Convex Optimization in Signal Processing and Communications

March 13, 2009

Contents

List of contributors

Amir Beck

Amir Beck was born in Israel in 1975. He received the B. Sc. degree in pure mathematics (*cum laude*) in 1991, the M.Sc. degree in operations research (*suma cum laude*), and the Ph.D degree in operations research—all from Tel-Aviv University (TAU), Tel-Aviv, israel. From 2003-2005, he was a Postdoctoral Fellow at the Minerva Optimization Center, Technion, Haifa, Israel. He is currently a Senior Lecturer in the Department of Industrial Engineering at the Technion—Israel Institute of Technology, Haifa, Israel. His research interests are in continuous optimization, in particular, algorithms for large-scale problems, convex optimization, as well as nonconvex quadratic optimization: theory, algorithms and applications in signal processing, image processing and communication. His research has been supported by various funding agencies including, the Israel Science Foundation, the German-Israeli Foundation and the 6-th framework programme of the European Commission.

Marc Teboulle

Marc Teboulle received his D.Sc. degree from the Technion, Israel Institute of Technology in 1985. He has held a position of applied mathematician at the Israel Aircraft Industries, and academic appointements at Dalhousie University, Canada, and the University of Maryland, USA. From 1999 to 2002, he served as Chairman of the Department of Statistics and Operations Research at the School of Mathematical Sciences of Tel-Aviv University. He is currently a Professor at the School of Mathematical Sciences of Tel-Aviv University, Israel. His research interests are in the area of continuous optimization, including theory, algorithmic analysis and its applications. He has published numerous papers and two books, and has given invited lectures at many international conferences. His research has been supported by various funding agencies including, the National Science Foundation, the French-Israeli Ministry of Sciences, the Bi-National Israel-United States Science Foundation, and the Israel Science Foundation. He currently serves on the editorial board of Mathematics of Operations Research; the European Series in Applied and Industrial Mathematics, Control, Optimisation and Calculus of Variations COCV, and is the Area Editor of Optimization for APJOR.

Part I

1 Gradient-Based Algorithms with Applications to Signal Recovery Problems

Amir Beck and Marc Teboulle

Amir Beck is with the Technion - Israel Institute of Technology, Haifa, Israel.
Marc Teboulle is with the Tel-Aviv University, Tel-Aviv, Israel.

This chapter presents in a self-contained manner recent advances in the design and analysis of gradient-based schemes for specially structured smooth and nonsmooth minimization problems. We focus on the mathematical elements and ideas for building fast gradient-based methods and derive their complexity bounds. Throughout the chapter, the resulting schemes and results are illustrated and applied on a variety of problems arising in several specific key applications such as sparse approximation of signals, total variation-based image processing problems, and sensor location problems.

1.1 Introduction

The gradient method is probably one of the oldest optimization algorithms going back as early as 1847 with the initial work of Cauchy. Nowadays, gradient-based methods[1] have attracted a revived and intensive interest among researches both in theoretical optimization and in scientific applications. Indeed, the very large-scale nature of problems arising in many scientific applications, combined with an increase power of computer technology have motivated a "return" to the "old and simple" methods that can overcome the curse of dimensionality, a task which is usually out of reach for the current more sophisticated algorithms.

One of the main drawbacks of gradient-based methods is their speed of convergence, which is known to be slow. However, with proper modeling of the problem at hand, combined with some key ideas, it turns out that it is possible to build fast gradient schemes for various classes of problems arising in applications and in particular signal recovery problems.

The purpose of this chapter is to present in a self-contained manner such recent advances. We focus on the essential tools needed to build and analyze

[1] We also use the term "gradient" instead of "subgradient" in case of nonsmooth functions.

fast gradient schemes, and present successful applications to some key scientific problems. To achieve these goals our emphasis will focus on:

- Optimization models/formulations.
- Building approximation models for gradient schemes.
- Fundamental mathematical tools for convergence and complexity analysis.
- Fast gradient schemes with better complexity.

On the application front, we review some recent and challenging problems that can benefit from the above theoretical and algorithmic framework and we include gradient-based methods applied to:

- Sparse approximation of signals.
- Total variation-based image processing problems.
- Sensor location problems.

The contents and organization of the chapter is well summarized by the two lists of items above. We will strive to provide a broad picture of the current research in this area as well as to motivate further research within the gradient-based framework.

To avoid cutting the flow of the chapter, we refrain from citing references within the text. Rather, the last section of this chapter includes bibliographical notes. While we did not attempt to give a complete bibliography on the covered topics (which is very large), we did try to include earlier works and influential papers, to cite all the sources for the results we used in this chapter, and to indicate some pointers on very recent developments that hopefully will motivate further research in the field. We apologize in advance for any possible omission.

1.2 The General Optimization Model

1.2.1 Generic Problem Formulation

Consider the following generic optimization model:

$$\text{(M)} \quad \min \{F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x}) : x \in \mathbb{E}\},$$

where

- $\mathbb{E}$ is a finite dimensional Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \| = \langle \cdot, \cdot \rangle^{1/2}$.
- $g : \mathbb{E} \to (-\infty, \infty]$ is a proper closed and convex function which is assumed subdifferentiable over $\operatorname{dom} g$.[2]
- $f : \mathbb{E} \to (-\infty, \infty)$ is a continuously differentiable function over $\mathbb{E}$.

[2] Throughout this paper all necessary notations/definitions/results from convex analysis not explicitly given are standard and can be found in the classical monograph [51].

The model (M) is rich enough to recover generic classes of smooth/nonsmooth convex minimization problems as well as smooth nonconvex problems. This is illustrated in the following examples.

Example 1.1: (a) Convex minimization problems.
Pick $f \equiv 0$ and $g = h_0 + \delta_C$ where $h_0 : \mathbb{E} \to (-\infty, \infty)$ is a convex function (possibly nonsmooth) and δ_C is the indicator function defined by

$$\delta_C(\mathbf{x}) = \begin{cases} 0 & \mathbf{x} \in C, \\ \infty & \mathbf{x} \notin C, \end{cases}$$

where $C \subseteq \mathbb{E}$ is a closed and convex set. The model (M) reduces to the generic convex optimization problem

$$\min \{h_0(\mathbf{x}) : \mathbf{x} \in C\}.$$

In particular, if C is described by convex inequality constraints, i.e., with

$$C = \{\mathbf{x} \in \mathbb{E} : h_i(\mathbf{x}) \leq 0, i = 1, \ldots, m\},$$

where h_i are some given proper closed convex functions on $\mathbb{E}$, we recover the functional form of the convex program:

$$\min \{h_0(\mathbf{x}) : h_i(\mathbf{x}) \leq 0, i = 1, \ldots, m\}.$$

(b) Smooth constrained minimization
Set $g = \delta_C$ with $C \subseteq \mathbb{E}$ being a closed convex set. Then (M) reduces to the problem of minimizing a smooth (possibly nonconvex) function over C, i.e.,

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

A more specific example that can be modelled by (M) is from the field of signal recovery and is now described.

1.2.2 Signal Recovery via Nonsmooth Regularization

A basic linear inverse problem is to estimate an unknown signal $\mathbf{x}$ satisfying the relation

$$\mathbf{A}\mathbf{x} = \mathbf{b} + \mathbf{w},$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$ are known, and $\mathbf{w}$ is an unknown noise vector. The basic problem is then to recover the signal $\mathbf{x}$ from the noisy measurements $\mathbf{b}$. A common approach for this estimation problem is to solve the *regularized least squares* (RLS) minimization problem

$$\text{(RLS)} \quad \min_{\mathbf{x}} \left\{ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \lambda R(\mathbf{x}) \right\}, \tag{1.1}$$

where $\|\mathbf{Ax}-\mathbf{b}\|^2$ is a least squares term that measures the distance between $\mathbf{b}$ and $\mathbf{Ax}$ in an l_2 norm sense[3], $R(\cdot)$ is a convex regularizer used to stabilize the solution, and $\lambda > 0$ is a regularization parameter providing the trade off between fidelity to measurements and noise sensitivity. Model (RLS) is of course a special case of model (M) by setting $f(\mathbf{x}) \equiv \|\mathbf{Ax}-\mathbf{b}\|^2$ and $g(\mathbf{x}) \equiv \lambda R(\mathbf{x})$.

Popular choices for $R(\cdot)$ are dictated from the application in mind and include for example the following model:

$$R(\mathbf{x}) = \sum_{i=1}^{s} \|\mathbf{L}_i\mathbf{x}\|_p^p, \tag{1.2}$$

where $s \geq 1$ is an integer number, $p \geq 1$ and $\mathbf{L}_i : \mathbb{R}^n \to \mathbb{R}^{d_i}$ ($d_1, \ldots, d_s$ being positive integers) are linear maps. Of particular interest for signal processing applications are the following cases:

1. **Tikhonov regularization**. By setting $s = 1, \mathbf{L}_i = \mathbf{L}, p = 2$, we obtain the standard Tikhonov regularization problem:
$$\min_{\mathbf{x}} \|\mathbf{Ax}-\mathbf{b}\|^2 + \lambda\|\mathbf{Lx}\|^2.$$
2. l_1 **regularization.** By setting $s = 1, \mathbf{L}_i = \mathbf{I}, p = 1$ we obtain the l_1 regularization problem
$$\min_{\mathbf{x}} \|\mathbf{Ax}-\mathbf{b}\|^2 + \lambda\|\mathbf{x}\|_1.$$
Other closely related problems include for example,
$$\min\{\|\mathbf{x}\|_1 : \|\mathbf{Ax}-\mathbf{b}\|^2 \leq \epsilon\} \text{ and } \min\{\|\mathbf{Ax}-\mathbf{b}\|^2 : \|\mathbf{x}\|_1 \leq \epsilon\}.$$
The above are typical formulations in statistic regression (LASSO, basis pursuit) as well as in the emerging technology of compressive sensing.
3. **Wavelet-based regularization.** By choosing $p = 1, s = 1, \mathbf{L}_i = \mathbf{W}$ where $\mathbf{W}$ is a wavelet transform matrix, we recover the wavelet-based regularization problem
$$\min_{\mathbf{x}} \|\mathbf{Ax}-\mathbf{b}\|^2 + \lambda\|\mathbf{Wx}\|_1.$$
4. **TV-based regularization.** When the set $\mathbb{E}$ is the set of $m \times n$ real-valued matrices representing the set of all $m \times n$ images, it is often the case that one chooses the regularizer to be a total variation function which has the form
$$R(\mathbf{x}) \equiv \mathrm{TV}(\mathbf{x}) = \sum_{i=1}^{m}\sum_{j=1}^{n} \|(\nabla\mathbf{x})_{i,j}\|.$$
A more precise definition of R will be given in Section 1.7.

[3] Throughout the chapter, unless otherwise stated, the norm $\|\cdot\|$ stands for the Euclidean norm associated with $\mathbb{E}$.

Note that the last three examples deal with nonsmooth regularizers. The reason for using such seemingly difficult regularization functions and not the more standard smooth quadratic Tikhonov regularization will be explained in Sections 1.6 and 1.7.

In the forthcoming sections, many of these problems will be described in more detail and gradient-based methods will be the focus of relevant schemes for their solution.

1.3 Building Gradient-Based Schemes

In this section we describe the elements needed to generate a gradient-based method for solving problems of the form (M). These rely mainly on building "good approximation models" and on applying fixed point methods on corresponding optimality conditions. In Sections 1.3.1 and 1.3.2 we will present two different derivations—corresponding to the two building techniques—of a method called the *proximal gradient* algorithm. We will then show in Section 1.3.3 the connection to the so-called majorization-minimization approach. Finally, in Section 1.3.4 we will explain the connection of the devised techniques to Weiszfeld's method for the Fermat-Weber location problem.

1.3.1 The Quadratic Approximation Model for (M)

Let us begin with the simplest unconstrained minimization problem of a continuously differentiable function f on $\mathbb{E}$ (i.e., we set $g \equiv 0$ in (M)):

$$\text{(U)} \quad \min\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}.$$

The well known basic gradient method generates a sequence $\{\mathbf{x}_k\}$ via

$$\mathbf{x}_0 \in \mathbb{E}, \quad \mathbf{x}_k = \mathbf{x}_{k-1} - t_k \nabla f(\mathbf{x}_{k-1}) \quad (k \geq 1), \tag{1.3}$$

where $t_k > 0$ is a suitable step-size. The gradient method thus takes at each iteration a step along the negative gradient direction, which is the direction of "steepest descent". This interpretation of the method, although straightforward and natural, cannot be extended to the more general model (M). Another simple way to interpret the above scheme is via an approximation model that would replace the original problem (U) with a "reasonable" approximation of the objective function. The simplest idea is to consider the quadratic model

$$q_t(\mathbf{x}, \mathbf{y}) := f(\mathbf{y}) + \langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y}) \rangle + \frac{1}{2t} \|\mathbf{x} - \mathbf{y}\|^2, \tag{1.4}$$

namely, the linearized part of f at some given point $\mathbf{y}$, regularized by a quadratic proximal term that would measure the "local error" in the approximation, and also results in a well defined, i.e., a strongly convex approximate minimization

problem for (U):

$$(\hat{U}_t) \quad \min\{q_t(\mathbf{x},\mathbf{y}) : \mathbf{x} \in \mathbb{E}\}.$$

For a fixed given point $\mathbf{y} := \mathbf{x}_{k-1} \in \mathbb{E}$, the unique minimizer $\mathbf{x}_k$ solving $(\hat{U}_{t_k})$ is

$$\mathbf{x}_k = \operatorname{argmin}\{q_{t_k}(\mathbf{x},\mathbf{x}_{k-1}) : \mathbf{x} \in \mathbb{E}\},$$

which yields the same gradient scheme (1.3).

Simple algebra also shows that (1.4) can be written as,

$$q_t(\mathbf{x},\mathbf{y}) = \frac{1}{2t}\|\mathbf{x} - (\mathbf{y} - t\nabla f(\mathbf{y}))\|^2 - \frac{t}{2}\|\nabla f(\mathbf{y})\|^2 + f(\mathbf{y}). \tag{1.5}$$

Using the above identity also allows us to easily pass from the unconstrained minimization problem (U) to an approximation model for the constrained model

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

where $C \subseteq \mathbb{E}$ is a given closed convex set. Ignoring the constant terms in (1.5) leads us to solve (P) via the scheme

$$\mathbf{x}_k = \operatorname*{argmin}_{\mathbf{x}\in C} \frac{1}{2}\left\|\mathbf{x} - (\mathbf{x}_{k-1} - t_k\nabla f(\mathbf{x}_{k-1}))\right\|^2, \tag{1.6}$$

which is the so-called gradient projection method (GPM):

$$\mathbf{x}_k = \Pi_C(\mathbf{x}_{k-1} - t_k\nabla f(\mathbf{x}_{k-1})).$$

Here Π_C denotes the orthogonal projection operator defined by

$$\Pi_C(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{z}\in C} \|\mathbf{z} - \mathbf{x}\|^2.$$

Turning back to our general model (M), one could naturally suggest to consider the following approximation in place of $f(\mathbf{x}) + g(\mathbf{x})$:

$$q(\mathbf{x},\mathbf{y}) = f(\mathbf{y}) + \langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y})\rangle + \frac{1}{2t}\|\mathbf{x} - \mathbf{y}\|^2 + g(\mathbf{x}).$$

That is, we leave the nonsmooth part $g(\cdot)$ untouched.

Indeed, in accordance with the previous framework, the corresponding scheme would then read:

$$\mathbf{x}_k = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ g(\mathbf{x}) + \frac{1}{2t_k}\|\mathbf{x} - (\mathbf{x}_{k-1} - t_k\nabla f(\mathbf{x}_{k-1}))\|^2 \right\}. \tag{1.7}$$

In fact, the latter leads to another interesting way to write the above scheme via the fundamental proximal operator. For any scalar $t > 0$, the proximal map associated with g is defined by

$$\operatorname{prox}_t(g)(\mathbf{z}) = \operatorname*{argmin}_{\mathbf{u}\in\mathbb{E}} \left\{ g(\mathbf{u}) + \frac{1}{2t}\|\mathbf{u} - \mathbf{z}\|^2 \right\}. \tag{1.8}$$

With this notation, the scheme (1.7), which consists of a proximal step at a resulting gradient point will be called the *proximal gradient* method, and reads

as:

$$\mathbf{x}_k = \mathrm{prox}_{t_k}(g)(\mathbf{x}_{k-1} - t_k \nabla f(\mathbf{x}_{k-1})). \tag{1.9}$$

An alternative and useful derivation of the proximal gradient method is via the *fixed point approach* developed next.

