A First Order Method for Finding Minimal Norm-Like Solutions of Convex Optimization Problems

Amir Beck and Shoham Sabach

January 14, 2014

Abstract

We consider a general class of convex optimization problems in which one seeks to minimize a strongly convex function over a closed and convex set which is by itself an optimal set of another convex problem. We introduce a gradient-based method, called the minimal norm gradient method, for solving this class of problems, and establish the convergence of the sequence generated by the algorithm as well as a rate of convergence of the sequence of function values. Several numerical examples are given in order to illustrate our results.

1 Introduction

1.1 Problem Formulation

Consider the general convex constrained optimization problem given by

$$\text{(P):} \quad \begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in X, \end{array}$$

where the following assumptions are made throughout the paper:

- X is a nonempty, closed and convex subset of $\mathbb{R}^n$.
- The objective function f is convex and continuously differentiable over $\mathbb{R}^n$, and its gradient is Lipschitz with constant L:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{1.1}$$

- The optimal set of (P), denoted by X^*, is nonempty. The optimal value is denoted by f^*.

Problem (P) might have multiple optimal solutions, and in this case it is natural to consider the *minimal norm solution problem* in which one seeks to find the optimal solution of (P) with a minimal Euclidean norm:

$$\text{(Q):} \quad \min \left\{ \frac{1}{2} \|\mathbf{x}\|^2 : \mathbf{x} \in X^* \right\}.$$

We will denote the optimal solution of (Q) by $\mathbf{x}^*_{\mathrm{Q}}$. A well-known approach to tackling problem (Q) is via the celebrated Tikhonov regularization. More precisely, for a given $\varepsilon > 0$, consider the convex problem defined by

$$(\mathrm{Q}_\varepsilon): \quad \min\left\{ f(\mathbf{x}) + \frac{\varepsilon}{2}\|\mathbf{x}\|^2 : \mathbf{x} \in X \right\}.$$

The above problem is the so-called *Tikhonov regularized problem* [**?**]. Let us denote the unique optimal solution of (Q_ε) by $\mathbf{x}^\varepsilon$. In [**?**], Tikhonov showed in the linear case – that is, when f is a linear function and X is an intersection of halfspaces – that $\mathbf{x}^\varepsilon \to \mathbf{x}^*_{\mathrm{Q}}$ as $\varepsilon \to 0^+$. Therefore, for a small enough $\varepsilon > 0$, the vector $\mathbf{x}^\varepsilon$ can be considered as an approximation of the minimal norm solution $\mathbf{x}^*_{\mathrm{Q}}$. A stronger result in the linear case showing that for a small enough ε, $\mathbf{x}^\varepsilon$ is in fact *exactly the same* as $\mathbf{x}^*_{\mathrm{Q}}$ was established in [**?**] and was later on generalized to the more general convex case in [**?**]. Further generalization for the convex case without differentiability assumptions can be found in [**?**] as well as a wealth of relevant references.

From a practical point of view, the connection just alluded between the minimal norm solution and the solutions of the Tikhonov regularized problems, does not yield an explicit algorithm for solving (Q). It is not clear how to choose an appropriate sequence of regularization parameters $\varepsilon_k \to 0^+$, and how to solve the emerging subproblems. An exception can be found in the work [**?**] where it was shown that if an associated optimization problem possess a Lagrange multiplier, then an explicit expression for an exact regularization parameter (in terms of the Lagrange multiplier) exists. A different approach for solving (Q) in the linear case was developed in [**?**] where it was suggested to invoke a Newton-type method for solving a reformulation of (Q) as an unconstrained smooth minimization problem.

The main contribution of this paper is the construction and analysis of a new first order method for solving a generalization of problem (Q), which we call *the minimal norm-like solution problem* (MNP). Problem (MNP) consists of finding the optimal solution of problem (P) which minimizes a given strongly convex function ω:

$$(\mathrm{MNP}): \quad \min\{\omega(\mathbf{x}) : \mathbf{x} \in X^*\}.$$

The function ω is assumed to satisfy the following:

- ω is a strongly convex function over $\mathbb{R}^n$ with parameter $\sigma > 0$.
- ω is a continuously differentiable function.

By the strong convexity of ω, problem (MNP) has a unique solution which will be denoted by $\mathbf{x}^*_{\mathrm{mn}}$.

For simplicity, problem (P) will be called the *core problem*, problem (MNP) will be called *the outer problem* and correspondingly, ω will be called the *outer objective function*. It is obvious that problem (Q) is a special case of problem (MNP) with the choice $\omega(\mathbf{x}) \equiv \frac{1}{2}\|\mathbf{x}\|^2$. The so-called *prox center* of ω is given by

$$\mathbf{a} \equiv \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \omega(\mathbf{x}).$$

We assume without loss of generality that $\omega(\mathbf{a}) = 0$. Under this setting we also have

$$\omega(\mathbf{x}) \geq \frac{\sigma}{2}\|\mathbf{x} - \mathbf{a}\|^2 \text{ for all } \mathbf{x} \in \mathbb{R}^n. \tag{1.2}$$

1.2 Stage by Stage Solution

It is important to note that the minimal norm-like solution optimization problem (MNP) can also be formally cast as the following convex optimization problem:

$$\begin{array}{ll} \min & \omega(\mathbf{x}) \\ \text{s.t.} & f(\mathbf{x}) \leq f^*, \\ & \mathbf{x} \in X. \end{array} \tag{1.3}$$

Of course, the optimal value of the core problem f^* is not known in advance, which suggests a solution method that consists of two stages: first find the optimal value of the core problem, and then solve problem (**??**). This two-stage solution technique has two main drawbacks. First, the optimal value f^* is often not found *exactly* but rather up to some tolerance, which causes the feasible set of the outer problem to be incorrect or even infeasible. Second, even if it would have been possible to compute f^* exactly, problem (**??**) inherently does not satisfy Slater's condition, which means that this two-stage approach will usually run into numerical problems. We note that the lack of regularity condition for problem (**??**) implies that known optimality conditions such as Karush-Kuhn-Tucker are not valid; see for example the work [**?**] where different optimality conditions are derived.

It is interesting to note that the minimal norm-like optimization problem is a special case of the more general class of *bilevel programming* problems, for more details see the survey paper [**?**] and many references therein.

1.3 Paper Layout

This paper presents a first order method, called *the minimal norm gradient method*, aimed at solving the minimal norm-like solution problem (MNP). As opposed to the above mentioned method, the suggested method is an iterative algorithm that solves problem (MNP) *directly* and not "stage by stage" or via a solution of a sequence of related optimization problems. In Section **??** the required mathematical background on orthogonal projections, gradient mappings and cutting planes is presented. The minimal norm gradient method is derived and analyzed in Section **??**. At each iteration, the required computations are (i) a gradient evaluation of the core objective function, (ii) an orthogonal projection onto the feasible set of the core problem and (iii) a solution of a problem consisting of minimizing the outer objective function subject to the intersection of two halfspaces. In Section **??** the convergence of the sequence generated by the method is established along with an $O(1/\sqrt{k})$ convergence of the sequence of function values (k being the iteration index). Finally, in Section **??**, a numerical example of a portfolio optimization is described as well as a set of test cases arising from ill-conditioned inverse problems arising from discretizations of Fredholm integral equations of the first kind.

2 Mathematical Toolbox

2.1 The Orthogonal Projection

The orthogonal projection operator onto a given closed and convex set $S \subseteq \mathbb{R}^n$ is denoted by

$$P_S(\mathbf{x}) \equiv \underset{\mathbf{y}\in S}{\operatorname{argmin}} \|\mathbf{x}-\mathbf{y}\|^2 .$$

The orthogonal projection operator possesses several important properties; two of them will be useful in our analysis, and are thus recalled here.

