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

Shoham Sabach

Faculty of Industrial Engineering and Management

Technion

25.04.2018

Based on joint works with

Amir Beck (Tel Aviv) and Shimrit Shtern (Technion)

Optimization and Discrete Geometry: Theory and Practice

24-26 April, 2018, Tel Aviv University, Israel

Bi-Level Optimization Problems

Consider the following convex **inner problem**

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

Bi-Level Optimization Problems

Consider the following convex **inner problem**

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

where

- $f : \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable.
- ∇f is Lipschitz continuous with constant L_f.
- $g : \mathbb{R}^n \to (-\infty, +\infty]$ is proper, convex and lower semicontinuous.

Bi-Level Optimization Problems

Consider the following convex **inner problem**

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

where

- $f : \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable.
- ∇f is Lipschitz continuous with constant L_f.
- $g : \mathbb{R}^n \to (-\infty, +\infty]$ is proper, convex and lower semicontinuous.

We denote by X^* the **optimal solutions set**.

Bi-Level Optimization Problems

Consider the following convex **inner problem**

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

where

- $f : \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable.
- ∇f is Lipschitz continuous with constant L_f.
- $g : \mathbb{R}^n \to (-\infty, +\infty]$ is proper, convex and lower semicontinuous.

We denote by X^* the **optimal solutions set**.

In this talk we are interested in the following **outer problem**

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

Bi-Level Optimization Problems

Consider the following convex **inner problem**

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

where

- $f : \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable.
- ∇f is Lipschitz continuous with constant L_f.
- $g : \mathbb{R}^n \to (-\infty, +\infty]$ is proper, convex and lower semicontinuous.

We denote by X^* the **optimal solutions set**.

In this talk we are interested in the following **outer problem**

$$(\mathrm{MNP}) \qquad \begin{array}{ll} \min & \omega(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in X^*, \end{array}$$

where

- $\omega : \mathbb{R}^n \to \mathbb{R}$ is strongly convex with parameter σ.
- $\nabla \omega$ is Lipschitz continuous with constant L_ω.

Bi-Level Optimization Problems

Consider the following convex **inner problem**

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

where

- $f : \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable.
- ∇f is Lipschitz continuous with constant L_f.
- $g : \mathbb{R}^n \to (-\infty, +\infty]$ is proper, convex and lower semicontinuous.

We denote by X^* the **optimal solutions set**.

In this talk we are interested in the following **outer problem**

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

where

- $\omega : \mathbb{R}^n \to \mathbb{R}$ is strongly convex with parameter σ.
- $\nabla \omega$ is Lipschitz continuous with constant L_ω.

A particular case: the classical **minimal norm solution problem** $\min \left\{ \frac{1}{2} \|\mathbf{x}\|^2 : \mathbf{x} \in X^* \right\}$.

Tikhonov Regularization

Given $\varepsilon > 0$, consider the **regularized** convex problem

$$(\mathrm{Q}_{\varepsilon}) \quad \min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \varphi\left(\mathbf{x}\right) + \varepsilon \omega\left(\mathbf{x}\right) \right\}.$$

Tikhonov Regularization

Given $\varepsilon > 0$, consider the **regularized** convex problem

$$(\mathrm{Q}_\varepsilon) \quad \min_{\mathbf{x}\in\mathbb{R}^n} \left\{\varphi\left(\mathbf{x}\right) + \varepsilon\omega\left(\mathbf{x}\right)\right\}.$$

The **unique optimal solution** of (Q_ε) is denoted by $\mathbf{x}^\varepsilon$.

Tikhonov Regularization

Given $\varepsilon > 0$, consider the **regularized** convex problem

$$(\mathrm{Q}_\varepsilon) \quad \min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \varphi(\mathbf{x}) + \varepsilon \omega(\mathbf{x}) \right\}.$$

The **unique optimal solution** of (Q_ε) is denoted by $\mathbf{x}^\varepsilon$.

Let $\emptyset \neq X$ be closed and convex. We consider here the case $g(\cdot) = \delta_X(\cdot)$

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\{ f(\mathbf{x}) + \varepsilon \omega(\mathbf{x}) : \ \mathbf{x} \in X \right\}.$$

Tikhonov Regularization

Given $\varepsilon > 0$, consider the **regularized** convex problem

$$(\mathrm{Q}_\varepsilon) \quad \min_{\mathbf{x}\in\mathbb{R}^n} \left\{\varphi\left(\mathbf{x}\right) + \varepsilon\omega\left(\mathbf{x}\right)\right\}.$$

The **unique optimal solution** of (Q_ε) is denoted by $\mathbf{x}^\varepsilon$.

Let $\emptyset \neq X$ be closed and convex. We consider here the case $g\left(\cdot\right) = \delta_X\left(\cdot\right)$

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{f\left(\mathbf{x}\right) + \varepsilon\omega\left(\mathbf{x}\right) : \ \mathbf{x} \in X\right\}.$$

For $\omega\left(\mathbf{x}\right) = (1/2)\left\|\mathbf{x}\right\|^2$ we have the following results:

Tikhonov Regularization

Given $\varepsilon > 0$, consider the **regularized** convex problem

$$(\mathrm{Q}_\varepsilon) \qquad \min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \varphi\left(\mathbf{x}\right) + \varepsilon \omega\left(\mathbf{x}\right) \right\}.$$

The **unique optimal solution** of (Q_ε) is denoted by $\mathbf{x}^\varepsilon$.

Let $\emptyset \neq X$ be closed and convex. We consider here the case $g\left(\cdot\right) = \delta_X\left(\cdot\right)$

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\{ f\left(\mathbf{x}\right) + \varepsilon \omega\left(\mathbf{x}\right) : \, \mathbf{x} \in X \right\}.$$

For $\omega\left(\mathbf{x}\right) = (1/2) \left\|\mathbf{x}\right\|^2$ we have the following results:

- Tikhonov (1977) showed, in the **linear case**, that $\mathbf{x}^\varepsilon \rightarrow \mathbf{x}^*_{\mathrm{mn}}$ as $\varepsilon \rightarrow 0^+$.

