Chapter 5

Proximal gradient techniques

In this chapter we discuss how the ideas developed in the previous chapter regarding unconstrained minimization via gradient descent of a convex function f with Lipschitz continuous gradient, easily extend to compositive problems of the form

$$\underset{\mathbf{x}}{\text{minimize}}\ f\left(\mathbf{x}\right)+g\left(\mathbf{x}\right), \tag{5.1}$$

where g is convex, but not necessarily smooth or even bounded. These simple extensions of gradient descent are referred to as *proximal gradient* methods.

For example, of particular interest to us is the choice $g\left(\mathbf{x}\right)=\lambda\left\Vert \mathbf{x}\right\Vert_1$, in which case (5.1) is the prototypical sparse recovery problem (e.g., the Lasso when $f\left(\mathbf{x}\right)=\left\Vert \mathbf{Ax}-\mathbf{b}\right\Vert_2^2$, sparse logistic regression when f is the logistic loss, etc.). More generally, we will see that many problems discussed so far take the form in (5.1) for some g, making proximal gradient techniques, and their associated accelerated gradient version, extremely valuable tools for solving sparse and low-rank problems.

5.1 Projections

We begin by discussing a subclass of problem (5.1) of the form

$$\underset{\mathbf{x}}{\text{minimize}}\ \ \left\Vert \mathbf{x}-\mathbf{y}\right\Vert_2^2+I_{\mathcal{S}}\left(\mathbf{x}\right), \tag{5.2}$$

where $I_{\mathcal{S}}\left(\mathbf{x}\right)=\begin{cases}0 & \text{if } \mathbf{x}\in\mathcal{S}\\ \infty & \text{else}\end{cases}$ is the indicator on a set onto which we may compute a *projection*. The (orthogonal) projection of a point $\mathbf{y}$ onto the set $\mathcal{S}$, denoted by $P_{\mathcal{S}}(\mathbf{y})$, is the closest point to $\mathbf{y}$ in $\mathcal{S}$ (see Figure 5.1).

We may also phrase (5.2) as the simple *constrained* optimization problem

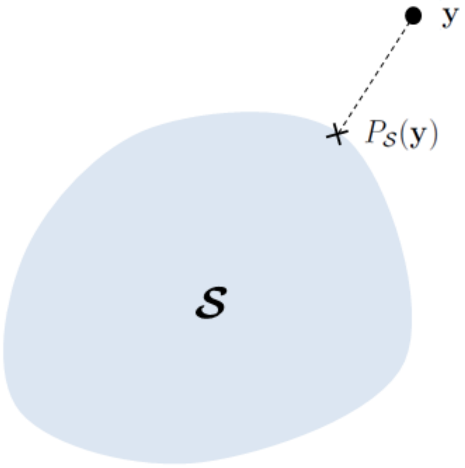

Figure 5.1: The projection of a point $\mathbf{y}$ onto the set $\mathcal{S}$, denoted by $P_{\mathcal{S}}(\mathbf{y})$, is the closest point to $\mathbf{y}$ in $\mathcal{S}$.

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}-\mathbf{y}\|_2^2 \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{S}. \end{aligned} \tag{5.3}$$

Notice that for a point $\mathbf{y}$ that is already in $\mathcal{S}$, the projection is simply the point $\mathbf{y}$ itself.

5.1.1 Analytic projections: Examples

There are a number of interesting sets $\mathcal{S}$ where (5.3) has an analytic solution. We discuss several important examples here (for a more comprehensive catalogue of projections see [?]). The analytic solutions to most of the following examples may be justified using simple geometric arguments, or the KKT conditions discussed in chapter **??**.

Example 5.1. The solid unit ℓ_2 ball

The solid unit ℓ_2 ball is defined as $\mathcal{S} = \{\mathbf{x} \mid \|\mathbf{x}\|_2 \leq 1\}$ with the associated projection problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}-\mathbf{y}\|_2^2 \\ \text{subject to} \quad & \|\mathbf{x}\|_2 \leq 1. \end{aligned} \tag{5.4}$$

The solution in this case is given by

$$P_{\mathcal{S}}\left(\mathbf{y}\right) = \begin{cases} \frac{\mathbf{y}}{\|\mathbf{y}\|_2} & \text{if } \|\mathbf{y}\|_2 > 1 \\ \mathbf{y} & \text{else.} \end{cases} \tag{5.5}$$

Example 5.2. The positive orthant

The positive orthant can be described by the set $\mathcal{S} = \{\mathbf{x} \,|\, \mathbf{x} \geq \mathbf{0}_{K\times 1}\}$, and hence the projection onto the positive orthant may be written as

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} && \|\mathbf{x} - \mathbf{y}\|_2^2 \\ &\text{subject to} && \mathbf{x} \geq \mathbf{0}_{K\times 1}. \end{aligned} \tag{5.6}$$

The solution is given by keeping only the non-negative entries of $\mathbf{y}$, and setting its negative terms to zero. That is, $P_{\mathcal{S}}\left(\mathbf{y}\right) = \mathbf{y}^+$ where $\mathbf{y}^+$ is the positive part of $\mathbf{y}$ defined entry-wise as

$$y_i^+ = \begin{cases} y_i & \text{if } y_i \geq 0 \\ 0 & \text{else.} \end{cases} \tag{5.7}$$

Example 5.3. A hyperplane

The set $\mathcal{S} = \left\{\mathbf{x} \,|\, \mathbf{a}^T\mathbf{x} = b\right\}$ defines a hyperplane whose corresponding projection problem is given as

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} && \|\mathbf{x} - \mathbf{y}\|_2^2 \\ &\text{subject to} && \mathbf{a}^T\mathbf{x} = b. \end{aligned} \tag{5.8}$$

A simple geometric argument gives $P_{\mathcal{S}}\left(\mathbf{y}\right) = \mathbf{y} - \frac{\mathbf{a}^T\mathbf{y} - b}{\|\mathbf{a}\|_2^2}\mathbf{a}$.

Example 5.4. A halfspace

A halfspace is defined by a single linear inequality $\mathcal{S} = \left\{\mathbf{x} \,|\, \mathbf{a}^T\mathbf{x} \leq b\right\}$, with the projection taking the form

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} && \|\mathbf{x} - \mathbf{y}\|_2^2 \\ &\text{subject to} && \mathbf{a}^T\mathbf{x} \leq b. \end{aligned} \tag{5.9}$$

Again, a simple geometric argument gives $P_{\mathcal{S}}\left(\mathbf{y}\right) = \mathbf{y} - \frac{\left(\mathbf{a}^T\mathbf{y} - b\right)^+}{\|\mathbf{a}\|_2^2}\mathbf{a}$.