1.3.2 The Fixed Point Approach

Consider the nonconvex and nonsmooth optimization model (M). If $\mathbf{x}^* \in \mathbb{E}$ is a local minimum of (M), then it is a stationary point of (M), i.e., one has

$$\mathbf{0} \in \nabla f(\mathbf{x}^*) + \partial g(\mathbf{x}^*), \tag{1.10}$$

where $\partial g(\cdot)$ is the subdifferential of g. Note that whenever f is also convex, the latter condition is necessary and sufficient for $\mathbf{x}^*$ to be a global minimum of (M).

Now, fix any $t > 0$, then (1.10) holds if and only if the following equivalent statements hold:

$$\begin{aligned} \mathbf{0} &\in t\nabla f(\mathbf{x}^*) + t\partial g(\mathbf{x}^*), \\ \mathbf{0} &\in t\nabla f(\mathbf{x}^*) - \mathbf{x}^* + \mathbf{x}^* + t\partial g(\mathbf{x}^*), \\ (I + t\partial g)(\mathbf{x}^*) &\in (I - t\nabla f)(\mathbf{x}^*), \\ \mathbf{x}^* &= (I + t\partial g)^{-1}(I - t\nabla f)(\mathbf{x}^*), \end{aligned}$$

where the last relation is an equality (and not an inclusion) thanks to the properties of the proximal map (c.f. the first part of Lemma 1.2 below). The last equation naturally calls for the *fixed point scheme* that generates a sequence $\{\mathbf{x}_k\}$ via:

$$\mathbf{x}_0 \in \mathbb{E}, \quad \mathbf{x}_k = (I + t_k \partial g)^{-1}(I - t_k \nabla f)(\mathbf{x}_{k-1}) \quad (t_k > 0). \tag{1.11}$$

Using the identity $(I + t_k \partial g)^{-1} = \mathrm{prox}_{t_k}(g)$ (c.f., first part of Lemma 1.2), it follows that the scheme (1.11) is nothing else but the proximal gradient method devised in Section 1.3.1. Note that the scheme (1.11) is in fact a special case of the so-called *proximal backward-forward* scheme, which was originally devised for finding a zero of the more general inclusion problem:

$$\mathbf{0} \in T_1(\mathbf{x}^*) + T_2(\mathbf{x}^*),$$

where T_1, T_2 are maximal monotone set valued maps (encompassing (1.10) with f, g convex and $T_1 := \nabla f, T_2 := \partial g$).

1.3.3 Majorization-Minimization Technique

The Idea

A popular technique to devise gradient-based methods in the statistical and engineering literature is the MM approach where the first M stands for majorization

and the second M for minimization[4] (maximization problems are similarly handled with minorization replacing majorization)

The MM technique follows in fact from the same idea of approximation models described in Section 1.3.1, except that the approximation model in the MM technique does not have to be quadratic.

The basic idea of MM relies on finding a "relevant" approximation to the objective function F of model (M) that satisfies:

(i) $M(\mathbf{x},\mathbf{x}) = F(\mathbf{x})$ for every $\mathbf{x} \in \mathbb{E}$.
(ii) $M(\mathbf{x},\mathbf{y}) \geq F(\mathbf{x})$ for every $\mathbf{x},\mathbf{y} \in \mathbb{E}$.

Geometrically, this means that $\mathbf{x} \mapsto M(\mathbf{x},\mathbf{y})$ lies above $F(\mathbf{x})$ and is tangent at $\mathbf{x} = \mathbf{y}$. From the above definition of $M(\cdot,\cdot)$, a natural and simple minimization scheme consists of solving

$$\mathbf{x}_k \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{E}} M(\mathbf{x},\mathbf{x}_{k-1}).$$

This scheme immediately implies that

$$M(\mathbf{x}_k,\mathbf{x}_{k-1}) \leq M(\mathbf{x},\mathbf{x}_{k-1}) \text{ for every } \mathbf{x} \in \mathbb{E}, \tag{1.12}$$

and hence from (i) and (ii) it follows that

$$F(\mathbf{x}_k) \overset{(ii)}{\leq} M(\mathbf{x}_k,\mathbf{x}_{k-1}) \overset{(1.12)}{\leq} M(\mathbf{x}_{k-1},\mathbf{x}_{k-1}) \overset{(i)}{=} F(\mathbf{x}_{k-1}) \text{ for every } k \geq 1,$$

thus naturally producing a descent scheme for minimizing problem (M).

Clearly, the key question is then how to generate a "good" upper bounding function $M(\cdot,\cdot)$ satisfying (i) and (ii). There does not exist a universal rule to determine the function M, and most often the structure of the problem at hand provides helpful hints to achieve this task. This will be illustrated below.

The MM Method for the RLS problem

An interesting example of the usage of MM methods is in the class of RLS problems described in Section 1.2.2. This example will also demonstrate the intimate relations between the MM approach and the basic approximation model.

Consider the RLS problem from Section 1.2.2 (problem (1.1)), that is, the general model (M) with $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2$ and $g(\mathbf{x}) = \lambda R(\mathbf{x})$. Since f is a quadratic function, easy algebra shows that for any $\mathbf{x},\mathbf{y}$:

$$f(\mathbf{x}) = f(\mathbf{y}) + 2\langle \mathbf{A}(\mathbf{x}-\mathbf{y}), \mathbf{A}\mathbf{y}-\mathbf{b}\rangle + \langle \mathbf{A}^T\mathbf{A}(\mathbf{x}-\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle.$$

Let $\mathbf{D}$ be any matrix satisfying $\mathbf{D} \succeq \mathbf{A}^T\mathbf{A}$. Then

$$f(\mathbf{x}) \leq f(\mathbf{y}) + 2\langle \mathbf{A}(\mathbf{x}-\mathbf{y}), \mathbf{A}\mathbf{y}-\mathbf{b}\rangle + \langle \mathbf{D}(\mathbf{x}-\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle,$$

[4] MM algorithms also appear under different terminology such as surrogate/transfer function approach and bound optimization algorithms.

and hence with

$$M(\mathbf{x},\mathbf{y}) := g(\mathbf{x}) + f(\mathbf{y}) + 2\langle \mathbf{A}(\mathbf{x}-\mathbf{y}), \mathbf{A}\mathbf{y}-\mathbf{b}\rangle + \langle \mathbf{D}(\mathbf{x}-\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle,$$

we have

$$M(\mathbf{x},\mathbf{x}) = F(\mathbf{x}) \text{ for every } \mathbf{x} \in \mathbb{E},$$
$$M(\mathbf{x},\mathbf{y}) \geq F(\mathbf{x}) \text{ for every } \mathbf{x},\mathbf{y} \in \mathbb{E}.$$

In particular, with $\mathbf{D} = \mathbf{I}$ ($\mathbf{I}$ being the identity matrix), the stated assumption on $\mathbf{D}$ reduces to $\lambda_{\max}(\mathbf{A}^T\mathbf{A}) \leq 1$ and a little algebra shows that M reduces in that case to

$$M(\mathbf{x},\mathbf{y}) = g(\mathbf{x}) + \|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 - \|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2.$$

The resulting iterative scheme is given by

$$\mathbf{x}_k = \operatorname*{argmin}_{\mathbf{x}} M(\mathbf{x},\mathbf{x}_{k-1}). \tag{1.13}$$

This scheme has been used extensively in the signal processing literature (see bibliographic notes) to devise convergent schemes for solving the RLS problem. A close inspection of the explanation above indicates that the MM approach for building iterative schemes to solve (RLS) is in fact equivalent to the basic approximation model discussed in Section 1.3.1. Indeed, opening the squares in $M(\cdot,\cdot)$ and collecting terms we obtain:

$$M(\mathbf{x},\mathbf{y}) = g(\mathbf{x}) + \|\mathbf{x} - \{\mathbf{y} - \mathbf{A}^T(\mathbf{A}\mathbf{y}-\mathbf{b})\}\|^2 + C(\mathbf{b},\mathbf{y}),$$

where $C(\mathbf{b},\mathbf{y})$ is constant with respect to $\mathbf{x}$. Since

$$\nabla f(\mathbf{y}) = 2\mathbf{A}^T(\mathbf{A}\mathbf{y}-\mathbf{b}),$$

it follows that (1.13) is just the scheme devised in (1.7) with constant step-size $t_k \equiv \frac{1}{2}$. Further ways to derive MM-based schemes that do not involve quadratic functions exploit tools and properties such as convexity of the objective; standard inequalities, e.g., Cauchy-Schwartz; topological properties of f, e.g., Lipschitz gradient; see the bibliographic notes.

1.3.4 Fermat-Weber Location Problem

In the early 17th century the French mathematician Pierre de Fermat challenged the mathematicians at the time (relax, this is not the "big" one!) with the following problem:

Fermat's problem: Given three points on the plane, find another point so that the sum of the distances to the existing points is minimum.

In the beginning of the 20th century, the German economist Weber studied an extension of this problem: Given n points on the plane, find another point such that the weighted sum of the Euclidean distances to these n points is minimal.

In mathematical terms, given m points $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$, we wish to find the location of $\mathbf{x} \in \mathbb{R}^n$ solving

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ f(\mathbf{x}) \equiv \sum_{i=1}^{m} \omega_i \|\mathbf{x} - \mathbf{a}_i\| \right\}.$$

In 1937, Weiszfeld proposed an algorithm for solving the Fermat-Weber problem. This algorithm, although not identical to the proximal gradient method, demonstrates well the two principles alluded in the previous sections for constructing gradient-based methods. On one hand, the algorithm can be viewed as a fixed point method employed on the optimality condition of the problem and on the other hand, each iteration can be equivalently constructed via minimization of a quadratic approximation of the problem at the previous iteration.

Let us begin with the first derivation of Weiszfeld's method. The optimality condition of the problem is $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Of course we may encounter problems if $\mathbf{x}^*$ happens to be one of the points $\mathbf{a}_i$, because $f(\mathbf{x})$ is not differentiable at these points, but for the moment, let us assume that this is not the case (see Section 1.8 for how to properly handle nonsmoothness). The gradient of the problem is given by

$$\nabla f(\mathbf{x}) = \sum_{i=1}^{m} \omega_i \frac{\mathbf{x} - \mathbf{a}_i}{\|\mathbf{x} - \mathbf{a}_i\|},$$

and thus after rearranging the terms, the optimality condition can be written as

$$\mathbf{x}^* \sum_{i=1}^{m} \omega_i \frac{1}{\|\mathbf{x}^* - \mathbf{a}_i\|} = \sum_{i=1}^{m} \omega_i \frac{\mathbf{a}_i}{\|\mathbf{x}^* - \mathbf{a}_i\|},$$

or equivalently as

$$\mathbf{x}^* = \frac{\sum_{i=1}^{m} \omega_i \frac{\mathbf{a}_i}{\|\mathbf{x}^* - \mathbf{a}_i\|}}{\sum_{i=1}^{m} \frac{\omega_i}{\|\mathbf{x}^* - \mathbf{a}_i\|}}.$$

Weiszfeld's method is nothing else but the fixed point iterations associated with the latter equation:

$$\mathbf{x}_k = \frac{\sum_{i=1}^{m} \omega_i \frac{\mathbf{a}_i}{\|\mathbf{x}_{k-1} - \mathbf{a}_i\|}}{\sum_{i=1}^{m} \frac{\omega_i}{\|\mathbf{x}_{k-1} - \mathbf{a}_i\|}} \tag{1.14}$$

with $\mathbf{x}_0$ a given arbitrary point.

The second derivation of Weiszfeld's method relies on the simple observation that the general step (1.14) can be equivalently written as

$$\mathbf{x}_k = \operatorname*{argmin}_{\mathbf{x}} \sum_{i=1}^{m} \omega_i \frac{\|\mathbf{x} - \mathbf{a}_i\|^2}{\|\mathbf{x}_{k-1} - \mathbf{a}_i\|}. \tag{1.15}$$

Therefore, the scheme (1.14) has the representation

$$\mathbf{x}_k = \operatorname*{argmin}_{\mathbf{x}} h(\mathbf{x}, \mathbf{x}_{k-1}), \tag{1.16}$$

where the auxiliary function $h(\cdot,\cdot)$ is defined by

$$h(\mathbf{x},\mathbf{y}) := \sum_{i=1}^{m} \omega_i \frac{\|\mathbf{x}-\mathbf{a}_i\|^2}{\|\mathbf{y}-\mathbf{a}_i\|}.$$

This approximation is completely different from the quadratic approximation described in Section 1.3.1 and it also cannot be considered as an MM method since the auxiliary function $h(\mathbf{x},\mathbf{y})$ is not an upper bound of the objective function $f(\mathbf{x})$.

Despite the above, Weiszfeld's method, like MM schemes, is a descent method. This is due to a nice property of the function h which is stated below.

Lemma 1.1. *For every* $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ *such that* $\mathbf{y} \notin \{\mathbf{a}_1,\ldots,\mathbf{a}_m\}$

$$h(\mathbf{x},\mathbf{x}) = f(\mathbf{x}), \tag{1.17}$$

$$h(\mathbf{x},\mathbf{y}) \geq 2f(\mathbf{x}) - f(\mathbf{y}). \tag{1.18}$$

Proof. The first property follows by substitution. To prove the second property (1.18), note that that for every two real numbers $a \in \mathbb{R}, b > 0$, the inequality

$$\frac{a^2}{b} \geq 2a - b,$$

holds true. Therefore, for every $i = 1,\ldots,m$

$$\frac{\|\mathbf{x}-\mathbf{a}_i\|^2}{\|\mathbf{y}-\mathbf{a}_i\|} \geq 2\|\mathbf{x}-\mathbf{a}_i\| - \|\mathbf{y}-\mathbf{a}_i\|.$$

Multiplying the latter inequality by ω_i and summing over $i = 1,\ldots,m$, (1.18) follows. □

Recall that in order to prove the descent property of an MM method we used the fact that the auxiliary function is an upper bound of the objective function. For the Fermat-Weber problem this is not the case, however the new property (1.18) is sufficient to prove the monotonicity. Indeed,

$$f(\mathbf{x}_{k-1}) \overset{(1.17)}{=} h(\mathbf{x}_{k-1},\mathbf{x}_{k-1}) \overset{(1.16)}{\geq} h(\mathbf{x}_k,\mathbf{x}_{k-1}) \overset{(1.18)}{\geq} 2f(\mathbf{x}_k) - f(\mathbf{x}_{k-1}).$$

Therefore, $f(\mathbf{x}_{k-1}) \geq 2f(\mathbf{x}_k) - f(\mathbf{x}_{k-1})$, implying the descent property $f(\mathbf{x}_k) \leq f(\mathbf{x}_{k-1})$.

As a final note, we mention the fact that Weiszfeld's method is in fact a gradient method

$$\mathbf{x}_k = \mathbf{x}_{k-1} - t_k \nabla f(\mathbf{x}_{k-1})$$

with a special choice of the step size t_k given by

$$t_k = \left(\sum_{i=1}^{m} \frac{\omega_i}{\|\mathbf{x}_{k-1}-\mathbf{a}_i\|} \right)^{-1}.$$

To conclude, Weiszfeld's method for the Fermat-Weber problem is one example of a gradient-based method that can be constructed by either fixed point ideas or by approximation models. The derivation of the method is different from what was described in previous sections, thus emphasizing the fact that the specific structure of the problem can and should be exploited. Another interesting example related to location and communication will be presented in Section 1.8.

In the forthcoming sections of this chapter we will focus on gradient-based methods emerging from the fixed point approach, and relying on the quadratic approximation. A special emphasis will be given to the proximal gradient method and its accelerations.

1.4 Convergence Results for the Proximal Gradient Method

In this section we make the setting more precise and introduce the main computational objects, study their properties and establish some key generic inequalities that serve as principal vehicle to establish convergence and rate of convergence results of the proximal gradient method and its extensions. The rate of convergence of the proximal gradient method will be established in this section while the analysis of extensions and/or accelerations of the method will be studied in the following sections. In the sequel, we make the standing assumption that there exists an optimal solution $\mathbf{x}^*$ to problem (M) and we set $F_* = F(\mathbf{x}^*)$.