Tikhonov Regularization

Given $\varepsilon > 0$, consider the **regularized** convex problem

$$(\mathrm{Q}_\varepsilon) \quad \min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \varphi\left(\mathbf{x}\right) + \varepsilon\omega\left(\mathbf{x}\right) \right\}.$$

The **unique optimal solution** of (Q_ε) is denoted by $\mathbf{x}^\varepsilon$.

Let $\emptyset \neq X$ be closed and convex. We consider here the case $g\left(\cdot\right) = \delta_X\left(\cdot\right)$

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\{ f\left(\mathbf{x}\right) + \varepsilon\omega\left(\mathbf{x}\right) : \ \mathbf{x} \in X \right\}.$$

For $\omega\left(\mathbf{x}\right) = (1/2)\left\|\mathbf{x}\right\|^2$ we have the following results:

- Tikhonov (1977) showed, in the **linear case**, that $\mathbf{x}^\varepsilon \to \mathbf{x}^*_{\mathrm{mn}}$ as $\varepsilon \to 0^+$.
- Mangasarian and Meyer (1979) showed, in the **linear case**, that for a small enough ε, $\mathbf{x}^\varepsilon$ is **exactly the same** as $\mathbf{x}^*_{\mathrm{mn}}$.

Tikhonov Regularization

Given $\varepsilon > 0$, consider the **regularized** convex problem

$$(\mathrm{Q}_{\varepsilon}) \quad \min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \varphi\left(\mathbf{x}\right) + \varepsilon \omega\left(\mathbf{x}\right) \right\}.$$

The **unique optimal solution** of $(\mathrm{Q}_{\varepsilon})$ is denoted by $\mathbf{x}^{\varepsilon}$.

Let $\emptyset \neq X$ be closed and convex. We consider here the case $g\left(\cdot\right) = \delta_X\left(\cdot\right)$

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\{ f\left(\mathbf{x}\right) + \varepsilon \omega\left(\mathbf{x}\right) : \; \mathbf{x} \in X \right\}.$$

For $\omega\left(\mathbf{x}\right) = (1/2)\left\|\mathbf{x}\right\|^2$ we have the following results:

- Tikhonov (1977) showed, in the **linear case**, that $\mathbf{x}^{\varepsilon} \to \mathbf{x}^*_{\mathrm{mn}}$ as $\varepsilon \to 0^+$.
- Mangasarian and Meyer (1979) showed, in the **linear case**, that for a small enough ε, $\mathbf{x}^{\varepsilon}$ is **exactly the same** as $\mathbf{x}^*_{\mathrm{mn}}$.
- Ferris and Mangasarian (1991) showed the same in a general **convex case**.

Tikhonov Regularization

Given $\varepsilon > 0$, consider the **regularized** convex problem

$$(\mathrm{Q}_\varepsilon) \qquad \min_{\mathbf{x} \in \mathbb{R}^n} \{\varphi(\mathbf{x}) + \varepsilon\omega(\mathbf{x})\}.$$

The **unique optimal solution** of (Q_ε) is denoted by $\mathbf{x}^\varepsilon$.

Let $\emptyset \neq X$ be closed and convex. We consider here the case $g(\cdot) = \delta_X(\cdot)$

$$\min_{\mathbf{x} \in \mathbb{R}^n} \{f(\mathbf{x}) + \varepsilon\omega(\mathbf{x}) : \mathbf{x} \in X\}.$$

For $\omega(\mathbf{x}) = (1/2)\|\mathbf{x}\|^2$ we have the following results:

- Tikhonov (1977) showed, in the **linear case**, that $\mathbf{x}^\varepsilon \to \mathbf{x}^*_{\mathrm{mn}}$ as $\varepsilon \to 0^+$.
- Mangasarian and Meyer (1979) showed, in the **linear case**, that for a small enough ε, $\mathbf{x}^\varepsilon$ is **exactly the same** as $\mathbf{x}^*_{\mathrm{mn}}$.
- Ferris and Mangasarian (1991) showed the same in a general **convex case**.

Solodov (2007) showed that the **projected gradient** when applied on $(\mathrm{Q}_{\varepsilon_k})$ with $\varepsilon_k \to 0$ and $\sum_{k=1}^\infty \varepsilon_k = \infty$, would generates a sequence which converges to $\mathbf{x}^*_{\mathrm{mn}}$.

Direct Algorithms

There are few more methods **BUT without** proven convergence rates

Direct Algorithms

There are few more methods **BUT without** proven convergence rates

Recently Beck-S. (2014) proposed the **Minimal Norm Gradient (MNG) method** for solving the (MNP) problem, when $g(\cdot) = \delta_X(\cdot)$.

Direct Algorithms

There are few more methods **BUT** **without** proven convergence rates

Recently Beck-S. (2014) proposed the **Minimal Norm Gradient (MNG) method** for solving the (MNP) problem, when $g(\cdot) = \delta_X(\cdot)$.

Input: L - a Lipschitz constant of ∇f.
Initialization: $\mathbf{x}^0 = \mathbf{a}$.
General Step (k = 1 , 2 , . . .):

$$\mathbf{x}^k = argmin\left\{\omega(\mathbf{x}) : \mathbf{x} \in Q^k \cap W^k\right\},$$

where

$$Q^k = \left\{\mathbf{z} \in \mathbb{R}^n : \left\langle G_L\left(\mathbf{x}^{k-1}\right), \mathbf{x}^{k-1} - \mathbf{z}\right\rangle \geq \frac{3}{4L}\left\|G_L\left(\mathbf{x}^{k-1}\right)\right\|^2\right\},$$

$$W^k = \left\{\mathbf{z} \in \mathbb{R}^n : \left\langle \nabla\omega\left(\mathbf{x}^{k-1}\right), \mathbf{z} - \mathbf{x}^{k-1}\right\rangle \geq 0\right\},$$

Direct Algorithms

There are few more methods **BUT**
without proven convergence rates

Recently Beck-S. (2014) proposed the **Minimal Norm Gradient (MNG) method** for solving the (MNP) problem, when $g(\cdot) = \delta_X(\cdot)$.

