Chapter 3

Minimizing Sums of Smooth and Simple Functions

In this chapter, we study minimization of the sum of a 'simple' and a smooth function. This modeling mechanism lets us incorporate prior information about the decision variable, including structure (e.g. sparsity or smoothness), and the feasible region (e.g. non-negativity or box constraints). The simple function must be convex, but is allowed to be non-smooth, and in particular can take on infinite values.

First-order methods are easily modified to account for the simple term. The modifications preserve the rates of convergence from Chapter 2, and can be analyzed using analogous techniques to those already presented. We gain flexibility at essentially no computational cost. We start with a few motivating examples, and then provide the analysis.

Example 3.1 (Optimization with Simple Constraints)**.** Consider a smooth model $f(x)$ from Chapter 2, e.g. any learning problem arising from a general linear model. Suppose you are also given side information about the domain of the predictors x. For example:

- some components of x are non-negative
- some components of x have lower and upper bounds
- x must be in the level set of some convex function, e.g. $\|x\|_2 \leq \tau$.
- x must be in a certain affine subspace, e.g. $Ax = b$.

All of these constraints can be concisely written as $x \in C$, where C is a closed convex set. The modified optimization problem is then

$$\min f(x) + \delta\left(x \mid C\right),$$

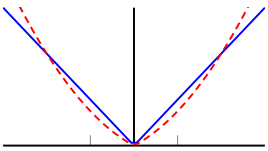

Figure 3.1: 1-norm (blue) and elastic net (red dashed) both have nonsmooth behavior at the origin.

where

$$\delta\left(x \mid C\right) := \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}.$$

is called the *convex indicator function* of C. We consider $\delta\left(\cdot \mid C\right)$ 'simple' when C admits an efficiently computable projection:

$$\mathrm{proj}_C(z) = \operatorname*{argmin}_{x \in C} \frac{1}{2}\|x - z\|^2.$$

Example 3.2 (Sparse regularization)**.** The notion of *sparsity* is fundamental to modern numerical analysis. Analogously to matrix sparsity, 'x is sparse' means either that most $x_i = 0$, or that the magnitudes of $|x_i|$ are quickly decaying. Modelers exploit sparsity in a range of settings.

1. **Compressive sensing.** Many signals are sparse in particular transform domains. For example, superpositions of periodic signals have a sparse Fourier representation. If a typical photograph is represented using *wavelets*, the magnitudes of the wavelet coefficients decay rapidly. Wavefields generated by earthquakes can be efficiently represented using *curvelets*. Applications such as image denoising and deblurring, seismic inverse problems, and image compression benefit from these ideas. The problems are captured by the formulation

   $$\min_x \|b - AWx\|^2 + r(x),$$

   where A is a specially designed measurement matrix, typically with far fewer rows than columns, W is the transform where the signal of interest admits a sparse representation (e.g. Fourier, wavelets or curvelets), and $r(\cdot)$ is a non-smooth function that promotes sparsity of the input. Two common convex examples are $r(x) = \|x\|_1$, and $r(x) = \alpha\|x\|_1 + (1-\alpha)\|x\|^2$, known as the *elastic net*, see Figure 3.1. The curvature of the elastic net helps it find groups of correlated predictors in practice.

2. **Statistical learning problems.** We cannot expect that general learning problems will have sparse solutions x. However, for many

models, we want to discover the most important predictors. We can therefore consider the parametrized family of solutions

$$x(\lambda) \in \arg\min_x f(x) + \lambda r(x),$$

with $r(x)$ a nonsmooth regularizer. When λ is larger than $\|\nabla f(0)\|_\infty$, $x(\lambda) = 0$. As λ decreases, x_i 'activate'. The earliest activated entries can indicate the most important predictors. This kind of analysis is known as the Lasso, and is used in conjunction with all general linear models.

Example 3.3 (More non-smooth regularizers)**.** While the 1-norm penalty is ubiquitously used to promote sparsity, many other related regularizers are also used in a range of learning and inverse problems.

- The *OWL norm* $r(x) = \alpha\|x\|_1 + (1-\alpha)\|x\|_\infty$ can detect groups of correlated predictors even better than the elastic net.
- The *group lasso* penalty $r(x) = \sum_j \|x_j\|$ forces pre-specified groups of indices x_j to be jointly included or excluded.
- The *total variation* penalty $r(x) = \|Dx\|_1$, gives piecewise constant signals along directions determined by differential operator D.

