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,* $\operatorname{dom}(f)$ *is convex,* $\operatorname{dom}(g) \subseteq \operatorname{int}(\operatorname{dom}(f))$*, and f is L_f-smooth over* $\operatorname{int}(\operatorname{dom}(f))$.

(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 $\operatorname{dom}(f) = \mathbb{E}$, 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 $\text{int}(\text{dom}(f))$ and $C \subseteq \text{int}(\text{dom}(f))$.

- **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)

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

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:

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

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

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

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

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

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

(a) pick $L_k > 0$;

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

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

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

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 $\text{prox}_{t_k g_0} = \mathcal{I}, \text{prox}_{t_k\delta_C} = P_C$ and $\text{prox}_{t_k\lambda\|\cdot\|_1} = \mathcal{T}_{\lambda t_k}$, 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 $\mathbf{x} \in \operatorname{int}(\operatorname{dom}(f))$ 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 $G_L^{f,g} : \operatorname{int}(\operatorname{dom}(f)) \to \mathbb{E}$ is the operator defined by $G_L^{f,g}(\mathbf{x}) = L(\mathbf{x} - T_L(\mathbf{x}))$ for all $\mathbf{x} \in \operatorname{int}(\operatorname{dom}(f))$.

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

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

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

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

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 $\mathbf{x}^* \in \text{dom}(g)$, $G_L^{f,g}(\mathbf{x}^*) = \mathbf{0}$ 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 $\mathbf{x} \in \text{int}(\text{dom}(f))$.

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:

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

Plugging $L = L_1, \mathbf{v} = \mathbf{x} - \frac{1}{L_1}\nabla f(\mathbf{x})$, and $\mathbf{w} = \text{prox}_{\frac{1}{L_2}g}\big(\mathbf{x} - \frac{1}{L_2}\nabla f(\mathbf{x})\big) = T_{L_2}(\mathbf{x})$ 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 $\text{int}(\text{dom}(f))$ is that $G_L(\cdot)$ is Lipschitz continuous with constant $2L + L_f$. Indeed, for any $\mathbf{x}, \mathbf{y} \in \text{int}(\text{dom}(f))$,

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

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* $\mathbf{x}, \mathbf{y} \in \text{int}(\text{dom}(f))$*;*

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

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 $G_{L_f}\equiv G_{L_f}^{f,g}$ and $T_{L_f}\equiv T_{L_f}^{f,g}$.

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 **a** and **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

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

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

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

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:

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

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 $\mathbf{y} \in \text{int}(\text{dom}(f))$, 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* $\mathbf{x} \in \text{int}(\text{dom}(f))$ *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

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

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 $F(\mathbf{x}^k) - F_{\text{opt}} \leq \varepsilon$, 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

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

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 ($\boldsymbol{k \geq 0}$):

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

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*

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

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

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),

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

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 $F(\mathbf{x}^k) - F_{\rm opt} \leq \varepsilon$*, 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 $F(\mathbf{x}^k) - F_{\rm opt} \leq \varepsilon$ 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

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

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_{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

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

  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. ∎

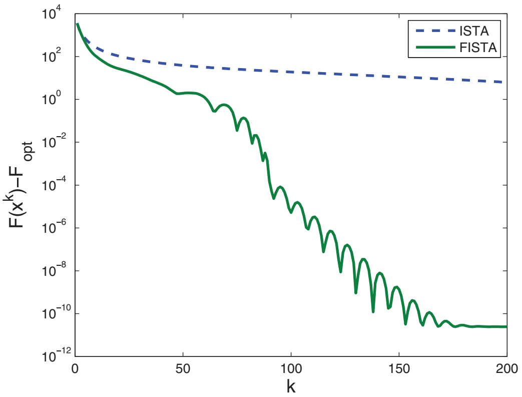

Figure 10.1. *Results of* 200 *iterations of ISTA and FISTA on an* l_1*-regularized least squares problem.*

