11 Proximal Operator

Proximal projection is a useful procedure for the minimization of nonsmooth convex functions. Its main power lies in transforming the minimization of convex functions into the equivalent problem of determining fixed points for contractive operators. The purpose of this chapter is to introduce proximal operators, highlight some of their properties, and explain the role that soft-thresholding or shrinkage plays in this context.

11.1 DEFINITION AND PROPERTIES

Let $h(w) : \mathbb{R}^M \rightarrow \mathbb{R}$ denote a convex function of real arguments, $w \in \mathbb{R}^M$. Since $h(w)$ is convex, it can only have minimizers and all of them will be global minimizers. We are interested in locating a minimizer for $h(w)$. The function $h(w)$ may be non-differentiable at some locations.

11.1.1 Definition

Let $z \in \mathbb{R}^M$ denote some given vector. We add a quadratic term to $h(w)$ that measures the squared distance from w to z and define:

$$h_p(w) \overset{\Delta}{=} h(w) \;+\; \frac{1}{2\mu}\|w-z\|^2, \;\; \forall\, w \in \text{dom}(h) \tag{11.1}$$

for some positive scalar $\mu > 0$ chosen by the designer. The proximal operator of $h(w)$, also called the *proximity operator*, is denoted by the notation $\text{prox}_{\mu h}(z)$ and defined as the mapping that transforms z into the vector $\widehat{w}$ computed as follows:

$$\widehat{w} \overset{\Delta}{=} \underset{w\in\mathbb{R}^M}{\text{argmin}} \;\; h_p(w) \tag{11.2}$$

We will refer to $\widehat{w}$ as the proximal projection of z relative to h. The function $h_p(w)$ is strongly-convex in w since $h(w)$ is convex and $\|w-z\|^2$ is strongly-convex. It follows that $h_p(w)$ has a unique global minimizer and, therefore, the proximal projection, $\widehat{w}$, exists and is unique. Obviously, the vector $\widehat{w}$ is a function of z and we express the transformation from z to $\widehat{w}$ by writing

$$\widehat{w} = \text{prox}_{\mu h}(z) \tag{11.3}$$

where the role of $h(w)$ is highlighted in the subscript notation, and where the proximal operator is defined by

$$\boxed{\text{prox}_{\mu h}(z) \;\stackrel{\Delta}{=}\; \underset{w\in\mathbb{R}^M}{\text{argmin}} \left\{ h(w) \;+\; \frac{1}{2\mu}\|w-z\|^2 \right\}} \tag{11.4}$$

In the trivial case when $h(w)=0$, we get $\widehat{w}=z$ so that

$$\text{prox}_0(z) = z \tag{11.5}$$

Using (11.3), the minimum value of $h_p(w)$, when w is replaced by $\widehat{w}$, is called the *Moreau envelope*, written as

$$\mathcal{M}_{\mu h}(z) \;\stackrel{\Delta}{=}\; h(\widehat{w}) \;+\; \frac{1}{2\mu}\|\widehat{w}-z\|^2 \tag{11.6}$$

Intuitively, the proximal construction (11.3) approximates z by the vector $\widehat{w}$ that is "close" to it under the squared Euclidean norm and subject to the "penalty" $h(w)$ on the "size" of w. The Moreau value at $\widehat{w}$ serves as a measure of a "generalized" distance between z and its proximal projection, $\widehat{w}$. The reason for the qualification "projection" is because the proximal operation (11.3) can be interpreted as a generalization of the notion of projection, as the following example illustrates for a particular choice of $h(w)$.

Example 11.1 **(Projection onto a convex set)** Let $\mathcal{C}$ denote some convex set and introduce the indicator function:

$$\mathbb{I}_{C,\infty}[w] \;\stackrel{\Delta}{=}\; \begin{cases} 0, & w\in\mathcal{C} \\ \infty, & \text{otherwise} \end{cases} \tag{11.7}$$

That is, the function assumes the value zero whenever w belongs to the set $\mathcal{C}$ and is infinite otherwise. It is straightforward to verify that when $h(w)=\mathbb{I}_{C,\infty}[w]$, the definition

$$\text{prox}_{\mu\,\mathbb{I}_{C,\infty}}(z) \;=\; \underset{w\in\mathbb{R}^M}{\text{argmin}} \left\{ \mathbb{I}_{C,\infty}(w) \;+\; \frac{1}{2\mu}\|w-z\|^2 \right\} \tag{11.8}$$

reduces to

$$\text{prox}_{\mu\,\mathbb{I}_{C,\infty}}(z) \;=\; \underset{w\in\mathcal{C}}{\text{argmin}}\; \|w-z\|^2 \tag{11.9}$$

In other words, the proximal operator (11.8) corresponds to projecting onto the set $\mathcal{C}$ and determining the closest element in $\mathcal{C}$ to the vector z. We express the result in the form:

$$\text{prox}_{\mu\,\mathbb{I}_{C,\infty}}(z) \;=\; \mathcal{P}_C(z) \tag{11.10}$$

If we compare (11.8) with the general definition (11.4), it becomes clear why the proximal operation is viewed as a generalization of the concept of projection onto convex sets. The generalization results from replacing the indicator function, $\mathbb{I}_{C,\infty}[w]$, by an arbitrary convex function, $h(w)$.

Optimality condition

When $h(w)$ is differentiable, the minimizer $\widehat{w}$ of (11.4) should satisfy

$$\nabla_{w^{\mathsf{T}}}\, h(\widehat{w}) + \frac{1}{\mu}(\widehat{w} - z) = 0 \iff \widehat{w} = z - \mu\, \nabla_{w^{\mathsf{T}}}\, h(\widehat{w}) \tag{11.11}$$

On the other hand, when $h(w)$ is not differentiable, the minimizer $\widehat{w}$ should satisfy

$$0 \in \, \partial_{w^{\mathsf{T}}}\, h(\widehat{w}) \; + \; \frac{1}{\mu}(\widehat{w} - z) \iff (z - \widehat{w}) \in \mu\, \partial_{w^{\mathsf{T}}}\, h(\widehat{w}) \tag{11.12}$$

This is a critical property and we rewrite it more generically as follows for ease of reference in terms of two vectors a and b:

$$\boxed{a = \text{prox}_{\mu h}(b) \iff (b - a) \in \mu\, \partial_{w^{\mathsf{T}}}\, h(a)} \tag{11.13}$$

11.1.2 Soft Thresholding

One useful choice for the proximal function is

$$h(w) = \alpha \|w\|_1 \tag{11.14}$$

in terms of the ℓ_1-norm of the vector w and where $\alpha > 0$. In this case, the function $h(w)$ is non-differentiable at $w = 0$ and the function $h_p(w)$ becomes

$$h_p(w) = \alpha \|w\|_1 \; + \; \frac{1}{2\mu}\|w - z\|^2 \tag{11.15}$$

It turns out that a closed form expression exists for the corresponding proximal operator:

$$\widehat{w} \; \stackrel{\Delta}{=} \; \text{prox}_{\mu\alpha\|w\|_1}(z) \tag{11.16}$$

We show below that $\widehat{w}$ is obtained by applying a soft-thresholding (or shrinkage) operation to z as follows. Let z_m denote the $m-$th entry of $z \in \mathbb{R}^M$. Then, the corresponding entry of $\widehat{w}$, denoted by $\widehat{w}_m$, is found by shrinking z_m in the following manner:

$$\widehat{w}_m \; = \; \mathbb{T}_{\mu\alpha}\left(z_m\right), \quad m = 1, 2, \ldots, M \tag{11.17}$$

where the soft-thresholding function, denoted by $\mathbb{T}_{\beta}(x) : \mathbb{R} \rightarrow \mathbb{R}$, with threshold $\beta \geq 0$, is defined as

$$\mathbb{T}_{\beta}(x) \; \stackrel{\Delta}{=} \; \begin{cases} x - \beta, & \text{if } x \geq \beta \\ 0, & \text{if } -\beta < x < \beta \\ x + \beta, & \text{if } x \leq -\beta \end{cases} \tag{11.18}$$

This can also be written in the alternative form:

$$\begin{aligned} \mathbb{T}_{\beta}(x) &= \text{sign}(x) \times \max\Big\{0, |x| - \beta\Big\} \\ &= \text{sign}(x) \Big(|x| - \beta\Big)_+ \end{aligned} \tag{11.19}$$

where $(a)_+ \stackrel{\Delta}{=} \max\{0, a\}$. Figure 11.1 plots the function $\mathbb{T}_\beta(x)$ defined by (11.18); observe how values of x outside the interval $(-\beta, \beta)$ have their magnitudes reduced by the amount β, while values of x within this interval are set to zero.

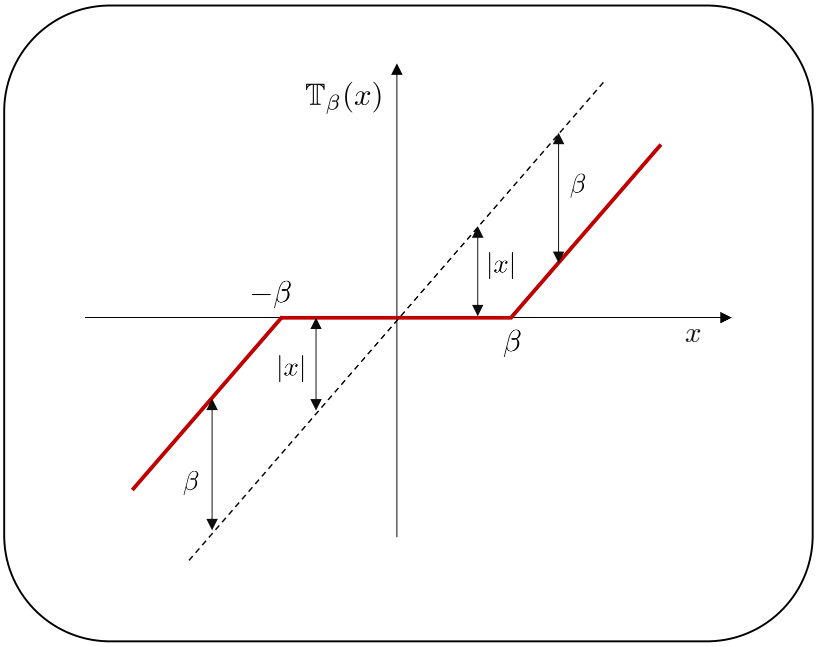

Figure 11.1 The soft-thresholding function, $\mathbb{T}_\beta(x)$, reduces the value of x gradually. Small values of x within the interval $(-\beta, \beta)$ are set to zero, while values of x outside this interval have their size reduced by an amount equal to β. The dotted curve corresponds to the line $y = x$, where y is the vertical coordinate.

We will replace notation (11.17) by the more compact representation:

$$\widehat{w} \;=\; \mathbb{T}_{\mu\alpha}(z) \tag{11.20}$$

in terms of the vector arguments $\{\widehat{w}, z\}$, with the understanding that $\mathbb{T}_{\mu\alpha}(\cdot)$ is applied to the individual entries of z to construct the corresponding entries of $\widehat{w}$ according to (11.18).

Proof of (11.17): To determine $\widehat{w}$ we need to minimize the function $h_p(w)$ defined by (11.15). We rewrite (11.15) in terms of the individual entries $\{w_m, z_m\}$ as follows:

$$h_p(w) = \sum_{m=1}^{M} \alpha|w_m| + \sum_{m=1}^{M} \frac{1}{2\mu}(w_m - z_m)^2 \tag{11.21}$$

It follows from this expression that the minimization of $h_p(w)$ over w decouples into M separate minimization problems:

$$\widehat{w}_m \stackrel{\Delta}{=} \underset{w_m \in \mathbb{R}}{\text{argmin}} \left\{\alpha|w_m| + \frac{1}{2\mu}(w_m - z_m)^2\right\} \tag{11.22}$$

For ease of reference, we denote the function that appears on the right-hand side by

$$h_m(w_m) \;\stackrel{\Delta}{=}\; \alpha|w_m| + \frac{1}{2\mu}(w_m - z_m)^2 \tag{11.23}$$

This cost is convex in w_m but is not differentiable at $w_m = 0$. We can arrive at a closed-form expression for the minimizer by examining the behavior of the function separately over the ranges $w_m \geq 0$ and $w_m \leq 0$:

(1) $\underline{w_m \geq 0}$: In this case, we can use the expression for $h_m(w_m)$ to write

$$\begin{aligned} 2\mu h_m(w_m) &= 2\mu\alpha w_m + (w_m - z_m)^2 \\ &= (w_m - (z_m - \mu\alpha))^2 \; + z_m^2 \; - \; (z_m - \mu\alpha)^2 \end{aligned} \tag{11.24}$$

It follows that the minimizer of $h_m(w_m)$ over the range $w_m \geq 0$ is given by

$$w_m^+ \;\stackrel{\Delta}{=}\; \underset{w_m \geq 0}{\text{argmin}}\; h_m(w_m) = \begin{cases} 0, & \text{if } \; z_m < \mu\alpha \\ z_m - \mu\alpha, & \text{if } \; z_m \geq \mu\alpha \end{cases} \tag{11.25}$$

(2) $\underline{w_m \leq 0}$: In this case, we get

$$\begin{aligned} 2\mu h_m(w_m) &= -2\mu\alpha w_m + (w_m - z_m)^2 \\ &= (w_m - (z_m + \mu\alpha))^2 \; + z_m^2 \; - \; (z_m + \mu\alpha)^2 \end{aligned} \tag{11.26}$$

It follows that the minimizer of $h_m(w_m)$ over the range $w_m \leq 0$ is given by

$$w_m^- \;\stackrel{\Delta}{=}\; \underset{w_m \leq 0}{\text{argmin}}\; h_m(w_m) = \begin{cases} 0, & \text{if } \; z_m > -\mu\alpha \\ z_m + \mu\alpha, & \text{if } \; z_m \leq -\mu\alpha \end{cases} \tag{11.27}$$

We conclude that the optimal value for each w_m is given by

$$\widehat{w}_m = \begin{cases} z_m - \mu\alpha, & \text{if } \; z_m \geq \mu\alpha \\ 0, & \text{if } \; -\mu\alpha < z_m < \mu\alpha \\ z_m + \mu\alpha, & \text{if } \; z_m \leq -\mu\alpha \end{cases} \tag{11.28}$$

which agrees with (11.17).

