# A First Order Method for Solving Convex Bi-Level Optimization Problems

Shoham Sabach* Shimrit Shtern†

March 22, 2017

**Abstract**

In this paper we study convex bi-level optimization problems for which the inner level consists of minimization of the sum of smooth and nonsmooth functions. The outer level aims at minimizing a smooth and strongly convex function over the optimal solutions set of the inner problem. We analyze a first order method which is based on an existing fixed-point algorithm. Global sublinear rate of convergence of the method is established in terms of the inner objective function values.

## 1 Introduction

In this paper we are interested in the following bi-level optimization problem (where we use the terminology of inner and outer levels). The *outer* level is given by the following constraint minimization problem

$$\min_{\mathbf{x}\in X^*} \omega(\mathbf{x}), \tag{MNP}$$

where $\omega$ is a strongly convex and differentiable function while $X^*$ is the, assumed nonempty, set of minimizers of the *inner* level problem, which is the classical convex composite model, given by

$$\min_{\mathbf{x}\in\mathbb{R}^n} \{\varphi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}, \tag{P}$$

where $f$ is a continuously differentiable function and $g$ is an extended valued (possibly nonsmooth) function, see next section for precise assumptions. We denote the unique optimal solution of problem (MNP) by $\mathbf{x}^*_{mn}$, following the notation used in [2].

The most known indirect method (the meaning of direct and indirect method will be made precise in the following lines) for solving problem (MNP) is by the well-known

*Department of Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa 3200003, Israel. E-mail: ssabach@ie.technion.ac.il.

†Department of Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa 3200003, Israel. E-mail: shimrits@tx.technion.ac.il.

Tikhonov regularization [14] which suggests solving the following alternative regularized problem, for some $\lambda > 0$,

$$\min_{\mathbf{x}\in\mathbb{R}^n} \{\varphi_\lambda(\mathbf{x}) := \varphi(\mathbf{x}) + \lambda\omega(\mathbf{x})\}. \tag{$Q_\lambda$}$$

In [8] the authors treat the case that $g$ is an indicator function of a closed and convex set $X$, and show that under some restrictive conditions including $X$ being a polyhedron, there exists a small enough $\lambda^* > 0$ such that the optimal solution of problem ($Q_{\lambda^*}$) is the optimal solution of problem (MNP), see [8, Theorem 9]. However, in practice, even for this specific case, the value of $\lambda^*$ is unknown, and so ($Q_\lambda$) must be solved for a sequence of regularizing parameters $\{\lambda_k\}_{k\in\mathbb{N}}$ for which $\lambda_k \to 0$ as $k \to \infty$. In [13] Solodov showed that, provided that $\sum_{k=1}^{\infty} \lambda_k = \infty$ and $g$ is again an indicator function of a closed and convex set, there is no need to find the optimal solution of problem ($Q_{\lambda_k}$), $k \in \mathbb{N}$, and it is sufficient to approximate its solution by performing single projected gradient step on $\varphi_{\lambda_k}$ for all $k \in \mathbb{N}$. In the case that both $f$ and $\omega$ are differentiable with Lipschitz continuous gradients, the generated sequence converges to the optimal solution of problem (MNP), even if $\omega$ is not strongly convex. Thus, the algorithm suggested in [13] provides a *direct* method for solving problem (MNP). Another direct approach to solve problem (MNP) is the *Hybrid Steepest Descent Method* (HSDM) presented in [16, Section 17.3.2], which was proved to converge to the optimal solution of problem (MNP) provided that $\lambda_k \to 0$ as $k \to \infty$ and $\sum_{k=1}^{\infty} \lambda_k = \infty$. In [11], an extension of the HSDM is suggested for the case where $g(\cdot) := 0$ and $\omega$ is not necessarily differentiable or strongly convex but has bounded subgradients on the optimal set.

The major missing part of these papers is that while convergence was proven the convergence rates of these algorithms are unknown. Very recently, a new direct first order method for solving problem (MNP), called the Minimal Norm Gradient (MNG) was proposed in [2], for which the authors proved an $O(1/\sqrt{k})$ rate of convergence result, in terms of the inner objective function values. Even though the authors of [2] deal with the specific case of problem (P) for which the nonsmooth function $g$ is assumed to be an indicator function of a nonempty, closed and convex set, it seems that their analysis carries over even in the more general setting of this paper, that is, for any convex and extended valued function $g$. The MNG method is based on the cutting plane idea which means that at each iteration of the algorithm two specific half-spaces are constructed and then a minimization of the outer objective function $\omega$ over the intersection of these half-spaces is solved (see more detail in Section 2.2). In some cases, computing a solution to this minimization task can be done analytically. However, for some choice of outer function $\omega$ obtaining a solution might be computationally expensive and require an additional nested algorithm to approximate its solution.

Inspired by [2] and motivated by the limitations of the MNG method (as discussed in Section 2.2) we are interested in pursuing the research on bi-level optimization problems

in the following way. We study and analyze a first order method[1] for solving problem (MNP) with a non-asymptotic $O(1/k)$ global rate of convergence in terms of the inner objective function values, which we call BiG-SAM. In addition to the improved rate of convergence, BiG-SAM seems to be simpler and cheaper than the MNG method in the following sense. The operation in the algorithm which relates to the inner problem is of the same complexity as in the MNG method. On the other hand, the operation with respect to the outer problem is very simple in our case, and consists of computing the gradient of the objective function $\omega$, in the MNG method two tasks are needed: computation of the gradient of $\omega$ and a minimization of $\omega$ over the intersection of two half-spaces, whose computational cost highly depend on the function $\omega$, as discussed below (see Section 3.1).

Another contribution of this paper is the fact that BiG-SAM can also be used in situations for which the outer objective function $\omega$ is strongly convex but not necessarily smooth. In this case, we show that BiG-SAM solves problem (MNP) but with respect to the Moreau envelope of $\omega$ instead of $\omega$ itself. In this case we offer a new concept of measuring rate of convergence. This property of BiG-SAM allows considering outer functions which are not necessarily smooth and include, for example, functions with sparsity terms (see more details in Section 4.1).

The paper is organized in the following way. In Section 2 we discuss the optimization framework of the class of bi-level problems, then we give all notations and auxiliary results that are needed for the forthcoming sections. We conclude this section with a short overview of the MNG method developed in [2] (see Section 2.2). Section 3 is devoted to an algorithm which forms the basis of BiG-SAM. We first discuss it in its generality for solving a certain class of fixed-point problems (see Section 3.1) and then we specify it for solving the bi-level problems described in Section 2.1. In Section 4 we prove rate of convergence results of BiG-SAM. This section also includes our results in the case where the outer function is not necessarily smooth. Section 5 contains numerical experiments comparing MNG to BiG-SAM and showing its computational superiority in obtaining faster rates.

Throughout the paper we denote vectors by boldface letters. The notation $\langle \cdot, \cdot \rangle$ is used to denote the inner product of two vectors and $\|\cdot\|$ is the norm associated with this inner product, unless stated otherwise.

# 2 Optimization Framework and Mathematical Tools

## 2.1 Convex Bi-Level Optimization

In this paper we are focusing on bi-level optimization problems which are formulated as follows. We first discuss the inner level problem which is given by,

$$\min_{\mathbf{x}\in\mathbb{R}^n} \{\varphi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}. \tag{P}$$

[1] which is based on an existing algorithm, proposed in [15], for solving a certain class of fixed point problems (see precise details in Section 3.1).

The standing assumption on the functions of the inner level problem is recorded now.

**Assumption A.** (i) $f : \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable such that its gradient is Lipschitz with constant $L_f$, that is,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L_f \|\mathbf{x} - \mathbf{y}\|, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

(ii) $g : \mathbb{R}^n \to (-\infty, \infty]$ is proper, lower semicontinuous and convex.

(iii) The set $X^*$ of all optimal solutions of problem (P) is nonempty, that is, $X^* \neq \emptyset$.

Problem (P), which consists of minimizing the sum of a smooth function $f$ and a possibly nonsmooth function $g$, is one of the most studied models in modern optimization with a huge body of literature (see, for instance, [4] and the references therein). The basic algorithm for solving problem (P) is the so called *Proximal Gradient* (PG) or proximal forward-backward method (see [7, 12] for the origin of the algorithm and [3] for the rate of convergence result including an accelerated version), which iteratively generates a sequence $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$ starting from an arbitrary point $\mathbf{x}^0 \in \mathbb{R}^n$ via the following rule

$$\mathbf{x}^{k+1} = \operatorname{prox}_{tg}\left(\mathbf{x}^k - t\nabla f\left(\mathbf{x}^k\right)\right), \quad k \in \mathbb{N}, \tag{2.1}$$

for some step-size $t > 0$. The main operation of this algorithm is the computation of the *Moreau proximal mapping* of a proper, lower semicontinuous and convex function $h : \mathbb{R}^n \to (-\infty, \infty]$ which is denoted and defined by

$$\operatorname{prox}_h(\mathbf{x}) := \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^n} \left\{ h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2 \right\}. \tag{2.2}$$

The PG method can also be seen as a fixed-point algorithm where the iterated mapping is given (using the notation of [2]) by

$$T_t(\mathbf{x}) := \operatorname{prox}_{tg}(\mathbf{x} - t\nabla f(\mathbf{x})), \tag{2.3}$$

and is called the *prox-grad mapping*. In the case where $g$ is the indicator function $\delta_X$ of a set $X$, defined to be zero on $X$ and $+\infty$ on $\mathbb{R}^n \setminus X$, the prox-grad mapping coincides with the *proj-grad mapping* (which is discussed in [2, 10]), since in this case the proximal mapping of $g$ is exactly the orthogonal projection onto $X$. The prox-grad mapping possesses the following two important properties which are relevant to our analysis (see [1, Proposition 12.27, Page 176] and [4, Section 2.3.2, Page 48]). The second property characterizes the set of all fixed points of $T_t$, which is denoted by $\operatorname{Fix}(T_t)$ and defined by $\operatorname{Fix}(T_t) = \{\mathbf{x} \in \mathbb{R}^n : T_t(\mathbf{x}) = \mathbf{x}\}$.