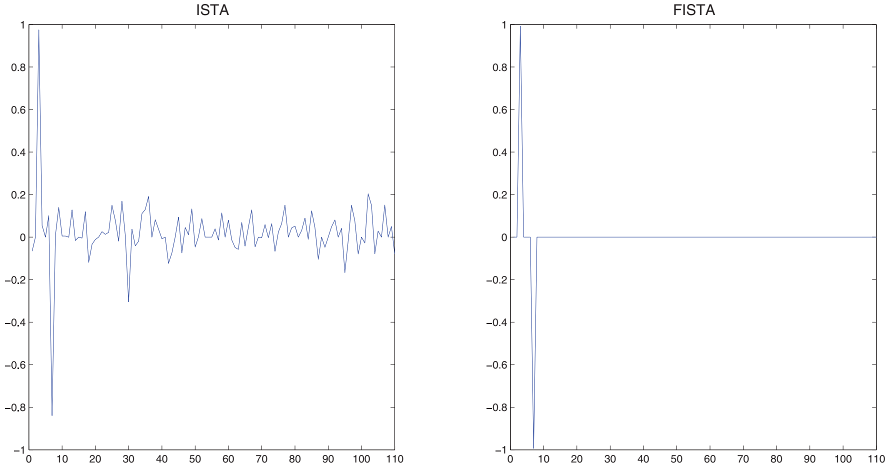

Figure 10.2. *Solutions obtained by ISTA (left) and FISTA (right).*

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 $v_n \equiv F(\mathbf{x}^n) - F_{\mathrm{opt}}$ 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 $k \geq 1$,

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

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

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

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,

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

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 **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 **Q**-norm, which we will denote by $L_f^{\mathbf{Q}}$. The constant is essentially defined by the relation

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

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 **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 **Q**-norm:

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

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 $k \geq 0$,*

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

(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,

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

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

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

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

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

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

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

[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 $t = \sqrt{\frac{L_f}{\sigma}}$)

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

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* **$(\boldsymbol{\alpha}, \boldsymbol{\beta})$-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 $\mathbf{x} \in \mathbb{R}^n$,

$$\begin{aligned} h_\mu(\mathbf{x}) &= \sqrt{\|\mathbf{x}\|_2^2 + \mu^2} - \mu \le \|\mathbf{x}\|_2 + \mu - \mu = \|\mathbf{x}\|_2 = h(\mathbf{x}), \\ h(\mathbf{x}) &= \|\mathbf{x}\|_2 \le \sqrt{\|\mathbf{x}\|_2^2 + \mu^2} = h_\mu(\mathbf{x}) + \mu, \end{aligned}$$

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 $\mathbf{x} \in \mathbb{R}^n$,

$$\begin{aligned} h_\mu(\mathbf{x}) &= \mu \log\left(\sum_{i=1}^n e^{x_i/\mu}\right) - \mu \log n \\ &\le \mu \log\left(n e^{\max_i\{x_i\}/\mu}\right) - \mu \log n = h(\mathbf{x}), && (10.60) \\ h(\mathbf{x}) &= \max_i\{x_i\} \le \mu \log\left(\sum_{i=1}^n e^{x_i/\mu}\right) = h_\mu(\mathbf{x}) + \mu \log n. && (10.61) \end{aligned}$$

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 $q(\mathbf{x}) = \max_{i=1,\ldots,m}\{\mathbf{a}_i^T\mathbf{x} + b_i\}$, where $\mathbf{a}_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$ for any $i = 1, 2, \ldots, m$. Then $q(\mathbf{x}) = g(\mathbf{Ax}+\mathbf{b})$, where $g(\mathbf{y}) = \max\{y_1, y_2, \ldots, y_m\}$, $\mathbf{A}$ is the matrix whose rows are $\mathbf{a}_1^T, \mathbf{a}_2^T, \ldots, \mathbf{a}_m^T$, and $\mathbf{b} = (b_1, b_2, \ldots, b_m)^T$. 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*

$$|h(\mathbf{x}) - h(\mathbf{y})| \le \ell_h \|\mathbf{x} - \mathbf{y}\| \textit{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{E}.$$

