On the Convergence of Alternating Minimization for Convex Programming with Applications to Iteratively Reweighted Least Squares and Decomposition Schemes

Amir Beck*

August 7, 2014

Abstract

This paper is concerned with the alternating minimization (AM) for solving convex minimization problems where the decision variables vector is split into two blocks. The objective function is a sum of a differentiable convex function and a separable (possibly) nonsmooth extended real-valued convex function, and consequently constraints can be incorporated. We analyze the convergence rate of the method and establish a nonasymptotic sublinear rate of convergence where the multiplicative constant depends on the minimal block Lipschitz constant. We then analyze the iteratively reweighted least squares (IRLS) method for solving convex problems involving sums of norms. Based on the results derived for the AM method, we establish a nonasymptotic sublinear rate of convergence of the IRLS method. In addition, we show an asymptotic rate of convergence whose efficiency estimate does not depend on the data of the problem. Finally, we study the convergence properties of a decomposition-based approach designed to solve a composite convex model.

1 Introduction and Problem/Model Formulation

In this paper we consider the following minimization problem:

$$\min_{\mathbf{y}\in\mathbb{R}^{n_1},\mathbf{z}\in\mathbb{R}^{n_2}} \{H(\mathbf{y},\mathbf{z}) \equiv f(\mathbf{y},\mathbf{z}) + g_1(\mathbf{y}) + g_2(\mathbf{z})\}, \tag{1.1}$$

where f, g_1, g_2 are assumed to satisfy the following two properties:

[A] The functions $g_1 : \mathbb{R}^{n_1} \to (-\infty,\infty]$ and $g_2 : \mathbb{R}^{n_2} \to (-\infty,\infty]$ are closed and proper convex functions assumed to be subdifferentiable over their domain.

[B] The function f is a continuously differentiable convex function over $\operatorname{dom} g_1 \times \operatorname{dom} g_2$.

*Department of Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa 32000, Israel. E-mail: becka@ie.technion.ac.il.

We will use the convention that the variables vector $\mathbf{x} \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ is composed of the vectors $\mathbf{y}$ and $\mathbf{z}$ as follows:

$$\mathbf{x} = (\mathbf{y}, \mathbf{z}).$$

The sum of the two functions g_1 and g_2 is denoted by the function $g : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to (-\infty, \infty]$:

$$g(\mathbf{x}) = g(\mathbf{y}, \mathbf{z}) = g_1(\mathbf{y}) + g_2(\mathbf{z}).$$

In this notation the objective function can be written as $H(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$. Since g_1 and g_2 are extended real-valued, the above formulation also encompasses the case of convex constraints. We will denote the vector of all partial derivatives corresponding to the variables vector $\mathbf{y}$ by $\nabla_1 f(\mathbf{x})$, and the vector of all partial derivatives corresponding to $\mathbf{z}$ by $\nabla_2 f(\mathbf{x})$, so that in particular $\nabla f(\mathbf{x}) = (\nabla_1 f(\mathbf{x}), \nabla_2 f(\mathbf{x})) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$. With this notation, we will also assume that the following property holds:

[C] The gradient of f is (uniformly) Lipschitz continuous with respect to the variables vector $\mathbf{y}$ over $\operatorname{dom} g_1$ with constant $L_1 \in (0, \infty)$:

$$\|\nabla_1 f(\mathbf{y} + \mathbf{d}_1, \mathbf{z}) - \nabla_1 f(\mathbf{y}, \mathbf{z})\| \leq L_1 \|\mathbf{d}_1\|$$

for any $\mathbf{y} \in \operatorname{dom} g_1, \mathbf{z} \in \operatorname{dom} g_2$ and $\mathbf{d}_1 \in \mathbb{R}^{n_1}$ such that $\mathbf{y} + \mathbf{d}_1 \in \operatorname{dom} g_1$.

In some cases, we will also assume that the gradient of f is Lipschitz continuous with respect to the variables vector $\mathbf{z}$. For that, we will also have the following assumption:

[D] The gradient of f is Lipschitz continuous with respect to the variables vector $\mathbf{z}$ over $\operatorname{dom} g_2$ with constant $L_2 \in (0, \infty]$:

$$\|\nabla_2 f(\mathbf{y}, \mathbf{z} + \mathbf{d}_2) - \nabla_2 f(\mathbf{y}, \mathbf{z})\| \leq L_2 \|\mathbf{d}_2\|$$

for all $\mathbf{y} \in \operatorname{dom} g_1, \mathbf{z} \in \operatorname{dom} g_2$ and $\mathbf{d}_2 \in \mathbb{R}^{n_2}$ such that $\mathbf{z} + \mathbf{d}_2 \in \operatorname{dom} g_2$.

We will always assume that properties [A]-[D] are satisfied, but note that we also allow L_2 to be ∞, which means that the gradient of the objective function might not be Lipschitz with respect to the second block of variables, unless the additional assumption $L_2 < \infty$ is explicitly made.

We will be interested in analyzing the *alternating minimization* (AM) method that is described explicitly below.

The Alternating Minimization Method

Initialization: $\mathbf{y}_0 \in \operatorname{dom} g_1, \mathbf{z}_0 \in \operatorname{dom} g_2$ such that $\mathbf{z}_0 \in \operatorname*{argmin}_{\mathbf{z} \in \mathbb{R}^{n_2}} f(\mathbf{y}_0, \mathbf{z}) + g_2(\mathbf{z})$.

General Step (k=0,1,...):

$$\begin{aligned} \mathbf{y}_{k+1} &\in \operatorname*{argmin}_{\mathbf{y} \in \mathbb{R}^{n_1}} f(\mathbf{y}, \mathbf{z}_k) + g_1(\mathbf{y}), \\ \mathbf{z}_{k+1} &\in \operatorname*{argmin}_{\mathbf{z} \in \mathbb{R}^{n_2}} f(\mathbf{y}_{k+1}, \mathbf{z}) + g_2(\mathbf{z}). \end{aligned}$$

Note that we assume that "half" an iteration was performed prior to the first iteration (that is, $\mathbf{z}_0 \in \operatorname{argmin}_{\mathbf{z}\in\mathbb{R}^{n_2}} f(\mathbf{y}_0, \mathbf{z}) + g_2(\mathbf{z})$). We could have defined the initial vector as $(\mathbf{y}_{-1}, \mathbf{z}_{-1})$ without the need to assume anything about the initial vector, but for sake of simplicity of notation, we keep this setting.

To make the method well-defined we will also make the following assumption throughout the paper.

[E] The optimal set of (1.1), denoted by X^*, is nonempty, and the corresponding optimal value is denoted by H^*. In addition, for any $\tilde{\mathbf{y}} \in \operatorname{dom} g_1$ and $\tilde{\mathbf{z}} \in \operatorname{dom} g_2$, the problems

$$\min_{\mathbf{z}\in\mathbb{R}^{n_2}} f(\tilde{\mathbf{y}}, \mathbf{z}) + g_2(\mathbf{z}), \min_{\mathbf{y}\in\mathbb{R}^{n_1}} f(\mathbf{y}, \tilde{\mathbf{z}}) + g_1(\mathbf{y})$$

have minimizers.

The k-th iterate will be denoted by $\mathbf{x}_k = (\mathbf{y}_k, \mathbf{z}_k)$, and we will also consider the "sequence in between" given by

$$\mathbf{x}_{k+\frac{1}{2}} = (\mathbf{y}_{k+1}, \mathbf{z}_k).$$

Obviously, the generated sequence is monotone and satisfies:

$$H(\mathbf{x}_0) \geq H(\mathbf{x}_{\frac{1}{2}}) \geq H(\mathbf{x}_1) \geq H(\mathbf{x}_{\frac{3}{2}}) \ldots \geq$$

The ability to employ the alternating minimization method relies on the capability of computing minimizers with respect to each of the blocks. In Sections 4 and 5 we will consider two classes of problems in which these partial minimizations can indeed be computed in a relatively simple manner.

The alternating minimization method in its general sense, that is, when the decision variables vector is decomposed into p sub-blocks (p being an integer greater than one) is a rather old and fundamental algorithm [6, 25]. It appears in the literature under various names such as the block-nonlinear Gauss-Seidel method or the block coordinate descent method (see e.g., [5]). Several results on the convergence of the method were established in the literature. Auslender studied in [1] the convergence of the method under a strong convexity assumption, but without assuming differentiability. In [5] Bertsekas showed that if the minimum with respect to each block of variables is unique, then any accumulation point of the sequence generated by the method is also a stationary point. Grippo and Sciandrone showed in [14] convergence results of the sequence generated by the method under different sets of assumptions such as strict quasiconvexity with respect to each block. Luo and Tseng proved in [19] that under the assumptions of strong convexity with respect to each block, existence of a local error bound of the objective function and proper separation of isocost surfaces, linear rate of convergence can be established. The only result available on the rate of convergence of the method under general convexity assumptions (and not strong convexity) is the result in [4] showing a sublinear rate of convergence, and that the multiplicative constant depends on the minimum of the block Lipschitz constants. However, the result in [4] is limited in the sense that it only holds for unconstrained problems with a smooth objective function. In Section 3 we show that a sublinear $O(1/k)$ rate of convergence can be obtained for the

alternating minimization method employed on the general problem (1.1).

In Section 4 we consider the iteratively reweighted least squares (IRLS) algorithm designed to solve problems involving sums of norms. It is well known that the method is essentially an alternating minimization method employed on an auxiliary problem. This scheme was used in the context of many applications such as robust regression [20], sparse recovery [10] and localization [29, 30]. There are only few works that discuss the rate of convergence of the IRLS method. Among them is the work [7] in which an asymptotic linear rate of convergence is established for a certain class of unconstrained convex problems. Asymptotic linear rate of convergence was also shown in [10] for the so-called "basis pursuit" problem that consists of minimizing the l_1 norm function over a set of linear equations. The underlying assumption required in the latter paper is that the matrix defining the constraints satisfies the restricted isometry property. In Section 4 we establish, based on the results from Section 3, a nonasymptotic sublinear rate of convergence of the IRLS problem under a very general setting. In addition we show that the asymptotic efficiency estimate depends only on the smoothing parameter and not on the data of the problem. Finally, in Section 5 we analyze a solution scheme based on the alternating minimization method for solving an approximation of a composite model, and derive a complexity result in terms of the original problem.

2 Mathematical Preliminaries

In this section we layout some of the mathematical background essential for our analysis. In particular, we recall the notions of the proximal operator and the gradient mapping and define their partial counterparts.