■

Example 11.2 **(A second useful choice)** Let

$$h(w) = \alpha\|w\|_1 + \frac{\rho}{2}\|w\|^2, \;\; \alpha > 0, \;\; \rho > 0 \tag{11.29}$$

which involves a combination of ℓ_1 and ℓ_2-norms. For this choice of $h(w)$, a similar derivation leads to the expression — see Prob. 11.4:

$$\widehat{w} \;=\; \mathbb{T}_{\frac{\mu\alpha}{1+\mu\rho}}\left(\frac{z}{1+\mu\rho}\right) \tag{11.30}$$

Compared with (11.20), we see that the value of z is scaled by $(1+\mu\rho)$ and the threshold value is also scaled by the same amount.

Example 11.3 **(General statement I)** It is useful to state the main conclusion that follows from the derivation leading to (11.17). This conclusion is of general interest and will be called upon later in different contexts, especially when we study regularization problems. Thus, consider a generic optimization problem of the following form (compare with (11.15)):

$$\boxed{\widehat{w} \;\stackrel{\Delta}{=}\; \underset{w\in \mathbb{R}^M}{\text{argmin}} \;\left\{\alpha\|w\|_1 \;+\; \frac{1}{2\mu}\|w-z\|^2 \;+\; \phi\right\}} \tag{11.31}$$

for some constants $\alpha \geq 0$, $\mu > 0$, ϕ, and vector $z \in \mathbb{R}^M$. Then, the solution is unique and given by the following expression:

$$\boxed{\widehat{w} \;=\; \mathbb{T}_{\mu\alpha}(z) \;=\; \text{sign}(z) \odot \Big(|z| - \mu\alpha\,\mathbb{1}\Big)_+} \tag{11.32}$$

in terms of the soft-thresholding function whose entry-wise operation is defined by (11.18), and where $\odot$ denotes the Hadamard elementwise product.

Example 11.4 (**General statement II**) Consider a diagonal scaling matrix

$$D = \text{diag}\Big\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_M^2\Big\} \tag{11.33}$$

with $\sigma_m^2 > 0$ and replace (11.31) by

$$\boxed{\widehat{w} \;\stackrel{\Delta}{=}\; \underset{w\in \mathbb{R}^M}{\text{argmin}} \;\left\{\alpha\|Dw\|_1 \;+\; \frac{1}{2\mu}\|w-z\|^2 \;+\; \phi\right\}} \tag{11.34}$$

Repeating the arguments that led to (11.32), it can be verified that — see Prob. 11.11:

$$\boxed{\widehat{w} \;=\; \text{sign}(z) \odot D\Big(D^{-1}|z| - \mu\alpha\,\mathbb{1}\Big)_+} \tag{11.35}$$

where the operations sign(x), $|x|$, and $(a)_+$ are applied element-wise.

We collect in Table 11.1 several proximal operators established before and in the problems at the end of the chapter. In the table, the notation $(a)_+ = \max\{0, a\}$ and $\mathbb{I}[x] = 1$ when statement x is true; otherwise, it is equal to zero.

11.1.3 Fixed Points

One main motivation for introducing proximal operators is the fact that fixed points for the proximal mapping coincide with global minimizers for the convex function $h(w)$. Specifically, if we let w^o denote any global minimum for $h(w)$, i.e., a point where

$$0 \in \partial_w\, h(w^o) \tag{11.36}$$

then w^o will be a fixed point for $\text{prox}_{\mu h}(z)$, namely, it holds that:

$$\boxed{\underbrace{w^o = \text{prox}_{\mu h}(w^o)}_{\textbf{fixed point}} \;\Longleftrightarrow\; \underbrace{0 \in \partial_w\, h(w^o)}_{\textbf{global minimum}}} \tag{11.37}$$

Table 11.1 Some useful proximal operators along with some properties.

	convex function, $h(w)$	**proximal operator**, $\widehat{w} = \text{prox}_{\mu h}(z)$	**reference**
1.	$\mathbb{I}_{C,\infty}[w]$ (C convex set)	$\widehat{w} = \mathcal{P}_C(z)$ (projection operator)	Eq. (11.10)
2.	$\alpha\|w\|_1$, $\alpha > 0$	$\widehat{w} = \mathbb{T}_{\mu\alpha}(z)$ (soft thresholding)	Eq. (11.20)
3.	$\alpha\|w\|_1 + \frac{\rho}{2}\|w\|^2$	$\widehat{w} = \mathbb{T}_{\frac{\mu\alpha}{1+\mu\rho}}\left(\dfrac{z}{1+\mu\rho}\right)$	Eq. (11.30)
4.	$\alpha\|w\|$	$\widehat{w} = \left(1 - \frac{\mu\alpha}{\|z\|}\right)_+ z$	Prob. 11.7
5.	$\alpha\|w\|_0$	$\widehat{w} = z\,\mathbb{I}\Big[\,\|z\| > \sqrt{2\mu\alpha}\,\mathbb{1}\Big]$	Prob. 11.8
6.	$g(w) = h(w) + c$	$\text{prox}_{\mu g}(z) = \text{prox}_{\mu h}(z)$	Prob. 11.1
7.	$g(w) = h(\alpha w + b)$	$\text{prox}_g(z) = \frac{1}{\alpha}\Big(\text{prox}_{\alpha^2 h}(\alpha z + b) - b\Big)$	Prob. 11.2
8.	$g(w) = h(w) + \frac{\rho}{2}\|w\|^2$	$\text{prox}_{\mu g}(z) = \text{prox}_{\frac{\mu h}{1+\mu\rho}}\left(\dfrac{z}{1+\mu\rho}\right)$	Prob. 11.6

This fact follows immediately from property (11.13). Consequently, iterative procedures for finding fixed points of $\text{prox}_{\mu h}(z)$ can be used to find minimizers for $h(w)$. Although unnecessary, we will generally be dealing with the case when there is a unique minimizer, w^o, for $h(w)$ and, correspondingly, a unique fixed point for $\text{prox}_{\mu h}(z)$.

11.2 PROXIMAL POINT ALGORITHM

We now exploit property (11.37) to motivate iterative procedures that converge to global minima w^o for $h(w)$. We know from (11.37) that these minima as fixed points for the proximal operator, i.e., they satisfy

$$w^o = \text{prox}_{\mu h}(w^o) \tag{11.38}$$

There are many iterative constructions that can be used to seek fixed points for operators of this type. We examine in this section the *proximal point algorithm*.

Let $w = f(z) : \mathbb{R}^M \rightarrow \mathbb{R}^M$ denote some generic strictly contractive operator, i.e., a mapping from z to w that satisfies

$$\|f(z_1) - f(z_2)\| \;<\; \lambda\,\|z_1 - z_2\|, \quad \text{for some } 0 \leq \lambda < 1 \tag{11.39}$$

with strict inequality for any vectors $z_1, z_2 \in \text{dom}(f)$. Then, a result known as the *Banach fixed-point theorem* ensures that such operators have unique fixed points, i.e., a unique w^o satisfying $w^o = f(w^o)$ and, moreover, this point can be determined by iterating:

$$w_n \;=\; f(w_{n-1}), \;\; n \geq 0 \tag{11.40}$$

starting from any initial condition w_{-1}. Then, $w_n \rightarrow w^o$ as $n \rightarrow \infty$ — see Prob. 11.18.

Remark 11.1. (Convention for initial conditions) Throughout our treatment, most recursive implementations will run over $n \geq 0$, as in (11.40), which means that the initial condition for the recursion will be specified at $n = -1$. This is simply a matter of convention; one can of course run recursions for $n > 0$ and specify the initial condition at $n = 0$. We adopt $n = -1$ as the time instant for the initial condition, w_{-1}.

■

The main challenge in using (11.40) for proximal operators is that the function $\text{prox}_{\mu h}(z)$ is *not* strictly contractive. It is shown in Prob. 11.19 that $\text{prox}_{\mu h}(z)$ satisfies

$$\|\text{prox}_{\mu h}(z_1) - \text{prox}_{\mu h}(z_2)\| \leq \; \|z_1 - z_2\| \tag{11.41}$$

with *inequality* rather than strict inequality as required by (11.39). In this case, we say that the operator $\text{prox}_{\mu h}(z)$ is *non-expansive*. Nevertheless, an iterative procedure can still be developed for determining fixed points for $\text{prox}_{\mu h}(z)$ by exploiting the fact that the operator is, in addition, *firmly* non-expansive. This means that $\text{prox}_{\mu h}(z)$ satisfies the stronger property — see again Prob. 11.19:

$$\|\text{prox}_{\mu h}(z_1) - \text{prox}_{\mu h}(z_2)\|^2 \leq \; (z_1 - z_2)^{\mathsf{T}}\Big(\text{prox}_{\mu h}(z_1) - \text{prox}_{\mu h}(z_2)\Big) \tag{11.42}$$

This relation implies (11.41); it also implies the following conclusion, in view of the Cauchy-Schwarz relation for inner products:

$$\begin{aligned} &\|z_1 - z_2\| = \|\text{prox}_{\mu h}(z_1) - \text{prox}_{\mu h}(z_2)\| \\ &\overset{(11.42)}{\Longleftrightarrow} \; z_1 - z_2 \;=\; \text{prox}_{\mu h}(z_1) - \text{prox}_{\mu h}(z_2) \end{aligned} \tag{11.43}$$

By using this fact, along with (11.41), it is shown in Prob. 11.20 that for such (firmly non-expansive) operators, an iteration similar to (11.40) will converge to a fixed point w^o of $\text{prox}_{\mu h}(z)$, namely,

$$\boxed{w_n \;=\; \text{prox}_{\mu h}(w_{n-1}), \;\; n > 0} \qquad \textbf{(proximal iteration)} \tag{11.44}$$

Recursion (11.44) is known as the *proximal point algorithm*, which is summarized in (11.45).

Proximal point algorithm for minimizing $h(w)$.

```
function $h(w)$ is convex;
given the proximal operator for $h(w)$;
start from an arbitrary initial condition $w_{-1}$.
repeat over $n \geq 0$ until convergence:
    $w_n \;=\; \text{prox}_{\mu h}(w_{n-1})$
end
return minimizer $w^o \leftarrow w_n$.
```

(11.45)

By appealing to cases where the proximal projection has a closed-form expression in terms of the entries of w_{n-1}, we arrive at a *realizable* implementation for (11.45). For example, when $h(w) = \alpha\|w\|_1$, we already know from (11.20) that the proximal step in (11.45) can be replaced by:

$$w_n \;=\; \mathbb{T}_{\mu\alpha}(w_{n-1}), \quad n > 0 \tag{11.46}$$

On the other hand, when $h(w) = \alpha\|w\|_1 + \frac{\rho}{2}\|w\|^2$, we would use instead:

$$w_n \;=\; \mathbb{T}_{\frac{\mu\alpha}{1+\mu\rho}}\left(\frac{w_{n-1}}{1+\mu\rho}\right), \quad n > 0 \tag{11.47}$$

based on the result of Prob. 11.6. Procedures of the form (11.46) and (11.47) are called *iterative soft-thresholding algorithms* for obvious reasons.

11.3 PROXIMAL GRADIENT ALGORITHM

The proximal iteration (11.45) is useful for seeking global minimizers of stand-alone convex functions $h(w)$. However, the algorithm requires the proximal operator for $h(w)$, which is not always available in closed form. We now consider an extension of the method for situations where the objective function $h(w)$ can be expressed as the sum of *two* components, and where the proximal operator for one of the components is available in closed form. The extension leads to the *proximal gradient algorithm.* In preparation for the notation used in subsequent chapters where the optimization objective is denoted by $P(w)$, we will henceforth replace the notation $h(w)$ by $P(w)$ and seek to minimize $P(w)$.

Thus, assume that we are faced with an optimization problem of the form:

$$\min_{w\in\mathbb{R}^M} \left\{P(w) \;\overset{\Delta}{=}\; q(w) + E(w)\right\} \tag{11.48}$$

involving the sum of two convex components, $E(w)$ and $q(w)$. Usually, $E(w)$ is differentiable and $q(w)$ is non-smooth. For example, we will encounter in later chapters, while studying the LASSO and basis pursuit problems, optimization problems of the following form:

$$\min_{w\in\mathbb{R}^M} \left\{\alpha\|w\|_1 \;+\; \|d-Hw\|^2\right\} \tag{11.49}$$

where $d \in \mathbb{R}^{N\times 1}$ and $H \in \mathbb{R}^{N\times M}$. For this case, we have $q(w) = \alpha\|w\|_1$ and $E(w) = \|d - Hw\|^2$. At first sight, we could consider applying the proximal iteration (11.44) directly to the aggregate function $P(w)$ and write

$$w_n \;=\; \text{prox}_{\mu P}(w_{n-1}),\;\; n \geq 0 \tag{11.50}$$

The difficulty with this approach is that, in general, the proximal operator for $P(w)$ may not be available in closed form. We now derive an alternative procedure for situations when the form of $\text{prox}_{\mu q}(z)$ is known.

Thus, note first that a minimizer for $P(w)$ is a fixed point for the operator $\text{prox}_{\mu q}(z - \mu\nabla_{w^{\mathsf{T}}} E(z))$. Indeed, let w^o denote a fixed point for this operator so that

$$w^o \;=\; \text{prox}_{\mu q}\Big(w^o - \mu\nabla_{w^{\mathsf{T}}} E(w^o)\Big) \tag{11.51a}$$

Then, we know from property (11.13) that

$$\Big(w^o \;-\; \mu\nabla_{w^{\mathsf{T}}} E(w^o) - w^o\Big) \;\in\; \mu\,\partial_{w^{\mathsf{T}}}\, q(w^o) \tag{11.51b}$$

from which we conclude that

$$0 \;\in\; \nabla_{w^{\mathsf{T}}} E(w^o) + \partial_{w^{\mathsf{T}}}\, q(w^o) \tag{11.51c}$$

or, equivalently,

$$0 \;\in\; \partial_{w^{\mathsf{T}}}\, P(w^o) \tag{11.51d}$$

We therefore find, as claimed, that the fixed point w^o is a minimizer for $P(w)$. We can then focus on finding fixed points for $\text{prox}_{\mu q}(z - \mu\nabla_{w^{\mathsf{T}}} E(z))$ by using the recursion (cf. (11.44)):

$$\boxed{w_n \;=\; \text{prox}_{\mu q}\Big(w_{n-1} - \mu\nabla_{w^{\mathsf{T}}} E(w_{n-1})\Big), \quad n \geq 0} \tag{11.52}$$

We refer to this implementation as the *proximal gradient algorithm*, which we rewrite in expanded form in (11.53) by introducing an intermediate variable z_n. The algorithm involves two steps: a gradient-descent step on the differentiable component $E(w)$ to obtain z_n, followed by a proximal projection step relative to the nonsmooth component $q(w)$.

