Smoothing and First Order Methods
A Unified Framework

Marc Teboulle

School of Mathematical Sciences
Tel Aviv University

Joint work with Amir Beck, Technion, Haifa, Israel

SIAM Imaging Sciences Conference, May 19-22, 2012 – Philadelphia, USA

NSO via Smooth Approximation

- A well known methodology for designing solution techniques to non smooth optimization (NSO for short) problems is to replace the original problem by a sequence of *approximating smooth* problems.
- Hopefully smoothed problem can be solved more efficiently than by using classical NSO-schemes such as subgradient/bundle type methods .

NSO via Smooth Approximation

- A well known methodology for designing solution techniques to non smooth optimization (NSO for short) problems is to replace the original problem by a sequence of *approximating smooth* problems.
- Hopefully smoothed problem can be solved more efficiently than by using classical NSO-schemes such as subgradient/bundle type methods .
- **Basic idea:** Given a convex nonsmooth objective J, transform

$$(NSO) \quad \min_{\mathbf{x}\in X} J(\mathbf{x}) \longrightarrow (SO) \quad \min_{\mathbf{x}\in X} J_\mu(\mathbf{x})$$

  $\mu > 0$ is a *smoothing parameter* s.t. $\lim_{\mu\to 0^+} J_\mu(\mathbf{x}) = J(\mathbf{x})$.
- Many approaches: some earlier works, Moreau (65), Bertsekas (75), BenTal-Teboulle (89), Attouch-Wets (89).....

NSO via Smooth Approximation

- A well known methodology for designing solution techniques to non smooth optimization (NSO for short) problems is to replace the original problem by a sequence of *approximating smooth* problems.
- Hopefully smoothed problem can be solved more efficiently than by using classical NSO-schemes such as subgradient/bundle type methods .
- **Basic idea:** Given a convex nonsmooth objective J, transform

$$(NSO) \quad \min_{\mathbf{x}\in X} J(\mathbf{x}) \longrightarrow (SO) \quad \min_{\mathbf{x}\in X} J_\mu(\mathbf{x})$$

  $\mu > 0$ is a *smoothing parameter* s.t. $\lim_{\mu\to 0^+} J_\mu(\mathbf{x}) = J(\mathbf{x})$.
- Many approaches: some earlier works, Moreau (65), Bertsekas (75), BenTal-Teboulle (89), Attouch-Wets (89).....
- **Good News:** Smooth problems can be solved efficiently

NSO via Smooth Approximation

- A well known methodology for designing solution techniques to non smooth optimization (NSO for short) problems is to replace the original problem by a sequence of *approximating smooth* problems.
- Hopefully smoothed problem can be solved more efficiently than by using classical NSO-schemes such as subgradient/bundle type methods .
- **Basic idea:** Given a convex nonsmooth objective J, transform

$$(NSO) \quad \min_{\mathbf{x}\in X} J(\mathbf{x}) \longrightarrow (SO) \quad \min_{\mathbf{x}\in X} J_\mu(\mathbf{x})$$

  $\mu > 0$ is a *smoothing parameter* s.t. $\lim_{\mu\to 0^+} J_\mu(\mathbf{x}) = J(\mathbf{x})$.
- Many approaches: some earlier works, Moreau (65), Bertsekas (75), BenTal-Teboulle (89), Attouch-Wets (89).....
- **Good News:** Smooth problems can be solved efficiently
- **Bad News:**

NSO via Smooth Approximation

- A well known methodology for designing solution techniques to non smooth optimization (NSO for short) problems is to replace the original problem by a sequence of *approximating smooth* problems.
- Hopefully smoothed problem can be solved more efficiently than by using classical NSO-schemes such as subgradient/bundle type methods .
- **Basic idea:** Given a convex nonsmooth objective J, transform

$$(NSO) \quad \min_{\mathbf{x}\in X} J(\mathbf{x}) \longrightarrow (SO) \quad \min_{\mathbf{x}\in X} J_\mu(\mathbf{x})$$

  $\mu > 0$ is a *smoothing parameter* s.t. $\lim_{\mu\to 0^+} J_\mu(\mathbf{x}) = J(\mathbf{x})$.
- Many approaches: some earlier works, Moreau (65), Bertsekas (75), BenTal-Teboulle (89), Attouch-Wets (89).....
- **Good News:** Smooth problems can be solved efficiently
- **Bad News:**
  - One needs to solve a *sequence* J_μ with $\mu \searrow 0$

NSO via Smooth Approximation

- A well known methodology for designing solution techniques to non smooth optimization (NSO for short) problems is to replace the original problem by a sequence of *approximating smooth* problems.
- Hopefully smoothed problem can be solved more efficiently than by using classical NSO-schemes such as subgradient/bundle type methods .
- **Basic idea:** Given a convex nonsmooth objective J, transform

$$(NSO) \quad \min_{\mathbf{x}\in X} J(\mathbf{x}) \longrightarrow (SO) \quad \min_{\mathbf{x}\in X} J_\mu(\mathbf{x})$$

  $\mu > 0$ is a *smoothing parameter* s.t. $\lim_{\mu\to 0^+} J_\mu(\mathbf{x}) = J(\mathbf{x})$.
- Many approaches: some earlier works, Moreau (65), Bertsekas (75), BenTal-Teboulle (89), Attouch-Wets (89).....
- **Good News:** Smooth problems can be solved efficiently
- **Bad News:**
  - One needs to solve a *sequence* J_μ with $\mu \searrow 0$
  - Convergence results/approximate solutions are **not** for original NSO J, but for the *approximated* smoothed problem J_μ.