- **Nonexpansive:**

$$\|P_S(\mathbf{x}) - P_S(\mathbf{y})\| \leq \|\mathbf{x}-\mathbf{y}\| \text{ for all } \mathbf{x},\mathbf{y} \in \mathbb{R}^n.$$

- **Firmly Nonexpansive:**

$$\langle P_S(\mathbf{x}) - P_S(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \geq \|P_S(\mathbf{x}) - P_S(\mathbf{y})\|^2 \text{ for all } \mathbf{x},\mathbf{y} \in \mathbb{R}^n. \tag{2.1}$$

In addition, the operator P_S is characterized by the following inequality (see, e.g., [**?**])

$$\langle \mathbf{x} - P_S(\mathbf{x}), \mathbf{y} - P_S(\mathbf{x})\rangle \leq 0 \text{ for all } \mathbf{x} \in \mathbb{R}^n,\ \mathbf{y} \in S. \tag{2.2}$$

2.2 Bregman Distances

The definition of a Bregman distance associated with a given strictly convex function is given below.

Definition 2.1 (Bregman distance)**.** Let $h : \mathbb{R}^n \to \mathbb{R}$ be a strictly convex function. The *Bregman distance* is the following bifunction

$$D_h(\mathbf{x},\mathbf{y}) := h(\mathbf{x}) - h(\mathbf{y}) - \langle \nabla h(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle . \tag{2.3}$$

Two basic properties of bregman distances are:

- $D_h(\mathbf{x},\mathbf{y}) \geq 0$ for any $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$.
- $D_h(\mathbf{x},\mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$.

If, in addition, h is *strongly convex* with parameter $\sigma > 0$, then

$$D_h(\mathbf{x},\mathbf{y}) \geq \frac{\sigma}{2}\|\mathbf{x}-\mathbf{y}\|^2 .$$

In particular, the strongly convex function ω defined in Section **??** whose prox center is $\mathbf{a}$ satisfies:

$$\omega(\mathbf{x}) = D_\omega(\mathbf{x},\mathbf{a}) \geq \frac{\sigma}{2}\|\mathbf{x}-\mathbf{a}\|^2 \text{ for any } \mathbf{x} \in \mathbb{R}^n$$

and

$$D_\omega(\mathbf{x},\mathbf{y}) \geq \frac{\sigma}{2}\|\mathbf{x}-\mathbf{y}\|^2 \text{ for any } \mathbf{x},\mathbf{y} \in \mathbb{R}^n. \tag{2.4}$$

One of the fundamental identities that will be used in our analysis is the following "three point identity" which is stated via the terminology used in this paper.

Lemma 2.1 (Three point identity [**?**])**.** *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a strongly convex function with strong convexity parameter $\sigma > 0$. Then for any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$:*

$$D_h(\mathbf{x}, \mathbf{y}) + D_h(\mathbf{y}, \mathbf{z}) - D_h(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x} - \mathbf{y}, \nabla h(\mathbf{z}) - \nabla h(\mathbf{y}) \rangle . \tag{2.5}$$

2.3 The Gradient Mapping

We define the following two mappings which are essential in our analysis of the proposed algorithm for solving (MNP).

Definition 2.2. For every $M > 0$,

(i) the *proj-grad mapping* is defined by

$$T_M(\mathbf{x}) \equiv P_X\left(\mathbf{x} - \frac{1}{M}\nabla f(\mathbf{x})\right) \text{ for all } \mathbf{x} \in \mathbb{R}^n.$$

(ii) the *gradient mapping* (see also [**?**]) is defined by

$$G_M(\mathbf{x}) \equiv M(\mathbf{x} - T_M(\mathbf{x})) = M\left[\mathbf{x} - P_X\left(\mathbf{x} - \frac{1}{M}\nabla f(\mathbf{x})\right)\right].$$

Remark 2.1 (Unconstrained case)**.** In the unconstrained setting, that is, when $X = \mathbb{R}^n$, the orthogonal projection is the identity operator and hence

(i) The proj-grad mapping T_M is equal to $I - \frac{1}{M}\nabla f$.

(ii) The gradient mapping G_M is equal to ∇f.

It is well known that $G_M(\mathbf{x}) = \mathbf{0}$ if and only if $\mathbf{x} \in X^*$. Another important and known property of the gradient mapping is the monotonicity of its norm with respect to M (see [**?**, Lemma 2.3.1, p. 236]).

Lemma 2.2. *For any $\mathbf{x} \in \mathbb{R}^n$, the function*

$$g(M) \equiv \|G_M(\mathbf{x})\| \qquad M > 0$$

is monotonically nondecreasing over $(0, \infty)$.

2.4 Cutting Planes

The notion of a *cutting plane* is a fundamental concept in optimization algorithms such as the ellipsoid and analytic cutting plane methods. As an illustration, let us first consider the unconstrained setting in which $X = \mathbb{R}^n$. Given a point $\mathbf{x} \in \mathbb{R}^n$, the idea is to find a hyperplane which separates $\mathbf{x}$ from X^*. For example, it is well known that for any $\mathbf{x} \in X$, the following inclusion holds:

$$X^* \subseteq \{\mathbf{z} \in \mathbb{R}^n : \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{z} \rangle \geq 0\}.$$

The importance of the above result is that it "eliminates" the open halfspace

$$\{\mathbf{z} \in \mathbb{R}^n : \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{z} \rangle < 0\}.$$

The same cut is also used in the ellipsoid method where in the nonsmooth case the gradient is replaced with a subgradient (see, e.g., [**?**, **?**]). Note that $\mathbf{x}$ belongs to the cut, that is, to the hyperplane given by:

$$H = \{\mathbf{z} \in \mathbb{R}^n : \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{z} \rangle = 0\},$$

which means that H is a so-called *neutral cut*. In a *deep* cut, the point $\mathbf{x}$ does not belong to the corresponding hyperplane. Deep cuts are at the core of the minimal norm gradient method that will be described in the sequel, and in this subsection we describe how to construct them in several scenarios (specifically, known/unknown Lipschitz constant, constrained/unconstrained versions). The halfspaces corresponding to the deep cuts are always of the form

$$Q_{M,\alpha,\mathbf{x}} \equiv \left\{\mathbf{z} \in \mathbb{R}^n : \langle G_M(\mathbf{x}), \mathbf{x} - \mathbf{z} \rangle \geq \frac{1}{\alpha M} \|G_M(\mathbf{x})\|^2\right\}, \tag{2.6}$$

where the values of α and M depend on the specific scenario. Of course, in the unconstrained case, $G_M(\mathbf{x}) \equiv \nabla f(\mathbf{x})$, and (**??**) reads as

$$Q_{M,\alpha,\mathbf{x}} \equiv \left\{\mathbf{z} \in \mathbb{R}^n : \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{z} \rangle \geq \frac{1}{\alpha M} \|\nabla f(\mathbf{x})\|^2\right\}.$$

We will now split the analysis into two scenarios. In the first one, the Lipschitz constant L is known, while in the second, it is not.

2.4.1 Known Lipschitz Constant

In the unconstrained case ($X = \mathbb{R}^n$), and when the Lipschitz constant L is known, we can use the following known inequality (see, e.g., [**?**]):

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \text{ for every } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{2.7}$$

By plugging $\mathbf{y} = \mathbf{x}^*$ for some $\mathbf{x}^* \in X^*$ in (**??**) and recalling that $\nabla f(\mathbf{x}^*) = \mathbf{0}$, we obtain that

$$\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle \geq \frac{1}{L} \|\nabla f(\mathbf{x})\|^2 \tag{2.8}$$

for every $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{x}^* \in X^*$. Thus, $X^* \subseteq Q_{L,1,\mathbf{x}}$ for any $\mathbf{x} \in \mathbb{R}^n$.