**Lemma 1.** (i) *The prox-grad mapping $T_t$ is nonexpansive for all $t \in (0, 1/L_f]$, that is,*

$$\|T_t(\mathbf{x}) - T_t(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{2.4}$$

(ii) *Fixed points of the prox-grad mapping* $T_t$ *are optimal solutions of problem* (P) *and vice versa, that is,*

$$\mathbf{x} \in X^* \quad \Leftrightarrow \quad \mathbf{x} = T_t(\mathbf{x}) = \operatorname{prox}_{tg}(\mathbf{x} - t\nabla f(\mathbf{x})). \tag{2.5}$$

*Therefore, we have that* $\operatorname{Fix}(T_t) = X^*$ *for all* $t > 0$.

The following result will be essential for the rate of convergence analysis presented in Section 4 (*cf.* [3, Lemma 2.3, Page 190]).

**Proposition 1.** *Suppose that Assumption A holds true. Let* $\mathbf{x} \in \mathbb{R}^n$ *and denote* $\mathbf{x}^+ = T_t(\mathbf{x})$. *Then, for any* $t \leq 1/L_f$ *and* $\mathbf{u} \in \mathbb{R}^n$, *we have*

$$\varphi(\mathbf{x}^+) - \varphi(\mathbf{u}) \leq \frac{1}{t}\langle \mathbf{x} - \mathbf{x}^+, \mathbf{x} - \mathbf{u}\rangle - \frac{1}{2t}\|\mathbf{x} - \mathbf{x}^+\|^2. \tag{2.6}$$

To conclude the discussion on the inner problem (P), we note that it is well-known that the PG method has an $O(1/k)$ rate of convergence in terms of the objective function values $\varphi$ (see [3]). In this respect, the method proposed in this paper shares the same rate of convergence as the PG method while capable of solving the more complicated bi-level problem described in detail now.

We now turn to discuss the outer problem. As mentioned in the introduction, the outer problem is given by the following convex constrained problem

$$\min_{\mathbf{x} \in X^*} \omega(\mathbf{x}), \tag{MNP}$$

where $X^*$ is the optimal solution set of problem (P). Here, using the same terminology as in [2], we refer to this outer problem as the *Minimal Norm Problem* (MNP). The standing assumption on the objective function $\omega$ of problem (MNP) is recorded now.

**Assumption B.** (i) $\omega : \mathbb{R}^n \to \mathbb{R}$ is strongly convex with parameter $\sigma > 0$.

(ii) $\omega$ is a continuously differentiable function such that $\nabla\omega$ is Lipschitz continuous with constant $L_\omega$.

It should be noted that Assumption B(i) here is slightly stronger than the corresponding assumption given in [2, Page 27], since we do not only assume differentiability of $\omega$, as assumed in [2], but additionally assume that its gradient is Lipschitz continuous. However, in practice, most of the interesting examples of $\omega$ do satisfy this additional assumption.

A function that would be essential in our paper is the well-known *Moreau envelope* of a given function $\omega$, which is denoted by $M_{s\omega}$, and defined by

$$M_{s\omega}(\mathbf{x}) = \min_{\mathbf{u} \in \mathbb{R}^n} \left\{ \omega(\mathbf{u}) + \frac{1}{2s}\|\mathbf{u} - \mathbf{x}\|^2 \right\}. \tag{2.7}$$

It is well-known that $M_{s\omega}$ is continuously differentiable on $\mathbb{R}^n$ with an $1/s$-Lipschitz continuous gradient (see [1, Proposition 12.29, Page 176]), which is given by

$$\nabla M_{s\omega}(\mathbf{x}) = \frac{1}{s}\left(\mathbf{x} - \operatorname{prox}_{s\omega}(\mathbf{x})\right). \tag{2.8}$$

Another property of the Moreau envelope that plays a central role in this paper is that, if the corresponding function $\omega$ is strongly convex then its Moreau envelope is also strongly convex as recorded in the following result (for completeness, the proof given in Appendix A).

**Proposition 2.** *Let $\omega : \mathbb{R}^n \to (-\infty, \infty]$ be a strongly convex function with strong convexity parameter $\sigma$ and let $s > 0$. Then, the Moreau envelope $M_{s\omega}$ is strongly convex with parameter $\sigma/(1 + s\sigma)$.*

See Section 4.1 for more details on the Moreau envelope relevant for our discussion.

A mapping $S : \mathbb{R}^n \to \mathbb{R}^n$ is said to be *$\beta$-contraction* if there exists $\beta < 1$ such that

$$\|S(\mathbf{x}) - S(\mathbf{y})\| \leq \beta \|\mathbf{x} - \mathbf{y}\|, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

For functions $\omega$ which satisfy Assumption B we have the following result which is crucial for our derivations. Although this result seem to be classic, we did not find an exact reference for its proof and therefore, for the sake of completeness, we provide a proof in Appendix B.

**Proposition 3.** *Suppose that Assumption B holds. Then, the mapping defined by $S_s = I - s\nabla\omega$, where $I$ is the identity operator, is a contraction for all $s \leq 2/(L_\omega + \sigma)$, that is,*

$$\|\mathbf{x} - s\nabla\omega(\mathbf{x}) - (\mathbf{y} - s\nabla\omega(\mathbf{y}))\| \leq \sqrt{1 - \frac{2s\sigma L_\omega}{\sigma + L_\omega}} \|\mathbf{x} - \mathbf{y}\|, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{2.9}$$

We now conclude this section by giving a short overview on the MNG method developed in [2].

## 2.2 The Minimal Norm Gradient Method

The MNG method of [2] was designed to tackle bi-level optimization problems for which $g(\cdot) := \delta_X(\cdot)$. In this case $\varphi(\mathbf{x}) = f(\mathbf{x})$, for all $\mathbf{x} \in X$.

Each iteration of the MNG method consists of three main computational tasks.

(i) Computing the proj-grad mapping $T_t$ in order to construct the first half-space.

(ii) Computing the gradient of $\omega$, which is needed to construct the second half-space.

(iii) Minimizing $\omega$ over the intersection of these two half spaces.

The first two tasks are standard in first order methods and consist of computing gradient and projections. On the other hand, the third task (which depends on $\omega$) is more involved and might requires a nested optimization algorithm . Thus, in many scenarios, we end up with *nested* schemes which implies that: (i) there is accumulation of computational error in each step, and (ii) the stopping criteria of the nested algorithm at each iteration is not well- defined. Therefore, the third task determines the computational complexity of the entire method, and thus the applicability of the MNG method for certain implementation.

In the case where $\omega(\cdot) := \|\cdot\|_{\mathbf{Q}}^2$, where $\mathbf{Q}$ is a positive definite matrix, the computation is easy and given by an explicit formula as noted in [2, Example 1, Page 36], although it may require some decomposition and inversion of matrix $\mathbf{Q}$ (see Section 5 for more details).

The main result derived in [2] (*cf.* [2, Theorems 4.1 and 4.2, Pages 37 and 39]) is valid when Assumptions A and B hold and when $g(\cdot) := \delta_X(\cdot)$. As we already mentioned, in [2] the authors did not assume that $\nabla\omega$ is Lipschitz continuous, only continuously differentiable. We state here the following result which deals with the case where the Lipschitz constant $L_f$ of $\nabla f$ is known (for a backtracking version see [2]).

**Proposition 4.** *Let $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$ be the sequence generated by the MNG method. Then, the sequence $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$ converges to the optimal solution $\mathbf{x}_{mn}^*$ of problem (MNP) and, for any $k \in \mathbb{N}$, we have that*

$$\min_{1\leq l\leq k} \varphi\left(T_{1/L_f}\left(\mathbf{x}^l\right)\right) - \varphi\left(\mathbf{x}_{mn}^*\right) \leq \frac{\rho L_f \left\|\mathbf{x}^0 - \mathbf{x}_{mn}^*\right\|^2}{\sqrt{k}},$$

*where $\rho = 1$ if $X = \mathbb{R}^n$ and $\rho = 4/3$ otherwise.*

It should be noted that the MNG method is not a feasible method in the sense that $\mathbf{x}^k$, $k \in \mathbb{N}$, does not necessarily belongs to the constraint set $X$ and therefore the rate of convergence result is obtained on the feasible sequence $\{T_{1/L_f}(\mathbf{x}^k)\}_{k\in\mathbb{N}}$, $k \in \mathbb{N}$. Furthermore, though in the original paper the authors only discuss the case where $g(\cdot) := \delta_X(\cdot)$ the result can actually be extended to the more general case given in the introduction.

Our main goal in this paper is to study a different algorithm for solving bi-level optimization problems, than the MNG method, for which *we prove a rate of convergence* in terms of the inner objective function values that is superior to the rate of the MNG method (given in Proposition 4 above). In addition to complexity aspect, the studied method, which is discussed in the next section, *is simpler and capable of tackling bi-level problems for which the outer objective function is not necessarily smooth*. This attractiveness is mainly due to the fact that the studied method does not require the minimization of the outer objective function $\omega$ over two half-spaces as needed in the MNG method.

# 3 The Sequential Averaging Method

## 3.1 The General Framework

Our approach here is based on taking an existing algorithm, that we call Sequential Averaging Method (SAM), which was developed in [15] for solving a certain class of fixed-point problems (see precise details below), and determining how it can be used in the setting of bi- level optimization problems as described in Section 2. It should be noted that in [15] it is already proved that SAM generates a sequence which converges to a solution of the corresponding fixed-point problem. Now we will discuss in detail the method proposed in [15].

The problem of main interest in [15] is finding a fixed-point of the nonexpansive mapping $T$, that is, $\mathbf{x}^* \in \mathrm{Fix}(T)$, which also satisfies certain property with respect to a contraction mapping $S$ over all points which belong to $\mathrm{Fix}(T)$. This property is formulated using the following variational equality

$$\langle \mathbf{x}^* - S(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \quad \forall\, \mathbf{x} \in \mathrm{Fix}(T). \tag{3.1}$$

This means that the problem here is to find a fixed-point of the mapping $T$ which is "better" than all other fixed-points of $T$ in the sense of inequality (3.1).