Black Box Schemes versus Specially Structured Problems

To compute an ε-optimal solution via black-box schemes:

1. Convex Nonsmooth: $\sim O(1/\varepsilon^2)$ iterations
2. Convex Smooth: $\sim O(1/\varepsilon)$ iterations..and even $\sim O(1/\sqrt{\varepsilon})$ iters., N(83).

Black Box Schemes versus Specially Structured Problems

To compute an ε-optimal solution via black-box schemes:

1. Convex Nonsmooth: $\sim O(1/\varepsilon^2)$ iterations
2. Convex Smooth: $\sim O(1/\varepsilon)$ iterations..and even $\sim O(1/\sqrt{\varepsilon})$ iters., N(83).

Structure Helps: For *nonsmooth composite* models $\min_{\mathbf{x}}\{f(\mathbf{x}) + g(\mathbf{x})\}$ with f smooth, we can also get schemes $\sim O(1/\sqrt{\varepsilon})$ iterations,[Nesterov (07), FISTA-Beck-Teboulle(09)],i.e., like there is **no nonsmooth term**.

Black Box Schemes versus Specially Structured Problems

To compute an ε-optimal solution via black-box schemes:

1. Convex Nonsmooth: $\sim O(1/\varepsilon^2)$ iterations
2. Convex Smooth: $\sim O(1/\varepsilon)$ iterations..and even $\sim O(1/\sqrt{\varepsilon})$ iters., N(83).

Structure Helps: For *nonsmooth composite* models $\min_{\mathbf{x}}\{f(\mathbf{x}) + g(\mathbf{x})\}$ with f smooth, we can also get schemes $\sim O(1/\sqrt{\varepsilon})$ iterations,[Nesterov (07), FISTA-Beck-Teboulle(09)],i.e., like there is **no nonsmooth term**.

.....**But what happens if *both* f,g are nonsmooth?....**

Black Box Schemes versus Specially Structured Problems

To compute an ε-optimal solution via black-box schemes:

1. Convex Nonsmooth: $\sim O(1/\varepsilon^2)$ iterations
2. Convex Smooth: $\sim O(1/\varepsilon)$ iterations..and even $\sim O(1/\sqrt{\varepsilon})$ iters., N(83).

Structure Helps: For *nonsmooth composite* models $\min_{\mathbf{x}}\{f(\mathbf{x}) + g(\mathbf{x})\}$ with f smooth, we can also get schemes $\sim O(1/\sqrt{\varepsilon})$ iterations,[Nesterov (07), FISTA-Beck-Teboulle(09)],i.e., like there is **no nonsmooth term**.

.....**But what happens if *both* f,g are nonsmooth?....**

Question 1: Can we solve NSO in $\sim O(1/\varepsilon)$ iterations via **one fixed smoothed** counterpart problem?

Black Box Schemes versus Specially Structured Problems

To compute an ε-optimal solution via black-box schemes:

1. Convex Nonsmooth: $\sim O(1/\varepsilon^2)$ iterations
2. Convex Smooth: $\sim O(1/\varepsilon)$ iterations..and even $\sim O(1/\sqrt{\varepsilon})$ iters., N(83).

Structure Helps: For *nonsmooth composite* models $\min_{\mathbf{x}}\{f(\mathbf{x}) + g(\mathbf{x})\}$ with f smooth, we can also get schemes $\sim O(1/\sqrt{\varepsilon})$ iterations,[Nesterov (07), FISTA-Beck-Teboulle(09)],i.e., like there is **no nonsmooth term**.

.....**But what happens if *both* f,g are nonsmooth?....**

Question 1: Can we solve NSO in $\sim O(1/\varepsilon)$ iterations via **one fixed smoothed** counterpart problem?

Answer 1: Yes, Nesterov-2005 $\min_{\mathbf{x}} \mathcal{F}(\mathbf{x})$

- For a particular class of NSO problems with **– a max-structure–**
- Applying a *peculiar* fast gradient scheme for the smoothed problem

$$\begin{array}{rcl} \mathcal{F}(\mathbf{x}) & := & \max\{\langle \mathbf{u}, A\mathbf{x}\rangle - \phi(\mathbf{u}) : \mathbf{u} \in Q\}, \ Q \text{ convex compact} \\ \phi & & \text{convex continuous on } Q \subset \operatorname{dom}\phi; \ A : \mathbb{E} \to \mathbb{V} \text{ a linear map} \end{array}$$

Black Box Schemes versus Specially Structured Problems

To compute an ε-optimal solution via black-box schemes:

1. Convex Nonsmooth: $\sim O(1/\varepsilon^2)$ iterations
2. Convex Smooth: $\sim O(1/\varepsilon)$ iterations..and even $\sim O(1/\sqrt{\varepsilon})$ iters., N(83).