1.4.1 The Prox-Grad Map

Following Section 1.3.1, we adopt the following approximation model for F. For any $L > 0$, and any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$, define

$$Q_L(\mathbf{x}, \mathbf{y}) := f(\mathbf{y}) + \langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y}) \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 + g(\mathbf{x}),$$

and

$$p_L^{f,g}(\mathbf{y}) := \operatorname{argmin} \{Q_L(\mathbf{x}, \mathbf{y}) : \mathbf{x} \in \mathbb{E}\}.$$

Ignoring the constant terms in $\mathbf{y}$, this reduces to (see also (1.7)):

$$\begin{aligned} p_L^{f,g}(\mathbf{y}) &= \operatorname*{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ g(\mathbf{x}) + \frac{L}{2}\|\mathbf{x} - (\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y}))\|^2 \right\} \\ &= \operatorname{prox}_{\frac{1}{L}}(g)\left(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\right). \end{aligned} \tag{1.19}$$

We call this composition of the proximal map with a gradient step of f the *prox-grad map* associated with f and g. The prox-grad map $p_L^{f,g}(\cdot)$ is well defined by the underlying assumptions on f and g. To simplify notation, we will omit the superscripts f and g and simply write p_L instead of $p_L^{f,g}$ whenever no confusion arises. First, we recall basic properties of Moreau's proximal map.

Lemma 1.2. *Let $g : \mathbb{E} \to (-\infty, \infty]$ be a closed proper convex function and for any $t > 0$, let*

$$g_t(\mathbf{z}) = \min_{\mathbf{u}} \left\{ g(\mathbf{u}) + \frac{1}{2t}\|\mathbf{u} - \mathbf{z}\|^2 \right\}. \tag{1.20}$$

Then,

1. *The minimum in (1.20) is attained at the unique point $prox_t(g)(\mathbf{z})$. As a consequence, the map $(I + t\partial g)^{-1}$ is single valued from $\mathbb{E}$ into itself and*

$$prox_t(g)(\mathbf{z}) = (I + t\partial g)^{-1}(\mathbf{z}) \text{ for every } \mathbf{z} \in \mathbb{E}.$$

2. *The function $g_t(\cdot)$ is continuously differentiable on $\mathbb{E}$ with a $\frac{1}{t}$-Lipschitz gradient given by*

$$\nabla g_t(\mathbf{z}) = \frac{1}{t}(I - prox_t(g)(\mathbf{z})) \text{ for every } \mathbf{z} \in \mathbb{E}.$$

In particular, if $g \equiv \delta_C$, with $C \subseteq \mathbb{E}$ closed and convex, then $prox_t(g) = (I + t\partial g)^{-1} = \Pi_C$, the orthogonal projection on C and we have

$$g_t(\mathbf{z}) = \frac{1}{2t}\|\mathbf{z} - \Pi_C(\mathbf{z})\|^2.$$

1.4.2 Fundamental Inequalities

We develop here some key inequalities that play a central role in the analysis of the proximal gradient method and in fact for any gradient-based method. Throughout the rest of this chapter we assume that ∇f is Lipschitz on $\mathbb{E}$, namely, there exists $L(f) > 0$ such that

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L(f)\|\mathbf{x} - \mathbf{y}\| \text{ for every } \mathbf{x}, \mathbf{y} \in \mathbb{E}.$$

For convenience we denote this class by $C^{1,1}_{L(f)}$. The first result is a well known important property of smooth functions.

Lemma 1.3 (Descent Lemma). *Let $f : \mathbb{E} \to (-\infty, \infty)$ be $C^{1,1}_{L(f)}$. Then for any $L \geq L(f)$,*

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y})\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 \text{ for every } \mathbf{x}, \mathbf{y} \in \mathbb{E}.$$

The next result gives a useful inequality for the prox-grad map which in turn can be used in the characterization of $p_L(\cdot)$. For a function f we define

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

Lemma 1.4. *Let $\boldsymbol{\xi} = prox_t(g)(\mathbf{z})$ for some $\mathbf{z} \in \mathbb{E}$ and let $t > 0$. Then*

$$2t(g(\boldsymbol{\xi}) - g(\mathbf{u})) \leq \|\mathbf{u} - \mathbf{z}\|^2 - \|\mathbf{u} - \boldsymbol{\xi}\|^2 - \|\boldsymbol{\xi} - \mathbf{z}\|^2 \text{ for every } \mathbf{u} \in dom\, g.$$

Proof. By definition of $\boldsymbol{\xi}$ we have

$$\boldsymbol{\xi} = \operatorname*{argmin}_{\mathbf{u}} \left\{ g(\mathbf{u}) + \frac{1}{2t}\|\mathbf{u}-\mathbf{z}\|^2 \right\}.$$

Writing the optimality condition for the above minimization problem yields

$$\langle \mathbf{u}-\boldsymbol{\xi}, \boldsymbol{\xi}-\mathbf{z}+t\boldsymbol{\gamma}\rangle \geq 0 \text{ for every } \mathbf{u} \in \operatorname{dom} g, \tag{1.21}$$

where $\boldsymbol{\gamma} \in \partial g(\boldsymbol{\xi})$. Since g is convex with $\boldsymbol{\gamma} \in \partial g(\boldsymbol{\xi})$, we also have

$$g(\boldsymbol{\xi}) - g(\mathbf{u}) \leq \langle \boldsymbol{\xi}-\mathbf{u}, \boldsymbol{\gamma}\rangle,$$

which combined with (1.21) and the fact that $t > 0$ yields

$$2t(g(\boldsymbol{\xi}) - g(\mathbf{u}) \leq 2\langle \mathbf{u}-\boldsymbol{\xi}, \boldsymbol{\xi}-\mathbf{z}\rangle,$$

and the desired result follows from the identity

$$2\langle \mathbf{u}-\boldsymbol{\xi}, \boldsymbol{\xi}-\mathbf{z}\rangle = \|\mathbf{u}-\mathbf{z}\|^2 - \|\mathbf{u}-\boldsymbol{\xi}\|^2 - \|\boldsymbol{\xi}-\mathbf{z}\|^2. \tag{1.22}$$

□

Since $p_L(\mathbf{y}) = \operatorname{prox}_{1/L}(g)\left(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\right)$, invoking Lemma 1.4, we now obtain a useful characterization of p_L. For further reference we denote for any $\mathbf{y} \in \mathbb{E}$:

$$\xi_L(\mathbf{y}) := \mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y}). \tag{1.23}$$

Lemma 1.5. *For any $\mathbf{x} \in \operatorname{dom} g, \mathbf{y} \in \mathbb{E}$, the prox-grad map p_L satisfies*

$$\frac{2}{L}\left[g(p_L(\mathbf{y})) - g(\mathbf{x})\right] \leq \|\mathbf{x}-\xi_L(\mathbf{y})\|^2 - \|\mathbf{x}-p_L(\mathbf{y})\|^2 - \|p_L(\mathbf{y})-\xi_L(\mathbf{y})\|^2, \tag{1.24}$$

where $\xi_L(\mathbf{y})$ is given in (1.23).

Proof. Follows from Lemma 1.4 with $t = \frac{1}{L}, \boldsymbol{\xi} = p_L(\mathbf{y})$ and $\mathbf{z} = \xi_L(\mathbf{y}) = \mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})$. □

Our last result combines all the above to produce the main pillar of the analysis.

Lemma 1.6. *Let $\mathbf{x} \in \operatorname{dom} g, \mathbf{y} \in \mathbb{E}$ and let $L > 0$ be such that the inequality*

$$F(p_L(\mathbf{y})) \leq Q(p_L(\mathbf{y}), \mathbf{y}). \tag{1.25}$$

is satisfied. Then

$$\frac{2}{L}(F(\mathbf{x}) - F(p_L(\mathbf{y})) \geq \frac{2}{L} l_f(\mathbf{x}, \mathbf{y}) + \|\mathbf{x}-p_L(\mathbf{y})\|^2 - \|\mathbf{x}-\mathbf{y}\|^2.$$

Furthermore, if f is also convex then

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

Proof. Recalling that

$$p_L(\mathbf{y}) = \operatorname*{argmin}_{\mathbf{x}} Q_L(\mathbf{x}, \mathbf{y}),$$

and using the definition of $Q_L(\cdot, \cdot)$ we have:

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

Therefore, using (1.25) it follows that

$$\begin{aligned} F(\mathbf{x}) - F(p_L(\mathbf{y})) &\geq F(\mathbf{x}) - Q_L(p_L(\mathbf{y}), \mathbf{y}) \\ &= f(\mathbf{x}) - f(\mathbf{y}) - \langle p_L(\mathbf{y}) - \mathbf{y}, \nabla f(\mathbf{y})\rangle \\ &\quad -\frac{L}{2}\|p_L(\mathbf{y}) - \mathbf{y}\|^2 + g(\mathbf{x}) - g(p_L(\mathbf{y})) \\ &= l_f(\mathbf{x}, \mathbf{y}) + \langle \mathbf{x} - p_L(\mathbf{y}), \nabla f(\mathbf{y})\rangle \\ &\quad -\frac{L}{2}\|p_L(\mathbf{y}) - \mathbf{y}\|^2 + g(\mathbf{x}) - g(p_L(\mathbf{y})). \end{aligned}$$

Now, invoking Lemma 1.5 and (1.22) we obtain,

$$\frac{2}{L}(g(\mathbf{x}) - g(p_L(\mathbf{y})) \geq 2\langle \mathbf{x} - p_L(\mathbf{y}), \xi_L(\mathbf{y}) - p_L(\mathbf{y})\rangle,$$

which substituted in the above inequality and recalling that $\xi_L(\mathbf{y}) = \mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})$ yields

$$\begin{aligned} \frac{2}{L}(F(\mathbf{x}) - F(p_L(\mathbf{y})) &\geq \frac{2}{L} l_f(\mathbf{x}, \mathbf{y}) + 2\langle \mathbf{x} - p_L(\mathbf{y}), \mathbf{y} - p_L(\mathbf{y})\rangle - \|p_L(\mathbf{y}) - \mathbf{y}\|^2 \\ &= \frac{2}{L} l_f(\mathbf{x}, \mathbf{y}) + \|p_L(\mathbf{y}) - \mathbf{y}\|^2 + 2\langle \mathbf{y} - \mathbf{x}, p_L(\mathbf{y}) - \mathbf{y}\rangle \\ &= \frac{2}{L} l_f(\mathbf{x}, \mathbf{y}) + \|\mathbf{x} - p_L(\mathbf{y})\|^2 - \|\mathbf{x} - \mathbf{y}\|^2, \end{aligned}$$

proving the first inequality. When f is convex we have $l_f(\mathbf{x}, \mathbf{y}) \geq 0$ and hence the second inequality follows. □

Note that condition (1.25) of Lemma 1.6 is always satisfied for $p_L(\mathbf{y})$ with $L \geq L(f)$, thanks to the descent lemma (Lemma 1.3).

1.4.3 Convergence of the Proximal Gradient Method: the Convex Case

We consider the proximal gradient method scheme for solving the model (M) when f is assumed convex. Since g is also assumed convex, the general model (M) is in this case convex. When $L(f) > 0$ is known, we can define the proximal gradient method with a constant stepsize rule.

Proximal Gradient Method with Constant Stepsize
Input: $L = L(f)$ - A Lipschitz constant of ∇f.
Step 0. Take $\mathbf{x}_0 \in \mathbb{E}$.
Step k. $(k \geq 1)$ Compute

$$\mathbf{x}_k = p_L(\mathbf{x}_{k-1})$$

An evident possible drawback of the above scheme is that the Lipschitz constant $L(f)$ is not always known or not easily computable (e.g., in the example of Section 1.6 where one needs to know $\lambda_{\max}(\mathbf{A}^T\mathbf{A})$). To overcome this potential difficulty, we also suggest and analyze the proximal gradient method with an easy backtracking stepsize rule. This is the next algorithm described below.

Proximal Gradient Method with Backtracking
Step 0. Take $L_0 > 0$, some $\eta > 1$ and $\mathbf{x}_0 \in \mathbb{E}$.
Step k. $(k \geq 1)$ Find the smallest nonnegative integer i_k such that with, $\bar{L} = \eta^{i_k} L_{k-1}$:

$$F(p_{\bar{L}}(\mathbf{x}_{k-1})) \leq Q_{\bar{L}}(p_{\bar{L}}(\mathbf{x}_{k-1}), \mathbf{x}_{k-1})). \tag{1.26}$$

Set $L_k = \eta^{i_k} L_{k-1}$ and compute

$$\mathbf{x}_k = p_{L_k}(\mathbf{x}_{k-1}). \tag{1.27}$$

Remark 1.1. *The sequence of function values $\{F(\mathbf{x}_k)\}$ produced by the proximal gradient method with either constant or backtracking stepsize rules is nonincreasing. Indeed, for every $k \geq 1$:*

$$F(\mathbf{x}_k) \leq Q_{L_k}(\mathbf{x}_k, \mathbf{x}_{k-1}) = Q_{L_k}(\mathbf{x}_{k-1}, \mathbf{x}_{k-1}) = F(\mathbf{x}_{k-1}),$$

where L_k is chosen by the backtracking rule or $L_k \equiv L(f)$ whenever the Lipschitz constant of ∇f is known.

Remark 1.2. *Since (1.26) holds for $\bar{L} \geq L(f)$, then for the proximal gradient method with backtracking it holds that $L_k \leq \eta L(f)$ for every $k \geq 1$ so that overall*

$$\beta L(f) \leq L_k \leq \alpha L(f), \tag{1.28}$$

where $\alpha = \beta = 1$ for the constant stepsize setting and $\alpha = \eta, \beta = \frac{L_0}{L(f)}$ for the backtracking case.

The next result shows that the proximal gradient method (with either of the two constant stepsize rules) converge at a sublinear rate in function values. Recall that since for $g \equiv 0$ and $g = \delta_C$, our model (M) recovers the basic gradient and gradient projection methods respectively, the result demonstrates that the presence of the nonsmooth term g in the model (M) does not deteriorate the same rate of convergence which is known to be valid for smooth problems.

Theorem 1.1 (Sublinear Rate Of Convergence of the Proximal Gradient Method). *Let $\{\mathbf{x}_k\}$ be the sequence generated by the proximal gradient method with either a constant or a backtracking stepsize rule. Then for every $k \geq 1$:*

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \frac{\alpha L(f)\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{2k}$$

for every optimal solution $\mathbf{x}^*$.

Proof. Invoking Lemma 1.6 with $\mathbf{x} = \mathbf{x}^*, \mathbf{y} = \mathbf{x}_n$ and $L = L_{n+1}$, we obtain

$$\frac{2}{L_{n+1}}\left(F(\mathbf{x}^*) - F(\mathbf{x}_{n+1})\right) \geq \|\mathbf{x}^* - \mathbf{x}_{n+1}\|^2 - \|\mathbf{x}^* - \mathbf{x}_n\|^2,$$

which combined with (1.28) and the fact that $F(\mathbf{x}^*) - F(\mathbf{x}_{n+1}) \leq 0$ yields

$$\frac{2}{\alpha L(f)}\left(F(\mathbf{x}^*) - F(\mathbf{x}_{n+1})\right) \geq \|\mathbf{x}^* - \mathbf{x}_{n+1}\|^2 - \|\mathbf{x}^* - \mathbf{x}_n\|^2. \tag{1.29}$$

Summing this inequality over $n = 0, \ldots, k-1$ gives

$$\frac{2}{\alpha L(f)}\left(kF(\mathbf{x}^*) - \sum_{n=0}^{k-1} F(\mathbf{x}_{n+1})\right) \geq \|\mathbf{x}^* - \mathbf{x}_k\|^2 - \|\mathbf{x}^* - \mathbf{x}_0\|^2. \tag{1.30}$$

Invoking Lemma 1.6 one more time with $\mathbf{x} = \mathbf{y} = \mathbf{x}_n, L = L_{n+1}$, yields

$$\frac{2}{L_{n+1}}\left(F(\mathbf{x}_n) - F(\mathbf{x}_{n+1})\right) \geq \|\mathbf{x}_n - \mathbf{x}_{n+1}\|^2.$$

Since we have $L_{n+1} \geq \beta L(f)$ (see (1.28)) and $F(\mathbf{x}_n) - F(\mathbf{x}_{n+1}) \geq 0$, it follows that

$$\frac{2}{\beta L(f)}\left(F(\mathbf{x}_n) - F(\mathbf{x}_{n+1})\right) \geq \|\mathbf{x}_n - \mathbf{x}_{n+1}\|^2.$$

Multiplying the last inequality by n and summing over $n = 0, \ldots k-1$, we obtain,

$$\frac{2}{\beta L(f)}\sum_{n=0}^{k-1}\left(nF(\mathbf{x}_n) - (n+1)F(\mathbf{x}_{n+1}) + F(\mathbf{x}_{n+1})\right) \geq \sum_{n=0}^{k-1} n\|\mathbf{x}_n - \mathbf{x}_{n+1}\|^2,$$

which simplifies to:

$$\frac{2}{\beta L(f)}\left(-kF(\mathbf{x}_k) + \sum_{n=0}^{k-1} F(\mathbf{x}_{n+1})\right) \geq \sum_{n=0}^{k-1} n\|\mathbf{x}_n - \mathbf{x}_{n+1}\|^2. \tag{1.31}$$

Adding (1.30) and (1.31) times β/α, we get

$$\frac{2k}{\alpha L(f)}\left(F(\mathbf{x}^*) - F(\mathbf{x}_k)\right) \geq \|\mathbf{x}^* - \mathbf{x}_k\|^2 + \frac{\beta}{\alpha}\sum_{n=0}^{k-1} n\|\mathbf{x}_n - \mathbf{x}_{n+1}\|^2 - \|\mathbf{x}^* - \mathbf{x}_0\|^2,$$

and hence it follows that

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \frac{\alpha L(f)\|\mathbf{x} - \mathbf{x}_0\|^2}{2k}.$$

□

This result demonstrates that in order to obtain an ϵ-optimal solution of (M), that is, a point $\hat{\mathbf{x}}$ such that $F(\hat{\mathbf{x}}) - F(\mathbf{x}^*) \leq \epsilon$, one requires at most $\left\lceil \frac{C}{\epsilon} \right\rceil$ iterations, where $C = \frac{\alpha L(f)\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{2}$. Thus, even for low accuracy requirements, the proximal gradient method can be very slow and inadequate for most applications. Later on, in Section 1.5 we will present an acceleration of the proximal gradient method that is equally simple, but possesses a significantly improved complexity rate.