Input: L - a Lipschitz constant of ∇f.
Initialization: $\mathbf{x}^0 = \mathbf{a}$.
General Step (k = 1 , 2 , . . .):

$$\mathbf{x}^k = argmin\left\{\omega(\mathbf{x}) : \mathbf{x} \in Q^k \cap W^k\right\},$$

where

$$Q^k = \left\{\mathbf{z} \in \mathbb{R}^n : \left\langle G_L\left(\mathbf{x}^{k-1}\right), \mathbf{x}^{k-1} - \mathbf{z}\right\rangle \geq \frac{3}{4L}\left\|G_L\left(\mathbf{x}^{k-1}\right)\right\|^2\right\},$$

$$W^k = \left\{\mathbf{z} \in \mathbb{R}^n : \left\langle \nabla\omega\left(\mathbf{x}^{k-1}\right), \mathbf{z} - \mathbf{x}^{k-1}\right\rangle \geq 0\right\},$$

The **gradient mapping** is defined by $G_L(\mathbf{x}) \equiv L\left[\mathbf{x} - P_X\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right)\right]$.

The Minimal Norm Gradient Method

Each iteration of the MNG method **consists of 3 main computational tasks**:

(i) Computing the **gradient** of f and the **projection** onto the set X.

The Minimal Norm Gradient Method

Each iteration of the MNG method **consists of** 3 **main computational tasks**:

(i) Computing the **gradient** of f and the **projection** onto the set X.

(ii) Computing the **gradient** of ω.

The Minimal Norm Gradient Method

Each iteration of the MNG method **consists of 3 main computational tasks**:

(i) Computing the **gradient** of f and the **projection** onto the set X.

(ii) Computing the **gradient** of ω.

(iii) **Minimizing** ω over the intersection of two (given) half spaces.

The Minimal Norm Gradient Method

Each iteration of the MNG method **consists of 3 main computational tasks**:

(i) Computing the **gradient** of f and the **projection** onto the set X.

(ii) Computing the **gradient** of ω.

(iii) **Minimizing** ω over the intersection of two (given) half spaces.

Proposition (Beck-S. (2014))

Let $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ be the sequence generated by the MNG method. Then, the sequence $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ converges to the optimal solution $\mathbf{x}_{mn}^*$ and, for any $k \in \mathbb{N}$, we have that

$$\min_{1\leq m\leq k} \varphi\left(T_{1/L_f}\left(\mathbf{x}^m\right)\right) - \varphi\left(\mathbf{x}_{mn}^*\right) \leq \frac{4L_f\left\|\mathbf{x}^0 - \mathbf{x}_{mn}^*\right\|^2}{3\sqrt{k}},$$

where $T_t\left(\mathbf{x}\right) := P_X\left(\mathbf{x} - t\nabla f\left(\mathbf{x}\right)\right)$ is the proj-grad mapping

The Minimal Norm Gradient Method

Each iteration of the MNG method **consists of 3 main computational tasks**:

(i) Computing the **gradient** of f and the **projection** onto the set X.

(ii) Computing the **gradient** of ω.

(iii) **Minimizing** ω over the intersection of two (given) half spaces.

Proposition (Beck-S. (2014))

Let $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ be the sequence generated by the MNG method. Then, the sequence $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ converges to the optimal solution $\mathbf{x}^*_{mn}$ and, for any $k \in \mathbb{N}$, we have that

$$\min_{1\leq m\leq k} \varphi\left(T_{1/L_f}\left(\mathbf{x}^m\right)\right) - \varphi\left(\mathbf{x}^*_{mn}\right) \leq \frac{4L_f\left\|\mathbf{x}^0 - \mathbf{x}^*_{mn}\right\|^2}{3\sqrt{k}},$$

where $T_t\left(\mathbf{x}\right) := P_X\left(\mathbf{x} - t\nabla f\left(\mathbf{x}\right)\right)$ is the proj-grad mapping

Note: In the case that the Lipschitz constant L is **unknown in advance**, a backtracking scheme can be incorporated **(rate remains the same).**

Goal and Outline

Study a new method for solving the (MNP) problem with better rate of convergence and lower computational cost

Goal and Outline

Study a new method for solving the (MNP) problem with better rate of convergence and lower computational cost

Outline

- The Sequential Averaging Method (SAM).

Goal and Outline

Study a new method for solving the (MNP) problem with better rate of convergence and lower computational cost

Outline

- The Sequential Averaging Method (SAM).
- The Bi-Level Gradient Sequential Averaging Method (BiG-SAM).

Goal and Outline

Study a new method for solving the (MNP) problem with better rate of convergence and lower computational cost

Outline

- The Sequential Averaging Method (SAM).
- The Bi-Level Gradient Sequential Averaging Method (BiG-SAM).
- Convergence analysis of BiG-SAM.

Goal and Outline

Study a new method for solving the (MNP) problem with better rate of convergence and lower computational cost

Outline

- The Sequential Averaging Method (SAM).
- The Bi-Level Gradient Sequential Averaging Method (BiG-SAM).
- Convergence analysis of BiG-SAM.
- BiG-SAM for nonsmooth ω.