The SAM iteratively generates a sequence $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ starting from any $\mathbf{x}^0 \in \mathbb{R}^n$ by averaging the two mappings $S$ and $T$ in the following way

$$\mathbf{x}^k = \alpha_k S\left(\mathbf{x}^{k-1}\right) + (1-\alpha_k)\, T\left(\mathbf{x}^{k-1}\right),$$

where $\{\alpha_k\}_{k\in\mathbb{N}}$ is a well-chosen sequence of real numbers from $(0,1]$ which satisfies the following assumption.

**Assumption C.** Let $\{\alpha_k\}_{k\in\mathbb{N}}$ be a sequence of real numbers in $(0,1]$ which satisfies $\lim_{k\to\infty} \alpha_k = 0$, $\sum_{k=1}^{\infty} \alpha_k = \infty$ and $\lim_{k\to\infty} \alpha_{k+1}/\alpha_k = 1$.

It should be noted that Assumption C holds true for several choices of sequences $\{\alpha_k\}_{k\in\mathbb{N}}$ which include, for example, $\alpha_k = \alpha/k$, $k \in \mathbb{N}$ for any choice of $\alpha \in (0,1]$.

The following lemma summarizes the main known results on SAM, which were proved in [15, Theorem 3.2] and formed the basis for this paper.

**Lemma 2.** *Assume that $S : \mathbb{R}^n \to \mathbb{R}^n$ is a $\beta$-contraction and that $T : \mathbb{R}^n \to \mathbb{R}^n$ is nonexpansive mapping, for which $\mathrm{Fix}(T) \neq \emptyset$. Let $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ be the sequence generated by SAM. If Assumption C holds true, then the following assertions are valid.*

(i) *The sequence $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ is bounded, in particular, for any $\tilde{\mathbf{x}} \in \mathrm{Fix}(T)$ we have, for all $k \in \mathbb{N}$, that*

$$\left\|\mathbf{x}^k - \tilde{\mathbf{x}}\right\| \leq C_{\tilde{\mathbf{x}}} := \max\left\{\left\|\mathbf{x}^0 - \tilde{\mathbf{x}}\right\|, \frac{1}{1-\beta}\left\|(I-S)\,\tilde{\mathbf{x}}\right\|\right\}. \tag{3.2}$$

*Moreover, for all $k \in \mathbb{N}$, we also have that*

$$\left\|T\left(\mathbf{x}^k\right) - \tilde{\mathbf{x}}\right\| \leq C_{\tilde{\mathbf{x}}} \quad \textit{and} \quad \left\|S\left(\mathbf{x}^k\right) - S\left(\tilde{\mathbf{x}}\right)\right\| \leq \beta C_{\tilde{\mathbf{x}}}.$$

(ii) *The sequence $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ converges to some $\mathbf{x}^* \in \mathrm{Fix}(T)$.*

(iii) *The limit point $\mathbf{x}^*$ of $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$, which the existence is ensured by (ii), satisfies the following variational inequality*

$$\left\langle \mathbf{x}^* - S\left(\mathbf{x}^*\right), \mathbf{x} - \mathbf{x}^* \right\rangle \geq 0, \quad \forall\ \mathbf{x} \in \mathrm{Fix}(T). \tag{3.3}$$

We conclude this part by highlighting and streamlining our contributions in this paper, which go beyond convergence of SAM as recorded in Lemma 2.

(i) We prove that under a specific choice of parameters, SAM generates a sequence $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ for which the sequence of the gaps between the iterator and its mapping by $T$, that is $\left\{\left\|T\left(\mathbf{x}^k\right) - \mathbf{x}^k\right\|\right\}_{k\in\mathbb{N}}$, converges with the non-asymptotic rate of $O(1/k)$. *This result gives a rate of convergence to a fixed point of $T$ for the first time.*

(ii) We study BiG-SAM for solving bi-level optimization problems, for which the functions $f$, $g$ and $\omega$ satisfy Assumptions A and B. For this variant, we prove an $O(1/k)$ rate of convergence for the sequence of inner objective function values (see details in Section 4). *This result affirmatively answers the question raised in [2] about a first order method for bi-level problems with an improved rate of convergence.*

(iii) We show that BiG-SAM can be also applied in situations where the outer objective function $\omega$ satisfies only Assumption B(i) but not B(ii), *i.e.*, it is strongly convex but not necessarily smooth. In this case we also prove a rate of convergence result in terms of the inner objective function values (see details in Section 4.1).

## 3.2 SAM for Smooth Bi-level Optimization Problem

We begin this part by connecting the fixed-point problem discussed above with the bi-level optimization problem described in Section 2.1. We will make this connection by linking the mappings $S$ and $T$ with problems (MNP) and (P), respectively. We begin by connecting the mapping $T$ with problem (P).

First of all, as explained in Section 3.1, the mapping $T$ and its fixed-point set $\mathrm{Fix}(T)$ are the inner part in the fixed-point problem, in the sense that we want to find a fixed-point of $T$ which is "better" than any other points in $\mathrm{Fix}(T)$ with respect to a criteria given by the mapping $S$ (see (3.1)). In the bi-level setting, the situation is similar, since the inner problem (P) is actually the inner part and from all its optimal solutions, *i.e.*, from the set $X^*$, we would like to find the one which satisfies the additional criteria, being minimizer of $\omega$ over $X^*$. Therefore, the following relations hold.

(i) The mapping $T$ and its fixed-point set $\mathrm{Fix}(T)$ are related to problem (P) with the composite function $\varphi = f + g$ and the optimal solution set $X^*$.

(ii) The mapping $S$ is related to problem (MNP) and the objective function $\omega$.

From now on, we set the mapping $T$ to be the prox-grad mapping defined in (2.3), that is, for some $t \in (0, 1/L_f]$ we have

$$T(\mathbf{x}) := T_t(\mathbf{x}) = \mathrm{prox}_{tg}(\mathbf{x} - t\nabla f(\mathbf{x})). \tag{3.4}$$

According to Lemma 1 and since Assumption A holds, we ensure that in this case $T$ is nonexpansive and $\mathrm{Fix}(T) = X^*$. We therefore fill all the requirements on the mapping $T$ in Lemma 2, and immediately obtain from Lemma 2(ii) that the sequence generated by SAM (with any $\beta$-contraction $S$) converges to a point in $X^*$. Thus, the only remaining part is to connect problem (MNP) with the criteria given in the variational inequality presented in Lemma 2(iii).

Taking into account Proposition 3 and given that Assumption B holds, a natural choice for the mapping $S$ is as follows

$$S(\mathbf{x}) := \mathbf{x} - s\nabla\omega(\mathbf{x}), \tag{3.5}$$

where $s \in (0, 2/(\sigma + L_\omega)]$. In this case we know, from Proposition 3, that $S$ is a $\beta$-contraction with $\beta = (1 - 2sL_\omega\sigma/(L_\omega + \sigma))^{1/2}$.

Therefore, SAM for solving the bi-level optimization problems (P) and (MNP) is given now.

**Bi-level Gradient SAM (BiG-SAM)**

(1) **Input:** $t \in (0, 1/L_f]$, $s \in (0, 2/(L_\omega + \sigma)]$, and $\{\alpha_k\}_{k\in\mathbb{N}}$ satisfying Assumption C.

(2) **Initialization**: $\mathbf{x}^0 \in \mathbb{R}^n$.

(3) **General Step** ($k = 1, 2, \ldots$):

$$\mathbf{y}^k = \mathrm{prox}_{tg}\left(\mathbf{x}^{k-1} - t\nabla f\left(\mathbf{x}^{k-1}\right)\right), \tag{3.6}$$
$$\mathbf{z}^k = \mathbf{x}^{k-1} - s\nabla\omega\left(\mathbf{x}^{k-1}\right), \tag{3.7}$$
$$\mathbf{x}^k = \alpha_k\mathbf{z}^k + (1 - \alpha_k)\mathbf{y}^k. \tag{3.8}$$

To conclude this section we would like to interpret the variational inequality given in Lemma 2(iii) in the setting of bi-level optimization, in which $S$ and $T$ are given by (3.5) and (3.4), respectively, for some $t \in (0, 1/L_f]$ and $s \in (0, 2/(L_\omega + \sigma)]$. In the following result we give the desired interpretation and prove that BiG-SAM generates a sequence which converges to the solution of problem (MNP).

**Proposition 5.** *Let* $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ *be a sequence generated by BiG-SAM and suppose that Assumptions A, B and C hold true. Then, the sequence* $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ *converges to* $\mathbf{x}^* \in X^*$ *which satisfies*

$$\left\langle \nabla\omega\left(\mathbf{x}^*\right), \mathbf{x}-\mathbf{x}^*\right\rangle \geq 0, \quad \forall\, \mathbf{x}\in X^*, \tag{3.9}$$

*and therefore,* $\mathbf{x}^* = \mathbf{x}^*_{mn}$ *is the optimal solution of problem* (MNP).

*Proof.* Since Assumptions A and B hold true, by Lemma 1 and Proposition 3, we have that $S$ and $T$ which defined in (3.5) and (3.4) are a contraction and a nonexpansive mapping, respectively. Thus, all the assumptions of Lemma 2 are valid and therefore we immediately obtain that $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ converges to $\mathbf{x}^* \in X^*$ (see Lemmas 1(ii) and 2(ii)). The only remaining part is showing that the variational inequality given in Lemma 2(iii) implies that (3.9) holds true. Indeed, using the fact that $S = I - s\nabla\omega$ we obtain that (3.3) is equivalent to

$$\left\langle \mathbf{x}^* - \left(\mathbf{x}^* - s\nabla\omega\left(\mathbf{x}^*\right)\right), \mathbf{x}-\mathbf{x}^*\right\rangle \geq 0, \quad \forall\, \mathbf{x}\in X^*,$$

which directly implies that (3.9) holds true, since $s > 0$. This means that $\mathbf{x}^*$ satisfies the first order optimality condition for constrained convex problems (see, for example, [6, Proposition 2.1.2, Page 194]) and therefore $\mathbf{x}^* = \mathbf{x}^*_{mn}$, as asserted. □

# 4 Rate of Convergence Analysis

In this section we will first prove a technical result about the rate of convergence of the gap between two successive iterations generated by SAM in its most generality, for solving the fixed-point problem, as described in Section 3.1. Then, we will use it to derive the main result of our paper which is a rate of convergence result for BiG-SAM in terms of the inner objective function values. This rate is superior to the one presented in [2] for the case of differentiable $\omega$, and holds true for any contraction mapping $S$ regardless of $\omega$.