It is also possible to prove the convergence of the sequence generated by the proximal gradient method and not only convergence of function values. This result relies on the Fejer monotonicity property of the generated sequence.

Theorem 1.2 (Convergence of the Sequence Generated by the Proximal Gradient Method). *Let $\{\mathbf{x}_k\}$ be the sequence generated by the proximal gradient method with either a constant or a backtracking stepsize rule. Then*

1. **Fejer monotonicity**. *For every optimal solution $\mathbf{x}^*$ of the model (M) and any $k \geq 1$:*

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \|\mathbf{x}_{k-1} - \mathbf{x}^*\|. \tag{1.32}$$

2. *The sequence $\{\mathbf{x}_k\}$ converges to an optimal solution of problem (M).*

Proof. 1. Invoking Lemma 1.6 with $\mathbf{x} = \mathbf{x}^*, \mathbf{y} = \mathbf{x}_{k-1}$ and $L = L_k$ (for the constant stepsize rule), $L_k \equiv L(f)$ we obtain

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

Since $F(\mathbf{x}^*) - F(\mathbf{x}_k) \leq 0$, property (1.32) follows.
2. By Fejer monotonicity it follows that for a given optimal solution $\mathbf{x}^*$

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \|\mathbf{x}_0 - \mathbf{x}^*\|.$$

Therefore, the sequence $\{\mathbf{x}_k\}$ is bounded. To prove the convergence of $\{\mathbf{x}_k\}$, it only remains to show that all converging subsequences have the same limit. Suppose in contradiction that there exist two subsequences $\{\mathbf{x}_{k_j}\}, \{\mathbf{x}_{n_j}\}$ converging to different limits $\mathbf{x}^\infty, \mathbf{y}^\infty$ respectively ($\mathbf{x}^\infty \neq \mathbf{y}^\infty$). Since $F(\mathbf{x}_{k_j}), F(\mathbf{x}_{n_j}) \to F_*$ (recalling that F_* is the optimal function value), it follows that $\mathbf{x}^\infty$ and $\mathbf{y}^\infty$ are optimal solutions of (M). Now, by Fejer monotonicity of the sequence $\{\mathbf{x}_k\}$, it follows that the sequence $\{\|\mathbf{x}_k - \mathbf{x}^\infty\|\}$ is bounded and nonincreasing and thus has a limit $\lim_{k\to\infty} \|\mathbf{x}_k - \mathbf{x}^\infty\| = l_1$. However, we also have $\lim_{k\to\infty} \|\mathbf{x}_k - \mathbf{x}^\infty\| = \lim_{j\to\infty} \|\mathbf{x}_{k_j} - \mathbf{x}^\infty\| = 0$, and $\lim_{k\to\infty} \|\mathbf{x}_k - \mathbf{x}^\infty\| = \lim_{j\to\infty} \|\mathbf{x}_{n_j} - \mathbf{x}^\infty\| = \|\mathbf{y}^\infty - \mathbf{x}^\infty\|$, so that $l_1 = 0 = \|\mathbf{x}^\infty - \mathbf{y}^\infty\|$, which is obviously a contradiction. □

1.4.4 The Nonconvex Case

When f is nonconvex, the convergence result is of course weaker. Convergence to a global minimum is out of reach. Recall that for a fixed $L > 0$, the condition $\mathbf{x}^* = p_L(\mathbf{x}^*)$ is a necessary condition for $\mathbf{x}^*$ to be an optimal solution of (M). Therefore, the convergence of the sequence to a stationary point can be measured by the quantity $\|\mathbf{x} - p_L(\mathbf{x})\|$. This is done in the next result.

Theorem 1.3 (Convergence of the Proximal Gradient Method in the Nonconvex Case). *Let $\{\mathbf{x}_k\}$ be the sequence generated by the proximal gradient method with either a constant or a backtracking stepsize rule. Then for every $n \geq 1$ we have*

$$\gamma_n \leq \frac{1}{\sqrt{n}} \left(\frac{2(F(\mathbf{x}_0) - F_*)}{\beta L(f)} \right)^{1/2},$$

where

$$\gamma_n := \min_{1 \leq k \leq n} \|\mathbf{x}_{k-1} - p_{L_k}(\mathbf{x}_{k-1})\|.$$

Moreover, $\|\mathbf{x}_{k-1} - p_{L_k}(\mathbf{x}_{k-1})\| \to 0$ as $k \to \infty$.

Proof. Invoking Lemma 1.6 with $\mathbf{x} = \mathbf{y} = \mathbf{x}_{k-1}, L = L_k$ and using the relation $\mathbf{x}_k = p_{L_k}(\mathbf{x}_{k-1})$, it follows that

$$\frac{2}{L_k}(F(\mathbf{x}_{k-1}) - F(\mathbf{x}_k)) \geq \|\mathbf{x}_{k-1} - \mathbf{x}_k\|^2, \tag{1.33}$$

where we also used the fact that $l_f(\mathbf{x}, \mathbf{x}) = 0$. By (1.28), $L_k \geq \beta L(f)$, which combined with (1.33) results with the inequality

$$\frac{\beta L(f)}{2}\|\mathbf{x}_{k-1} - \mathbf{x}_k\|^2 \leq F(\mathbf{x}_{k-1}) - F(\mathbf{x}_k).$$

Summing over $k = 1, \ldots, n$ we obtain

$$\frac{\beta L(f)}{2} \sum_{k=1}^{n} \|\mathbf{x}_{k-1} - \mathbf{x}_k\|^2 \leq F(\mathbf{x}_0) - F(\mathbf{x}_n),$$

which readily implies that $\|\mathbf{x}_{k-1} - p_{L_k}(\mathbf{x}_{k-1})\| \to 0$ and that

$$\min_{1 \leq k \leq n} \|\mathbf{x}_{k-1} - p_{L_k}(\mathbf{x}_{k-1})\|^2 \leq \frac{2(F(\mathbf{x}_0) - F_*)}{\beta L(f) n}.$$

□

Remark 1.3. *If $g \equiv 0$, the proximal gradient method reduces to the gradient method for the unconstrained nonconvex problem*

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

In this case

$$\mathbf{x}_{k-1} - p_{L_k}(\mathbf{x}_{k-1}) = \mathbf{x}_{k-1} - \left(\mathbf{x}_{k-1} - \frac{1}{L_k}\nabla f(\mathbf{x}_{k-1})\right) = \frac{1}{L_k}\nabla f(\mathbf{x}_{k-1}),$$

and Theorem 1.3 reduces to

$$\min_{1\le k\le n} \|\nabla f(\mathbf{x}_{k-1})\| \le \frac{1}{\sqrt{n}}\left(\frac{2\alpha^2 L(f)(F(\mathbf{x}_0) - F_*)}{\beta}\right)^{1/2},$$

recovering the classical rate of convergence of the gradient method, i.e., $\nabla f(\mathbf{x}_k) \to 0$ *at a rate of* $O(1/\sqrt{k})$.

1.5 A Fast Proximal Gradient Method

1.5.1 Idea of the method

In this section we return to the convex scenario, that is we assume that f is convex. The basic gradient method relies on using information on the previous iterate only. On the other hand, the so-called conjugate gradient method does use "memory", i.e., it generates steps which exploit the *two* previous iterates, and has been known to often improve the performance of basic gradient methods. Similar ideas have been followed to handle *nonsmooth* problems, in particular the so-called R-algorithm of Shor (see bibliography notes).

However, such methods have not been proven to exhibit a better complexity rate than $O(1/k)$, furthermore, they also often involve some matrix operations that can be problematic in large-scale applications.

Therefore, here the objective is double, namely to build a gradient-based method that

i. keeps the simplicity of the proximal gradient method to solve model (M).
ii. is proven to be significantly faster, both theoretically and practically.

Both tasks will be achieved by considering again the basic model (M) of Section 1.2 in the convex case. Specifically, we will build a method that is very similar to the proximal gradient method and is of the form

$$\mathbf{x}_k = p_L(\mathbf{y}_k),$$

where the new point $\mathbf{y}_k$ will be smartly chosen in terms of the two previous iterates $\{\mathbf{x}_{k-1}, \mathbf{x}_{k-2}\}$ and is very easy to compute. Thus, here also we follow the idea of building a scheme with memory, but which is much simpler than the methods alluded above, and as shown below will be proven to exhibit a faster rate of convergence.

1.5.2 A Fast Proximal Gradient Method using Two Past Iterations

We begin by presenting the algorithm with a constant stepsize.

Fast Proximal Gradient Method with Constant Stepsize
Input: $L = L(f)$ - A Lipschitz constant of ∇f.
Step 0. Take $\mathbf{y}_1 = \mathbf{x}_0 \in \mathbb{E}$, $t_1 = 1$.
Step k. ($k \geq 1$) Compute

$$\mathbf{x}_k = p_L(\mathbf{y}_k), \tag{1.34}$$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \tag{1.35}$$

$$\mathbf{y}_{k+1} = \mathbf{x}_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}_k - \mathbf{x}_{k-1}). \tag{1.36}$$

The main difference between the above algorithm and the proximal gradient method, is that that the prox-grad operation $p_L(\cdot)$ is not employed on the previous point $\mathbf{x}_{k-1}$, but rather at the point $\mathbf{y}_k$ which uses a very specific linear combination of the previous two points $\{\mathbf{x}_{k-1}, \mathbf{x}_{k-2}\}$. Obviously the main computational effort in both the basic and fast versions of the proximal gradient method remains the same, namely in the operator p_L. The requested additional computation for the fast proximal gradient method in the steps (1.35) and (1.36) is clearly marginal. The specific formula for (1.35) emerges from the recursive relation that will be established below in Lemma 1.7.

For the same reasons already explained in Section 1.4.3, we will also analyze the fast proximal gradient method with a backtracking stepsize rule, which we now explicitly state.

Fast Proximal Gradient Method with Backtracking
Step 0. Take $L_0 > 0$, some $\eta > 1$ and $\mathbf{x}_0 \in \mathbb{E}$. Set $\mathbf{y}_1 = \mathbf{x}_0$, $t_1 = 1$.
Step k. ($k \geq 1$) Find the smallest nonnegative integer i_k such that with $\bar{L} = \eta^{i_k} L_{k-1}$:

$$F(p_{\bar{L}}(\mathbf{y}_k)) \leq Q_{\bar{L}}(p_{\bar{L}}(\mathbf{y}_k), \mathbf{y}_k).$$

Set $L_k = \eta^{i_k} L_{k-1}$ and compute

$$\mathbf{x}_k = p_{L_k}(\mathbf{y}_k),$$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2},$$

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

Note that the upper and lower bounds on L_k given in Remark 1.2 still hold true for the fast proximal gradient method, namely

$$\beta L(f) \leq L_k \leq \alpha L(f).$$

The next result provides the key recursive relation for the sequence $\{F(\mathbf{x}_k) - F(\mathbf{x}^*)\}$ that will imply the better complexity rate $O(1/k^2)$. As we shall see, Lemma 1.6 of Section 1.4.2 plays a central role in the proofs.

Lemma 1.7. *The sequences $\{\mathbf{x}_k, \mathbf{y}_k\}$ generated via the fast proximal gradient method with either a constant or backtracking stepsize rule satisfy for every $k \geq 1$*

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

where

$$v_k := F(\mathbf{x}_k) - F(\mathbf{x}^*), \tag{1.37}$$

$$\mathbf{u}_k := t_k \mathbf{x}_k - (t_k - 1)\mathbf{x}_{k-1} - \mathbf{x}^*. \tag{1.38}$$

Proof. Invoking Lemma 1.6 with $\mathbf{x} = t_{k+1}^{-1}\mathbf{x}^* + (1 - t_{k+1}^{-1})\mathbf{x}_k, \mathbf{y} = \mathbf{y}_{k+1}$ and $L = L_{k+1}$ we have

$$\frac{2}{L_{k+1}}(F(t_{k+1}^{-1}\mathbf{x}^* + (1 - t_{k+1}^{-1})\mathbf{x}_k) - F(\mathbf{x}_{k+1})) \tag{1.39}$$

$$\geq \frac{1}{t_{k+1}^2}\{\|t_{k+1}\mathbf{x}_{k+1} - (\mathbf{x}^* + (t_{k+1} - 1)\mathbf{x}_k)\|^2 - \|t_{k+1}\mathbf{y}_{k+1} - (\mathbf{x}^* + (t_{k+1} - 1)\mathbf{x}_k)\|^2\}.$$

By the convexity of F we also have

$$F(t_{k+1}^{-1}\mathbf{x}^* + (1 - t_{k+1}^{-1})\mathbf{x}_k) \leq t_{k+1}^{-1} F(\mathbf{x}^*) + (1 - t_{k+1}^{-1})F(\mathbf{x}_k),$$

which combined with (1.39) yields

$$\frac{2}{L_{k+1}}((1 - t_{k+1}^{-1})v_k - v_{k+1}) \geq \frac{1}{t_{k+1}^2}\{\|t_{k+1}\mathbf{x}_{k+1} - (\mathbf{x}^* + (t_{k+1} - 1)\mathbf{x}_k)\|^2 - \|t_{k+1}\mathbf{y}_{k+1} - (\mathbf{x}^* + (t_{k+1} - 1)\mathbf{x}_k)\|^2\}.$$

Using the relation $t_k^2 = t_{k+1}^2 - t_{k+1}$, the latter is equivalent to

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

where we used the definition of $\mathbf{u}_k$ (1.38) and the definition of $\mathbf{y}_{k+1}$ (1.36) to simplify the righthand side. Since $L_{k+1} \geq L_k$, the desired result follows. □

We also need the following trivial facts.

Lemma 1.8. *Let $\{a_k, b_k\}$ be positive sequences of reals satisfying*

$$a_k - a_{k+1} \geq b_{k+1} - b_k, \forall k \geq 1, \text{ with } a_1 + b_1 \leq c,\ c > 0.$$

Then, $a_k \leq c$ for every $k \geq 1$.

Lemma 1.9. *The positive sequence $\{t_k\}$ generated by the fast proximal gradient method via (1.35) with $t_1 = 1$ satisfies $t_k \geq (k+1)/2$ for all $k \geq 1$.*

We are now ready to prove the promised improved complexity result for the fast proximal gradient method.

Theorem 1.4. *Let $\{\mathbf{x}_k\}, \{\mathbf{y}_k\}$ be generated by the fast proximal gradient method with either a constant or a backtracking stepsize rule. Then for any $k \geq 1$*

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \frac{2\alpha L(f)\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{(k+1)^2}, \quad \forall \mathbf{x}^* \in X_*, \tag{1.40}$$

where $\alpha = 1$ for the constant stepsize setting and $\alpha = \eta$ for the backtracking stepsize setting.