Joint work with Shimrit Shtern (Technion)

Sequential Averaging Method (SAM)

Suppose we are given two mappings:

- A **nonexpansive** mapping T: $\|T(\mathbf{x}) - T(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Sequential Averaging Method (SAM)

Suppose we are given two mappings:

- A **nonexpansive** mapping T: $\|T(\mathbf{x}) - T(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
- A β**-contraction** mapping S ($\beta < 1$): $\|S(\mathbf{x}) - S(\mathbf{y})\| \leq \beta \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Sequential Averaging Method (SAM)

Suppose we are given two mappings:

- A **nonexpansive** mapping T: $\|T(\mathbf{x}) - T(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
- A β**-contraction** mapping S ($\beta < 1$): $\|S(\mathbf{x}) - S(\mathbf{y})\| \leq \beta \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Goal: find $\mathbf{x}^* \in \mathrm{Fix}(T) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} = T(\mathbf{x})\}$, which satisfies

$$\langle \mathbf{x}^* - S(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \quad \forall \mathbf{x} \in \mathrm{Fix}(T). \tag{1}$$

Sequential Averaging Method (SAM)

Suppose we are given two mappings:

- A **nonexpansive** mapping T: $\|T(\mathbf{x}) - T(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
- A β**-contraction** mapping S ($\beta < 1$): $\|S(\mathbf{x}) - S(\mathbf{y})\| \leq \beta \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Goal: find $\mathbf{x}^* \in \mathrm{Fix}(T) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} = T(\mathbf{x})\}$, which satisfies

$$\langle \mathbf{x}^* - S(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \quad \forall\ \mathbf{x} \in \mathrm{Fix}(T). \tag{1}$$

Algorithm: $\mathbf{x}^{k+1} = \alpha_{k+1} S\left(\mathbf{x}^k\right) + (1 - \alpha_{k+1}) T\left(\mathbf{x}^k\right)$.

Sequential Averaging Method (SAM)

Suppose we are given two mappings:

- A **nonexpansive** mapping T: $\|T(\mathbf{x}) - T(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
- A β**-contraction** mapping S ($\beta < 1$): $\|S(\mathbf{x}) - S(\mathbf{y})\| \leq \beta \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Goal: find $\mathbf{x}^* \in \mathrm{Fix}(T) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} = T(\mathbf{x})\}$, which satisfies

$$\langle \mathbf{x}^* - S(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \quad \forall\, \mathbf{x} \in \mathrm{Fix}(T). \tag{1}$$

Algorithm: $\mathbf{x}^{k+1} = \alpha_{k+1} S\left(\mathbf{x}^k\right) + (1 - \alpha_{k+1}) T\left(\mathbf{x}^k\right)$.

We say that $\{\alpha_k\}_{k \in \mathbb{N}}$ is **“well-chosen" sequence** of real numbers from $(0, 1]$ if

$$\lim_{k \to \infty} \alpha_k = 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{and} \quad \lim_{k \to \infty} \alpha_{k+1}/\alpha_k = 1.$$

Sequential Averaging Method (SAM)

Suppose we are given two mappings:

- A **nonexpansive** mapping T: $\|T(\mathbf{x}) - T(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
- A **β-contraction** mapping S ($\beta < 1$): $\|S(\mathbf{x}) - S(\mathbf{y})\| \leq \beta \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Goal: find $\mathbf{x}^* \in \mathrm{Fix}(T) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} = T(\mathbf{x})\}$, which satisfies

$$\langle \mathbf{x}^* - S(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \quad \forall\, \mathbf{x} \in \mathrm{Fix}(T). \tag{1}$$

Algorithm: $\mathbf{x}^{k+1} = \alpha_{k+1} S(\mathbf{x}^k) + (1 - \alpha_{k+1}) T(\mathbf{x}^k)$.

We say that $\{\alpha_k\}_{k\in\mathbb{N}}$ is **"well-chosen" sequence** of real numbers from $(0, 1]$ if

$$\lim_{k\to\infty} \alpha_k = 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{and} \quad \lim_{k\to\infty} \alpha_{k+1}/\alpha_k = 1.$$

Theorem (Xu (2004))

Given a "well-chosen" sequence $\{\alpha_k\}_{k\in\mathbb{N}}$. Then

- *The sequence $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$ is* **bounded**.
- *The sequence $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$* **converges** *to a point $\mathbf{x}^* \in \mathrm{Fix}(T)$.*
- *The limit point $\mathbf{x}^*$* **satisfies** *(1).*

Bi-Level Gradient Sequential Averaging Method (BiG-SAM)

$$\langle \mathbf{x}^* - S(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \quad \forall\, \mathbf{x} \in \mathrm{Fix}(T). \tag{1}$$

- We will **connect problem** (1) **to an optimality condition of problem (MNP).**
- Meaning of (1): $\mathbf{x}^* \in \mathrm{Fix}(T)$ is better (w.r.t criterion (1)) than any other $\mathbf{x} \in \mathrm{Fix}(T)$.

Bi-Level Gradient Sequential Averaging Method (BiG-SAM)

$$\langle \mathbf{x}^* - S(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \quad \forall\, \mathbf{x} \in \mathbf{X}^*. \tag{1}$$

- We will **connect problem** (1) **to an optimality condition of problem (MNP).**
- Meaning of (1): $\mathbf{x}^* \in \mathrm{Fix}(T)$ is better (w.r.t criterion (1)) than any other $\mathbf{x} \in \mathrm{Fix}(T)$.
- Choosing T such that $\mathrm{Fix}(\mathbf{T}) \Leftrightarrow \mathrm{argmin}_{\mathbf{x}}\varphi(\mathbf{x}) = \mathbf{X}^*$.

Bi-Level Gradient Sequential Averaging Method (BiG-SAM)

$$\langle \mathbf{x}^* - S(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \quad \forall\, \mathbf{x} \in \mathbf{X}^*. \tag{1}$$

- We will **connect problem** (1) **to an optimality condition of problem (MNP).**
- Meaning of (1): $\mathbf{x}^* \in \mathrm{Fix}(T)$ is better (w.r.t criterion (1)) than any other $\mathbf{x} \in \mathrm{Fix}(T)$.
- Choosing T such that $\mathrm{Fix}(\mathbf{T}) \Leftrightarrow \mathrm{argmin}_{\mathbf{x}} \varphi(\mathbf{x}) = \mathbf{X}^*$.
- This holds true for the prox-grad mapping

$$T(\mathbf{x}) \equiv T_t(\mathbf{x}) = \mathrm{prox}_{tg}(\mathbf{x} - t\nabla f(\mathbf{x})),$$

which is nonexpansive for any $t \in (0, 1/L_f]$.