We first present a technical lemma which will assist us in the rate of convergence proof. The proof of this lemma is given in Appendix C.

**Lemma 3.** *Let* $M > 0$. *Assume that* $\{a_k\}_{k\in\mathbb{N}}$ *is a sequence of nonnegative real numbers which satisfy* $a_1 \leq M$ *and*

$$a_{k+1} \leq \left(1-\gamma b_{k+1}\right) a_k + \left(b_k - b_{k+1}\right) c_k, \quad k \geq 1,$$

*where* $\gamma \in (0,1]$, $\{b_k\}_{k\in\mathbb{N}}$ *is a sequence which is defined by* $b_k = \min\left\{2/\left(\gamma k\right), 1\right\}$, *and* $\{c_k\}_{k\in\mathbb{N}}$ *is a sequence of real numbers such that* $c_k \leq M < \infty$. *Then, the sequence* $\{a_k\}_{k\in\mathbb{N}}$ *satisfies*

$$a_k \leq \frac{MJ}{\gamma k}, \quad k \geq 1,$$

*where* $J = \lfloor 2/\gamma \rfloor$.

For ease of notation, from this point onward, we will denote, for any $k \in \mathbb{N}$, $\mathbf{y}^k = T\left(\mathbf{x}^{k-1}\right)$ and $\mathbf{z}^k = S\left(\mathbf{x}^{k-1}\right)$. For convenience we will split the rate analysis into two technical results which will lead us to the main result given in Theorem 1. In Lemma 4 we present some useful inequalities, and in Lemma 5 we show that by choosing an appropriate sequence $\{\alpha_k\}_{k\in\mathbb{N}}$ we can bound the distance between two successive elements of the sequence $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ by $O(1/k)$ and to show that the sequence $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ converges, with the same rate, to a fixed-point of $T$.

**Lemma 4.** *Assume that $S : \mathbb{R}^n \to \mathbb{R}^n$ is a $\beta$-contraction and that $T : \mathbb{R}^n \to \mathbb{R}^n$ is nonexpansive mapping, for which $\mathrm{Fix}(T) \neq \emptyset$. Let $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$, $\left\{\mathbf{y}^k\right\}_{k\in\mathbb{N}}$ and $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$ be sequences generated by SAM. Then, for any $k \geq 1$ and any $\tilde{\mathbf{x}} \in \mathrm{Fix}(T)$, defining $\tilde{\mathbf{z}} = S\left(\tilde{\mathbf{x}}\right)$ the following inequalities hold true*

$$\left\|\mathbf{y}^{k+1} - \mathbf{y}^k\right\| \leq \left\|\mathbf{x}^k - \mathbf{x}^{k-1}\right\|, \tag{4.1}$$

$$\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\| \leq \beta\left\|\mathbf{x}^k - \mathbf{x}^{k-1}\right\|, \tag{4.2}$$

$$\left\|\mathbf{y}^k - \tilde{\mathbf{x}}\right\| \leq \left\|\mathbf{x}^{k-1} - \tilde{\mathbf{x}}\right\|, \tag{4.3}$$

$$\left\|\mathbf{z}^k - \tilde{\mathbf{z}}\right\| \leq \beta\left\|\mathbf{x}^{k-1} - \tilde{\mathbf{x}}\right\|. \tag{4.4}$$

*Proof.* All the required inequalities are a direct consequence of the nonexpansivity of $T$ and the contraction property of $S$. □

Now we prove the rate of convergence to zero of the sequence $\left\{\left\|\mathbf{x}^k - \mathbf{x}^{k-1}\right\|\right\}_{k\in\mathbb{N}}$, where $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ is generated by SAM and the averaging parameters $\alpha_k$, $k \in \mathbb{N}$, are chosen as follows

$$\alpha_k = \min\left\{\frac{2\gamma}{k\left(1-\beta\right)}, 1\right\}, \quad k \geq 1, \tag{4.5}$$

where $\gamma \in (0, 1]$. For the simplicity of the developments, we will prove our results when $\gamma = 1$. It should be noted that all the results below remains valid also when $\gamma$ is chosen arbitrarily from the interval $(0, 1]$.

**Lemma 5.** *Let $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$, $\left\{\mathbf{y}^k\right\}_{k\in\mathbb{N}}$ and $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$ be sequences generated by SAM where $\{\alpha_k\}_{k\in\mathbb{N}}$ is defined by (4.5). Then, for any $\tilde{\mathbf{x}} \in \mathrm{Fix}(T)$, the two sequences $\left\{\left\|\mathbf{x}^k - \mathbf{x}^{k-1}\right\|\right\}_{k\in\mathbb{N}}$ and $\left\{\left\|\mathbf{y}^k - \mathbf{x}^{k-1}\right\|\right\}_{k\in\mathbb{N}}$ converge to $\mathbf{0}$, and the rates of convergence are given by*

$$\left\|\mathbf{x}^k - \mathbf{x}^{k-1}\right\| \leq \frac{2C_{\tilde{\mathbf{x}}}J}{\left(1-\beta\right)k}, \quad k \geq 1, \tag{4.6}$$

*and*

$$\left\|\mathbf{y}^k - \mathbf{x}^{k-1}\right\| \leq \frac{2C_{\tilde{\mathbf{x}}}\left(J+2\right)}{\left(1-\beta\right)k}, \quad k \geq 1, \tag{4.7}$$

*where $C_{\tilde{\mathbf{x}}}$ is defined in (3.2), and $J = \lfloor 2/\left(1-\beta\right)\rfloor$.*

*Proof.* By the definition of $\mathbf{x}^k$ and $\mathbf{x}^{k+1}$ we immediately obtain

$$\begin{aligned}
\left\|\mathbf{x}^{k+1}-\mathbf{x}^k\right\| &= \left\|(1-\alpha_{k+1})\,\mathbf{y}^{k+1}+\alpha_{k+1}\mathbf{z}^{k+1}-\left((1-\alpha_k)\,\mathbf{y}^k+\alpha_k\mathbf{z}^k\right)\right\| \\
&= \left\|(1-\alpha_{k+1})\left(\mathbf{y}^{k+1}-\mathbf{y}^k\right)+\alpha_{k+1}\left(\mathbf{z}^{k+1}-\mathbf{z}^k\right)+(\alpha_k-\alpha_{k+1})\left(\mathbf{y}^k-\mathbf{z}^k\right)\right\| \\
&\leq (1-\alpha_{k+1})\left\|\mathbf{y}^{k+1}-\mathbf{y}^k\right\|+\alpha_{k+1}\left\|\mathbf{z}^{k+1}-\mathbf{z}^k\right\|+(\alpha_k-\alpha_{k+1})\left\|\mathbf{y}^k-\mathbf{z}^k\right\| \\
&\leq (1-\alpha_{k+1})\left\|\mathbf{x}^k-\mathbf{x}^{k-1}\right\|+\alpha_{k+1}\beta\left\|\mathbf{x}^k-\mathbf{x}^{k-1}\right\|+(\alpha_k-\alpha_{k+1})\left\|\mathbf{y}^k-\mathbf{z}^k\right\| \\
&= (1-\alpha_{k+1}\,(1-\beta))\left\|\mathbf{x}^k-\mathbf{x}^{k-1}\right\|+(\alpha_k-\alpha_{k+1})\left\|\mathbf{y}^k-\mathbf{z}^k\right\|, \qquad (4.8)
\end{aligned}$$

where the second inequality follows from (4.1) and (4.2). Now, let $\tilde{\mathbf{x}}\in\mathrm{Fix}(T)$ and let $\tilde{\mathbf{z}}=S\left(\tilde{\mathbf{x}}\right)$, then

$$\begin{aligned}
\left\|\mathbf{y}^k-\mathbf{z}^k\right\| &= \left\|\mathbf{y}^k-\tilde{\mathbf{x}}+\tilde{\mathbf{x}}-\tilde{\mathbf{z}}+\tilde{\mathbf{z}}-\mathbf{z}^k\right\| \\
&\leq \left\|\mathbf{y}^k-\tilde{\mathbf{x}}\right\|+\left\|\tilde{\mathbf{x}}-\tilde{\mathbf{z}}\right\|+\left\|\tilde{\mathbf{z}}-\mathbf{z}^k\right\| \\
&\leq \left\|\mathbf{x}^{k-1}-\tilde{\mathbf{x}}\right\|+\left\|(I-S)\,\tilde{\mathbf{x}}\right\|+\beta\left\|\mathbf{x}^{k-1}-\tilde{\mathbf{x}}\right\| \\
&\leq C_{\tilde{\mathbf{x}}}+(1-\beta)\,C_{\tilde{\mathbf{x}}}+\beta C_{\tilde{\mathbf{x}}}=2C_{\tilde{\mathbf{x}}}, \qquad (4.9)
\end{aligned}$$

where the second inequality follows from (4.3) and (4.4), as well as the definition of $\tilde{\mathbf{z}}$, and the last inequality follows from Lemma 2(i). Additionally, we have that

$$\left\|\mathbf{x}^1-\mathbf{x}^0\right\|=\left\|\mathbf{x}^1-\tilde{\mathbf{x}}+\tilde{\mathbf{x}}-\mathbf{x}^0\right\|\leq\left\|\mathbf{x}^1-\tilde{\mathbf{x}}\right\|+\left\|\mathbf{x}^0-\tilde{\mathbf{x}}\right\|\leq 2C_{\tilde{\mathbf{x}}}, \qquad (4.10)$$

where the second inequality follows from Lemma 2(i). The convergence rate for the sequence $\left\{\left\|\mathbf{x}^k-\mathbf{x}^{k-1}\right\|\right\}_{k\in\mathbb{N}}$ is now an immediate result of applying Lemma 3 on (4.8) with $a_k:=\left\|\mathbf{x}^k-\mathbf{x}^{k-1}\right\|$, $b_k:=\alpha_k$, $\gamma:=1-\beta$ and $c_k:=\left\|\mathbf{y}^k-\mathbf{z}^k\right\|$ and using (4.9) and (4.10) where $M:=2C_{\tilde{\mathbf{x}}}$. This proves (4.6).