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 $\mu = 0.2$). ■

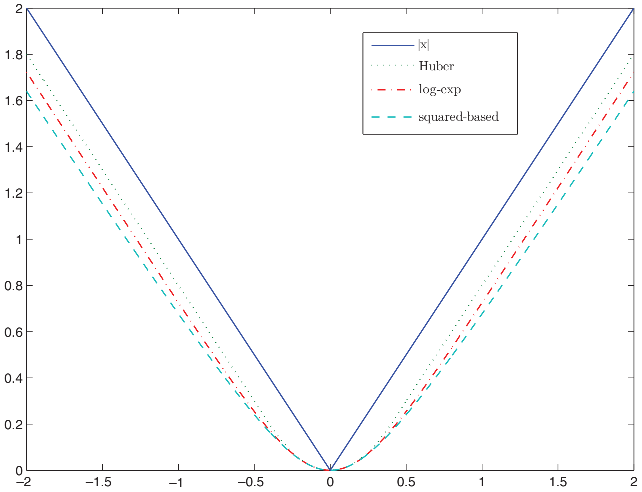

Figure 10.3. *The absolute value function along with its three* 5*-smooth approximations* ($\mu = 0.2$)*. "squared-based" is the function* $h^1_\mu(x) = \sqrt{x^2+\mu^2} - \mu$*, "log-exp" is* $h^2_\mu(x) = \mu\log(e^{x/\mu} + e^{-x/\mu}) - \mu\log 2$*, and "Huber" is* $h^3_\mu(x) = H_\mu(x)$*.*

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 $h_\mu$—a $\frac{1}{\mu}$-smooth approximation of $h$ with parameters $(\alpha, \beta)$; 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 $C \equiv \{\mathbf{x} \in \mathbb{E} : H_\mu(\mathbf{x}) \le H_\mu(\mathbf{x}^0)\}$ 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 $h(\mathbf{x}) \le h_\mu(\mathbf{x}) + \beta\mu$ for all $\mathbf{x} \in \mathbb{E}$, and consequently $H(\mathbf{x}) \le H_\mu(\mathbf{x}) + \beta\mu$ for all $\mathbf{x} \in \mathbb{E}$. Thus,

$$C \subseteq \{\mathbf{x} \in \mathbb{E} : H(\mathbf{x}) \le H_\mu(\mathbf{x}^0) + \beta\mu\},$$

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,

$$H_\mu(\mathbf{x}^k) - H_{\mu,\text{opt}} \le 2\left(L_f + \frac{\alpha}{\mu}\right)\frac{\|\mathbf{x}^0 - \mathbf{x}_\mu^*\|^2}{(k+1)^2} = 2\left(L_f + \frac{\alpha}{\mu}\right)\frac{\Lambda}{(k+1)^2}, \tag{10.67}$$

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 $\mathbf{x} \in \mathbb{E}$,

$$H_\mu(\mathbf{x}) \le H(\mathbf{x}) \le H_\mu(\mathbf{x}) + \beta\mu. \tag{10.68}$$

In particular, the following two inequalities hold:

$$H_{\text{opt}} \ge H_{\mu,\text{opt}} \quad \text{and} \quad H(\mathbf{x}^k) \le H_\mu(\mathbf{x}^k) + \beta\mu,\ k = 0, 1, \ldots, \tag{10.69}$$

which, combined with (10.67), yields

$$\begin{aligned} H(\mathbf{x}^k) - H_{\text{opt}} \le H_\mu(\mathbf{x}^k) + \beta\mu - H_{\mu,\text{opt}} &\le 2L_f\frac{\Lambda}{(k+1)^2} + \left(\frac{2\alpha\Lambda}{(k+1)^2}\right)\frac{1}{\mu} + \beta\mu \\ &\le 2L_f\frac{\Lambda}{k^2} + \left(\frac{2\alpha\Lambda}{k^2}\right)\frac{1}{\mu} + \beta\mu. \end{aligned}$$