Example 5.5. A box

A box is defined by $\mathcal{S} = \{\mathbf{x} \,|\, \mathbf{a} \leq \mathbf{x} \leq \mathbf{b}\}$ where $\mathbf{a}$ and $\mathbf{b}$ are given vectors. The projection is then written as

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}-\mathbf{y}\|_2^2 \\ \text{subject to} \quad & \mathbf{a} \leq \mathbf{x} \leq \mathbf{b}. \end{aligned} \tag{5.10}$$

This problem is seperable into the individual components of $\mathbf{x}$. In particular, the i^{th} component of the projection can be found by solving

$$\begin{aligned} \underset{x_i}{\text{minimize}} \quad & (x_i - y_i)^2 \\ \text{subject to} \quad & a_i \leq x_i \leq b_i. \end{aligned} \tag{5.11}$$

The projection operation is straightforward: if necessary, we adjust y_i to lie in the interval $[a_i, b_i]$ as

$$P_{\mathcal{S}}(y_i) = \begin{cases} b_i & \text{if } y_i > b_i \\ y_i & \text{if} \, a_i \leq y_i \leq b_i \\ a_i & \text{if} \, y_i < a_i. \end{cases} \tag{5.12}$$

Example 5.6. An affine set

The affine set defined by $\mathcal{S} = \{\mathbf{x} \,|\, \mathbf{A}\mathbf{x} = \mathbf{b}\}$ gives the projection problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}-\mathbf{y}\|_2^2 \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{b}. \end{aligned} \tag{5.13}$$

In complete analogy with the hyperplane projection, the solution is given by

$$P_{\mathcal{S}}(\mathbf{y}) = \mathbf{y} - \mathbf{A}^{\dagger}(\mathbf{A}\mathbf{y} - \mathbf{b}). \tag{5.14}$$

Example 5.7. The set of k-sparse vectors

The set of k-sparse vectors given by $\mathcal{S} = \{\mathbf{x} \,|\, \|\mathbf{x}\|_0 \leq k\}$ defines the projection problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}-\mathbf{y}\|_2^2 \\ \text{subject to} \quad & \|\mathbf{x}\|_0 \leq k. \end{aligned} \tag{5.15}$$

To form the solution we keep the k largest (in magnitude) entries of $\mathbf{y}$, and set the remaining entries to zero. This operation is referred to as *hard thresholding* and can be written entrywise as

$$P_{\mathcal{S}}(y_i) = \begin{cases} y_i & \text{if } |y_i| \text{ is one of } k \text{ largest} \\ 0 & \text{else.} \end{cases} \tag{5.16}$$

When $\mathbf{y}$ has repeated values as entries the projection may not be unique.

Example 5.8. A subset of entries

For given matrices $\mathbf{B}$ and $\mathbf{Y}$, and the index set Ω, the projection

$$\begin{aligned} & \underset{\mathbf{A}}{\text{minimize}} \quad \|\mathbf{A}-\mathbf{B}\|_F^2 \\ & \text{subject to } \mathbf{A}_{ij}=\mathbf{Y}_{ij} \quad (i,j)\in\Omega, \end{aligned} \tag{5.17}$$

takes the closed form solution defined entry-wise as

$$\mathbf{A}_{ij}=\begin{cases}\mathbf{Y}_{ij} & \text{if } (i,j)\in\Omega \\ \mathbf{B}_{ij} & \text{else.}\end{cases} \tag{5.18}$$

Example 5.9. The set of rank-r matrices

Recall the discussion of Singular Value Decomposition (SVD) in **SECTION SVD LOW-RANK APPROXIMATION** where we approximated an $N \times P$ matrix $\mathbf{B}$ by a rank $r \leq \min(N,P)$ matrix $\mathbf{A}$ as the solution to

$$\begin{aligned} & \underset{\mathbf{A}}{\text{minimize}} \quad \|\mathbf{A}-\mathbf{B}\|_F^2 \\ & \text{subject to } \operatorname{rank}(\mathbf{A}) \leq r. \end{aligned} \tag{5.19}$$

This is precisely the projection of $\mathbf{B}$ onto the set of rank-r matrices given by $\mathcal{S}=\{\mathbf{A} \mid \operatorname{rank}(\mathbf{A}) \leq r\}$. As we saw, a solution here is given by keeping only the first r singular values in the SVD of $\mathbf{B}=\sum_{i=1}^{N}\mathbf{u}_i\sigma_i\mathbf{v}_i^T$, where we have assumed $N \leq P$, as

$$P_{\mathcal{S}}(\mathbf{B})=\sum_{i=1}^{r}\mathbf{u}_i\sigma_i\mathbf{v}_i^T. \tag{5.20}$$

Note that this action is a *hard thresholding* of the singular values of $\mathbf{B}$, and so singular value decomposition as a matrix operation is analogous to the vector hard thresholding projection in (5.15).

Example 5.10. The cone of positive semi-definite matrices

Denoted by $\mathcal{S}=\{\mathbf{A} \mid \mathbf{A} \succeq \mathbf{0}_{N\times N}\}$ the set of $N \times N$ symmetric positive semi-definite matrices, we have the projection problem

$$\begin{aligned} & \underset{\mathbf{A}}{\text{minimize}} \quad \|\mathbf{A}-\mathbf{B}\|_F^2 \\ & \text{subject to } \mathbf{A} \succeq \mathbf{0}_{N\times N}. \end{aligned} \tag{5.21}$$

Assuming the spectral decomposition of $\mathbf{B}$ is given as $\mathbf{B}=\sum_{i=1}^{N}\mathbf{e}_i d_i \mathbf{e}_i^T$, the solution is obtained by keeping only the nonnegative eigenvalues of $\mathbf{B}$, i.e.,

$$P_{\mathcal{S}}(\mathbf{B})=\sum_{i=1}^{N}\mathbf{e}_i d_i^{+}\mathbf{e}_i^T. \tag{5.22}$$

Example 5.11. Orthogonal matrices

The set of orthogonal $N \times N$ matrices is given by $\mathcal{S} = \left\{ \mathbf{A} \mid \mathbf{A}\mathbf{A}^T = \mathbf{I}_{N\times N} \right\}$, and the projection of a matrix $\mathbf{B}$ onto $\mathcal{S}$ may be written as

$$\begin{aligned} &\underset{\mathbf{A}}{\text{minimize}} \quad \|\mathbf{A} - \mathbf{B}\|_F^2 \\ &\text{subject to} \;\; \mathbf{A}\mathbf{A}^T = \mathbf{I}_{N\times N}. \end{aligned} \tag{5.23}$$

The projection may be written in closed form as

$$P_{\mathcal{S}}\left(\mathbf{B}\right) = \left(\mathbf{B}\mathbf{B}^T\right)^{\frac{\dagger}{2}} \mathbf{B}. \tag{5.24}$$

5.1.2 Convexity and uniqueness of projections

With the exception of the sets of k-sparse vectors , rank-r matrices, and orthogonal matrices, the remainder of the projections listed in the previous section are *unique*. This is because the sets onto which we project are *convex*. A set $\mathcal{S}$ is convex if for any two points $\mathbf{x}$ and $\mathbf{y}$ in it, the line connecting the two completely lies inside the set. An example of a convex and non-convex set are illustrated in Figure 5.2.