When X is not the entire space $\mathbb{R}^n$, the generalization of (**??**) is a bit intricate and in fact the result we can prove is the slightly "weaker" inclusion $X^* \subseteq Q_{L,\frac{4}{3},\mathbf{x}}$. The result is based on the following property of the gradient mapping G_L which was proven in the thesis [**?**] and is given here for the sake of completeness.

Lemma 2.3. *The gradient mapping G_L satisfies the following relation:*

$$\langle G_L(\mathbf{x}) - G_L(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{3}{4L} \|G_L(\mathbf{x}) - G_L(\mathbf{y})\|^2 \tag{2.9}$$

for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Proof. From (**??**) it follows that

$$\left\langle T_L(\mathbf{x}) - T_L(\mathbf{y}), \left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right) - \left(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\right)\right\rangle \geq \|T_L(\mathbf{x}) - T_L(\mathbf{y})\|^2.$$

Since $T_L = I - \frac{1}{L}G_L$, we obtain that

$$\begin{aligned}&\left\langle \left(\mathbf{x} - \frac{1}{L}G_L(\mathbf{x})\right) - \left(\mathbf{y} - \frac{1}{L}G_L(\mathbf{y})\right), \left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right) - \left(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\right)\right\rangle \\ &\geq \left\| \left(\mathbf{x} - \frac{1}{L}G_L(\mathbf{x})\right) - \left(\mathbf{y} - \frac{1}{L}G_L(\mathbf{y})\right)\right\|^2,\end{aligned}$$

which is equivalent to

$$\left\langle \left(\mathbf{x} - \frac{1}{L}G_L(\mathbf{x})\right) - \left(\mathbf{y} - \frac{1}{L}G_L(\mathbf{y})\right), (G_L(\mathbf{x}) - \nabla f(\mathbf{x})) - (G_L(\mathbf{y}) - \nabla f(\mathbf{y}))\right\rangle \geq 0. \tag{2.10}$$

Thence

$$\begin{aligned}\langle G_L(\mathbf{x}) - G_L(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \;\geq\; &\frac{1}{L}\|G_L(\mathbf{x}) - G_L(\mathbf{y})\|^2 + \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \\ &-\frac{1}{L}\langle G_L(\mathbf{x}) - G_L(\mathbf{y}), \nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\rangle.\end{aligned}$$

Now it follows from (**??**) that

$$\begin{aligned}L\langle G_L(\mathbf{x}) - G_L(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \;\geq\; &\|G_L(\mathbf{x}) - G_L(\mathbf{y})\|^2 + \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \\ &\langle G_L(\mathbf{x}) - G_L(\mathbf{y}), \nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\rangle.\end{aligned}$$

From the Cauchy-Schwarz inequality we get

$$\begin{aligned}L\langle G_L(\mathbf{x}) - G_L(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \;\geq\; &\|G_L(\mathbf{x}) - G_L(\mathbf{y})\|^2 + \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \\ &- \|G_L(\mathbf{x}) - G_L(\mathbf{y})\| \cdot \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|.\end{aligned} \tag{2.11}$$

By denoting $\alpha = \|G_L(\mathbf{x}) - G_L(\mathbf{y})\|$ and $\beta = \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|$, the right-hand side of (**??**) reads as $\alpha^2 + \beta^2 - \alpha\beta$ and satisfies:

$$\alpha^2 + \beta^2 - \alpha\beta = \frac{3}{4}\alpha^2 + \left(\frac{\alpha}{2} - \beta\right)^2 \geq \frac{3}{4}\alpha^2,$$

which combined with (**??**) yields the inequality

$$L\langle G_L(\mathbf{x}) - G_L(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \geq \frac{3}{4}\|G_L(\mathbf{x}) - G_L(\mathbf{y})\|^2.$$

Thus, (**??**) holds. □

By plugging $\mathbf{y} = \mathbf{x}^*$, for some $\mathbf{x}^* \in X^*$ in (**??**) we obtain that indeed

$$X^* \subseteq Q_{L,\frac{4}{3},\mathbf{x}}.$$

We summarize the above discussion in the following lemma which describes the deep cuts in the case when the Lipschitz constant is known.

Lemma 2.4. *For any $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{x}^* \in X^*$, we have*

$$\langle G_L(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle \geq \frac{3}{4L} \|G_L(\mathbf{x})\|^2, \tag{2.12}$$

that is,

$$X^* \subseteq Q_{L,\frac{4}{3},\mathbf{x}}.$$

If, in addition, $X = \mathbb{R}^n$ then

$$\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle \geq \frac{1}{L} \|\nabla f(\mathbf{x})\|^2, \tag{2.13}$$

that is,

$$X^* \subseteq Q_{L,1,\mathbf{x}}.$$

2.4.2 Unknown Lipschitz Constant

When the Lipschitz constant is not known, the following result is most useful.

Lemma 2.5. *Let $\mathbf{x} \in \mathbb{R}^n$ be a vector satisfying the inequality*

$$f(T_M(\mathbf{x})) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), T_M(\mathbf{x}) - \mathbf{x} \rangle + \frac{M}{2} \|T_M(\mathbf{x}) - \mathbf{x}\|^2. \tag{2.14}$$

Then, for any $\mathbf{x}^* \in X^*$, the inequality

$$\langle G_M(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle \geq \frac{1}{2M} \|G_M(\mathbf{x})\|^2 \tag{2.15}$$

holds true, that is,

$$X^* \subseteq Q_{M,2,\mathbf{x}}.$$

Proof. Let $\mathbf{x}^* \in X^*$. By (**??**) it follows that

$$0 \leq f(T_M(\mathbf{x})) - f(\mathbf{x}^*) \leq f(\mathbf{x}) - f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}), T_M(\mathbf{x}) - \mathbf{x} \rangle + \frac{M}{2} \|T_M(\mathbf{x}) - \mathbf{x}\|^2. \tag{2.16}$$

Since f is convex, $f(\mathbf{x}) - f(\mathbf{x}^*) \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle$, which combined with (**??**) yields

$$0 \leq \langle \nabla f(\mathbf{x}), T_M(\mathbf{x}) - \mathbf{x}^* \rangle + \frac{M}{2} \|T_M(\mathbf{x}) - \mathbf{x}\|^2. \tag{2.17}$$

In addition, by the definition of T_M and (**??**) we have the following inequality:

$$\left\langle \mathbf{x} - \frac{1}{M} \nabla f(\mathbf{x}) - T_M(\mathbf{x}), T_M(\mathbf{x}) - \mathbf{x}^* \right\rangle \geq 0.$$

Summing the latter inequality multiplied by M with (**??**) yields the inequality

$$M \langle \mathbf{x} - T_M(\mathbf{x}), T_M(\mathbf{x}) - \mathbf{x}^* \rangle + \frac{M}{2} \|T_M(\mathbf{x}) - \mathbf{x}\|^2 \geq 0,$$

which after some simple algebraic manipulation, can be shown to be equivalent to the desired result (**??**). □

When $M \geq L$, the inequality (**??**) is satisfied due to the so-called descent lemma, which is now recalled as it will also be essential in our analysis (see [**?**]).

Lemma 2.6 (Descent Lemma)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function whose gradient is Lipschitz with constant L. Then for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,*

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2. \tag{2.18}$$

Remark 2.2. The inequality (**??**) for $M \geq L$ is well known, see for example [**?**].