Proof. Let us define the following quantities:

$$a_k := \frac{2}{L_k} t_k^2 v_k, \quad b_k := \|\mathbf{u}_k\|^2, \quad c := \|\mathbf{y}_1 - \mathbf{x}^*\|^2 = \|\mathbf{x}_0 - \mathbf{x}^*\|^2,$$

and recall (c.f. Lemma 1.7) that $v_k := F(\mathbf{x}_k) - F(\mathbf{x}^*)$. Then, by Lemma 1.7 we have for every $k \geq 1$

$$a_k - a_{k+1} \geq b_{k+1} - b_k,$$

and hence assuming that $a_1 + b_1 \leq c$ holds true, invoking Lemma 1.8, we obtain that

$$\frac{2}{L_k} t_k^2 v_k \leq \|\mathbf{x}_0 - \mathbf{x}^*\|^2,$$

which combined with $t_k \geq (k+1)/2$ (by Lemma 1.9) yields

$$v_k \leq \frac{2L_k\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{(k+1)^2}.$$

Utilizing the upper bound on L_k given in (1.28), the desired result (1.40) follows. Thus, all that remains is to prove the validity of the relation $a_1 + b_1 \leq c$. Since $t_1 = 1$, and using the definition of $\mathbf{u}_k$ (1.38), we have here:

$$a_1 = \frac{2}{L_1} t_1 v_1 = \frac{2}{L_1} v_1, \quad b_1 = \|\mathbf{u}_1\|^2 = \|\mathbf{x}_1 - \mathbf{x}^*\|^2.$$

Applying Lemma 1.6 to the points $\mathbf{x} := \mathbf{x}^*, \mathbf{y} := \mathbf{y}_1$ with $L = L_1$, we get

$$\frac{2}{L_1}(F(\mathbf{x}^*) - F(\mathbf{x}_1)) \geq \|\mathbf{x}_1 - \mathbf{x}^*\|^2 - \|\mathbf{y}_1 - \mathbf{x}^*\|^2, \tag{1.41}$$

namely

$$\frac{2}{L_1} v_1 \leq \|\mathbf{y}_1 - \mathbf{x}^*\|^2 - \|\mathbf{x}_1 - \mathbf{x}^*\|^2,$$

that is, $a_1 + b_1 \leq c$ holds true. □

The number of iterations of the proximal gradient method required to obtain an ε-optimal solution, that is an $\tilde{\mathbf{x}}$ such that $F(\tilde{\mathbf{x}}) - F_* \leq \varepsilon$, is at most $\lceil C/\sqrt{\varepsilon} - 1 \rceil$ where $C = \sqrt{2\alpha L(f)\|\mathbf{x}_0 - \mathbf{x}^*\|^2}$, and which clearly improves the convergence of

the basic proximal gradient method. In Section 1.6 we illustrate the practical value of this theoretical global convergence rate estimate on the l_1-based regularization problem and demonstrate its applicability in wavelet-based image deblurring.

1.5.3 Monotone versus Nonmonotone

The fast proximal gradient method, as opposed to the standard proximal gradient one, *is not* a monotone algorithm, that is, the function values are not guaranteed to be nonincreasing. Monotonicity seems to be a desirable property of minimization algorithms, but it is not required in the proof of convergence of the fast proximal gradient method. Moreover, numerical simulations show that in fact the algorithm is "almost monotone", that is, except for very few iterations the algorithm exhibits a monotonicity property.

However, for some applications the prox operation cannot be computed exactly, see for example the total-variation based deblurring example in Section 1.7. In these situations, monotonicity becomes an important issue. It might happen that due to the inexact computations of the prox map, the algorithm might become extremely non-monotone and in fact can even diverge! This is illustrated in the numerical examples of Section 1.7.3. This is one motivation to introduce a monotone version of the fast proximal gradient method, which is now explicitly stated in the constant stepsize rule setting.

Monotone Fast Proximal Gradient Method
Input: $L \geq L(f)$ - An upper bound on the Lipschitz constant of ∇f.
Step 0. Take $\mathbf{y}_1 = \mathbf{x}_0 \in \mathbb{E},\ t_1 = 1$.
Step k. $(k \geq 1)$ Compute

$$\mathbf{z}_k = p_L(\mathbf{y}_k),$$

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}, \tag{1.42}$$

$$\mathbf{x}_k = \operatorname{argmin}\{F(\mathbf{x}) :\ \mathbf{x} = \mathbf{z}_k, \mathbf{x}_{k-1}\} \tag{1.43}$$

$$\mathbf{y}_{k+1} = \mathbf{x}_k + \left(\frac{t_k}{t_{k+1}}\right)(\mathbf{z}_k - \mathbf{x}_k) + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}_k - \mathbf{x}_{k-1}). \tag{1.44}$$

Clearly, with this modification, we have now a monotone algorithm which is easily seen to be as simple as the fast proximal gradient method regarding its computational steps. Moreover, it turns out that this modification does not affect the theoretical rate of convergence. Indeed, the convergence rate result for the monotone version remains the same as the convergence rate result of the non-monotone method:

Theorem 1.5. *Let* $\{\mathbf{x}_k\}$ *be generated by the monotone proximal gradient method. Then for any* $k \geq 1$ *and any optimal solution* $\mathbf{x}^*$*:*

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \frac{2L(f)\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{(k+1)^2}.$$

1.6 Algorithms for l_1-based Regularization Problems

1.6.1 Problem Formulation

In this section we return to the RLS problem discussed in Section 1.2.2. We concentrate on the on the l_1-based regularization problem in which one seeks to find the solution of

$$\min_{\mathbf{x}}\{F(\mathbf{x}) \equiv \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \lambda\|\mathbf{x}\|_1\}, \tag{1.45}$$

which is the general model (M) with $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2, g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$. In image deblurring applications, and in particular in wavelet-based restoration methods, $\mathbf{A}$ is often chosen as $\mathbf{A} = \mathbf{R}\mathbf{W}$ where $\mathbf{R}$ is the blurring matrix and $\mathbf{W}$ contains a wavelet basis (i.e., multiplying by $\mathbf{W}$ corresponds to performing inverse wavelet transform). The vector $\mathbf{x}$ contains the coefficients of the unknown image. The underlying philosophy in dealing with the l_1 norm regularization criterion is that most images have a sparse representation in the wavelet domain. The presence of the l_1 term in (1.45) is used to induce sparsity of the solution. Another important advantage of the l_1-based regularization (1.45) over the l_2-based Tikhonov regularization is that as opposed to the latter, l_1 regularization is less sensitive to outliers, which in image processing applications correspond to sharp edges.

The convex optimization problem (1.45) can be cast as a second order cone programming problem and thus could be solved via interior point methods. However, in most applications, e.g., in image deblurring, the problem is not only large scale (can reach millions of decision variables), but also involves dense matrix data, which often precludes the use and potential advantage of sophisticated interior point methods. This motivated the search for simpler gradient-based algorithms for solving (1.45), where the dominant computational effort is relatively cheap matrix-vector multiplications involving $\mathbf{A}$ and $\mathbf{A}^T$.

1.6.2 ISTA: Iterative Shrinkage/Thresholding Algorithm

One popular method to solve problem (1.45) is to employ the proximal gradient method. The proximal map associated with $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$ can be analytically computed:

$$\text{prox}_t(g)(\mathbf{y}) = \underset{\mathbf{u}}{\text{argmin}} \left\{ \frac{1}{2t}\|\mathbf{u} - \mathbf{y}\|^2 + \lambda\|\mathbf{u}\|_1 \right\} = \mathcal{T}_{\lambda t}(\mathbf{y}),$$

where $\mathcal{T}_\alpha : \mathbb{R}^n \to \mathbb{R}^n$ is the shrinkage or soft threshold operator defined by

$$\mathcal{T}_\alpha(\mathbf{x})_i = (|x_i| - \alpha)_+ \text{sgn}\,(x_i). \tag{1.46}$$

The arising method is the so-called *iterative shrinkage/thresholding* algorithm (ISTA)[5], which we now explicitly write for the constant stepsize setting.

ISTA with Constant Stepsize
Input: $L = L(f)$ - A Lipschitz constant of ∇f.
Step 0. Take $\mathbf{x}_0 \in \mathbb{E}$.
Step k. ($k \geq 1$) Compute

$$\mathbf{x}_k = \mathcal{T}_{\lambda/L}\left(\mathbf{x}_{k-1} - \frac{2}{L}\mathbf{A}^T(\mathbf{A}\mathbf{x}_{k-1} - \mathbf{b})\right)$$

We note that the Lipschitz constant of the gradient of $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ is given by $L(f) = 2\lambda_{\max}(\mathbf{A}^T\mathbf{A})$. It is of course also possible to incorporate a backtracking stepsize rule in ISTA as defined in Section 1.4.3.

1.6.3 FISTA: Fast ISTA

The function values of the sequence generated by ISTA, which is just a special case of the proximal gradient method, converge to the optimal function value at a rate of $O(1/k)$, k being the iteration index. An acceleration of ISTA can be achieved by invoking the fast proximal gradient method for the l_1-based regularization problem (1.45) discussed in Section 1.5. The fast version ISTA algorithm is called FISTA and is now explicitly stated.

FISTA with constant stepsize
Input: $L = L(f)$ - A Lipschitz constant of ∇f.
Step 0. Take $\mathbf{y}_1 = \mathbf{x}_0 \in \mathbb{E}$, $t_1 = 1$.
Step k. ($k \geq 1$) Compute

$$\begin{aligned}
\mathbf{x}_k &= \mathcal{T}_{\lambda/L}\left(\mathbf{y}_k - \frac{2}{L}\mathbf{A}^T(\mathbf{A}\mathbf{y}_k - \mathbf{b})\right), \\
t_{k+1} &= \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \\
\mathbf{y}_{k+1} &= \mathbf{x}_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}_k - \mathbf{x}_{k-1}).
\end{aligned}$$

Invoking Theorem 1.4, the rate of convergence of FISTA is $O(1/k^2)$ – a substantial improvement of the rate of convergence of ISTA. Next, we demonstrate

[5] Other names in the signal processing literature include for example threshold Landweber method, iterative denoising, deconvolution algorithms.

Figure 1.1 Blurring of the cameraman.

through representative examples the practical value of this theoretical global convergence rate estimate derived for FISTA on the l_1 wavelet-based regularization problem (1.45).

1.6.4 Numerical Examples

Consider the 256×256 camerman test image whose pixels were scaled into the range between 0 and 1. The image went through a Gaussian blur of size 9×9 and standard deviation 4 followed by a an additive zero-mean white Gaussian noise with standard deviation 10^{-3}. The original and observed images are given in Figure 1.1.

For these experiments we assume reflexive (Neumann) boundary conditions. We then tested ISTA and FISTA for solving problem (1.45) where $\mathbf{b}$ represents the (vectorized) observed image and $\mathbf{A} = \mathbf{RW}$ where $\mathbf{R}$ is the matrix representing the blur operator and $\mathbf{W}$ is the inverse of a three stage Haar wavelet transform. The regularization parameter was chosen to be $\lambda =$ 2e-5 and the initial image was the blurred image. The Lipschitz constant was computable in this example (and those in the sequel) since the eigenvalues of the matrix $\mathbf{A}^T\mathbf{A}$ can be easily calculated using the two dimensional cosine transform.

Iterations 100 and 200 are described in Figure 1.2. The function value at iteration k is denoted by F_k. The images produced by FISTA are of a better quality than those created by ISTA. The function value of FISTA was consistently lower than the function value of ISTA. We also computed the function values produced after 1000 iterations for ISTA and FISTA which were respectively 2.45e-1 and 2.23e-1. Note that the function value of ISTA after 1000 iterations is still worse (that is, larger) than the function value of FISTA after 100 iterations.

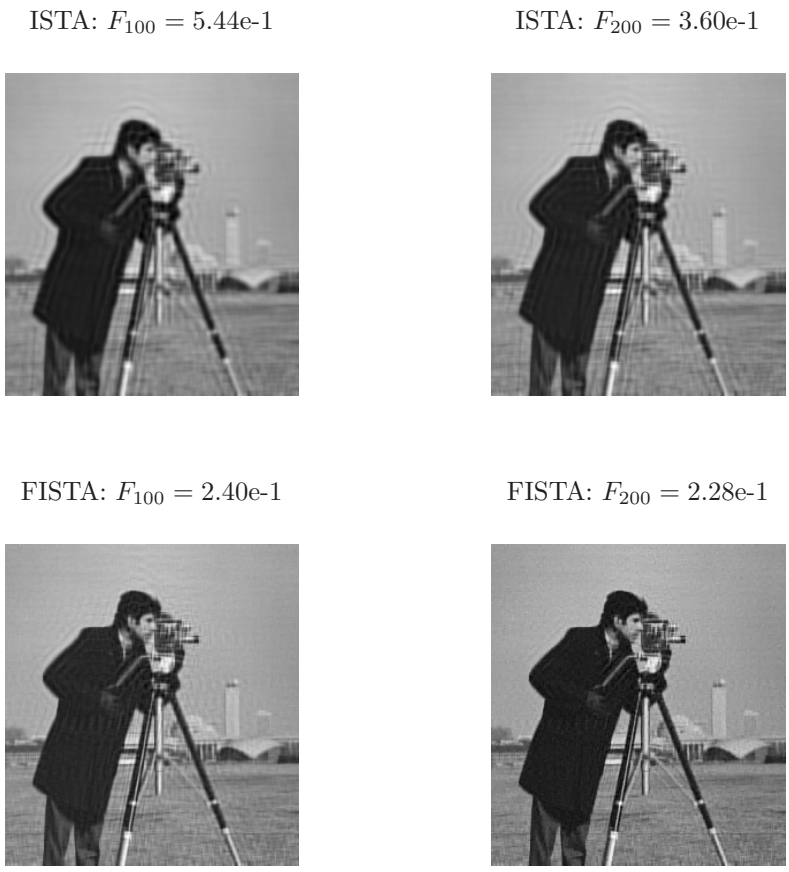

Figure 1.2 Iterations of ISTA and FISTA methods for deblurring of the cameraman.

From the previous example it seems that practically FISTA is able to reach accuracies that are beyond the capabilities of ISTA. To test this hypothesis we also considered an example in which the optimal solution is known. For that sake we considered a 64×64 image which undergoes the same blur operator as in the previous example. No noise was added and we solved the least squares problem, that is $\lambda = 0$. The optimal solution of this problem is zero. The function values of the two methods for 10000 iterations are described in Figure 1.3. The results produced by FISTA are better than those produced by ISTA by several orders of magnitude and clearly demonstrate the effective performance of FISTA. One can see that after 10000 iterations FISTA reaches accuracy of approximately 10^{-7} while ISTA reach accuracy of only 10^{-3}. Finally, we observe that the values obtained by ISTA at iteration 10000 was already obtained by FISTA at iterations 275.

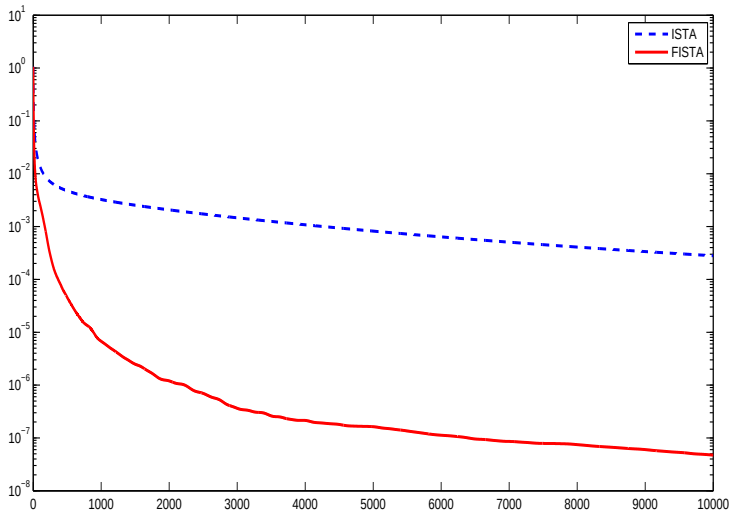

Figure 1.3 Comparison of function values errors $F(\mathbf{x}_k) - F(\mathbf{x}^*)$ of ISTA and FISTA.