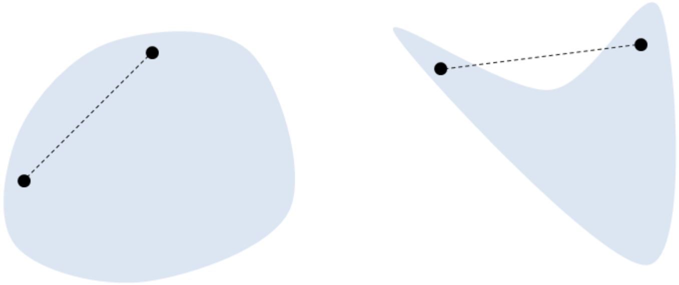

Figure 5.2: (left) A convex set contains the line connecting any two points inside it. (right) A non-convex set does not satisfy this property.

Put formally, a set $\mathcal{S}$ is convex if for any two points $\mathbf{x}$ and $\mathbf{y}$ in $\mathcal{S}$, so too is the line segment connecting them, characterized by $t\mathbf{x} + (1-t)\,\mathbf{y}$ for all $t \in (0,1)$.

The projection onto a convex set is unique, intuitively because such a set is always "bulging outwards" in every direction, and so there is always only a single closest point to $\mathbf{y}$. The same reasoning does not apply to non-convex sets as illustrated in Figure 5.3.

It is easy to show that if two sets $\mathcal{S}_1$ and $\mathcal{S}_2$ are convex, then so too is the set formed by their intersection $\mathcal{S} = \mathcal{S}_1 \cap \mathcal{S}_2$. In order to project a point onto such an $\mathcal{S}$

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \quad \|\mathbf{x} - \mathbf{y}\|_2^2 \\ &\text{subject to} \;\; \mathbf{x} \in \mathcal{S}_1 \cap \mathcal{S}_2, \end{aligned} \tag{5.25}$$

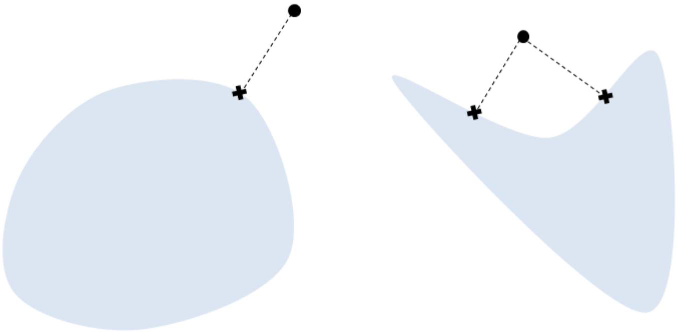

Figure 5.3: (left) A projection onto a convex set is always unique. (right) A projection onto a non-convex set is not necessarily unique.

we can alternate projecting onto either set. Denoting by $\mathbf{y}^0 = \mathbf{y}$ we compute

$$\mathbf{y}^k = P_{\mathcal{S}_1}\left(P_{\mathcal{S}_2}\left(\mathbf{y}^{k-1}\right)\right), \tag{5.26}$$

until, for example, $\frac{\left\|\mathbf{y}^k - \mathbf{y}^{k-1}\right\|_2}{\left\|\mathbf{y}^k\right\|_2}$ is small enough.

Alternatively, we can average projections onto both sets as

$$\mathbf{y}^k = \frac{P_{\mathcal{S}_1}\left(\mathbf{y}^{k-1}\right) + P_{\mathcal{S}_2}\left(\mathbf{y}^{k-1}\right)}{2}, \tag{5.27}$$

where again $\mathbf{y}^0 = \mathbf{y}$ and we iterate until the difference in subsequent iterates is small. Both alternating and averaging projections are intuitively reasonable and rigorously justifiable (see e.g., [?]) methods for projecting $\mathbf{y}$ onto $\mathcal{S}$.

A typically faster method for projecting onto $\mathcal{S}$ is referred to as *Dykstra's projection algorithm* (see [?] and references therein), and adds a correction term to the alternating projection scheme, giving the updates

$$\begin{aligned} \mathbf{z}^k &= P_{\mathcal{S}_2}\left(\mathbf{y}^{k-1} - \mathbf{d}^{k-1}\right) \\ \mathbf{y}^k &= P_{\mathcal{S}_1}\left(\mathbf{z}^k + \mathbf{d}^{k-1}\right) \\ \mathbf{d}^k &= \mathbf{d}^{k-1} + \mathbf{z}^k - \mathbf{y}^k, \end{aligned} \tag{5.28}$$

where we initialize $\mathbf{y}^0 = \mathbf{y}$ and $\mathbf{d}^0$ may be assigned random values. We will see Dykstra's projection algorithm arise naturally in the context of primal-dual optimization in Chapter ??.

5.2 Projected gradient

We now discuss a broader subclass of problem (5.1) of the form

$$\underset{\mathbf{x}}{\text{minimize}}\ f\left(\mathbf{x}\right)+I_{\mathcal{S}}\left(\mathbf{x}\right), \tag{5.29}$$

where $I_{\mathcal{S}}\left(\mathbf{x}\right)=\begin{cases}0 & \text{if}\,\mathbf{x}\in\mathcal{S}\\ \infty & \text{else}\end{cases}$ is the indicator on a set onto which we may compute a projection, and f is convex with Lipschitz continuous gradient. Equivlaently, we may write this problem as the constrained problem

$$\begin{aligned}\underset{\mathbf{x}}{\text{minimize}}\quad & f\left(\mathbf{x}\right)\\ \text{subject to}\quad & \mathbf{x}\in\mathcal{S}.\end{aligned} \tag{5.30}$$

In this section we will see how the gradient descent sceheme for unconstrained minimization of f, discussed in the previous chapter, can be easily extended to solve these sorts of problems. This extension, known as the *projected gradient* algorithm, consists of repeating two simple pieces as illustrated in Figure 5.4: first we take a gradient descent (or accelerated gradient descent) step towards a minimum of f, and then project it back onto the feasible set $\mathcal{S}$.

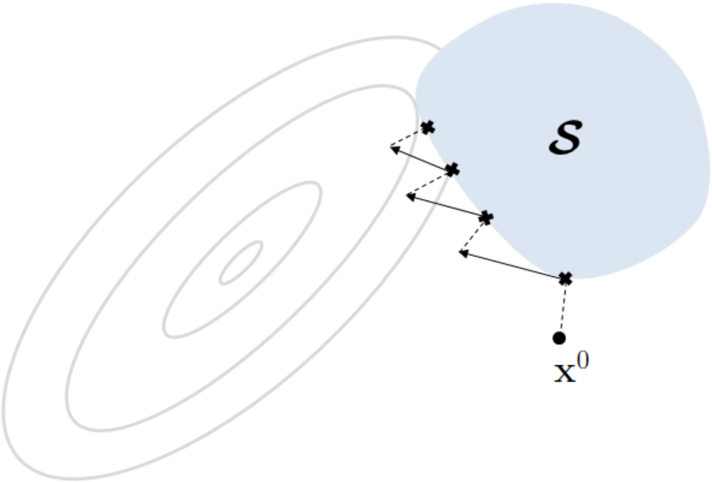

Figure 5.4: Projected gradient descent illustrated. We repeat the following two steps: take a gradient descent step towards a minimum of f, and then project it back onto $\mathcal{S}$.

