First-Order Algorithms for Solving Simple Convex Bilevel Optimization Problems

Shimrit Shtern
Joint work with Lior Doron (Technion)

One World Optimization Seminar
February 28th 2022

Simple Bilevel Optimization

A simple bilevel optimization problem is defined as:

$$\omega^* = \min_{\mathbf{x} \in X^*} \omega(\mathbf{x}) \tag{BLP}$$

where X^* is the set of minimizers of the convex problem (P)

$$\varphi^* = \min_{\mathbf{x} \in \mathbb{R}^n} \varphi(\mathbf{x}) \tag{P}$$

Simple Bilevel Optimization

A simple bilevel optimization problem is defined as:

$$\omega^* = \min_{\mathbf{x} \in X^*} \omega(\mathbf{x}) \tag{BLP}$$

where X^* is the set of minimizers of the convex problem (P)

$$\varphi^* = \min_{\mathbf{x} \in \mathbb{R}^n} \varphi(\mathbf{x}) \tag{P}$$

Background:

- We are concerned with the case where both ω and φ are convex.

Simple Bilevel Optimization

A simple bilevel optimization problem is defined as:

$$\omega^* = \min_{\mathbf{x} \in X^*} \omega(\mathbf{x}) \tag{BLP}$$

where X^* is the set of minimizers of the convex problem (P)

$$\varphi^* = \min_{\mathbf{x} \in \mathbb{R}^n} \varphi(\mathbf{x}) \tag{P}$$

Background:

- We are concerned with the case where both ω and φ are convex.
- Used to solve underdetermined problems in ML and signal processing.

Simple Bilevel Optimization

A simple bilevel optimization problem is defined as:

$$\omega^* = \min_{\mathbf{x} \in X^*} \omega(\mathbf{x}) \tag{BLP}$$

where X^* is the set of minimizers of the convex problem (P)

$$\varphi^* = \min_{\mathbf{x} \in \mathbb{R}^n} \varphi(\mathbf{x}) \tag{P}$$

Background:

- We are concerned with the case where both ω and φ are convex.
- Used to solve underdetermined problems in ML and signal processing.
- Example: Finding an optimal solution to

$$\min_{\mathbf{x} \in \mathbb{R}^n} \varphi(x) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$$

  which is the sparsest: $\omega(\mathbf{x}) = \|\mathbf{x}\|_1$, the densest: $\omega(\mathbf{x}) = \|\mathbf{x}\|_2^2$.

Challenges

Challenges

- The (BLP) is equivalent to:

$$\begin{aligned} \min \quad & \omega(\mathbf{x}) \\ \text{s.t.} \quad & \varphi(\mathbf{x}) \le \varphi^* \end{aligned} \qquad \text{(BLP')}$$

Challenges

- The (BLP) is equivalent to:

$$\begin{aligned} \min \quad & \omega(\mathbf{x}) \\ \text{s.t.} \quad & \varphi(\mathbf{x}) \leq \varphi^* \end{aligned} \tag{BLP'}$$

  - φ is usually not "simple", first-order methods such as (sub-)gradient projection cannot be used.

Challenges

- The (BLP) is equivalent to:

$$\begin{aligned} \min \quad & \omega(\mathbf{x}) \\ \text{s.t.} \quad & \varphi(\mathbf{x}) \leq \varphi^* \end{aligned} \tag{BLP'}$$

  - φ is usually not "simple", first-order methods such as (sub-)gradient projection cannot be used.
  - This problem does not satisfy regularity conditions.
  - Therefore strong duality and KKT conditions cannot be used.

Challenges

- The (BLP) is equivalent to:

$$\begin{aligned} \min \quad & \omega(\mathbf{x}) \\ \text{s.t.} \quad & \varphi(\mathbf{x}) \leq \varphi^* \end{aligned} \tag{BLP'}$$

  - φ is usually not "simple", first-order methods such as (sub-)gradient projection cannot be used.
  - This problem does not satisfy regularity conditions.
  - Therefore strong duality and KKT conditions cannot be used.
  - Even if φ^* is only approximated to high accuracy, the problem will be "almost irregular", which leads to numerical issues.

Regularization

- One of the well known methods to approximate bilevel problems is via regularization

$$\min_{x \in \mathbb{R}^n} \varphi(\mathbf{x}) + \alpha\omega(\mathbf{x}) \qquad (R_\alpha)$$

Regularization

- One of the well known methods to approximate bilevel problems is via regularization

$$\min_{x \in \mathbb{R}^n} \varphi(\mathbf{x}) + \alpha\omega(\mathbf{x}) \qquad (R_\alpha)$$

- For example, for the case $\varphi(\mathbf{x}) = \|\mathbf{Ax} - \mathbf{b}\|^2$
  - When $\omega(\mathbf{x}) = \|\mathbf{x}\|^2$ (Tikhonov regularization) - ridge regression.
  - When $\omega(\mathbf{x}) = \|\mathbf{x}\|_1$ - LASSO.

Regularization

- One of the well known methods to approximate bilevel problems is via regularization
$$\min_{x\in\mathbb{R}^n} \varphi(\mathbf{x}) + \alpha\omega(\mathbf{x}) \tag{R_α}$$

- For example, for the case $\varphi(\mathbf{x}) = \|\mathbf{Ax} - \mathbf{b}\|^2$
  - When $\omega(\mathbf{x}) = \|\mathbf{x}\|^2$ (Tikhonov regularization) - ridge regression.
  - When $\omega(\mathbf{x}) = \|\mathbf{x}\|_1$ - LASSO.

- Equivalent to a Lagrangian relaxation of the problem
$$\min_{\mathbf{x}\in\mathbb{R}^n}\{\varphi(\mathbf{x}) : \omega(\mathbf{x}) \leq \omega^*\}.$$

Regularization

- One of the well known methods to approximate bilevel problems is via regularization

$$\min_{x \in \mathbb{R}^n} \varphi(\mathbf{x}) + \alpha\omega(\mathbf{x}) \qquad (R_\alpha)$$

- For example, for the case $\varphi(\mathbf{x}) = \|\mathbf{Ax} - \mathbf{b}\|^2$
  - When $\omega(\mathbf{x}) = \|\mathbf{x}\|^2$ (Tikhonov regularization) - ridge regression.
  - When $\omega(\mathbf{x}) = \|\mathbf{x}\|_1$ - LASSO.

- Equivalent to a Lagrangian relaxation of the problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \{\varphi(\mathbf{x}) : \omega(\mathbf{x}) \leq \omega^*\}.$$

- Unclear how to find the right $\alpha > 0$ when ω^* is unknown.

Regularization

- One of the well known methods to approximate bilevel problems is via regularization

$$\min_{x \in \mathbb{R}^n} \varphi(\mathbf{x}) + \alpha\omega(\mathbf{x}) \qquad (R_\alpha)$$

- For example, for the case $\varphi(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$
  - When $\omega(\mathbf{x}) = \|\mathbf{x}\|^2$ (Tikhonov regularization) - ridge regression.
  - When $\omega(\mathbf{x}) = \|\mathbf{x}\|_1$ - LASSO.
- Equivalent to a Lagrangian relaxation of the problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \{\varphi(\mathbf{x}) : \omega(\mathbf{x}) \leq \omega^*\}.$$

- Unclear how to find the right $\alpha > 0$ when ω^* is unknown.
- Solving a sequence of (R_α) for decreasing values of α may be computationally demanding.

First-Order Methods for Iterative Regularization

- A class of methods that at iteration k perform one step of an iterative optimization algorithm on the problem (R_{α_k})

$$\min_{x \in \mathbb{R}^n} \varphi(\mathbf{x}) + \alpha_k \omega(\mathbf{x})$$

  where $\alpha_k \to 0$ as $k \to \infty$.

First-Order Methods for Iterative Regularization

- A class of methods that at iteration k perform one step of an iterative optimization algorithm on the problem (R_{α_k})

$$\min_{x \in \mathbb{R}^n} \varphi(\mathbf{x}) + \alpha_k \omega(\mathbf{x})$$

  where $\alpha_k \to 0$ as $k \to \infty$.

- The methods differ by the assumptions on the problem and the type of step performed.

First-Order Methods for Iterative Regularization

- A class of methods that at iteration k perform one step of an iterative optimization algorithm on the problem (R_{α_k})
$$\min_{x\in\mathbb{R}^n} \varphi(\mathbf{x}) + \alpha_k\omega(\mathbf{x})$$
where $\alpha_k \to 0$ as $k \to \infty$.

- The methods differ by the assumptions on the problem and the type of step performed.
    - **IR-PG**[Solodov 2007]: Asymptotic convergence to the solution of (BLP)

      Assumptions: $\varphi(\mathbf{x}) = f(\mathbf{x}) + \delta_C(\mathbf{x})$ where $f(\mathbf{x})$ is L_f-smooth, C closed and convex, and ω is L_ω-smooth.
      Step: Projected gradient $\mathbf{x}^{k+1} = \text{Proj}_C(\mathbf{x}^k - t_k(\nabla f(\mathbf{x}^k) + \alpha_k\nabla\omega(\mathbf{x}^k)))$, $t_k \leq \frac{1}{L_f+\alpha_k L_\omega}$.