The convergence rate for $\left\{\left\|\mathbf{y}^k-\mathbf{x}^{k-1}\right\|\right\}_{k\in\mathbb{N}}$ is derived by the following arguments

$$\begin{aligned}
\left\|\mathbf{y}^k-\mathbf{x}^{k-1}\right\| &= \left\|\mathbf{y}^k-\mathbf{x}^k+\mathbf{x}^k-\mathbf{x}^{k-1}\right\| \\
&\leq \left\|\mathbf{y}^k-\mathbf{x}^k\right\|+\left\|\mathbf{x}^k-\mathbf{x}^{k-1}\right\| \\
&= \alpha_k\left\|\mathbf{y}^k-\mathbf{z}^k\right\|+\left\|\mathbf{x}^k-\mathbf{x}^{k-1}\right\| \\
&\leq \frac{2}{(1-\beta)\,k}2C_{\tilde{\mathbf{x}}}+\frac{2C_{\tilde{\mathbf{x}}}J}{(1-\beta)\,k} \\
&= \frac{2C_{\tilde{\mathbf{x}}}\,(J+2)}{(1-\beta)\,k},
\end{aligned}$$

where the second inequality is due to the previous result as well as (4.9), which was already proven. □

It should be noted that in the case of BiG-SAM, the rate of convergence proven in (4.7) means that the sequence $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ converges to an optimal solution of the inner problem (P) with a rate of $O(1/k)$.

Now we turn to discuss the main result which is convergence of the sequence $\left\{\varphi\left(\mathbf{y}^k\right)\right\}_{k\in\mathbb{N}}$. We discuss the convergence of this sequence rather than the convergence of the sequence $\left\{\varphi\left(\mathbf{x}^k\right)\right\}_{k\in\mathbb{N}}$ since the latter might be an infeasible in terms of the domain of the function $\varphi$ (see also [2]). Moreover, since we proved that $\left\|\mathbf{y}^k-\mathbf{x}^{k-1}\right\|\to 0$ as $k\to\infty$ and $\varphi$ is lower semicontinuous it follows that proving convergence of the sequence $\left\{\varphi\left(\mathbf{y}^k\right)\right\}_{k\in\mathbb{N}}$ to the optimal value also implies convergence of the sequence $\left\{\varphi\left(\mathbf{x}^k\right)\right\}_{k\in\mathbb{N}}$ to the same value.

**Theorem 1.** *Let $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$, $\left\{\mathbf{y}^k\right\}_{k\in\mathbb{N}}$ and $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$ be sequences generated by BiG-SAM where $\{\alpha_k\}_{k\in\mathbb{N}}$ is defined by (4.5). Then, for all $t\leq 1/L_f$ and $k\in\mathbb{N}$, we have that*

$$\varphi\left(\mathbf{y}^k\right)-\varphi\left(\mathbf{x}^*_{mn}\right)\leq\frac{2C^2_{\mathbf{x}^*_{mn}}\left(J+2\right)}{\left(k+1\right)\left(1-\beta\right)t},$$

*where $C_{\mathbf{x}^*_{mn}}$ is defined in (3.2) and $J=\lfloor 2/\left(1-\beta\right)\rfloor$.*

*Proof.* From Proposition 1 we have, for any step-size $t\leq 1/L_f$, that the following inequality holds true

$$\varphi\left(\mathbf{y}^{k+1}\right)-\varphi\left(\mathbf{x}^*_{mn}\right)\leq\frac{1}{t}\left\langle\mathbf{x}^k-\mathbf{y}^{k+1},\mathbf{x}^k-\mathbf{x}^*_{mn}\right\rangle-\frac{1}{2t}\left\|\mathbf{x}^k-\mathbf{y}^{k+1}\right\|^2. \tag{4.11}$$

Applying Lemmas 2(i) and 5 for $\mathbf{x}^*_{mn}\in X^*=\mathrm{Fix}(T_t)$ we obtain that

$$\left\langle\mathbf{x}^k-\mathbf{y}^{k+1},\mathbf{x}^k-\mathbf{x}^*_{mn}\right\rangle\leq\left\|\mathbf{x}^k-\mathbf{y}^{k+1}\right\|\cdot\left\|\mathbf{x}^k-\mathbf{x}^*_{mn}\right\|\leq\frac{2C^2_{\mathbf{x}^*_{mn}}\left(J+2\right)}{\left(1-\beta\right)\left(k+1\right)}. \tag{4.12}$$

Substituting (4.11) back into (4.12) we obtain that

$$\varphi\left(\mathbf{y}^{k+1}\right)-\varphi\left(\mathbf{x}^*_{mn}\right)\leq\frac{2C^2_{\mathbf{x}^*_{mn}}\left(J+2\right)}{\left(k+1\right)\left(1-\beta\right)t},$$

which proves the desired result. □

**Remark 1.** The step-size $s$, which is used in step (3.7), should be chosen such that the mapping $S$ is a contraction. According to Proposition 3 the step-size $s$ depends on the knowledge of $L_\omega$ and $\sigma$, or at least on an upper bound on $L_\omega+\sigma$. Moreover, in order to calculate $\alpha_k$, $k\in\mathbb{N}$, we need an upper bound on the contraction parameter $\beta$ which also depends on $L_\omega$ and $\sigma$ or a lower bound on $\sigma L_\omega$.

## 4.1 SAM for Nonsmooth Bi-level Optimization Problems

In this section we focus on problem (MNP) as described in Section 2.1 for which the objective function $\omega$ does not necessarily satisfy Assumption B. Here we replace it by the following milder assumption.

**Assumption B'.** $\omega : \mathbb{R}^n \to \mathbb{R}$ is strongly convex with parameter $\sigma > 0$ and $\ell_\omega$-Lipschitz continuous.

It is clear that BiG-SAM can not be applied to bi-level problems for which $\omega$ satisfies Assumption B' instead of Assumption B, since $\omega$ is not necessarily differentiable. However, due to the strong convexity of $\omega$ we may use BiG-SAM in the following way.

We will use the Moreau envelope $M_{s\omega}$ of $\omega$ as a smooth replacement of the original objective function $\omega$. As we have already mentioned in Section 2.1, the Moreau envelope is continuously differentiable, its gradient is Lipschitz continuous with constant $1/s$ and strongly convex (see Proposition 2). Based on these facts we obtain that $M_{s\omega}$ satisfies Assumption B and therefore BiG-SAM can be applied in this case on the Moreau envelope $M_{s\omega}$. It should be noted that in this case step (3.7) is given by

$$\mathbf{z}^k = \mathbf{x}^{k-1} - s\nabla M_{s\omega}\left(\mathbf{x}^{k-1}\right) = \mathbf{x}^{k-1} - s\frac{1}{s}\left(\mathbf{x}^{k-1} - \operatorname{prox}_{s\omega}\left(\mathbf{x}^{k-1}\right)\right) = \operatorname{prox}_{s\omega}\left(\mathbf{x}^{k-1}\right), \quad (4.13)$$

where the second equality follows from (2.8). This means that in order to obtain $\mathbf{z}^k$, $k \in \mathbb{N}$, we need to compute the proximal mapping of $\omega$.

**Remark 2.** Based on the equality given in (4.13) it can be seen that BiG-SAM applied on the Moreau envelope $M_{s\omega}$ is exactly SAM which is applied to the bi-level problem with $S$ being the proximal mapping of $\omega$. In this respect it should be noted that the proximal mapping of a strongly convex function is a contraction (see Lemma 6 in Appendix A) and therefore all the theory presented in Section 3.1 is valid in this setting too.

The following result is an immediate consequence of Proposition 5 applies on the following mappings

$$S(\mathbf{x}) = \mathbf{x} - s\nabla M_{s\omega}(\mathbf{x}) \quad \text{and} \quad T(\mathbf{x}) = \operatorname{prox}_{tg}\left(\mathbf{x} - t\nabla f(\mathbf{x})\right),$$

where $s > 0$ and $t \in (0, 1/L_f]$.

**Proposition 6.** *Let $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ be a sequence generated by BiG-SAM and suppose that Assumptions A, B' and C hold true and let $s > 0$. Then, the sequence $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ converges to $\mathbf{x}_s^* \in X^*$ which satisfies*

$$\left\langle \nabla M_{s\omega}\left(\mathbf{x}_s^*\right), \mathbf{x} - \mathbf{x}_s^* \right\rangle \geq 0, \quad \forall\, \mathbf{x} \in X^*. \quad (4.14)$$

*Therefore, $\mathbf{x}_s^*$ is the optimal solution of problem (MNP) with respect to the Moreau envelope $M_{s\omega}$,* i.e.,

$$\mathbf{x}_s^* = \operatorname*{argmin}_{\mathbf{x}\in X^*} M_{s\omega}(\mathbf{x}),$$

*where $X^*$ is the optimal solutions set of problem (P).*

*Proof.* Similar to the proof of Proposition 5 using (4.13). □

The section began with the goal of solving problem (MNP) for which $\omega$ is not necessarily smooth. To this end we suggested to apply BiG-SAM on the Moreau envelope $M_{s\omega}$ for some step-size $s > 0$ and as a result we get $\mathbf{x}_s^*$ which minimizes $M_{s\omega}$ over $X^*$. The step-size $s$ also plays an important role in controlling the distance between $\mathbf{x}_s^*$ and the original solution $\mathbf{x}_{mn}^*$, this will be made precise below.

The fact that we smoothed the outer objective function $\omega$ seems to not influence the rate of convergence result which is in terms of the inner objective function. However, a careful inspection shows that this is not really the case since the rate of convergence result (see Theorem 1) depends on the contraction parameter $\beta$ which in this case depends on the smoothing parameter $s$. Indeed, from Lemma 6 (see Appendix A), we have that

$$\beta = \frac{1}{1 + s\sigma}.$$

Therefore, we suggest the following concept of rate of convergence result which is different than the classical one, but seems to be relevant when discussing algorithms for solving bi-level problems.