3 The Minimal Norm Gradient Algorithm

Before describing the algorithm, we require the following notation for the optimal solution of the problem consisting of minimizing ω over a given closed and convex set S:

$$\Omega(S) \equiv \operatorname*{argmin}_{\mathbf{x} \in S} \omega(\mathbf{x}). \tag{3.1}$$

By the optimality condition in problem (**??**), it follows that

$$\tilde{\mathbf{x}} = \Omega(S) \Leftrightarrow \langle \nabla \omega(\tilde{\mathbf{x}}), \mathbf{x} - \tilde{\mathbf{x}} \rangle \geq 0 \text{ for all } \mathbf{x} \in S. \tag{3.2}$$

When $\omega(\mathbf{x}) = \frac{1}{2} \|\mathbf{x} - \mathbf{a}\|^2$, then $\Omega(S) = P_S(\mathbf{a})$.

We are now ready to describe the algorithm in the case when the Lipschitz constant L is known.

The Minimal Norm Gradient Method (Known Lipschitz Constant)

Input: L - a Lipschitz constant of ∇f.
Initialization: $\mathbf{x}_0 = \mathbf{a}$.
General Step ($\mathbf{k = 1, 2, \ldots}$):

$$\mathbf{x}_k = \Omega\left(Q^k \cap W^k\right),$$

where

$$\begin{aligned} Q^k &= Q_{L,\beta,\mathbf{x}_{k-1}}, \\ W^k &= \{\mathbf{z} \in \mathbb{R}^n : \langle \nabla \omega(\mathbf{x}_{k-1}), \mathbf{z} - \mathbf{x}_{k-1} \rangle \geq 0\}, \end{aligned}$$

and β is equal to $\frac{4}{3}$ if $X \neq \mathbb{R}^n$ and to 1 if $X = \mathbb{R}^n$.

When the Lipschitz constant is unknown, then a backtracking procedure should be incorporated into the method.

The Minimal Norm Gradient Method (Unknown Lipschitz Constant)

Input: $L_0 > 0, \eta > 1$.
Initialization: $\mathbf{x}_0 = \mathbf{a}$.
General Step ($\mathbf{k = 1\ ,\ 2\ ,\ \ldots}$):

- Find the smallest nonnegative integer number i_k such that with $\bar{L} = \eta^{i_k} L_{k-1}$ the inequality

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

  is satisfied and set $L_k = \bar{L}$.

- Set

$$\mathbf{x}_k = \Omega\left(Q^k \cap W^k\right),$$

  where

$$\begin{aligned} Q^k &= Q_{L_k, 2, \mathbf{x}_{k-1}}, \\ W^k &= \left\{\mathbf{z} \in \mathbb{R}^n : \left\langle \nabla \omega\left(\mathbf{x}_{k-1}\right), \mathbf{z} - \mathbf{x}_{k-1} \right\rangle \geq 0\right\}. \end{aligned}$$

To unify the analysis, in the constant stepsize setting we will artificially define $L_k = L$ for any k and $\eta = 1$. In this notation the definition of the halfspace Q^k in both the constant and backtracking stepsize rules can be described as

$$Q^k = Q_{L_k, \beta, \mathbf{x}_{k-1}}, \tag{3.3}$$

where β is given by

$$\beta \equiv \begin{cases} \frac{4}{3} & X \neq \mathbb{R}^n, \text{known Lipschitz const.} \\ 1 & X = \mathbb{R}^n, \text{known Lipschitz const.} \\ 2 & \text{unknown Lipschitz const.} \end{cases} \tag{3.4}$$

Remark 3.1. By the definition of the backtracking rule it follows that

$$L_0 \leq L_k \leq \eta L, \quad k = 0, 1, 2, \ldots. \tag{3.5}$$

Therefore, it follows from Lemma **??** that for any $\mathbf{x} \in \mathbb{R}^n$,

$$\left\| G_{L_0}\left(\mathbf{x}\right) \right\| \leq \left\| G_{L_k}\left(\mathbf{x}\right) \right\| \leq \left\| G_{\eta L}\left(\mathbf{x}\right) \right\|. \tag{3.6}$$

The following example shows that in the Euclidean setting, the main step has a simple and explicit formula.

Example 1. In the Euclidean setting when $\omega = \frac{1}{2}\|\cdot\|^2$, we have $\Omega(S) = P_S$ and the computation of the main step

$$\mathbf{x}_k = \Omega\left(Q^k \cap W^k\right)$$

boils down to finding the orthogonal projection onto an intersection of two halfspaces. This is a simple task, since the orthogonal projection onto the intersection of two halfspaces:

$$T = \{\mathbf{x} \in \mathbb{R}^n : \langle \mathbf{a}_1, \mathbf{x}\rangle \leq b_1, \langle \mathbf{a}_2, \mathbf{x}\rangle \leq b_2\} \quad (\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^n, b_1, b_2 \in \mathbb{R})$$

is given by the following explicit formula:

$$P_T(\mathbf{x}) = \begin{cases} \mathbf{x}, & \alpha \leq 0 \text{ and } \beta \leq 0, \\ \mathbf{x} - (\beta/\nu)\,\mathbf{a}_2, & \alpha \leq \pi\,(\beta/\nu) \text{ and } \beta > 0, \\ \mathbf{x} - (\alpha/\mu)\,\mathbf{a}_1, & \beta \leq \pi\,(\alpha/\mu) \text{ and } \alpha > 0, \\ \mathbf{x} + (\alpha/\rho)\,(\pi\mathbf{a}_2 - \nu\mathbf{a}_1) + (\beta/\rho)\,(\pi\mathbf{a}_1 - \mu\mathbf{a}_2), & \text{otherwise,} \end{cases}$$

where here

$$\pi = \langle \mathbf{a}_1, \mathbf{a}_2\rangle,\ \mu = \|\mathbf{a}_1\|^2,\ \nu = \|\mathbf{a}_2\|^2,\ \rho = \mu\nu - \pi^2$$

and

$$\alpha = \langle \mathbf{a}_1, \mathbf{x}\rangle - b_1 \text{ and } \beta = \langle \mathbf{a}_2, \mathbf{x}\rangle - b_2.$$

Note that the algorithm is well defined as long as the set $Q^k \cap W^k$ is nonempty. The latter property does hold true and we will now show a stronger result stating that in fact $X^* \subseteq Q^k \cap W^k$ for all k.

Lemma 3.1. *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the minimal norm gradient method with either a constant or a backtracking stepsize rule. Then*

$$X^* \subseteq Q^k \cap W^k \tag{3.7}$$

for any $k = 1, 2, \ldots$.

Proof. By Lemmata **??** and **??** it follows that $X^* \subseteq Q^k$ for every $k = 1, 2, \ldots$ and we will now prove by induction on k that $X^* \subseteq W^k$. Since $W^1 = \mathbb{R}^n$, the claim is trivial for $k = 1$. Suppose that the claim holds for $k = n$, that is, we assume that $X^* \subseteq W^n$. To prove that $X^* \subseteq Q^{n+1} \cap W^{n+1}$, let us take $\mathbf{u} \in X^*$. Note that $X^* \subseteq Q^n \cap W^n$, and thus, since $\mathbf{x}_n = \Omega(Q^n \cap W^n)$, it follows from (**??**) that

$$\langle \nabla\omega(\mathbf{x}_n), \mathbf{x}_n - \mathbf{u}\rangle \leq 0.$$

This implies that $\mathbf{u} \in W^{n+1}$ and the claim that $X^* \subseteq Q^k \cap W^k$ for all k is proven. □

Remark 3.2. We note that the minimal norm gradient method requires the computation of the gradient mapping at each iteration, meaning in particular that the orthogonal projection onto the set X is computed at each iteration. Therefore, to apply the method, it is assumed that the set X is "simple" enough so that computing the orthogonal projection onto it is an easy task.