Bi-Level Gradient Sequential Averaging Method (BiG-SAM)

$$\langle \nabla \omega (\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \quad \forall\, \mathbf{x} \in \mathbf{X}^*. \tag{1}$$

- We will **connect problem** (1) **to an optimality condition of problem (MNP).**
- Meaning of (1): $\mathbf{x}^* \in \mathrm{Fix}(T)$ is better (w.r.t criterion (1)) than any other $\mathbf{x} \in \mathrm{Fix}(T)$.
- Choosing T such that $\mathrm{Fix}(\mathbf{T}) \Leftrightarrow \mathrm{argmin}_{\mathbf{x}} \varphi(\mathbf{x}) = \mathbf{X}^*$.
- This holds true for the prox-grad mapping

$$T(\mathbf{x}) \equiv T_t(\mathbf{x}) = \mathrm{prox}_{tg}(\mathbf{x} - t\nabla f(\mathbf{x})),$$

  which is nonexpansive for any $t \in (0, 1/L_f]$.
- To complete the connection, we will chose $S(\cdot)$ as

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

  which is a contraction with parameter $\beta = \left(1 - \frac{2sL_\omega \sigma}{L_\omega + \sigma}\right)^{1/2}$, whenever $s \in (0, 2/(\sigma + L_\omega)]$.

Bi-Level Gradient Sequential Averaging Method (BiG-SAM)

$$\langle \nabla \omega (\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \quad \forall \, \mathbf{x} \in \mathbf{X}^*. \tag{1}$$

- We will **connect problem** (1) **to an optimality condition of problem (MNP).**
- Meaning of (1): $\mathbf{x}^* \in \mathrm{Fix}(T)$ is better (w.r.t criterion (1)) than any other $\mathbf{x} \in \mathrm{Fix}(T)$.
- Choosing T such that $\mathrm{Fix}(\mathbf{T}) \Leftrightarrow \mathrm{argmin}_{\mathbf{x}} \varphi(\mathbf{x}) = \mathbf{X}^*$.
- This holds true for the prox-grad mapping

  $$T(\mathbf{x}) \equiv T_t(\mathbf{x}) = \mathrm{prox}_{tg}(\mathbf{x} - t\nabla f(\mathbf{x})),$$

  which is nonexpansive for any $t \in (0, 1/L_f]$.
- To complete the connection, we will chose $S(\cdot)$ as

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

  which is a contraction with parameter $\beta = \left(1 - \frac{2sL_\omega \sigma}{L_\omega + \sigma}\right)^{1/2}$, whenever $s \in (0, 2/(\sigma + L_\omega)]$.
- Thus, (1) **reduces to an optimality condition of problem (MNP).**

Bi-Level Gradient Sequential Averaging Method (BiG-SAM)

We take

$$T(\mathbf{x}) \equiv \mathrm{prox}_{tg}(\mathbf{x} - t\nabla f(\mathbf{x})) \quad \text{and} \quad S(\mathbf{x}) = \mathbf{x} - s\nabla\omega(\mathbf{x}).$$

Bi-Level Gradient Sequential Averaging Method (BiG-SAM)

We take

$$T(\mathbf{x}) \equiv \mathrm{prox}_{tg}(\mathbf{x} - t\nabla f(\mathbf{x})) \quad \text{and} \quad S(\mathbf{x}) = \mathbf{x} - s\nabla\omega(\mathbf{x}).$$

(i) **Input:** $t \in (0, 1/L_f]$ and $s \in (0, 2/(L_\omega + \sigma)]$.

(ii) **Initialization**: Start with any $\mathbf{x}^0 \in \mathbb{R}^n$.

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

$$\begin{aligned}
\mathbf{y}^k &= \mathrm{prox}_{tg}\left(\mathbf{x}^{k-1} - t\nabla f\left(\mathbf{x}^{k-1}\right)\right), \\
\mathbf{z}^k &= \mathbf{x}^{k-1} - s\nabla\omega\left(\mathbf{x}^{k-1}\right), \\
\mathbf{x}^k &= \alpha_{k+1}\mathbf{z}^k + (1 - \alpha_{k+1})\mathbf{y}^k.
\end{aligned}$$

Bi-Level Gradient Sequential Averaging Method (BiG-SAM)

We take

$$T(\mathbf{x}) \equiv \mathrm{prox}_{tg}(\mathbf{x} - t\nabla f(\mathbf{x})) \quad \text{and} \quad S(\mathbf{x}) = \mathbf{x} - s\nabla\omega(\mathbf{x}).$$

(i) **Input:** $t \in (0, 1/L_f]$ and $s \in (0, 2/(L_\omega + \sigma)]$.

(ii) **Initialization**: Start with any $\mathbf{x}^0 \in \mathbb{R}^n$.

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

$$\mathbf{y}^k = \mathrm{prox}_{tg}\left(\mathbf{x}^{k-1} - t\nabla f\left(\mathbf{x}^{k-1}\right)\right),$$

$$\boxed{\mathbf{z}^k = \mathbf{x}^{k-1} - s\nabla\omega\left(\mathbf{x}^{k-1}\right)}, \Rightarrow \text{no need to optimize } \omega \text{ over 2 half spaces}$$

$$\mathbf{x}^k = \alpha_{k+1}\mathbf{z}^k + (1 - \alpha_{k+1})\,\mathbf{y}^k.$$

Bi-Level Gradient Sequential Averaging Method (BiG-SAM)

We take

$$T(\mathbf{x}) \equiv \mathrm{prox}_{tg}(\mathbf{x} - t\nabla f(\mathbf{x})) \quad \text{and} \quad S(\mathbf{x}) = \mathbf{x} - s\nabla\omega(\mathbf{x}).$$

(i) **Input:** $t \in (0, 1/L_f]$ and $s \in (0, 2/(L_\omega + \sigma)]$.