2.1 The proximal mapping

For a given closed, proper extended real-valued convex function $h : \mathbb{R}^n \to (-\infty, \infty]$, the proximal operator is defined by

$$\operatorname{prox}_h(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{u}} \left\{ h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2 \right\},$$

where $\|\cdot\|$ is the l_2 norm defined on $\mathbb{R}^n$. The operator and many of its properties were introduced and studied by Moreau in his seminar work [22]. An important and useful characterization of the prox operation is given in the following result (see for example [2])

Lemma 2.1. *Let $h : \mathbb{R}^n \to (-\infty, \infty]$ be a proper, closed and convex function, and let $M > 0$. Then*

$$\mathbf{w} = \operatorname{prox}_{\frac{1}{M}h}(\mathbf{x})$$

if and only if for any $\mathbf{u} \in \operatorname{dom} h$:

$$h(\mathbf{u}) \geq h(\mathbf{w}) + M\langle \mathbf{x} - \mathbf{w}, \mathbf{u} - \mathbf{w} \rangle. \tag{2.1}$$

2.2 The gradient mapping

Another important concept is that of the *gradient mapping* [23]. Given $M > 0$, the *prox-grad* mapping associated with problem (1.1) is defined as

$$T_M(\mathbf{x}) = \text{prox}_{\frac{1}{M}g}\left(\mathbf{x} - \frac{1}{M}\nabla f(\mathbf{x})\right).$$

We also define the prox-grad operators with respect to $\mathbf{y}$ and $\mathbf{z}$ as

$$\begin{aligned} T_M^1(\mathbf{x}) &= \text{prox}_{\frac{1}{M}g_1}\left(\mathbf{y} - \frac{1}{M}\nabla_1 f(\mathbf{y},\mathbf{z})\right), \\ T_M^2(\mathbf{x}) &= \text{prox}_{\frac{1}{M}g_2}\left(\mathbf{z} - \frac{1}{M}\nabla_2 f(\mathbf{y},\mathbf{z})\right), \end{aligned}$$

repectively. Obviously we have

$$T_M(\mathbf{x}) = (T_M^1(\mathbf{x}), T_M^2(\mathbf{x}))$$

for any $\mathbf{x} \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$. The gradient mapping, as defined in [23], is given by (for any $M > 0$):

$$G_M(\mathbf{x}) = M\left(\mathbf{x} - T_M(\mathbf{x})\right) = M\left(\mathbf{x} - \text{prox}_{\frac{1}{M}g}\left[\mathbf{x} - \frac{1}{M}\nabla_{\mathbf{x}} f(\mathbf{x})\right]\right).$$

The partial gradient mappings are correspondingly defined as

$$\begin{aligned} G_M^1(\mathbf{x}) &= M\left(\mathbf{y} - \text{prox}_{\frac{1}{M}g_1}\left[\mathbf{y} - \frac{1}{M}\nabla_1 f(\mathbf{y},\mathbf{z})\right]\right), \\ G_M^2(\mathbf{x}) &= M\left(\mathbf{z} - \text{prox}_{\frac{1}{M}g_2}\left[\mathbf{z} - \frac{1}{M}\nabla_2 f(\mathbf{y},\mathbf{z})\right]\right), \end{aligned}$$

and we have

$$G_M(\mathbf{x}) = (G_M^1(\mathbf{x}), G_M^2(\mathbf{x})).$$

The gradient mapping is an optimality measure in the sense that $G_M(\mathbf{x}) = \mathbf{0}$ for some $M > 0$ if and only if $\mathbf{x}$ is an optimal solution of (1.1). Note that if $G_M(\mathbf{x}) = \mathbf{0}$ for *some* $M > 0$, then $G_M(\mathbf{x}) = \mathbf{0}$ for *all* $M > 0$. In addition, $\mathbf{x}$ is an optimal solution of (1.1) if and only if $G_{M_1}(\mathbf{x}) = \mathbf{0}, G_{M_2}(\mathbf{x}) = \mathbf{0}$ for some $M_1, M_2 > 0$. If we remove the assumption of convexity of f, then the equation $G_M(\mathbf{x}) = \mathbf{0}$ holds if and only if $\mathbf{x}$ is a stationary point of problem (1.1).

2.3 The block descent lemma

Under the block-Lipschitz assumption described in properties [C] and [D], we can write the following block-version of the so-called descent lemma [5].

Lemma 2.2 (block descent lemma). *(i) Let $\mathbf{y} \in \text{dom}\, g_1$ and $\mathbf{z} \in \text{dom}\, g_2$, and let $\mathbf{h} \in \mathbb{R}^{n_1}$ be such that $\mathbf{y} + \mathbf{h} \in \text{dom}\, g_1$. Then for any $M \geq L_1$, it holds that*

$$f(\mathbf{y}+\mathbf{h},\mathbf{z}) \leq f(\mathbf{y},\mathbf{z}) + \langle \nabla_1 f(\mathbf{y},\mathbf{z}), \mathbf{h}\rangle + \frac{M}{2}\|\mathbf{h}\|^2. \tag{2.2}$$

(ii) Let $\mathbf{y} \in \operatorname{dom} g_1$ *and* $\mathbf{z} \in \operatorname{dom} g_2$*, and let* $\mathbf{h} \in \mathbb{R}^{n_2}$ *be such that* $\mathbf{z}+\mathbf{h} \in \operatorname{dom} g_2$*. Then assuming that* $L_2 < \infty$ *and* $M \geq L_2$*, it holds that*

$$f(\mathbf{y}, \mathbf{z}+\mathbf{h}) \leq f(\mathbf{y}, \mathbf{z}) + \langle \nabla_2 f(\mathbf{y}, \mathbf{z}), \mathbf{h} \rangle + \frac{M}{2}\|\mathbf{h}\|^2. \tag{2.3}$$

2.4 The sufficient decrease property

Suppose that $s : \mathbb{R}^p \to \mathbb{R}$ is a continuously differentiable function with Lipschitz gradient with constant $L > 0$, and that $h : \mathbb{R}^p \to [-\infty, \infty)$ is a closed, proper and convex function. Then it is well known (see e.g., [2, Lemma 2.6]) that the following sufficient decrease property holds for the function $S(\mathbf{x}) \equiv s(\mathbf{x}) + h(\mathbf{x})$:

$$S(\mathbf{x}) - S\left(\operatorname{prox}_h\left(\mathbf{x} - \frac{1}{L}\nabla s(\mathbf{x})\right)\right) \geq \frac{1}{2L}\left\|L\left(\mathbf{x} - \operatorname{prox}_h\left(\mathbf{x} - \frac{1}{L}\nabla s(\mathbf{x})\right)\right)\right\|^2$$

for any $\mathbf{x} \in \operatorname{dom} h$. Noting that for any $\mathbf{z} \in \operatorname{dom} g_2$ the function $f(\cdot, \mathbf{z})$ is continuously differentiable with Lipschitz gradient with constant L_1, and recalling the definition of the partial gradient mapping, we can conclude that

$$H(\mathbf{y}, \mathbf{z}) - H(T^1_{L_1}(\mathbf{x}), \mathbf{z}) \geq \frac{1}{2L_1}\|G^1_{L_1}(\mathbf{y}, \mathbf{z})\|^2 \tag{2.4}$$

for any $\mathbf{y} \in \operatorname{dom} g_1$. Similarly, if $L_2 < \infty$, then for any $\mathbf{y} \in \operatorname{dom} g_1$ the function $f(\mathbf{y}, \cdot)$ is continuously differentiable with Lipschitz gradient with constant L_2, and hence

$$H(\mathbf{y}, \mathbf{z}) - H(\mathbf{y}, T^1_{L_2}(\mathbf{x})) \geq \frac{1}{2L_2}\|G^2_{L_2}(\mathbf{y}, \mathbf{z})\|^2 \tag{2.5}$$

for any $\mathbf{z} \in \operatorname{dom} g_2$.

3 Convergence Analysis

3.1 The nonconvex case

In this subsection (and only in this subsection) we remove the convexity assumption of f and show that the limit points of the sequence generated by the alternating minimization method are stationary points of the problem. We begin with the following direct consequence of the sufficient decrease property given in (2.4) and (2.5).

Lemma 3.1. *Let* $\{\mathbf{x}_k\}_{k\geq 0}$ *be the sequence generated by the alternating minimization method. Then for any* $k \geq 0$ *the following inequality holds:*

$$H(\mathbf{x}_k) - H(\mathbf{x}_{k+\frac{1}{2}}) \geq \frac{1}{2L_1}\|G^1_{L_1}(\mathbf{x}_k)\|^2. \tag{3.1}$$

If in addition $L_2 < \infty$*, then*

$$H(\mathbf{x}_{k+\frac{1}{2}}) - H(\mathbf{x}_{k+1}) \geq \frac{1}{2L_2}\|G^2_{L_2}(\mathbf{x}_{k+\frac{1}{2}})\|^2. \tag{3.2}$$

Proof. Plugging $\mathbf{y} = \mathbf{y}_k$ and $\mathbf{z} = \mathbf{z}_k$ into (2.4) we have

$$H(\mathbf{y}_k, \mathbf{z}_k) - H(T^1_{L_1}(\mathbf{x}_k), \mathbf{z}_k) \geq \frac{1}{2L_1}\|G^1_{L_1}(\mathbf{y}_k, \mathbf{z}_k)\|^2.$$

The inequality (3.1) now follows from the inequality $H(\mathbf{x}_{k+\frac{1}{2}}) \leq H(T^1_{L_1}(\mathbf{x}_k), \mathbf{z}_k)$ and the fact that $\mathbf{x}_k = (\mathbf{y}_k, \mathbf{z}_k)$. The inequality (3.2) follows by plugging $\mathbf{y} = \mathbf{y}_{k+1}, \mathbf{z} = \mathbf{z}_k$ into (2.5) and using the inequality $H(\mathbf{x}_{k+1}) \leq H(\mathbf{y}_{k+1}, T^2_{L_2}(\mathbf{x}_{k+\frac{1}{2}}))$. □

We are now ready to prove the main convergence result for the nonconvex case: a rate of convergence of the norms of the partial gradient mappings to zero.

Theorem 3.1 (rate of convergence of partial gradient mappings)**.** *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the alternating minimization method. Then for any $n \geq 1$*

$$\min_{k=0,1,\ldots n} \|G^1_{L_1}(\mathbf{x}_k)\| \leq \frac{\sqrt{2L_1(H(\mathbf{x}_0) - H^*)}}{\sqrt{n+1}}. \tag{3.3}$$

If in addition $L_2 < \infty$ then

$$\min_{k=0,1,\ldots n} \|G^2_{L_2}(\mathbf{x}_{k+\frac{1}{2}})\| \leq \frac{\sqrt{2L_2(H(\mathbf{x}_0) - H^*)}}{\sqrt{n+1}}.$$

Proof. By Lemma 3.1 we have for any $k \geq 0$:

$$H(\mathbf{x}_k) - H(\mathbf{x}_{k+1}) \geq H(\mathbf{x}_k) - H(\mathbf{x}_{k+\frac{1}{2}}) \geq \frac{1}{2L_1}\|G^1_{L_1}(\mathbf{x}_k)\|^2.$$

Summing the above inequality for $k = 0, 1, \ldots, n$ we obtain

$$H(\mathbf{x}_0) - H(\mathbf{x}_{n+1}) \geq \frac{1}{2L_1}\sum_{k=0}^{n} \|G^1_{L_1}(\mathbf{x}_k)\|^2.$$

Hence, since

$$\sum_{k=0}^{n} \|G^1_{L_1}(\mathbf{x}_k)\|^2 \geq (n+1) \min_{k=0,1,\ldots n} \|G^1_{L_1}(\mathbf{x}_k)\|^2$$

and $H(\mathbf{x}_{n+1}) \geq H^*$, it follows that

$$\min_{k=0,1,\ldots n} \|G^1_{L_1}(\mathbf{x}_k)\| \leq \frac{\sqrt{2L_1(H(\mathbf{x}_0) - H^*)}}{\sqrt{n+1}}.$$

If in addition $L_2 < \infty$, then by (3.2) we have that for any $k \geq 0$

$$H(\mathbf{x}_k) - H(\mathbf{x}_{k+1}) \geq H(\mathbf{x}_{k+\frac{1}{2}}) - H(\mathbf{x}_{k+1}) \geq \frac{1}{2L_2}\|G^2_{L_2}(\mathbf{x}_{k+\frac{1}{2}})\|^2.$$

Summing the above inequality over $k = 0, 1, \ldots, n$ and using the same type of arguments as those invoked in the first part, we obtain that

$$\min_{k=0,1,\ldots n} \|G^2_{L_2}(\mathbf{x}_{k+\frac{1}{2}})\| \leq \frac{\sqrt{2L_2(H(\mathbf{x}_0) - H^*)}}{\sqrt{n+1}}.$$

□

Lemma 3.2. *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the alternating minimization method. Then any accumulation point of $\{\mathbf{x}_k\}$ is a stationary point of problem (1.1).*

Proof. Suppose that $\tilde{\mathbf{x}}$ is an accumulation point of the sequence. Then there exists a subsequence $\{\mathbf{x}_{k_n}\}_{n\geq 0}$ that converges to $\tilde{\mathbf{x}}$. By the definition of the sequence we have $G^2_{L_2}(\mathbf{x}_{k_n}) = \mathbf{0}$, and hence by the continuity of the gradient mapping, $G^2_{L_2}(\tilde{\mathbf{x}}) = \mathbf{0}$. Also, since $\{H(\mathbf{x}_{\frac{k}{2}})\}_{k\geq 0}$ is a bounded below nonincreasing sequence, it converges to some finite value and hence $H(\mathbf{x}_k) - H(\mathbf{x}_{k+\frac{1}{2}}) \to 0$, and we conclude by (3.1) that $G^1_{L_1}(\mathbf{x}_{k_n}) \to \mathbf{0}$ as n tends to ∞, which implies by the continuity of $G^1_{L_1}$, that $G^1_{L_1}(\tilde{\mathbf{x}}) = \mathbf{0}$. Thus, since $G^1_{L_1}(\tilde{\mathbf{x}}) = \mathbf{0}$ and $G^2_{L_2}(\tilde{\mathbf{x}}) = \mathbf{0}$, we obtain that $\tilde{\mathbf{x}}$ is a stationary point of problem (1.1). □

3.2 The convex case

We now bring back the convexity assumption on f. Our main objective will be to prove a rate of convergence result for the sequence of function values. We begin with the following technical lemma.