4 Convergence Analysis

Our first claim is that the minimal norm gradient method generates a sequence $\{\mathbf{x}_k\}_{k\geq 0}$ which converges to $\mathbf{x}^*_{\text{mn}} = \Omega(X^*)$.

Theorem 4.1 (Sequence Convergence)**.** *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the minimal norm gradient method with either a constant or a backtracking stepsize rule. Then*

(i). The sequence $\{\mathbf{x}_k\}_{k\geq 0}$ is bounded.

(ii). The following inequality holds for any $k = 1, 2, \ldots$:

$$D_\omega(\mathbf{x}_k, \mathbf{x}_{k-1}) + D_\omega(\mathbf{x}_{k-1}, \mathbf{a}) \leq D_\omega(\mathbf{x}_k, \mathbf{a}). \tag{4.1}$$

(iii). $\mathbf{x}_k \to \mathbf{x}^*_{\text{mn}}$ as $k \to \infty$.

Proof. **(i).** Since $\mathbf{x}_k = \Omega\left(Q^k \cap W^k\right)$, it follows that for any $\mathbf{u} \in Q^k \cap W^k$, and in particular for any $\mathbf{u} \in X^*$

$$\omega(\mathbf{x}_k) \leq \omega(\mathbf{u}), \tag{4.2}$$

which combined with (**??**) establishes the boundedness of $\{\mathbf{x}_k\}_{k\geq 0}$.
(ii). By the three point identity (see Lemma **??**) we have

$$D_\omega(\mathbf{x}_k, \mathbf{x}_{k-1}) + D_\omega(\mathbf{x}_{k-1}, \mathbf{a}) - D_\omega(\mathbf{x}_k, \mathbf{a}) = \langle -\nabla\omega(\mathbf{x}_{k-1}), \mathbf{x}_k - \mathbf{x}_{k-1}\rangle.$$

By the definition of W^k we have $\mathbf{x}_{k-1} = \Omega\left(W^k\right)$. In addition, $\mathbf{x}_k \in W^k$, and hence by (**??**) it follows that

$$\langle \nabla\omega(\mathbf{x}_{k-1}), \mathbf{x}_k - \mathbf{x}_{k-1}\rangle \geq 0$$

and therefore (**??**) follows.
(iii). Recall that for any $\mathbf{x} \in \mathbb{R}^n$, we have $D_\omega(\mathbf{x}, \mathbf{a}) = \omega(\mathbf{x})$. By (**??**) it follows that the sequence $\{\omega(\mathbf{x}_k)\}_{k\geq 0} = \{D_\omega(\mathbf{x}_k, \mathbf{a})\}_{k\geq 0}$ is nondecreasing and bounded, and hence $\lim_{k\to\infty} \omega(\mathbf{x}_k)$ exists. This, together with (**??**) implies that

$$\lim_{k\to\infty} D_\omega(\mathbf{x}_k, \mathbf{x}_{k-1}) = 0,$$

and hence, since $D_\omega(\mathbf{x}_k, \mathbf{x}_{k-1}) \geq \frac{\sigma}{2}\|\mathbf{x}_k - \mathbf{x}_{k-1}\|^2$, it follows that

$$\lim_{k\to\infty} \|\mathbf{x}_k - \mathbf{x}_{k-1}\| = 0. \tag{4.3}$$

Since $\mathbf{x}_k \in Q^k$ we have

$$\langle G_{L_k}(\mathbf{x}_{k-1}), \mathbf{x}_{k-1} - \mathbf{x}_k\rangle \geq \frac{1}{\beta L_k}\|G_{L_k}(\mathbf{x}_{k-1})\|^2,$$

which by the Cauchy-Schwarz inequality, implies that

$$\frac{1}{\beta L_k}\|G_{L_k}(\mathbf{x}_{k-1})\| \leq \|\mathbf{x}_{k-1} - \mathbf{x}_k\|.$$

Now, from (**??**) and (**??**) it follows that

$$\frac{1}{\eta L}\left\|G_{L_0}\left(\mathbf{x}_{k-1}\right)\right\| \leq \left\|\mathbf{x}_{k-1}-\mathbf{x}_k\right\|. \tag{4.4}$$

To show that $\{\mathbf{x}_k\}_{k\geq 0}$ converges to $\mathbf{x}^*_{\text{mn}}$, it is enough to show that any convergent subsequence converges to $\mathbf{x}^*_{\text{mn}}$. Let then $\{\mathbf{x}_{k_n}\}_{n\geq 0}$ be a convergent subsequence whose limit is $\mathbf{w}$. From (**??**) and (**??**) along with the continuity of G_{L_0}, it follows that $G_{L_0}(\mathbf{w})=0$, so that $\mathbf{w}\in X^*$. Finally, we will prove that $\mathbf{w}=\Omega(X^*)=\mathbf{x}^*_{\text{mn}}$. Since $\mathbf{x}_{k_n}=\Omega\left(Q^{k_n}\cap W^{k_n}\right)$, it follows by (**??**):

$$\left\langle \nabla\omega\left(\mathbf{x}_{k_n}\right), \mathbf{z}-\mathbf{x}_{k_n}\right\rangle \geq 0 \quad \text{for all} \quad \mathbf{z}\in Q^{k_n}\cap W^{k_n}.$$

Since $X^*\subseteq Q^{k_n}\cap W^{k_n}$ (see Lemma **??**), we obtain that

$$\left\langle \nabla\omega\left(\mathbf{x}_{k_n}\right), \mathbf{z}-\mathbf{x}_{k_n}\right\rangle \geq 0 \quad \text{for all} \quad \mathbf{z}\in X^*.$$

Taking the limit as $n\to\infty$, and using the continuity of $\nabla\omega$, we get

$$\left\langle \nabla\omega\left(\mathbf{w}\right), \mathbf{z}-\mathbf{w}\right\rangle \geq 0 \quad \text{for all} \quad \mathbf{z}\in X^*.$$

Therefore, it follows from (**??**) that $\mathbf{w}=\Omega(X^*)=\mathbf{x}^*_{\text{mn}}$, and the result is proven. □

The next result shows that in the unconstrained case ($X=\mathbb{R}^n$), the function values of the sequence generated by the minimal norm gradient method, $\{f(\mathbf{x}_k)\}$, converges with a rate of $O\left(1/\sqrt{k}\right)$ (k being the iteration index) to the optimal value of the core problem. In the constrained case, the value $f(\mathbf{x}_k)$ is by no means a measure of the quality of the iterate $\mathbf{x}_k$ as it is not necessarily feasible. Instead, we will show that the rate of convergence of the function values of the *feasible* sequence $T_{L_k}(\mathbf{x}_k)$ (which in any case is computed by the algorithm), is also $O\left(1/\sqrt{k}\right)$. We also note that since the minimal norm gradient method is non-monotone, the convergence results are with respect to the minimal function value obtained until iteration k.