Structure Helps: For *nonsmooth composite* models $\min_{\mathbf{x}}\{f(\mathbf{x}) + g(\mathbf{x})\}$ with f smooth, we can also get schemes $\sim O(1/\sqrt{\varepsilon})$ iterations,[Nesterov (07), FISTA-Beck-Teboulle(09)],i.e., like there is **no nonsmooth term**.

.....**But what happens if *both* f,g are nonsmooth?....**

Question 1: Can we solve NSO in $\sim O(1/\varepsilon)$ iterations via **one fixed smoothed** counterpart problem?

Answer 1: Yes, Nesterov-2005 $\min_{\mathbf{x}} \mathcal{F}(\mathbf{x})$

- For a particular class of NSO problems with **– a max-structure–**
- Applying a *peculiar* fast gradient scheme for the smoothed problem

$$\begin{array}{rcl} \mathcal{F}(\mathbf{x}) & := & \max\{\langle \mathbf{u}, A\mathbf{x}\rangle - \phi(\mathbf{u}) : \mathbf{u} \in Q\}, \ Q \text{ convex compact} \\ \phi & & \text{convex continuous on } Q \subset \operatorname{dom}\phi; \ A : \mathbb{E} \to \mathbb{V} \text{ a linear map} \end{array}$$

Question 2: Can we extend such a result ***independently*** of the structure ***and*** of the fast scheme involved?

Black Box Schemes versus Specially Structured Problems

To compute an ε-optimal solution via black-box schemes:

1. Convex Nonsmooth: $\sim O(1/\varepsilon^2)$ iterations
2. Convex Smooth: $\sim O(1/\varepsilon)$ iterations..and even $\sim O(1/\sqrt{\varepsilon})$ iters., N(83).

Structure Helps: For *nonsmooth composite* models $\min_{\mathbf{x}}\{f(\mathbf{x}) + g(\mathbf{x})\}$ with f smooth, we can also get schemes $\sim O(1/\sqrt{\varepsilon})$ iterations,[Nesterov (07), FISTA-Beck-Teboulle(09)],i.e., like there is **no nonsmooth term**.

.....**But what happens if *both* f,g are nonsmooth?....**

Question 1: Can we solve NSO in $\sim O(1/\varepsilon)$ iterations via **one fixed smoothed** counterpart problem?

Answer 1: Yes, Nesterov-2005 $\min_{\mathbf{x}} \mathcal{F}(\mathbf{x})$

- For a particular class of NSO problems with **– a max-structure–**
- Applying a *peculiar* fast gradient scheme for the smoothed problem

$$\begin{array}{rcl} \mathcal{F}(\mathbf{x}) & := & \max\{\langle \mathbf{u}, A\mathbf{x}\rangle - \phi(\mathbf{u}) : \mathbf{u} \in Q\}, \ Q \text{ convex compact} \\ \phi & & \text{convex continuous on } Q \subset \operatorname{dom}\phi; \ A : \mathbb{E} \to \mathbb{V} \text{ a linear map} \end{array}$$

Question 2: Can we extend such a result ***independently*** of the structure ***and*** of the fast scheme involved?

Answer 2: ...Yes... this talk..!..

Outline

Goal – Independently of

1. Structure of the convex nonsmooth objective
2. A given fast first order iterative scheme,

An ε-optimal solution of the ***original*** nonsmooth problem can be obtained with an $O(\varepsilon^{-1})$ efficiency estimate, by solving ***one*** adequate smoothed counterpart.

Outline

Goal – Independently of

1. Structure of the convex nonsmooth objective
2. A given fast first order iterative scheme,

An ε-optimal solution of the ***original*** nonsmooth problem can be obtained with an $O(\varepsilon^{-1})$ efficiency estimate, by solving ***one*** adequate smoothed counterpart.

- A Broad Concept: *Smoothable Functions* for general (convex) functions
- The Partial Smoothing Optimization Approach
- A Fast Scheme and Smoothing Method with Complexity $O(\varepsilon^{-1})$
- Generating Smoothable Convex Functions
- To Smooth or not to Smooth?