First-Order Methods for Iterative Regularization

- A class of methods that at iteration k perform one step of an iterative optimization algorithm on the problem (R_{α_k})

$$\min_{x\in\mathbb{R}^n} \varphi(\mathbf{x}) + \alpha_k\omega(\mathbf{x})$$

  where $\alpha_k \to 0$ as $k \to \infty$.

- The methods differ by the assumptions on the problem and the type of step performed.
  - **IR-PG**[Solodov 2007]: Asymptotic convergence to the solution of (BLP)
  - **IR-IG** [Amini and Yousefian 2019]: $O(1/k^{0.5-\beta})$, $\beta \in (0, 0.5)$ convergence of $\varphi(\mathbf{x})$.

    Assumptions: $\varphi(\mathbf{x}) = \sum_{i=1}^{m} f_i(\mathbf{x}) + \delta_C(\mathbf{x})$, f_i proper, closed, and convex, C convex and compact, ω is strongly convex.
    Step: Incremental projected subgradient.

First-Order Methods for Iterative Regularization

- A class of methods that at iteration k perform one step of an iterative optimization algorithm on the problem (R_{α_k})

$$\min_{x\in\mathbb{R}^n} \varphi(\mathbf{x}) + \alpha_k \omega(\mathbf{x})$$

where $\alpha_k \to 0$ as $k \to \infty$.

- The methods differ by the assumptions on the problem and the type of step performed.
  - **IR-PG**[Solodov 2007]: Asymptotic convergence to the solution of (BLP)
  - **IR-IG** [Amini and Yousefian 2019]: $O(1/k^{0.5-\beta}),\ \beta \in (0, 0.5)$ convergence of $\varphi(\mathbf{x})$.
  - **SBP** [Dutta and Pandit 2020]: Asymptotic.

    Assumptions: Convexity.

    Step: Proximal point (limited applicability)

Other First-order Methods

- Assuming
  - ω is smooth and strongly convex.
  - $\varphi(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$ is composite function.

Other First-order Methods

- Assuming
  - ω is smooth and strongly convex.
  - $\varphi(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$ is composite function.

- The following methods provide a rate of convergence of $\varphi(\mathbf{x})$ to φ^* and asymptotic convergence to the solution of (BLP)

Other First-order Methods

- Assuming
  - ω is smooth and strongly convex.
  - $\varphi(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$ is composite function.

- The following methods provide a rate of convergence of $\varphi(\mathbf{x})$ to φ^* and asymptotic convergence to the solution of (BLP)
  - **MNG**[Beck and Sabach 2014]: Convergence rate of $O(1/\sqrt{k})$.

    Based on the notion of cutting-planes.
    Requires optimizing ω on the intersection of two half spaces in each iteration.

Other First-order Methods

- Assuming
  - ω is smooth and strongly convex.
  - $\varphi(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$ is composite function.

- The following methods provide a rate of convergence of $\varphi(\mathbf{x})$ to φ^* and asymptotic convergence to the solution of (BLP)
  - **MNG**[Beck and Sabach 2014]: Convergence rate of $O(1/\sqrt{k})$.
  - **BiG-SAM**[Sabach and Shtern 2017]: Convergence rate of $O(1/k)$.

    Based sequential averaging of the gradient step for ω and proximal gradient step for φ.
    Extension to cases where ω is a sum of Lipschitz continuous and smooth functions.

Other First-order Methods

- Assuming
  - ω is smooth and strongly convex.
  - $\varphi(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$ is composite function.

- The following methods provide a rate of convergence of $\varphi(\mathbf{x})$ to φ^* and asymptotic convergence to the solution of (BLP)
  - **MNG**[Beck and Sabach 2014]: Convergence rate of $O(1/\sqrt{k})$.
  - **BiG-SAM**[Sabach and Shtern 2017]: Convergence rate of $O(1/k)$.
  - **iBiG-SAM**[Shehu, Vuong, and Zemkoho 2021]: Asymptotic convergence. Running an inertial extrapolation over BiG-SAM steps.

Contribution

- Motivation - $\omega(\cdot) = \|\cdot\|_1$

Contribution

- Motivation - $\omega(\cdot) = \|\cdot\|_1$
- **IT***erative* **A***pproximation and* **L***evel-set* **EX***pansion* (ITALEX) scheme to solve (BLP):

Contribution

- Motivation - $\omega(\cdot) = \|\cdot\|_1$
- **IT***erative* **A***pproximation and* **L***evel-set* **EX***pansion* (ITALEX) scheme to solve (BLP):
  - We do not require ω to be neither smooth nor strongly-convex.

Contribution

- Motivation - $\omega(\cdot) = \|\cdot\|_1$
- **IT***erative* **A***pproximation and* **L***evel-set* **EX***pansion* (ITALEX) scheme to solve (BLP):
  - We do not require ω to be neither smooth nor strongly-convex.
  - Easily applied to l_p norms.

Contribution

- Motivation - $\omega(\cdot) = \|\cdot\|_1$
- ***IT**erative **A**pproximation and **L**evel-set **EX**pansion* (ITALEX) scheme to solve (BLP):
  - We do not require ω to be neither smooth nor strongly-convex.
  - Easily applied to l_p norms.
  - For any $\varepsilon > 0$ produces a solution $\mathbf{x}^k$ such that

$$\varphi(\mathbf{x}^k) \le \varphi^* + \varepsilon, \quad \omega(\mathbf{x}^k) - \omega^* \le O(\sqrt{\epsilon}).$$

  where $\varepsilon = O(1/k)$.

Bilevel methods - comparison

Method	$\varphi = f + g$ properties	ω properties	Convergence to $\varphi*$	Convergence to ω^*
IR-PG [Solodov 2007]	Classical composite	Smooth	Asymptotic	Asymptotic
MNG [Beck and Sabach 2014]	Classical composite	Smooth, strongly convex	$O\left(\frac{1}{\sqrt{k}}\right)$	Asymptotic
BiG-SAM [Sabach and Shtern 2017]	Classical composite	Smooth, strongly convex	$O\left(\frac{1}{k}\right)$	Asymptotic
IR-IG [Amini and Yousefian 2019]	f is a finite sum, $g = \delta_C$, C compact	Strongly convex	$O\left(\frac{1}{k^{0.5-\beta}}\right)$ $\beta \in (0, 0.5)$	Asymptotic
SBP [Dutta and Pandit 2020]	General	General	Asymptotic	Asymptotic
ITALEX [This paper]	Classical composite	Norm-like function	$O\left(\frac{1}{k}\right)$	$O\left(\frac{1}{\sqrt{k}}\right)$
	$g = 0$			Super-optimal

Reformulating (BLP)

- The key idea: if ω is a simple function we can compute projection/linear oracle on its level set.

Reformulating (BLP)

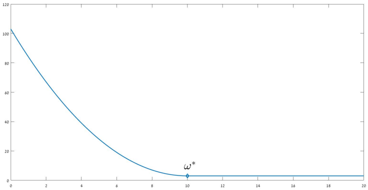
- The key idea: if ω is a simple function we can compute projection/linear oracle on its level set.
- For any $\alpha \in \mathbb{R}$ we can define the extended valued function

$$h(\alpha) = \min_{\mathbf{x},\mathbf{z}} \{\varphi(\mathbf{x}) + \|\mathbf{x} - \mathbf{z}\|^2 : \omega(\mathbf{z}) \leq \alpha\} \qquad (P_\alpha)$$

Reformulating (BLP)

- The key idea: if ω is a simple function we can compute projection/linear oracle on its level set.
- For any $\alpha \in \mathbb{R}$ we can define the extended valued function

$$h(\alpha) = \min_{\mathbf{x},\mathbf{z}} \{\varphi(\mathbf{x}) + \|\mathbf{x} - \mathbf{z}\|^2 : \omega(\mathbf{z}) \leq \alpha\} \qquad (P_\alpha)$$



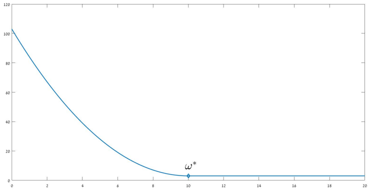
Reformulating (BLP)

- The key idea: if ω is a simple function we can compute projection/linear oracle on its level set.

- For any $\alpha \in \mathbb{R}$ we can define the extended valued function

$$h(\alpha) = \min_{\mathbf{x},\mathbf{z}} \{\varphi(\mathbf{x}) + \|\mathbf{x} - \mathbf{z}\|^2 : \omega(\mathbf{z}) \leq \alpha\} \qquad (P_\alpha)$$

- We will approximately solve a sequence of (P_α).