5.2.1 The gradient descent step as a projection

Recall that f is majorized at each point in its domain by a quadratic function, and hopping along the minima of these quadratic majorizers is *equivalent* to performing gradient descent on f with fixed step length $\frac{1}{L}$. This equivalence was proven in Chapter **??** by showing that the gradient descent step $\mathbf{x}^k=\mathbf{x}^{k-1}-\frac{1}{L}\nabla f\left(\mathbf{x}^{k-1}\right)$ is in fact the minimizer of

$$\underset{\mathbf{x}}{\text{minimize}}\ f\left(\mathbf{x}^{k-1}\right)+\nabla f\left(\mathbf{x}^{k-1}\right)^T\left(\mathbf{x}-\mathbf{x}^{k-1}\right)+\frac{L}{2}\|\mathbf{x}-\mathbf{x}^{k-1}\|_2^2\ , \tag{5.31}$$

where the objective is the quadratic majorizer of f at $\mathbf{x}^{k-1}$. This alternative characterization of the gradient descent step for convex functions with Lipschitz continuous gradient, as the minimum of a quadratic majorizer, is often referred to as the *proximal definition* of the gradient. This is named as such because the quadratic term in (5.31) is forcing the update $\mathbf{x}^k$ to be in *proximity* to the previous step $\mathbf{x}^{k-1}$.

To solve (5.31) we previously used calculus by setting the gradient of the objective equal to zero. Alternatively, it is easy to show by *completing the square* that (5.31) can be written equivalently[1] as

$$\underset{\mathbf{x}}{\text{minimize}}\ \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|_2^2, \tag{5.32}$$

where $\mathbf{y} = \mathbf{x}^{k-1} - \frac{1}{L}\nabla f\left(\mathbf{x}^{k-1}\right)$.

5.2.2 The projected gradient descent step

Now, turning to the constrained problem

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \quad f\left(\mathbf{x}\right) \\ &\text{subject to } \mathbf{x} \in \mathcal{S}, \end{aligned} \tag{5.33}$$

we can easily see how to define the k^{th} projected gradient descent step. Like the unconstrained case, we again desire to move to the minimum of the quadratic majorizer, only now we constrain the admittable search space to $\mathcal{S}$ as in

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \quad f\left(\mathbf{x}^{k-1}\right) + \nabla f\left(\mathbf{x}^{k-1}\right)^T\left(\mathbf{x}-\mathbf{x}^{k-1}\right) + \frac{L}{2}\|\mathbf{x}-\mathbf{x}^{k-1}\|_2^2 \\ &\text{subject to } \mathbf{x} \in \mathcal{S}. \end{aligned} \tag{5.34}$$

From the previous section we know that the objective may be simplified, giving the equivalent problem

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \quad \frac{L}{2}\left\|\mathbf{x}-\mathbf{y}\right\|_2^2 \\ &\text{subject to } \ \mathbf{x} \in \mathcal{S}, \end{aligned} \tag{5.35}$$

where $\mathbf{y} = \mathbf{x}^{k-1} - \frac{1}{L}\nabla f\left(\mathbf{x}^{k-1}\right)$. Since the constant $\frac{L}{2}$ in the objective may be ignored without changing the problem, this is simply the canonical form of the projection of $\mathbf{y}$ onto $\mathcal{S}$, or in other words the projection of the k^{th} gradient step onto $\mathcal{S}$. This is referred to as the projected gradient step and may be written compactly as

[1] Since the objective is minimized over $\mathbf{x}$ we may ignore the constant term $f\left(\mathbf{x}^{k-1}\right)$ in (5.31) and add the constant term $\frac{L}{2}\|\frac{1}{L}\nabla f\left(\mathbf{x}^{k-1}\right)\|_2^2$, without changing the problem. We can then wrap up $\nabla f\left(\mathbf{x}^{k-1}\right)^T\left(\mathbf{x}-\mathbf{x}^{k-1}\right) + \frac{L}{2}\|\mathbf{x}-\mathbf{x}^{k-1}\|_2^2 + \frac{L}{2}\|\frac{1}{L}\nabla f\left(\mathbf{x}^{k-1}\right)\|_2^2$ into the objective of (5.32).

$$\mathbf{x}^k = P_{\mathcal{S}}\left(\mathbf{y}\right) = P_{\mathcal{S}}\left(\mathbf{x}^{k-1} - \frac{1}{L}\nabla f\left(\mathbf{x}^{k-1}\right)\right). \tag{5.36}$$

For convenience, we summarize the projected gradient descent algorithm in Algorithm 5.1.

Algorithm 5.1 Projected gradient descent

```
Input: function $f$ with Lipschitz constant $L$, projection operator $P_{\mathcal{S}}(\cdot)$ for the convex constraint set $\mathcal{S}$, and initial point $\mathbf{x}^0$
$k = 1$
Repeat until convergence:
    $\mathbf{x}^k = P_{\mathcal{S}}\left(\mathbf{x}^{k-1} - \frac{1}{L}\nabla f\left(\mathbf{x}^{k-1}\right)\right)$
    $k \leftarrow k + 1$
```

5.2.3 Projected gradient: Examples

Example 5.12. The Lasso

Recall that the Lasso problem defined as

$$\underset{\mathbf{x}}{\text{minimize}}\ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1 \quad , \tag{5.37}$$

can be reformulated as a Quadratic Program (see Chapter ?)

$$\begin{aligned} &\underset{\mathbf{z}}{\text{minimize}} \quad \mathbf{z}^T\mathbf{B}\mathbf{z} + \mathbf{c}^T\mathbf{z} \\ &\text{subject to} \ \ \mathbf{z} \geq \mathbf{0}, \end{aligned} \tag{5.38}$$

where $\mathbf{z} = \begin{bmatrix} \mathbf{p} \\ \mathbf{n} \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} \mathbf{A}^T\mathbf{A} & -\mathbf{A}^T\mathbf{A} \\ -\mathbf{A}^T\mathbf{A} & \mathbf{A}^T\mathbf{A} \end{bmatrix}$, $\mathbf{c} = \lambda\mathbf{1}_{2M\times 1} + \begin{bmatrix} -\mathbf{A}^T\mathbf{b} \\ \mathbf{A}^T\mathbf{b} \end{bmatrix}$, and $\mathbf{p}$ and $\mathbf{n}$ are the positive and negative parts of $\mathbf{x}$, respectively. We can write out the projected gradient steps for this problem as

$$\mathbf{z}^k = \left(\mathbf{z}^{k-1} - \frac{1}{\|\mathbf{B}\|_2^2}\left(\mathbf{B}\mathbf{z}^{k-1} + \mathbf{c}\right)\right)^+, \tag{5.39}$$

where $\|\mathbf{B}\|_2^2$ is the largest eigenvalue of $\mathbf{B}$.