Example 3.4 (Sparse covariance estimation)**.** Suppose we are given a symmetric positive definite sample covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$. Its inverse F is the (Fisher) information matrix, and the equality $F_{ij} = 0$ implies conditional independence of variables i and j. The *graphical Lasso* problem looks for sparse information by solving the problem

$$\min_{X\geq 0} \left\{\log\det(X) + \operatorname{tr}(\Sigma X) + \lambda\|X\|_1\right\}.$$

Example 3.5 (Convex matrix completion)**.** Suppose we observe some entries a_{ij} of a large matrix $A \in \mathbf{R}^{m\times n}$, with ij ranging over some small index set $\mathcal{I}$, and wish to to recover A (i.e. fill in the missing entries). A classic approach is to penalize the *nuclear norm*, leading to the problem

$$\min_X \ \tfrac{1}{2}\sum_{ij\in\mathcal{I}} \|X_{ij} - a_{ij}\|^2 + \|X\|_*.$$

Compare this formulation to the smooth factorization approach.

Example 3.6 (Portfolio Estimation with Simplex Constraints)**.** Markowitz portfolio estimation is a foundational topic in computational finance. Given a set of N stocks, we consider their returns over T time steps, encoded by a matrix $F \in \mathbb{R}^{N\times T}$. From this information it is straightforward to compute a vector of mean returns $\mu \in \mathbb{R}^N$ and a covariance matrix $\Sigma \in \mathbb{R}^{N\times N}$,

assuming the returns process is stationary. We want to choose investment weights for the N assets to minimize a measure of risk for a given return α. A common risk measure is the variance of the portfolio, $w^T \Sigma w$, so we have

$$\min_{w \in \Delta} w^T \Sigma w \quad \text{such that} \quad w^T \mu = \alpha.$$

The set $\Delta = \{w : w_i \in [0,1], 1^T w_i = 1\}$ is the unit simplex, which forces purchases must be non-negative (no shorting) and investment of all assets (one asset is typically a 'safe' option such as a bond or index fund).

3.1 Proximal Gradient Method

Consider the problem

$$\min_x f(x) = g(x) + h(x),$$

with g, h convex and g a β-smooth map. Analogously to steepest descent, we can design an iterative method by minimizing a simple upper bound obtained from g:

$$\begin{aligned} x^+ &= \operatorname*{argmin}_y g(x) + \langle \nabla g(x), y - x \rangle + \frac{\beta}{2}\|y - x\|^2 + h(y) \\ &= \operatorname*{argmin}_y \frac{\beta}{2}\|y - (x - \beta^{-1}\nabla g(x))\|^2 + h(y) \end{aligned}$$

Minimizing the sum of $h(y)$ and a small quadratic can be viewed as an atomic operation.

Definition 3.7 (Proximity Operator)**.** For a convex function $h(y) : \mathbb{R}^n \to \mathbb{R} \cup \infty$, define the *proximity* operator $\text{prox}_{\alpha h} : \mathbb{R}^n \to \mathbb{R}^n$ by

$$\operatorname*{prox}_{\gamma h}(z) = \operatorname*{argmin}_x \frac{1}{2\gamma}\|x - z\|^2 + h(x).$$

Note that the optimization problem defining $\text{prox}_{\gamma h}$ is strongly convex, so the solution is unique. The iteration for x^+ can therefore be written more compactly as

$$x^+ = \operatorname*{prox}_{\beta^{-1} h}(x - \beta^{-1}\nabla g(x)).$$

To analyze this algorithm, we introduce the proximal gradient map

$$G_t(x) := \frac{1}{t}\left(x - \text{prox}_{th}(x - t\nabla g(x))\right),$$

which behaves similarly to the gradient of a smooth function. For example, the proximal gradient iteration is written

$$x^+ = x - \beta^{-1} G_{\beta^{-1}}(x).$$

To understand the map G and its consequences, we first need to extend the notion of derivative to nonsmooth convex functions.

Definition 3.8 (Subgradient and Subdifferential)**.** Let $h : U \to \mathbb{R}$ be a convex function. A *subgradient* of h at x is a vector $v \in \mathbb{R}^n$ that satisfies

$$h(y) \geq h(x) + \langle v, y - x\rangle \quad \text{for all } y \in U.$$

The *subdifferential* of h at x is the set of all subgradients, and is denoted by $\partial h(x)$. Equivalently,

$$\partial h(x) := \{v \in \mathbb{R}^n : h(y) \geq h(x) + \langle v, y - x\rangle \text{ for all } y \in U.\}$$

When there is only one point in $\partial h(x)$, then h is differentiable at x and $\partial h(x) = \nabla f(x)$. If $0 \in \partial h(x)$, then immediately from the definition we have $h(y) \geq h(x)$ for all $y \in U$ so x must be a global minimizer.