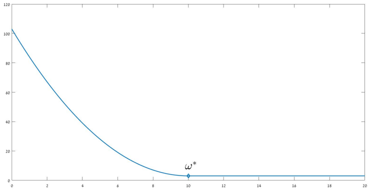


Reformulating (BLP)

- The key idea: if ω is a simple function we can compute projection/linear oracle on its level set.
- For any $\alpha \in \mathbb{R}$ we can define the extended valued function

$$h(\alpha) = \min_{\mathbf{x},\mathbf{z}} \{\varphi(\mathbf{x}) + \|\mathbf{x} - \mathbf{z}\|^2 : \omega(\mathbf{z}) \leq \alpha\} \qquad (P_\alpha)$$

- We will approximately solve a sequence of (P_α).
- We will look for the smallest α such that $h(\alpha)$ is ε close to φ^*.



Approach

- **IT***erative* **A***pproximation and* **L***evel-set* **EX***pansion* is based on two main operations:

Approach

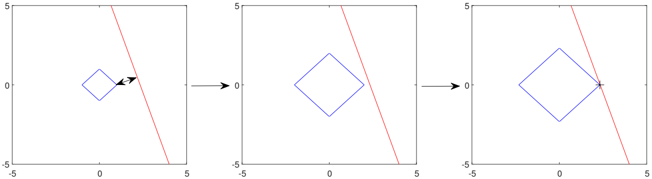
- **IT***erative* **A***pproximation and* **L***evel-set* **EX***pansion* is based on two main operations:
  1. Approximate $h(\alpha)$ - the optimal value of (P_α)

Approach

- **IT***erative* **A***pproximation and* **L***evel-set* **EX***pansion* is based on two main operations:
  1. Approximate $h(\alpha)$ - the optimal value of (P_α)
  2. If $h(\alpha)$ is too big, then increase α
     Expansion of the level set while maintaining $\alpha \leq \omega^*$.

Approach

- **IT***erative* **A***pproximation and* **L***evel-set* **EX***pansion* is based on two main operations:
  1. Approximate $h(\alpha)$ - the optimal value of (P_α)
  2. If $h(\alpha)$ is too big, then increase α
     Expansion of the level set while maintaining $\alpha \leq \omega^*$.



ITALEX - General algorithm

Algorithm 1: ITALEX- General Scheme

Input: ε, $\bar{\varphi} \in [\varphi^*, \varphi^* + \frac{\varepsilon}{2}]$,
$\alpha_0 \leq \omega^*$, $\mathbf{x}^0 \in \text{dom}(\varphi)$, $\mathbf{z}^0 \in \text{Lev}_\omega(\alpha_0)$
Approximation oracle $\mathcal{O}^{\omega,\varphi}$, Expansion oracle $\mathcal{E}^\omega$,

for all $k = 1, 2, \ldots$ **do**

$(\rho_k, (\mathbf{x}^k, \mathbf{z}^k)) = \mathcal{O}^{\omega,\varphi}((\mathbf{x}^{k-1}, \mathbf{z}^{k-1}), \alpha_{k-1}, \bar{\varphi}, \frac{\varepsilon}{2})$

if $\varphi(\mathbf{x}^k) + \|\mathbf{x}^k - \mathbf{z}^k\|^2 \leq \bar{\varphi} + \frac{\varepsilon}{2}$ **then**

return $\mathbf{x}^k$

else

$\alpha_k = \mathcal{E}^\omega(\alpha_{k-1}, \bar{\varphi}, \rho_k)$

end if

end for

ITALEX - General algorithm

Algorithm 2: ITALEX- General Scheme

Input: ε, $\bar{\varphi} \in [\varphi^*, \varphi^* + \frac{\varepsilon}{2}]$,
$\alpha_0 \leq \omega^*, \; \mathbf{x}^0 \in \mathrm{dom}(\varphi), \; \mathbf{z}^0 \in \mathrm{Lev}_\omega(\alpha_0)$
Approximation oracle $\mathcal{O}^{\omega,\varphi}$, Expansion oracle $\mathcal{E}^\omega$,

for all $k = 1, 2, \ldots$ **do**
$(\rho_k, (\mathbf{x}^k, \mathbf{z}^k)) = \mathcal{O}^{\omega,\varphi}((\mathbf{x}^{k-1}, \mathbf{z}^{k-1}), \alpha_{k-1}, \bar{\varphi}, \frac{\varepsilon}{2})$
if $\varphi(\mathbf{x}^k) + \|\mathbf{x}^k - \mathbf{z}^k\|^2 \leq \bar{\varphi} + \frac{\varepsilon}{2}$ **then**
return $\mathbf{x}^k$
else
$\alpha_k = \mathcal{E}^\omega(\alpha_{k-1}, \bar{\varphi}, \rho_k)$
end if
end for

What should we require from these oracles to guarantee ITALEX converges to the solution of (BLP)?

Expansion Oracle

Expansion Oracle

Definition (Expansion Oracle)

An operator $\mathcal{E}^{\omega,\varphi}(\alpha, \bar{\varphi}, \rho)$ which for any $\rho \leq h(\alpha) - \bar{\varphi}$ returns $\alpha < \beta \leq \omega^*$

Expansion Oracle

Definition (Expansion Oracle)

An operator $\mathcal{E}^{\omega,\varphi}(\alpha, \bar{\varphi}, \rho)$ which for any $\rho \leq h(\alpha) - \bar{\varphi}$ returns $\alpha < \beta \leq \omega^*$

How do we construct such an operator?

Constructing an Expansion Oracle - Assumptions

Assumption (Norm-like function)

$\omega : \mathbb{R}^n \to \mathbb{R}$ is convex and satisfies the following properties.

- (i) For any $\alpha \in \mathbb{R}$, The level set $\mathrm{Lev}_\omega(\alpha)$ is compact.
- (ii) There exists a γ-global error-bound of ω, *i.e.,*

$$\exists \gamma > 0 : \forall \mathbf{x} \in \mathbb{R}^n, \mathrm{dist}(\mathbf{x}, \mathrm{Lev}_\omega(\alpha)) \leq \gamma[\omega(\mathbf{x}) - \alpha]_+.$$

Constructing an Expansion Oracle - Assumptions

Assumption (Norm-like function)

$\omega : \mathbb{R}^n \to \mathbb{R}$ is convex and satisfies the following properties.

1. (i) For any $\alpha \in \mathbb{R}$, The level set $\text{Lev}_\omega(\alpha)$ is compact.
2. (ii) There exists a γ-global error-bound of ω, *i.e.*,

$$\exists \gamma > 0 : \forall \mathbf{x} \in \mathbb{R}^n, \text{dist}(\mathbf{x}, \text{Lev}_\omega(\alpha)) \leq \gamma[\omega(\mathbf{x}) - \alpha]_+.$$

- (i) holds if ω is coercive (*e.g.*, any norm).

Constructing an Expansion Oracle - Assumptions

Assumption (Norm-like function)

$\omega : \mathbb{R}^n \to \mathbb{R}$ is convex and satisfies the following properties.

(i) For any $\alpha \in \mathbb{R}$, The level set $\text{Lev}_\omega(\alpha)$ is compact.

(ii) There exists a γ-global error-bound of ω, *i.e.*,

$$\exists \gamma > 0 : \forall \mathbf{x} \in \mathbb{R}^n, \text{dist}(\mathbf{x}, \text{Lev}_\omega(\alpha)) \leq \gamma[\omega(\mathbf{x}) - \alpha]_+.$$

- (i) holds if ω is coercive (*e.g.*, any norm).
- Using [Lewis and Pang 1998, Theorem 1], (ii) can be verified for various functions by calculating

$$\gamma^{-1} = \inf_{\mathbf{v},\mathbf{x}} \{\|\mathbf{v}\| : \mathbf{v} \in \partial\omega(\mathbf{x}), \omega(\mathbf{x}) > \alpha\}.$$

Constructing an Expansion Oracle - Assumptions

Assumption (Norm-like function)

$\omega : \mathbb{R}^n \to \mathbb{R}$ is convex and satisfies the following properties.

(i) For any $\alpha \in \mathbb{R}$, The level set $\text{Lev}_\omega(\alpha)$ is compact.

(ii) There exists a γ-global error-bound of ω, *i.e.,*

$$\exists \gamma > 0 : \forall \mathbf{x} \in \mathbb{R}^n, \text{dist}(\mathbf{x}, \text{Lev}_\omega(\alpha)) \leq \gamma[\omega(\mathbf{x}) - \alpha]_+.$$

- (i) holds if ω is coercive (*e.g.*, any norm).
- Using [Lewis and Pang 1998, Theorem 1], (ii) can be verified for various functions by calculating

$$\gamma^{-1} = \inf_{\mathbf{v},\mathbf{x}} \{\|\mathbf{v}\| : \mathbf{v} \in \partial\omega(\mathbf{x}), \omega(\mathbf{x}) > \alpha\}.$$

- Examples: ℓ_p-norm, Q-norm, Elastic net $(\|\mathbf{x}\|_1 + t\|x\|_2^2)$.

Constructing an Expansion Oracle - cont.