(ii) **Initialization**: Start with any $\mathbf{x}^0 \in \mathbb{R}^n$.

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

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

Proposition (S.-Shtern (2015))

Let $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$ be a sequence **generated by BiG-SAM** *and a let $\{\alpha_k\}_{k\in\mathbb{N}}$ be a "well-chosen" sequence. Then, the sequence $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$* **converges to $\mathbf{x}^* \in X^*$** *and*

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

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

Rate of Convergence of BiG-SAM

Proposition (S.-Shtern (2015))

Let $\left\{ \left(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k \right) \right\}_{k \in \mathbb{N}}$ *be a sequence generated by SAM where* $\{\alpha_k\}_{k \in \mathbb{N}} \in (0, 1]$ *such that* $\alpha_k = \min\left\{ \frac{2}{k(1-\beta)}, 1 \right\}$. *Then, for any* $\tilde{\mathbf{x}} \in \mathrm{Fix}(T)$ *we have*

$$\left\| \mathbf{x}^k - \mathbf{x}^{k-1} \right\| \leq \frac{C_{\tilde{\mathbf{x}}}}{k} \quad \textit{and} \quad \left\| \mathbf{y}^k - \mathbf{x}^{k-1} \right\| \leq \frac{C_{\tilde{\mathbf{x}}}}{k}, \quad k \geq 1,$$

where

$$C_{\mathbf{x}} = \frac{2\left(J+2\right)}{1-\beta} \max\left\{ \left\| \mathbf{x}^0 - \mathbf{x} \right\|, \frac{\left\| \nabla \omega\left(\mathbf{x} \right) \right\|}{1-\beta} \right\} \quad \textit{and} \quad J = \left\lfloor \frac{2}{1-\beta} \right\rfloor .$$

Rate of Convergence of BiG-SAM

Proposition (S.-Shtern (2015))

Let $\left\{\left(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k\right)\right\}_{k\in\mathbb{N}}$ *be a sequence generated by SAM where* $\{\alpha_k\}_{k\in\mathbb{N}} \in (0,1]$ *such that* $\alpha_k = \min\left\{\frac{2}{k(1-\beta)}, 1\right\}$. *Then, for any* $\tilde{\mathbf{x}} \in \mathrm{Fix}(T)$ *we have*

$$\left\|\mathbf{x}^k - \mathbf{x}^{k-1}\right\| \leq \frac{C_{\tilde{\mathbf{x}}}}{k} \quad \textit{and} \quad \left\|\mathbf{y}^k - \mathbf{x}^{k-1}\right\| \leq \frac{C_{\tilde{\mathbf{x}}}}{k}, \quad k \geq 1,$$

where

$$C_{\mathbf{x}} = \frac{2(J+2)}{1-\beta} \max\left\{\left\|\mathbf{x}^0 - \mathbf{x}\right\|, \frac{\|\nabla\omega(\mathbf{x})\|}{1-\beta}\right\} \quad \textit{and} \quad J = \left\lfloor \frac{2}{1-\beta} \right\rfloor.$$

Theorem (S.-Shtern (2015))

Let $\left\{\left(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k\right)\right\}_{k\in\mathbb{N}}$ *be a sequence generated by BiG-SAM. Then*

$$\varphi\left(\mathbf{y}^k\right) - \varphi\left(\mathbf{x}^*_{mn}\right) \leq \frac{C^2_{\mathbf{x}^*_{mn}}}{t(k+1)}, \quad t \in \left(0, \frac{1}{L_f}\right].$$

Rate of Convergence of BiG-SAM

Proposition (S.-Shtern (2015))

Let $\{(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\}_{k\in\mathbb{N}}$ *be a sequence generated by SAM where* $\{\alpha_k\}_{k\in\mathbb{N}} \in (0, 1]$ *such that* $\alpha_k = \min\left\{\frac{2}{k(1-\beta)}, 1\right\}$. *Then, for any* $\tilde{\mathbf{x}} \in \mathrm{Fix}(T)$ *we have*

$$\left\|\mathbf{x}^k - \mathbf{x}^{k-1}\right\| \leq \frac{C_{\tilde{\mathbf{x}}}}{k} \quad \textit{and} \quad \left\|\mathbf{y}^k - \mathbf{x}^{k-1}\right\| \leq \frac{C_{\tilde{\mathbf{x}}}}{k}, \quad k \geq 1,$$

where

$$C_{\mathbf{x}} = \frac{2(J+2)}{1-\beta} \max\left\{\left\|\mathbf{x}^0 - \mathbf{x}\right\|, \frac{\|\nabla\omega(\mathbf{x})\|}{1-\beta}\right\} \quad \textit{and} \quad J = \left\lfloor \frac{2}{1-\beta} \right\rfloor.$$

Theorem (S.-Shtern (2015))

Let $\{(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\}_{k\in\mathbb{N}}$ *be a sequence generated by BiG-SAM. Then*

$$\varphi\left(\mathbf{y}^k\right) - \varphi\left(\mathbf{x}^*_{mn}\right) \leq \frac{C^2_{\mathbf{x}^*_{mn}}}{t(k+1)}, \quad t \in \left(0, \frac{1}{L_f}\right].$$

Since $\mathbf{x}^k$ is **not necessarily feasible** for the inner problem, the convergence rate is given in terms of $\mathbf{y}^k$

BiG-SAM for Nonsmooth ω

- If ω is **nonsmooth** we can not use BiG-SAM directly ($\nabla\omega$ is not available).

BiG-SAM for Nonsmooth ω

- If ω is **nonsmooth** we can not use BiG-SAM directly ($\nabla\omega$ is not available).
- Instead we will use the Moreau envelope $M_{s\omega}$ of ω

$$M_{s\omega}(\mathbf{x}) = \min_{\mathbf{z}} \left\{ \omega(\mathbf{z}) + \frac{1}{2s}\|\mathbf{x} - \mathbf{z}\|^2 \right\},$$

which is

BiG-SAM for Nonsmooth ω

- If ω is **nonsmooth** we can not use BiG-SAM directly ($\nabla\omega$ is not available).