Some Notations

- Finite dimensional normed vector spaces are denoted by $\mathbb{E}, \mathbb{F}, \mathbb{V}$ etc..
- For a vector space $\mathbb{E}$, the endowed norm is denoted by $\|\cdot\|_{\mathbb{E}}$
- The space of linear functionals is denoted by $\mathbb{E}^*$
- The dual norm is denoted by either $\|\cdot\|_{\mathbb{E}}^*$ or $\|\cdot\|_{\mathbb{E}^*}$ defined via

$$\|\mathbf{x}\|_{\mathbb{E}}^* = \|\mathbf{x}\|_{\mathbb{E}^*} = \max\{\langle \mathbf{u}, \mathbf{x}\rangle : \|\mathbf{u}\|_{\mathbb{E}} \leq 1\} \text{ for any } \mathbf{x} \in \mathbb{E}^*$$

- The norm of a linear transformation $A : \mathbb{E} \to \mathbb{V}$, where $\mathbb{E}$ and $\mathbb{V}$ are finite dimensional vector spaces with endowed norms $\|\cdot\|_{\mathbb{E}}$ and $\|\cdot\|_{\mathbb{V}}$ respectively is given by

$$\|A\|_{\mathbb{E},\mathbb{V}} = \max\{\|A\mathbf{x}\|_{\mathbb{V}} : \|\mathbf{x}\|_{\mathbb{E}} = 1\}.$$

- $C_L^{1,1}(S)$ class of functions continuously differentiable with L-Lipschitz gradient on convex S.
- The conjugate of $f : \mathbb{E} \to (-\infty, \infty]$ at $\mathbf{y}$:

$$f^*(\mathbf{y}) = \sup_{\mathbf{x}\in\mathbb{E}} \{\langle \mathbf{x}, \mathbf{y}\rangle - f(\mathbf{x})\} = \sup \{\langle \mathbf{x}, \mathbf{y}\rangle - f(\mathbf{x}) : \ \mathbf{x} \in \operatorname{dom} f\}.$$

Smoothable Convex Functions

We begin by defining the concept of a *smoothable* function and the corresponding *smooth approximation* of a given nonsmooth convex function.

Definition (smoothable functions)

Let $g : \mathbb{E} \to (-\infty, \infty]$ be a closed and proper convex function and let $X \subseteq \operatorname{dom} g$ be a closed convex set.

The function g is called "(α, β, K)-*smoothable*" over X if there exist β_1, β_2 satisfying $\beta_1 + \beta_2 = \beta > 0$ such that for every $\mu > 0$ there exists a continuously differentiable convex function $g_\mu : X \to (-\infty, \infty)$ such that

(i) $g(\mathbf{x}) - \beta_1 \mu \leq g_\mu(\mathbf{x}) \leq g(\mathbf{x}) + \beta_2 \mu$ for every $\mathbf{x} \in X$.

(ii) $g_\mu \in C_L^{1,1}(X)$ with $L \leq K + \frac{\alpha}{\mu}$.i.e., there exists $K \geq 0, \alpha > 0$, s.t.

$$\|\nabla g_\mu(\mathbf{x}) - \nabla g_\mu(\mathbf{y})\|^* \leq \left(K + \frac{\alpha}{\mu}\right) \|\mathbf{x} - \mathbf{y}\| \text{ for every } \mathbf{x}, \mathbf{y} \in X.$$

Smoothable Convex Functions

We begin by defining the concept of a *smoothable* function and the corresponding *smooth approximation* of a given nonsmooth convex function.

Definition (smoothable functions)

Let $g : \mathbb{E} \to (-\infty, \infty]$ be a closed and proper convex function and let $X \subseteq \operatorname{dom} g$ be a closed convex set.

The function g is called "(α, β, K)-*smoothable*" over X if there exist β_1, β_2 satisfying $\beta_1 + \beta_2 = \beta > 0$ such that for every $\mu > 0$ there exists a continuously differentiable convex function $g_\mu : X \to (-\infty, \infty)$ such that

(i) $g(\mathbf{x}) - \beta_1 \mu \leq g_\mu(\mathbf{x}) \leq g(\mathbf{x}) + \beta_2 \mu$ for every $\mathbf{x} \in X$.

(ii) $g_\mu \in C_L^{1,1}(X)$ with $L \leq K + \frac{\alpha}{\mu}$.i.e., there exists $K \geq 0, \alpha > 0$, s.t.

$$\|\nabla g_\mu(\mathbf{x}) - \nabla g_\mu(\mathbf{y})\|^* \leq \left(K + \frac{\alpha}{\mu}\right) \|\mathbf{x} - \mathbf{y}\| \text{ for every } \mathbf{x}, \mathbf{y} \in X.$$

- The function g_μ is called a "$\mu-$*smooth approximation*" of g over X with parameters (α, β, K).
- If a function is smoothable over the entire vector space $\mathbb{E}$, then it will just be called (α, β, K)-smoothable.