Lemma 3.3. *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the alternating minimization method. Then for any $k \geq 0$:*

$$H(\mathbf{x}_{k+\frac{1}{2}}) - H(\mathbf{x}^*) \leq \|G^1_{L_1}(\mathbf{x}_k)\| \cdot \|\mathbf{x}_k - \mathbf{x}^*\|.$$

If in addition $L_2 < \infty$, then the following inequality also holds for any $k \geq 0$:

$$H(\mathbf{x}_{k+1}) - H(\mathbf{x}^*) \leq \|G^2_{L_2}(\mathbf{x}_{k+\frac{1}{2}})\| \cdot \|\mathbf{x}_{k+\frac{1}{2}} - \mathbf{x}^*\|. \tag{3.4}$$

Proof. Note that

$$T_{L_1}(\mathbf{x}_k) = (T^1_{L_1}(\mathbf{x}_k), T^2_{L_1}(\mathbf{x}_k)) = \left(T^1_{L_1}(\mathbf{x}_k), \mathbf{z}_k - \frac{1}{L_1} G^2_{L_1}(\mathbf{x}_k)\right) = (T^1_{L_1}(\mathbf{x}_k), \mathbf{z}_k),$$

where in the last equality we used the fact that by the definition of the alternating minimization method $G^2_M(\mathbf{x}_k) = \mathbf{0}$ for any $M > 0$ and $k = 0, 1, 2, \ldots$. Combining this with the block descent lemma (Lemma 2.2), we obtain that

$$\begin{aligned} f(T_{L_1}(\mathbf{x}_k)) - f(\mathbf{x}^*) &\leq f(\mathbf{x}_k) + \langle \nabla_1 f(\mathbf{x}_k), T^1_{L_1}(\mathbf{x}_k) - \mathbf{y}_k \rangle + \frac{L_1}{2}\|T^1_{L_1}(\mathbf{x}_k) - \mathbf{y}_k\|^2 - f(\mathbf{x}^*) \\ &= f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), T_{L_1}(\mathbf{x}_k) - \mathbf{x}_k \rangle + \frac{L_1}{2}\|T^1_{L_1}(\mathbf{x}_k) - \mathbf{y}_k\|^2 - f(\mathbf{x}^*) \end{aligned} \tag{3.5}$$

Since f is convex, it follows that $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle$, which combined with (3.5) yields

$$f(T_{L_1}(\mathbf{x}_k)) - f(\mathbf{x}^*) \leq \langle \nabla f(\mathbf{x}_k), T_{L_1}(\mathbf{x}_k) - \mathbf{x}^* \rangle + \frac{L_1}{2}\|T^1_{L_1}(\mathbf{x}_k) - \mathbf{y}_k\|^2. \tag{3.6}$$

Since

$$T_{L_1}(\mathbf{x}_k) = \operatorname{prox}_{\frac{1}{L_1} g}\left(\mathbf{x}_k - \frac{1}{L_1}\nabla f(\mathbf{x}_k)\right),$$

then by invoking Lemma 2.1 with $h = g, M = L_1, \mathbf{x} = \mathbf{x}_k - \frac{1}{L_1}\nabla f(\mathbf{x}_k), \mathbf{w} = T_{L_1}(\mathbf{x}_k)$ and $\mathbf{u} = \mathbf{x}^*$, we have

$$g(\mathbf{x}^*) \geq g(T_{L_1}(\mathbf{x}_k)) + L_1 \left\langle \mathbf{x}_k - \frac{1}{L_1}\nabla f(\mathbf{x}_k) - T_{L_1}(\mathbf{x}_k), \mathbf{x}^* - T_{L_1}(\mathbf{x}_k) \right\rangle. \tag{3.7}$$

Combining inequalities (3.6) and (3.7), along with the fact that $H(\mathbf{x}_{k+\frac{1}{2}}) \leq H(T_{L_1}(\mathbf{x}_k))$, we finally have

$$\begin{aligned}
H(\mathbf{x}_{k+\frac{1}{2}}) - H(\mathbf{x}^*) &\leq H(T_{L_1}(\mathbf{x}_k)) - H(\mathbf{x}^*) \\
&= f(T_{L_1}(\mathbf{x}_k)) + g(T_{L_1}(\mathbf{x}_k)) - f(\mathbf{x}^*) - g(\mathbf{x}^*) \\
&\leq L_1\langle \mathbf{x}_k - T_{L_1}(\mathbf{x}_k), T_{L_1}(\mathbf{x}_k) - \mathbf{x}^*\rangle + \frac{L_1}{2}\|T^1_{L_1}(\mathbf{x}_k) - \mathbf{y}_k\|^2 \\
&= \langle G_{L_1}(\mathbf{x}_k), T_{L_1}(\mathbf{x}_k) - \mathbf{x}^*\rangle + \frac{1}{2L_1}\|G_{L_1}(\mathbf{x}_k)\|^2 \\
&= \langle G_{L_1}(\mathbf{x}_k), T_{L_1}(\mathbf{x}_k) - \mathbf{x}_k\rangle + \langle G_{L_1}(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^*\rangle + \frac{1}{2L_1}\|G_{L_1}(\mathbf{x}_k)\|^2 \\
&= -\frac{1}{L_1}\|G_{L_1}(\mathbf{x}_k)\|^2 + \langle G_{L_1}(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^*\rangle + \frac{1}{2L_1}\|G_{L_1}(\mathbf{x}_k)\|^2 \\
&\leq \langle G_{L_1}(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^*\rangle \\
&\leq \|G_{L_1}(\mathbf{x}_k)\| \cdot \|\mathbf{x}_k - \mathbf{x}^*\| \\
&= \|G^1_{L_1}(\mathbf{x}_k)\| \cdot \|\mathbf{x}_k - \mathbf{x}^*\|.
\end{aligned}$$

When $L_2 < \infty$, the same argument on the sequence generated by the alternating minimization method defined on the function

$$\tilde{f}(\mathbf{z}, \mathbf{y}) \equiv f(\mathbf{y}, \mathbf{z}),$$

with starting point $(\mathbf{z}^0, \mathbf{y}^1)$ gives the required result. □

Remark 3.1. The inequality (3.4) is proven only under the condition that $L_2 < \infty$. However, if the function $\mathbf{z} \mapsto f(\mathbf{y}_{k+1}, \mathbf{z})$ has a Lipschitz gradient with constant $M(\mathbf{y}_{k+1})$, then the exact same argument shows that

$$H(\mathbf{x}_{k+1}) - H(\mathbf{x}^*) \leq \|G_{M(\mathbf{y}_{k+1})}(\mathbf{x}_{k+\frac{1}{2}})\| \cdot \|\mathbf{x}_{k+\frac{1}{2}} - \mathbf{x}^*\|.$$

This result also holds in the case $L_2 = \infty$.

We will assume that the level set

$$S = \{\mathbf{x} \in \operatorname{dom} g_1 \times \operatorname{dom} g_2 : H(\mathbf{x}) \leq H(\mathbf{x}_0)\}$$

is compact and we denote by R the following "diameter":

$$R = \max_{\mathbf{x} \in \mathbb{R}^{n_1 \times n_2}} \max_{\mathbf{x}^* \in X^*} \{\|\mathbf{x} - \mathbf{x}^*\| : H(\mathbf{x}) \leq H(\mathbf{x}_0)\}. \tag{3.8}$$

In particular, by the monotonicity of $\{H(\mathbf{x}_{\frac{k}{2}})\}_{k \geq 0}$:

$$\|\mathbf{x}_k - \mathbf{x}^*\|, \|\mathbf{x}_{k+\frac{1}{2}} - \mathbf{x}^*\| \leq R \text{ for every } k = 0, 1, \ldots \tag{3.9}$$

In this terminology we can write the following recurrence inequality relation of the function values of the generated sequence.

Lemma 3.4. *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the alternating minimization method. Then*

$$H(\mathbf{x}_k) - H(\mathbf{x}_{k+1}) \geq \frac{1}{2\min\{L_1, L_2\}R^2}(H(\mathbf{x}_{k+1}) - H^*)^2 \tag{3.10}$$

for all $k \geq 0$.

Proof. By Lemma 3.3 and (3.9) we have

$$H(\mathbf{x}_{k+\frac{1}{2}}) - H^* \leq \|G^1_{L_1}(\mathbf{x}_k)\|R.$$

Now, by Lemma 3.1,

$$\begin{aligned} H(\mathbf{x}_k) - H(\mathbf{x}_{k+1}) &\geq H(\mathbf{x}_k) - H(\mathbf{x}_{k+\frac{1}{2}}) \\ &\geq \frac{1}{2L_1}\|G^1_{L_1}(\mathbf{x}_k)\|^2 \\ &\geq \frac{(H(\mathbf{x}_{k+\frac{1}{2}}) - H^*)^2}{2L_1R^2} \\ &\geq \frac{1}{2L_1R^2}(H(\mathbf{x}_{k+1}) - H^*)^2 \end{aligned} \tag{3.11}$$

If $L_2 = \infty$, then obviously (3.10) holds. If $L_2 < \infty$, then by Lemma 3.3 and (3.9) we have

$$H(\mathbf{x}_{k+1}) - H^* \leq \|G^2_{L_2}(\mathbf{x}_{k+\frac{1}{2}})\|R.$$

Now,

$$H(\mathbf{x}_k) - H(\mathbf{x}_{k+1}) \geq H(\mathbf{x}_{k+\frac{1}{2}}) - H(\mathbf{x}_{k+1}) \geq \frac{1}{2L_2}\|G^2_{L_2}(\mathbf{x}_{k+\frac{1}{2}})\|^2 \geq \frac{(H(\mathbf{x}_{k+1}) - H^*)^2}{2L_2R^2},$$

which combined with (3.11) yields the desired result. □

Theorem 3.2 below establishes the sublinear rate of convergence of the sequence of function values generated by the alternating minimization method. We will also require the following simple lemma on sequences of nonnegative numbers that was proven in [4, Lemma 6.2].

