • *www.mathopt.org*

MO25

ISBN 978-1-611974-98-0
90000
9781611974980