Basic Properties/Operations Apply For Smoothable Functions

- The choice of the decomposition of β as $\beta_1 + \beta_2$ is arbitrary
- The nonnegative linear combination of smoothable functions is another smoothable function:
  if g_1 and g_2 are (α_1, β_1, K_1) and (α_2, β_2, K_2) smoothable, then $\gamma_1 g_1 + \gamma_2 g_2 (\gamma_1, \gamma_2 \geq 0)$ is $(\gamma_1\alpha_1 + \gamma_2\alpha_2, \gamma_1\beta_1 + \gamma_2\beta_2, \gamma_1 K_1 + \gamma_2 K_2)$ smoothable.
- Linear transformation on the variables of a smoothable function yields another smoothable function (with appropriate scaling of parameters):
  if g is (α, β, K) smoothable, then $q(x) = g(Ax + b)$ is $(\alpha\|A\|^2, \beta, K\|A\|^2)$ smoothable.

There exist various ways to generate **smoothable functions**, this issue will be addressed later.

The Nonsmooth Optimization Model

We are interested in solving the convex problem (G) given by

$$(G) \quad H^* = \min\{H(\mathbf{x}) \equiv g(\mathbf{x}) + f(\mathbf{x}) + h(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}$$

where the assumptions on the underlying functions are:

- $h : \mathbb{E} \to (-\infty, \infty]$ closed proper convex function which is subdifferentiable over its domain which is denoted by $X = \operatorname{dom} h$.
- $f : X \to (-\infty, \infty)$ is a continuously differentiable function over X whose gradient is Lipschitz with constant L_f
- $g : X \to (\infty, \infty]$ is a (α, β, K)-smoothable function over X.

Problem (G) is rich enough to cover many interesting generic optimization models by appropriate choices of (f, g, h).

Why 3 Functions?

The Approach: Partial Smoothing

$$(G) \quad H^* = \min\{H(\mathbf{x}) \equiv g(\mathbf{x}) + f(\mathbf{x}) + h(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}$$

We will invoke what we call **partial smoothing**, namely, only the function g is smoothed.

The Approach: Partial Smoothing

$$(G) \quad H^* = \min\{H(\mathbf{x}) \equiv g(\mathbf{x}) + f(\mathbf{x}) + h(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}$$

We will invoke what we call **partial smoothing**, namely, only the function g is smoothed. The *partially smoothed* problem is thus the composite NSO:

$$(G_\mu) \quad H^*_\mu = \min\{H_\mu(\mathbf{x}) \equiv g_\mu(\mathbf{x}) + f(\mathbf{x}) + h(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\},$$

where g_μ is a μ-smooth approximation of g over X with parameters (α, β, K) for an appropriately chosen μ.

The Motivation of 3 functions is Double:

The Approach: Partial Smoothing

$$(G) \quad H^* = \min\{H(\mathbf{x}) \equiv g(\mathbf{x}) + f(\mathbf{x}) + h(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}$$

We will invoke what we call **partial smoothing**, namely, only the function g is smoothed. The *partially smoothed* problem is thus the composite NSO:

$$(G_\mu) \quad H^*_\mu = \min\{H_\mu(\mathbf{x}) \equiv g_\mu(\mathbf{x}) + f(\mathbf{x}) + h(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\},$$

where g_μ is a μ-smooth approximation of g over X with parameters (α, β, K) for an appropriately chosen μ.

The Motivation of 3 functions is Double:

- First, as previously mentioned, *composite NSO* can be solved by fast FOM with $O(1/k^2)$ rate of convergence.

The Approach: Partial Smoothing

$$(G) \quad H^* = \min\{H(\mathbf{x}) \equiv g(\mathbf{x}) + f(\mathbf{x}) + h(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}$$

We will invoke what we call **partial smoothing**, namely, only the function g is smoothed. The *partially smoothed* problem is thus the composite NSO:

$$(G_\mu) \quad H^*_\mu = \min\{H_\mu(\mathbf{x}) \equiv g_\mu(\mathbf{x}) + f(\mathbf{x}) + h(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\},$$

where g_μ is a μ-smooth approximation of g over X with parameters (α, β, K) for an appropriately chosen μ.

The Motivation of 3 functions is Double:

- First, as previously mentioned, *composite NSO* can be solved by fast FOM with $O(1/k^2)$ rate of convergence.
- Secondly, in many applications, one of the nonsmooth term in the model plays a key role in describing a desirable property of $\mathbf{x}$ which otherwise **could be destroyed by smoothing!** (e.g., sparsity).

Formal Setting and Definitions

To make our approach general and independent of the iterative schemes involved we introduce two formal definitions:

1. The input convex optimization model
2. A Fast Iterative Method $\mathcal{M}$

Input Convex Optimization Model

The *partially smoothed* problem

$$(G_\mu) \quad H_\mu^* = \min\{H_\mu(\mathbf{x}) \equiv \underbrace{g_\mu(\mathbf{x}) + f(\mathbf{x})}_{F(\mathbf{x}) \in C^{1,1}} + h(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\},$$

The idea is now to be able to use **any** adequate algorithm for solving (G_μ) which admits a composite form,

$$(C) \qquad \min\{F(\mathbf{x}) + h(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}$$

This problem will be called the ***input convex optimization model*** and is characterized by the triplet (F, h, L_F)

- $h : \mathbb{E} \to (-\infty, \infty]$ is proper, closed convex function, and subdifferentiable over $\operatorname{dom} h$.
- F is a convex function in $C^{1,1}_{L_F}(\operatorname{dom} h)$

Problem (G_μ) is of this form with $F := f + g_\mu$.