- Instead we will use the Moreau envelope $M_{s\omega}$ of ω

$$M_{s\omega}(\mathbf{x}) = \min_{\mathbf{z}} \left\{ \omega(\mathbf{z}) + \frac{1}{2s} \|\mathbf{x} - \mathbf{z}\|^2 \right\},$$

  which is
  - $\sigma/(1 + s\sigma)$ strongly convex,

BiG-SAM for Nonsmooth ω

- If ω is **nonsmooth** we can not use BiG-SAM directly ($\nabla \omega$ is not available).

- Instead we will use the Moreau envelope $M_{s\omega}$ of ω

$$M_{s\omega}(\mathbf{x}) = \min_{\mathbf{z}} \left\{ \omega(\mathbf{z}) + \frac{1}{2s} \|\mathbf{x} - \mathbf{z}\|^2 \right\},$$

which is
  - $\sigma/(1 + s\sigma)$ strongly convex,
  - Lipschitz continuous gradient with constant $1/s$ and

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

BiG-SAM for Nonsmooth ω

- If ω is **nonsmooth** we can not use BiG-SAM directly ($\nabla\omega$ is not available).

- Instead we will use the Moreau envelope $M_{s\omega}$ of ω

$$M_{s\omega}(\mathbf{x}) = \min_{\mathbf{z}} \left\{ \omega(\mathbf{z}) + \frac{1}{2s} \|\mathbf{x} - \mathbf{z}\|^2 \right\},$$

  which is
  - $\sigma/(1 + s\sigma)$ strongly convex,
  - Lipschitz continuous gradient with constant $1/s$ and

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

We can **apply Big-SAM** on the following bi-level problem

$$(\mathrm{MNP}_s) \qquad \begin{array}{ll} \min & M_{s\omega}(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in X^*. \end{array}$$

BiG-SAM for Nonsmooth ω

- If ω is **nonsmooth** we can not use BiG-SAM directly ($\nabla\omega$ is not available).
- Instead we will use the Moreau envelope $M_{s\omega}$ of ω

$$M_{s\omega}(\mathbf{x}) = \min_{\mathbf{z}} \left\{ \omega(\mathbf{z}) + \frac{1}{2s} \|\mathbf{x} - \mathbf{z}\|^2 \right\},$$

which is
  - $\sigma/(1 + s\sigma)$ strongly convex,
  - Lipschitz continuous gradient with constant $1/s$ and

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

We can **apply Big-SAM** on the following bi-level problem

$$(\mathrm{MNP}_s) \qquad \begin{array}{ll} \min & M_{s\omega}(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in X^*. \end{array}$$

Choosing $S(\mathbf{x}) = \mathbf{x} - s\nabla M_{s\omega}(\mathbf{x}) = \mathrm{prox}_{s\omega}(\mathbf{x})$ which is a $\beta = 1/(1 + s\sigma)$ contraction.

BiG-SAM for Nonsmooth ω

(1) **Input:** $t \in (0, 1/L_f]$ and $s > 0$.

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

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

$$\mathbf{y}^k = \mathrm{prox}_{tg}\left(\mathbf{x}^{k-1} - t\nabla f\left(\mathbf{x}^{k-1}\right)\right),$$

$$\mathbf{z}^k = \mathrm{prox}_{s\omega}\left(\mathbf{x}^{k-1}\right),$$

$$\mathbf{x}^k = \alpha_{k+1}\mathbf{z}^k + (1 - \alpha_{k+1})\,\mathbf{y}^k.$$

BiG-SAM for Nonsmooth ω

(1) **Input:** $t \in (0, 1/L_f]$ and $s > 0$.

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

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

$$\mathbf{y}^k = \mathrm{prox}_{tg}\left(\mathbf{x}^{k-1} - t\nabla f\left(\mathbf{x}^{k-1}\right)\right),$$

$$\mathbf{z}^k = \mathrm{prox}_{s\omega}\left(\mathbf{x}^{k-1}\right),$$

$$\mathbf{x}^k = \alpha_{k+1}\mathbf{z}^k + (1 - \alpha_{k+1})\,\mathbf{y}^k.$$

Proposition (S.-Shtern (2015))

Let $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ *be a sequence* **generated by BiG-SAM** *and a let* $\{\alpha_k\}_{k\in\mathbb{N}}$ *be a "well-chosen" sequence. Then, the sequence* $\left\{\mathbf{x}^k\right\}_{k\in\mathbb{N}}$ **converges to** $\mathbf{x}_s^* \in X^*$ *and*

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

Therefore $\mathbf{x}_s^*$ *is the* **optimal solution of problem (MNP$_s$).**

BiG-SAM for Nonsmooth ω

Our goal is to solve the following problem

$$\text{(MNP)} \qquad \begin{array}{ll} \min & \omega(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in X^*. \end{array}$$

BiG-SAM for Nonsmooth ω

Our goal is to solve the following problem

$$\text{(MNP)} \qquad \begin{array}{ll} \min & \omega(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in X^*. \end{array}$$

We assume that ω is **Lipschitz continuous** with constant ℓ_ω.

BiG-SAM for Nonsmooth ω

Our goal is to solve the following problem

$$\text{(MNP)} \qquad \begin{array}{ll} \min & \omega\left(\mathbf{x}\right) \\ \text{s.t.} & \mathbf{x} \in X^{*}. \end{array}$$

We assume that ω is **Lipschitz continuous** with constant ℓ_{ω}.

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

$$\omega\left(\mathbf{x}^{k}\right) - M_{s\omega}\left(\mathbf{x}^{k}\right) \leq \delta, \quad \forall\, k \geq 1.$$

BiG-SAM for Nonsmooth ω

Our goal is to solve the following problem

$$\text{(MNP)} \qquad \begin{array}{ll} \min & \omega(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in X^*. \end{array}$$

We assume that ω is **Lipschitz continuous** with constant ℓ_ω.