Therefore, for a given $K > 0$, it holds that for any $k \ge K$,

$$H(\mathbf{x}^k) - H_{\text{opt}} \le 2L_f\frac{\Lambda}{K^2} + \left(\frac{2\alpha\Lambda}{K^2}\right)\frac{1}{\mu} + \beta\mu. \tag{10.70}$$

Minimizing the right-hand side w.r.t. μ, we obtain

$$\mu = \sqrt{\frac{2\alpha\Lambda}{\beta}}\frac{1}{K}. \tag{10.71}$$

Plugging the above expression into (10.70), we conclude that for any $k \ge K$,

$$H(\mathbf{x}^k) - H_{\text{opt}} \le 2L_f\frac{\Lambda}{K^2} + 2\sqrt{2\alpha\beta\Lambda}\frac{1}{K}.$$

Thus, to guarantee that $\mathbf{x}^k$ is an ε-optimal solution for any $k \ge K$, it is enough that K will satisfy

$$2L_f\frac{\Lambda}{K^2} + 2\sqrt{2\alpha\beta\Lambda}\frac{1}{K} \le \varepsilon.$$

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, $\nabla M_h^\mu(\mathbf{x}) = \frac{1}{\mu}(\mathbf{x} - \operatorname{prox}_{\mu h}(\mathbf{x}))$. 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:

$$\begin{aligned}\mathbf{x}^{k+1} &= \operatorname{prox}_{\frac{1}{\tilde{L}}g}\left(\mathbf{y}^k - \frac{1}{\tilde{L}}\nabla F_\mu(\mathbf{y}^k)\right) = P_C\left(\mathbf{y}^k - \frac{1}{\tilde{L}\mu}(\mathbf{y}^k - \operatorname{prox}_{\mu h}(\mathbf{y}^k))\right)\\ &= P_C(\operatorname{prox}_{\mu h}(\mathbf{y}^k)).\end{aligned}$$

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) $\mathbf{x}^{k+1} = P_C(\operatorname{prox}_{\mu h}(\mathbf{y}^k))$;

(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

$$\begin{aligned}\nabla F_\mu(\mathbf{x}) &= \nabla f(\mathbf{x}) + \mathbf{D}^T \nabla M_q^\mu(\mathbf{Dx}) \\ &= \nabla f(\mathbf{x}) + \tfrac{1}{\mu}\mathbf{D}^T(\mathbf{Dx} - \operatorname{prox}_{\mu q}(\mathbf{Dx})) \quad \text{[Theorem 6.60]} \\ &= \nabla f(\mathbf{x}) + \tfrac{1}{\mu}\mathbf{D}^T(\mathbf{Dx} - \mathcal{T}_\mu(\mathbf{Dx})). \quad \text{[Example 6.8]}\end{aligned}$$

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) $\mathbf{x}^{k+1} = \mathcal{T}_{\lambda/\tilde{L}}\left(\mathbf{y}^k - \frac{1}{\tilde{L}}(\mathbf{A}^T(\mathbf{Ay}^k - \mathbf{b}) + \frac{1}{\mu}\mathbf{D}^T(\mathbf{Dy}^k - \mathcal{T}_\mu(\mathbf{Dy}^k)))\right)$;