Proximal gradient algorithm for minimizing $P(w) = q(w) + E(w)$.

```
$q(w)$ and $E(w)$ are convex functions;
given the proximal operator for the nonsmooth component, $q(w)$;
given the gradient operator for the smooth component, $E(w)$;
start from an arbitrary initial condition $w_{-1}$.                    (11.53)
repeat over $n \geq 0$ until convergence:
    $z_n = w_{n-1} - \mu \nabla_{w^\mathsf{T}} E(w_{n-1})$
    $w_n = \text{prox}_{\mu q}(z_n)$
end
return minimizer $w^o \leftarrow w_n$.
```

Remark 11.2. (Forward-backward splitting) In view of property (11.13), the proximal step in (11.53) implies the following relation:

$$w_n = \text{prox}_{\mu q}(z_n) \iff (z_n - w_n) \in \mu\, \partial_{w^\mathsf{T}} q(w_n) \tag{11.54}$$

which in turn implies that we can rewrite the algorithm in the form:

$$\text{(forward-backward splitting)} \quad \left\{ \begin{array}{lcl} z_n & = & w_{n-1} - \mu \nabla_{w^\mathsf{T}} E(w_{n-1}) \\ w_n & = & z_n - \mu\, \partial_{w^\mathsf{T}} q(w_n) \end{array} \right. \tag{11.55}$$

where $\partial_{w^\mathsf{T}} q(w_n)$ denotes *some* subgradient for $q(w)$ at location w_n. In this form, the first step corresponds to a forward update step moving from w_{n-1} to z_n, while the second step corresponds to a backward (or implicit) update since it involves w_n on both sides. For these reasons, recursion (11.53) is sometimes referred to as a *forward-backward splitting* implementation.

■

Example 11.5 **(Two useful cases)** Let us consider two special instances of formulation (11.48). When $q(w) = \alpha\|w\|_1$, the proximal gradient algorithm reduces to

$$\left\{ \begin{array}{lcl} z_n & = & w_{n-1} - \mu \nabla_{w^\mathsf{T}} E(w_{n-1}) \\ w_n & = & \mathbb{T}_{\mu\alpha}(z_n) \end{array} \right. \tag{11.56}$$

and when $q(w) = \alpha\|w\|_1 + \frac{\rho}{2}\|w\|^2$, we get

$$\left\{ \begin{array}{lcl} z_n & = & w_{n-1} - \mu \nabla_{w^\mathsf{T}} E(w_{n-1}) \\ w_n & = & \mathbb{T}_{\frac{\mu\alpha}{1+\mu\rho}}\left(\dfrac{z_n}{1+\mu\rho}\right) \end{array} \right. \tag{11.57}$$

Example 11.6 **(Logistic cost function)** Consider a situation where $E(w)$ is chosen as the logistic loss function, i.e.,

$$E(w) = \ln\left(1 + e^{-\gamma h^\mathsf{T} w}\right) \tag{11.58}$$

where $\gamma \in \mathbb{R}$ and $h \in \mathbb{R}^M$. Note that we are using here the notation h to refer to the column vectors that appear in the exponent expression; this is in line with our future notation for feature vectors in subsequent chapters. Let

$$q(w) = \alpha\|w\|_1 + \frac{\rho}{2}\|w\|^2 \tag{11.59}$$

so that the function that we wish to minimize is

$$P(w) = \alpha\|w\|_1 + \frac{\rho}{2}\|w\|^2 + \ln\left(1 + e^{-\gamma h^{\mathsf{T}} w}\right) \tag{11.60}$$

The function $E(w)$ is differentiable everywhere with

$$\nabla_{w^{\mathsf{T}}} E(w) \;=\; -\gamma h \times \frac{1}{1 + e^{\gamma h^{\mathsf{T}} w}} \tag{11.61}$$

and we find that the proximal gradient recursion (11.53) reduces to

$$\begin{cases} z_n = w_{n-1} \;+\; \mu\gamma h \times \frac{1}{1+e^{\gamma h^{\mathsf{T}} w_{n-1}}} \\ w_n = \mathbb{T}_{\frac{\mu\alpha}{1+\mu\rho}}\left(\dfrac{z_n}{1+\mu\rho}\right) \end{cases} \tag{11.62}$$

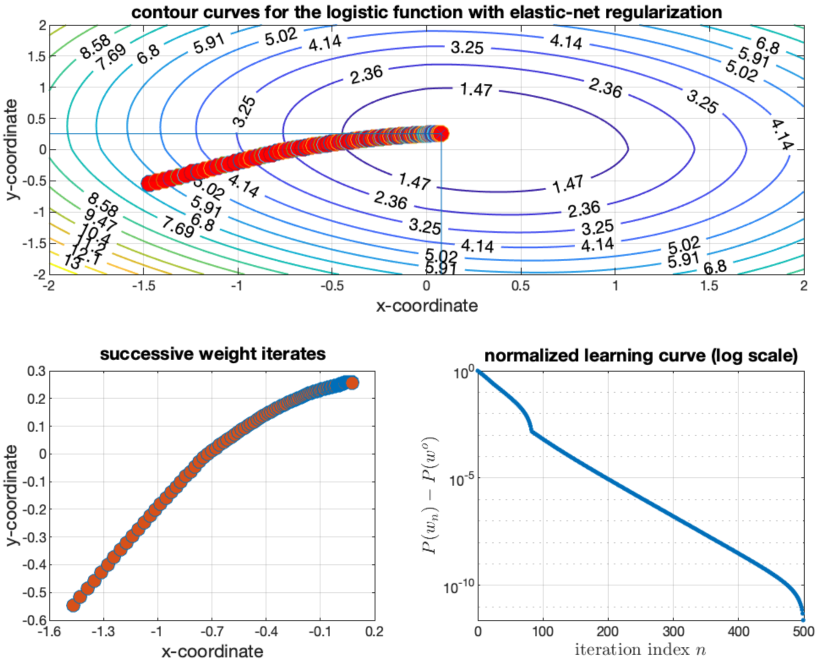

Figure 11.2 (*Top*) Contour curves of a regularized logistic function $P(w)$ of the form (11.60) with $\alpha = 0.2$, $\rho = 2$, $\gamma = 1$ and $h = \text{col}\{1, 2\}$. The successive locations of the weight iterates generated by the proximal gradient recursion (11.62) are shown in circles moving from left to right towards the minimizer location of $P(w)$. (*Bottom left*) Trajectory of the successive weight iterates in $\mathbb{R}^2$, moving from left to right, as they approach the location of the minimizer w^o of $P(w)$. (*Bottom right*) Normalized learning curve in logarithmic scale. The curve illustrates the expected "linear convergence" rate for the successive iterate values towards $P(w^o)$.

Figure 11.2 plots the contour curves for the regularized logistic function $P(w) : \mathbb{R}^2 \to \mathbb{R}$ with parameters

$$\gamma = 1, \quad h = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \alpha = 0.2, \quad \rho = 2 \tag{11.63}$$

The location of the minimizer w^o for $P(w)$ and the corresponding minimum value are determined to be approximately:

$$w^o \approx \begin{bmatrix} 0.0782 \\ 0.2564 \end{bmatrix}, \quad P(w^o) \approx 0.5795 \tag{11.64}$$

These values are obtained by running the proximal gradient recursion (11.62) for 500 iterations starting from a random initial condition w_{-1} and using $\mu = 0.01$. The top plot in the figure illustrates the trajectory of the successive weight iterates, moving from left to right, in relation to the contour curves of $P(w)$; this same trajectory is shown in the lower left plot of the same figure. The lower right plot shows the evolution of the learning curve $P(w_n) - P(w^o)$ over the iteration index n in logarithmic scale for the vertical axis. Specifically, the figure shows the evolution of the *normalized* quantity

$$\ln\left(\frac{P(w_n) - P(w^o)}{\max_n\{P(w_n) - P(w^o)\}}\right) \tag{11.65}$$

where the curve is normalized by its maximum value so that the peak value in the logarithmic scale is zero. This will be our standing assumption in plots for learning curves in the logarithmic scale; the peak values will be normalized to zero. It is clear from the figure, due to its decaying “linear form” that the convergence of $P(w_n)$ towards $P(w^o)$ occurs at an exponential rate, as predicted further ahead by result (11.74).

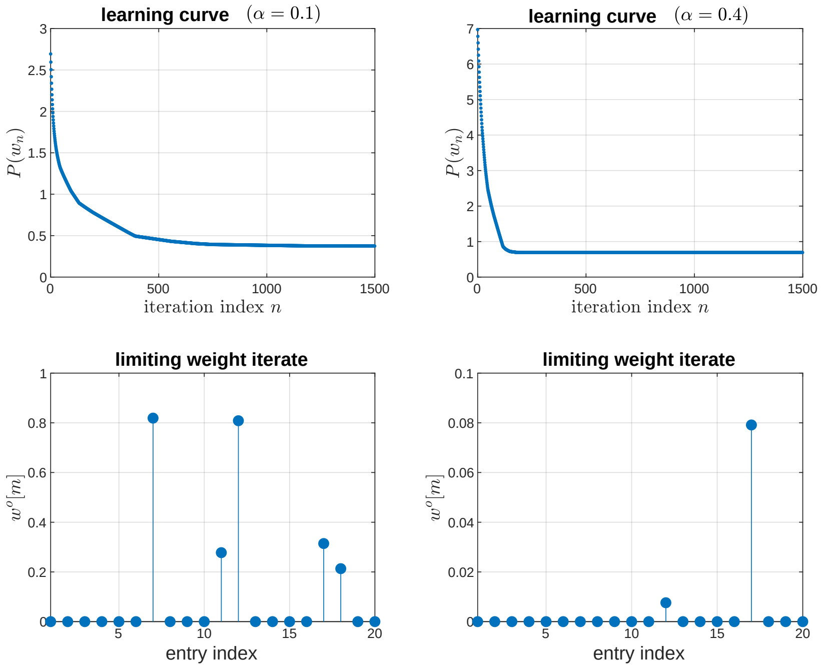

Figure 11.3 (*Top*) Learning curves for the proximal gradient recursion (11.62) using $\mu = 0.03$ when applied to $P(w) = \alpha\|w\|_1 + \ln(1 + e^{-\gamma h^\mathsf{T} w})$ using $\alpha = 0.1$ and $\alpha = 0.4$. We use $\gamma = 1$ and generate a random h with entries selected uniformly from within the interval $[-1, 1]$. (*Bottom*) Form of resulting minimizers, w^o, for both choices of α. Observe how the minimizer w^o is more sparse for $\alpha = 0.4$ than for $\alpha = 0.1$.

Figure 11.3 illustrates the influence of the parameter α on the sparsity of the minimizer w^o. We set $\rho = 0$ and plot the learning curves and the resulting minimizers for $P(w)$ with $w \in \mathbb{R}^{20}$, i.e., $M = 20$. We continue to use $\gamma = 1$ but generate a random h with entries selected uniformly from within the interval $[-1, 1]$. The curves in the figure are the result of running the proximal gradient recursion (11.62) for 1500 iterations using $\mu = 0.03$. The learning curves tend towards the minimum values $P(w^o) \approx 0.3747$ for $\alpha = 0.1$ and $P(w^o) \approx 0.6924$ for $\alpha = 0.4$. Observe how the minimizer w^o is more sparse for $\alpha = 0.4$ than for $\alpha = 0.1$.

Example 11.7 **(LASSO or basis pursuit)** Consider the optimization problem

$$w^o = \underset{w\in\mathbb{R}^M}{\text{argmin}} \left\{ P(w) \overset{\Delta}{=} \alpha\|w\|_1 \; + \; \|d - Hw\|^2 \right\} \tag{11.66}$$

where $d \in \mathbb{R}^{N\times 1}$ and $H \in \mathbb{R}^{N\times M}$. We will encounter this problem later when we study sparsity-inducing solutions and, in particular, the LASSO and basis pursuit algorithms. The function $E(w) = \|d - Hw\|^2$ is differentiable everywhere with

$$\nabla_{w^{\mathsf{T}}} E(w) \; = \; -2H^{\mathsf{T}}(d - Hw) \tag{11.67}$$

so that the proximal gradient recursion (11.53) reduces to

$$\left\{ \begin{array}{lcl} z_n & = & w_{n-1} \; + \; 2\mu H^{\mathsf{T}}(d - Hw_{n-1}) \\ w_n & = & \mathbb{T}_{\mu\alpha}(z_n) \end{array} \right. \tag{11.68}$$

Figure 11.4 plots the contour curves for $P(w) : \mathbb{R}^2 \to \mathbb{R}$ with $\alpha = 0.5$, $M = 2$, and $N = 50$. The quantities $\{d, H\}$ are generated randomly; their entries are zero-mean Gaussian distributed with unit variance. The location of the minimizer w^o for $P(w)$ and the corresponding minimum value are determined to be approximately

$$w^o \approx \begin{bmatrix} 0.0205 \\ 0 \end{bmatrix}, \quad P(w^o) \approx 42.0375 \tag{11.69}$$

These values are obtained by running the proximal gradient recursion (11.68) for 400 iterations starting from a random initial condition w_0 and using $\mu = 0.001$. The top plot in the figure illustrates the trajectory of the successive weight iterates, moving from left to right, in relation to the contour curves of $P(w)$; this same trajectory is shown in the lower left plot of the same figure. The lower right plot shows the evolution of the normalized learning curve $P(w_n) - P(w^o)$ over the iteration index n in logarithmic scale according to the same construction from (11.65).

11.4 CONVERGENCE RESULTS

We list two convergence results for the proximal gradient algorithm (11.53) for small enough μ. We examine two cases: **(a)** $E(w)$ is convex and **(b)** $E(w)$ is strongly convex. The latter case has faster convergence rate. The proof of the following first result appears in Appendix 11.A.