Lemma 3.5. *Let $\{A_k\}_{k\geq 0}$ be a nonnegative sequence of real numbers satisfying*

$$A_k - A_{k+1} \geq \gamma A^2_{k+1}, \quad k = 0, 1, \ldots \tag{3.12}$$

and

$$A_1 \leq \frac{1.5}{\gamma},\ A_2 \leq \frac{1.5}{2\gamma}$$

for some positive γ. Then

$$A_k \leq \frac{1.5}{\gamma}\frac{1}{k}, \quad k = 1, 2, \ldots \tag{3.13}$$

Theorem 3.2. *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the alternating minimization method. Then for any $k \geq 1$*

$$H(\mathbf{x}_k) - H^* \leq \frac{3\max\{H(\mathbf{x}_0) - H^*, \min\{L_1, L_2\}R^2\}}{k}. \tag{3.14}$$

Proof. Denoting $A_k = H(\mathbf{x}_k) - H^*$ and $\tilde{\gamma} = \frac{1}{2\min\{L_1,L_2\}R^2}$, we obtain by (3.10) that the following inequality holds:

$$A_k - A_{k+1} \geq \tilde{\gamma} A_{k+1}^2.$$

Obviously,

$$\begin{aligned} A_1 &= H(\mathbf{x}_1) - H^* \leq H(\mathbf{x}_0) - H^*, \\ A_2 &\leq H(\mathbf{x}_0) - H^*, \end{aligned}$$

and therefore in particular

$$A_1 \leq \frac{1.5}{\gamma}, A_2 \leq \frac{1.5}{2\gamma},$$

where we take

$$\gamma = \frac{1}{2}\min\left\{\frac{1}{H(\mathbf{x}_0) - H^*}, \frac{1}{\min\{L_1, L_2\}R^2}\right\}.$$

Since $\gamma \leq \tilde{\gamma}$, it follows that

$$A_k - A_{k+1} \geq \gamma A_{k+1}^2$$

for all $k \geq 1$, and hence by Lemma 3.5 we conclude that for any $k \geq 1$

$$A_k \leq \frac{1.5}{\gamma} \cdot \frac{1}{k},$$

establishing the desired result. □

Remark 3.2. The constant in the efficiency estimate (3.14) depends on $\min\{L_1, L_2\}$, and not on the maximum of the block Lipschitz constants, or on the global Lipschitz constant. This means that the convergence of the alternating minimization method is dictated by smoother block of the function, that is, the smallest Lipschitz constant, which is a rather optimistic result. This corresponds to the result obtained in the smooth and unconstrained case in [4] (that is, $g_1 = 0, g_2 = 0$) where it was shown that

$$H(\mathbf{x}_k) - H^* \leq \frac{2\min\{L_1, L_2\}R^2(\mathbf{x}_0)}{k-1}.$$

Note that the constant in the efficiency estimate (3.14) also depends on $H(\mathbf{x}_0) - H^*$ which potentially can mean that there is an implicit dependence on some global Lipschitz constant, which is obviously a potential drawback. However, we will show that in fact the dependency on $H(\mathbf{x}_0) - H^*$ is rather mild and does not have a significant effect on the number of iterations required to obtain a predescribed accuracy. For that, we will require a finer analysis of sequences satisfying the inequality (3.12). This is done in the following lemma.

Lemma 3.6. *Let $\{A_k\}_{k\geq 0}$ be a nonnegative sequence of real numbers satisfying*

$$A_k - A_{k+1} \geq \gamma A_{k+1}^2, \quad k = 0, 1, \ldots$$

Then for any $n \geq 2$

$$A_n \leq \max\left\{\left(\frac{1}{2}\right)^{(n-1)/2} A_0, \frac{4}{\gamma(n-1)}\right\}. \tag{3.15}$$

In addition, for any $\varepsilon > 0$, if

$$n \geq \max\left\{\frac{2}{\ln(2)}(\ln(A_0) + \ln(1/\varepsilon)), \frac{4}{\gamma\varepsilon}\right\} + 1,$$

then $A_n \leq \varepsilon$.

Proof. Note that for any $k \geq 0$ we have

$$\frac{1}{A_{k+1}} - \frac{1}{A_k} = \frac{A_k - A_{k+1}}{A_k A_{k+1}} \geq \gamma \frac{A_{k+1}}{A_k}.$$

For each k, there are two options:

(i) $A_{k+1} \leq \frac{1}{2} A_k$.

(ii) $A_{k+1} > \frac{1}{2} A_k$.

Under option (ii) we have

$$\frac{1}{A_{k+1}} - \frac{1}{A_k} \geq \frac{\gamma}{2}.$$

Suppose that n is even. If there are at least $\frac{n}{2}$ indices for which option (ii) occurs, then

$$\frac{1}{A_n} \geq \frac{\gamma n}{4},$$

and hence

$$A_n \leq \frac{4}{\gamma n}.$$

On the other hand, if this is not the case, then there are at least $\frac{n}{2}$ indices for which option (i) occurs, and consequently

$$A_n \leq \left(\frac{1}{2}\right)^{n/2} A_0.$$

We therefore obtain that in any case

$$A_n \leq \max\left\{\left(\frac{1}{2}\right)^{n/2} A_0, \frac{4}{\gamma n}\right\}. \tag{3.16}$$

If n is odd, then we can conclude that

$$A_n \leq A_{n-1} \leq \max\left\{\left(\frac{1}{2}\right)^{(n-1)/2} A_0, \frac{4}{\gamma(n-1)}\right\}. \tag{3.17}$$

Since the right-hand side of (3.17) is larger than the right-hand side of (3.16), the result (3.15) follows. In order to guarantee that the inequality $A_n \leq \varepsilon$ holds, it is sufficient that the inequality

$$\max\left\{\left(\frac{1}{2}\right)^{(n-1)/2} A_0, \frac{4}{\gamma(n-1)}\right\} \leq \varepsilon$$

holds. The latter is equivalent to the set of two inequalities

$$\begin{aligned}\left(\frac{1}{2}\right)^{(n-1)/2} A_0 &\leq \varepsilon,\\ \frac{4}{\gamma(n-1)} &\leq \varepsilon,\end{aligned}$$

which is the same as

$$\begin{aligned}n &\geq \frac{2}{\ln(2)}(\ln(A_0)+\ln(1/\varepsilon))+1,\\ n &\geq \frac{4}{\gamma\varepsilon}+1.\end{aligned}$$

Therefore, if

$$n \geq \max\left\{\frac{2}{\ln(2)}(\ln(A_0)+\ln(1/\varepsilon)), \frac{4}{\gamma\varepsilon}\right\}+1,$$

then the inequality $A_n \leq \varepsilon$ is guaranteed. □

We are now ready to prove a refined rate of convergence result for the sequence of function values generated by the alternating minimization method. In this result the number of iterations depend mildly on $H(\mathbf{x}_0)-H^*$ in the sense the the required number of iteration depends on $\ln(H(\mathbf{x}_0)-H^*)$ rather than on $H(\mathbf{x}_0)-H^*$.

Theorem 3.3. *Let $\{\mathbf{x}_n\}_{n\geq 0}$ be the sequence generated by the alternating minimization method. Then for all $n \geq 2$*

$$H(\mathbf{x}_n)-H^* \leq \max\left\{\left(\frac{1}{2}\right)^{\frac{n-1}{2}}(H(\mathbf{x}_0)-H^*), \frac{8\min\{L_1,L_2\}R^2}{n-1}\right\}.$$

In addition, an ε-optimal solution is obtained after at most

$$\max\left\{\frac{2}{\ln(2)}(\ln(H(\mathbf{x}_0)-H^*)+\ln(1/\varepsilon)), \frac{8\min\{L_1,L_2\}R^2}{\varepsilon}\right\}+2$$

iterations.

Proof. By Lemma 3.4 we have

$$A_k - A_{k+1} \geq \gamma A_{k+1}^2,$$

where $A_k = H(\mathbf{x}_k)-H^*$ and $\gamma = \frac{1}{2\min\{L_1,L_2\}R^2}$. The result now follows by invoking Lemma 3.6. □

4 Iteratively Reweighted Least Squares

In this section we will consider a well known method for solving problems involving sums of norms – the iteratively reweighted least squares method. We will recall its connection to the alternating minimization method. Based on the results obtained for the alternating minimization method, we will derive a non-asymptotic sublinear rate of convergence of the method. We will further study the method and show that we can derive an asymptotic rate of convergence that does not depend on the data of the problem, but rather on the smoothing parameter.

4.1 Problem formulation

Consider the general problem of minimizing the sum of a continuously differentiable function and sum of norms of affine mappings:

$$\text{(P)} \quad \begin{array}{ll} \min & s(\mathbf{y}) + \sum_{i=1}^{m} \|\mathbf{A}_i\mathbf{y} + \mathbf{b}_i\|_2 \\ \text{s.t.} & \mathbf{y} \in X, \end{array}$$

where $\mathbf{A}_i \in \mathbb{R}^{k_i \times n}, \mathbf{b}_i \in \mathbb{R}^{k_i}, i = 1, 2, \ldots, m$ and s is a continuously differentiable function over the closed and convex set $X \subseteq \mathbb{R}^n$; we further assume that ∇s is Lipschitz continuous with parameter $L_{\nabla s}$. This is a rather general model encompassing several important models and applications – some of them we now describe.

- l_1**-norm linear regression**. The problem is

  $$\min \|\mathbf{B}\mathbf{y} - \mathbf{c}\|_1,$$

  where $\mathbf{B} \in \mathbb{R}^{m \times n}, \mathbf{c} \in \mathbb{R}^m$. The problem obviously fits model (P) with $s \equiv 0, X = \mathbb{R}^n, \mathbf{A}_i = \mathbf{e}_i^T\mathbf{B}, \mathbf{b}_i = -c_i$.

- **The Fermat-Weber roblem.** Given m points $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m \in \mathbb{R}^n$ called *anchors* and weights $\omega_1, \omega_2, \ldots, \omega_m > 0$, the Fermat-Weber problem seeks $\mathbf{x} \in \mathbb{R}^n$ that minimizes the weighted sum of distances between $\mathbf{y}$ and the m anchors:

  $$\text{(FW)} \quad \min_{\mathbf{y} \in \mathbb{R}^n} \sum_{i=1}^{m} \omega_i \|\mathbf{y} - \mathbf{a}_i\|.$$

  The problem fits model (P) with $s = 0, X = \mathbb{R}^n, \mathbf{A}_i = \mathbf{I}_n, \mathbf{b}_i = -\mathbf{a}_i$. The Fermat-Weber problem has a long history and was investigated for many years in the optimization as well as location communities. More details on the history of the Fermat-Weber problem can be found for example in [12] as well as in the survey paper [11].

- l_1**-regularized least squares problem.** In the l_1-regularized least squares problem we minimize a sum of a least squares term and an l_1-based penalty:

  $$\min \|\mathbf{B}\mathbf{y} - \mathbf{c}\|_2^2 + \lambda\|\mathbf{D}\mathbf{y}\|_1.$$

  Here $\mathbf{B} \in \mathbb{R}^{m \times n}, \mathbf{c} \in \mathbb{R}^m, \mathbf{D} \in \mathbb{R}^{k \times n}$ and $\lambda > 0$. This model fits problem (P) with $s(\mathbf{y}) = \|\mathbf{B}\mathbf{y} - \mathbf{c}\|_2^2, \mathbf{A}_i = \lambda\mathbf{e}_i^T\mathbf{D}$ and $\mathbf{b}_i = 0$. This type of least squares problems has many applications in diverse areas such as statistics [16] and signal/image processing [15].