(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 $H(\mathbf{x}) \leq \alpha$, then

$$\lambda\|\mathbf{x}\|_1 \leq \frac{1}{2}\|\mathbf{Ax} - \mathbf{b}\|_2^2 + \|\mathbf{Dx}\|_1 + \lambda\|\mathbf{x}\|_1 \leq \alpha,$$

and since $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1$, 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,

$$\Lambda_{\mathbf{a}} = \left\{\mathbf{z} \in \mathbb{R}^n : z_i = \operatorname{sgn}(a_i), i \in I_{\neq}(\mathbf{a}), |z_j| \leq 1, j \in I_0(\mathbf{a})\right\},$$

where

$$I_{\neq}(\mathbf{a}) = \{i \in \{1, 2, \ldots, n\} : a_i \neq 0\}, I_0(\mathbf{a}) = \{i \in \{1, 2, \ldots, n\} : a_i = 0\}. \qquad \blacksquare$$

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

$$M = \begin{cases} \frac{\bar{L} - \frac{L_f}{2}}{(\bar{L})^2}, & \textit{constant stepsize}, \\ \frac{\gamma}{\max\left\{s, \frac{\eta L_f}{2(1-\gamma)}\right\}}, & \textit{backtracking}, \\ \frac{1}{2L_f}, & \textit{exact line search}. \end{cases} \tag{10.80}$$

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, $f(\mathbf{x}^{k+1}) \le f(\tilde{\mathbf{x}}^k)$, where $\tilde{\mathbf{x}}^k = \mathbf{x}^k - \frac{\|\nabla f(\mathbf{x}^k)\|_*}{L_f}\nabla f(\mathbf{x}^k)^\dagger$, and hence

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \ge f(\mathbf{x}^k) - f(\tilde{\mathbf{x}}^k) \ge \frac{1}{2L_f}\|\nabla f(\mathbf{x}^k)\|_*^2,$$

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

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \ge \frac{\gamma}{L_k}\|\nabla f(\mathbf{x}^k)\|_*^2 \ge \frac{\gamma}{\max\left\{s, \frac{\eta L_f}{2(1-\gamma)}\right\}}\|\nabla f(\mathbf{x}^k)\|_*^2. \qquad \square$$

Theorem 10.67 (convergence of the non-Euclidean gradient method—nonconvex case). *Suppose that f is an L_f-smooth function. Let $\{\mathbf{x}^k\}_{k\ge 0}$ 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 $\{f(\mathbf{x}^k)\}_{k\ge 0}$ is nonincreasing; in addition, $f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k)$ if and only if $\nabla f(\mathbf{x}^k) \ne \mathbf{0}$;*

(b) *if the sequence $\{f(\mathbf{x}^k)\}_{k\ge 0}$ 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_{opt}, then*

$$\min_{n=0,1,\ldots,k} \|\nabla f(\mathbf{x}^k)\|_* \le \frac{\sqrt{f(\mathbf{x}^0) - f_{\text{opt}}}}{\sqrt{M(k+1)}}, \tag{10.82}$$

where M is given in (10.80)*;*

(d) *all limit points of the sequence $\{\mathbf{x}^k\}_{k\ge 0}$ are stationary points of problem* (10.81).

Proof. (a) By Lemma 10.66,

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \ge M\|\nabla f(\mathbf{x}^k)\|_*^2, \tag{10.83}$$

where $M > 0$ is given in (10.80). The inequality (10.83) readily implies that $f(\mathbf{x}^k) \ge f(\mathbf{x}^{k+1})$ and that if $\nabla f(\mathbf{x}^k) \ne \mathbf{0}$, 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 $\{f(\mathbf{x}^k)\}_{k\ge 0}$ 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

$$L_f^{(p)} = \|\mathbf{A}\|_{p,q} = \max_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}\|_q : \|\mathbf{x}\|_p \le 1\}$$

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 ($k \ge 0$):** $\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L_f^{(2)}}(\mathbf{A}\mathbf{x}^k + \mathbf{b})$.