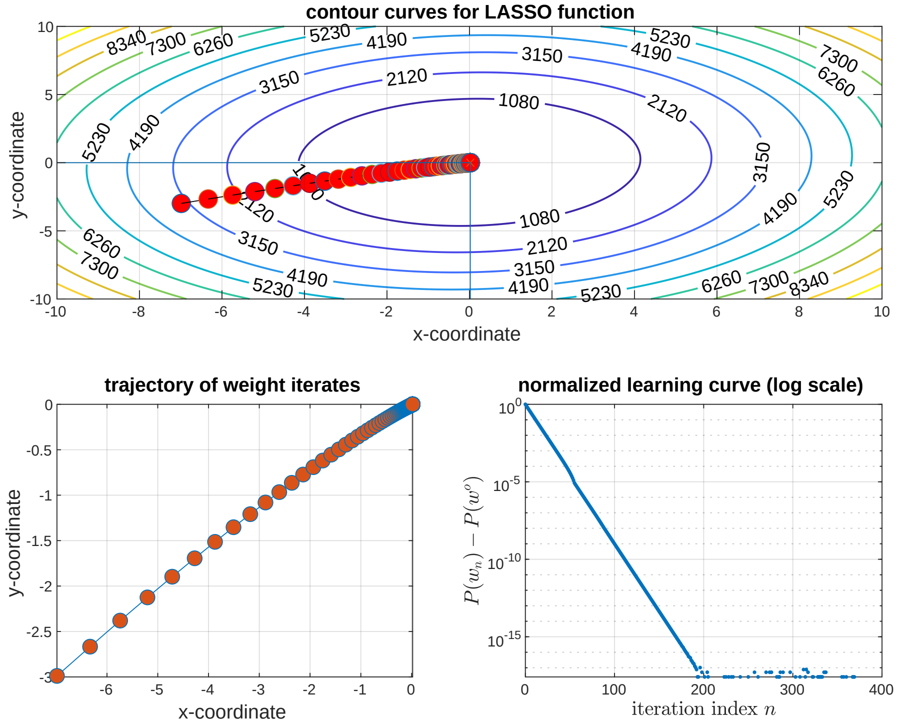

Figure 11.4 (*Top*) Contour curves of a regularized LASSO function $P(w)$ of the form (11.66) with $\alpha = 0.5$, $M = 2$, and $N = 50$. The quantities (d, H) are generated randomly; their entries are zero-mean Gaussian distributed with unit variance. The successive locations of the weight iterates generated by the proximal gradient recursion (11.68) are shown in circles moving from left to right towards the minimizer location of $P(w)$. (*Bottom left*) Trajectory of the successive weight iterates in $\mathbb{R}^2$, moving from left to right, as they approach the location of the minimizer w^o of $P(w)$. (*Bottom right*) Normalized learning curve in logarithmic scale.

Theorem 11.1. (Convergence of proximal gradient algorithm) *Consider the problem of minimizing $P(w) = q(w) + E(w)$ where $q(w)$ and $E(w)$ are both convex functions, with $E(w)$ differentiable and having $\delta-$Lipschitz gradients. If*

$$\mu < 1/\delta \tag{11.70}$$

then the proximal gradient algorithm (11.53) converges to a global minimizer w^o of $P(w)$ at the following rate

$$P(w_n) - P(w^o) \leq \frac{1}{2(n+1)\mu}\|w^o - w_{-1}\|^2 \;=\; O(1/n) \tag{11.71}$$

where w_{-1} is an arbitrary initial condition.

Remark 11.3 (Big-O notation). Statement (11.71) uses the big-O notation, which we already encountered in the earlier expression (3.226). This notation will appear regularly in our presentation and it is used to compare the asymptotic growth rates of sequences. Thus, recall that writing $a_n = O(b_n)$ means $|a_n| \leq cb_n$ for some constant $c > 0$ and for large enough n, say, $n > n_o$ for some n_o. For example, writing $a_n = O(1/n)$ means that the sequence a_n decays asymptotically at a rate that is comparable to or faster than $1/n$.

■

Theorem 11.1 establishes that for convex $E(w)$, the cost $P(w_n)$ approaches the minimum value $P(w^o)$ at the rate $O(1/n)$. Faster convergence at exponential rate is possible when $E(w)$ is $\nu-$strongly convex, i.e., when

$$E(w_2) \;\geq\; E(w_1) \;+\; \left(\nabla_{w^{\mathsf{T}}} E(w_1)\right)^{\mathsf{T}} (w_2 - w_1) \;+\; \frac{\nu}{2}\|w_2 - w_1\|^2 \tag{11.72}$$

for any $w_1, w_2 \in \text{dom}(E)$. The proof of this second result appears in Appendix 11.B.

Theorem 11.2. (Exponential convergence under strong convexity) *Consider the same setting of Theorem 11.1 except that $E(w)$ is now assumed to be $\nu-$strongly convex. If*

$$\mu < 2\nu/\delta^2 \tag{11.73}$$

then the proximal gradient algorithm (11.53) converges to the global minimizer w^o of $P(w)$ at the exponential rate

$$P(w_n) - P(w^o) \leq \beta\lambda^n \|w^o - w_{-1}\|^2 \;=\; O(\lambda^n) \tag{11.74}$$

for some constant β and where

$$\lambda \;\stackrel{\Delta}{=}\; 1 - 2\mu\nu + \mu^2\delta^2 \;\in [0, 1) \tag{11.75}$$

Remark 11.4. (**A more relaxed bound on μ**) The result of Theorem 11.2 establishes the exponential convergence of the excess risk to zero for sufficiently small step-sizes, μ. In most instances, this result is sufficient for our purposes since our objective will generally be to verify whether the iterative algorithms approach their desired limit. This conclusion is established in Theorem 11.2 under the bound $\mu < 2\nu/\delta^2$. We can relax the bound and show that convergence will continue to occur for the more relaxed bound $\mu < 2/\delta$ at the rate $O((\lambda')^n)$, where $\lambda' = 1 - 2\mu\nu + \mu^2\nu\delta$. We explain in the same Appendix 11.B. that this can be achieved by exploiting a certain *co-coercivity* property that is satisfied by convex functions with $\delta-$Lipschitz gradients — see Example 11.8.

■

11.5 DOUGLAS-RACHFORD ALGORITHM

We continue with the optimization problem (11.48) except that we now allow for both functions $q(w)$ and $E(w)$ to be non-smooth and present a splitting algorithm for minimizing their aggregate sum. This second algorithm will involve the proximal operators for *both* functions and is therefore suitable when these proximal operators can be determined beforehand. There are several variations of the Douglas–Rachford algorithm — see Probs. 11.22 and 11.23. We list one form in (11.76).

Douglas–Rachford algorithm for minimizing $P(w) = q(w) + E(w)$.

```
$q(w)$ and $E(w)$ are (possibly nonsmooth) convex functions;
given the proximal operator for $q(w)$;
given the proximal operator for $E(w)$;
start from an arbitrary initial condition $z_{-1}$;
repeat over $n \geq 0$ until convergence:                          (11.76)
    $w_n = \text{prox}_{\mu q}(z_{n-1})$
    $t_n = \text{prox}_{\mu E}(2w_n - z_{n-1})$
    $z_n = t_n - w_n + z_{n-1}$
end
return minimizer $w^o \leftarrow w_n$.
```

The main motivation for the algorithm lies in the fact that the mapping from z_{n-1} to z_n in (11.76) can be shown to be firmly non-expansive and that fixed points for this mapping determine the desired minimizer(s) for (11.48). Let us comment on the second property first. For this purpose, we write down the fixed-point relations:

$$\begin{cases} w^o &= \text{prox}_{\mu q}(z^o) \\ t^o &= \text{prox}_{\mu E}(2w^o - z^o) \\ z^o &= t^o - w^o + z^o \end{cases} \tag{11.77}$$

where we replaced the variables $\{w_n, w_{n-1}, z_n, z_{n-1}, t_n\}$ by fixed-point values $\{w^o, z^o, t^o\}$. Using property (11.13) for proximal projections, these relations translate into:

$$\begin{cases} (z^o - w^o) &\in \mu\, \partial_{w^\mathsf{T}}\, q(w^o) \\ (2w^o - z^o - t^o) &\in \mu\, \partial_{w^\mathsf{T}}\, E(t^o) \\ w^o &= t^o \end{cases} \tag{11.78}$$

which imply that

$$(w^o - z^o) \in \mu\, \partial_{w^\mathsf{T}} E(w^o) \quad \text{and} \quad (z^o - w^o) \in \mu\, \partial_{w^\mathsf{T}} q(w^o) \tag{11.79}$$

Using the result of Prob. 8.31 we conclude that

$$0 \in \partial_{w^\mathsf{T}} \Big(E(w^o) + q(w^o) \Big) \tag{11.80}$$

which confirms that w^o is a minimizer for the sum $E(w) + q(w)$, as claimed. This argument establishes that if z^o is a fixed point for the mapping from z_{n-1} to z_n in (11.76), then $w^o = \text{prox}_{\mu q}(z^o)$ is a minimizer for the $P(w)$. We still need to establish that the mapping from z_{n-1} to z_n is firmly non-expansive, in which case algorithm (11.76) would correspond to a fixed-point iteration applied to this mapping. If we group the three relations appearing in (11.76) we find that the mapping from z_{n-1} to z_n is given by

$$z_n = z_{n-1} + \text{prox}_{\mu E}\Big(2\text{prox}_{\mu q}(z_{n-1}) - z_{n-1}\Big) \; - \; \text{prox}_{\mu q}(z_{n-1}) \tag{11.81}$$

which we denote more compactly by writing $z_n = R(z_{n-1})$ with the mapping $R(z)$ defined by

$$R(z) \; \stackrel{\Delta}{=} \; z + \text{prox}_{\mu E}\Big(2\text{prox}_{\mu q}(z) - z\Big) \; - \; \text{prox}_{\mu q}(z) \tag{11.82}$$

By exploiting the fact that proximal operators are themselves firmly non-expansive, it is shown in Prob. 11.21 that $R(z)$ is also firmly non-expansive, meaning that it satisfies:

$$\|R(z_1) - R(z_2)\|^2 \leq \; (z_1 - z_2)^{\mathsf{T}} \left(R(z_1) - R(z_2)\right) \tag{11.83}$$

for any z_1, z_2. In this case, recursion (11.81) can be viewed as a fixed-point iteration for this mapping and implementation (11.76) amounts to unfolding this iteration into three successive steps.

11.6 COMMENTARIES AND DISCUSSION

Proximal operators. The proximal projection of a convex function $h(w)$ was defined in (11.4) as the mapping that transforms a vector $z \in \mathbb{R}^M$ into the vector:

$$\widehat{w} \; = \; \text{prox}_{\mu h}(z) \; \stackrel{\Delta}{=} \; \underset{w\in\mathbb{R}^M}{\text{argmin}} \; \left\{ h(w) \; + \; \frac{1}{2\mu}\|w - z\|^2 \right\} \tag{11.84}$$

with the resulting minimum value given by its Moreau envelope from (11.6):

$$\mathcal{M}_{\mu h}(z) \; \stackrel{\Delta}{=} \; \min_{w\in\mathbb{R}^M} \left\{ h(w) \; + \; \frac{1}{2\mu}\|w - z\|^2 \right\} \tag{11.85}$$

This envelope is also called the Moreau-Yosida envelope in recognition of the contributions by Moreau (1965) and Yosida (1968). Intuitively, the proximal construction approximates z by some vector, $\widehat{w}$, that is close to it under the squared Euclidean norm but subject to the penalty $h(w)$. The Moreau value at $\widehat{w}$ serves as a measure of a (generalized) distance between z and its proximal projection.

We discussed several properties of proximal projections in Sec. 11.1. The form of the soft-thresholding operator (11.18) appeared in the works by Donoho and Johnstone (1994,1995) on the recovery of signals embedded in additive Laplace-distributed noise. Other useful interpretations and properties of proximal operators can be found in the treatments by Lemaire (1989a,b), Rockafellar and Wets (1998), Combettes and Pesquet (2011), and Parikh and Boyd (2013). In Prob. 11.7 we highlight one useful connection of proximal operators to the Huber function, which is a popular tool in robust statistics used to reduce the effect of data outliers — see Huber (1981). We explain in that problem that the Moreau envelope that corresponds to the choice $h(w) = \|w\|$ is the Huber loss function, denoted by

$$H_\mu(z) = \begin{cases} \frac{1}{2\mu}\|z\|^2, & \|z\| \leq \mu \\ \|z\| - \frac{\mu}{2}, & \|z\| > \mu \end{cases} \tag{11.86}$$

where $z, w \in \mathbb{R}^M$. The Huber function is linear in $\|z\|$ over the range $\|z\| > \mu$, for some parameter $\mu > 0$ and, therefore, it penalizes less drastically large values for $\|z\|$ in comparison to the quadratic loss, $\|z\|^2$.

Moreau decomposition. The concept of the proximal operator (11.84) and its envelope were introduced and studied in a series of works by Moreau (1962,1963a,b,1965). One of the main driving themes in the work by Moreau (1965) was to establish the following interesting decomposition. Consider a convex function $h(w)$ defined over $w \in \mathbb{R}^M$, and let $h^\star(x)$ denote its conjugate function defined earlier by (8.83), i.e.,

$$h^\star(x) \;\overset{\Delta}{=}\; \sup_{w} \Big(x^\mathsf{T} w - h(w)\Big), \quad x \in \mathcal{X} \tag{11.87}$$

where $\mathcal{X}$ denotes the set of all x where the supremum operation is finite. It is shown in Prob. 8.47 that $\mathcal{X}$ is a convex set. Then, any vector $z \in \mathbb{R}^M$ can be decomposed as — see Prob. 11.24:

$$z = \text{prox}_h(z) \; + \; \text{prox}_{h^\star}(z) \tag{11.88a}$$

$$\frac{1}{2}\|z\|^2 = \mathcal{M}_h(z) + \mathcal{M}_{h^\star}(z) \tag{11.88b}$$

These expressions provide an interesting generalization of the orthogonal projection decomposition that is familiar from Euclidean geometry, namely, for any closed vector space $\mathcal{C} \subset \mathbb{R}^M$, every vector z can be decomposed as

$$z = \mathcal{P}_C(z) \; + \; \mathcal{P}_{C^\perp}(z) \tag{11.89a}$$

$$\|z\|^2 = \|\mathcal{P}_C(z)\|^2 \; + \; \|\mathcal{P}_{C^\perp}(z)\|^2 \tag{11.89b}$$

where $\mathcal{P}_C(z)$ denotes the orthogonal projection of z onto $\mathcal{C}$, and similarly $\mathcal{P}_{C^\perp}(z)$ denotes the orthogonal projection of z onto the orthogonal complement space, $\mathcal{C}^\perp$, namely,

$$\mathcal{P}_C(z) \;\overset{\Delta}{=}\; \min_{w \in \mathcal{C}} \|w - z\|^2 \tag{11.90}$$

where for any $x \in \mathcal{C}$ and $y \in \mathcal{C}^\perp$, it holds that $x^\mathsf{T} y = 0$.