Theorem

Let ω be a norm-like function. Then for any $\rho \leq h(\alpha) - \bar{\varphi}$, the operator

$$\mathcal{E}^{\omega}(\alpha, \bar{\varphi}, \rho) = \alpha + \frac{\sqrt{\rho}}{\gamma}$$

is a valid expansion oracle.

Constructing an Expansion Oracle - cont.

Theorem

Let ω be a norm-like function. Then for any $\rho \leq h(\alpha) - \bar{\varphi}$, the operator

$$\mathcal{E}^{\omega}(\alpha, \bar{\varphi}, \rho) = \alpha + \frac{\sqrt{\rho}}{\gamma}$$

is a valid expansion oracle.

Proof sketch:

Constructing an Expansion Oracle - cont.

Theorem

Let ω be a norm-like function. Then for any $\rho \leq h(\alpha) - \bar{\varphi}$, the operator

$$\mathcal{E}^{\omega}(\alpha, \bar{\varphi}, \rho) = \alpha + \frac{\sqrt{\rho}}{\gamma}$$

is a valid expansion oracle.

Proof sketch:

- Let $\mathbf{x}^*$ be an optimal solution of (BLP).

Constructing an Expansion Oracle - cont.

Theorem

Let ω be a norm-like function. Then for any $\rho \leq h(\alpha) - \bar{\varphi}$, the operator

$$\mathcal{E}^{\omega}(\alpha, \bar{\varphi}, \rho) = \alpha + \frac{\sqrt{\rho}}{\gamma}$$

is a valid expansion oracle.

Proof sketch:

- Let $\mathbf{x}^*$ be an optimal solution of (BLP).
- Then $(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^*, \mathrm{Proj}_{\mathrm{Lev}_{\omega}(\alpha)}(\mathbf{x}^*))$ is sub-optimal for (P_α).

  $\rho \leq h(\alpha) - \bar{\varphi}$

Constructing an Expansion Oracle - cont.

Theorem

Let ω be a norm-like function. Then for any $\rho \leq h(\alpha) - \bar{\varphi}$, the operator

$$\mathcal{E}^{\omega}(\alpha, \bar{\varphi}, \rho) = \alpha + \frac{\sqrt{\rho}}{\gamma}$$

is a valid expansion oracle.

Proof sketch:

- Let $\mathbf{x}^*$ be an optimal solution of (BLP).
- Then $(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^*, \mathrm{Proj}_{\mathrm{Lev}_\omega(\alpha)}(\mathbf{x}^*))$ is sub-optimal for (P_α).

$$\rho \leq h(\alpha) - \bar{\varphi} \leq \varphi(\mathbf{x}^*) + \mathrm{dist}(\mathbf{x}^*, \mathrm{Lev}_\omega(\alpha))^2 - \bar{\varphi} \leq \mathrm{dist}(\mathbf{x}^*, \mathrm{Lev}_\omega(\alpha))^2.$$

Constructing an Expansion Oracle - cont.

Theorem

Let ω be a norm-like function. Then for any $\rho \le h(\alpha) - \bar{\varphi}$, the operator

$$\mathcal{E}^{\omega}(\alpha, \bar{\varphi}, \rho) = \alpha + \frac{\sqrt{\rho}}{\gamma}$$

is a valid expansion oracle.

Proof sketch:

- Let $\mathbf{x}^*$ be an optimal solution of (BLP).
- Then $(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^*, \text{Proj}_{\text{Lev}_{\omega}(\alpha)}(\mathbf{x}^*))$ is sub-optimal for (P_{α}).

$$\rho \le h(\alpha) - \bar{\varphi} \le \varphi(\mathbf{x}^*) + \text{dist}(\mathbf{x}^*, \text{Lev}_{\omega}(\alpha))^2 - \bar{\varphi} \le \text{dist}(\mathbf{x}^*, \text{Lev}_{\omega}(\alpha))^2.$$

- Since ω is norm-like

$$\text{dist}(\mathbf{x}^*, \text{Lev}_{\omega}(\alpha)) \le \gamma(\omega^* - \alpha).$$

Constructing an Expansion Oracle - cont.

Theorem

Let ω be a norm-like function. Then for any $\rho \leq h(\alpha) - \bar{\varphi}$, the operator

$$\mathcal{E}^{\omega}(\alpha, \bar{\varphi}, \rho) = \alpha + \frac{\sqrt{\rho}}{\gamma}$$

is a valid expansion oracle.

Proof sketch:

- Let $\mathbf{x}^*$ be an optimal solution of (BLP).
- Then $(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^*, \text{Proj}_{\text{Lev}_{\omega}(\alpha)}(\mathbf{x}^*))$ is sub-optimal for (P_{α}).

$$\rho \leq h(\alpha) - \bar{\varphi} \leq \varphi(\mathbf{x}^*) + \text{dist}(\mathbf{x}^*, \text{Lev}_{\omega}(\alpha))^2 - \bar{\varphi} \leq \text{dist}(\mathbf{x}^*, \text{Lev}_{\omega}(\alpha))^2.$$

- Since ω is norm-like

$$\text{dist}(\mathbf{x}^*, \text{Lev}_{\omega}(\alpha)) \leq \gamma(\omega^* - \alpha).$$

- Thus, $\mathcal{E}^{\omega}(\alpha, \bar{\varphi}, \rho) \leq \omega^*$.

Convergence.

Convergence.

We can now bound N (the number of ITALEX outer iterations)

Convergence.

We can now bound N (the number of ITALEX outer iterations)

Corollary

Let ω be a norm-like function, and $\varepsilon > 0$. Then ITALEX with the above expansion oracle has at most N iterations where

$$N \leq \left\lceil \frac{\gamma\left(\omega^* - \omega(\mathbf{z}^0)\right)}{\varepsilon} \right\rceil .$$

Moreover,

$$\omega(\mathbf{x}^N) - \omega^* \leq \ell_{\omega,0}\sqrt{\epsilon}$$

where $\ell_{\omega,0}$ is the Lipschitz constant of ω on the compact set

$$\mathcal{W}^0 = \left\{\mathbf{x} \in \mathbb{R}^n : \mathrm{dist}(\mathbf{x}, \mathrm{Lev}_\omega(\alpha_0)) \leq \gamma(\bar{\omega} - \omega(\mathbf{z}^0)) + \sqrt{\epsilon}\right\}.$$

Approximation Oracle

Approximation Oracle

Definition (Approximation Oracle)

An operator $\mathcal{O}^{\omega,\varphi}((\mathbf{x},\mathbf{z}),\alpha,\bar{\varphi},\varepsilon)$ for any $\varepsilon > 0$, $\bar{\varphi} \geq \varphi^*, \alpha \geq \min_{\mathbf{x}\in\mathbb{R}^n}\{\omega(\mathbf{x})\} \equiv \underline{\omega}$ which determines

1. If $h(\alpha) - \bar{\varphi} \geq \frac{\varepsilon}{2}$ and returns $\frac{\varepsilon}{2} \leq \rho \leq h(\alpha) - \bar{\varphi}$.
2. If we found $\mathbf{x}$ such that $\varphi(\mathbf{x}) + \|\mathbf{x} - \mathbf{z}\|^2 - \bar{\varphi} \leq \varepsilon$ returns $(\mathbf{x},\mathbf{z})$.

Approximation Oracle

Definition (Approximation Oracle)

An operator $\mathcal{O}^{\omega,\varphi}((\mathbf{x},\mathbf{z}),\alpha,\bar{\varphi},\varepsilon)$ for any $\varepsilon > 0$, $\bar{\varphi} \geq \varphi^*, \alpha \geq \min_{\mathbf{x}\in\mathbb{R}^n}\{\omega(\mathbf{x})\} \equiv \underline{\omega}$ which determines

1. If $h(\alpha) - \bar{\varphi} \geq \frac{\varepsilon}{2}$ and returns $\frac{\varepsilon}{2} \leq \rho \leq h(\alpha) - \bar{\varphi}$.
2. If we found $\mathbf{x}$ such that $\varphi(\mathbf{x}) + \|\mathbf{x} - \mathbf{z}\|^2 - \bar{\varphi} \leq \varepsilon$ returns $(\mathbf{x},\mathbf{z})$.

- There is an overlap between the two possible outputs if $\frac{\varepsilon}{2} \leq h(\alpha) - \bar{\varphi} \leq \varepsilon$.

Approximation Oracle

Definition (Approximation Oracle)

An operator $\mathcal{O}^{\omega,\varphi}((\mathbf{x},\mathbf{z}),\alpha,\bar{\varphi},\varepsilon)$ for any $\varepsilon > 0$, $\bar{\varphi} \geq \varphi^*, \alpha \geq \min_{\mathbf{x}\in\mathbb{R}^n}\{\omega(\mathbf{x})\} \equiv \underline{\omega}$ which determines

1. If $h(\alpha) - \bar{\varphi} \geq \frac{\varepsilon}{2}$ and returns $\frac{\varepsilon}{2} \leq \rho \leq h(\alpha) - \bar{\varphi}$.
2. If we found $\mathbf{x}$ such that $\varphi(\mathbf{x}) + \|\mathbf{x}-\mathbf{z}\|^2 - \bar{\varphi} \leq \varepsilon$ returns $(\mathbf{x},\mathbf{z})$.