It is sometimes useful to consider the smooth approximation of the general problem (P) given by:

$$\text{(P}_\eta\text{)} \quad \begin{array}{ll} \min & S_\eta(\mathbf{y}) \equiv s(\mathbf{y}) + \sum_{i=1}^{m} \sqrt{\|\mathbf{A}_i\mathbf{y} + \mathbf{b}_i\|^2 + \eta^2} \\ \text{s.t.} & \mathbf{y} \in X. \end{array}$$

Here $\eta > 0$ is a smoothing parameter. The optimal value of problem (P_η) is denoted by S^*_η. Problem (P_η) can be interpreted as a smooth approximation of problem (P) since

$$\text{val(P)} \leq \text{val}(P_\eta) \leq \text{val(P)} + m\eta, \tag{4.1}$$

where for a minimization problem (D), val(D) denotes the optimal value of the problem. The relation (4.1) is a direct result from the fact that for any $\alpha \in \mathbb{R}$ the inequality $|\alpha| \leq \sqrt{\alpha^2 + \eta^2} \leq |\alpha| + \eta$ holds.

4.2 The iteratively reweighted least squares method

The *iteratively reweighted least squares* method – abbreviated IRLS – for solving (P_η) is the method defined below.

The Iteratively Reweighted Least Squares Method

Input: $\eta > 0$ - a given parameter.
Initialization: $\mathbf{y}_0 \in X$.
General Step (k=0,1,...):

$$\mathbf{y}_{k+1} \in \operatorname*{argmin}_{\mathbf{y} \in X} \left\{ s(\mathbf{y}) + \frac{1}{2} \sum_{i=1}^{m} \frac{\|\mathbf{A}_i \mathbf{y} + \mathbf{b}_i\|^2}{\sqrt{\|\mathbf{A}_i \mathbf{y}_k + \mathbf{b}_i\|^2 + \eta^2}} \right\}. \tag{4.2}$$

The IRLS method is a rather popular scheme used for example in robust regression [20], sparse recovery [10], but perhaps the most famous and oldest example of the IRLS method is Weiszfeld's method for solving the Fermat-Weber problem [29, 30] – an algorithm that was introduced in 1937. Since then, this method was extensively studied, see e.g., [18, 8, 28]. Despite the fact that the positivity of η is essential, since otherwise the method will not be well defined, the method with $\eta = 0$ was considered in the literature. For example, Weiszfeld's method is the IRLS method with $\eta = 0$ and much research was performed to analyze the conditions under which the method is indeed well-defined, see e.g., [8, 28].

It is well known that the IRLS method is actually the alternating minimization method applied to an auxiliary function, see e.g., [10]. We will now recall this connection, and for that we consider the following auxiliary problem:

$$\begin{array}{ll} \min & h_\eta(\mathbf{y}, \mathbf{z}) \equiv s(\mathbf{y}) + \frac{1}{2} \sum_{i=1}^{m} \left(\frac{\|\mathbf{A}_i \mathbf{y} + \mathbf{b}_i\|^2 + \eta^2}{z_i} + z_i \right) \\ \text{s.t.} & \mathbf{y} \in X \\ & \mathbf{z} \in [\eta/2, \infty)^m, \end{array} . \tag{4.3}$$

which fits into the general model (1.1) with

$$\begin{aligned} f(\mathbf{y}, \mathbf{z}) &= h_\eta(\mathbf{y}, \mathbf{z}), \\ g_1(\mathbf{y}) &= \delta(\mathbf{y}, X), \\ g_2(\mathbf{z}) &= \delta(\mathbf{z}, [\eta/2, \infty)^m), \end{aligned}$$

where for a set S, the indicator function $\delta(\cdot, S)$ is defined by

$$\delta(\mathbf{x}, S) = \begin{cases} 0 & \mathbf{x} \in S, \\ \infty & \text{else.} \end{cases}$$

The equivalence between problems (4.3) and (P_η) is in the sense that minimizing h_η with respect to $\mathbf{z}$ (while fixing $\mathbf{y}$) results with the function S_η. The following lemma states this property along with the explicit relation between the optimal solutions of the two problems.

Lemma 4.1. *(i) For any $\mathbf{y} \in X$, it holds that*

$$\min_{\mathbf{z} \in [\eta/2, \infty)^m} h_\eta(\mathbf{y}, \mathbf{z}) = S_\eta(\mathbf{y})$$

and the minimum is attained at $\mathbf{z}$ given by

$$z_i = \sqrt{\|\mathbf{A}_i \mathbf{y} - \mathbf{b}\|^2 + \eta^2}, \quad i = 1, 2, \ldots, m. \tag{4.4}$$

(ii) $\min_{\mathbf{y} \in X, \mathbf{z} \in [\eta/2, \infty)^m} h_\eta(\mathbf{y}, \mathbf{z}) = S_\eta^*$.

Proof. (i) Follows by the fact that given $\mathbf{y} \in X$ and $i \in \{1, 2, \ldots, m\}$, we have by the arithmetic-geometric mean inequality that

$$\frac{1}{2}\left(\frac{\|\mathbf{A}_i \mathbf{y} + \mathbf{b}_i\|^2 + \eta^2}{z_i} + z_i\right) \geq \sqrt{\|\mathbf{A}_i \mathbf{y} + \mathbf{b}_i\|^2 + \eta^2}$$

for all $z_i > 0$, and equality is satisfied if and only if $z_i = \sqrt{\|\mathbf{A}_i \mathbf{y} + \mathbf{b}_i\|^2 + \eta^2}$.
(ii) By part (i),

$$\min_{\mathbf{y} \in X, \mathbf{z} \in [\eta/2, \infty)^m} h_\eta(\mathbf{y}, \mathbf{z}) = \min_{\mathbf{y} \in X} \left\{ \min_{\mathbf{z} \in [\eta/2, \infty)^m} h_\eta(\mathbf{y}, \mathbf{z}) \right\} = \min_{\mathbf{y} \in X} S_\eta(\mathbf{y}) = S_\eta^*.$$

□

Remark 4.1. The constraint $\mathbf{z} \in [\eta/2, \infty)^m$ could have been replaced with the constraint $\mathbf{z} > \mathbf{0}$. However, due to reasons related to the theoretical analysis, we consider a feasible set which is also closed.

The alternating minimization method employed on problem (4.3) takes the following form: the $\mathbf{z}$-step at the k-iteration just consists of evaluating $z_i = \sqrt{\|\mathbf{A}_i \mathbf{y}_k - \mathbf{b}_i\|^2 + \eta^2}$, and the $\mathbf{y}$-step is exactly the one defined by (4.2). We therefore obtain that the IRLS method is exactly the alternating minimization method applied to the function

$$H(\mathbf{y}, \mathbf{z}) = h_\eta(\mathbf{y}, \mathbf{z}) + g_1(\mathbf{y}) + g_2(\mathbf{z})$$

with initial point $(\mathbf{y}_0, \mathbf{z}_0)$, where $\mathbf{z}_0$ is defined by

$$[\mathbf{z}_0]_i = \sqrt{\|\mathbf{A}_i \mathbf{y}_0 + \mathbf{b}_i\|^2 + \eta^2}.$$

A direct consequence of Lemma 4.1 is that

$$h_\eta(\mathbf{y}_k, \mathbf{z}_k) = \min_{\mathbf{z} \in [\eta/2, \infty)^m} h_\eta(\mathbf{y}_k, \mathbf{z}) = S_\eta(\mathbf{y}_k).$$

From this we can also conclude that the IRLS method produces a nonincreasing sequence with respect to S_η. Indeed, for any $k \geq 0$:

$$S_\eta(\mathbf{y}_k) = h_\eta(\mathbf{y}_k, \mathbf{z}_k) \geq h_\eta(\mathbf{y}_{k+1}, \mathbf{z}_{k+1}) = S_\eta(\mathbf{y}_{k+1}).$$

4.3 Nonasymptotic Sublinear Rate of Convergence

To derive a nonasymptotic sublinear rate of convergence result, we can invoke Theorem 3.3. For that, we need to compute the block Lipschitz constants of the function $h_\eta(\mathbf{y}, \mathbf{z})$. The gradient of h_η with respect to $\mathbf{z}$ is in fact not Lipschitz continuous. Therefore, $L_2 = \infty$ and we only need to compute L_1 - the Lipschitz constant of $\nabla_{\mathbf{y}} h_\eta(\cdot, \mathbf{z})$, which is given by

$$L_1 = L_{\nabla s} + \frac{1}{\eta}\lambda_{\max}\left(\sum_{i=1}^{m} \mathbf{A}_i^T \mathbf{A}_i\right).$$

Plugging the above expression of the block Lipschitz constant in Theorem 3.3, we obtain the following result on the sublinear convergence of the IRLS method.

Theorem 4.1 (sublinear rate of convergence of the IRLS method)**.** *Let $\{\mathbf{y}_k\}_{k\geq 0}$ be the sequence generated by the IRLS method with smoothing parameter $\eta > 0$. Then for any $n \geq 2$*

$$S_\eta(\mathbf{y}_n) - S_\eta^* \leq \max\left\{\left(\frac{1}{2}\right)^{\frac{n-1}{2}} (S_\eta(\mathbf{y}_0) - S_\eta^*), \frac{8(L_{\nabla s} + \frac{1}{\eta}\lambda_{\max}\left(\sum_{i=1}^m \mathbf{A}_i^T \mathbf{A}_i\right))R^2}{n-1}\right\}, \quad (4.5)$$

where R is given in (3.8).

Proof. Invoking Theorem 3.3, we obtain that

$$h_\eta(\mathbf{y}_n, \mathbf{z}_n) - S_\eta^* \leq \max\left\{\left(\frac{1}{2}\right)^{\frac{n-1}{2}} (h_\eta(\mathbf{y}_0, \mathbf{z}_0) - S_\eta^*), \frac{8(L_{\nabla s} + \frac{1}{\eta}\lambda_{\max}\left(\sum_{i=1}^m \mathbf{A}_i^T \mathbf{A}_i\right))R^2}{n-1}\right\},$$

where here we used the fact stated in Lemma 4.1 that $S_\eta^* = \min_{\mathbf{y}\in X, \mathbf{z}\in[\eta/2,\infty)^m} h_\eta(\mathbf{y}, \mathbf{z})$. The result now follows by noting that $h_\eta(\mathbf{y}_n, \mathbf{z}_n) = S_\eta(\mathbf{y}_n)$. □

4.4 Convergence of the sequence

We can now prove that the accumulation points of the sequence generated by the IRLS method are optimal solutions of problem (P_η). Although convergence of the entire sequence is not established, we are able to prove a result, that will be useful later on in the analysis of the asymptotic rate of convergence, that for any i, the sequence $\{\|\mathbf{A}_i\mathbf{y}_k + \mathbf{b}_i\|\}_{k\geq 0}$ converges. For that, we will require the following elementary fact on convex problems.

Lemma 4.2. *Let f_1, f_2 be two proper extended real-valued closed and convex functions over $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively and assume in addition that f_2 is strictly convex over its domain. Let $\mathbf{C} \in \mathbb{R}^{m\times n}$. Assume that $\mathbf{y}_1^*$ and $\mathbf{y}_2^*$ are optimal solutions of*

$$\min\{F(\mathbf{y}) \equiv f_1(\mathbf{y}) + f_2(\mathbf{C}\mathbf{y}) : \mathbf{y} \in \mathbb{R}^n\}.$$

Then $\mathbf{C}\mathbf{y}_1^* = \mathbf{C}\mathbf{y}_2^*$.