Let $\delta > 0$ be the required uniform accuracy in terms of the outer objective function, that is,

$$\omega\left(\mathbf{x}^k\right) - M_{s\omega}\left(\mathbf{x}^k\right) \leq \delta, \quad \forall\, k \in \mathbb{N}, \tag{4.15}$$

where it should be remembered that $\omega\left(\mathbf{x}^k\right) - M_{s\omega}\left(\mathbf{x}^k\right) \geq 0$ for all $k \in \mathbb{N}$. Now, we would like do determine the number of iterations $K$ that is needed to achieve $\varepsilon$-optimal solution of the inner problem, that is,

$$\varphi\left(\mathbf{x}^K\right) - \varphi\left(\mathbf{x}_{mn}^*\right) \leq \varepsilon,$$

while keeping the uniform accuracy as given in (4.15). This means that $K$ depends on both $\varepsilon$ and $\delta$.

**Proposition 7.** *Let $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ be a sequence generated by BiG-SAM and suppose that Assumptions A, B' and C hold true. In addition, suppose that the smoothing parameter is chosen by*

$$s = \frac{2\delta}{\ell_\omega^2}.$$

*Let $t \in (0, 1/L_f]$. Then, (4.15) holds true and for*

$$k \geq \frac{4C_{\mathbf{x}_{mn}^*}^2}{t\varepsilon}\left(2 + \frac{3\ell_\omega^2}{2\sigma\delta} + \frac{\ell_\omega^4}{4\sigma^2\delta^2}\right) - 1,$$

*it holds that $\varphi\left(\mathbf{x}^k\right) - \varphi\left(\mathbf{x}_{mn}^*\right) \leq \varepsilon$.*

*Proof.* Since $\omega$ is $\ell_\omega$-Lipschitz continuous (see Assumption B') it follows that the norms of the subgradients of $\omega$ are bounded from above by $\ell_\omega$. Thus, from [5, Lemma 4.2] it follows, for all $\mathbf{x} \in \mathbb{R}^n$, that

$$\omega\left(\mathbf{x}\right) - \frac{s\ell_\omega^2}{2} \leq M_{s\omega}\left(\mathbf{x}\right) \leq \omega\left(\mathbf{x}\right). \tag{4.16}$$

Therefore, for $s = 2\delta/\ell_\omega^2$, we obtain that

$$\omega\left(\mathbf{x}^k\right) - M_{s\omega}\left(\mathbf{x}^k\right) \leq \delta, \quad \forall\, k \in \mathbb{N}.$$

Using Theorem 1 we have that

$$\varphi\left(\mathbf{y}^{k+1}\right) - \varphi\left(\mathbf{x}_{mn}^*\right) \leq \frac{2C_{\mathbf{x}_{mn}^*}^2\,(J+2)}{(k+1)\,(1-\beta)\,t},$$

where $J = \lfloor 2/(1-\beta) \rfloor$. Substituting $\beta = 1/(1+s\sigma)$ in the above bound yields that

$$\begin{aligned}
\varphi\left(\mathbf{y}^{k+1}\right) - \varphi\left(\mathbf{x}_{mn}^*\right) &\leq \frac{2C_{\mathbf{x}_{mn}^*}^2\left(\frac{2}{1-\beta}+2\right)}{(k+1)\,(1-\beta)\,t} \\
&= \frac{4C_{\mathbf{x}_{mn}^*}^2}{(k+1)\,t} \cdot \frac{2-\beta}{(1-\beta)^2} \\
&= \frac{4C_{\mathbf{x}_{mn}^*}^2}{(k+1)\,t}\left(\frac{(1+s\sigma)^2}{(s\sigma)^2} + \frac{1+s\sigma}{s\sigma}\right) \\
&= \frac{4C_{\mathbf{x}_{mn}^*}^2}{(k+1)\,t}\left(2 + \frac{3}{s\sigma} + \frac{1}{(s\sigma)^2}\right).
\end{aligned}$$

Now, we use the smoothing parameter that we found above to obtain that

$$\varphi\left(\mathbf{y}^{k+1}\right) - \varphi\left(\mathbf{x}_{mn}^*\right) \leq \frac{4C_{\mathbf{x}_{mn}^*}^2}{(k+1)\,t}\left(2 + \frac{3\ell_\omega^2}{2\sigma\delta} + \frac{\ell_\omega^4}{4\sigma^2\delta^2}\right).$$

Thus, given $\varepsilon > 0$, in order to obtain $\varphi\left(\mathbf{y}^{k+1}\right) - \varphi\left(\mathbf{x}_{mn}^*\right) \leq \varepsilon$ it remains to find values of $k$ for which

$$\frac{4C_{\mathbf{x}_{mn}^*}^2}{(k+1)\,t}\left(2 + \frac{3\ell_\omega^2}{2\sigma\delta} + \frac{\ell_\omega^4}{4\sigma^2\delta^2}\right) \leq \varepsilon,$$

which is equivalent to

$$k \geq \frac{4C_{\mathbf{x}_{mn}^*}^2}{t\varepsilon}\left(2 + \frac{3\ell_\omega^2}{2\sigma\delta} + \frac{\ell_\omega^4}{4\sigma^2\delta^2}\right) - 1.$$

The desired result is obtained by choosing $k$ to be the upper bound just obtained. □

Two remarks on the just obtained result.

**Remark 3.** (i) The uniform accuracy property mentioned in (4.15) yields that the limit point $\mathbf{x}_s^*$ of the sequence generated by BiG-SAM satisfies that

$$\omega\left(\mathbf{x}_s^*\right) - \omega\left(\mathbf{x}_{mn}^*\right) \leq M_{s\omega}\left(\mathbf{x}_s^*\right) + \delta - M_{s\omega}\left(\mathbf{x}_{mn}^*\right) \leq \delta,$$

where the first inequality follows by using the two inequalities given in (4.15) while the second inequality follows from the fact that $\mathbf{x}_s^*$ is a minimizer of $M_{s\omega}$ over $X^*$ and obviously $\mathbf{x}_{mn}^* \in X^*$. This means that $\delta$ also controls the gap between the wished optimal value of $\omega$, that is, $\omega\left(\mathbf{x}_{mn}^*\right)$, and the value of $\omega$ evaluated at the optimal solution of the smoothed problem.

(ii) The number of iterations needed to achieve a desired inner function accuracy $\varepsilon$ is therefore $O(1/\varepsilon\delta^2)$. Consequently, increasing the uniform accuracy parameter $\delta$ by an order of magnitude results in increment of two orders of magnitude in the number of iterations. For example, by taking $\delta = \sqrt{\varepsilon}$ one will results with a rate of $O(1/\varepsilon^2)$ in terms of the inner objective function values.

# 5 Numerical Experiments

In this section we consider the inverse problems tested in [2, Section 5.2.2] and present a numerical comparison between the MNG and BiG-SAM methods. Linear inverse problems seeks to reconstruct a vector $\mathbf{x} \in \mathbb{R}^n$ from a set of measurements $\mathbf{b} \in \mathbb{R}^m$ which satisfy the following relation $\mathbf{b} = \mathbf{A}\mathbf{x} + \rho\epsilon$ where $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ is a given linear mapping, $\epsilon \in \mathbb{R}^m$ denotes an unknown noise vector and $\rho > 0$ denotes its magnitude.

There are several ways to solve linear inverse problems using optimization techniques, but here we will focus on the following bi-level formulation. In this case, the inner objective function is defined by

$$\varphi(\mathbf{x}) := \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \delta_X(\mathbf{x}),$$

where $\delta_X$ is the indicator function over the non-negative orthant $X = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq 0\}$. The outer objective function is given by

$$\omega(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x},$$

where $\mathbf{Q}$ is a certain positive definite matrix.

Following [2] we consider three inverse problems phillips, baart, and foxgood which can be found in the "regularization tools" website[2].

For each of these inverse problems we generated the corresponding $1,000 \times 1,000$ exact linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ by applying the relevant function ('philips','baart','foxgood'). We then performed 100 Monte-Carlo simulations by adding normally distributed noise with zero mean to the right-hand side vector $\mathbf{b}$, using three different choices of standard deviation: $\rho = 10^{-1}, 10^{-2}, 10^{-3}$. The matrix $\mathbf{Q}$ is defined by $\mathbf{Q} = \mathbf{L}\mathbf{L}' + \mathbf{I}$ where $\mathbf{L}$ is generated by the function `get_l(1,000,1)` from the "regularization tools" and approximates the first-derivative operator.

In order to implement the MNG method, we need to compute $\mathbf{Q}^{-1/2}$ and $\mathbf{Q}^{1/2}$ before the algorithm starts. However, note that while $\mathbf{Q}$ may be a sparse matrix, the matrices $\mathbf{Q}^{-1/2}$ and $\mathbf{Q}^{1/2}$ may not be, even if we use other decompositions, such as the Cholesky decomposition. Since the MNG method requires the starting point to be the optimal solution of the unconstrained minimization of $\omega$, we start both algorithms from the point $\mathbf{0}$.

[2] see http://www2.imm.dtu.dk/~pcha/Regutools/

We tested BiG-SAM with three different choices of the parameter $\gamma$ which are $0.1, 0.5$ and 1 (see the discussion before Lemma 5 about the parameter $\gamma$). All experiments were ran on a Unix server with 32 Intel Xeon CPUs E5-2690 @2.9GHz and 250GB RAM, using MATLAB R2016a, with no parallelization.

In Table 1 we present the mean time (out of 100 runs) until the algorithm (MNG and BiG-SAM) reach the stopping criteria $\left(\varphi\left(\mathbf{y}^k\right) - \varphi^*\right)/\varphi^* < 10^{-2}$, where $\varphi^*$ is the optimal value of the inner problem. If this stopping criteria was not achieved, then we stopped the algorithm after 500 seconds. In Table 2 we present the mean relative feasibility gap (RFG) given by $\Delta\varphi = \left(\varphi\left(\mathbf{y}^k\right) - \varphi^*\right)/\varphi^*$ and the mean relative optimality gap (ROG) given by $\Delta\omega = \left|\omega\left(\mathbf{y}^k\right) - \omega^*\right|/\omega^*$, for each algorithm after a running time of 250 seconds.