- There is an overlap between the two possible outputs if $\frac{\varepsilon}{2} \leq h(\alpha) - \bar{\varphi} \leq \varepsilon$.

How do we construct such an operator?

Approximation Oracle

Assumption

The inner function $\varphi \equiv f + g$ satisfies the following:

1. $f : \mathbb{R}^n \to \mathbb{R}$ is closed, convex, continuously differentiable with a Lipschitz-continuous gradient with constant L_f, *i.e.*,
$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L_f \|\mathbf{x} - \mathbf{y}\|, \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$
2. $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a proper, closed, and convex function.

Approximation Oracle

Assumption

The inner function $\varphi \equiv f + g$ satisfies the following:

1. $f : \mathbb{R}^n \to \mathbb{R}$ is closed, convex, continuously differentiable with a Lipschitz-continuous gradient with constant L_f, *i.e.*,

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

2. $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a proper, closed, and convex function.

- For $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2) \in \mathbb{R}^n \times \mathbb{R}^n$ defining

$$\hat{\varphi}^{\alpha}(\mathbf{y}) = \varphi(\mathbf{y}_1) + \|\mathbf{y}_1 - \mathbf{y}_2\|^2 + \delta_{\mathrm{Lev}_{\omega}(\alpha)}(\mathbf{y}_2)$$

Approximation Oracle

Assumption

The inner function $\varphi \equiv f + g$ satisfies the following:

1. $f : \mathbb{R}^n \to \mathbb{R}$ is closed, convex, continuously differentiable with a Lipschitz-continuous gradient with constant L_f, *i.e.*,
$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L_f \|\mathbf{x} - \mathbf{y}\|, \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$
2. $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a proper, closed, and convex function.

- For $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2) \in \mathbb{R}^n \times \mathbb{R}^n$ defining
$$\hat{\varphi}^\alpha(\mathbf{y}) = \varphi(\mathbf{y}_1) + \|\mathbf{y}_1 - \mathbf{y}_2\|^2 + \delta_{\text{Lev}_\omega(\alpha)}(\mathbf{y}_2)$$
- $\hat{\varphi}^\alpha = \hat{f} + \hat{g}^\alpha$ is a composite function.
  - $\hat{f}(\mathbf{y}) = f(\mathbf{y}_1) + \|\mathbf{y}_1 - \mathbf{y}_2\|^2$ has an $(L_f + 2)$-Lipschitz continuous gradient.
  - $\hat{g}^\alpha(\mathbf{y}) = g(\mathbf{y}_1) + \delta_{\text{Lev}_\omega(\alpha)}(\mathbf{y}_2)$ is separable.

Generalized Conditional Gradient

Generalized Conditional Gradient

- Generalized Conditional Gradeint (GCG) composite functions:
  - GCG step

$$\mathbf{y}^{k+1} = \mathbf{y}^k + t_k(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k),$$

where

$$\mathbf{p}(\mathbf{y}) \in \arg\min \{\langle \nabla f(\mathbf{y}), \mathbf{p}\rangle + g(\mathbf{p})\}$$

Generalized Conditional Gradient

- Generalized Conditional Gradeint (GCG) composite functions:
  - GCG step
    $$\mathbf{y}^{k+1} = \mathbf{y}^k + t_k(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k),$$
    where
    $$\mathbf{p}(\mathbf{y}) \in \arg\min \{\langle \nabla f(\mathbf{y}), \mathbf{p} \rangle + g(\mathbf{p})\}$$
  - Bound on the optimality gap:
    $$S(\mathbf{y}) = \langle \nabla f(\mathbf{y}), \mathbf{y} - \mathbf{p}(\mathbf{y}) \rangle + g(\mathbf{y}) - g(\mathbf{p}(\mathbf{y}))$$

Generalized Conditional Gradient

- Generalized Conditional Gradeint (GCG) composite functions:
  - GCG step
    $$\mathbf{y}^{k+1} = \mathbf{y}^k + t_k(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k),$$
    where
    $$\mathbf{p}(\mathbf{y}) \in \arg\min \{\langle \nabla f(\mathbf{y}), \mathbf{p}\rangle + g(\mathbf{p})\}$$
  - Bound on the optimality gap:
    $$S(\mathbf{y}) = \langle \nabla f(\mathbf{y}), \mathbf{y} - \mathbf{p}(\mathbf{y})\rangle + g(\mathbf{y}) - g(\mathbf{p}(\mathbf{y})) \geq \varphi(\mathbf{y}) - \varphi(\mathbf{p}(\mathbf{y})) \geq \varphi(\mathbf{y}) - \varphi^*$$

Generalized Conditional Gradient

- Generalized Conditional Gradeint (GCG) composite functions:
  - GCG step
$$\mathbf{y}^{k+1} = \mathbf{y}^k + t_k(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k),$$
where
$$\mathbf{p}(\mathbf{y}) \in \arg\min \{\langle \nabla f(\mathbf{y}), \mathbf{p}\rangle + g(\mathbf{p})\}$$
  - Bound on the optimality gap:
$$S(\mathbf{y}) = \langle \nabla f(\mathbf{y}), \mathbf{y} - \mathbf{p}(\mathbf{y})\rangle + g(\mathbf{y}) - g(\mathbf{p}(\mathbf{y})) \geq \varphi(\mathbf{y}) - \varphi(\mathbf{p}(\mathbf{y})) \geq \varphi(\mathbf{y}) - \varphi^*$$
  - For a proper choice of step-size, admits sufficient decrease
$$\varphi(\mathbf{y}^k) - \varphi(\mathbf{y}^{k+1}) \geq \frac{1}{2}\min\left\{S(\mathbf{y}^k), \frac{(S(\mathbf{y}^k))^2}{L_f D^2}\right\},$$
where D is an upper bound on the diameter of dom(g)

Generalized Conditional Gradient

- Generalized Conditional Gradeint (GCG) composite functions:
  - GCG step
  $$\mathbf{y}^{k+1} = \mathbf{y}^k + t_k(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k),$$
  where
  $$\mathbf{p}(\mathbf{y}) \in \arg\min \{\langle \nabla f(\mathbf{y}), \mathbf{p}\rangle + g(\mathbf{p})\}$$
  - Bound on the optimality gap:
  $$S(\mathbf{y}) = \langle \nabla f(\mathbf{y}), \mathbf{y} - \mathbf{p}(\mathbf{y})\rangle + g(\mathbf{y}) - g(\mathbf{p}(\mathbf{y})) \geq \varphi(\mathbf{y}) - \varphi(\mathbf{p}(\mathbf{y})) \geq \varphi(\mathbf{y}) - \varphi^*$$
  - For a proper choice of step-size, admits sufficient decrease
  $$\varphi(\mathbf{y}^k) - \varphi(\mathbf{y}^{k+1}) \geq \frac{1}{2}\min\left\{S(\mathbf{y}^k), \frac{(S(\mathbf{y}^k))^2}{L_f D^2}\right\},$$
  where D is an upper bound on the diameter of $\mathrm{dom}(g)$
  - Leads to $O(1/k)$ convergence.

Generalized Conditional Gradient

- Generalized Conditional Gradeint (GCG) composite functions:
  - GCG step
    $$\mathbf{y}^{k+1} = \mathbf{y}^k + t_k(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k),$$
    where
    $$\mathbf{p}(\mathbf{y}) \in \arg\min \{\langle \nabla f(\mathbf{y}), \mathbf{p} \rangle + g(\mathbf{p})\}$$
  - Bound on the optimality gap:
    $$S(\mathbf{y}) = \langle \nabla f(\mathbf{y}), \mathbf{y} - \mathbf{p}(\mathbf{y}) \rangle + g(\mathbf{y}) - g(\mathbf{p}(\mathbf{y})) \geq \varphi(\mathbf{y}) - \varphi(\mathbf{p}(\mathbf{y})) \geq \varphi(\mathbf{y}) - \varphi^*$$
  - For a proper choice of step-size, admits sufficient decrease
    $$\varphi(\mathbf{y}^k) - \varphi(\mathbf{y}^{k+1}) \geq \frac{1}{2} \min\left\{ S(\mathbf{y}^k), \frac{(S(\mathbf{y}^k))^2}{L_f D^2} \right\},$$
    where D is an upper bound on the diameter of $\mathrm{dom}(g)$
  - Leads to $O(1/k)$ convergence.

Applying the algorithm to $\hat{\varphi}^\alpha$.