Example 5.13. Non-negative Matrix Factorization

The Non-negative Matrix Factorization problem is given by

$$\begin{aligned} \underset{\mathbf{A},\mathbf{X}}{\text{minimize}} \quad & \|\mathbf{A}\mathbf{X}-\mathbf{B}\|_F^2 \\ \text{subject to } & \mathbf{A},\mathbf{X} \geq \mathbf{0}. \end{aligned} \tag{5.40}$$

While the objective here is non-convex we can find local minima by applying *alternating minimization,* switching back and forth between minimizing over $\mathbf{X}$ with $\mathbf{A}$ fixed, and over $\mathbf{A}$ with $\mathbf{X}$ fixed.

Starting with the subproblem in $\mathbf{A}$, notice that the objective is separable in the *rows* of $\mathbf{A}$ which we can see by expanding the objective row-wise

$$\|\mathbf{A}\mathbf{X}-\mathbf{B}\|_F^2 = \sum_i \|\mathbf{a}^i\mathbf{X}-\mathbf{b}^i\|_2^2, \tag{5.41}$$

where $\mathbf{a}^i$ and $\mathbf{b}^i$ are the i^{th} rows of $\mathbf{A}$ and $\mathbf{B}$, respectively. This means we can parallelize the update for the matrix $\mathbf{A}$ via the row-wise problems of the form

$$\begin{aligned} \underset{\mathbf{a}^i}{\text{minimize}} \quad & \|\mathbf{a}^i\mathbf{X}-\mathbf{b}^i\|_2^2 \\ \text{subject to } & \mathbf{a}^i \geq \mathbf{0}. \end{aligned} \tag{5.42}$$

The projected gradient update step is then given by

$$\left(\mathbf{a}^i\right)^k = \left[\left(\mathbf{a}^i\right)^{k-1} - \frac{1}{\|\mathbf{X}\|_2^2}\left(\left(\mathbf{a}^i\right)^{k-1}\mathbf{X}-\mathbf{b}^i\right)\mathbf{X}^T\right]^+. \tag{5.43}$$

Note that we can write all the row updates together as a single matrix update of the similar shape

$$\mathbf{A}^k = \left[\mathbf{A}^{k-1} - \frac{1}{\|\mathbf{X}\|_2^2}\left(\mathbf{A}^{k-1}\mathbf{X}-\mathbf{B}\right)\mathbf{X}^T\right]^+. \tag{5.44}$$

Next, the update in $\mathbf{X}$ is seperable in the *columns* of $\mathbf{X}$, which we can see by expanding the original objective column-wise

$$\|\mathbf{A}\mathbf{X}-\mathbf{B}\|_F^2 = \sum_j \|\mathbf{A}\mathbf{x}_j-\mathbf{b}_j\|_2^2, \tag{5.45}$$

where $\mathbf{x}_j$ and $\mathbf{b}_j$ are the j^{th} columns of $\mathbf{X}$ and $\mathbf{B}$, respectively. Again this problem is seperable, and thus the $\mathbf{X}$ update is parallelizable. The update problem for the j^{th} column of $\mathbf{X}$ is given by

$$\begin{aligned} &\underset{\mathbf{x}_j}{\text{minimize}} \quad \|\mathbf{A}\mathbf{x}_j - \mathbf{b}_j\|_2^2 \\ &\text{subject to } \mathbf{x}_j \geq \mathbf{0}. \end{aligned} \tag{5.46}$$

The projected gradient step corresponding to this problem is then given by

$$\mathbf{x}_j^k = \left[\mathbf{x}_j^{k-1} - \frac{1}{\|\mathbf{A}\|_2^2}\mathbf{A}^T\left(\mathbf{A}\mathbf{x}_j^{k-1} - \mathbf{b}_j\right)\right]^+. \tag{5.47}$$

Once again we can write all the column updates together as a single matrix update

$$\mathbf{X}^k = \left[\mathbf{X}^{k-1} - \frac{1}{\|\mathbf{A}\|_2^2}\mathbf{A}^T\left(\mathbf{A}\mathbf{X}^{k-1} - \mathbf{B}\right)\right]^+. \tag{5.48}$$

Finally, notice that in both directions **A** and **X** we may apply the accelerated projected gradient scheme.

Example 5.14. Dictionary update in sparse coding

We can apply projected gradient to updating the dictionary **A** in the sparse coding problem taking the form

$$\begin{aligned} &\underset{\mathbf{A},\mathbf{X}}{\text{minimize}} \quad \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F^2 + \lambda\|\mathbf{X}\|_1 \\ &\text{subject to } \ \|\mathbf{A}\|_F \leq 1. \end{aligned} \tag{5.49}$$

Minimizing over **A** gives the subproblem

$$\begin{aligned} &\underset{\mathbf{A}}{\text{minimize}} \quad \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F^2 \\ &\text{subject to } \ \|\mathbf{A}\|_F \leq 1, \end{aligned} \tag{5.50}$$

to which we may apply projected gradient, with the k^{th} step taking the form

$$\mathbf{A}^k = P_{\mathcal{S}}\left(\mathbf{A}^{k-1} - \frac{1}{\|\mathbf{X}\|_2^2}\left(\mathbf{A}^{k-1}\mathbf{X} - \mathbf{B}\right)\mathbf{X}^T\right), \tag{5.51}$$

where $P_{\mathcal{S}}(\cdot)$ is the matrix version of the vector projection onto the unit ball, given by

$$P_{\mathcal{S}}(\mathbf{A}) = \begin{cases} \frac{\mathbf{A}}{\|\mathbf{A}\|_F} & \text{if } \|\mathbf{A}\|_F > 1 \\ \mathbf{A} & \text{else.} \end{cases} \tag{5.52}$$

Example 5.15. Sparse Least Squares (ℓ_0 constrained)

Under certain conditions on the matrix $\mathbf{A}$, known as a Restricted Isometry Property, the ℓ_0 constrained problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{Ax}-\mathbf{b}\|_2^2 \\ \text{subject to} \quad & \|\mathbf{x}\|_0 \leq k, \end{aligned} \tag{5.53}$$

can be solved using hard thresholding gradient projection [?]. This kind of $\mathbf{A}$ arises often in the context of *compressive sensing* (see e.g., [?]). Gradient projection takes the update form

$$\mathbf{x}^k = P_{\mathcal{S}}\left(\mathbf{x}^{k-1} - \alpha\mathbf{A}^T\left(\mathbf{Ax}^{k-1}-\mathbf{b}\right)\right), \tag{5.54}$$

where $P_{\mathcal{S}}(\cdot)$ is now the hard thresholding operator keeping only the k largest (in magnitude) entries of the input, and α is a step length. A similiar idea can be applied to applying projected gradient to the rank constrained matrix completion problem [?].

5.3 Proximal operators

We have now seen how the *proximal definition* of the gradient descent step is naturally adjusted to give closed form *projected gradient* steps when f is constrained by a set for which we can compute a projection. In this section we will see how a simple generalization of a *projection*, referred to as the *proximal operator*, can similarly generalize the proximal definition of the gradient step to produce what is known as *proximal gradient* step. This broadens the projected gradient approach previously discussed, providing a more general framework.