Two other critical properties established by Moreau (1965) are that **(a)** the proximal operator is a firmly non-expansive mapping (cf. (11.42)) and **(b)** the Moreau envelope is *differentiable* over z, regardless of whether the original function $h(w)$ is differentiable or not over w. Actually, the gradient vector of the Moreau envelope is given by — see Prob. 11.25:

$$\nabla_{z^\mathsf{T}} \mathcal{M}_{\mu h}(z) \;=\; \frac{1}{\mu}\big(z - \text{prox}_{\mu h}(z)\big) \tag{11.91}$$

It is then clear from property (11.37) relating minimizers of $h(w)$ to fixed points of $\text{prox}_{\mu h}(z)$ that it also holds

$$\underbrace{w^o = \text{prox}_{\mu h}(w^o)}_{\textbf{fixed point}} \;\Longleftrightarrow\; \underbrace{0 \in \partial_w\, h(w^o)}_{\textbf{global minimum}} \;\Longleftrightarrow\; \nabla_z\, \mathcal{M}_{\mu h}(w^o) = 0 \tag{11.92}$$

In this way, the problem of determining the minimizer(s) of a possibly non-smooth (i.e., non-differentiable) convex function $h(w)$ can be reduced to the equivalent problems of determining the fixed point(s) of a firmly non-expansive proximal operator or the stationary point(s) of a smooth (i.e., differentiable) Moreau envelope. This observation is further reinforced by noting that we can rewrite (11.91) as

$$\text{prox}_{\mu h}(z) \;=\; z - \mu \nabla_{z^\mathsf{T}} \mathcal{M}_{\mu h}(z) \tag{11.93}$$

which shows that the proximal operation amounts to performing a gradient-descent step over the Moreau envelope.

Fixed-point iterations. In future chapters, we will derive several algorithms for online learning by relying on the use of proximal projections to deal with optimization problems that involve non-smooth components. There are two key properties that make proximal operators particularly suitable for solving non-smooth optimization problems. The first property is the fact, established in (11.37), that their fixed points coincide with the minimizers of the functions defining them, namely,

$$\underbrace{w^o = \text{prox}_{\mu h}(w^o)}_{\textbf{fixed point}} \iff \underbrace{0 \in \partial_w h(w^o)}_{\textbf{global minimum}} \tag{11.94}$$

The second property is the fact that proximal operators are firmly non-expansive, meaning that they satisfy property (11.42). Consequently, as shown by (11.44), a convergent iteration can be used to determine their fixed points, which leads to the proximal point algorithm:

$$w_n \;=\; \text{prox}_{\mu h}(w_{n-1}), \quad n \geq 0 \tag{11.95}$$

These observations have motivated broad research efforts into constructing new families of algorithms for the solution of non-smooth convex optimization problems by exploiting the theory of fixed-point iterations for firmly non-expansive operators — see, e.g., the works by Minty (1962), Browder (1965,1967), Bruck and Reich (1977), and Combettes (2004), as well as the texts by Brezis (1973), Granas and Dugundji (2003), and Bauschke and Combettes (2011). One of the earliest contributions in this regard is the proximal point algorithm (11.95). It was proposed by Martinet (1970,1972) as a way to construct iterates for minimizing a convex function $h(w)$ by solving successive problems of the type (using our notation):

$$w_n \;=\; \underset{w\in\mathbb{R}^M}{\text{argmin}} \; \left\{ h(w) + \frac{1}{2\mu}\|w - w_{n-1}\|^2 \right\} \tag{11.96}$$

where the variable z is replaced by the prior iterate w_{n-1}. According to (11.84), this leads to $w_n = \text{prox}_{\mu h}(w_{n-1})$, which is the proximal iteration (11.95). Two other early influential works in the area of proximal operators for non-smooth optimization, with stronger convergence results, are the articles by Rockafellar (1976a,b) on the proximal point algorithm and generalizations. Two early works on the proximal gradient method (11.52) are Sibony (1970) and Mercier (1979). Since then, several important advances have occurred in the development of techniques for the optimization of non-smooth problems. The presentation and convergence analysis in the chapter and appendices are motivated by the useful overviews given by Polyak (1987), Combettes and Pesquet (2011), Parikh and Boyd (2013), Polson, Scott, and Willard (2015), and Beck (2017), in addition to the contributions by Luo and Tseng (1992b,1993), Nesterov (2004,2005), Combettes and Wajs (2005), Figueiredo, Bioucas-Dias, and Nowak (2007), and Beck and Teboulle (2009a,2012).

Resolvents and splitting techniques. Motivated by expression (11.12) for $\widehat{w}$, it is customary in the literature to express the proximal operator of a convex function $h(w)$ by writing

$$\text{prox}_{\mu h}(z) \;=\; (I + \mu\, \partial_{w^\mathsf{T}} h)^{-1}(z) \tag{11.97}$$

The notation on the right-hand side means that the point z is mapped to its proximal projection $\widehat{w}$. The operation that performs this mapping is denoted either by $\text{prox}_{\mu h}(z)$, which is our standard notation, or more broadly by the operator notation $(I+\mu\partial_{w^\mathsf{T}} h)^{-1}$. This latter notation is called the *resolvent* of the operator $\mu\partial_{w^\mathsf{T}} h$. Observe that although the subdifferential of a function at any particular location is not uniquely defined (i.e., it generally maps one point in space to multiple points since there can be many choices for the subgradient vector), the result of the resolvent operation is always unique since we already know that proximal projections are unique. In this way, notation (11.97) maps z to a unique point $\widehat{w}$.

We can employ the resolvent notation to motivate splitting algorithms, along the lines of Lions and Mercier (1979) and Eckstein and Bertsekas (1992). For instance, one other way to motivate the Douglas–Rachford splitting procedure (11.76) is to consider initially the simpler but related problem of determining vectors $w \in \mathbb{R}^M$ that lie in the nullspace of the sum of two nonnegative-definite matrices, i.e., vectors w that satisfy

$$(A+B)w = 0 \iff Aw + Bw = 0 \tag{11.98}$$

for some matrices $A \geq 0, B \geq 0$. If we consider the equivalent form

$$w = \mu Aw + \mu Bw + w \tag{11.99}$$

for some $\mu > 0$, or

$$w = (I + \mu(A+B))^{-1} w \tag{11.100}$$

then the solution can be pursued by considering a fixed-point iteration of the form:

$$w_n = (I + \mu(A+B))^{-1} w_{n-1} \tag{11.101}$$

This recursion requires inverting a matrix consisting of the sum $A+B$. We would like to replace it by an alternative procedure that requires inverting the terms $(I+\mu A)$ and $(I+\mu B)$ separately. This can be achieved by introducing auxiliary variables as follows. From (11.99), the vector w is a fixed-point for the equation:

$$2w = (I+\mu A)w + (I+\mu B)w \tag{11.102}$$

We next introduce the variable:

$$z \;\stackrel{\Delta}{=}\; (I+\mu B)w \iff w = (I+\mu B)^{-1} z \tag{11.103}$$

and note from (11.102) that

$$2w - z = (I+\mu A)w \iff w = (I+\mu A)^{-1}(2w - z) \tag{11.104}$$

We therefore have two separate expressions in (11.103)–(11.104) involving the inverses $(I+\mu A)^{-1}$ and $(I+\mu B)^{-1}$. Now observe from the trivial equality:

$$z = z + w - w \tag{11.105}$$

that we can write using (11.103)–(11.104):

$$\begin{aligned} z &= z \;+\; (I+\mu A)^{-1}(2w - z) \;-\; (I+\mu B)^{-1} z \\ &= z \;+\; (I+\mu A)^{-1}\big(2(I+\mu B)^{-1} z - z\big) \;-\; (I+\mu B)^{-1} z \end{aligned} \tag{11.106}$$

The mapping on the right-hand side has a form similar to (11.82) if we make the identifications:

$$(I+\mu A)^{-1} \leftarrow (I + \mu \partial_{w^\mathsf{T}} E)^{-1} \;=\; \text{prox}_{\mu E}(\cdot) \tag{11.107}$$

$$(I+\mu B)^{-1} \leftarrow (I + \mu \partial_{w^\mathsf{T}} q)^{-1} \;=\; \text{prox}_{\mu q}(\cdot) \tag{11.108}$$

Indeed, the same argument can be repeated to arrive at (11.82) by noting that minimizers of (11.48) should satisfy, for any nonzero μ:

$$0 \in \mu \partial_{w^\mathsf{T}} E(w) + \mu \partial_{w^\mathsf{T}} q(w) \tag{11.109}$$

or, equivalently,

$$2w \in (I + \mu \partial_{w^\mathsf{T}} E)(w) + (I + \mu \partial_{w^\mathsf{T}} q)(w) \tag{11.110}$$

If we now introduce the variable:

$$z \;\stackrel{\Delta}{=}\; (I + \mu \partial_{w^\mathsf{T}} q)(w) \iff w = (I + \mu \partial_{w^\mathsf{T}} q)^{-1}(z) \;=\; \text{prox}_{\mu q}(z) \tag{11.111}$$

and note from (11.110) that

$$2w - z = (I + \mu\partial_{w^{\mathsf{T}}} E)(w) \iff w = (I + \mu\partial_{w^{\mathsf{T}}} E)^{-1}(2w - z) = \text{prox}_{\mu E}(2w - z) \tag{11.112}$$

then the same argument ends up leading to the mapping (11.82).

Discussions on, and variations of, the forward-backward proximal splitting technique (11.55) can be found in Combettes and Wajs (2005), Combettes and Pesquet (2011), and the many references therein, as well as in the articles by Lions and Mercier (1979), Passty (1979), Fukushima and Mine (1981), Guler (1991), Tseng (1991), Chen and Rockafellar (1997), Daubechies, Defrise, and De Mol (2004), Hale, Yin, and Zhang (2008), Bredies (2009), Beck and Teboulle (2009a,b), and Duchi and Singer (2009). The proximal Douglas–Rachford splitting algorithm (11.76) was introduced by Lions and Mercier (1979), who were motivated by the original work of Douglas and Rachford (1956) on a numerical discretized solution for the heat conduction problem. More discussion on this algorithm can be found in Eckstein and Bertsekas (1992), Combettes (2004), Combettes and Pesquet (2011), and O'Connor and Vandenberghe (2014), as well as in the lecture notes by Vandenderghe (2010).

PROBLEMS

11.1 Consider two convex functions $h(w)$ and $f(w)$ related by $h(w) = f(w) + c$, for some constant c. Show that $\text{prox}_{\mu h}(z) = \text{prox}_{\mu f}(z)$ for any $z \in \mathbb{R}^M$ and $\mu > 0$.

11.2 Let $h(w) : \mathbb{R}^M \to \mathbb{R}$ denote a convex function and introduce the transformation $g(w) = h(\alpha w + b)$, where $\alpha \neq 0$ and $b \in \mathbb{R}^M$. Show that

$$\text{prox}_g(z) \;=\; \frac{1}{\alpha}\Big(\text{prox}_{\alpha^2 h}(\alpha z + b) - b\Big)$$

11.3 Let $h(w) = \alpha\|w\|$. Show that

$$\text{prox}_{\mu h}(z) \;=\; \begin{cases} \left(1 - \frac{\mu\alpha}{\|z\|}\right) z, & \text{if } \|z\| \geq \mu\alpha \\ 0, & \text{otherwise} \end{cases}$$

11.4 Establish the validity of expression (11.30).

11.5 Show that the soft-thresholding function (11.18) satisfies the property $\mathbb{T}_{\rho\beta}(\rho x) = \rho\mathbb{T}_\beta(x)$, for any scalars $\rho > 0, \beta > 0$.

11.6 Let $h(w) : \mathbb{R}^M \to \mathbb{R}$ denote a convex function and introduce the transformation $g(w) = h(w) + \frac{\rho}{2}\|w\|^2$, where $\rho \neq 0$. Show that

$$\text{prox}_{\mu g}(z) \;=\; \text{prox}_{\frac{\mu h}{1+\mu\rho}}\left(\frac{z}{1+\mu\rho}\right)$$

Conclude that the proximal operator of $f(w) = \alpha\|w\|_1 + \frac{\rho}{2}\|w\|^2$ is given by

$$\text{prox}_{\mu f}(z) \;=\; \mathbb{T}_{\frac{\mu\alpha}{1+\mu\rho}}\left(\frac{z}{1+\mu\rho}\right)$$

11.7 Assume we select $h(w) = \|w\|$, where $w \in \mathbb{R}^M$. Show that

$$\text{prox}_{\mu h}(z) = \left(1 - \frac{\mu}{\|z\|}\right)_+ z, \qquad \mathcal{M}_{\mu h}(z) = \begin{cases} \frac{1}{2\mu}\|z\|^2, & \|z\| \leq \mu \\ \|z\| - \frac{\mu}{2}, & \|z\| > \mu \end{cases}$$

where $(x)_+ = \max\{x, 0\}$. Verify that when $M = 1$, the above expression for the proximal projection reduces to the soft-thresholding operation (11.18); the corresponding Moreau envelope will be the Huber function.

11.8 Let $h(w) = \alpha\|w\|_0$, where the notation $\|x\|_0$ counts the number of nonzero elements in vector x. Show that $\text{prox}_h(z) = z\,\mathbb{I}[\,|z| > \sqrt{2\alpha}\,]$. That is, all values of z larger in magnitude than $\sqrt{2\alpha}$ are retained otherwise they are set to zero. This function is sometimes referred to as the *hard-thresholding* mapping as opposed to soft-thresholding.

11.9 Let $w \in \mathbb{R}^M$ with entries $\{w(m)\}$ and M even. Consider the function

$$h(w) = |w(1) - w(2)| \;+\; |w(3) - w(4)| \;+\; \ldots \;+\; |w(M-1) - w(M)|$$

Introduce the $\frac{M}{2} \times M$ matrix (e.g., for $M = 10$)

$$D = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

Verify that $DD^{\mathsf{T}} = 2I_{M/2}$ and $h(w) = \|Dw\|_1$. Show that the proximal operator can be expressed in terms of the soft-thresholding operator as follows:

$$\text{prox}_{\mu h}(z) = z + \frac{1}{2\mu}D^{\mathsf{T}}\Big(\text{prox}_{2\mu^2\|w\|_1}(\mu Dz) - \mu Dz\Big)$$

Remark. For more information on this problem and the next, the reader may refer to Beck (2017, Ch. 6).