1.7 TV-based Restoration Problems

1.7.1 Problem Formulation

Consider images that are defined on rectangle domains. Let $\mathbf{b} \in \mathbb{R}^{m\times n}$ be an observed noisy image, $\mathbf{x} \in \mathbb{R}^{m\times n}$ the true (original) image to be recovered, $\mathcal{A}$ an affine map representing a blurring operator, and $\mathbf{w} \in \mathbb{R}^{m\times n}$ a corresponding additive unknown noise satisfying the relation:

$$\mathbf{b} = \mathcal{A}(\mathbf{x}) + \mathbf{w}. \tag{1.47}$$

The problem of finding an $\mathbf{x}$ from the above relation is a special case of the basic discrete linear inverse problem discussed in Section 1.2.2. Here we are concerned with total variation (TV)-based regularization, which, given $\mathcal{A}$ and $\mathbf{b}$ seeks to recover $\mathbf{x}$ by solving the convex nonsmooth minimization problem

$$\min_{\mathbf{x}}\{\|\mathcal{A}(\mathbf{x}) - \mathbf{b}\|_F^2 + 2\lambda \mathrm{TV}(\mathbf{x})\}, \tag{1.48}$$

where $\lambda > 0$ and $\mathrm{TV}(\cdot)$ stands for the discrete total variation function. The underlying Euclidean space $\mathbb{E}$ comprises all $m \times n$ matrices with the usual inner product: $\langle \mathbf{a}, \mathbf{b}\rangle = \mathrm{Tr}(\mathbf{b}^T\mathbf{a})$ and the induced Frobenius norm $\|\cdot\|_F$. The identity map will be denoted by $\mathcal{I}$ and with $\mathcal{A} \equiv \mathcal{I}$ problem (1.48) reduces to the so-called *denoising* problem.

Two popular choices for the discrete TV are the isotropic TV defined by

$$\mathbf{x} \in \mathbb{R}^{m\times n}, \quad \mathrm{TV}_I(\mathbf{x}) = \sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\sqrt{(x_{i,j}-x_{i+1,j})^2+(x_{i,j}-x_{i,j+1})^2} + \sum_{i=1}^{m-1}|x_{i,n}-x_{i+1,n}| + \sum_{j=1}^{n-1}|x_{m,j}-x_{m,j+1}|$$

and the l_1-based, anisotropic TV defined by

$$\mathbf{x} \in \mathbb{R}^{m\times n}, \quad \mathrm{TV}_{l_1}(\mathbf{x}) = \sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\{|x_{i,j}-x_{i+1,j}|+|x_{i,j}-x_{i,j+1}|\} + \sum_{i=1}^{m-1}|x_{i,n}-x_{i+1,n}| + \sum_{j=1}^{n-1}|x_{m,j}-x_{m,j+1}|,$$

where in the above formulas we assumed the (standard) reflexive boundary conditions:

$$x_{m+1,j}-x_{m,j}=0,\ \forall\, j \text{ and } x_{i,n+1}-x_{i,n}=0,\ \forall i.$$

1.7.2 TV-based Denoising

As was already mentioned, when $\mathcal{A}=\mathcal{I}$, problem (1.48) reduces to the denoising problem

$$\min \|\mathbf{x}-\mathbf{b}\|_F^2 + 2\lambda \mathrm{TV}(\mathbf{x}), \tag{1.49}$$

where the nonsmooth regularizer function TV is either the isotropic TV_I or anisotropic TV_{l_1} function. Although this problem can be viewed as a special case of the general model (M) by substituting $f(\mathbf{x})=\|\mathbf{x}-\mathbf{b}\|^2$ and $g(\mathbf{x})=2\lambda\mathrm{TV}(\mathbf{x})$, it is not possible to solve it via the proximal gradient method (or its extensions). This is due to the fact that computation of the prox map amounts to solving a denoising problem of the exact same form.

A common approach for solving the denoising problem is to formulate its dual problem and solve it via a gradient-based method. In order to define the dual problem, some notation is in order

- $\mathcal{P}$ is the set of matrix-pairs $(\mathbf{p},\mathbf{q})$ where $\mathbf{p}\in\mathbb{R}^{(m-1)\times n}$ and $\mathbf{q}\in\mathbb{R}^{m\times(n-1)}$ that satisfy

$$\begin{aligned} p_{i,j}^2+q_{i,j}^2 &\le 1, \quad i=1,\dots,m-1, j=1,\dots,n-1,\\ |p_{i,n}| &\le 1, \quad i=1,\dots,m-1,\\ |q_{m,j}| &\le 1, \quad j=1,\dots,n-1. \end{aligned}$$

- The linear operation $\mathcal{L}:\mathbb{R}^{(m-1)\times n}\times\mathbb{R}^{m\times(n-1)}\to\mathbb{R}^{m\times n}$ is defined by the formula

$$\mathcal{L}(\mathbf{p},\mathbf{q})_{i,j} = p_{i,j}+q_{i,j}-p_{i-1,j}-q_{i,j-1}, \quad i=1,\dots,m, j=1,\dots,n,$$

where we assume that $p_{0,j}=p_{m,j}=q_{i,0}=q_{i,n}\equiv 0$ for every $i=1,\dots,m$ and $j=1,\dots,n$.

The formulation of the dual problem is now recalled.

Proposition 1.1. *Let $(\mathbf{p},\mathbf{q}) \in \mathcal{P}$ be the optimal solution of the problem*

$$\max_{(\mathbf{p},\mathbf{q})\in\mathcal{P}} -\|\mathbf{b} - \lambda\mathcal{L}(\mathbf{p},\mathbf{q})\|_F^2. \tag{1.50}$$

Then the optimal solution of (1.49) with $TV = TV_I$ is given by

$$\mathbf{x} = \mathbf{b} - \lambda\mathcal{L}(\mathbf{p},\mathbf{q})). \tag{1.51}$$

Proof. First note that the relations

$$\sqrt{x^2+y^2} = \max_{p_1,p_2}\{p_1x + p_2y : p_1^2 + p_2^2 \le 1\},$$
$$|x| = \max_{p}\{px : |p| \le 1\}$$

hold true. Hence, we can write

$$\mathrm{TV}_I(\mathbf{x}) = \max_{(\mathbf{p},\mathbf{q})\in\mathcal{P}} T(\mathbf{x},\mathbf{p},\mathbf{q}),$$

where

$$\begin{aligned} T(\mathbf{x},\mathbf{p},\mathbf{q}) = &\textstyle\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}[p_{i,j}(x_{i,j} - x_{i+1,j}) + q_{i,j}(x_{i,j} - x_{i,j+1})] \\ &+ \textstyle\sum_{i=1}^{m-1} p_{i,n}(x_{i,n} - x_{i+1,n}) + \sum_{j=1}^{n-1} q_{m,j}(x_{m,j} - x_{m,j+1}). \end{aligned}$$

With this notation we have

$$T(\mathbf{x},\mathbf{p},\mathbf{q}) = \mathrm{Tr}(\mathcal{L}(\mathbf{p},\mathbf{q})^T\mathbf{x}).$$

The problem (1.49) therefore becomes

$$\min_{\mathbf{x}} \max_{(\mathbf{p},\mathbf{q})\in\mathcal{P}} \left\{\|\mathbf{x} - \mathbf{b}\|_F^2 + 2\lambda\mathrm{Tr}(\mathcal{L}(\mathbf{p},\mathbf{q})^T\mathbf{x})\right\}. \tag{1.52}$$

Since the objective function is convex in $\mathbf{x}$ and concave in $\mathbf{p},\mathbf{q}$, we can exchange the order of the minimum and maximum and obtain the equivalent formulation

$$\max_{(\mathbf{p},\mathbf{q})\in\mathcal{P}} \min_{\mathbf{x}} \left\{\|\mathbf{x} - \mathbf{b}\|_F^2 + 2\lambda\mathrm{Tr}(\mathcal{L}(\mathbf{p},\mathbf{q})^T\mathbf{x})\right\},$$

The optimal solution of the inner minimization problem is

$$\mathbf{x} = \mathbf{b} - \lambda\mathcal{L}(\mathbf{p},\mathbf{q}).$$

Plugging the above expression for $\mathbf{x}$ back into (1.52), and omitting constant terms, we obtain the dual problem (1.50). □

Remark 1.4. *The only difference in the dual problem corresponding to the case $TV = TV_{l_1}$ (in comparison to the case $TV = TV_I$), is that the minimization in the dual problem is not done over the set $\mathcal{P}$, but over the set $\mathcal{P}_1$ which consists of all pairs of matrices $(\mathbf{p},\mathbf{q})$ where $\mathbf{p} \in \mathbb{R}^{(m-1)\times n}$ and $\mathbf{q} \in \mathbb{R}^{m\times(n-1)}$ satisfying*

$$|p_{i,j}| \le 1, i = 1,\dots,m-1, j = 1,\dots,n,$$
$$|q_{i,j}| \le 1, i = 1,\dots,m, j = 1,\dots,n-1.$$

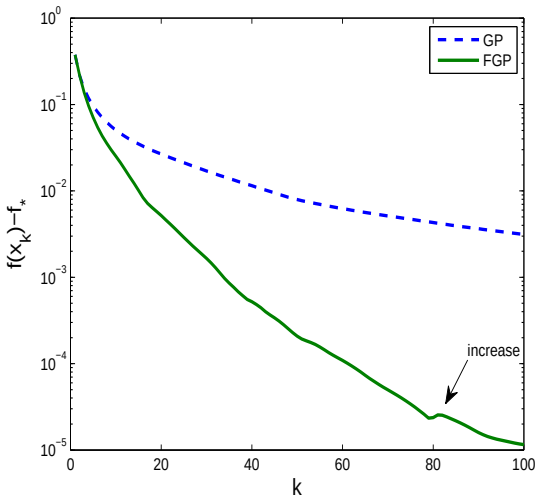

Figure 1.4 Accuracy of FGP compared with GP.

The dual problem (1.50), when formulated as a minimization problem:

$$\min_{(\mathbf{p},\mathbf{q})\in\mathcal{P}} \|\mathbf{b} - \lambda\mathcal{L}(\mathbf{p},\mathbf{q})\|_F^2 \tag{1.53}$$

falls into the category of model (M) by taking f to be the objective function of (1.53) and $g \equiv \delta_{\mathcal{P}}$ – the indicator function of $\mathcal{P}$. The objective function, being quadratic, has a Lipschitz gradient and as a result we can invoke either the proximal gradient method, which coincides with the gradient projection method in this case, or the fast proximal gradient method. The exact details of computations of the Lipschitz constant and of the gradient are omitted. The slower method will be called GP (for gradient projection) and the faster method will be called FGP (for fast gradient projection).

To demonstrate the advantage of FGP over GP, we have taken a small 10×10 image for which we added normally distributed white noise with standard deviation 0.1. The parameter λ was chosen as 0.1. Since the problem is small, we were able to find its exact solution. Figure 1.4 shows the the difference $F(\mathbf{x}_k) - F_*$ (in log scale) for $k = 1, \ldots, 100$.

Clearly, FGP reaches greater accuracies than those obtain by GP. After 100 iterations FGP reached an accuracy of 10^{-5} while GP reached an accuracy of only $10^{-2.5}$. Moreover, the function value reached by GP at iteration 100 was already obtained by GP after 25 iterations. Another interesting phenomena can be seen at iterations 81 and 82 and is marked on the figure. As opposed to the GP method, FGP is not a monotone method. This does not have an influence on the convergence of the sequence and we see that in most iterations there is a decrease in the function value. In the next section, we will see that this non-monotonicity

phenomena can have a severe impact on the convergence of a related two-steps method for the image deblurring problem.

1.7.3 TV-based Deblurring

Consider now the TV-based deblurring optimization model

$$\min \|\mathcal{A}(\mathbf{x}) - \mathbf{b}\|_F^2 + 2\lambda \mathrm{TV}(\mathbf{x}), \tag{1.54}$$

where $\mathbf{x} \in \mathbb{R}^{m\times n}$ is the original image to be restored, $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ is a linear transformation representing some blurring operator, $\mathbf{b}$ is the noisy and blurred image, and $\lambda > 0$ is a regularization parameter. Obviously problem (1.54) is within the setting of the general model (M) with

$$f(\mathbf{x}) = \|\mathcal{A}(\mathbf{x}) - \mathbf{b}\|^2, \quad g(\mathbf{x}) = 2\lambda \mathrm{TV}(\mathbf{x}), \text{ and } \mathbb{E} = \mathbb{R}^{m\times n}.$$

Deblurring is of course more challenging than denoising. Indeed, to construct an equivalent smooth optimization problem for (1.54) via its dual along the approach of Section 1.7.2, it is easy to realize that one would need to invert the operator $\mathcal{A}^T\mathcal{A}$, which is clearly an ill-posed problem, i.e., such an approach is not viable. This is in sharp contrast to the denoising problem, where a smooth dual problem was constructed, and was the basis of efficient solution methods. Instead, we suggest to solve the deblurring problem by the fast proximal gradient method. Each iteration of the method will require the computation of the prox map which in this case amounts to solving a denoising problem. More precisely, if denote the optimal solution of the constrained *denoising* problem (1.49) with observed image $\mathbf{b}$, regularization parameter λ by $D_C(\mathbf{b}, \lambda)$, then with this notation, the prox-grad map $p_L(\cdot)$ can be simply written as:

$$p_L(\mathbf{Y}) = D_C\left(\mathbf{Y} - \frac{2}{L}\mathcal{A}^T(\mathcal{A}(\mathbf{Y}) - \mathbf{b}), \frac{2\lambda}{L}\right).$$

Thus, each iteration involves the solution of a subproblem that should be solved using an iterative method such as GP or FGP. Note also that this is in contrast to the situation with the simpler l_1-based regularization problem where ISTA or FISTA requires only the computation of a gradient step and a shrinkage, which in that case is an *explicit* operation, see Section 1.6. The fact that the prox operation does not have an explicit expression but is rather computed via an iterative algorithm can have a profound impact on the performance of the method. This is illustrated in the following section.

1.7.4 Numerical Example

Consider a 64×64 image that was cut from the cameraman test image (whose pixels are scaled to be between 0 and 1). The image goes through a Gaussian blur of size 9×9 and standard deviation 4 followed by a an additive zero-mean white Gaussian noise with standard deviation 10^{-2}. The regularization parameter

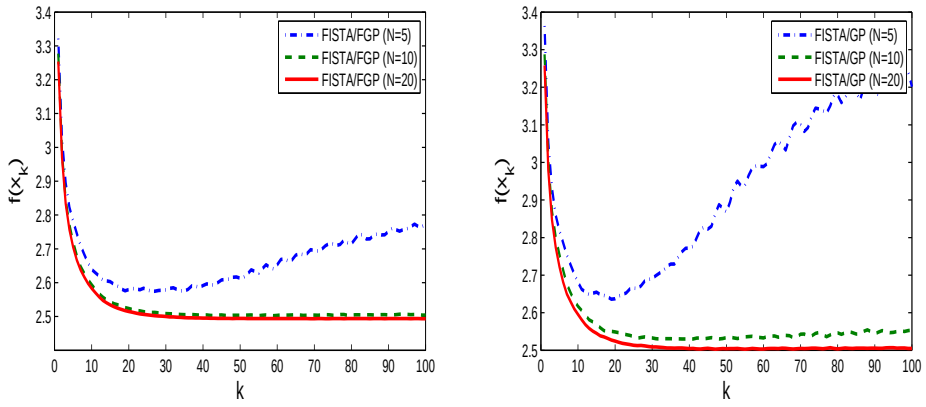

Figure 1.5 Function values of the first 100 iterations of FISTA. The denoising subproblems are solved using FGP (left image) or GP (right image) with $N = 5, 10, 20$.

λ is chosen to be 0.01. We adopt the same terminology used for the l_1-based regularization and use the name ISTA the proximal gradient method and the name FISTA for the fast proximal gradient method.

Figure 1.5 presents three graphs showing the function values of the FISTA method applied to (1.54) in which the denoising subproblems are solved using FGP with number of FGP iterations, denoted by N, taking the values $5, 10, 20$. In the left image the denoising subproblems are solved using FGP and in the right image the denoising subproblems are solved using GP. Clearly FISTA in combination with either GP or FGP diverges when $N = 5$, although it seems that the combination FISTA/GP is worse than FISTA/FGP. For $N = 10$ FISTA/FGP seems to converge to a value which is a bit higher than the one obtained by the same method with $N = 20$ and FISTA/GP with $N = 10$ is still very much erratic and does not seem to converge.

From this example we can conclude that (1) FISTA can diverge when the subproblems are not solved exactly and (2) the combination FISTA/FGP seems to be better than FISTA/GP. The latter conclusion is another numerical evidence (in addition to the results of Section 1.6) to the superiority of FGP over GP. The first conclusion motivates us to use the monotone version of FISTA, which we term MFISTA and which was introduced in Section 1.5.3 for the general model (M). We ran MFISTA on the exact same problem and the results are shown in Figure 1.6. Clearly the monotone version of FISTA seems much more robust and stable. Therefore, it seems that there is a clear advantage in using MFISTA instead of FISTA when the prox map cannot be computed exactly.