Theorem 4.2 (Rate of Convergence)**.** *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the minimal norm gradient method with either a constant or backtracking stepsize rules. Then for every $k\geq 1$, one has*

$$\min_{1\leq n\leq k} f\left(T_{L_n}\left(\mathbf{x}_n\right)\right)-f^* \leq \frac{\beta\eta L\left\|\mathbf{a}-\mathbf{x}^*_{\text{mn}}\right\|^2}{\sqrt{k}}, \tag{4.5}$$

*where β is given in (**??**). If $X=\mathbb{R}^n$, then in addition*

$$\min_{1\leq n\leq k} f\left(\mathbf{x}_n\right)-f^* \leq \frac{\beta\eta L\left\|\mathbf{a}-\mathbf{x}^*_{\text{mn}}\right\|^2}{\sqrt{k}}. \tag{4.6}$$

Proof. Let n be a nonnegative integer. Since $\mathbf{x}_{n+1}\in Q^{n+1}$, we have by the Cauchy-Schwarz inequality:

$$\left\|G_{L_{n+1}}\left(\mathbf{x}_n\right)\right\|^2 \leq \beta L_{n+1}\left\langle G_{L_{n+1}}\left(\mathbf{x}_n\right), \mathbf{x}_n-\mathbf{x}_{n+1}\right\rangle \leq \beta L_{n+1}\left\|G_{L_{n+1}}\left(\mathbf{x}_n\right)\right\|\left\|\mathbf{x}_n-\mathbf{x}_{n+1}\right\|.$$

Therefore,

$$\left\|G_{L_{n+1}}\left(\mathbf{x}_n\right)\right\| \leq \beta L_{n+1}\left\|\mathbf{x}_n-\mathbf{x}_{n+1}\right\|. \tag{4.7}$$

Squaring (**??**) and summing over $n=1,2,\ldots,k$, one obtains

$$\sum_{n=1}^{k}\left\|G_{L_{n+1}}\left(\mathbf{x}_n\right)\right\|^2 \leq \beta^2 L_{n+1}^2 \sum_{n=1}^{N}\left\|\mathbf{x}_{n+1}-\mathbf{x}_n\right\|^2 \leq \beta^2\eta^2 L^2 \sum_{n=1}^{N}\left\|\mathbf{x}_{n+1}-\mathbf{x}_n\right\|^2. \tag{4.8}$$

Taking into account (**??**) and (**??**), then from (**??**) we get

$$\begin{aligned}
\sum_{n=1}^{k}\left\|G_{L_{n+1}}\left(\mathbf{x}_n\right)\right\|^2 &\leq \beta^2\eta^2 L^2 \sum_{n=1}^{k}\left\|\mathbf{x}_{n+1}-\mathbf{x}_n\right\|^2 \\
&\leq \frac{2\beta^2\eta^2 L^2}{\sigma}\sum_{n=1}^{k} D_\omega\left(\mathbf{x}_{n+1},\mathbf{x}_n\right) \\
&\leq \frac{2\beta^2\eta^2 L^2}{\sigma}\sum_{n=1}^{k}\left(D_\omega\left(\mathbf{x}_{n+1},\mathbf{a}\right)-D_\omega\left(\mathbf{x}_n,\mathbf{a}\right)\right) \\
&= \frac{2\beta^2\eta^2 L^2}{\sigma} D_\omega\left(\mathbf{x}_{k+1},\mathbf{a}\right) = \frac{2\beta^2\eta^2 L^2}{\sigma}\omega\left(\mathbf{x}_{k+1}\right) \\
&\leq \frac{2\beta^2\eta^2 L^2}{\sigma}\omega\left(\mathbf{x}_{\mathrm{mn}}^*\right).
\end{aligned} \tag{4.9}$$

From the definition of L_n,

$$f\left(T_{L_n}\left(\mathbf{x}_n\right)\right)-f^* \leq f\left(\mathbf{x}_n\right)-f^*+\left\langle\nabla f\left(\mathbf{x}_n\right),T_{L_n}\left(\mathbf{x}_n\right)-\mathbf{x}_n\right\rangle+\frac{L_n}{2}\left\|T_{L_n}\left(\mathbf{x}_n\right)-\mathbf{x}_n\right\|^2. \tag{4.10}$$

Since the function f is convex, it follows that $f\left(\mathbf{x}_n\right)-f^* \leq \left\langle\nabla f\left(\mathbf{x}_n\right),\mathbf{x}_n-\mathbf{x}_{\mathrm{mn}}^*\right\rangle$, which combined with (**??**) yields

$$f\left(T_{L_n}\left(\mathbf{x}_n\right)\right)-f^* \leq \left\langle\nabla f\left(\mathbf{x}_n\right),T_{L_n}\left(\mathbf{x}_n\right)-\mathbf{x}_{\mathrm{mn}}^*\right\rangle+\frac{L_n}{2}\left\|T_{L_n}\left(\mathbf{x}_n\right)-\mathbf{x}_n\right\|^2. \tag{4.11}$$

By the characterization of the projection operator given in (**??**) with $\mathbf{x}=\mathbf{x}_n-\frac{1}{L_n}\nabla f\left(\mathbf{x}_n\right)$ and $\mathbf{y}=\mathbf{x}_{\mathrm{mn}}^*$, we have that

$$\left\langle \mathbf{x}_n-\frac{1}{L_n}\nabla f\left(\mathbf{x}_n\right)-T_{L_n}\left(\mathbf{x}_n\right),\mathbf{x}_{\mathrm{mn}}^*-T_{L_n}\left(\mathbf{x}_n\right)\right\rangle \leq 0,$$

which combined with (**??**) gives

$$\begin{aligned}
f\left(T_{L_n}\left(\mathbf{x}_n\right)\right)-f^* &\leq L_n\left\langle\mathbf{x}_n-T_{L_n}\left(\mathbf{x}_n\right),T_{L_n}\left(\mathbf{x}_n\right)-\mathbf{x}_{\mathrm{mn}}^*\right\rangle+\frac{L_n}{2}\left\|T_{L_n}\left(\mathbf{x}_n\right)-\mathbf{x}_n\right\|^2 \\
&= \left\langle G_{L_n}\left(\mathbf{x}_n\right),T_{L_n}\left(\mathbf{x}_n\right)-\mathbf{x}_{\mathrm{mn}}^*\right\rangle+\frac{1}{2L_n}\left\|G_{L_n}(\,\mathbf{x}_n\,)\right\|^2 \\
&= \left\langle G_{L_n}\left(\mathbf{x}_n\right),T_{L_n}\left(\mathbf{x}_n\right)-\mathbf{x}_n\right\rangle+\left\langle G_{L_n}(\mathbf{x}_n),\mathbf{x}_n-\mathbf{x}_{\mathrm{mn}}^*\right\rangle+\frac{1}{2L_n}\left\|G_{L_n}\left(\mathbf{x}_n\right)\right\|^2 \\
&= \left\langle G_{L_n}\left(\mathbf{x}_n\right),\mathbf{x}_n-\mathbf{x}_{\mathrm{mn}}^*\right\rangle-\frac{1}{2L_n}\left\|G_{L_n}\left(\mathbf{x}_n\right)\right\|^2 \\
&\leq \left\langle G_{L_n}\left(\mathbf{x}_n\right),\mathbf{x}_n-\mathbf{x}_{\mathrm{mn}}^*\right\rangle \\
&\leq \left\|G_{L_n}\left(\mathbf{x}_n\right)\right\|\cdot\left\|\mathbf{x}_n-\mathbf{x}_{\mathrm{mn}}^*\right\|.
\end{aligned}$$

Squaring the above inequality and summing over $n = 1, \ldots, k$, we get

$$\sum_{n=1}^{k} \left(f\left(T_{L_n}\left(\mathbf{x}_n\right)\right) - f^*\right)^2 \leq \sum_{n=1}^{k} \left\|G_{L_n}\left(\mathbf{x}_n\right)\right\|^2 \cdot \left\|\mathbf{x}_n - \mathbf{x}_{\mathrm{mn}}^*\right\|^2. \tag{4.12}$$