11.10 Let $w \in \mathbb{R}^M$ with entries $\{w(m)\}$ and M even. Consider the function

$$h(w) = |\sqrt{2}w(1) - 1| \;+\; |w(2) - w(3)| \;+\; |w(4) - w(5)| + \;\ldots\; + \;|w(M-2) - w(M-1)|$$

Introduce the $\frac{M}{2} \times M$ matrix D and basis vector $e_1 \in \mathbb{R}^{M/2}$ (e.g., for $M = 10$)

$$D = \begin{bmatrix} \sqrt{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \end{bmatrix}, \quad e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Verify that $DD^{\mathsf{T}} = 2I_{M/2}$ and $h(w) = \|Dw - e_1\|_1$. Show that the proximal operator can be expressed in terms of the soft-thresholding operator as follows:

$$\text{prox}_{\mu h}(z) = z + \frac{1}{2\mu}D^{\mathsf{T}}\Big(\text{prox}_{2\mu^2\|w\|_1}(\mu Dz - e_1) - \mu(Dz - e_1)\Big)$$

11.11 Establish the validity of expression (11.35).

11.12 Consider a matrix $W \in \mathbb{R}^{N\times M}$ and introduce the following matrix-based proximal function definition:

$$\text{prox}_{\mu}(Z) \;\stackrel{\Delta}{=}\; \underset{W\in\mathbb{R}^{N\times M}}{\text{argmin}} \left\{\alpha\,\|W\|_{\star} \;+\; \frac{1}{2}\|W - Z\|_{\text{F}}^2\right\}, \quad \alpha, \mu > 0$$

where $\|W\|_{\star}$ denotes the nuclear norm of W (sum of its nonzero singular values). Introduce the singular value decomposition $Z = U\Sigma V^{\mathsf{T}}$, and replace Σ by a new matrix Σ_α whose diagonal entries are computed from the nonzero singular values in Σ as follows:

$$[\Sigma_\alpha]_{kk} \;=\; \max\Big\{0, \Sigma_{kk} - \alpha\Big\}$$

That is, nonzero singular values in Σ larger than α are reduced by α, while nonzero singular values smaller than α are set to zero. Show that the proximal solution is given by

$$\widehat{W} = \text{prox}(Z) \;=\; U\Sigma_\alpha V^\mathsf{T} \;\stackrel{\Delta}{=}\; \mathbb{S}_\alpha(Z)$$

where we are also using the notation $\mathbb{S}_\alpha(Z)$ to refer to the singular value soft-thresholding operation defined above. *Remark.* This result appears in Cai, Candes, and Shen (2010), Mazumder, Hastie, and Tibshirani (2010), and Ma, Goldfarb and Chen (2011).

11.13 Consider a full rank matrix $H \in \mathbb{R}^{N\times M}$ with $N \geq M$ and a vector $d \in \mathbb{R}^M$. We introduce the least-squares problem

$$\widehat{w} \;=\; \underset{w\in\mathbb{R}^M}{\text{argmin}} \;\left\{\|w-\bar{w}\|^2 \;+\; \|d-Hw\|^2\right\}$$

where $\bar{w} \in \mathbb{R}^M$ is some given vector.

(a) Determine the solution $\widehat{w}$.

(b) Determine the proximal projection of $\bar{w}$ using $h(w) = \frac{1}{2}\|d-Hw\|^2$ and $\mu = 1$. Show that the result agrees with the solution to part (a).

11.14 Consider the proximal projection problem

$$\widehat{w} = \underset{w\in\mathbb{R}^M}{\text{argmin}} \;\left\{\frac{1}{2}w^\mathsf{T}Aw \;+\; \frac{1}{2}\|w-z\|^2\right\}$$

where $A \geq 0$. Show that the solution is given by $\widehat{w} = (I_M + A)^{-1}z$.

11.15 Consider the proximal projection problem

$$\underset{w^e\in\mathbb{R}^{M+1}}{\text{argmin}} \;\left\{\frac{1}{2}\left(w^e\right)^\mathsf{T}Aw^e \;+\; \frac{1}{2}\|w^e - z^e\|^2\right\}$$

where $A = \text{diag}\{0, \rho I_M\}$ with $\rho > 0$, $w^e = \text{col}\{-\theta, w\}$, and $z^e = \text{col}\{-\phi, z\}$. Both w^e and z^e are extended vectors of size $M+1$ each, and $\{\theta, \phi\}$ are scalars. Show that the solution is given by

$$\widehat{\theta} = \phi, \quad \widehat{w} = \frac{z}{1+\rho}$$

11.16 Consider the proximal projection problem

$$\underset{w^e\in\mathbb{R}^{M+1}}{\text{argmin}} \;\left\{\alpha\|Aw^e\|_1 \;+\; \frac{\rho}{2}\left(w^e\right)^\mathsf{T}Aw^e \;+\; \frac{1}{2}\|w^e - z^e\|^2\right\}$$

where $A = \text{diag}\{0, \rho I_M\}$ with $\rho > 0$, $w^e = \text{col}\{-\theta, w\}$, and $z^e = \text{col}\{-\phi, z\}$. Both w^e and z^e are extended vectors of size $M+1$ each, and $\{\theta, \phi\}$ are scalars. Show that the solution is given by

$$\widehat{\theta} = \phi, \quad \widehat{w} = \mathbb{T}_{\frac{\alpha}{1+\rho}}\left(\frac{z}{1+\rho}\right)$$

11.17 **(True or False)**. A firmly non-expansive operator is also non-expansive.

11.18 Refer to the fixed-point iteration (11.40) for a strictly contractive operator, $f(z)$. Show that $f(z)$ has a unique fixed point, z^o, and that $z_n \to z^o$ as $n \to \infty$.

11.19 Let $h(w) : \mathbb{R}^M \to \mathbb{R}$ denote a convex function. Establish the following properties for the proximal operator of $h(w)$, for any vectors $a, b \in \mathbb{R}^M$:

(a) $\|\text{prox}_{\mu h}(a) - \text{prox}_{\mu h}(b)\| \leq \|a-b\|$.

(b) $\|\text{prox}_{\mu h}(a) - \text{prox}_{\mu h}(b)\|^2 \leq (a-b)^\mathsf{T}\big(\text{prox}_{\mu h}(a) - \text{prox}_{\mu h}(b)\big)$.

(c) $\|\text{prox}_{\mu h}(a) - \text{prox}_{\mu h}(b)\|^2 + \|(a - \text{prox}_{\mu h}(a)) - (b - \text{prox}_{\mu h}(b))\|^2 \leq \|a-b\|^2$.

(d) $\|a-b\| = \|\text{prox}_{\mu h}(a) - \text{prox}_{\mu h}(b)\| \Longleftrightarrow a - b = \text{prox}_{\mu h}(a) - \text{prox}_{\mu h}(b)$.

Property (a) means that the proximal operator is non-expansive. Property (b) means that the operator is firmly non-expansive. Property (c) is equivalent to (b). Property (d) follows from (c).

11.20 Refer to the proximal iteration (11.44).

(a) Let w^o denote a fixed point for the proximal operator, i.e., $w^o = \text{prox}_{\mu h}(w^o)$. Show that $\|w^o - w_n\| \le \|w^o - w_{n-1}\|$.

(b) Let $a(n) = \|w^o - w_n\|$. Since the sequence $\{a(n)\}$ is bounded from below, conclude from the monotone convergence theorem that $a(n)$ converges to some limit value $\bar{a}$ as $n \to \infty$. Conclude further that $\|\text{prox}_{\mu h}(w^o) - \text{prox}_{\mu h}(w_{n-1})\|$ converges to the same value $\bar{a}$.

(c) Use the analogue of property (11.43) for $\mu q(w)$ and the result of part (b) to conclude that w_n converges to a fixed point of $\text{prox}_{\mu h}(w)$.

11.21 Refer to the Douglas–Rachford algorithm (11.76) and the mapping $R(z)$ defined by (11.82). The purpose of this problem is to establish that $R(z)$ is firmly non-expansive. For any z_1, z_2, introduce the variables:

$$w_1 = \text{prox}_{\mu q}(z_1), \qquad t_1 = \text{prox}_{\mu E}(2w_1 - z_1)$$
$$w_2 = \text{prox}_{\mu q}(z_2), \qquad t_2 = \text{prox}_{\mu E}(2w_2 - z_2)$$

(a) Use the fact that proximal operators are firmly non-expansive to conclude that

$$\|w_1 - w_2\|^2 \le (z_1 - z_2)^\mathsf{T}(w_1 - w_2)$$
$$\|t_1 - t_2\|^2 \le (2w_1 - z_1 - 2w_2 + z_2)^\mathsf{T}(t_1 - t_2)$$

(b) Verify that

$$(z_1 - z_2)^\mathsf{T}\left(R(z_1) - R(z_2)\right) \ge (z_1 - z_2)^\mathsf{T}(t_1 - w_1 + z_1 - t_2 + w_2 - z_2) + \|w_1 - w_2\|^2 - (z_1 - z_2)^\mathsf{T}(w_1 - w_2)$$

(c) Simplify the expression in part (b) to verify that

$$(z_1 - z_2)^\mathsf{T}\left(R(z_1) - R(z_2)\right) \ge \|R(z_1) - R(z_2)\|^2 + (2w_1 - z_1 - 2w_2 + z_2)^\mathsf{T}(t_1 - t_2) - \|t_1 - t_2\|^2$$

Conclude that $R(z)$ is firmly non-expansive.

11.22 Consider the following variation of the Douglas-Rachford algorithm, starting from any w_{-1} and z_{-1}:

$$\left\{\begin{array}{lcl} t_n & = & \text{prox}_{\mu E}(2w_{n-1} - z_{n-1}) \\ z_n & = & t_n + z_{n-1} - w_{n-1} \\ w_n & = & \text{prox}_{\mu q}(z_n) \end{array}\right.$$

Show that fixed points of the mapping from w_{n-1} to w_n are minimizers of the aggregate cost in (11.48). Show further that the mapping from w_{n-1} to w_n is firmly non-expansive.

11.23 Consider the following variation of the Douglas-Rachford algorithm:

$$\left\{\begin{array}{lcl} w_n & = & \text{prox}_{\mu q}(z_{n-1}) \\ t_n & = & \text{prox}_{\mu E}(2w_n - z_{n-1}) \\ z_n & = & (1-\rho)z_{n-1} + \rho(t_n - w_n + z_{n-1}) \end{array}\right.$$

where $0 < \rho < 2$ is called a relaxation parameter. Comparing with (11.76), we see that the last step is now a linear combination of z_{n-1} with the original quantity $t_n - w_n + z_{n-1}$ with the combination coefficients adding up to one. Show that w_n converges to a minimizer of (11.48).

11.24 Establish the validity of decomposition (11.88a)–(11.88b). More generally, show that

$$z = \text{prox}_{\mu h}(z) + \mu\, \text{prox}_{\frac{1}{\mu}h^\star}(z/\mu)$$

11.25 Refer to the definition of the Moreau envelope in (11.85). Is the Moreau envelope a convex function over z? Show that $\mathcal{M}_{\mu h}(z)$ is differentiable with respect to z and that its gradient vector is given by expression (11.91).

11.26 Consider a convex function $h(w)$ and its Fenchel conjugate $h^\star(x)$ as defined by (8.83). Show that $\text{prox}_h(z) \;=\; \nabla_{z^\mathsf{T}}\, \mathcal{M}_{h^\star}(z)$.

11.27 Consider a convex function $h(w)$ and its Fenchel conjugate $h^\star(x)$ as defined by (8.83). Show that the Moreau envelope satisfies:

$$\mathcal{M}_{\mu h}(z) \;=\; \Big(h^\star(w) + \frac{\mu}{2}\|w\|^2\Big)^\star$$

That is, the Moreau envelope is obtained by perturbing the Fenchel conjugate of $h(w)$ by a quadratic term and then computing the Fenchel conjugate of the result.

11.28 Using the Bregman divergence, assume we extend the definition of the proximal operator by replacing the quadratic measure in the original definition (11.4) by

$$\text{prox}_{\mu h}(z) \;\overset{\Delta}{=}\; \underset{w\in \mathbb{R}^M}{\text{argmin}} \left\{ h(w) \;+\; \frac{1}{\mu} D_\phi(w,z)\right\}$$

(a) Verify that the optimality condition (11.13) is replaced by

$$a = \text{prox}_{\mu h}(b) \quad\Longleftrightarrow\quad \Big(\nabla_{w^\mathsf{T}}\phi(a) - \nabla_{w^\mathsf{T}}\phi(b)\Big) \in \partial_{w^\mathsf{T}}\, h(a)$$

(b) Let $\mathcal{C}$ denote some closed convex set and introduce its indicator function $\mathbb{I}_{C,\infty}[w]$, which is equal to zero when $w \in \mathcal{C}$ and $+\infty$ otherwise. Verify that

$$\text{prox}_{\mathbb{I}_C}(z) \;=\; \underset{w\in\mathcal{C}}{\text{argmin}}\;\; D_\phi(w,z)$$

which amounts to finding the projection of z onto $\mathcal{C}$ using the Bregman measure.

11.A CONVERGENCE UNDER CONVEXITY

The convergence analyses in the two appendices to this chapter benefit from the presentations in Polyak (1987), Combettes and Wajs (2005), Combettes and Pesquet (2011), Polson, Scott, and Willard (2015), and Beck (2017). In this first appendix, we establish the statement of Theorem 11.1, which relates to the convergence of the proximal gradient algorithm under convexity of the smooth component, $E(w)$. In preparation for the argument we establish three useful facts. First, we rewrite algorithm (11.53) in the form:

$$w_n = w_{n-1} - \mu g_\mu(w_{n-1}) \tag{11.113}$$

where

$$g_\mu(w) \;\overset{\Delta}{=}\; \frac{1}{\mu}\Big(w - \text{prox}_{\mu q}(w - \mu\,\nabla_{w^\mathsf{T}} E(w))\Big) \tag{11.114}$$

Form (11.113) shows that the proximal gradient algorithm adjusts w_{n-1} by adding a correction along the direction of $-g_\mu(w_{n-1})$; the size of the correction is modulated by μ. Observe in particular that evaluating $g_\mu(w)$ at a minimizer w^o for $P(w)$ we get

$$g_\mu(w^o) = 0 \tag{11.115}$$

This is because w^o is a fixed point for the proximal operator by (11.51a).

Second, we have from (11.114) that

$$w - \mu\, g_\mu(w) \;=\; \text{prox}_{\mu q}\Big(w - \mu\, \nabla_{w^\mathsf{T}} E(w)\Big) \tag{11.116}$$

Therefore, using (11.13) with the identifications

$$a \leftarrow w - \mu\, g_\mu(w), \quad b \leftarrow w - \mu\, \nabla_{w^\mathsf{T}} E(w) \tag{11.117}$$

we find that

$$g_\mu(w) - \nabla_{w^\mathsf{T}} E(w) \;\in \partial_{w^\mathsf{T}}\, q(w - \mu g_\mu(w)) \tag{11.118}$$

Third, the smooth component $E(w) : \mathbb{R}^M \to \mathbb{R}$ is assumed to be convex differentiable with $\delta-$Lipschitz gradients, i.e.,

$$\|\nabla_w\, E(a) - \nabla_w\, E(b)\| \;\leq\; \delta\, \|a - b\| \tag{11.119}$$

for any $a, b \in \text{dom}(E)$ and some $\delta \geq 0$. It then follows from property (10.13) that

$$E(a) \leq E(b) + \nabla_w\, E(b)\,(a - b) + \frac{\delta}{2}\|a - b\|^2 \tag{11.120}$$

We are now ready to establish Theorem 11.1.