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 ($\boldsymbol{k \geq 0}$):**
  - 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]

$$f(\mathbf{x}^k) - f_{\mathrm{opt}} \leq \frac{2L_f^{(p)} R_{f(\mathbf{x}^0)}^2}{k}.$$

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 $\mathbf{Ax}^k$ 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 $\mathbf{g}^k \equiv \mathbf{Ax}^k + \mathbf{b}$ using the relation $\mathbf{g}^{k+1} = \mathbf{g}^k - \frac{\mathbf{A}_{i_k}\mathbf{x}^k + b_{i_k}}{L_f^{(1)}}\mathbf{Ae}_{i_k}$ ($\mathbf{Ae}_{i_k}$ 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 $f(\mathbf{x}^k) - f_{\mathrm{opt}}$ 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.

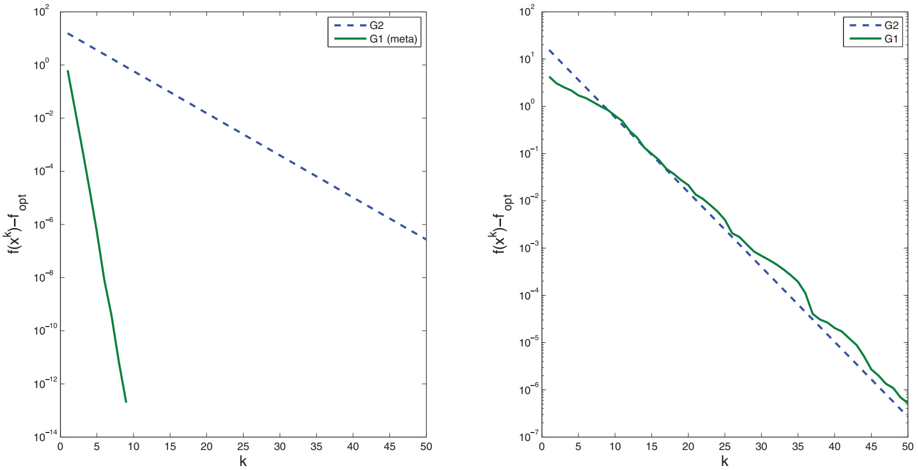

Figure 10.4. *Comparison of the Euclidean gradient method* (G2) *with the non-Euclidean gradient method* (G1) *applied on the problem from Example* 10.76 *with* $n = 10$. *The left image considers "meta-iterations" of* G1, *meaning that* 10 *iterations of* G1 *are counted as one iteration, while the right image counts each coordinate update as one iteration.*

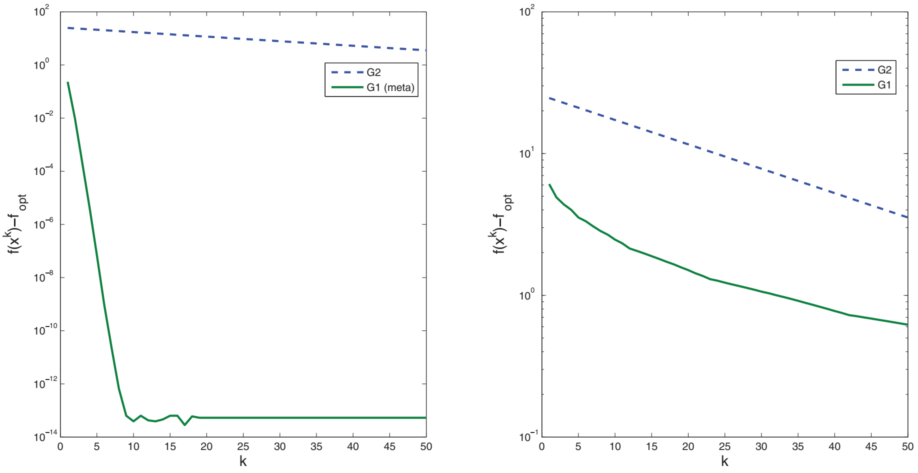

Figure 10.5. *Comparison of the Euclidean gradient method* (G2) *with the non-Euclidean gradient method* (G1) *applied on the problem from Example* 10.76 *with* $n = 100$. *The left image considers "meta-iterations" of* G1, *meaning that* 100 *iterations of* G1 *are counted as one iteration, while the 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 $n = 100$. $\blacksquare$