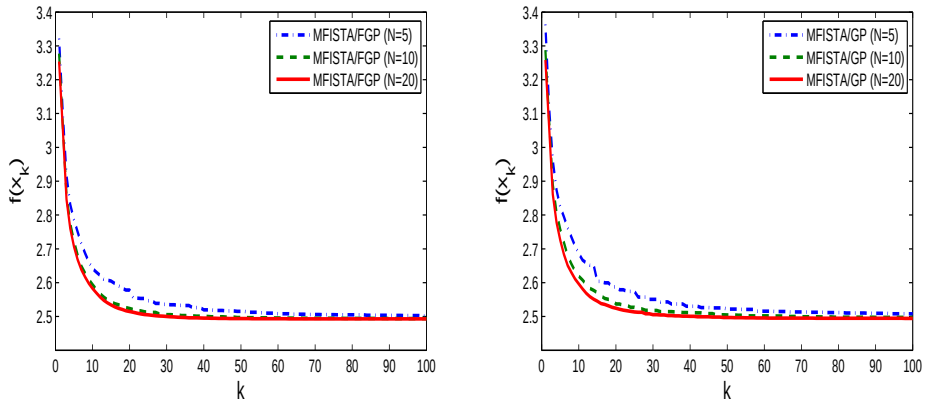

Figure 1.6 Function values of the first 100 iterations of MFISTA. The denoising subproblems are solved using either FGP (left image) or GP (right image) with $N = 5, 10, 20$.

1.8 The Source Localization Problem

1.8.1 Problem Formulation

Consider the problem of locating a single radiating source from noisy range measurements collected using a network of passive sensors. More precisely, consider an array of m sensors, and let $\mathbf{a}_j \in \mathbb{R}^n$ denote the coordinates of the jth sensor[6]. Let $\mathbf{x} \in \mathbb{R}^n$ denote the unknown source's coordinate vector, and let $d_j > 0$ be a noisy observation of the range between the source and the jth sensor:

$$d_j = \|\mathbf{x} - \mathbf{a}_j\| + \varepsilon_j, \quad j = 1, \ldots, m, \tag{1.55}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)^T$ denotes the unknown noise vector. Such observations can be obtained for example from the time-of-arrival measurements in a constant-velocity propagation medium. The source localization problem is the following:

The Source Localization Problem: Given the observed range measurements $d_j > 0$, find a "good" approximation of the source $\mathbf{x}$.

The source localization problem has received significant attention in the signal processing literature and specifically in the field of mobile phones localization.

There are many possible mathematical formulations for the source localization problem. A natural and common approach is to consider a least squares criterion

[6] in practical applications $n = 2$ or 3.

in which the optimization problems seeks to minimize the squared sum of errors:

$$\text{(SL):} \quad \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ f(\mathbf{x}) \equiv \sum_{j=1}^{m} (\|\mathbf{x}-\mathbf{a}_j\| - d_j)^2 \right\}. \tag{1.56}$$

The above criterion also has a statistical interpretation. When $\boldsymbol{\varepsilon}$ follows a Gaussian distribution with a covariance matrix proportional to the identity matrix, the optimal solution of (SL) is in fact the maximum likelihood estimate.

The SL problem is a nonsmooth nonconvex problem and as such is not an easy problem to solve. In the following we will show how to construct two simple methods using the concepts explained in Section 1.3. The derivation of the algorithms is inspired by Weiszfeld's algorithm for the Fermat-Weber problem which was described in Section 1.3.4. Throughout this section, we denote the set of sensors by $\mathcal{A} := \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$.

1.8.2 The Simple Fixed Point Algorithm: Definition and Analysis

Similarly to Weiszfeld's method, our starting point for constructing a fixed point algorithm to solve the SL problem is to write the optimality condition and "extract" $\mathbf{x}$. Assuming that $\mathbf{x} \notin \mathcal{A}$ we have that $\mathbf{x}$ is a stationary point for problem (SL) if and only if

$$\nabla f(\mathbf{x}) = 2\sum_{j=1}^{m} (\|\mathbf{x}-\mathbf{a}_j\| - d_j)\frac{\mathbf{x}-\mathbf{a}_j}{\|\mathbf{x}-\mathbf{a}_j\|} = \mathbf{0}, \tag{1.57}$$

which can be written as

$$\mathbf{x} = \frac{1}{m}\left\{ \sum_{j=1}^{m} \mathbf{a}_j + \sum_{j=1}^{m} d_j \frac{\mathbf{x}-\mathbf{a}_j}{\|\mathbf{x}-\mathbf{a}_j\|} \right\}.$$

The latter relation calls for the following fixed point algorithm which we term the *standard fixed point (SFP) scheme*:

Algorithm SFP:

$$\mathbf{x}_k = \frac{1}{m}\left\{ \sum_{j=1}^{m} \mathbf{a}_j + \sum_{j=1}^{m} d_j \frac{\mathbf{x}_{k-1}-\mathbf{a}_j}{\|\mathbf{x}_{k-1}-\mathbf{a}_j\|} \right\}, \quad k \geq 1. \tag{1.58}$$

Like in Weiszfeld's algorithm, the SFP scheme is not well defined if $\mathbf{x}_k \in \mathcal{A}$ for some k. In the sequel we will state a result claiming that by carefully selecting the initial vector $\mathbf{x}_0$ we can *guarantee* that the iterates are not in the sensors set $\mathcal{A}$, therefore establishing that the method is well defined.
Before proceeding with the analysis of the SFP method, we record the fact that the SFP scheme is actually a gradient method with a fixed step size.

Proposition 1.2. *Let $\{\mathbf{x}_k\}$ be the sequence generated by the SFP method (1.58) and suppose that $\mathbf{x}_k \notin \mathcal{A}$ for all $k \geq 0$. Then for every $k \geq 1$:*

$$\mathbf{x}_k = \mathbf{x}_{k-1} - \frac{1}{2m}\nabla f(\mathbf{x}_{k-1}). \tag{1.59}$$

Proof. Follows by a straightforward calculation, using the gradient of f computed in (1.57). □

It is interesting to note that the SFP method belongs to the class of MM methods (see Section 1.3.3). That is, there exists a function $h(\cdot,\cdot)$ such that $h(\mathbf{x},\mathbf{y}) \geq f(\mathbf{x})$ and $h(\mathbf{x},\mathbf{x}) = f(\mathbf{x})$ for which

$$\mathbf{x}_k = \operatorname*{argmin}_{\mathbf{x}} h(\mathbf{x}, \mathbf{x}_{k-1}).$$

The only departure from the philosophy of MM methods is that special care should be given to the sensors set $\mathcal{A}$. We define the auxiliary function h as

$$h(\mathbf{x},\mathbf{y}) \equiv \sum_{j=1}^{m} \|\mathbf{x} - \mathbf{a}_j - d_j r_j(\mathbf{y})\|^2, \quad \text{for every } \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^n \setminus \mathcal{A}, \tag{1.60}$$

where

$$r_j(\mathbf{y}) \equiv \frac{\mathbf{y} - \mathbf{a}_j}{\|\mathbf{y} - \mathbf{a}_j\|}, \quad j = 1, \ldots, m.$$

Note that for every $\mathbf{y} \notin \mathcal{A}$, the following relations hold for every $j = 1, \ldots, m$:

$$\|r_j(\mathbf{y})\| = 1, \tag{1.61}$$

$$(\mathbf{y} - \mathbf{a}_j)^T r_j(\mathbf{y}) = \|\mathbf{y} - \mathbf{a}_j\|. \tag{1.62}$$

In Lemma 1.10 below, we prove the key properties of the auxiliary function h defined in (1.60). These properties verify the fact that this is in fact an MM method.

Lemma 1.10. *(a) $h(\mathbf{x},\mathbf{x}) = f(\mathbf{x})$ for every $\mathbf{x} \notin \mathcal{A}$.*
(b) $h(\mathbf{x},\mathbf{y}) \geq f(\mathbf{x})$ for every $\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^n \setminus \mathcal{A}$.
(c) If $\mathbf{y} \notin \mathcal{A}$ then

$$\mathbf{y} - \frac{1}{2m}\nabla f(\mathbf{y}) = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} h(\mathbf{x},\mathbf{y}). \tag{1.63}$$

Proof. (a) For every $\mathbf{x} \notin \mathcal{A}$,

$$\begin{aligned} f(\mathbf{x}) &= \sum_{j=1}^{m} (\|\mathbf{x}-\mathbf{a}_j\| - d_j)^2 \\ &= \sum_{j=1}^{m} (\|\mathbf{x}-\mathbf{a}_j\|^2 - 2d_j\|\mathbf{x}-\mathbf{a}_j\| + d_j^2) \\ &\overset{(1.61),(1.62)}{=} \sum_{j=1}^{m} (\|\mathbf{x}-\mathbf{a}_j\|^2 - 2d_j(\mathbf{x}-\mathbf{a}_j)^T r_j(\mathbf{x}) + d_j^2\|r_j(\mathbf{x})\|^2) = h(\mathbf{x},\mathbf{x}), \end{aligned}$$

where the last equation follows from (1.60).
(b) Using the definition of f and h given respectively in (1.56),(1.60), and the fact (1.61), a short computation shows that for every $\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^n \setminus \mathcal{A}$,

$$h(\mathbf{x},\mathbf{y}) - f(\mathbf{x}) = 2\sum_{j=1}^{m} d_j \left(\|\mathbf{x}-\mathbf{a}_j\| - (\mathbf{x}-\mathbf{a}_j)^T r_j(\mathbf{y})\right) \geq 0,$$

where the last inequality follows from Cauchy-Schwartz inequality and using again (1.61).
(c) For any $\mathbf{y} \in \mathbb{R}^n \setminus \mathcal{A}$, the function $\mathbf{x} \mapsto h(\mathbf{x},\mathbf{y})$ is strictly convex on $\mathbb{R}^n$, and consequently admits a unique minimizer $\mathbf{x}^*$ satisfying

$$\nabla_{\mathbf{x}} h(\mathbf{x}^*,\mathbf{y}) = \mathbf{0}.$$

Using the definition of h given in (1.60), the latter identity can be explicitly written as

$$\sum_{j=1}^{m} (\mathbf{x}^* - \mathbf{a}_j - d_j r_j(\mathbf{y})) = \mathbf{0},$$

which by simple algebraic manipulation can be shown to be equivalent to $\mathbf{x}^* = \mathbf{y} - \frac{1}{2m}\nabla f(\mathbf{y})$. □

By the properties just established it follows that the SFP method is an MM method and as such is a descent scheme. It is also possible to prove the convergence result given in Theorem 1.6 below. Not surprisingly, since the problem is nonconvex, only convergence to stationary points is established.

Theorem 1.6 (Convergence of the SFP Method). *Let $\{\mathbf{x}_k\}$ be generated by (1.58) such that $\mathbf{x}_0$ satisfies*

$$f(\mathbf{x}_0) < \min_{j=1,\ldots,m} f(\mathbf{a}_j). \tag{1.64}$$

Then,

(a) $\mathbf{x}_k \notin \mathcal{A}$ for every $k \geq 0$.
(b) For every $k \geq 1$, $f(\mathbf{x}_k) \leq f(\mathbf{x}_{k-1})$ and equality is satisfied if and only if $\mathbf{x}_k = \mathbf{x}_{k-1}$.
(c) The sequence of function values $\{f(\mathbf{x}_k)\}$ converges.

(d) The sequence $\{\mathbf{x}_k\}$ *is bounded.*
(e) Any limit point of $\{\mathbf{x}_k\}$ *is a stationary point of* f*.*

The condition (1.64) is very mild in the sense that it is not difficult to find an initial vector $\mathbf{x}_0$ satisfying it, see bibliographic notes for details.

Next we show how to construct a different method for solving the source localization problem using a completely different approximating auxiliary function.

1.8.3 The SWLS Algorithm

To motivate the construction of the second method, let us first go back again to Weiszfeld's scheme, and recall that Weiszfeld's method can also be written as (see also (1.15)):

$$\mathbf{x}_k = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} h(\mathbf{x}, \mathbf{x}_{k-1}),$$

where

$$h(\mathbf{x}, \mathbf{y}) \equiv \sum_{j=1}^{m} \omega_j \frac{\|\mathbf{x}-\mathbf{a}_j\|^2}{\|\mathbf{y}-\mathbf{a}_j\|} \text{ for every } \mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^n\setminus\mathcal{A}.$$

The auxiliary function h was essentially constructed from the objective function of the Fermat-Weber location problem, by replacing the norm terms $\|\mathbf{x}-\mathbf{a}_j\|$ with $\frac{\|\mathbf{x}-\mathbf{a}_j\|^2}{\|\mathbf{y}-\mathbf{a}_j\|}$. Mimicking this observation for the SL problem under study, we will use an auxiliary function in which each norm term $\|\mathbf{x}-\mathbf{a}_j\|$ in the objective function (1.56) is replaced with $\frac{\|\mathbf{x}-\mathbf{a}_j\|^2}{\|\mathbf{y}-\mathbf{a}_j\|}$, resulting in the following auxiliary function:

$$g(\mathbf{x}, \mathbf{y}) \equiv \sum_{i=1}^{m} \left(\frac{\|\mathbf{x}-\mathbf{a}_i\|^2}{\|\mathbf{y}-\mathbf{a}_i\|} - d_i \right)^2, \quad \mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^n\setminus\mathcal{A}. \tag{1.65}$$

The general step of the algorithm for solving problem (SL), termed *the sequential weighted least squares* (SWLS) method, is now given by

$$\mathbf{x}_k \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} g(\mathbf{x}, \mathbf{x}_{k-1}).$$

or more explicitly by

Algorithm SWLS:

$$\mathbf{x}_k \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \sum_{j=1}^{m} \left(\frac{\|\mathbf{x}-\mathbf{a}_j\|^2}{\|\mathbf{x}_{k-1}-\mathbf{a}_j\|} - d_j \right)^2. \tag{1.66}$$

The name SWLS stems from the fact that at each iteration k we are required to solve the following weighted nonlinear least squares problem:

$$\text{(NLS):} \quad \min_{\mathbf{x}} \sum_{j=1}^{m} \omega_j^k (\|\mathbf{x} - \mathbf{c}_j\|^2 - \beta_j^k)^2, \tag{1.67}$$

with

$$\mathbf{c}_j = \mathbf{a}_j, \beta_j^k = d_j \|\mathbf{x}_{k-1} - \mathbf{a}_j\|, \omega_j^k = \frac{1}{\|\mathbf{x}_{k-1} - \mathbf{a}_j\|^2}. \tag{1.68}$$

Note that the SWLS algorithm as presented above is not defined for iterations in which $\mathbf{x}_{k-1} \in \mathcal{A}$. However, like in the SFP method, it is possible to find an initial point ensuring that the iterates are not in the sensor set $\mathcal{A}$.

The NLS problem is a nonconvex problem, but it can still be solved globally and efficiently by transforming it into a problem of minimizing a quadratic function subject to a single quadratic constraint. Indeed, for a given fixed k, we can transform (1.67) into a constrained minimization problem (the index k is omitted):

$$\min_{\mathbf{x}\in\mathbb{R}^n, \alpha\in\mathbb{R}} \left\{ \sum_{j=1}^{m} \omega_j (\alpha - 2\mathbf{c}_j^T \mathbf{x} + \|\mathbf{c}_j\|^2 - \beta_j)^2 : \|\mathbf{x}\|^2 = \alpha \right\}, \tag{1.69}$$

which can also be written as (using the substitution $\mathbf{y} = (\mathbf{x}^T, \alpha)^T$)

$$\min_{\mathbf{y}\in\mathbb{R}^{n+1}} \left\{ \|\mathbf{A}\mathbf{y} - \mathbf{b}\|^2 : \mathbf{y}^T \mathbf{D} \mathbf{y} + 2\mathbf{f}^T \mathbf{y} = 0 \right\}, \tag{1.70}$$

where

$$\mathbf{A} = \begin{pmatrix} -2\sqrt{\omega_1}\mathbf{c}_1^T & \sqrt{\omega_1} \\ \vdots & \vdots \\ -2\sqrt{\omega_m}\mathbf{c}_m^T & \sqrt{\omega_m} \end{pmatrix}, \mathbf{b} = \begin{pmatrix} \sqrt{\omega_1}(\beta_1 - \|\mathbf{c}_1\|^2) \\ \vdots \\ \sqrt{\omega_m}(\beta_m - \|\mathbf{c}_m\|^2) \end{pmatrix}$$

and

$$\mathbf{D} = \begin{pmatrix} \mathbf{I}_n & \mathbf{0}_{n\times 1} \\ \mathbf{0}_{1\times n} & 0 \end{pmatrix}, \mathbf{f} = \begin{pmatrix} 0 \\ -0.5 \end{pmatrix}.$$

Problem(1.70) belongs to the class of problems consisting of minimizing a quadratic function subject to a single quadratic constraint (without any convexity assumptions). Problems of this type are called generalized trust region subproblems (GTRS). GTRS problems possess necessary and sufficient optimality conditions from which efficient solution methods can be derived.