Proof. Assume in contradiction that $\mathbf{C}\mathbf{y}_1^* \neq \mathbf{C}\mathbf{y}_2^*$, and denote the optimal value by α. Then by Jensen's inequality and the strict convexity of f_2, it follows that for $\mathbf{z} = \frac{1}{2}\mathbf{y}_1^* + \frac{1}{2}\mathbf{y}_2^*$ we have

$$\begin{aligned} f_1(\mathbf{z}) &\leq \frac{1}{2}f_1(\mathbf{y}_1^*) + \frac{1}{2}f_1(\mathbf{y}_2^*), \\ f_2(\mathbf{C}\mathbf{z}) &< \frac{1}{2}f_2(\mathbf{C}\mathbf{y}_1^*) + \frac{1}{2}f_2(\mathbf{C}\mathbf{y}_2^*), \end{aligned}$$

and hence

$$\begin{aligned} F(\mathbf{z}) &= f_1(\mathbf{z}) + f_2(\mathbf{C}\mathbf{z}) \\ &< \frac{1}{2}(f_1(\mathbf{y}_1^*) + f_2(\mathbf{C}\mathbf{y}_1^*)) + \frac{1}{2}(f_1(\mathbf{y}_2^*) + f_2(\mathbf{C}\mathbf{y}_2^*)) \\ &= \frac{\alpha}{2} + \frac{\alpha}{2} \\ &= \alpha, \end{aligned}$$

contradicting the optimality of $\mathbf{y}_1^*$ and $\mathbf{y}_2^*$. □

We can now conclude the following property of the optimal solutions of (P_η).

Corollary 4.1. *Let Y^* be the set of optimal solutions of problem (P_η). Then for any $i \in \{1, 2, \ldots, m\}$ the set*

$$\mathbf{A}_i Y^* = \{\mathbf{A}_i \mathbf{y} : \mathbf{y} \in Y^*\}$$

is a singleton.

Proof. The function $t : \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_m} \to \mathbb{R}$ defined by

$$t((\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m)) = \sum_{i=1}^{m} \sqrt{\|\mathbf{w}_i + \mathbf{b}_i\|^2 + \eta^2}$$

is strictly convex since its Hessian is a positive definite matrix as a block-diagonal matrix whose i-th block is the $k_i \times k_i$ matrix given by

$$\begin{aligned} \nabla^2_{\mathbf{z}_i, \mathbf{z}_i} t(\mathbf{z}_1, \ldots, \mathbf{z}_m) &= \frac{1}{(\|\mathbf{z}_i + \mathbf{b}_i\|^2 + \eta^2)^{3/2}} \left[(\|\mathbf{z}_i + \mathbf{b}_i\|^2 \mathbf{I} - (\mathbf{z}_i + \mathbf{b}_i)(\mathbf{z}_i + \mathbf{b}_i)^T) + \eta^2 \mathbf{I}\right] \\ &\succeq \frac{\eta^2}{(\|\mathbf{z}_i + \mathbf{b}_i\|^2 + \eta^2)^{3/2}} \mathbf{I} \succ \mathbf{0}. \end{aligned}$$

Therefore, since the objective function in problem (P_η) is of the form

$$s(\mathbf{y}) + t(\mathbf{A}_1 \mathbf{y}, \mathbf{A}_2 \mathbf{y}, \ldots, \mathbf{A}_m \mathbf{y}),$$

with s being convex and t being strictly convex. It follows by Lemma 4.2 that the value of $(\mathbf{A}_1 \mathbf{y}, \ldots, \mathbf{A}_m \mathbf{y})$ is constant for all optimal solutions $\mathbf{y}$. □

Lemma 4.3. *Let $\{\mathbf{y}_k\}_{k \geq 0}$ be the sequence generated by the IRLS method with smoothing parameter $\eta > 0$. Then*

(i) any accumulation point of $\{\mathbf{y}_k\}_{k\geq 0}$ *is an optimal solution of* (P_η).

(ii) for any $i \in \{1, 2, \ldots, m\}$ *the sequence* $\{\|\mathbf{A}_i\mathbf{y}_k + \mathbf{b}_i\|\}_{k\geq 0}$ *converges.*

Proof. (i) Let $\mathbf{y}^*$ be an accumulation point of $\{\mathbf{y}_k\}_{k\geq 0}$. By Theorem 4.1, the closedness of X and the continuity of S_η, it follows that $S_\eta(\mathbf{y}^*) = S_\eta^*$, which shows that $\mathbf{y}^*$ is an optimal solution of (P_η).
(ii) To show the convergence of $\|\mathbf{A}_i\mathbf{y}_k + \mathbf{b}_i\|$ for any $i \in \{1, 2, \ldots, m\}$, define by $\mathbf{d}_i$ the vector which is equal to $\mathbf{A}_i\mathbf{y}^*$ for all optimal solutions $\mathbf{y}^*$ (the uniqueness follows from Corollary 4.1). Take a convergent subsequence $\{\|\mathbf{A}_i\mathbf{y}_{k_n} + \mathbf{b}_i\|\}_{n\geq 1}$. By part (i), $\{\mathbf{y}_{k_n}\}_{n\geq 0}$ converges to an optimal solution $\mathbf{y}^*$, and therefore as $n \to \infty$ we have

$$\|\mathbf{A}_i\mathbf{y}_{k_n} + \mathbf{b}_i\| \to \|\mathbf{A}_i\mathbf{y}^* + \mathbf{b}_i\| = \|\mathbf{d}_i + \mathbf{b}_i\|.$$

Since we showed that all subsequences of $\{\|\mathbf{A}_i\mathbf{y}_k + \mathbf{b}_i\|\}$ converge to the same value, it follows that it is a convergent sequence. □

4.5 Asymptotic sublinear rate of convergence

The result of Theorem 4.1 does not reveal the full strength of the IRLS method since the multiplicative constant in the efficiency estimate strongly depends on the data of the problem, that is, the Lipschitz constant of ∇s and the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_m$, and can be rather large. We will show in this section that we can establish an asymptotic sublinear rate of convergence that actually does not depends on either $L_{\nabla s}$ or $\mathbf{A}_1, \ldots, \mathbf{A}_m$, but only on the smoothing parameter η and the diameter R. We begin by proving the following simple lemma on the difference between the arithmetic and geometric means of two numbers:

Lemma 4.4. *Let* $a, b > 0$. *Then*

$$\frac{a+b}{2} - \sqrt{ab} \geq \frac{1}{16\max\{a,b\}}(a-b)^2. \tag{4.6}$$

Proof. Consider the function

$$f(x) = \frac{1}{2}(1+x) - \sqrt{x} - \frac{1}{16}(x-1)^2.$$

Note that

$$\begin{aligned} f'(x) &= \frac{1}{2} - \frac{1}{2\sqrt{x}} - \frac{1}{8}(x-1), \\ f''(x) &= \frac{1}{4x^{3/2}} - \frac{1}{8}, \end{aligned}$$

and hence

$$f(1) = f'(1) = 0,$$

as well as

$$f''(x) \geq 0 \text{ for all } 0 \leq x \leq 1.$$

Therefore, since f is convex over $[0,1]$ and since $f'(1) = 0$, it follows that $x = 1$ is the minimizer of f over $[0,1]$, and consequently for all $x \in [0,1]$ we have

$$\frac{1}{2}(1+x) - \sqrt{x} - \frac{1}{16}(x-1)^2 = f(x) \geq f(1) = 0. \tag{4.7}$$

Assume now that $0 < a \leq b$. Then using the inequality (4.7) with $x = \frac{a}{b}$ we obtain that

$$\frac{a+b}{2} - \sqrt{ab} = b\left(\frac{1}{2}\left[\frac{a}{b}+1\right] - \sqrt{\frac{a}{b}}\right) \geq \frac{b}{16}\left(\frac{a}{b}-1\right)^2 = \frac{1}{16b}(a-b)^2 = \frac{1}{16\max\{a,b\}}(a-b)^2.$$

If $a > b > 0$, then the same type of computation shows that

$$\frac{a+b}{2} - \sqrt{ab} \geq \frac{1}{16a}(a-b)^2 = \frac{1}{16\max\{a,b\}}(a-b)^2,$$

showing that the desired inequality (4.6) holds for any $a, b > 0$. □

We can use the latter lemma in order to show an important recurrence relation satisfied by the sequence of objective function values defined by the IRLS method.

Lemma 4.5. *Let $\{\mathbf{y}_k\}_{k\geq 0}$ be the sequence generated by the IRLS method with smoothing parameter $\eta > 0$. Then*

$$S_\eta(\mathbf{y}_k) - S_\eta(\mathbf{y}_{k+1}) \geq \frac{\eta}{16R^2} \min_{i=1,2,\ldots,m}\left\{1, \frac{\|\mathbf{A}_i\mathbf{y}_k + \mathbf{b}_i\|^2 + \eta^2}{\|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \eta^2}\right\}(S_\eta(\mathbf{y}_{k+1}) - S_\eta^*)^2.$$

Proof. First note that by (4.6) we have for all $\mathbf{w} \in \mathbb{R}^p$ and $z > 0$:

$$\begin{aligned}
\frac{1}{2}\left(\frac{\|\mathbf{w}\|^2+\eta^2}{z} + z\right) - \sqrt{\|\mathbf{w}\|^2+\eta^2} &\geq \frac{1}{16\max\left\{z, \frac{\|\mathbf{w}\|^2+\eta^2}{z}\right\}}\left(\frac{\|\mathbf{w}\|^2+\eta^2}{z} - z\right)^2 \\
&= \frac{z^2}{16\max\left\{z, \frac{\|\mathbf{w}\|^2+\eta^2}{z}\right\}}\left(\frac{\|\mathbf{w}\|^2+\eta^2}{z^2} - 1\right)^2 \\
&= \frac{1}{16}\min\left\{z, \frac{z^3}{\|\mathbf{w}\|^2+\eta^2}\right\}\left(\frac{\|\mathbf{w}\|^2+\eta^2}{z^2} - 1\right)^2 \\
&= \frac{z}{16}\min\left\{1, \frac{z^2}{\|\mathbf{w}\|^2+\eta^2}\right\}\left(\frac{\|\mathbf{w}\|^2+\eta^2}{z^2} - 1\right)^2.
\end{aligned}$$

Therefore, (denoting the i-th component of $\mathbf{z}_k$ by $[\mathbf{z}_k]_i$):

$$\begin{aligned}
h_\eta(\mathbf{x}_{k+\frac{1}{2}}) - h_\eta(\mathbf{x}_{k+1}) &= h_\eta(\mathbf{y}_{k+1}, \mathbf{z}_k) - h_\eta(\mathbf{y}_{k+1}, \mathbf{z}_{k+1}) \\
&= h_\eta(\mathbf{y}_{k+1}, \mathbf{z}_k) - S_\eta(\mathbf{y}_{k+1}) \\
&= \sum_{i=1}^m \left(\frac{1}{2} \left[\frac{\|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \eta^2}{[\mathbf{z}_k]_i} + [\mathbf{z}_k]_i \right] - \sqrt{\|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \eta^2} \right) \\
&\geq \frac{1}{16} \sum_{i=1}^m [\mathbf{z}_k]_i \min\left\{1, \frac{[\mathbf{z}_k]_i^2}{\|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \eta^2}\right\} \left(\frac{\|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \eta^2}{[\mathbf{z}_k]_i^2} - 1 \right)^2 \\
&= \frac{1}{16} \sum_{i=1}^m [\mathbf{z}_k]_i \min\left\{1, \frac{\|\mathbf{A}_i\mathbf{y}_k + \mathbf{b}_i\|^2 + \eta^2}{\|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \eta^2}\right\} \left(\frac{\|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \eta^2}{[\mathbf{z}_k]_i^2} - 1 \right)^2 \\
&\overset{[\mathbf{z}_k]_i \geq \eta}{\geq} \frac{\eta}{16} \sum_{i=1}^m \min\left\{1, \frac{\|\mathbf{A}_i\mathbf{y}_k + \mathbf{b}_i\|^2 + \eta^2}{\|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \eta^2}\right\} \left(\frac{\|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \eta^2}{[\mathbf{z}_k]_i^2} - 1 \right)^2 \\
&\geq \frac{\eta}{16} \min_{i=1,2,\ldots,m} \left\{1, \frac{\|\mathbf{A}_i\mathbf{y}_k + \mathbf{b}_i\|^2 + \eta^2}{\|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \eta^2}\right\} \sum_{i=1}^m \left(\frac{\|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \eta^2}{[\mathbf{z}_k]_i^2} - 1 \right)^2.
\end{aligned}$$