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

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

  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

$$B_\omega(\mathbf{x}, \mathbf{x}^{n+1}) + B_\omega(\mathbf{x}^{n+1}, \mathbf{x}^n) - B_\omega(\mathbf{x}, \mathbf{x}^n) \le \frac{m(\mathbf{x}, \mathbf{x}^n) - m(\mathbf{x}^{n+1}, \mathbf{x}^n) + g(\mathbf{x}) - g(\mathbf{x}^{n+1})}{L_n}.$$

Rearranging terms, we obtain that

$$\begin{aligned} m(\mathbf{x}^{n+1}, \mathbf{x}^n) + g(\mathbf{x}^{n+1}) + L_n B_\omega(\mathbf{x}^{n+1}, \mathbf{x}^n) \le m(\mathbf{x}, \mathbf{x}^n) &+ g(\mathbf{x}) + L_n B_\omega(\mathbf{x}, \mathbf{x}^n) \\ &- L_n B_\omega(\mathbf{x}, \mathbf{x}^{n+1}), \end{aligned}$$

which, combined with (10.94), yields the inequality

$$F(\mathbf{x}^{n+1}) \le m(\mathbf{x}, \mathbf{x}^n) + g(\mathbf{x}) + L_n B_\omega(\mathbf{x}, \mathbf{x}^n) - L_n B_\omega(\mathbf{x}, \mathbf{x}^{n+1}).$$

Since f is convex, $m(\mathbf{x}, \mathbf{x}^n) \le f(\mathbf{x})$, and hence

$$F(\mathbf{x}^{n+1}) - F(\mathbf{x}) \le L_n B_\omega(\mathbf{x}, \mathbf{x}^n) - L_n B_\omega(\mathbf{x}, \mathbf{x}^{n+1}).$$

Plugging in $\mathbf{x} = \mathbf{x}^*$ and dividing by L_n, we obtain

$$\frac{F(\mathbf{x}^{n+1}) - F(\mathbf{x}^*)}{L_n} \le B_\omega(\mathbf{x}^*, \mathbf{x}^n) - B_\omega(\mathbf{x}^*, \mathbf{x}^{n+1}).$$

Using the bound $L_n \le \alpha L_f$ (see Remark 10.80),

$$\frac{F(\mathbf{x}^{n+1}) - F(\mathbf{x}^*)}{\alpha L_f} \le B_\omega(\mathbf{x}^*, \mathbf{x}^n) - B_\omega(\mathbf{x}^*, \mathbf{x}^{n+1}),$$

and hence

$$F(\mathbf{x}^{n+1}) - F_{\text{opt}} \le \alpha L_f B_\omega(\mathbf{x}^*, \mathbf{x}^n) - \alpha L_f B_\omega(\mathbf{x}^*, \mathbf{x}^{n+1}).$$

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

$$\sum_{n=0}^{k-1} (F(\mathbf{x}^{n+1}) - F_{\text{opt}}) \le \alpha L_f B_\omega(\mathbf{x}^*, \mathbf{x}^0) - \alpha L_f B_\omega(\mathbf{x}^*, \mathbf{x}^k) \le \alpha L_f B_\omega(\mathbf{x}^*, \mathbf{x}^0).$$

Using the monotonicity of the sequence of function values, we conclude that

$$k(F(\mathbf{x}^k) - F_{\text{opt}}) \le \alpha L_f B_\omega(\mathbf{x}^*, \mathbf{x}^0),$$

thus obtaining the result

$$F(\mathbf{x}^k) - F_{\text{opt}} \le \frac{\alpha L_f B_\omega(\mathbf{x}^*, \mathbf{x}^0)}{k}. \qquad \square$$