Generalized Conditional Gradient

- Generalized Conditional Gradeint (GCG) composite functions:
  - GCG step
    $$\mathbf{y}^{k+1} = \mathbf{y}^k + t_k(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k),$$
    where
    $$\mathbf{p}(\mathbf{y}) \in \arg\min \left\{ \langle \nabla \hat{f}(\mathbf{y}), \mathbf{p} \rangle + \hat{g}^\alpha(\mathbf{p}) \right\}$$
  - Bound on the optimality gap:

$$S^\alpha(\mathbf{y}) = \langle \nabla \hat{f}(\mathbf{y}), \mathbf{y} - \mathbf{p}(\mathbf{y}) \rangle + \hat{g}^\alpha(\mathbf{y}) - \hat{g}^\alpha(\mathbf{p}(\mathbf{y})) \geq \hat{\varphi}^\alpha(\mathbf{y}) - \hat{\varphi}^\alpha(\mathbf{p}(\mathbf{y})) \geq \hat{\varphi}^\alpha(\mathbf{y}) - h(\alpha)$$

  - For a proper choice of step-size, admits sufficient decrease
    $$\hat{\varphi}^\alpha(\mathbf{y}^k) - \hat{\varphi}^\alpha(\mathbf{y}^{k+1}) \geq \frac{1}{2} \min \left\{ S^\alpha(\mathbf{y}^k), \frac{(S^\alpha(\mathbf{y}^k))^2}{(L_f + 2)L_f D_\alpha^2} \right\},$$
    where D_α is an upper bound on the diameter of $\mathrm{dom}(g) \times \mathrm{Lev}_\omega(\omega^*)$
  - Leads to $O(1/k)$ convergence.

Applying the algorithm to $\hat{\varphi}^\alpha$.

Generalized Conditional Gradient

- Generalized Conditional Gradeint (GCG) composite functions:
  - GCG step
$$\mathbf{y}^{k+1} = \mathbf{y}^k + t_k(\mathbf{p}(\mathbf{y}^k) - \mathbf{y}^k),$$
    where
$$\mathbf{p}(\mathbf{y}) = (\mathbf{p}_1(\mathbf{y}), \mathbf{p}_2(\mathbf{y})) \begin{cases} \mathbf{p}_1(\mathbf{y}) & = \arg\min\{\langle \nabla f(\mathbf{y}_1) + 2(\mathbf{y}_1 - \mathbf{y}_2), \mathbf{p}_1 \rangle + g(\mathbf{p}_1)\} \\ \mathbf{p}_2(\mathbf{y}) & = \underset{\mathbf{p}_2 \in \mathrm{Lev}_\omega(\alpha)}{\arg\min} \{\langle 2(\mathbf{y}_2 - \mathbf{y}_1), \mathbf{p}_2 \rangle\} \end{cases}$$
  - Bound on the optimality gap:
$$S^\alpha(\mathbf{y}) = \langle \nabla \hat{f}(\mathbf{y}), \mathbf{y} - \mathbf{p}(\mathbf{y}) \rangle + \hat{g}^\alpha(\mathbf{y}) - \hat{g}^\alpha(\mathbf{p}(\mathbf{y})) \geq \hat{\varphi}^\alpha(\mathbf{y}) - \hat{\varphi}^\alpha(\mathbf{p}(\mathbf{y})) \geq \hat{\varphi}^\alpha(\mathbf{y}) - h(\alpha)$$
  - For a proper choice of step-size, admits sufficient decrease
$$\hat{\varphi}^\alpha(\mathbf{y}^k) - \hat{\varphi}^\alpha(\mathbf{y}^{k+1}) \geq \frac{1}{2} \min\left\{ S^\alpha(\mathbf{y}^k), \frac{(S^\alpha(\mathbf{y}^k))^2}{(L_f + 2) L_f D_\alpha^2} \right\},$$
    where D_α is an upper bound on the diameter of $\mathrm{dom}(g) \times \mathrm{Lev}_\omega(\omega^*)$
  - Leads to $O(1/k)$ convergence. Is this convergence rate maintained?

Applying the algorithm to $\hat{\varphi}^\alpha$.

GCG based Approximation Oracle

Algorithm 3: A GCG based Approximation Algorithm

Input: Initial point $\mathbf{y}^0 \equiv \mathbf{x} \in C \cap \text{Lev}_\omega(\alpha)$, $\alpha \leq \omega^*$, $\bar{\varphi} \geq \varphi^*$, ε,
for $j = 0, 1, 2, \ldots$ **do**
  Apply one iteration of GCG at point $\mathbf{y}^j$ to obtain $\mathbf{y}^{j+1}$ and $S^\alpha(\mathbf{y}^j)$.
  if $\hat{\varphi}^\alpha(\mathbf{y}^j) - \bar{\varphi} \leq \varepsilon$ **then**
    Exit algorithm and return $(\rho, \mathbf{y}) = (0, \mathbf{y}^j)$
  end if
  if $\hat{\varphi}^\alpha(\mathbf{y}^j) - \bar{\varphi} - S^\alpha(\mathbf{y}^j) \geq \frac{\varepsilon}{2}$ **then**
    Exit and return $(\rho, \mathbf{y}) = (\hat{\varphi}^\alpha(\mathbf{y}^j) - \bar{\varphi} - S^\alpha(\mathbf{y}^j), \mathbf{y}^j)$
    $\left(\text{Note that } \frac{\varepsilon}{2} \leq \rho = \hat{\varphi}^\alpha(\mathbf{y}^j) - \bar{\varphi} - S^\alpha(\mathbf{y}^j) \leq \hat{\varphi}^\alpha(\mathbf{y}^j) - \bar{\varphi} - \hat{\varphi}^\alpha(\mathbf{y}^j) + h(\alpha) = h(\alpha) - \bar{\varphi} \leq h(\alpha) - \varphi^*\right)$
  end if
end for

GCG based Approximation Oracle

Algorithm 4: A GCG based Approximation Algorithm

Input: Initial point $\mathbf{y}^0 \equiv \mathbf{x} \in C \cap \mathrm{Lev}_\omega(\alpha)$, $\alpha \leq \omega^*$, $\bar{\varphi} \geq \varphi^*$, ε,
for $j = 0, 1, 2, \ldots$ **do**
  Apply one iteration of GCG at point $\mathbf{y}^j$ to obtain $\mathbf{y}^{j+1}$ and $S^\alpha(\mathbf{y}^j)$.
  if $\hat{\varphi}^\alpha(\mathbf{y}^j) - \bar{\varphi} \leq \varepsilon$ **then**
    Exit algorithm and return $(\rho, \mathbf{y}) = (0, \mathbf{y}^j)$
  end if
  if $\hat{\varphi}^\alpha(\mathbf{y}^j) - \bar{\varphi} - S^\alpha(\mathbf{y}^j) \geq \frac{\varepsilon}{2}$ **then**
    Exit and return $(\rho, \mathbf{y}) = (\hat{\varphi}^\alpha(\mathbf{y}^j) - \bar{\varphi} - S^\alpha(\mathbf{y}^j), \mathbf{y}^j)$
    $\left(\text{Note that } \frac{\varepsilon}{2} \leq \rho = \hat{\varphi}^\alpha(\mathbf{y}^j) - \bar{\varphi} - S^\alpha(\mathbf{y}^j) \leq \hat{\varphi}^\alpha(\mathbf{y}^j) - \bar{\varphi} - \hat{\varphi}^\alpha(\mathbf{y}^j) + h(\alpha) = h(\alpha) - \bar{\varphi} \leq h(\alpha) - \varphi^*\right)$
  end if
end for

Theorem

During a run of ITALEX using the GCG based approximation oracle, the total number of GCG iterations (inner iterations) is at most $K + N$, where $K = O(1/\varepsilon)$ and N is the number of calls to the expansion oracle (outer iterations).

Flexibility

- For the above oracle implementation the inner iteration complexity is $K + N = O(1/\varepsilon)$.

Flexibility

- For the above oracle implementation the inner iteration complexity is $K + N = O(1/\varepsilon)$.
- L_f can be approximated locally. [Pedregosa et al. 2020]

Flexibility

- For the above oracle implementation the inner iteration complexity is $K + N = O(1/\varepsilon)$.
- L_f can be approximated locally. [Pedregosa et al. 2020]
- Our methodology is more general and other oracle implementations may be considered.

Flexibility

- For the above oracle implementation the inner iteration complexity is $K + N = O(1/\varepsilon)$.
- L_f can be approximated locally. [Pedregosa et al. 2020]
- Our methodology is more general and other oracle implementations may be considered.
- Specifically, instead of GCG we can use the proximal gradient (PG) method and get similar guarantees.

Flexibility

- For the above oracle implementation the inner iteration complexity is $K + N = O(1/\varepsilon)$.
- L_f can be approximated locally. [Pedregosa et al. 2020]
- Our methodology is more general and other oracle implementations may be considered.
- Specifically, instead of GCG we can use the proximal gradient (PG) method and get similar guarantees.
- On one hand, we note that $S^{\alpha}(\mathbf{y})$ is not computed during the run of PG.

Flexibility

- For the above oracle implementation the inner iteration complexity is $K + N = O(1/\varepsilon)$.
- L_f can be approximated locally. [Pedregosa et al. 2020]
- Our methodology is more general and other oracle implementations may be considered.
- Specifically, instead of GCG we can use the proximal gradient (PG) method and get similar guarantees.
- On one hand, we note that $S^{\alpha}(\mathbf{y})$ is not computed during the run of PG.
- On the other hand, PG generates a decreasing sequence and does not require $\mathrm{dom}(g)$ to be compact.