On the other hand,

$$\nabla_2 h_\eta(\mathbf{y}_{k+1}, \mathbf{z}_k) = \left(-\frac{\|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \eta^2}{[\mathbf{z}_k]_i^2} + 1 \right)_{i=1}^m.$$

Hence,

$$\|\nabla_2 h_\eta(\mathbf{y}_{k+1}, \mathbf{z}_k)\|^2 = \sum_{i=1}^m \left(\frac{\|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \eta^2}{[\mathbf{z}_k]_i^2} - 1 \right)^2.$$

We thus obtain that

$$h_\eta(\mathbf{x}_{k+\frac{1}{2}}) - h_\eta(\mathbf{x}_{k+1}) \geq \frac{\eta}{16} \min_{i=1,2,\ldots,m} \left\{1, \frac{\|\mathbf{A}_i\mathbf{y}_k + \mathbf{b}_i\|^2 + \eta^2}{\|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \eta^2}\right\} \|\nabla_2 h_\eta(\mathbf{x}_{k+\frac{1}{2}})\|^2. \tag{4.8}$$

Since $\mathbf{z} \mapsto h_\eta(\mathbf{y}_{k+1}, \mathbf{z})$ has a Lipschitz gradient with constant

$$M(\mathbf{y}_{k+1}) = \frac{1}{\eta^3} \max_{i=1,2,\ldots,m} \|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \frac{1}{\eta},$$

by Remark 4.1, it follows that

$$h_\eta(\mathbf{x}_{k+1}) - S_\eta^* \leq \|G^2_{M(\mathbf{y}_{k+1})}(\mathbf{x}_{k+\frac{1}{2}})\| \cdot \|\mathbf{x}_{k+\frac{1}{2}} - \mathbf{x}^*\| \leq \|G^2_L(\mathbf{x}_{k+\frac{1}{2}})\| R \tag{4.9}$$

for any $L \geq M(\mathbf{y}_{k+1})$. Note that

$$G^2_L(\mathbf{x}_{k+\frac{1}{2}}) = G^2_L(\mathbf{y}_{k+1}, \mathbf{z}_k) = L\left(\mathbf{z}_k - P_{[\frac{\eta}{2}, \infty)^m} \left[\mathbf{z}_k - \frac{1}{L} \nabla_2 h_\eta(\mathbf{y}_{k+1}, \mathbf{z}_k) \right] \right).$$

Since $\mathbf{z}_k \geq \eta\mathbf{e}$, it follows that for large enough L we have that $\mathbf{z}_k - \frac{1}{L}\nabla_2 h_\eta(\mathbf{y}_{k+1}, \mathbf{z}_k) \geq \frac{\eta}{2}\mathbf{e}$ and hence for such L we have

$$G^2_L(\mathbf{y}_{k+1}, \mathbf{z}_k) = L\left(\mathbf{z}_k - \left[\mathbf{z}_k - \frac{1}{L} \nabla_2 h_\eta(\mathbf{y}_{k+1}, \mathbf{z}_k \right]) \right) = \nabla_2 h_\eta(\mathbf{y}_{k+1}, \mathbf{z}_k),$$

and by (4.9) that

$$h_\eta(\mathbf{x}_{k+1}) - S_\eta^* \le \|\nabla_2 h_\eta(\mathbf{y}_{k+1}, \mathbf{z}_k)\| R,$$

which combined with (4.8) yields the required inequality:

$$\begin{aligned} S_\eta(\mathbf{y}_k) - S_\eta(\mathbf{y}_{k+1}) &= h_\eta(\mathbf{x}_k) - h_\eta(\mathbf{x}_{k+1}) \\ &\ge h_\eta(\mathbf{x}_{k+\frac{1}{2}}) - h_\eta(\mathbf{x}_{k+1}) \\ &\ge \frac{\eta}{16R^2} \min_{i=1,2,\ldots,m} \left\{ 1, \frac{\|\mathbf{A}_i\mathbf{y}_k + \mathbf{b}_i\|^2 + \eta^2}{\|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \eta^2} \right\} (S_\eta(\mathbf{y}_{k+1}) - S_\eta^*)^2. \end{aligned}$$

□

The latter lemma is the basis for the main asymptotic convergence result

Theorem 4.2 (asymptotic rate of convergence of the IRLS method)**.** *Let* $\{\mathbf{y}_k\}_{k\ge 0}$ *be the sequence generated by the alternating minimization. Then there exists* $K > 0$ *such that*

$$S_\eta(\mathbf{y}_n) - S_\eta^* \le \frac{48R^2}{\eta(n-K)}$$

for all $n \ge K+1$.

Proof. By Lemma 4.3 the expression $\|\mathbf{A}_i\mathbf{y}_k + \mathbf{b}_i\|$ converges for any $i \in \{1, 2, \ldots, m\}$. Therefore, for any i

$$\frac{\|\mathbf{A}_i\mathbf{y}_k + \mathbf{b}_i\|^2 + \eta^2}{\|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \eta^2} \to 1$$

as $k \to \infty$. Therefore, there exist $K_1 > 0$ such that for all $k > K_1$:

$$\frac{\|\mathbf{A}_i\mathbf{y}_k + \mathbf{b}_i\|^2 + \eta^2}{\|\mathbf{A}_i\mathbf{y}_{k+1} + \mathbf{b}_i\|^2 + \eta^2} \ge \frac{1}{2}.$$

Invoking Lemma 4.5, we thus obtain that for all $k > K_1$ the following inequality holds:

$$S_\eta(\mathbf{y}_k) - S_\eta(\mathbf{y}_{k+1}) \ge \frac{\eta}{32R^2}(S_\eta(\mathbf{y}_{k+1}) - S_\eta^*)^2.$$

In addition, since $S_\eta(\mathbf{y}_k) \to S_\eta^*$ (by Theorem 4.1), there exist $K_2 > 0$ such that for all $k > K_2$

$$S_\eta(\mathbf{y}_k) - S_\eta^* < \frac{16R^2}{\eta}.$$

Define $K = \lfloor \max\{K_1, K_2\} \rfloor + 1$. Then denoting $\gamma = \frac{\eta}{32R^2}$, we obtain that the relation $a_k - a_{k+1} \ge \gamma a_{k+1}^2$ holds for all $k \ge 0$ with $a_k = S_\eta(\mathbf{y}_{k+K}) - S_{\eta^*}$. In addition, $a_1 < \frac{1}{2\gamma} < \frac{1.5}{\gamma}, a_2 < \frac{1}{2\gamma} < \frac{1.5}{2\gamma}$, and hence, invoking Lemma 3.5, we obtain that for all $k \ge 1$ the inequality

$$a_k = S_\eta(\mathbf{y}_{k+K}) - S_\eta^* \le \frac{48\eta R^2}{k},$$

holds, which after the change of indices $n = k + K$, establishes the desired result. □

5 Solution of a Composite Model via Alternating Minimization

5.1 The composite model

Consider the problem

$$T^* = \min \{T(\mathbf{y}) \equiv q(\mathbf{y}) + r(\mathbf{A}\mathbf{y})\}, \tag{5.1}$$

where

- $q : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a closed, proper convex extended real-valued convex function.
- $r : \mathbb{R}^m \to \mathbb{R}$ is a real-valued convex function which is Lipschitz with constant L_r:

$$|r(\mathbf{z}) - r(\mathbf{w})| \leq L_r \|\mathbf{z} - \mathbf{w}\|, \quad \mathbf{z}, \mathbf{w} \in \mathbb{R}^m.$$

- $\mathbf{A}$ is an $m \times n$ matrix.

Several approaches have been devised to solve the problem under different types of assumptions, see for example [13, 9, 27]. A popular approach for solving the problem is to use penalty and to consider the following auxiliary problem:

$$T_\rho^* = \min_{\mathbf{y},\mathbf{z}} \left\{ T_\rho(\mathbf{y}, \mathbf{z}) = q(\mathbf{y}) + r(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{z} - \mathbf{A}\mathbf{y}\|^2 \right\}. \tag{5.2}$$

where $\rho > 0$ is a penalization parameter. Problem (5.2) fits into the general model (1.1) with

$$\begin{aligned} f(\mathbf{y}, \mathbf{z}) &= \frac{\rho}{2}\|\mathbf{z} - \mathbf{A}\mathbf{y}\|^2, \\ g_1(\mathbf{y}) &= q(\mathbf{y}), \\ g_2(\mathbf{z}) &= r(\mathbf{z}). \end{aligned}$$

In many scenarios, it is relatively simple to perform minimization with respect to either $\mathbf{y}$ or $\mathbf{z}$, but not with respect to the two vectors $\mathbf{y}, \mathbf{z}$ simultaneously. Therefore, it is quite natural to invoke the alternating minimization method, which is given now in details.

Alternating Minimization for Solving (5.2)

Input: $\rho > 0$ - a given parameter.
Initialization: $\mathbf{y}_0 \in \mathbb{R}^n, \mathbf{z}_0 \in \operatorname{argmin} \left\{ r(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{z} - \mathbf{A}\mathbf{y}_0\|^2 \right\}$.
General Step (k=0,1,...):

$$\begin{aligned} \mathbf{y}_{k+1} &\in \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^n} \left\{ q(\mathbf{y}) + \frac{\rho}{2}\|\mathbf{z}_k - \mathbf{A}\mathbf{y}\|^2 \right\}, \\ \mathbf{z}_{k+1} &= \operatorname*{argmin}_{\mathbf{z}\in\mathbb{R}^m} \left\{ r(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{z} - \mathbf{A}\mathbf{y}_{k+1}\|^2 \right\}. \end{aligned}$$

An important relation between problems (5.1) and (5.2) is that the optimal value of the auxiliary problem (5.2) is smaller or equal to that of the original problem (5.1).

Lemma 5.1. *$T_\rho^* \leq T^*$ for any $\rho \geq 0$.*

Proof. Indeed,

$$\begin{aligned} T_\rho^* &= \min_{\mathbf{y},\mathbf{z}} \left\{ q(\mathbf{y}) + r(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{y} - \mathbf{z}\|^2 \right\} \\ &\leq \min_{\mathbf{y},\mathbf{z}:\mathbf{A}\mathbf{y}=\mathbf{z}} \{q(\mathbf{y}) + r(\mathbf{z})\} \\ &= \min_{\mathbf{y}} \{q(\mathbf{y}) + r(\mathbf{A}\mathbf{y})\} = T^*. \end{aligned}$$

□

Before proceeding, we would like to further investigate the connections between problems (5.1) and (5.2). Fixing $\mathbf{y}$, and minimizing with respect to $\mathbf{z}$, we obtain that problem (5.2) is equivalent to

$$\min_{\mathbf{y}} \{q(\mathbf{y}) + r_\rho(\mathbf{A}\mathbf{y})\}, \tag{5.3}$$

where r_ρ is the so-called Moreau envelope [22] of r given by

$$r_\rho(\mathbf{w}) = \operatorname{argmin} \left\{ r(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{w} - \mathbf{z}\|^2 \right\}.$$

It is well known that r_ρ is a continuously differentiable function with Lipschitz constant $L_{\nabla r_\rho} = \rho$, and that the following inequality holds (see e.g., [3]):

$$r(\mathbf{w}) - \frac{L_r}{\rho} \leq r_\rho(\mathbf{w}) \leq r(\mathbf{w}) \text{ for all } \mathbf{w} \in \mathbb{R}^m,$$

and consequently we have

$$T^* - \frac{L_r}{\rho} \leq T_\rho^* \leq T^*. \tag{5.4}$$

5.2 Convergence Analysis

Invoking Theorem 3.3, we can establish the following result on the $O(1/k)$ rate of convergence of the auxiliary function T_ρ to the approximate optimal value T_ρ^*.