A Formal Fast Iterative Method $\mathcal{M}$

$$(C) \qquad \min\{D(\mathbf{x}) \equiv F(\mathbf{x}) + h(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}$$

Definition (Fast Iterative Method $\mathcal{M}$)

Let (F, h, L_F) be a given input convex optimization model with an optimal solution $\mathbf{x}^*$, and let $\mathbf{x}_0 \in \mathbb{E}$ be any given initial point.

An iterative method $\mathcal{M}$ for solving problem (C) is called a *fast* method with constant $0 < \Lambda < \infty$, (which possibly depends on $(\mathbf{x}_0, \mathbf{x}^*)$), if it generates a sequence $\{\mathbf{x}_k\}_{k \geq 0}$ satisfying for all $k \geq 1$,

$$\spadesuit \qquad D(\mathbf{x}_k) - D^* \leq \frac{L_F \Lambda}{k^2}.$$

A Formal Fast Iterative Method $\mathcal{M}$

$$(C) \qquad \min\{D(\mathbf{x}) \equiv F(\mathbf{x}) + h(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}$$

Definition (Fast Iterative Method $\mathcal{M}$)

Let (F, h, L_F) be a given input convex optimization model with an optimal solution $\mathbf{x}^*$, and let $\mathbf{x}_0 \in \mathbb{E}$ be any given initial point.

An iterative method $\mathcal{M}$ for solving problem (C) is called a *fast* method with constant $0 < \Lambda < \infty$, (which possibly depends on $(\mathbf{x}_0, \mathbf{x}^*)$), if it generates a sequence $\{\mathbf{x}_k\}_{k \geq 0}$ satisfying for all $k \geq 1$,

$$\spadesuit \qquad D(\mathbf{x}_k) - D^* \leq \frac{L_F \Lambda}{k^2}.$$

- Algorithmic steps of $\mathcal{M}$ play *no role*. **Any** method with $\spadesuit$ will do.

A Formal Fast Iterative Method $\mathcal{M}$

$$(C) \qquad \min\{D(\mathbf{x}) \equiv F(\mathbf{x}) + h(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}$$

Definition (Fast Iterative Method $\mathcal{M}$)

Let (F, h, L_F) be a given input convex optimization model with an optimal solution $\mathbf{x}^*$, and let $\mathbf{x}_0 \in \mathbb{E}$ be any given initial point.

An iterative method $\mathcal{M}$ for solving problem (C) is called a *fast* method with constant $0 < \Lambda < \infty$, (which possibly depends on $(\mathbf{x}_0, \mathbf{x}^*)$), if it generates a sequence $\{\mathbf{x}_k\}_{k \geq 0}$ satisfying for all $k \geq 1$,

$$\spadesuit \qquad D(\mathbf{x}_k) - D^* \leq \frac{L_F \Lambda}{k^2}.$$

- Algorithmic steps of $\mathcal{M}$ play *no role*. **Any** method with $\spadesuit$ will do.

A Formal Fast Iterative Method $\mathcal{M}$

$$(C) \qquad \min\{D(\mathbf{x}) \equiv F(\mathbf{x}) + h(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}$$

Definition (Fast Iterative Method $\mathcal{M}$)

Let (F, h, L_F) be a given input convex optimization model with an optimal solution $\mathbf{x}^*$, and let $\mathbf{x}_0 \in \mathbb{E}$ be any given initial point.

An iterative method $\mathcal{M}$ for solving problem (C) is called a *fast* method with constant $0 < \Lambda < \infty$, (which possibly depends on $(\mathbf{x}_0, \mathbf{x}^*)$), if it generates a sequence $\{\mathbf{x}_k\}_{k \geq 0}$ satisfying for all $k \geq 1$,

$$\spadesuit \qquad D(\mathbf{x}_k) - D^* \leq \frac{L_F \Lambda}{k^2}.$$

- Algorithmic steps of $\mathcal{M}$ play *no role*. **Any** method with $\spadesuit$ will do.
- Thanks to the concept of smoothable functions, we now establish that applying *any* fast method $\mathcal{M}$ on the *partially smoothed* problem (G_μ), there exists an *explicit* smoothing parameter μ, such that an ε optimal solution of **the original nonsmooth problem** can be obtained in no more than $O(1/\varepsilon)$ iterations.