Numerical experiments

Numerical experiments

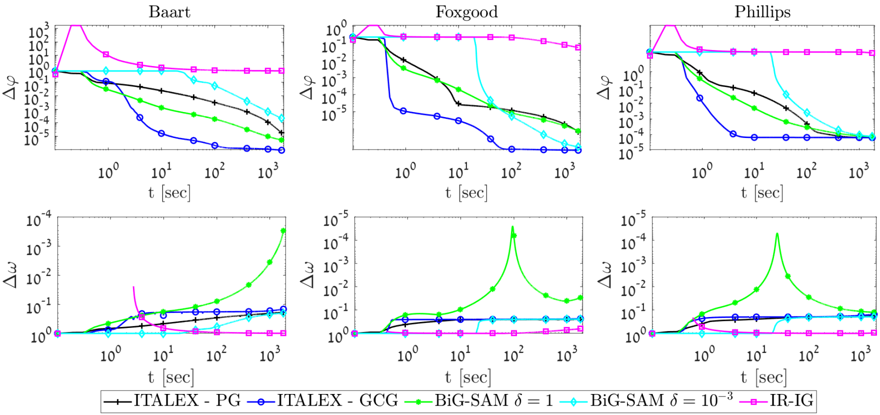
- Given a sparse $\mathbf{x}^{\text{true}} \in \mathbb{R}^{1000}$ we create $\mathbf{b} = \mathbf{A}\mathbf{x}^{\text{true}} + \nu$.

Numerical experiments

- Given a sparse $\mathbf{x}^{\text{true}} \in \mathbb{R}^{1000}$ we create $\mathbf{b} = \mathbf{A}\mathbf{x}^{\text{true}} + \nu$.
- $\varphi = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$, $\omega(\mathbf{x}) = \|\mathbf{x}\|_1 + \rho\|\mathbf{x}\|_2^2$ with $\rho = 0.5$.

Numerical experiments

- Given a sparse $\mathbf{x}^{\text{true}} \in \mathbb{R}^{1000}$ we create $\mathbf{b} = \mathbf{A}\mathbf{x}^{\text{true}} + \nu$.
- $\varphi = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$, $\omega(\mathbf{x}) = \|\mathbf{x}\|_1 + \rho\|\mathbf{x}\|_2^2$ with $\rho = 0.5$.
- Averaged over 100 simulations of ν.



Numerical experiments -

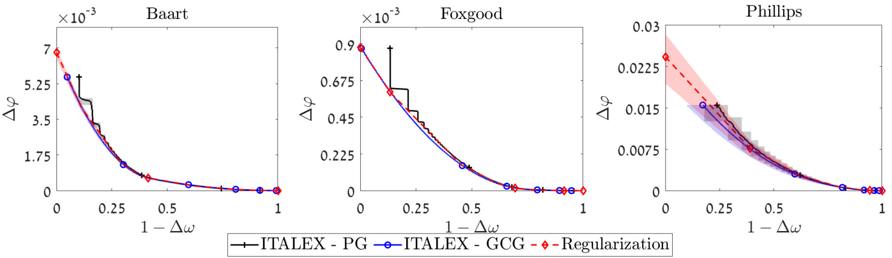
- $\omega(\mathbf{x}) = \|\mathbf{x}\|_1$.
- PG faster than GCG

Numerical experiments -

- $\omega(\mathbf{x}) = \|\mathbf{x}\|_1$.
- PG faster than GCG
- Benchmark: iterative regularization where with regularization parameter $\frac{1}{2^\ell}\lambda_{\max}(\mathbf{A}^\top\mathbf{A})$ for $\ell \in [15]$.

Numerical experiments -

- $\omega(\mathbf{x}) = \|\mathbf{x}\|_1$.
- PG faster than GCG
- Benchmark: iterative regularization where with regularization parameter $\frac{1}{2^\ell}\lambda_{\max}(\mathbf{A}^\top\mathbf{A})$ for $\ell \in [15]$.



Summary

- ITALEX has proven $O(1/k)$ feasibility and $O(1/\sqrt{k})$ optimality rate for (BLP) with norm-like ω.

Summary

- ITALEX has proven $O(1/k)$ feasibility and $O(1/\sqrt{k})$ optimality rate for (BLP) with norm-like ω.
- More on ITALEX project:
  - ε does not need to be fixed in advance.
  - Getting super-optimal solutions when $g = 0$.
  - Accelerated rates under additional conditions on φ and ω.
  - Allowing outer function of the form $\omega(\mathbf{Lx})$.

Thank you for listening!

Bibliography I

[AY19] M. Amini and F. Yousefian. "An Iterative Regularized Incremental Projected Subgradient Method for a Class of Bilevel Optimization Problems". In: 2019 American Control Conference (ACC). July 2019, pp. 4069–4074. DOI: https://doi.org/10.23919/ACC.2019.8814637.

[BS14] Amir Beck and Shoham Sabach. "A first order method for finding minimal norm-like solutions of convex optimization problems". In: Math Program 147.1 (Oct. 2014), pp. 25–46. DOI: https://doi.org/10.1007/s10107-013-0708-2.

[DP20] Joydeep Dutta and Tanushree Pandit. "Algorithms for Simple Bilevel Programming". In: Bilevel Optimization: Advances and Next Challenges. Ed. by Stephan Dempe and Alain Zemkoho. Cham: Springer International Publishing, 2020, pp. 253–291. DOI: https://doi.org/10.1007/978-3-030-52119-6_9.

[LP98] Adrian S. Lewis and Jong-Shi Pang. "Error Bounds for Convex Inequality Systems". In: Generalized Convexity, Generalized Monotonicity. Ed. by Martinez-Legaz J. Crouziex J. and Volle M. Kluwer Academic Publishers, 1998. Chap. 3, pp. 75–110.

[Ped+20] Fabian Pedregosa et al. "Linearly convergent Frank-Wolfe with backtracking line-search". In: International Conference on Artificial Intelligence and Statistics. PMLR. 2020, pp. 1–10.

[Sol07] M. Solodov. "An Explicit Descent Method for Bilevel Convex Optimization". In: J Convex Anal 14 (Jan. 2007), pp. 227–237.

Bibliography II

[SS17] Shoham Sabach and Shimrit Shtern. "A First Order Method for Solving Convex Bilevel Optimization Problems". In: SIAM J. Optim. 27.2 (2017), pp. 640–660. DOI: https://doi.org/10.1137/16M105592X.

[SVZ21] Yekini Shehu, Phan Tu Vuong, and Alain Zemkoho. "An inertial extrapolation method for convex simple bilevel optimization". In: Optim Methods Softw 36.1 (2021), pp. 1–19. DOI: https://doi.org/10.1080/10556788.2019.1619729.

Proximal Gradient

Proximal Gradient

- Proximal Gradient for composite functions:
  - PG step $\mathbf{y}^{k+1} = T_{L_f}(\mathbf{y}^k)$ where

$$T_{L_f}(\mathbf{y}) = \arg\min_{\mathbf{u}} \left\{ g(\mathbf{x}) + \frac{L_f}{2} \|\mathbf{y} - \frac{1}{L_f} \nabla f(\mathbf{y}) - \mathbf{u}\|^2 \right\}$$

Proximal Gradient

- Proximal Gradient for composite functions:
  - PG step $\mathbf{y}^{k+1} = T_{L_f}(\mathbf{y}^k)$ where
  $$T_{L_f}(\mathbf{y}) = \arg\min_{\mathbf{u}} \left\{ g(\mathbf{x}) + \frac{L_f}{2}\left\|\mathbf{y} - \frac{1}{L_f}\nabla f(\mathbf{y}) - \mathbf{u}\right\|^2 \right\}$$
  - Assuming that $\text{Lev}_\varphi(\varphi(\mathbf{y})) \leq D(\mathbf{y})$:
  $$\tilde{S}(\mathbf{y}) = 2\max\left\{ \varphi(\mathbf{y}) - \varphi\left(T_{L_f}(\mathbf{y})\right), \sqrt{\frac{L_f}{2}D(\mathbf{y})^2(\varphi(\mathbf{y}) - \varphi\left(T_{L_f}(\mathbf{y})\right))} \right\}$$

Proximal Gradient

- Proximal Gradient for composite functions:
  - PG step $\mathbf{y}^{k+1} = T_{L_f}(\mathbf{y}^k)$ where
  $$T_{L_f}(\mathbf{y}) = \arg\min_{\mathbf{u}} \left\{ g(\mathbf{x}) + \frac{L_f}{2}\left\|\mathbf{y} - \frac{1}{L_f}\nabla f(\mathbf{y}) - \mathbf{u}\right\|^2 \right\}$$
  - Assuming that $\mathrm{Lev}_{\varphi}(\varphi(\mathbf{y})) \leq D(\mathbf{y})$:

$$\tilde{S}(\mathbf{y}) = 2\max\left\{ \varphi(\mathbf{y}) - \varphi\left(T_{L_f}(\mathbf{y})\right), \sqrt{\frac{L_f}{2} D(\mathbf{y})^2 (\varphi(\mathbf{y}) - \varphi\left(T_{L_f}(\mathbf{y})\right))} \right\}$$
$$\geq S_{D(\mathbf{y})}(\mathbf{y})$$