Recall that the *indicator function* of a set $\mathcal{S}$ is given by

$$I_{\mathcal{S}}(\mathbf{x}) = \begin{cases} 0 & \text{if}\, \mathbf{x} \in \mathcal{S} \\ \infty & \text{else.} \end{cases} \tag{5.55}$$

Using this notation we can reformulate the standard projection of $\mathbf{y}$ onto $\mathcal{S}$

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}-\mathbf{y}\|_2^2 \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{S}, \end{aligned} \tag{5.56}$$

equivalently, as the unconstrained problem

$$\underset{\mathbf{x}}{\text{minimize}}\, \|\mathbf{x}-\mathbf{y}\|_2^2 + I_{\mathcal{S}}(\mathbf{x}). \tag{5.57}$$

This is a simple, and at first seemingly arbitrary, thing to do. But it rephrases the projection problem in a useful, and generalizable way.

We have seen through a variety of examples in Section 5.1.1 that there are a host of projection problems whose solutions are given by $P_{\mathcal{S}}(\mathbf{y})$, i.e., some function of $\mathbf{y}$. This naturally raises the question: can we replace the indicator function $I_{\mathcal{S}}(\cdot)$ with some other function $g(\cdot)$, and expect an output that is some simple function of $\mathbf{y}$ as well? In other words, what sorts of functions $g(\cdot)$ in the problem

$$\underset{\mathbf{x}}{\text{minimize}}\ \|\mathbf{x}-\mathbf{y}\|_2^2+g(\mathbf{x}), \tag{5.58}$$

will produce a closed form solution? Many such g indeed exist, in which case the solution to (5.58) is referred to as the *proximal operator* of g at point $\mathbf{y}$, denoted by $\text{prox}_g(\mathbf{y})$.

5.3.1 Proximal operators: Examples

The following are some important examples of functions which have analytic proximal evaluations. For a more comprehensive catalogue as well as a host of additional interpretations and valuable connections between projections and proximal operators, see [**?**].

Example 5.16. Projections

All of the examples in Section 5.1.1 are proximal operators of the form

$$\underset{\mathbf{x}}{\text{minimize}}\ \|\mathbf{x}-\mathbf{y}\|_2^2+I_{\mathcal{S}}(\mathbf{x})\ , \tag{5.59}$$

where $I_{\mathcal{S}}(\mathbf{x})$ is the indicator function on the set $\mathcal{S}$.

Example 5.17. ℓ_1 norm

The proximal evaluation of $g(\mathbf{x})=\lambda\|\mathbf{x}\|_1$ takes the form

$$\underset{\mathbf{x}}{\text{minimize}}\ \|\mathbf{x}-\mathbf{y}\|_2^2+\lambda\|\mathbf{x}\|_1, \tag{5.60}$$

where $\lambda>0$ is a constant. This problem is separable in each entry of $\mathbf{x}$ and hence we can solve for the i^{th} component x_i individually as

$$\underset{x_i}{\text{minimize}}\ \left(x_i-y_i\right)^2+\lambda|x_i|. \tag{5.61}$$

With a little algebra one can easily show[2] that there exists a closed form solution to this subproblem, and hence to the problem over the entire vector $\mathbf{x}$. This solution is denoted

$$\mathbf{x}^\star=\mathcal{T}_\lambda(\mathbf{y}), \tag{5.62}$$

where $\mathcal{T}_\lambda(\mathbf{y})$ is defined entry-wise as $\mathcal{T}_\lambda(y_i)=\left[1-\frac{\lambda}{2|y_i|}\right]^+y_i$, and is referred to as the *shrinkage* or *soft thresholding operator*.

[2] Work it out in cases involving the sign of y_i. First notice that if $y_i=0$ then clearly $x_i^\star=0$. Now suppose $y_i>0$. You can show (by contradiction) that if this is true then we must have $x_i^\star\geq0$. This implies that the optimal $x_i^\star$ is recovered via solving the simple projection problem $\underset{x_i\geq0}{\text{minimize}}\ \left(x_i-y_i\right)^2+\lambda x_i$, so $x_i^\star=\left[y_i-\frac{\lambda}{2}\right]^+$. Likewise, assume that $y_i<0$ and you can show that we must have $x_i^\star=-\left[-y_i-\frac{\lambda}{2}\right]^+$. Combined, this gives for general y_i that $x_i^\star=\left[|y_i|-\frac{\lambda}{2}\right]^+\text{sign}(y_i)$, or equivalently $x_i^\star=\left[1-\frac{\lambda}{2|y_i|}\right]^+y_i$.

Example 5.18. ℓ_2 norm squared

The proximal evaluation of $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_2^2$ takes the form

$$\underset{\mathbf{x}}{\text{minimize}}\ \|\mathbf{x}-\mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_2^2\ . \tag{5.63}$$

Setting the gradient equal to zero and solving for $\mathbf{x}$ gives the solution

$$\mathbf{x}^\star = \frac{1}{1+\lambda}\mathbf{y}. \tag{5.64}$$

Example 5.19. ℓ_1/ℓ_2 norm

Assume we want to count the number of non-zero columns (i.e., columns that have at least one nonzero entry) of an $N \times K$ matrix $\mathbf{A}$. One way to do so is by forming a vector $\mathbf{a}$ containing the ℓ_2 norm of each column of $\mathbf{A}$ as its entries

$$\mathbf{a} = \left[\|\mathbf{a}_1\|_2,\ \|\mathbf{a}_2\|_2,\ ...,\ \|\mathbf{a}_K\|_2\right]^T . \tag{5.65}$$

Now the ℓ_0 norm of this vector $\|\mathbf{a}\|_0$ gives the number of non-zero columns in $\mathbf{A}$. Replacing the ℓ_0 with the ℓ_1 norm gives the analogous *convexified* version of this count as

$$\|\mathbf{a}\|_1 = \sum_{k=1}^{K}\|\mathbf{a}_k\|_2. \tag{5.66}$$

This is in fact a *matrix norm* on $\mathbf{A}$ known as the ℓ_1/ℓ_2 norm, and denoted as $\|\mathbf{A}\|_{1,2}$. The proximal problem associated to $g(\mathbf{A}) = \lambda\|\mathbf{A}\|_{1,2}$ takes the form

$$\underset{\mathbf{A}}{\text{minimize}}\ \|\mathbf{A}-\mathbf{B}\|_F^2 + \lambda\|\mathbf{A}\|_{1,2}. \tag{5.67}$$

This says we are trying to find a matrix to approximate $\mathbf{B}$ with zero columns, the number of which is dependendent on the value of λ. Since this problem is seperable in columns of $\mathbf{A}$, we can solve for each column $\mathbf{a}_k$ individually as

$$\underset{\mathbf{a}_k}{\text{minimize}}\ \|\mathbf{a}_k-\mathbf{b}_k\|_2^2 + \lambda\|\mathbf{a}_k\|_2. \tag{5.68}$$

The optimum $\mathbf{a}_k^\star$ can be calculated via a simple sub-differential optimality condition that is a generalization of the standard first order condition (see e.g., [?])

$$\mathbf{a}_k^\star = \left(1 - \frac{\lambda}{2\|\mathbf{b}_k\|_2}\right)^+ \mathbf{b}_k. \tag{5.69}$$

Notice the similarity between the form of (5.69) and the soft thresholding operator.