An $O(1/\varepsilon)$ for The Original Nonsmooth Optimization Problem

$$(G) \quad H^* = \min_{\mathbf{x}}\{H(\mathbf{x}) \equiv g(\mathbf{x}) + f(\mathbf{x}) + h(\mathbf{x})\}, \quad (G_\mu) \quad \min_{\mathbf{x}}\{g_\mu(\mathbf{x}) + f(\mathbf{x}) + h(\mathbf{x})\}$$

Theorem

Let $\{\mathbf{x}_k\}$ be the sequence generated by a fast iterative method $\mathcal{M}$ when applied to problem (G_μ), that is, to the input optimization problem $(f + g_\mu, h, L_{f+g_\mu})$. Suppose that the smoothing parameter is chosen as

$$\mu = \sqrt{\frac{\alpha}{\beta}} \frac{\varepsilon}{\sqrt{\alpha\beta} + \sqrt{\alpha\beta + (L_f + K)\varepsilon}}.$$

Then for

$$k \geq 2\sqrt{\alpha\beta\Lambda}\frac{1}{\varepsilon} + \sqrt{(L_f + K)\Lambda}\frac{1}{\sqrt{\varepsilon}},$$

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

An $O(1/\varepsilon)$ for The Original Nonsmooth Optimization Problem

$$(G) \quad H^* = \min_{\mathbf{x}}\{H(\mathbf{x}) \equiv g(\mathbf{x}) + f(\mathbf{x}) + h(\mathbf{x})\}, \quad (G_\mu) \quad \min_{\mathbf{x}}\{g_\mu(\mathbf{x}) + f(\mathbf{x}) + h(\mathbf{x})\}$$

Theorem

Let $\{\mathbf{x}_k\}$ be the sequence generated by a fast iterative method $\mathcal{M}$ when applied to problem (G_μ), that is, to the input optimization problem $(f + g_\mu, h, L_{f+g_\mu})$. Suppose that the smoothing parameter is chosen as

$$\mu = \sqrt{\frac{\alpha}{\beta}} \frac{\varepsilon}{\sqrt{\alpha\beta} + \sqrt{\alpha\beta + (L_f + K)\varepsilon}}.$$

Then for

$$k \geq 2\sqrt{\alpha\beta\Lambda}\frac{1}{\varepsilon} + \sqrt{(L_f + K)\Lambda}\frac{1}{\sqrt{\varepsilon}},$$

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

Remarks

- The smoothing parameter μ *does not depend* on the constant Λ of the method.
- As usual, Λ appears in the complexity bound to find the number of iters.
- If $\mathrm{dom}\, h$ is bounded, one can show that $\Lambda \in (0, \infty)$. But, even when $\mathrm{dom}\, h$ is not bounded, this works, as μ is *independent* of Λ *and* the method $\mathcal{M}$.

Smoothing Convex Functions

Nondifferentiable convex functions can be approximated by smooth functions by various techniques.

1. The so-called proximal map introduced by Moreau (1964)
2. Asymptotic (recession) functions

Building on these two fundamental convex analytic tools we propose a natural and unifying framework to smooth a general class of nonsmooth convex functions.

The Moreau Proximal Smoothing

One of the most popular approaches in the Euclidean setting (that is, when $\mathbb{E}$ is an Euclidean space with norm $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$) is the celebrated Moreau proximal approximation yields a family of approximations $\{g_\mu^{\rm px}\}_{\mu>0}$ via

$$g_\mu^{\rm px}(\mathbf{x}) = \inf_{\mathbf{u}\in\mathbb{E}} \left\{ g(\mathbf{u}) + \frac{1}{2\mu}\|\mathbf{u}-\mathbf{x}\|^2 \right\}, \quad g : \mathbb{E} \to (-\infty, \infty] \text{ closed and proper convex.}$$

As proven by Moreau (1965) for any $\mu > 0$:

1. $g_\mu^{\rm px}$ is convex continuous and finite-valued
2. $g_\mu^{\rm px}$ is differentiable with gradient $\nabla g_\mu^{\rm px}$ which is Lipschitz continuous with Lipschitz constant $1/\mu$.

The Moreau Proximal Smoothing

One of the most popular approaches in the Euclidean setting (that is, when $\mathbb{E}$ is an Euclidean space with norm $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$) is the celebrated Moreau proximal approximation yields a family of approximations $\{g_\mu^{\rm px}\}_{\mu>0}$ via

$$g_\mu^{\rm px}(\mathbf{x}) = \inf_{\mathbf{u}\in\mathbb{E}} \left\{ g(\mathbf{u}) + \frac{1}{2\mu}\|\mathbf{u}-\mathbf{x}\|^2 \right\}, \quad g : \mathbb{E} \to (-\infty, \infty] \text{ closed and proper convex.}$$

As proven by Moreau (1965) for any $\mu > 0$:

1. $g_\mu^{\rm px}$ is convex continuous and finite-valued
2. $g_\mu^{\rm px}$ is differentiable with gradient $\nabla g_\mu^{\rm px}$ which is Lipschitz continuous with Lipschitz constant $1/\mu$.

A less popular, representation of Moreau proximal smoothing is through its dual formulation,

$$g_\mu^{\rm px}(\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{E}^*} \left\{ \langle \mathbf{y}, \mathbf{x} \rangle - g^*(\mathbf{y}) - \frac{\mu}{2}\|\mathbf{y}\|^2 \right\}.$$

In essence, the above shows that Moreau smoothing is a natural tool to also smooth *conjugate functions*. This provides the starting point of the forthcoming results.