The analysis of the SWLS method is more complicated than the analysis of the SFP method and we only state the main results. For the theoretical convergence analysis two rather mild assumptions are required:

Assumption 1. *The matrix*

$$\mathbf{A} = \begin{pmatrix} 1 & \mathbf{a}_1^T \\ 1 & \mathbf{a}_2^T \\ \vdots & \vdots \\ 1 & \mathbf{a}_m^T \end{pmatrix}$$

is of full column rank.

For example, when $n = 2$, the assumption states that $\mathbf{a}_1, \ldots, \mathbf{a}_m$ are not on the same line. The second assumption states that the value of the initial vector $\mathbf{x}_0$ is "small enough".

Assumption 2. $f(\mathbf{x}_0) < \frac{\min_j \{d_j\}^2}{4}$.

A similar assumption was made for the SFP method (see condition (1.64)). Note that for the true source location $\mathbf{x}_{\text{true}}$ one has $f(\mathbf{x}_{\text{true}}) = \sum_{j=1}^m \varepsilon_j^2$. Therefore, $\mathbf{x}_{\text{true}}$ satisfies Assumption 2 if the errors ε_j are smaller in some sense from the range measurements d_j. This is a very reasonable assumption since in real applications the errors ε_i are often smaller in an order of magnitude than d_i. Now, if the initial point $\mathbf{x}_0$ is "good enough" in the sense that it is close to the true source location, then Assumption 2 will be satisfied.

Under the above assumption it is possible to prove the following key properties of the auxiliary function g:

$$\begin{aligned} g(\mathbf{x}, \mathbf{x}) &= f(\mathbf{x}), \text{ for every } \mathbf{x} \in \mathbb{R}^n, \\ g(\mathbf{x}, \mathbf{x}_{k-1}) &\geq 2f(\mathbf{x}) - f(\mathbf{x}_{k-1}), \text{ for every } \mathbf{x} \in \mathbb{R}^n, k \geq 1. \end{aligned} \tag{1.71}$$

Therefore, the SWLS method, as opposed to the SFP method, is not an MM method since the auxiliary function $g(\cdot, \cdot)$ is not an upper bound on the objective function. However, similarly to Weiszfeld's method for the Fermat-Weber problem (see Section 1.15), the property (1.71) implies the descent property of the SWLS method and it can also be used in order to prove the convergence result of the method which is given below.

Theorem 1.7 (Convergence of the SWLS Method). *Let $\{\mathbf{x}_k\}$ be the sequence generated by the SWLS method. Suppose that Assumptions 1 and 2 hold true. Then*

(a) $\mathbf{x}_k \notin \mathcal{A}$ *for* $k \geq 0$.
(b) For every $k \geq 1$, $f(\mathbf{x}_k) \leq f(\mathbf{x}_{k-1})$ *and equality holds if and only if* $\mathbf{x}_k = \mathbf{x}_{k-1}$.
(c) The sequence of function values $\{f(\mathbf{x}^k)\}$ *converges.*
(d) The sequence $\{\mathbf{x}^k\}$ *is bounded.*
(e) Any limit point of $\{\mathbf{x}^k\}$ *is a stationary point of* f.

1.9 Bibliographic Notes

Section 1.2 The class of optimization problems (M) has been first studied in [2] and provide a natural vehicle to study various generic optimization models under a common framework. Linear inverse problems arise in a wide range of diverse applications, and the literature is vast, see the monograph [26] and references therein. A popular regularization technique is the Tikhonov smooth quadratic regularization [56] which has been extensively studied and extended, see for instance [35, 36, 33]. Early works promoting the use of the convex nonsmooth l_1 regularization appear for example in [20, 34, 18]. The l_1 regularization has now attracted an intensive revived interest in the signal processing literature, in particular in compressed sensing, and which has led to a large amount of literature see e.g., the recent works [11, 24, 30] and references therein.

Section 1.3 The gradient method is one of the very first method for unconstrained minimization going back to 1847 with the work of Cauchy [12]. Gradient methods and their variants have been studied by many authors. We mention in particular the classical works developed in the 60's and 70's by [32, 1, 41, 47, 22], and for more modern presentations with many results, including the extension to problems with constraints and the resulting gradient projection method given in 1.3.1, see the books of [49, 9, 44] and references therein. The quadratic approximation model in 1.3.1 is a very well known interpretation of the gradient method as a proximal regularization of the linearized part of a differentiable function f [49]. The proximal map was introduced by Moreau [42]. The extension to handle the nonsmooth model (M), as given in Sections 1.3.1 and 1.3.2 is a special case of the proximal-forward backward method for finding the zero of the sum of two maximal monotone operators, originally proposed by [48]. The terminology "proximal gradient" is used to emphasize the specific composite operation when applied to a minimization problem of the form (M). The majorization-minimization idea discussed in 1.3.3 has been developed by many authors, and for a recent tutorial on MM algorithms, applications and many references see [38], and for its use in signal processing see for instance, [23, 28]. The material on the Fermat-Weber location problem presented in Section 1.3.4 is classical. The original Weiszfeld algorithm can be found in [58] and further analyzed in [40]. It has been intensively and further studied in the location theory literature, see the monograph [50].

Section 1.4 For simplicity of exposition, we assumed the existence of an optimal solution for the optimization model (M). For classical and more advanced techniques to handle existence of minimizers, see [3]. The proximal map and regularization of a closed proper convex function is due to Moreau, see [42] for the proof of Lemma 1.2. The results of Section 1.4.2 are well known and can be found in [49, 9]. Lemma 1.6 is a slight modification of a recent result proven in [7]. The material of Section 1.4.3 follows [7], except for the pointwise convergence Theorem 1.2. More general convergence results of the sequence $\mathbf{x}_k$ can be

found in [27], and in particular we refer to the comprehensive recent work [21]. The nonconvex case in Theorem 1.3 seems to be new and naturally extends the known result [44] for the smooth unconstrained case, cf. Remark 1.3.

Section 1.5 A well known and popular gradient method based on two steps memory is the conjugate gradient algorithm, see e.g. [49, 9]. In the nonsmooth case, a similar idea was developed by Shor with the R-algorithm [54]. However, such methods do not appear to improve the complexity rate of basic gradient-like methods. This goal was achieved by Nesterov [45] who was the first to introduce a new idea and algorithm for minimizing a smooth convex function proven to be an *optimal gradient* method in the sense of complexity analysis [43]. This algorithm was recently extended to the convex nonsmooth model (M) in [7] and all the material in this section is from [7], except for the nonmomotone case and Theorem 1.5 which was very recently developed in [6]. For recent alternative gradient-based methods, including methods based on two or more gradient steps, and that could speed-up the proximal gradient method for solving the special case of model (M) with f being the least squares objective, see for instance [10, 25, 30]. The speedup gained by these methods has been shown through numerical experiments, but global nonasymptotic rate of convergence results have not been established. In the recent work of [46], a multistep fast gradient method that solves model (M) has been developed and proven to share the same complexity rate $O(1/k^2)$ as derived here. The new method of [46] is remarkably different conceptually and computationally from the fast proximal gradient method; it uses accumulated history of past iterates, and requires two projection-like operations per iteration, see [46, 7] for more details. For a recent study on a gradient scheme based on non Euclidean distances for solving smooth conic convex problems and which shares the same fast complexity rate, see the recent work [4].

Section 1.6 The quadratic l_1-based regularization model has attracted a considerable amount of attention in the signal processing literature, see for example [15, 29, 23] for the iterative shrinkage/thresholding algorithm (ISTA) and for more recent works, including new algorithms, applications and many pointers to relevant literature, [21, 25, 30, 39]. The results and examples presented in Section 1.6.3 are from the recent work [7] where more details, references and examples are given.

Section 1.7 The total variation (TV) based model has been introduced by [52]. The literature on numerical methods for solving model (1.48) abounds. To mention just few, see for instance [57, 16, 17, 31, 37]. This list is just given as an indicator of the intense research in the field and far from being comprehensive. The work of Chambolle [13, 14] is of particular interest. There, he introduced and developed a globally convergent gradient dual based algorithm for the denoising problem, which was shown faster than primal-based schemes. His works motivated our recent analysis and algorithmic developments given in [6] for the more involved constrained TV-based deblurring problem, which when combined with

FISTA produces fast gradient methods. This section have presented some results and numerical examples from [6] to which we refer the reader for further reading.

Section 1.8 The single source localization problem has received significant attention in the field of signal processing, see e.g. [5, 19, 55, 53] and references therein. The algorithms and results given in this section are taken from the recent work [8], where more details, results and proofs of the theorems can be found.

References

[1] L. Armijo. Minimization of functions having continuous partial derivatives. *Pacific J. Math.*, 16:1–3, 1966.

[2] A. Auslender. *Minimisation de fonctions localement Lipschitziennes: applications a la programmation mi-convexe, mi-differentiable.* in Nonlnear Programming 3, Eds. O. L. Mangasarian and R. R. Meyer and S. M. Robinson, Academic Press, New York, pp. 429–460, 1978.

[3] A. Auslender and M. Teboulle. *Asymptotic Cones and Functions in Optimization and Variational Inequalities.* Springer Monographs in Mathematics. New York: Springer, 2003.

[4] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optimization*, 16(3):697–725, 2006.

[5] A. Beck, P. Stoica, and J. Li. Exact and approximate solutions of source localization problems. *IEEE Trans. Signal Processing*, 56(5):1770–1778, 2008.

[6] A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *Submitted for Publication*, 2008.

[7] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences, accepted for publication*, 2008.

[8] A. Beck, M. Teboulle, and Z. Chikichev. Iterative minimization schemes for solving the single source localization problem. *SIAM J. Optimization, to appear*, 2008.

[9] D. P. Bertsekas. *Nonlinear Programming.* Belmont MA: Athena Scientific, second edition, 1999.

[10] J. Bioucas-Dias and M. Figueiredo. A new TwIST: two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Trans. on Image Processing*, 16:2992–3004, 2007.

[11] E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006.

[12] A-L. Cauchy. Méthode generales pour la résolution des systèmes d'equations simultanées. *Comptes Rendues Acad. Sc. Paris*, 25:536–538, 1847.

[13] A. Chambolle. An algorithm for total variation minimization and applications. *J. Math. Imaging Vision*, 20(1-2):89–97, 2004. Special issue on mathematics and image analysis.

[14] A. Chambolle. Total variation minimization and a class of binary MRF models. In *Lecture Notes in Computer Sciences*, volume 3757, pages 136–152, 2005.

[15] A. Chambolle, R. A. DeVore, N. Y. Lee, and B. J. Lucier. Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Trans. Image Processing*, 7:319–335, 1998.

[16] A. Chambolle and P. L. Lions. Image recovery via total variation minimization and related problmes. *Numerische Mathematik*, 76:167–188, 1997.

[17] T. F. Chan, G. H. Golub, and P. Mulet. A nonlinear primal-dual method for total variation-based image restoration. *SIAM J. Sci. Comput.*, 20(6):1964–1977 (electronic), 1999.

[18] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61 (electronic), 1998.

[19] K. W. Cheung, W. K. Ma, and H. C. So. Accurate approximation algorithm for TOA-based maximum likelihood mobile location using semidefinite programming. In *Proc. ICASSP*, volume 2, pages 145–148, 2004.

[20] J. Claerbout and F. Muir. Robust modelling of erratic data. *Geophysics*, 38:826–844, 1973.

[21] P. Combettes and V. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation*, 4:1168–1200, 2005.

[22] J. W. Daniel. *The Approximate Minimization of Functionals*. Prentice-Hall, Englewood Cliffs, N.J., 1971.

[23] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004.

[24] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[25] M. Elad, B. Matalon, J. Shtok, and M. Zibulevsky. A wide-angle view at iterated shrinkage algorithms. In *SPIE (Wavelet XII) 2007, San-Diego CA*, August 26-29, 2007.

[26] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[27] F. Facchinei and J. S. Pang. *Finite-dimensional variational inequalities and complementarity problems, Vol. II*. Springer Series in Operations Research. Springer-Verlag, New York, 2003.

[28] M. A. T. Figueiredo, J. Bioucas-Dias, and R. Nowak. Majorization-minimization algorithms for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 16(12):2980–2991.

[29] M. A. T. Figueiredo and R. D. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Trans. Image Processing*, 12(8):906–916, 2003.

[30] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. To appear in IEEE J. Selected Topics in Signal Processing, 2007.

[31] D. Goldfarb and W. Yin. Second-order cone programming methods for total variation-based image restoration. *SIAM Journal on Scientific Computing*, pages 622–645, 2005.

[32] A. A. Goldstein. Cauchy's method for minimization. *Numerisch Math.*, 4:146–150, 1962.

[33] G. H. Golub, P. C. Hansen, and D. P. O'Leary. Tikhonov regularization and total least squares. *SIAM J. Matrix Anal. Appl.*, 21(2):185–194, 1999.

[34] S. Bank H. Taylor and J. McCoy. Deconvolution with the l_1-norm. *Geophysics*, 44:39–52, 1979.

[35] M. Hanke and P. C. Hansen. Regularization methods for large-scale problems. *Surveys Math. Indust.*, 3(4):253–315, 1993.

[36] P. C. Hansen. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Stat. Comput.*, 14:1487–1503, 1993.

[37] M. Hintermuller and G. Stadler. An infeasible primal-dual algorithm for tv-based infconvolution-type image restoration. *SIAM Journal on Scientific Computing*, 28:1–23, 2006.

[38] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004.

[39] S. J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. *A Method for Large-Scale l1-Regularized Least Squares*, volume 1. December 2007.

[40] H. W. Kuhn. A note on Fermat's problem. *Math. Programming*, 4:98–107, 1973.

[41] E. S. Levitin and B. T. Polyak. Constrained minimization methods. *USSR Computational Math. and Math. Phys.*, 6:787–823, 1966.

[42] J. J. Moreau. Proximitéet dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93:273–299, 1965.

[43] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization.* A Wiley-Interscience Publication. John Wiley & Sons Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[44] Y. Nesterov. *Introductory Lectures on Convex Optimization.* Kluwer, Boston, 2004.

[45] Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983.

[46] Y. E. Nesterov. Gradient methods for minimizing composite objective function. 2007. CORE Report. Available at http://www.ecore.beDPs/dp1191313936.pdf.

[47] J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*, volume 30 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000. Reprint of the 1970 original.

[48] G. B. Passty. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *J. Math. Anal. Appl.*, 72(2):383–390, 1979.

[49] B. T. Polyak. *Introduction to optimization*. Translations Series in Mathematics and Engineering. Optimization Software Inc. Publications Division, New York, 1987. Translated from the Russian, With a foreword by Dimitri P. Bertsekas.

[50] J. G. Morris R. F. Love and G. O. Wesolowsky. *Facilities location: Models and Methods*. North-Holland Publishing Co., New York, 1988.

[51] R. T. Rockafellar. *Convex Analysis*. Princeton NJ: Princeton Univ. Press, 1970.

[52] L. I. Rudin, S. J. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[53] A. H. Sayed, A. Tarighat, and N. Khajehnouri. Network-based wireless location. *IEEE Signal Processing Mag.*, 22(4):24–40, July 2005.

[54] N. Z. Shor. *Minimization Methods for Nondifferentiable Functions*, volume 3 of *Springer Series in Computational Mathematics*. Springer-Verlag, 1985.

[55] J. O. Smith and J. S. Abel. Closed-form least-squares source location estimation from range-difference measurements. *IEEE Trans. Acoustics, Speech and Signal Processing*, 12:1661–1669, Dec. 1987.

[56] A. N. Tikhonov and V. Y. Arsenin. *Solution of Ill-Posed Problems*. Washington, DC: V.H. Winston, 1977.

[57] C. R. Vogel and M. E. Oman. Iterative methods for total variation denoising. *SIAM Journal of Scientific Computing*, 17:227–238, 1996.

[58] E. Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal*, 43:355–386, 1937.

Index