Proximal Gradient

- Proximal Gradient for composite functions:
  - PG step $\mathbf{y}^{k+1} = T_{L_f}(\mathbf{y}^k)$ where
$$T_{L_f}(\mathbf{y}) = \arg\min_{\mathbf{u}} \left\{ g(\mathbf{x}) + \frac{L_f}{2}\left\|\mathbf{y} - \frac{1}{L_f}\nabla f(\mathbf{y}) - \mathbf{u}\right\|^2 \right\}$$
  - Assuming that $\text{Lev}_{\varphi}(\varphi(\mathbf{y})) \leq D(\mathbf{y})$:
$$\begin{aligned}\tilde{S}(\mathbf{y}) &= 2\max\left\{ \varphi(\mathbf{y}) - \varphi\left(T_{L_f}(\mathbf{y})\right), \sqrt{\frac{L_f}{2}D(\mathbf{y})^2(\varphi(\mathbf{y}) - \varphi\left(T_{L_f}(\mathbf{y})\right))} \right\} \\ &\geq S_{D(\mathbf{y})}(\mathbf{y})\end{aligned}$$

Lemma

$\tilde{S}(\mathbf{y})$ *satisfies:*

- $\tilde{S}(\mathbf{y}) \geq \varphi(\mathbf{y}) - \varphi^*$
- $\varphi(\mathbf{y}) - \varphi(T_{L_f}(\mathbf{y})) \geq \frac{1}{2}\min\left\{\tilde{S}(\mathbf{y}), \frac{2\tilde{S}(\mathbf{y})^2}{L_f D(\mathbf{y})^2}\right\}$

Proximal Gradient

- Proximal Gradient for composite functions:
  - PG step $\mathbf{y}^{k+1} = T^{\alpha}_{L_f+2}(\mathbf{y}^k)$ where
    $$T^{\alpha}_{L_f+2}(\mathbf{y}) = \arg\min_{\mathbf{u}} \left\{ \hat{g}^{\alpha}(\mathbf{x}) + \frac{L_f+2}{2}\left\|\mathbf{y} - \frac{1}{L_f+2}\nabla \hat{f}(\mathbf{y}) - \mathbf{u}\right\|^2 \right\}$$
  - ~~Assuming that~~ $\text{Lev}_{\hat{\varphi}^{\alpha}}(\hat{\varphi}^{\alpha}(\mathbf{y})) \leq D_{\alpha}(\mathbf{y})$:

$$\tilde{S}^{\alpha}(\mathbf{y}) = 2\max\left\{ \hat{\varphi}^{\alpha}(\mathbf{y}) - \hat{\varphi}^{\alpha}\left(T^{\alpha}_{L_f+2}(\mathbf{y})\right), \sqrt{\frac{L_f+2}{2} D_{\alpha}(\mathbf{y})^2 (\hat{\varphi}^{\alpha}(\mathbf{y}) - \hat{\varphi}^{\alpha}\left(T^{\alpha}_{L_f+2}(\mathbf{y})\right))} \right\}$$
$$\geq S^{\alpha}_{D(\mathbf{y})}(\mathbf{y})$$

Lemma

$\tilde{S}^{\alpha}(\mathbf{y})$ *satisfies:*

- $\tilde{S}^{\alpha}(\mathbf{y}) \geq \hat{\varphi}^{\alpha}(\mathbf{y}) - h(\alpha)$
- $\hat{\varphi}^{\alpha}(\mathbf{y}) - \hat{\varphi}^{\alpha}(T^{\alpha}_{L_f+2}(\mathbf{y})) \geq \frac{1}{2}\min\left\{ \tilde{S}^{\alpha}(\mathbf{y}), \frac{2\tilde{S}(\mathbf{y})^2}{(L_f+2)D_{\alpha}(\mathbf{y})^2} \right\}$

Proximal Gradient

- Proximal Gradient for composite functions:
  - PG step $\mathbf{y}^{k+1} = T^{\alpha}_{L_f+2}(\mathbf{y}^k)$ where

$$T^{\alpha}_{L_f+2}(\mathbf{y}) = (T^{\alpha}_1(\mathbf{y}), T^{\alpha}_2(\mathbf{y})), \begin{cases} T^{\alpha}_1(\mathbf{y}) & = \text{prox}_{\frac{1}{L_f+2}g}\left(\mathbf{y}_1 - \frac{1}{L_f+2}(\nabla f(\mathbf{y}_1) + 2(\mathbf{y}_1 - \mathbf{y}_2))\right) \\ T^{\alpha}_2(\mathbf{y}) & = \text{Proj}_{\text{Lev}_\omega(\alpha)}\left(\frac{L_f\mathbf{y}_2+2\mathbf{y}_1}{L_f+2}\right) \end{cases}$$

  - ~~Assuming that~~ $\text{Lev}_{\hat{\varphi}^\alpha}(\hat{\varphi}^\alpha(\mathbf{y})) \leq D_\alpha(\mathbf{y})$:

$$\tilde{S}^\alpha(\mathbf{y}) = 2\max\left\{\hat{\varphi}^\alpha(\mathbf{y}) - \hat{\varphi}^\alpha\left(T^{\alpha}_{L_f+2}(\mathbf{y})\right), \sqrt{\frac{L_f+2}{2}D_\alpha(\mathbf{y})^2(\hat{\varphi}^\alpha(\mathbf{y}) - \hat{\varphi}^\alpha\left(T^{\alpha}_{L_f+2}(\mathbf{y})\right))}\right\}$$

$$\geq S^\alpha_{D(\mathbf{y})}(\mathbf{y})$$

Lemma

$\tilde{S}^\alpha(\mathbf{y})$ satisfies:

- $\tilde{S}^\alpha(\mathbf{y}) \geq \hat{\varphi}^\alpha(\mathbf{y}) - h(\alpha)$
- $\hat{\varphi}^\alpha(\mathbf{y}) - \hat{\varphi}^\alpha(T^{\alpha}_{L_f+2}(\mathbf{y})) \geq \frac{1}{2}\min\left\{\tilde{S}^\alpha(\mathbf{y}), \frac{2\tilde{S}(\mathbf{y})^2}{(L_f+2)D_\alpha(\mathbf{y})^2}\right\}$

Proximal Gradient

- Proximal Gradient for composite functions:
  - PG step $\mathbf{y}^{k+1} = T^{\alpha}_{L_f+2}(\mathbf{y}^k)$ where

$$T^{\alpha}_{L_f+2}(\mathbf{y}) = (T^{\alpha}_1(\mathbf{y}), T^{\alpha}_2(\mathbf{y})), \begin{cases} T^{\alpha}_1(\mathbf{y}) & = \text{prox}_{\frac{1}{L_f+2}g}\left(\mathbf{y}_1 - \frac{1}{L_f+2}(\nabla f(\mathbf{y}_1) + 2(\mathbf{y}_1 - \mathbf{y}_2))\right) \\ T^{\alpha}_2(\mathbf{y}) & = \text{Proj}_{\text{Lev}_\omega(\alpha)}\left(\frac{L_f\mathbf{y}_2+2\mathbf{y}_1}{L_f+2}\right) \end{cases}$$

  - ~~Assuming that~~ $\text{Lev}_{\hat{\varphi}^\alpha}(\hat{\varphi}^\alpha(\mathbf{y})) \leq D_\alpha(\mathbf{y})$:

$$\tilde{S}^\alpha(\mathbf{y}) = 2\max\left\{\hat{\varphi}^\alpha(\mathbf{y}) - \hat{\varphi}^\alpha\left(T^{\alpha}_{L_f+2}(\mathbf{y})\right), \sqrt{\frac{L_f+2}{2}D_\alpha(\mathbf{y})^2(\hat{\varphi}^\alpha(\mathbf{y}) - \hat{\varphi}^\alpha\left(T^{\alpha}_{L_f+2}(\mathbf{y})\right))}\right\}$$
$$\geq S^{\alpha}_{D(\mathbf{y})}(\mathbf{y})$$

Lemma

$\tilde{S}^\alpha(\mathbf{y})$ satisfies:

- $\tilde{S}^\alpha(\mathbf{y}) \geq \hat{\varphi}^\alpha(\mathbf{y}) - h(\alpha)$ *- enables early stopping*
- $\hat{\varphi}^\alpha(\mathbf{y}) - \hat{\varphi}^\alpha(T^{\alpha}_{L_f+2}(\mathbf{y})) \geq \frac{1}{2}\min\left\{\tilde{S}^\alpha(\mathbf{y}), \frac{2\tilde{S}^\alpha(\mathbf{y})^2}{(L_f+2)D_\alpha(\mathbf{y})^2}\right\}$

  $O(\frac{1}{\epsilon})$ *convergence*