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

$$\omega\left(\mathbf{x}^k\right) - M_{s\omega}\left(\mathbf{x}^k\right) \leq \delta, \quad \forall\, k \geq 1.$$

The rate of convergence of Big-SAM, in this case, depends on δ

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

BiG-SAM for Nonsmooth ω

Our goal is to solve the following problem

$$\text{(MNP)} \qquad \begin{array}{ll} \min & \omega(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in X^*. \end{array}$$

We assume that ω is **Lipschitz continuous** with constant ℓ_ω.

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

$$\omega\left(\mathbf{x}^k\right) - M_{s\omega}\left(\mathbf{x}^k\right) \leq \delta, \quad \forall\, k \geq 1.$$

The rate of convergence of Big-SAM, in this case, depends on δ

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

Therefore the convergence rate is $O(1/(\varepsilon\delta^2))$, where ε is the desired inner function accuracy.

BiG-SAM for Nonsmooth ω

Our goal is to solve the following problem

$$\text{(MNP)} \qquad \begin{array}{ll} \min & \omega(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in X^*. \end{array}$$

We assume that ω is **Lipschitz continuous** with constant ℓ_ω.

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

$$\omega\left(\mathbf{x}^k\right) - M_{s\omega}\left(\mathbf{x}^k\right) \leq \delta, \quad \forall\, k \geq 1.$$

The rate of convergence of Big-SAM, in this case, depends on δ

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

Therefore the convergence rate is $O(1/(\varepsilon\delta^2))$, where ε is the desired inner function accuracy.

It should be noted that outer accuracy parameter δ also controls the following gap

$$\omega\left(\mathbf{x}_s^*\right) - \omega\left(\mathbf{x}_{mn}^*\right) \leq \delta.$$

Discretizations of Fredholm integral equations

The Phillips problem of estimating a function $f(t)$ that solves the integral equation

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

Discretizations of Fredholm integral equations

The Phillips problem of estimating a function $f(t)$ that solves the integral equation

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

where

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

Discretizations of Fredholm integral equations

The Phillips problem of estimating a function $f(t)$ that solves the integral equation

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

where

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

and

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

Discretizations of Fredholm integral equations

The Phillips problem of estimating a function $f(t)$ that solves the integral equation

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

where

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

and

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

(i) Discretize and reduce it to a **linear system** of the form $\mathbf{A}\mathbf{x}_{\mathrm{T}} = \mathbf{b}_{\mathrm{T}}$ using Galerkin method ($n = 1000$).

Discretizations of Fredholm integral equations

The Phillips problem of estimating a function $f(t)$ that solves the integral equation

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

where

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

and

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

(i) Discretize and reduce it to a **linear system** of the form $\mathbf{A}\mathbf{x}_{\mathrm{T}} = \mathbf{b}_{\mathrm{T}}$ using Galerkin method ($n = 1000$).

(ii) The observed right-hand side vector is given by $\mathbf{b} = \mathbf{b}_{\mathrm{T}} + \sigma\mathbf{w}$ (each component of $\mathbf{w}$ generated from a standard normal distribution and $\rho = 10^{-1}, 10^{-2}, 10^{-3}$).

Comparison between MNG and BiG-SAM

We are interested in the following **least squares core problem**

$$\min_{\mathbf{x} \geq 0} \|\mathbf{Ax} - \mathbf{b}\|^2 .$$

Comparison between MNG and BiG-SAM

We are interested in the following **least squares core problem**

$$\min_{\mathbf{x} \geq 0} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 .$$

Since the matrix $\mathbf{A}$ has **zero eigenvalues**, we consider outer objective function

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

where $\mathbf{Q} = \mathbf{L}^T\mathbf{L} + \mathbf{I}$ and $\mathbf{L}$ approximates the first-derivative operator.

Comparison between MNG and BiG-SAM

We are interested in the following **least squares core problem**

$$\min_{\mathbf{x} \geq 0} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 .$$

Since the matrix **A** has **zero eigenvalues**, we consider outer objective function

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

where $\mathbf{Q} = \mathbf{L}^T\mathbf{L} + \mathbf{I}$ and $\mathbf{L}$ approximates the first-derivative operator.

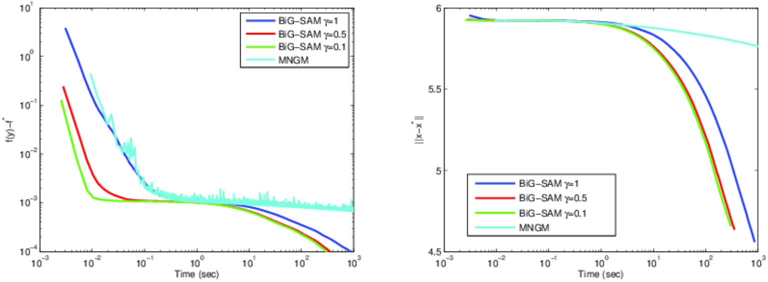

Figure : The progress of the algorithms in time for a Phillips example with $\rho = 0.01$ and $n = 100$

Comparison between MNG and BiG-SAM

Problem	ρ	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 : Averaged over 100 realization for each instance of problem and noise magnitude ρ (number of realizations terminated because of the time limit of 500 seconds).

For the MNG method see

Beck, A. and Sabach, S., **A first order method for finding minimal norm-like solutions of convex optimization problems**, *Mathematical Programming (Ser. A)* **147** (2014), 25–46.

For the MNG method see

Beck, A. and Sabach, S., **A first order method for finding minimal norm-like solutions of convex optimization problems**, *Mathematical Programming (Ser. A)* **147** (2014), 25–46.

For the BiG-SAM method see

Sabach, S. and Shtern, S., **A first order method for solving convex bi-level optimization problems**. Accepted in *SIAM Journal on Optimization* (2017).

Many thanks for your attention!

Email: ssabach@ie.technion.ac.il

Website: http://ssabach.net.technion.ac.il/