Now, from the three point identity (see Lemma **??**), we obtain that

$$D_\omega\left(\mathbf{x}_{\mathrm{mn}}^*, \mathbf{x}_n\right) + D_\omega\left(\mathbf{x}_n, \mathbf{a}\right) - D_\omega\left(\mathbf{x}_{\mathrm{mn}}^*, \mathbf{a}\right) = -\left\langle \nabla\omega\left(\mathbf{x}_n\right), \mathbf{x}_{\mathrm{mn}}^* - \mathbf{x}_n\right\rangle \leq 0$$

and hence

$$D_\omega\left(\mathbf{x}_{\mathrm{mn}}^*, \mathbf{x}_n\right) \leq D_\omega\left(\mathbf{x}_{\mathrm{mn}}^*, \mathbf{a}\right) = \omega\left(\mathbf{x}_{\mathrm{mn}}^*\right),$$

so that

$$\left\|\mathbf{x}_n - \mathbf{x}_{\mathrm{mn}}^*\right\|^2 \leq \frac{2\omega\left(\mathbf{x}_{\mathrm{mn}}^*\right)}{\sigma}. \tag{4.13}$$

Combining (**??**) and (**??**) along with (**??**) we get that

$$\sum_{n=1}^{k} \left(f\left(T_{L_n}\left(\mathbf{x}_n\right)\right) - f^*\right)^2 \leq \frac{2\omega\left(\mathbf{x}_{\mathrm{mn}}^*\right)}{\sigma} \sum_{n=1}^{k} \left\|G_{L_n}\left(\mathbf{x}_n\right)\right\|^2 \leq \frac{4\beta^2\eta^2L^2}{\sigma^2}\omega\left(\mathbf{x}_{\mathrm{mn}}^*\right)^2$$

from which we obtain that

$$k \min_{n=1,2,\ldots,k} \left(f\left(T_{L_n}\left(\mathbf{x}_n\right)\right) - f^*\right)^2 \leq \frac{4\beta^2\eta^2L^2}{\sigma^2}\omega\left(\mathbf{x}_{\mathrm{mn}}^*\right)^2,$$

proving the result (**??**). The result (**??**) in the case when $X = \mathbb{R}^n$ is established by following the same line of proof along with the observation that due to the convexity of f

$$f\left(\mathbf{x}_n\right) - f^* \leq \left\|\nabla f\left(\mathbf{x}_n\right)\right\| \cdot \left\|\mathbf{x}_n - \mathbf{x}_{\mathrm{mn}}^*\right\| = \left\|G_{L_n}\left(\mathbf{x}_n\right)\right\| \cdot \left\|\mathbf{x}_n - \mathbf{x}_{\mathrm{mn}}^*\right\|.$$

□

5 Numerical Examples

5.1 Markowitz Portfolio Optimization Model

Consider the Markowitz portfolio optimization problem [**?**]. Suppose that we are given N assets numbered $1, 2, \ldots, N$ for which a vector of expected returns $\boldsymbol{\mu} \in \mathbb{R}^N$ and a positive semidefinite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N}$ are known. In the Markowitz portfolio optimization problem we seek to find a minimum variance portfolio subject to the constraint that the expected return is greater or equal to a certain predefined minimal value r_0:

$$\begin{array}{ll} \min & \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \\ \text{s.t.} & \sum_{i=1}^{N} w_i = 1, \\ & \mathbf{w}^T\boldsymbol{\mu} \geq r_0, \\ & \mathbf{w} \geq \mathbf{0}. \end{array} \tag{5.1}$$

The decision variables vector $\mathbf{w}$ describes the allocation of the given resource to the different assets. When the covariance matrix is rank deficient (that is, positive semidefinite but not positive definite), the optimal solution is not unique, and a natural issue in this scenario is to find one portfolio among all the optimal portfolios that is "best" with respect to an objective function different than the portfolio variance. This is, of course, a minimal norm-like solution optimization problem. We note that the situation in which the covariance matrix is rank deficient is quite common since the covariance matrix is usually estimated from the past trading price data and when the number of sampled periods is smaller than the number of assets, the covariance matrix is surely rank deficient. As a specific example, consider the portfolio optimization problem given by (**??**), where the expected returns vector $\boldsymbol{\mu}$ and convariance matrix $\boldsymbol{\Sigma}$ are both estimated from real data on 8 types of assets ($N = 8$): US 3 month treasury bills, US government long bonds, SP 500, Wilshire 500, NASDAQ composite, corporate bond index, EAFE and Gold. The yearly returns are from 1973 to 1994. The data can be found at http://www.princeton.edu/ rvdb/ampl/nlmodels/markowitz/ and we have used the data between the years 1974 and 1977 in order to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ which are given below:

$$\boldsymbol{\mu} = (1.0630, 1.0633, 1.0670, 1.0853, 1.0882, 1.0778, 1.0820, 1.1605)^T$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.0002 & -0.0005 & -0.0028 & -0.0032 & -0.0039 & -0.0007 & -0.0024 & 0.0048 \\ -0.0005 & 0.0061 & 0.0132 & 0.0136 & 0.0126 & 0.0049 & -0.0003 & -0.0154 \\ -0.0028 & 0.0132 & 0.0837 & 0.0866 & 0.0810 & 0.0196 & 0.0544 & -0.1159 \\ -0.0032 & 0.0136 & 0.0866 & 0.0904 & 0.0868 & 0.0203 & 0.0587 & -0.1227 \\ -0.0039 & 0.0126 & 0.0810 & 0.0868 & 0.0904 & 0.0192 & 0.0620 & -0.1232 \\ -0.0007 & 0.0049 & 0.0196 & 0.0203 & 0.0192 & 0.0054 & 0.0090 & -0.0261 \\ -0.0024 & -0.0003 & 0.0544 & 0.0587 & 0.0620 & 0.0090 & 0.0619 & -0.0900 \\ 0.0048 & -0.0154 & -0.1159 & -0.1227 & -0.1232 & -0.0261 & -0.0900 & 0.1725 \end{pmatrix}.$$

The sampled covariance matrix was computed via the following known formula for an unbiased estimator of the covariance matrix:

$$\boldsymbol{\Sigma} := \frac{1}{T-1}\mathbf{R}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}\mathbf{1}^T\right)\mathbf{R}^T.$$

Here $T = 4$ (number of periods) and $\mathbf{R}$ is the 8×4 matrix containing the assets' returns for each of the 4 years. The rank of the matrix $\boldsymbol{\Sigma}$ is at most 4, thus it is rank deficient. We have chosen the minimal return as $r_0 = 1.05$. In this case the portfolio problem (**??**) has multiple optimal solution, and we therefore consider problem (**??**) as the core problem and introduce a second objective function for the outer problem. Here we choose

$$\omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{a}\|^2.$$

Suppose that we wish to invest as much as possible in gold. Then we can choose $\mathbf{a} = (0, 0, 0, 0, 0, 0, 0, 1)^T$ and in this case the minimal norm gradient method gives the solution

$$(0.0000, 0.0000, 0.0995, 0.1421, 0.2323, 0.0000, 0.1261, 0.3999)^T.$$

If we wish a portfolio which is as dispersed as possible, then we can choose

$$\mathbf{a} = (1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8)^T,$$

and in this case the algorithm produces the following optimal solution:

$$(0.1531, 0.1214, 0.0457, 0.0545, 0.1004, 0.1227, 0.1558, 0.2466)^T,$$

which is very much different from the first optimal solution. Note that in the second optimal solution the investment in gold is much smaller and that the allocation of the resources is indeed much more scattered.