Example 5.20. Nuclear norm

The proximal problem associated with $g(\mathbf{A}) = \lambda\|\mathbf{A}\|_*$ (see **SECTION RANK AND NUCLEAR**) may be written as

$$\underset{\mathbf{A}}{\text{minimize}}\ \|\mathbf{A}-\mathbf{B}\|_F^2 + \lambda\|\mathbf{A}\|_*. \tag{5.70}$$

By analogy, since the nuclear norm is simply the ℓ_1 norm of $\boldsymbol{\sigma} = \text{diag}\left(\boldsymbol{\Sigma}\right)$, the vector containing the singular values of $\mathbf{B} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, it is reasonable to guess that the solution to (5.70) is given by applying the soft thresholding operator to $\boldsymbol{\Sigma}$, as in

$$\mathcal{D}_\lambda\left(\mathbf{B}\right) = \mathbf{U}\mathcal{T}_\lambda\left(\boldsymbol{\Sigma}\right)\mathbf{V}^T. \tag{5.71}$$

This is indeed correct, and can be proven rigorously (see e.g., [?]). Writing this as a sum of rank-1 singular matrices, we have

$$\mathcal{D}_\lambda\left(\mathbf{B}\right) = \sum_i \left(\sigma_i - \frac{\lambda}{2}\right)^+ \mathbf{u}_i\mathbf{v}_i^T, \tag{5.72}$$

where we have used the fact that singular values are always non-negative.

5.4 Proximal gradient algorithms

In thinking about solving

$$\underset{\mathbf{x}}{\text{minimize}}\ f\left(\mathbf{x}\right) + g\left(\mathbf{x}\right), \tag{5.73}$$

where g is for instance one of the functions discussed in the previous section, we can apply the same idea to get at a gradient descent step here as we did in the projected gradient case in Section 5.2.1. Again, at the k^{th} gradient step we think of minimizing the sum of the majorizing quadratic of f at $\mathbf{x}^{k-1}$ and $g\left(\mathbf{x}\right)$, as

$$\underset{\mathbf{x}}{\text{minimize}}\ f\left(\mathbf{x}^{k-1}\right) + \nabla f\left(\mathbf{x}^{k-1}\right)^T\left(\mathbf{x}-\mathbf{x}^{k-1}\right) + \tfrac{L}{2}\|\mathbf{x}-\mathbf{x}^{k-1}\|_2^2 + g\left(\mathbf{x}\right)\ . \tag{5.74}$$

Wrapping up the quadratic in the same way we did in Section 5.2.1 gives the equivalent problem

$$\underset{\mathbf{x}}{\text{minimize}}\ \tfrac{L}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 + g\left(\mathbf{x}\right), \tag{5.75}$$

where once again $\mathbf{y} = \mathbf{x}^{k-1} - \frac{1}{L}\nabla f\left(\mathbf{x}^{k-1}\right)$. Multiplying the objective by $\frac{2}{L}$ gives precisely a proximal form

$$\underset{\mathbf{x}}{\text{minimize}}\ \|\mathbf{x}-\mathbf{y}\|_2^2 + \tfrac{2}{L}g\left(\mathbf{x}\right), \tag{5.76}$$

examples of which we discussed in the previous section. Denoting by $\text{prox}_{\frac{2}{L}g}\left(\cdot\right)$ the solution to this proximal problem for general g, the k^{th} proximal gradient step is written as

$$\mathbf{x}^k = \text{prox}_{\frac{2}{L}g}\left(\mathbf{x}^{k-1} - \frac{1}{L}\nabla f\left(\mathbf{x}^{k-1}\right)\right). \tag{5.77}$$

For convenience, we summarize the proximal gradient descent algorithm in Algorithm 5.2.

Algorithm 5.2 Proximal gradient descent

```
Input: function $f$ with Lipschitz constant $L$, proximal operator $\text{prox}_g(\cdot)$, and initial point $\mathbf{x}^0$
$k = 1$
Repeat until convergence:
    $\mathbf{x}^k = \text{prox}_{\frac{2}{L}g}\left(\mathbf{x}^{k-1} - \frac{1}{L}\nabla f\left(\mathbf{x}^{k-1}\right)\right)$
    $k \leftarrow k + 1$
```

5.4.1 Accelerated proximal gradient

We have now seen how the fixed length gradient descent step for convex f with Lipschitz continuous gradient is carried over, essentially unaltered, from the unconstrained form to the proximal case. Therefore it is not unreasonable to expect that we might be able to carry over the *optimal gradient step* for unconstrained minimization discussed in Section **??** to the proximal case as well. This in fact can be done, giving the *accelerated proximal gradient* scheme

$$\begin{array}{r}\mathbf{x}^k = \text{prox}_{\frac{2}{L}g}\left[\mathbf{y}^{k-1} - \frac{1}{L}\nabla f\left(\mathbf{y}^{k-1}\right)\right] \\ \mathbf{y}^k = \mathbf{x}^k + \frac{k}{k+3}\left(\mathbf{x}^k - \mathbf{x}^{k-1}\right).\end{array} \tag{5.78}$$

We again achieve the same speed-up here as we did in the unconstrained version: accelerated proximal gradient provides an order of magnitued improvement over standard proximal gradient, reaching within $\frac{1}{k^2}$ of a global minimum of (5.73) in $\mathcal{O}(k)$ steps. For convenience, we summarize the accelerated proximal gradient descent algorithm in Algorithm 5.3.

Algorithm 5.3 Accelerated proximal gradient descent

```
Input: function $f$ with Lipschitz constant $L$, proximal operator $\text{prox}_g(\cdot)$, and initial points $\mathbf{x}^0$ and $\mathbf{y}^0$
$k = 1$
Repeat until convergence:
    $\mathbf{x}^k = \text{prox}_{\frac{2}{L}g}\left(\mathbf{y}^{k-1} - \frac{1}{L}\nabla f\left(\mathbf{y}^{k-1}\right)\right)$
    $\mathbf{y}^k = \mathbf{x}^k + \frac{k}{k+3}\left(\mathbf{x}^k - \mathbf{x}^{k-1}\right)$
    $k \leftarrow k + 1$
```

Example 5.21. The speed-up advantage of accelerated proximal gradient descent

In Figure 5.5 we illustrate the speed-up provided by the accelerated proximal gradient descent procedure applied to the Lasso problem

$$\underset{\mathbf{x}}{\text{minimize}}\ \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2+\lambda\|\mathbf{x}\|_1 \quad , \tag{5.79}$$

through a simple experiment. Here we have created an overcomplete matrix $\mathbf{A}$ of size 100×200 with standard normal entries, as well as an instance of a 200×1 zero vector $\mathbf{x}$ with 10 randomly placed ± 1. We then make $\mathbf{b} = \mathbf{A}\mathbf{x}$, and look to recover the $\mathbf{x}$ we started with using the standard and accelerated proximal gradient descent algorithms. As shown in Figure 5.5, convergence is achieved in fewer iterations using the accelerated version of the proximal gradient algorithm.