The values of $\varphi^*$ and $\omega^*$ were calculated in advance using CVX [9] for MATLAB and GUROBI version 7.0.1 solver, which we will refer to hereafter as the standard solver. The value of $\varphi^*$ was computed as the optimal solution of the inner problem. The value $\omega^*$ is a lower bound on the optimal value of the outer problem, obtained by solving the following convex problem

$$\min\left\{\omega\left(\mathbf{x}\right) : \mathbf{x} \geq 0,\, \varphi\left(\mathbf{x}\right) \leq \varphi^*(1+\mu)\right\},$$

where $\mu$ is a small number for which the problem was solvable (we used $10^{-4}$).

| Problem | $\rho$ | Mean time (Number of realization terminated at time limit) | | | |
|---|---|---|---|---|---|
| | | BiG-SAM | | | MNG |
| | | $\gamma = 0.1$ | $\gamma = 0.5$ | $\gamma = 1$ | |
| Baart | $10^{-1}$ | **5.37e−3** (0) | 3.62e−2 (0) | 6.08e−2 (0) | 2.92e−1 (0) |
| | $10^{-2}$ | **1.51e−1** (0) | 5.03e−1 (0) | 8.26e−1 (0) | 4.40 (0) |
| | $10^{-3}$ | **9.78** (0) | 2.23e+1 (0) | 3.57e+1 (0) | 4.18e+2 (31) |
| Foxgood | $10^{-1}$ | **1.51e−2** (0) | 6.88e−2 (0) | 1.06e−1 (0) | 3.33e−1 (0) |
| | $10^{-2}$ | **4.47e−1** (0) | 1.20 (0) | 2.17 (0) | 3.65 (0) |
| | $10^{-3}$ | **1.30e+1** (1) | 2.99e+1 (0) | 4.43e+1 (1) | 2.93e+1 (1) |
| Phillips | $10^{-1}$ | **1.13e−2** (0) | 3.90e−2 (0) | 6.58e−2 (0) | 4.02e−1 (0) |
| | $10^{-2}$ | **2.44** (0) | 6.77 (0) | 9.83 (0) | 1.67e+2 (5) |
| | $10^{-3}$ | **4.93e+2** (97) | 4.98e+2 (98) | 4.99e+2 (99) | 5.00e+2 (100) |

Table 1: Comparison between MNG and BiG-SAM (3 versions) of mean running times (in seconds) until termination and the number of realizations terminated because of the time limit (of 500 seconds) over 100 realization. The comparison is across the different problem instances and noise magnitude $\rho$.

The average number of iterations needed to reach these results for the MNG method was usually higher than that of BiG-SAM with $\gamma = 0.1$ but lower than that of BiG-SAM with $\gamma = 0.5$. However, as we can clearly see in the table above the higher iteration cost of

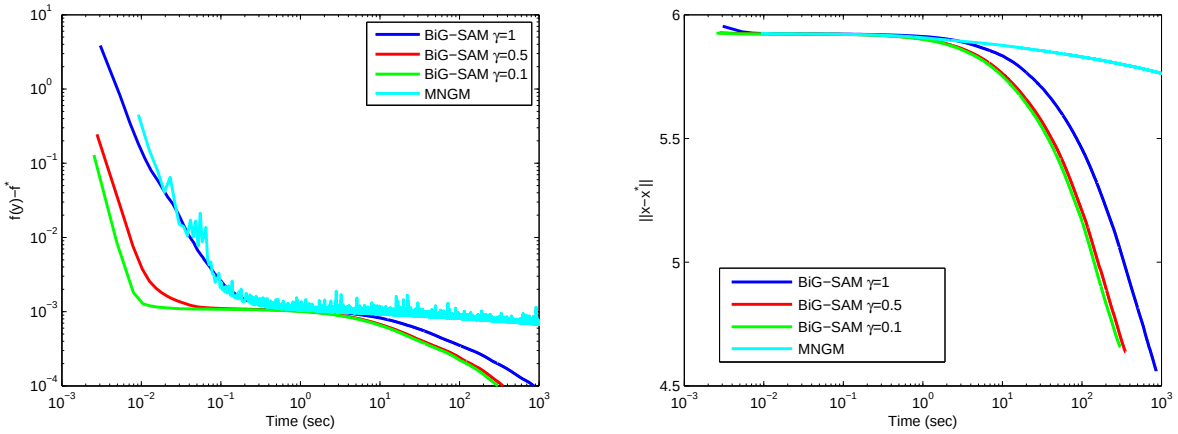
this method causes the mean time of the MNG method to be the highest, in most cases, and causes the algorithm to stop because of time limit rather than because it reached the termination criteria.

In Table 2 we see that when all methods are ran for the same amount of time, all BiG-SAM variants (except for BiG-SAM with $\gamma = 1$ for the Foxgood with $\rho = 0.001$) obtain superior $\Delta\varphi$ values compared to the MNG method, up to 2 orders of magnitude better. Moreover, in most cases all BiG-SAM variants obtain slightly better $\Delta\omega$ values compared to the MNG method, a fact which is more pronounced for higher values of the noise $\rho$.

| Problem | $\rho$ | Mean RFG ($\Delta\varphi$ in %) | | | | Mean ROG ($\Delta\omega$ in %) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | BiG-SAM | | | MNG | BiG-SAM | | | MNG |
| | | $\gamma = 0.1$ | $\gamma = 0.5$ | $\gamma = 1$ | | $\gamma = 0.1$ | $\gamma = 0.5$ | $\gamma = 1$ | |
| Baart | $10^{-1}$ | **6.93e-4** | 1.31e-3 | 1.80e-3 | 3.11e-2 | 26.12 | **23.87** | 24.88 | 58.08 |
| | $10^{-2}$ | **4.61e-2** | 5.42e-2 | 5.64e-2 | 1.37e-1 | **58.13** | 58.83 | 59.05 | 61.16 |
| | $10^{-3}$ | **1.475e-1** | 1.73e-1 | 2.05e-1 | 3.80 | **53.26** | 53.40 | 53.50 | 54.94 |
| Foxgood | $10^{-1}$ | **1.11e-2** | 1.52e-2 | 1.74e-2 | 5.19e-2 | **48.37** | 51.13 | 52.51 | 59.98 |
| | $10^{-2}$ | **3.01e-2** | 3.47e-2 | 3.82e-2 | 5.40e-2 | 15.42 | 15.29 | **15.25** | 15.28 |
| | $10^{-3}$ | **1.88e-2** | 2.46e-2 | 3.55e-2 | 2.56e-2 | 1.65 | 1.50 | **1.41** | 1.45 |
| Phillips | $10^{-1}$ | **3.84e-2** | 5.11e-2 | 6.00e-2 | 2.40e-1 | **78.95** | 81.73 | 83.14 | 90.03 |
| | $10^{-2}$ | **4.51e-1** | 4.98e-1 | 5.26e-1 | 7.44e-1 | **93.91** | 94.01 | 94.06 | 94.16 |
| | $10^{-3}$ | **1.75** | 1.82 | 1.87 | 2.19 | **90.16** | 90.16 | 90.16 | 90.17 |

Table 2: Comparison of relative feasible gap (RFG) and relative optimal gap (ROG) after 250 seconds for MNG and BiG-SAM with various parameters, averaged over 100 realization for each instance of problem and noise magnitude $\rho$.

In order to better understand this comparison we look at a specific realization of size 100 for a problem of Phillips type with $\rho = 0.01$. In Figure 1a we can see that BiG-SAM (with $\gamma = 1$) and MNG are very close in the first 10 seconds, but then BiG-SAM starts to improve much faster than the MNG method. Moreover, we can see clearly that lower value of $\gamma$ yields a faster convergence. In Figure 1b we see the distance between the iteration $\mathbf{x}^k$ and the optimal solution $\mathbf{x}^*$ (which was evaluated via the same procedure we used to find $\omega^*$). We can see the same behavior here as in the first figure, which means, a faster convergence of the BiG-SAM variants.




(a) Value of the inner objective function vs. time (b) Distance to optimal solution vs. time

Figure 1: The progress of the algorithms in time for a Phillips example with $\rho = 0.01$ and $n = 100$.

# Appendix A Proof of Proposition 2

We provide two proofs. One is short and based on well-known facts of convex analysis. The second proof is more complicated but provide useful properties of the mappings that play a central role in this work and will be useful in other contexts. We begin with first proof.

*Proof.* The proof is simple and based on the notion of infimal convolution and its properties. First of all, by its definition we have that the Moreau envelope is the infimal convolution of $\omega$ with the the squared norm function $h(\cdot) = (1/2s)\|\cdot\|^2$. Thus, for all $\mathbf{x} \in \mathbb{R}^n$, we have that $M_{s\omega}(\mathbf{x}) = (\omega \square h)(\mathbf{x})$. A well-known fact (see [1, Proposition 13.21(i), Page 187]) yields that $(\omega \square h)^* = \omega^* + h^*$, and therefore $M^*_{s\omega} = \omega^* + h^*$. Now, the rest of the proof follows from the following known fact: function $\vartheta : \mathbb{R}^n \to (-\infty, \infty]$ is strongly convex with strong convexity parameter $t$ if and only if its conjugate $\vartheta^*$ is a continuously differentiable function whose gradient is Lipschitz continuous with constant $1/t$. Using this fact twice yields that $M^*_{s\omega}$ has Lipschitz continuous gradient with constant $s + 1/\sigma$. Now, using the converse implication of the fact gives the desired result. □

Before we give the second proof, we will prove that the proximal mapping of a strongly convex function is a $\beta$-contraction, where $\beta = 1/(1 + s\sigma)$.