Lemma 5.2. *Let $\{(\mathbf{y}_k, \mathbf{z}_k)\}_{k\geq 0}$ the sequence generated by the alternating minimization method employed on problem (5.2). Then*

$$T_\rho(\mathbf{y}_k, \mathbf{z}_k) - T_\rho^* \leq \max\left\{ \left(\frac{1}{2}\right)^{\frac{k-1}{2}} (T_\rho(\mathbf{y}_0, \mathbf{z}_0) - T_\rho^*), \frac{8R^2\rho}{k-1} \right\}.$$

Proof. Note that since $f(\mathbf{y}, \mathbf{z}) = \frac{\rho}{2}\|\mathbf{A}\mathbf{y} - \mathbf{z}\|^2$, it follows that

$$L_1 = \frac{\rho}{2}\lambda_{\max}(\mathbf{A}^T\mathbf{A}), L_2 = \frac{\rho}{2},$$

and hence,

$$\min\{L_1, L_2\} = \frac{\rho}{2}\min\{\lambda_{\max}(\mathbf{A}^T\mathbf{A}), 1\} \leq \frac{\rho}{2}.$$

Therefore, invoking Theorem 3.3, we obtain the desired result. □

The latter lemma considers the rate of convergence of the approximate problem to the approximate optimal value. Our objective is to derive a complexity result that establishes the rate of convergence of the original function $T(\mathbf{y}_k)$ to the exact optimal value T^*. For that, we begin by proving a lemma that bounds the difference $T(\mathbf{y}_k) - T^*$ by an expression that is bounded away from zero by a term that depends on ρ.

Lemma 5.3. *Let $\{(\mathbf{y}_k, \mathbf{z}_k)\}_{k\geq 0}$ be the sequence generated by the alternating minimization methods employed on problem (5.2). Then for any $k \geq 2$:*

$$T(\mathbf{y}_k) - T^* \leq \max\left\{\left(\frac{1}{2}\right)^{\frac{k-1}{2}}\left(T(\mathbf{y}_0) - T^* + \frac{L_r}{\rho}\right), \frac{8R^2\rho}{k-1}\right\} + \frac{L_r^2}{2\rho},$$

where R is given in (3.8).

Proof. We have

$$\begin{aligned}
T(\mathbf{y}_k) &= q(\mathbf{y}_k) + r(\mathbf{A}\mathbf{y}_k) \\
&= q(\mathbf{y}_k) + r(\mathbf{z}_k) + r(\mathbf{A}\mathbf{y}_k) - r(\mathbf{z}_k) \\
&= T_\rho(\mathbf{y}_k, \mathbf{z}_k) - \frac{\rho}{2}\|\mathbf{A}\mathbf{y}_k - \mathbf{z}_k\|^2 + r(\mathbf{A}\mathbf{y}_k) - r(\mathbf{z}_k) \\
&\leq T_\rho(\mathbf{y}_k, \mathbf{z}_k) - \frac{\rho}{2}\|\mathbf{A}\mathbf{y}_k - \mathbf{z}_k\|^2 + L_r\|\mathbf{A}\mathbf{y}_k - \mathbf{z}_k\| \\
&\leq T_\rho(\mathbf{y}_k, \mathbf{z}_k) + \max_{t\geq 0}\left\{-\frac{\rho}{2}t^2 + L_r t\right\} \\
&= T_\rho(\mathbf{y}_k, \mathbf{z}_k) + \frac{L_r^2}{2\rho} \\
&\leq T_\rho^* + \max\left\{\left(\frac{1}{2}\right)^{\frac{k-1}{2}}(T_\rho(\mathbf{y}_0, \mathbf{z}_0) - T_\rho^*), \frac{8R^2\rho}{k-1}\right\} + \frac{L_r^2}{2\rho} \\
&\overset{T_\rho^*\leq T^*}{\leq} T^* + \max\left\{\left(\frac{1}{2}\right)^{\frac{k-1}{2}}(T_\rho(\mathbf{y}_0, \mathbf{z}_0) - T_\rho^*), \frac{8R^2\rho}{k-1}\right\} + \frac{L_r^2}{2\rho}
\end{aligned}$$

The result now follows from the facts that $T_\rho(\mathbf{y}_0, \mathbf{z}_0) \leq T(\mathbf{y}_0)$ and $T_\rho^* \geq T^* - \frac{L_r}{\rho}$. □

We are now ready to prove that an ε-optimal solution can be obtained after $O(1/\varepsilon^2)$ iterations.

Theorem 5.1. *Let $\varepsilon > 0$ and let $\{(\mathbf{y}_k, \mathbf{z}_k)\}_{k\geq 0}$ be the sequence generated by the alternating minimization method employed on problem (5.2) with $\rho = \frac{L_r^2}{\varepsilon}$. Then for any*

$$k \geq \max\left\{\frac{16R^2L_r^2}{\varepsilon^2}, \frac{2}{\ln(2)}\left[\ln\left(T(\mathbf{y}_0) - T^* + \frac{\varepsilon}{L_r}\right) + \ln\left(\frac{1}{\varepsilon}\right)\right]\right\} + 3 \tag{5.5}$$

it holds that $T(\mathbf{y}_k) - T^* \leq \varepsilon$.

Proof. Taking $\rho = \frac{L_r^2}{\varepsilon}$, we have by Lemma 5.3 that

$$T(\mathbf{y}_k) - T^* \leq \max\left\{\left(\frac{1}{2}\right)^{\frac{k-1}{2}}\left(T(\mathbf{y}_0) - T^* + \frac{\varepsilon}{L_r}\right), \frac{8R^2L_r^2}{\varepsilon(k-1)}\right\} + \frac{\varepsilon}{2}.$$

Therefore, to guarantee the inequality $T(\mathbf{y}_k) - T^* \leq \varepsilon$, it is sufficient that the following two inequalities hold:

$$\begin{aligned} \frac{8R^2L_r^2}{\varepsilon(k-1)} &\leq \frac{\varepsilon}{2} \\ \left(\frac{1}{2}\right)^{\frac{k-1}{2}} \left(T(\mathbf{y}_0) - T^* + \frac{\varepsilon}{L_r}\right) &\leq \frac{\varepsilon}{2}. \end{aligned}$$

The above two inequalities are the same as

$$\begin{aligned} k &\geq \frac{16R^2L_r^2}{\varepsilon^2} + 1, \\ k &\geq \frac{2}{\ln(2)} \left[\ln\left(T(\mathbf{y}_0) - T^* + \frac{\varepsilon}{L_r}\right) + \ln\left(\frac{1}{\varepsilon}\right)\right] + 3. \end{aligned}$$

Since the above two inequalities are satisfied if condition (5.5) holds, the result follows. □

Remark 5.1. Another solution methodology is to employ the smoothing approach to the composite model (5.1). This means that an optimal gradient method is employed on the smooth problem (5.3). By [3, Theorem 3.1], it follows that if we choose $\rho = \frac{L_r^2}{\varepsilon}$ (exactly the same as the choice in the alternating minimization method), then an an ε-optimal solution is attained after at most

$$\left\lceil \frac{\sqrt{2}\|\mathbf{A}\|L_r\Lambda}{\varepsilon} \right\rceil$$

iterations, where Λ is another type of diameter. By this analysis, it seems that the smoothing approach is preferable since it requires only $O(1/\varepsilon)$ iterations and not $O(1/\varepsilon^2)$. However, there is one advantage to the alternating minimization approach here since its efficiency estimate does not depend on the norm of $\mathbf{A}$, which might be a large number. A different methodology for solving the composite model is through the *alternating direction method* (ADM) of multipliers. It was shown in [17], and later on in [21], that an ergodic sublinear $O(1/\varepsilon)$ rate of convergence can be established, where the corresponding constant also depends on the norm of $\mathbf{A}$, see also [26] for further extensions.

References

[1] A. Auslender. *Optimisation*. Masson, Paris, 1976. Méthodes numériques, Maîtrise de Mathématiques et Applications Fondamentales.

[2] A. Beck and M. Teboulle. Gradient-based algorithms with applications to signal recovery problems. In D. Palomar and Y. Eldar, editors, *Convex Optimization in Signal Processing and Communications*, pages 139–162. Cambridge University Press, 2009.

[3] A. Beck and M. Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22(2):557–580, 2012.

[4] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM J. Optim*, 23(2):2037–2060, 2013.

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

[6] D. P. Bertsekas and J. N. Tsitsiklis. Parallel and distributed computation. *Prentice-Hall International Editions, Englewood Cliffs,NJ*, 1989.

[7] N. Bissantz, L. Dümbgen, A. Munk, and B. Stratmann. Convergence analysis of generalized iteratively reweighted least squares algorithms on convex function spaces. *SIAM J. Optim.*, 19(4):1828–1845, 2008.

[8] R. Chandrasekaran and A. Tamir. Open questions concerning Weiszfeld's algorithm for the Fermat-Weber location problem. *Math. Programming*, 44(3):293–295, 1989.

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

[10] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Comm. Pure Appl. Math.*, 63(1):1–38, 2010.

[11] Z. Drezner, K. Klamroth, A. Schobel, and G. O. Wesolowsky. The weber problem. In Z. Drezner and H. W. Hamacher, editors, *Facility Location: Applications and Theory*, pages 1–36.

[12] H. A. Eiselt and V. Marianov. Pioneering developments in location analysis. In H. A. Eiselt and V. Marianov, editors, *Foundations of Location Analysis*, International Series in Operations Research & Management Science, pages 3–22. Springer, 2011.

[13] E. Esser, X. Zhang, and T. F. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM J. Imaging Sci.*, 3(4):1015–1046, 2010.

[14] L. Grippo and M. Sciandrone. Globally convergent block-coordinate techniques for unconstrained optimization. *Optimization Methods and Software*, 10:587–637, 1999.

[15] P. C. Hansen, J. G. Nagy, and D. P. O'Leary. *Deblurring images*, volume 3 of *Fundamentals of Algorithms*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006. Matrices, spectra, and filtering.

[16] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

[17] B. He and X. Yuan. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM J. Numer. Anal.*, 50(2):700–709, 2012.

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

[19] T. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Reasearch*, 46:157–178, 1993.

[20] P. McCullagh and J. A. Nelder. *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1983.

[21] R. D. C. Monteiro and B. F. Svaiter. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM J. Optim.*, 23(1):475–507, 2013.

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

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

[24] Y. E. Nesterov. Gradient methods for minimizing composite objective functions. *Mathematical Programming*, 140(1):125–161, 2013.

[25] J. M. Ortega and W. C. Rheinboldt. Iterative solution of nonlinear equations in several variables. *Academic Press, New York*, 1970.

[26] R. Shefi and M. Teboulle. Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. 2013. SIAM J. Optimization, to appear.

[27] P. Tseng. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM J. Control Optim.*, 29(1):119–138, 1991.

[28] Y. Vardi and C. H. Zhang. A modified Weiszfeld algorithm for the Fermat-Weber location problem. *Math. Program.*, 90(3):559–566, 2001.

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

[30] E. Weiszfeld. On the point for which the sum of the distances to n given points is minimum. *Ann. Oper. Res.*, 167:7–41, 2009. Translated from the French original [Tohoku Math. J. **43** (1937), 355–386] and annotated by Frank Plastria.