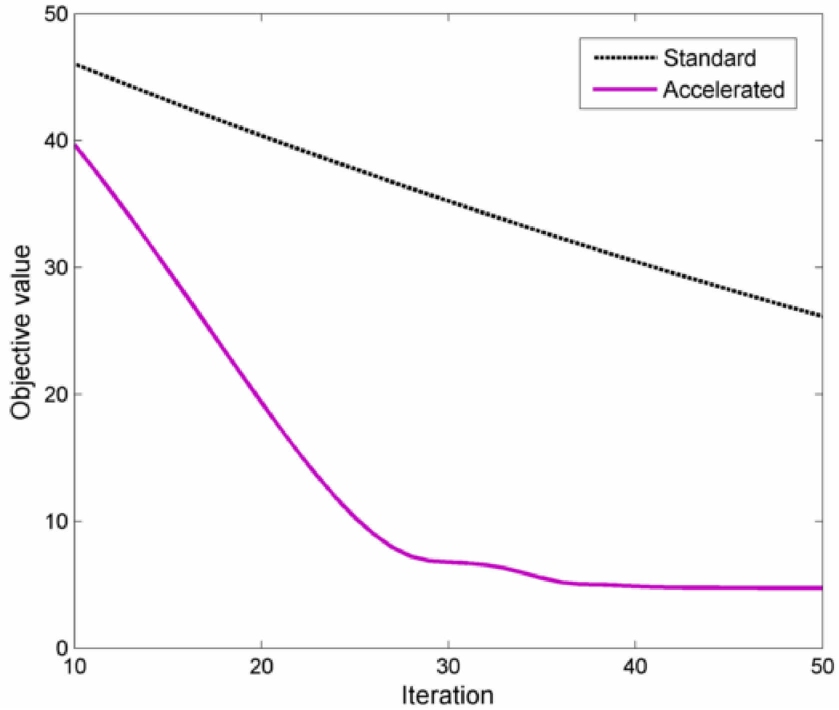

Figure 5.5: Comparison of standard and accelerated proximal gradient applied to the Lasso problem.

5.4.2 Proximal gradient: Examples

Here we discuss the (accelerated) proximal gradient algorithms for a variety of problems we saw in previous chapters. For clarity of exposition, we report only the proximal gradient steps in each instance. However, the accelerated versions can and should be used as in Algorithm 5.3.

Example 5.22. The Lasso

The $\ell_1$regularized Least Squares problem, typically referred to as the Lasso, is written formally as

$$\underset{\mathbf{x}}{\text{minimize}} \; \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1 \quad . \tag{5.80}$$

The Least Squares part of the objective, i.e., $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ is Lipschitz with constant $L = 2\|\mathbf{A}\|_2^2$. With $\nabla f(\mathbf{x}) = 2\mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{b})$ we then have that the k^{th} proximal gradient step for this problem takes the form

$$\mathbf{x}^k = \mathcal{T}_{\frac{2\lambda}{\|\mathbf{A}\|_2^2}} \left[\mathbf{x}^{k-1} - \frac{1}{\|\mathbf{A}\|_2^2} \mathbf{A}^T \left(\mathbf{A}\mathbf{x}^{k-1} - \mathbf{b} \right) \right], \tag{5.81}$$

where $\mathcal{T}_{\frac{2\lambda}{\|\mathbf{A}\|_2^2}}(\cdot)$ is the *soft thresholding* operator from Section 5.3.

Example 5.23. Sparse logistic regression

With f as the cross-entrpy loss (see Section **??**) we can phrase sparse logistic regression problem as

$$\underset{\mathbf{x}}{\text{minimize}} \; f(\mathbf{x}) + \lambda \|\mathbf{x}\|_1 \quad . \tag{5.82}$$

Using the Lipschitz constant L derived in Example **??** we may perform proximal gradient descent taking steps of the form

$$\mathbf{x}^k = \mathcal{T}_{\frac{2\lambda}{L}} \left[\mathbf{x}^{k-1} - \frac{1}{L} \nabla f\left(\mathbf{x}^{k-1}\right) \right]. \tag{5.83}$$

Example 5.24. Sparse PCA

Recall from Section **??** that the sparse PCA model takes the form

$$\underset{\mathbf{A},\mathbf{X}}{\text{minimize}} \sum_i \|\mathbf{A}\mathbf{x}_i - \mathbf{b}_i\|_2^2 + \lambda_i \sum_i \|\mathbf{x}_i\|_1 + \mu \|\mathbf{A}\|_F^2 \quad . \tag{5.84}$$

The most straight-forward approach to solving this problem is through alternatingly minimizing over the columns of $\mathbf{X}$ and $\mathbf{A}$.

The $\mathbf{X}$ update requires proximal gradient steps of the form

$$\mathbf{x}_i^k = \mathcal{T}_{\frac{2\lambda}{\|\mathbf{A}\|_2^2}} \left[\mathbf{x}_i^{k-1} - \frac{1}{\|\mathbf{A}\|_2^2} \mathbf{A}^T \left(\mathbf{A}\mathbf{x}_i^{k-1} - \mathbf{b}_i \right) \right], \tag{5.85}$$

which may easily be parallelized. Minimizing over $\mathbf{A}$, on the other hand, gives the closed form update

$$\mathbf{A} = \mathbf{B}\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T + \mu \mathbf{I}_{N\times N} \right)^{-1}. \tag{5.86}$$

Alternatively, we can solve the following constrained version of the sparse PCA problem

$$\begin{aligned} \underset{\mathbf{A},\mathbf{X}}{\text{minimize}} \quad & \sum_i \|\mathbf{A}\mathbf{x}_i - \mathbf{b}_i\|_2^2 + \lambda_i \sum_i \|\mathbf{x}_i\|_1 \\ \text{subject to} \quad & \|\mathbf{A}\|_F \leq C, \end{aligned} \tag{5.87}$$

by alternatingly minimizing over $\mathbf{X}$ using the same approach, and over $\mathbf{A}$ by using (accelerated) projected gradient as detailed in Section (5.2.3).

Example 5.25. Multiple Measurement Vectors (MMV)

Multiple Measurement Vectors (a variant of which is known as the group Lasso [?, ?]) is a sparse Least Squares problem that enforces *fully-zero columns* in the recovered sparse assignment matrix through the use of the ℓ_1/ℓ_2 norm

$$\underset{\mathbf{X}}{\text{minimize}} \ \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F^2 + \lambda \|\mathbf{X}\|_{1,2} \quad . \tag{5.88}$$

A proximal gradient step then takes the form

$$\mathbf{X}^k = \text{prox}_{1,2} \left[\mathbf{X}^{k-1} - \frac{1}{\|\mathbf{A}\|_2^2} \mathbf{A}^T \left(\mathbf{A}\mathbf{X}^{k-1} - \mathbf{B} \right) \right]$$

where $\text{prox}_{1,2}(\cdot)$ is the proximal operator of the ℓ_1/ℓ_2 norm as described in Section 5.3.