**Lemma 6.** *Let $\omega : \mathbb{R}^n \to (-\infty, +\infty]$ be a strongly convex function with parameter $\sigma$. Then, for any $s > 0$, it follows $\mathrm{prox}_{s\omega}$ is a $\beta$-contraction, where $\beta = 1/(1 + s\sigma)$, that is,*

$$\|\mathrm{prox}_{s\omega}(\mathbf{x}) - \mathrm{prox}_{s\omega}(\mathbf{y})\| \leq \frac{1}{1 + s\sigma}\|\mathbf{x} - \mathbf{y}\|, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

*Proof.* We first define an auxiliary function $\phi$ by $\phi(\mathbf{x}) = s\omega(\mathbf{x}) - (s\sigma/2)\|\mathbf{x}\|^2$. Since $\omega$ is strongly convex with parameter $\sigma$ it follows that $\varphi$ is convex. Hence, by the definition of the proximal mapping, we obtain

$$\begin{aligned}
\operatorname{prox}_{s\omega}(\mathbf{x}) &= \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^n}\left\{s\omega(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2\right\} \\
&= \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^n}\left\{\phi(\mathbf{u}) + \frac{s\sigma}{2}\|\mathbf{u}\|^2 + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2\right\} \\
&= \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^n}\left\{\phi(\mathbf{u}) + \frac{1+s\sigma}{2}\left\|\mathbf{u}-\frac{1}{1+s\sigma}\mathbf{x}\right\|^2\right\} \\
&= \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^n}\left\{\frac{1}{1+s\sigma}\phi(\mathbf{u}) + \frac{1}{2}\left\|\mathbf{u}-\frac{1}{1+s\sigma}\mathbf{x}\right\|^2\right\} \\
&= \operatorname{prox}_{\frac{1}{1+s\sigma}\phi}\left(\frac{\mathbf{x}}{1+s\sigma}\right). \qquad \text{(A.1)}
\end{aligned}$$

From (A.1) and the non-expensiveness of the proximal mapping (see [1, Proposition 12.27, Page 176]) we get

$$\begin{aligned}
\|\operatorname{prox}_{s\omega}(\mathbf{x}) - \operatorname{prox}_{s\omega}(\mathbf{y})\| &= \left\|\operatorname{prox}_{\frac{1}{1+s\sigma}\phi}\left(\frac{\mathbf{x}}{1+s\sigma}\right) - \operatorname{prox}_{\frac{1}{1+s\sigma}\phi}\left(\frac{\mathbf{y}}{1+s\sigma}\right)\right\| \\
&\leq \frac{1}{1+s\sigma}\|\mathbf{x}-\mathbf{y}\|.
\end{aligned}$$

This proves that the proximal mapping is $1/(1+s\sigma)$-contraction. □

Now we can provide the second proof of Proposition 2.

*Proof.* By (2.8) and using the Cauchy-Schwartz inequality we have that

$$\begin{aligned}
\langle \nabla M_{s\omega}(\mathbf{x}) - \nabla M_{s\omega}(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle &= \frac{1}{s}\langle \mathbf{x}-\mathbf{y} - (\operatorname{prox}_{s\omega}(\mathbf{x}) - \operatorname{prox}_{s\omega}(\mathbf{y})), \mathbf{x}-\mathbf{y}\rangle \\
&= \frac{1}{s}\|\mathbf{x}-\mathbf{y}\|^2 - \frac{1}{s}\langle \operatorname{prox}_{s\omega}(\mathbf{x}) - \operatorname{prox}_{s\omega}(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \\
&\geq \frac{1}{s}\|\mathbf{x}-\mathbf{y}\|^2 - \frac{1}{s}\|\mathbf{x}-\mathbf{y}\| \cdot \|\operatorname{prox}_{s\omega}(\mathbf{x}) - \operatorname{prox}_{s\omega}(\mathbf{y})\|.
\end{aligned}$$

By Lemma 6 we have that

$$\|\operatorname{prox}_{s\omega}(\mathbf{x}) - \operatorname{prox}_{s\omega}(\mathbf{y})\| \leq \frac{1}{1+s\sigma}\|\mathbf{x}-\mathbf{y}\|.$$

Thus combining the two inequalities we obtain that

$$\langle \nabla M_{s\omega}(\mathbf{x}) - \nabla M_{s\omega}(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \geq \frac{1}{s}\left(1 - \frac{1}{1+s\sigma}\right)\|\mathbf{x}-\mathbf{y}\|^2 = \frac{\sigma}{1+s\sigma}\|\mathbf{x}-\mathbf{y}\|^2.$$

Thus we conclude that $M_{s\omega}$ is strongly convex with parameter $\sigma/(1+s\sigma)$. □

# Appendix B Proof of Proposition 3

Denote $\tilde{\mathbf{x}} := \mathbf{x} - s\nabla\omega(\mathbf{x})$ and $\tilde{\mathbf{y}} := \mathbf{y} - s\nabla\omega(\mathbf{y})$. By the definition of $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ we have that

$$\begin{aligned}\|\tilde{\mathbf{x}} - \tilde{\mathbf{y}}\|^2 &= \|\mathbf{x} - s\nabla\omega(\mathbf{x}) - (\mathbf{y} - s\nabla\omega(\mathbf{y}))\|^2 \\ &= \|\mathbf{x} - \mathbf{y}\|^2 - 2s\langle\nabla\omega(\mathbf{x}) - \nabla\omega(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + s^2\|\nabla\omega(\mathbf{x}) - \nabla\omega(\mathbf{y})\|^2. \quad \text{(B.1)}\end{aligned}$$

Since $\omega$ is $\sigma$-strongly convex then by [10, Theorem 2.1.12, Page 66] we have that

$$\langle\nabla\omega(\mathbf{x}) - \nabla\omega(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \geq \frac{\sigma L_\omega}{\sigma + L_\omega}\|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{\sigma + L_\omega}\|\nabla\omega(\mathbf{x}) - \nabla\omega(\mathbf{y})\|^2. \quad \text{(B.2)}$$

By combining (B.1) and (B.2) we obtain that

$$\|\tilde{\mathbf{x}} - \tilde{\mathbf{y}}\|^2 \leq \left(1 - \frac{2s\sigma L_\omega}{\sigma + L_\omega}\right)\|\mathbf{x} - \mathbf{y}\|^2 + \left(s^2 - \frac{2s}{\sigma + L_\omega}\right)\|\nabla\omega(\mathbf{x}) - \nabla\omega(\mathbf{y})\|^2.$$

Therefore, for any $s \leq 2/(\sigma + L_\omega)$, we have that the second term is negative and so

$$\|\mathbf{x} - s\nabla\omega(\mathbf{x}) - (\mathbf{y} - s\nabla\omega(\mathbf{y}))\| \leq \sqrt{1 - \frac{2s\sigma L_\omega}{\sigma + L_\omega}}\|\mathbf{x} - \mathbf{y}\|,$$

which proves the desired result. □

# Appendix C Proof of Lemma 3

We split the proof into two cases: $k \leq J$ and $k > J$. We will start with proving the desired result for $k \leq J$.

**Case 1:**

Since $J \geq 2$ and $\gamma \leq 1$ it follows that $a_1 \leq M < 2M \leq MJ/\gamma$, and since $b_k = 1$ for any $k \leq J$ we have that

$$a_k = (1 - \gamma)a_{k-1}, \quad k = 2, 3, \ldots, J.$$

Now, using the fact that $\gamma k \leq J$, we obtain

$$a_k = (1 - \gamma)^{k-1}a_1 \leq a_1 \leq \frac{Ja_1}{\gamma k} \leq \frac{MJ}{\gamma k}, \quad k = 2, 3, \ldots, J.$$

This proves that the desired result hols true for all $k \leq J$.

**Case 2:**

We will assume that the claim is true for all $l = 1, 2, \ldots, k$ where $k \geq J$ and prove that it is true for $k + 1$. In this case, it is clear that $b_{k+1} = 2/(\gamma(k + 1))$ and $b_k \leq 2/(\gamma k)$.

Since $c_k \leq M$ and using the induction assumption we obtain

$$\begin{aligned}
a_{k+1} &\leq (1 - \gamma b_{k+1})\, a_k + (b_k - b_{k+1})\, c_k \\
&\leq \left(1 - \frac{2\gamma}{\gamma\,(k+1)}\right) \frac{JM}{\gamma k} + \left(b_k - \frac{2}{\gamma\,(k+1)}\right) M \\
&\leq \left(1 - \frac{2}{k+1}\right) \frac{JM}{\gamma k} + \left(\frac{2}{k} - \frac{2}{k+1}\right) \frac{M}{\gamma} \\
&= \frac{k-1}{k+1} \cdot \frac{JM}{\gamma k} + \frac{2M}{\gamma k\,(k+1)} \\
&\leq \frac{JM\,(k-1)}{\gamma k\,(k+1)} + \frac{JM}{\gamma k\,(k+1)} \\
&= \frac{JM}{\gamma\,(k+1)},
\end{aligned}$$

where the last inequality follows from the fact that $k > J \geq 2$. Thus the claim for $k > J$ is also proven. This completes the proof of the desired result. □

# References


[1] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, New York, 2011.

[2] A. Beck and S. Sabach. A first order method for finding minimal norm-like solutions of convex optimization problems. *Math. Program.*, 147(1-2, Ser. A):25–46, 2014.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[4] A. Beck and M. Teboulle. Gradient-based algorithms with applications to signal-recovery problems. In *Convex optimization in signal processing and communications*, pages 42–88. Cambridge Univ. Press, Cambridge, 2010.

[5] A. Beck and M. Teboulle. Smoothing and first order methods: a unified framework. *SIAM J. Optim.*, 22(2):557–580, 2012.

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

[7] Jr. R. E. Bruck. On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space. *J. Math. Anal. Appl.*, 61(1):159–164, 1977.

[8] M. C. Ferris and O. L. Mangasarian. Finite perturbation of convex programs. *Appl. Math. Optim.*, 23(3):263–273, 1991.

[9] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2014.

[10] Y. Nesterov. *Introductory Lectures on Convex Optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.

[11] E. S. H. Neto and Álvaro A. R. De Pierro. On perturbed steepest descent methods with inexact line search for bilevel convex optimization. *Optimization*, 60(8-9):991–1008, 2011.

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

[13] M. Solodov. An explicit descent method for bilevel convex optimization. *J. Convex Anal.*, 4(2):227–237, 2007.

[14] A. N. Tikhonov and V. Y. Arsenin. *Solutions of ill-posed problems*. V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York-Toronto, Ont.-London, 1977. Translated from the Russian, Preface by translation editor Fritz John, Scripta Series in Mathematics.

[15] H.-K. Xu. Viscosity approximation methods for nonexpansive mappings. *J. Math. Anal. Appl.*, 298(1):279–291, 2004.

[16] I. Yamada, M. Yukawa, and M. Yamagishi. Minimizing the Moreau envelope of nonsmooth convex functions over the fixed point set of certain quasi-nonexpansive mappings. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 345–390. Springer New York, 2011.