Example 3.9 (Subdifferential of $\|\cdot\|_1$)**.** Suppose $h(x) = |x|$. Then

$$\partial h(x) = \begin{cases} \{1\} & \text{if } x > 0 \\ [-1, 1] & \text{if } x = 0 \\ \{-1\} & \text{if } x < 0 \end{cases}$$

Since $\|x\|_1 = \sum |x_i|$, the subdifferential of the 1-norm can be computed by applying the above formula to each coordinate.

Example 3.10 (Subdifferential of an indicator function)**.** Suppose $h(x) = \delta(x \mid C)$ where C is a closed convex set. If $x \in C$, $v \in \partial h(x)$ is characterized by

$$\delta(y \mid C) \geq \langle v, y - x\rangle + \delta(x \mid C) = \langle v, y - x\rangle.$$

This inequality always holds for $y \notin C$; if $y \in C$, it gives $0 \geq \langle v, y - x\rangle$. Therefore

$$\partial h(x) = \{v : 0 \geq \langle v, y - x\rangle \quad \text{for all } y \in C\},$$

which is called the *normal cone* to C at x.

Coming back to the map G, we show it is analogous to the gradient map in the smooth case.

Remark 3.11. $G_t(x) - \nabla g(x) \in \partial h(x^+)$, where $x^+ = \text{prox}_{th}(x - t\nabla g(x))$

Proof. Observe

$$x^+ = \underset{u}{\text{argmin}} \ \{g(x) + \langle \nabla g(x), u - x\rangle + \frac{1}{2t}\|u - x\|^2 + h(u)\}$$

Then differentiating the RHS of the above expression with respect to u at $u = x^+$ gives

$$0 \in \nabla g(x) + \underbrace{\frac{1}{t}(x^+ - x)}_{G_t(x)} + \partial h(x^+)$$

That is,

$$G_t(x) - \nabla g(x) \in \partial h(x^+).$$

□

It immediately follows that $G_t(x) = 0$ if and only if x minimizes $g + h$.

Theorem 3.12. *Suppose that g is β-smooth and α-convex, where α can be 0, and define $x^+ := \mathrm{prox}_{th}(x - t\nabla g(x))$, and assume that h is convex. Then we have*

$$f(y) \geq f(x^+) + \langle G_t(x), y - x\rangle + t\left(1 - \frac{\beta t}{2}\right)\|G_t(x)\|^2 + \frac{\alpha}{2}\|y - x\|^2. \quad (3.1)$$

Proof.

$$\begin{aligned}
f(x^+) &= g(x - tG_t(x)) + h(x^+) \\
&\leq g(x) - t\langle \nabla g(x), G_t(x)\rangle + \frac{\beta t^2}{2}\|G_t(x)\|^2 + h(x^+) \\
&\leq g(y) + \langle x - y, \nabla g(x)\rangle - \frac{\alpha}{2}\|y - x\|^2 - t\langle \nabla g(x), G_t(x)\rangle + \frac{\beta t^2}{2}\|G_t(x)\|^2 h(x^+) \\
&= g(y) + \langle x^+ - y, \nabla g(x)\rangle - \frac{\alpha}{2}\|y - x\|^2 + \frac{\beta t^2}{2}\|G_t(x)\|^2 + h(x^+) \\
&\leq f(y) + \langle x^+ - y, \nabla g(x)\rangle - \frac{\alpha}{2}\|y - x\|^2 + \frac{\beta t^2}{2}\|G_t(x)\|^2 + \langle G_t(x) - \nabla g(x), x^+ - y\rangle \\
&\leq f(y) + \langle x^+ - y, G_t(x)\rangle - \frac{\alpha}{2}\|y - x\|^2 + \frac{\beta t^2}{2}\|G_t(x)\|^2 \\
&= f(y) - \langle y - x, G_t(x)\rangle - \frac{\alpha}{2}\|y - x\|^2 - \langle x - x^+, G_t(x)\rangle + \frac{\beta t^2}{2}\|G_t(x)\|^2 \\
&= f(y) - \langle y - x, G_t(x)\rangle - \frac{\alpha}{2}\|y - x\|^2 - \left(t - \frac{\beta t^2}{2}\right)\|G_t(x)\|^2.
\end{aligned}$$

□

Remarks:

1. If $\alpha = 0$, taking $t = \frac{1}{\beta}$ and $y = x$, we have

$$f(x^+) \leq f(x) - \frac{1}{2\beta}\|G_t(x)\|^2.$$

2. Letting $y = x^*$ and $t = \frac{1}{\beta}$, we have

$$0 \geq f(x^+) - f(x^*) + \langle G_t(x), x^* - x\rangle + \frac{1}{2t}\|G_t(x)\|^2 + \frac{\alpha}{2}\|x^* - x\|^2$$

and in particular

$$\langle G_t(x), x - x^*\rangle \geq \frac{1}{2t}\|G_t(x)\|^2 + \frac{\alpha}{2}\|x^* - x\|^2$$

Rate for convex problems: The proximal gradient method with $\frac{1}{\beta}$ step satisfies

$$f(x_k) - f(x^*) \leq \frac{\beta}{2k}\|x_1 - x^*\|^2.$$

The key inequality is

$$f(x_{k+1}) - f(x^*) \leq -\langle G_t(x), x^* - x_k\rangle - \frac{1}{2\beta}\|G_t(x_k)\|^2 \leq \frac{\beta}{2}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right).$$

Exercise 3.13. Derive the above inequality.

We immediately have Theorem 2.25 for the prox-gradient method.

Theorem 3.14 (Prox-gradient descent and convexity)**.** *Suppose that $h(x) = f(x) + g(x)$, with $f\colon \mathbf{E} \to \mathbf{R}$ is convex and β-smooth, and g convex. Then the iterates generated by the prox-gradient descent method satisfy*

$$f(x_k) - f^* \leq \frac{\beta\|x_0 - x^*\|^2}{2k}.$$

Rate for strongly convex problems: If in addition f is α-convex, then

$$f(x_{k+1}) - f(x^*) \leq \frac{\beta}{2}\left(1 - \frac{\alpha}{\beta}\right)^k \|x_1 - x^*\|^2$$

and

$$\|x_{k+1} - x^*\|^2 \leq \left(1 - \frac{\alpha}{\beta}\right)^k \|x_1 - x^*\|^2.$$

Proof :

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 - 2t\,\langle G_t(x_k), x_k - x^*\rangle + t^2\|G_t(x_k)\|^2 \\
&\leq \|x_k - x^*\|^2 - 2t\left(\frac{t}{2}\|G_t(x_k)\|^2 + \frac{\alpha}{2}\|x_k - x^*\|^2\right) + t^2\|G_t(x_k)\|^2 \\
&= \|x_k - x^*\|^2 - \frac{\alpha}{\beta}\|x_k - x^*\|^2
\end{aligned}$$

Again, we can immediately state a theorem analogous to Theorem 3.15.

Theorem 3.15 (Prox-gradient descent and strong convexity)**.**
Suppose that $h(x) = f(x) + g(x)$, with $f\colon \mathbf{E} \to \mathbf{R}$ is α-strongly convex and β-smooth, and g convex. Then the iterates generated by the proximal gradient descent method satisfy

$$\|x_k - x^*\|^2 \leq \left(\frac{Q-1}{Q+1}\right)^k \|x_0 - x^*\|^2,$$

where $Q := \beta/\alpha$ is the condition number of f.

In other words, adding a 'prox-friendly' convex function g preserves the rates of first order methods for f alone. If g is strongly convex but f is not, proximal gradient still has a linear rate (see the exercises).

Exercise 3.16. Show that if g is α_2-convex, then (3.1) can be strengthened to

$$f(y) \geq f(x^+)+(1+t\alpha_2)\langle G_t(x), y-x\rangle+t\left(1-\frac{\beta t+\alpha_2 t}{2}\right)\|G_t(x)\|^2+\frac{\alpha+\alpha_2}{2}\|y-x\|^2.$$

Exercise 3.17. Show that if g is α_2-convex, then with step $t = \frac{1}{\beta+\alpha_2}$ in Remark 2 we have

$$0 \geq f(x^+) - f(x^*) + \langle G_t(x), x^* - x\rangle + \frac{1}{2t}\|G_t(x)\|^2 + \frac{\alpha+\alpha_2}{2}\|x^* - x\|^2$$

and in particular

$$\langle G_t(x), x - x^*\rangle \geq \frac{1}{2t}\|G_t(x)\|^2 + \frac{\alpha+\alpha_2}{2}\|x^* - x\|^2$$

Exercise 3.18. State and prove the convergence rate under the additional assumption that g is α_2 convex.