Proof of Theorem 11.1: The argument involves several steps:

(**step 1**) We use property (11.120) and select $a \leftarrow w_n$ and $b \leftarrow w_{n-1}$ to write

$$\begin{aligned}
E(w_n) \quad &\leq \quad E(w_{n-1}) + \nabla_w\, E(w_{n-1})(w_n - w_{n-1}) + \frac{\delta}{2}\|w_n - w_{n-1}\|^2 \\
&\overset{(11.113)}{=} \quad E(w_{n-1}) - \mu \nabla_w\, E(w_{n-1})\, g_\mu(w_{n-1}) + \frac{\mu^2 \delta}{2}\|g_\mu(w_{n-1})\|^2 \\
&\overset{(11.70)}{\leq} \quad E(w_{n-1}) - \mu \nabla_w\, E(w_{n-1})\, g_\mu(w_{n-1}) + \frac{\mu}{2}\|g_\mu(w_{n-1})\|^2
\end{aligned} \tag{11.121}$$

(**step 2**) We use the inequality from the first step to establish that, for any $z \in \text{dom}(h)$,

$$P(w_n) \leq P(z) + (g_\mu(w_{n-1}))^\mathsf{T}\,(w_{n-1} - z) - \frac{\mu}{2}\|g_\mu(w_{n-1})\|^2 \tag{11.122}$$

Indeed, note that

$$\begin{aligned}
P(w_n) \quad &\overset{\Delta}{=} \quad q(w_n) + E(w_n) \\
&\overset{(11.121)}{\leq} \quad q(w_n) + E(w_{n-1}) - \mu \nabla_w\, E(w_{n-1})\, g_\mu(w_{n-1}) + \frac{\mu}{2}\|g_\mu(w_{n-1})\|^2 \\
&\overset{(a)}{\leq} \quad q(z) + (s(w_n))^\mathsf{T}(w_n - z) + E(z) + \nabla_w\, E(w_{n-1})(w_{n-1} - z) - \\
&\qquad\qquad \mu \nabla_w E(w_{n-1}) g_\mu(w_{n-1}) + \frac{\mu}{2}\|g_\mu(w_{n-1})\|^2
\end{aligned} \tag{11.123}$$

where in step (a) we used the convexity of $q(w)$ and $E(w)$ and properties (8.4) and (8.43). Moreover, the notation $s(w_n)$ refers to a subgradient of $q(w)$ relative to w^T:

$$s(w_n) \;\in \partial_{w^\mathsf{T}}\, q(w_n) \tag{11.124}$$

Now, appealing to (11.118) and letting $w = w_{n-1}$ we find that

$$g_\mu(w_{n-1}) - \nabla_{w^\mathsf{T}} E(w_{n-1}) \;\in \partial_{w^\mathsf{T}} q(w_n) \tag{11.125}$$

where the argument of $q(\cdot)$ becomes w_n under (11.113). This relation implies that we can select the subgradient $s(w_n)$ as the difference on the left so that

$$\nabla_{w^{\mathsf{T}}} E(w_{n-1}) = g_\mu(w_{n-1}) - s(w_n) \tag{11.126}$$

We can now evaluate three terms appearing (11.123) as follows:

$$\begin{aligned}(s(w_n))^{\mathsf{T}}(w_n - z) &= (s(w_n))^{\mathsf{T}}(w_{n-1} - \mu g_\mu(w_{n-1}) - z) \qquad (11.127\text{a})\\ &= (s(w_n))^{\mathsf{T}}(w_{n-1} - z) \;-\; \mu(s(w_n))^{\mathsf{T}} g_\mu(w_{n-1})\end{aligned}$$

and

$$\begin{aligned}\nabla_w\, E(w_{n-1})(w_{n-1} - z) &= \Big(g_\mu(w_{n-1}) - s(w_n)\Big)^{\mathsf{T}}(w_{n-1} - z) \qquad (11.127\text{b})\\ &= \Big(g_\mu(w_{n-1})\Big)^{\mathsf{T}}(w_{n-1} - z) \;-\; (s(w_n))^{\mathsf{T}}(w_{n-1} - z)\end{aligned}$$

and

$$\begin{aligned}-\mu\nabla_w E(w_{n-1}) g_\mu(w_{n-1}) &= -\mu\Big((g_\mu(w_{n-1} - s(w_n)\Big)^{\mathsf{T}} g_\mu(w_{n-1}) \qquad (11.127\text{c})\\ &= -\mu\|g_\mu(w_{n-1}\|^2 \;+\; \mu(s(w_n))^{\mathsf{T}} g_\mu(w_{n-1})\end{aligned}$$

Substituting into (11.123) and simplifying gives

$$\begin{aligned}P(w_n) &\le q(z) + E(z) + (g_\mu(w_{n-1}))^{\mathsf{T}}(w_{n-1} - z) - \frac{\mu}{2}\|g_\mu(w_{n-1})\|^2\\ &= P(z) + (g_\mu(w_{n-1}))^{\mathsf{T}}(w_{n-1} - z) - \frac{\mu}{2}\|g_\mu(w_{n-1})\|^2 \qquad (11.128)\end{aligned}$$

(**step 3**) Using $z = w_{n-1}$ in the last inequality we get

$$P(w_n) \le P(w_{n-1}) - \frac{\mu}{2}\|g_\mu(w_{n-1})\|^2 \tag{11.129}$$

which shows that $P(w_n)$ is a non-increasing sequence. If we use instead $z = w^o$ in (11.128) we get

$$\begin{aligned}0 \le P(w_n) - P(w^o) \quad &\le \quad (g_\mu(w_{n-1}))^{\mathsf{T}}(w_{n-1} - w^o) - \frac{\mu}{2}\|g_\mu(w_{n-1})\|^2\\ &\overset{(a)}{=} \quad \frac{1}{2\mu}\Big(\|w_{n-1} - w^o\|^2 - \|w_{n-1} - w^o - \mu g_\mu(w_{n-1})\|^2\Big)\\ &\overset{(11.113)}{\le} \quad \frac{1}{2\mu}(\|w_{n-1} - w^o\|^2 - \|w_n - w^o\|^2) \qquad (11.130)\end{aligned}$$

where step (a) follows by expanding the terms in the second line and noting that they coincide with those in the first line. It follows that

$$\|w^o - w_n\| \le \|w^o - w_{n-1}\| \tag{11.131}$$

so that $\|w^o - w_n\|$ is a non-increasing sequence.

(**step 4**) Adding from $n = 0$ up to some $N - 1$ gives

$$\begin{aligned}\sum_{n=0}^{N-1}(P(w_n) - P(w^o)) &\le \frac{1}{2\mu}\sum_{n=0}^{N-1}\left(\|w^o - w_{n-1}\|^2 - \|w^o - w_n\|^2\right)\\ &= \frac{1}{2\mu}(\|w^o - w_{-1}\|^2 - \|w^o - w_{N-1}\|^2)\\ &\le \frac{1}{2\mu}\|w^o - w_{-1}\|^2 \qquad (11.132)\end{aligned}$$

Now since $P(w_n)$ is non-increasing we get

$$P(w_{N-1}) - P(w^o) \leq \frac{1}{N}\sum_{n=0}^{N-1}(P(w_n) - P(w^o)) \leq \frac{1}{2\mu N}\|w^o - w_{-1}\|^2 \qquad (11.133)$$

11.B CONVERGENCE UNDER STRONG CONVEXITY

In this appendix, we establish the statement of Theorem 11.2, which relates to the convergence of the proximal gradient algorithm under strong convexity of the smooth component, $E(w)$. Theorem 11.1 shows that under a *convexity* condition on $E(w)$, the cost value $P(w_n)$ approaches the minimum $P(w^o)$ at the rate $O(1/n)$. Faster convergence at an exponential rate is possible when $E(w)$ is $\nu-$strongly convex.

Proof of Theorem 11.2: From the proximal gradient recursion (11.53) and the fixed-point relation (11.51a) we have

$$w_n = \text{prox}_{\mu q}\Big(w_{n-1} - \mu\nabla_{w^\mathsf{T}} E(w_{n-1})\Big) \qquad (11.134a)$$

$$w^o = \text{prox}_{\mu q}\Big(w^o - \mu\nabla_{w^\mathsf{T}} E(w^o)\Big) \qquad (11.134b)$$

Applying the non-expansive property (11.41) of the proximal operator we get

$$\begin{aligned}
&\|w^o - w_n\|^2 \\
&= \left\|\text{prox}_{\mu q}\left(w^o - \mu\nabla_{w^\mathsf{T}} E(w^o)\right) - \text{prox}_{\mu q}\left(w_{n-1} - \mu\nabla_{w^\mathsf{T}} E(w_{n-1})\right)\right\|^2 \\
&\overset{(11.41)}{\leq} \|w^o - \mu\nabla_{w^\mathsf{T}} E(w^o) - w_{n-1} + \mu\nabla_{w^\mathsf{T}} E(w_{n-1})\|^2 \\
&= \|w^o - w_{n-1}\|^2 - \\
&\quad 2\mu\left(\nabla_{w^\mathsf{T}} E(w^o) - \nabla_{w^\mathsf{T}} E(w_{n-1})\right)^\mathsf{T}(w^o - w_{n-1}) + \\
&\quad \mu^2\|\nabla_{w^\mathsf{T}} E(w^o) - \nabla_{w^\mathsf{T}} E(w_{n-1})\|^2 \\
&\overset{(a)}{\leq} \|w^o - w_{n-1}\|^2 - 2\mu\nu\|w^o - w_{n-1}\|^2 + \mu^2\delta^2\|w^o - w_{n-1}\|^2 \\
&= (1 - 2\mu\nu + \mu^2\delta^2)\|w^o - w_{n-1}\|^2 \\
&\triangleq \lambda^{n+1}\|w^o - w_{-1}\|^2
\end{aligned} \qquad (11.135)$$

where λ is defined by (11.75) and step (a) is because of the Lipschitz condition (11.119) and the assumed strong-convexity of $E(w)$, which implies in view of property (8.24) the relation:

$$\Big(\nabla_{w^\mathsf{T}} E(w^o) - \nabla_{w^\mathsf{T}} E(w_{n-1})\Big)^\mathsf{T}(w^o - w_{n-1}) \geq \nu\|w^o - w_{n-1}\|^2 \qquad (11.136)$$

We conclude from (11.135) that the squared error vector converges exponentially fast to zero at a rate dictated by λ. To verify that condition (11.73) ensures $0 \leq \lambda < 1$, we refer to Figure 11.5 where we plot the coefficient $\lambda(\mu)$ as a function of μ. The minimum value of $\lambda(\mu)$, which occurs at the location $\mu = \nu/\delta^2$ and is equal to $1 - \nu^2/\delta^2$, is nonnegative since $0 < \nu \leq \delta$. It is clear from the figure that $0 \leq \lambda < 1$ for $\mu \in (0, \frac{2\nu}{\delta^2})$.

We next establish the exponential convergence of the excess cost value as shown by (11.74). For this purpose, we first note from the convexity of $P(w)$ and property (8.43) for non-smooth convex functions that

$$P(w_n) - P(w^o) \leq (s(w_n))^\mathsf{T}(w_n - w^o) \qquad (11.137)$$

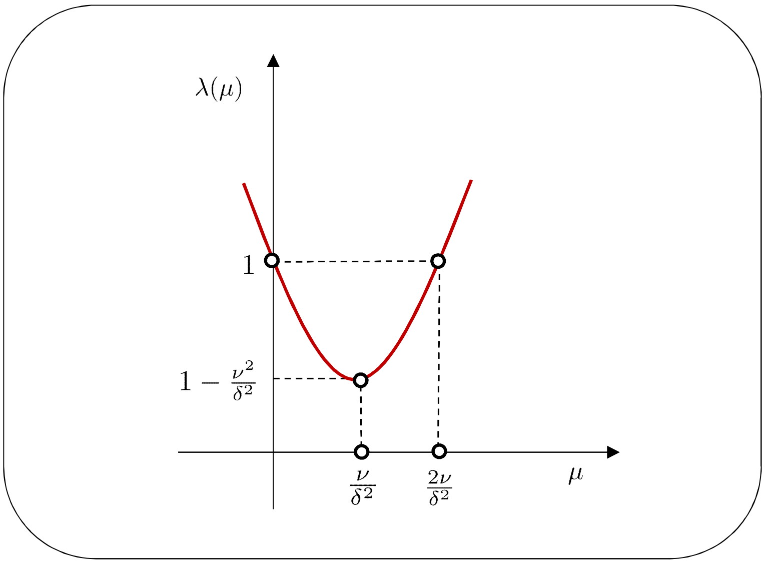

Figure 11.5 Plot of the function $\lambda(\mu) = 1 - 2\nu\mu + \mu^2\delta^2$ given by (11.75). It shows that the function $\lambda(\mu)$ assumes values below one in the range $0 < \mu < 2\nu/\delta^2$.

where $s(w_n)$ denotes a subgradient vector for $P(w)$ relative to w^{T} at location w_n. Using (11.54), and the fact that $P(w) = q(w) + E(w)$, we know that this subgradient vector can be chosen as:

$$s(w_n) = \frac{1}{\mu}(z_n - w_n) \; + \; \nabla_{w^{\mathsf{T}}} E(w_n) \tag{11.138}$$

Substituting into (11.137) gives

$$\begin{aligned}
&P(w_n) - P(w^o) \\
&\leq \left(\frac{1}{\mu}(z_n - w_n) \; + \; \nabla_{w^{\mathsf{T}}} E(w_n)\right)^{\mathsf{T}} (w_n - w^o) \\
&= \left(\frac{1}{\mu}(z_n - \text{prox}_{\mu q}(z_n)) \; + \; \nabla_{w^{\mathsf{T}}} E(w_n)\right)^{\mathsf{T}} (w_n - w^o) \\
&= \left(\frac{1}{\mu}(z_n - \text{prox}_{\mu q}(z_n)) + \nabla_{w^{\mathsf{T}}} E(w^o) + \nabla_{w^{\mathsf{T}}} E(w_n) - \nabla_{w^{\mathsf{T}}} E(w^o)\right)^{\mathsf{T}} (w_n - w^o)
\end{aligned} \tag{11.139}$$

so that using (11.119):

$$\begin{aligned}
&P(w_n) - P(w^o) \\
&\leq \left(\frac{1}{\mu}(z_n - \text{prox}_{\mu q}(z_n)) + \nabla_{w^{\mathsf{T}}} E(w^o)\right)^{\mathsf{T}} (w_n - w^o) + \delta\|w_n - w^o\|^2 \\
&\leq \left\|\frac{1}{\mu}(z_n - \text{prox}_{\mu q}(z_n)) + \nabla_{w^{\mathsf{T}}} E(w^o)\right\| \|w_n - w^o\| + \delta\|w_n - w^o\|^2
\end{aligned} \tag{11.140}$$