5.2 Solution of Integral Equations

The minimal norm-like solution can be considered as a type of regularized solution of the core problem. In this section we illustrate this stabilization effect of the minimal norm solution on ill-conditioned inverse problems arising from discretizations of Fredholm integral equations of the first kind. The tested problems were taken from the "regularization tools" package; see [**?**] for a complete description. We begin by looking at the famous Phillips problem [**?**] of estimating a function $f(t)$ that solves the integral equation

$$\int_{-6}^{6} k(s-t)f(t) = g(s),$$

where

$$k(t) = \begin{cases} 1 + \cos\left(\frac{\pi t}{3}\right) & |t| < 3, \\ 0 & \text{else} \end{cases}$$

and

$$g(s) = (6 - |s|)\left(1 + \frac{1}{2}\cos\left(\frac{\pi s}{3}\right)\right) + \frac{9}{2\pi}\sin\left(\frac{\pi|s|}{3}\right).$$

Using Galkerin method with orthonormal basis functions, the system is discretized and reduces to a linear system of the form $\mathbf{A}\mathbf{x}_{\mathrm{T}} = \mathbf{b}_{\mathrm{T}}$. The system and its solution are implemented in the function `phillips(n)` from [**?**]. In this example we choose $n = 1000$, so that the number of decision variables is 1000. The observed right-hand side vector is given by

$$\mathbf{b} = \mathbf{b}_{\mathrm{T}} + \sigma \cdot \mathbf{w}, \tag{5.2}$$

where $\sigma = 0.02$ and each component of $\mathbf{w}$ is generated from a standard normal distribution. The core problem we consider is the least squares problem

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2. \tag{5.3}$$

The matrix $\mathbf{A}$ has zero eigenvalues and thus the core problem consists of multiple optimal solutions. The outer objective function is chosen as

$$\omega(\mathbf{x}) = \mathbf{x}^T\mathbf{Q}\mathbf{x},$$

where $\mathbf{Q} = \mathbf{L}^T\mathbf{L} + \mathbf{I}$ and $\mathbf{L}$ approximates the first-derivative operator implemented in the function get `get_l(1000,1)` from [**?**]. In addition to the solution obtained by the minimal norm gradient method, we will consider several solutions of the corresponding Tikhonov problem:

$$\mathbf{x}_\lambda = \operatorname*{argmin}_{\mathbf{x}} \left\{ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \lambda \mathbf{x}^t\mathbf{Q}\mathbf{x} \right\},$$

where λ is chosen by either the L-curve strategy [**?**] or the GCV (short for "generalized cross validation") [**?**] criterion. In addition, we also consider $\lambda = 10^{-3}$ as a "representative" of small values of λ. The result are shown in Figure **??**. Clearly, the result produced by the minimal norm gradient method is of a significantly better quality than the other three alternatives. The choice $\lambda = 10^{-3}$ was proven to be extremely poor while the L-curve also gave a rather poor reconstruction. The choice of regularization parameter dictated

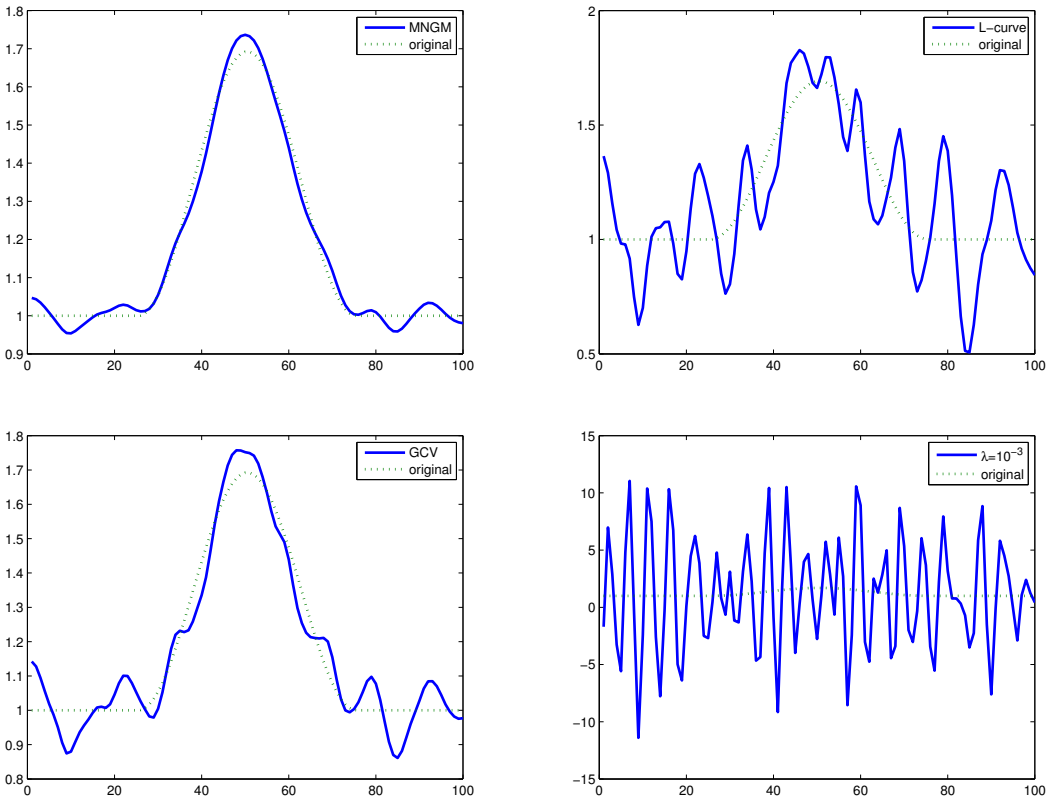

Figure 1: Results of different regularization techniques for the inverse Phillips problem.

by the GCV strategy gave a reasonable reconstruction, but obviously worse than the one generated by the minimal norm gradient method. To illustrate the attractiveness of the minimal norm solution over different choices of Tikhonov solutions for a large amount of runs, we performed Monte-Carlo simulations on three different inverse problems from the "regularization" toolbox: phillips, baart and foxgood. For each of these inverse problems – like in the Phillips example, we generated the corresponding exact linear system $\mathbf{A}\mathbf{x}_{\mathrm{T}} = \mathbf{b}_{\mathrm{T}}$; noise was added to the righthand side as in (**??**) for three different choices of standard deviation: $\sigma = 10^{-1}, 10^{-2}, 10^{-3}$. The core problem problem is the least squares problem (**??**) and the outer objective function is $\omega(\mathbf{x}) = \mathbf{x}^T\mathbf{Q}\mathbf{x}$. Table **??** describes the average of the squared error residual $\|\mathbf{x}_{\mathrm{T}} - \hat{\mathbf{x}}\|^2$ over 100 realizations of $\mathbf{w}$ where $\hat{\mathbf{x}}$ is the solution obtained by one of the four examined methods. The best results in each row are marked by boldface. Clearly, the minimal gradient method is competitive to the L-curve and GCV approaches and is better in the majority of cases.

Problem	σ	Squared Estimation Error			
		MNGM	L-CURVE	GCV	$\lambda = 10^{-3}$
phillips	10^{-3}	**1.06e-2**	8.30	1.3e-2	5.83
phillips	10^{-2}	**1.22e-1**	3.69	1.86e-1	5.83e+2
phillips	10^{-1}	**1.68**	2.52	3.94	5.82e+4
baart	10^{-3}	**3.63e-2**	4.14e-2	2.07e-2	4.26e-2
baart	10^{-2}	3.82e-2	**3.09e-2**	2.14e-2	1.18
baart	10^{-1}	4.30e-2	**3.69e-2**	6.92e-2	1.03e+2
foxgood	10^{-3}	**6.02e-3**	7.5e-2	8.65e-3	7.04e-1
foxgood	10^{-2}	**8.61e-2**	9.42e-2	1.77e-1	9.01
foxgood	10^{-1}	5.71e-1	5.19e-1	**3.88e-1**	9.23e+2

Table 1: Comparison between the different regularization techniques on 100 realizations of the error vector.