To continue, we substitute $\nabla_{w^{\mathsf{T}}} E(w^o)$ by an equivalent expression as follows. We know from the first part of the proof that w_n converges to w^o, which satisfies the fixed-point relation (11.134b). Let

$$\begin{aligned} z^o &\triangleq w^o - \mu \nabla_{w^\mathsf{T}} E(w^o) && \text{(11.141a)} \\ w^o &= \text{prox}_{\mu q}(z^o) && \text{(11.141b)} \end{aligned}$$

Combining these two relations we get

$$\nabla_{w^\mathsf{T}} E(w^o) = -\frac{1}{\mu}(z^o - w^o) \; = \; -\frac{1}{\mu}\left(z^o - \text{prox}_{\mu q}(z^o)\right) \tag{11.142}$$

Substituting into (11.140) we obtain

$$\begin{aligned} & P(w_n) - P(w^o) \\ & \leq \frac{1}{\mu}\Big\| z_n - \text{prox}_{\mu q}(z_n) - z^o + \text{prox}_{\mu q}(z^o)\Big\| \, \|w_n - w^o\| + \delta\|w_n - w^o\|^2 \\ & \overset{(11.41)}{\leq} \frac{2}{\mu}\|z_n - z^o\| \, \|w_n - w^o\| + \delta\|w_n - w^o\|^2 \\ & \overset{(a)}{\leq} \frac{2}{\mu}\|w_{n-1} - w^o\| \, \|w_n - w^o\| + 2\|\nabla_{w^\mathsf{T}} E(w_{n-1}) - \nabla_{w^\mathsf{T}} E(w^o)\| \, \|w_n - w^o\| + \\ & \qquad \delta\|w_n - w^o\|^2 \\ & \overset{(11.119)}{\leq} \frac{2}{\mu}\|w_{n-1} - w^o\| \, \|w_n - w^o\| + 2\delta\|w_{n-1} - w^o\| \, \|w_n - w^o\| + \delta\|w_n - w^o\|^2 \\ & \overset{(11.135)}{\leq} \frac{2\sqrt{\lambda}}{\mu}\|w_{n-1} - w^o\|^2 + 2\sqrt{\lambda}\delta\|w_{n-1} - w^o\|^2 + \lambda\delta\|w_{n-1} - w^o\|^2 \\ & = \underbrace{\left(\frac{2\,(1+\mu\delta)}{\mu\sqrt{\lambda}} + \delta\right)\lambda}_{\triangleq\, \beta} \times \, \|w_{n-1} - w^o\|^2 \\ & \overset{(b)}{=} \beta\|w_{n-1} - w^o\|^2 \\ & \overset{(11.135)}{\leq} \beta\lambda^n\|w^o - w_{-1}\|^2 \end{aligned} \tag{11.143}$$

where in step (a) we used the relation

$$z_n - z^o = (w_{n-1} - \mu\nabla_{w^\mathsf{T}} E(w_{n-1})) \; - \; (w^o - \mu\nabla_{w^\mathsf{T}} E(w^o)) \tag{11.144}$$

and in step (b) we introduced the scalar β.

■

Example 11.8 **(A more relaxed bound on μ)** We revisit Remark 11.4 and explain that the bound on μ for convergence can be relaxed to $\mu < 2/\delta$. For this purpose, we exploit the *co-coercivity* property of convex functions with $\delta-$Lipschitz gradients. We know from Prob. 10.4 that:

$$\Big(\nabla_{w^\mathsf{T}} E(w_2) - \nabla_{w^\mathsf{T}} E(w_1)\Big)^\mathsf{T}(w_2 - w_1) \geq \frac{1}{\delta}\,\|\nabla_w E(w_2) - \nabla_w E(w_1)\|^2 \tag{11.145}$$

We use this inequality in the third line of (11.135) as follows:

$$\begin{aligned}
\|\widetilde{w}_n\|^2 &\overset{(11.145)}{\leq} \|\widetilde{w}_{n-1}\|^2 - 2\mu\Big(\nabla_{w^\mathsf{T}} E(w^o) - \nabla_{w^\mathsf{T}} E(w_{n-1})\Big)^\mathsf{T} \widetilde{w}_{n-1} + \\
&\qquad\qquad + \mu^2\delta\Big(\nabla_{w^\mathsf{T}} E(w^o) - \nabla_{w^\mathsf{T}} E(w_{n-1})\Big)^\mathsf{T} \widetilde{w}_{n-1} \\
&= \|\widetilde{w}_{n-1}\|^2 - (2\mu - \mu^2\delta)\Big(\nabla_{w^\mathsf{T}} E(w^o) - \nabla_{w^\mathsf{T}} E(w_{n-1})\Big)^\mathsf{T} \widetilde{w}_{n-1} \\
&\overset{(11.136)}{\leq} \|\widetilde{w}_{n-1}\|^2 - (2\mu - \mu^2\delta)\nu\|\widetilde{w}_{n-1}\|^2 \\
&= \underbrace{(1 - 2\mu\nu + \mu^2\nu\delta)}_{\triangleq\,\lambda'} \|\widetilde{w}_{n-1}\|^2 \\
&= (\lambda')^{n+1}\|w^o - w_{-1}\|^2 \qquad\qquad (11.146)
\end{aligned}$$

This result is consistent with (11.135) since $\lambda' \leq \lambda$ in view of $\nu \leq \delta$. Working with λ' instead of λ and repeating the argument leading to (11.143), we will arrive at the bound $0 < \mu < 2/\delta$ for stability with convergence occurring at $O((\lambda')^n)$.

REFERENCES

Bach, F., R. Jenatton, J. Mairal, and G. Obozinski (2012), "Optimization with sparsity-inducing penalties," *Foundations and Trends in Machine Learning*, vol. 4, no. 1, pp. 1–106.

Bauschke, H. H. and P. L. Combettes (2011), *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, NY.

Beck, A. (2017), *First-Order Methods in Optimization*, SIAM, PA.

Beck, A. and M. Teboulle (2009a), "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202.

Beck, A. and M. Teboulle (2009b), "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2419–2434.

Beck, A. and M. Teboulle (2012), "Smoothing and first order methods: A unified framework," *SIAM J. Optim.*, vol. 22, no. 2, pp. 557–580.

Bredies, K. (2009), "A forward-backward splitting algorithm for the minimization of non-smooth convex functionals in Banach space," *Inverse Problems*, vol. 25, Art. 015005.

Brezis, H. (1973), *Opérateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert*, North-Holland, Amsterdam.

Browder, F. (1965), "Nonlinear monotone operators and convex sets in Banach spaces," *Bull. Amer. Math. Soc.*, vol. 71, no. 5, pp. 780–785.

Browder, F. (1967), "Convergence theorems for sequences of nonlinear operators in Banach spaces," *Mathematische Zeitschrift*, vol. 100, no. 3, pp. 201–225.

Bruck, R. E. and S. Reich (1977), "Nonexpansive projections and resolvents of accretive operators in Banach spaces," *Houston J. Math.*, vol. 3, pp. 459–470.

Cai, J., E. J. Candes, and Z. Shen (2010), "A singular value thresholding algorithm for matrix completion," SIAM J. Optimization, vol. 20, no. 4, pp. 1956–1982.

Chen, G. H. G. and R. T. Rockafellar (1997), "Convergence rates in forward-backward splitting," *SIAM J. Optim.* vol. 7, pp. 421–444.

Combettes, P. L. (2004), "Solving monotone inclusions via compositions of nonexpansive averaged operators," *Optimization*, vol. 53, no. 5–6, 2004.

Combettes, P. L. and J.-C. Pesquet (2011), "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, H. H. Bauschke *et al.*, *Eds.*, pp. 185-212. Springer, NY.

Combettes, P. L. and V. R. Wajs (2005), "Signal recovery by proximal forward-backward splitting," *Multiscale Model. Simul.*, vol. 4, no. 4., pp. 1168-1200.

Daubechies, I., M. Defrise, and C. De Mol (2004), "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. LVII, pp. 1413–1457.

Donoho, D. L. and I. M. Johnstone (1994), "Ideal spatial adaptation via wavelet shrinkage," *Biomefrika*, vol. 81, pp. 425–455.

Donoho, D. L. and I. M. Johnstone (1995), "Adapting to unknown smoothness via wavelet shrinkage," *J. American Stat. Assoc.*, vol. 90, pp. 1200–1224.

Douglas, J. and H. H. Rachford (1956), "On the numerical solution of heat conduction problems in two and three space variables," *Trans. Amer. Math. Soc.*, vol. 82, pp. 421–439.

Duchi, J. and Y. Singer (2009), "Efficient online and batch learning using forward backward splitting,' *J. Mach. Learn. Res.*, vol. 10, pp. 2873–2908.

Eckstein, J. and D. P. Bertsekas (1992), "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Programming*, vol. 55, pp. 293–318.

Figueiredo, M., J. Bioucas-Dias, and R. Nowak (2007), "Majorization-minimization algorithms for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2980–2991.

Fukushima, M. and H. Mine (1981), "A generalized proximal point algorithm for certain non-convex minimization problems," *Inter. J. Systems Sci.*, vol. 12, no. 8, pp. 989–1000.

Granas, A. and J. Dugundji (2003), *Fixed Point Theory*, Springer, NY.

Guler, O. (1991), "On the convergence of the proximal point algorithm for convex minimization," *SIAM J. Control Optim.*, vol. 20, pp. 403–419.

Hale, E. T., M. Yin, and Y. Zhang (2008), "Fixed-point continuation for ℓ_1-minimization: Methodology and convergence," *SIAM J. Optim.*, vol. 19, pp. 1107–1130.

Huber, P. J. (1981), *Robust Statistics*, Wiley, NY.

Lemaire, B. (1989a), "The proximal algorithm," *Intern. Series Numer. Math.*, pp. 73–87.

Lemaire, B. (1989b), "New methods in optimization and their industrial uses," *Inter. Ser. Numer. Math.*, vol. 87, pp. 73–87.

Lions, P. and B. Mercier (1979), "Splitting algorithms for the sum of two nonlinear operators," *SIAM J. Numer. Anal.*, vol. 16, pp. 964–979.

Luo, Z. Q. and P. Tseng (1992b), "On the linear convergence of descent methods for convex essentially smooth minimization," *SIAM J. Control and Optimization*, vol. 30, pp. 408–425.

Luo, Z. Q. and P. Tseng (1993), "Error bounds and convergence analysis of feasible descent methods: A general approach," *Annals of Operations Research*, vol. 46, pp. 157–178.

Ma, S., D. Goldfarb, and L. Chen (2011), "Fixed point and Bregman iterative methods for matrix rank minimization," *Mathematical Programming*, vol. 128, pp. 321–353.

Martinet, B. (1970), "Régularisation d'inéquations variationnelles par approximations successives," *Rev. Francaise Informat. Recherche Opérationnelle*, vol. 4, no. 3, pp. 154–158.

Martinet, B. (1972), "Determination approchtfe d'un point fixe d'une application pseudo-contractante," *Comptes Rendus de l'Académie des Sciences de Paris*, vol. 274, pp. 163–165.

Mazumder, R., T. Hastie, and R. Tibshirani (2010), "Spectral regularization algorithms for learning large incomplete matrices," *J. Machine Learning Research*, vol. 11, pp. 2287–2322.

Mercier, B. (1979), *Topics in Finite Element Solution of Elliptic Problems*, Lectures on Mathematics, Tata Institute of Fundamental Research, Bombay, India.

Minty, G. J. (1962), "Monotone (nonlinear) operators in Hilbert space," *Duke Math. Journal*, vol. 29, pp. 341–346.

Moreau, J. J. (1962), "Fonctions convexes duales et points proximaux dans un espace hilbertien," *Comptes Rendus de l'Académie des Sciences de Paris*, vol. A255, pp. 2897–2899.

Moreau, J. J. (1963a), "Propriétés des applications prox," *Comptes Rendus de l'Académie des Sciences de Paris*, vol. A256, pp. 1069–1071.

Moreau, J. J. (1963b), "Fonctionnelles sous-différentiables," *Comptes Rendus de l'Académie des Sciences de Paris*, vol. A257, pp. 4117–4119.

Moreau, J. J. (1965), "Proximité et dualité dans un espace hilbertien," *Bull. Soc. Math. de France*, vol. 93, pp. 273–299.

Nesterov, Y. (2004), *Introductory Lectures on Convex Optimization*, Springer, NY.

Nesterov, Y. (2005), "Smooth minimization of non-smooth functions," *Math. Programming*, vol. 103, no. 1, pp. 127–152.

O'Connor, D. and L. Vandenberghe (2014), "Primal-dual decomposition by operator splitting and applications to image deblurring," *SIAM J. Imag. Sci.*, vol. 7, no. 3, pp. 1724–1754.

Parikh, N. and S. Boyd (2013), "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239.

Passty, G. (1979), "Ergodic convergence to a zero of the sum of monotone operators in Hilbert space," *J. Math. Anal. Appl.*, vol. 72, no. 2, pp. 383–390.

Polson, N. G., J. G. Scott, and B. T. Willard (2015), "Proximal algorithms in statistics and machine learning," *Statistical Science*, vol. 30, np. 4, pp. 559–581.

Polyak, B. T. (1987), *Introduction to Optimization*, Optimization Software, NY.

Rockafellar, R. T. (1976a), "Monotone operators and the proximal point algorithm," *SIAM J. Control Opt.*, vol. 14, no. 5, pp. 877–898.

Rockafellar, R. T. (1976b), "Augmented Lagrangians and applications of the proximal point algorithm in convex programming," *Math. Oper. Res.*, vol. 1, no. 2, pp. 97–116.

Rockafellar, R. T. and R. Wets (1998), *Variational Analysis*, Springer.

Sibony, M. (1970), "Méthodes itératives pour les équations et inéquations aux dérivées partielles non linéaires de type monotone," *Calcolo*, vol. 7, pp. 65–183.

Tseng, P. (1991), "Applications of a splitting algorithm to decomposition in convex programming and variational inequalities," *SIAM J. Control Optim.*, vol. 29, pp. 119–138.

Yosida, K. (1968), *Functional Analysis*, Springer, NY.