A Natural Analogue to Moreau Approximation

- The Moreau approximation $g_\mu^{\rm px}$ is the so-called infimal convolution of f with the quadratic function $q(\mathbf{x}) = \frac{1}{2\mu}\|\mathbf{x}\|^2$, i.e.,

$$g_\mu^{\rm px}(\mathbf{x}) = \inf_{\mathbf{x}_1,\mathbf{x}_2}\{g(\mathbf{x}_1)+q(\mathbf{x}_2) : \ \mathbf{x}_1+\mathbf{x}_2 = \mathbf{x}\} = \inf_{\mathbf{u}\in\mathbb{E}}\{g(\mathbf{u})+q(\mathbf{x}-\mathbf{u})\} \equiv (g\square q)(\mathbf{x}).$$

A Natural Analogue to Moreau Approximation

- The Moreau approximation $g_\mu^{\rm px}$ is the so-called infimal convolution of f with the quadratic function $q(\mathbf{x}) = \frac{1}{2\mu}\|\mathbf{x}\|^2$, i.e.,

$$g_\mu^{\rm px}(\mathbf{x}) = \inf_{\mathbf{x}_1,\mathbf{x}_2}\{g(\mathbf{x}_1)+q(\mathbf{x}_2) : \mathbf{x}_1+\mathbf{x}_2 = \mathbf{x}\} = \inf_{\mathbf{u}\in\mathbb{E}}\{g(\mathbf{u})+q(\mathbf{x}-\mathbf{u})\} \equiv (g\square q)(\mathbf{x}).$$

- The infimal convolution operation remains the key player in smoothing *any* convex function without requiring any a-priori special structure of the function to be smoothed, like e.g., *its dual form*.

A Natural Analogue to Moreau Approximation

- The Moreau approximation g_μ^{px} is the so-called infimal convolution of f with the quadratic function $q(\mathbf{x}) = \frac{1}{2\mu}\|\mathbf{x}\|^2$, i.e.,

$$g_\mu^{\mathrm{px}}(\mathbf{x}) = \inf_{\mathbf{x}_1,\mathbf{x}_2}\{g(\mathbf{x}_1)+q(\mathbf{x}_2) : \mathbf{x}_1+\mathbf{x}_2 = \mathbf{x}\} = \inf_{\mathbf{u}\in\mathbb{E}}\{g(\mathbf{u})+q(\mathbf{x}-\mathbf{u})\} \equiv (g\square q)(\mathbf{x}).$$

- The infimal convolution operation remains the key player in smoothing *any* convex function without requiring any a-priori special structure of the function to be smoothed, like e.g., *its dual form*.
- Imitating Moreau, we define the inf-conv μ-smooth approximation of a convex function via an infimal convolution with a $C^{1,1}$ convex function.

A Natural Analogue to Moreau Approximation

- The Moreau approximation $g_\mu^{\rm px}$ is the so-called infimal convolution of f with the quadratic function $q(\mathbf{x}) = \frac{1}{2\mu}\|\mathbf{x}\|^2$, i.e.,

$$g_\mu^{\rm px}(\mathbf{x}) = \inf_{\mathbf{x}_1,\mathbf{x}_2}\{g(\mathbf{x}_1)+q(\mathbf{x}_2) : \mathbf{x}_1+\mathbf{x}_2 = \mathbf{x}\} = \inf_{\mathbf{u}\in\mathbb{E}}\{g(\mathbf{u})+q(\mathbf{x}-\mathbf{u})\} \equiv (g\square q)(\mathbf{x}).$$

- The infimal convolution operation remains the key player in smoothing *any* convex function without requiring any a-priori special structure of the function to be smoothed, like e.g., *its dual form*.
- Imitating Moreau, we define the inf-conv μ-smooth approximation of a convex function via an infimal convolution with a $C^{1,1}$ convex function.

Definition (inf-conv μ-smooth approximation)

Let $g : \mathbb{E} \to (-\infty, \infty]$ be a closed proper convex function.
Let $\omega : \mathbb{E} \to \mathbb{R}$ be a $C^{1,1}_{1/\sigma}$ convex function ($\sigma > 0$).
Suppose that for any $\mu > 0$ and any $\mathbf{x} \in \mathbb{E}$, the following infimal convolution is finite:

$$g_\mu^{\rm ic}(\mathbf{x}) = \inf_{\mathbf{u}\in\mathbb{E}}\left\{g(\mathbf{u}) + \mu\omega\left(\frac{\mathbf{x}-\mathbf{u}}{\mu}\right)\right\} = (g\square\omega_\mu)(\mathbf{x}),\ \omega_\mu(\cdot) \equiv \mu\omega\left(\frac{\cdot}{\mu}\right).$$

Then $g_\mu^{\rm ic}$ is called the *inf-conv μ-smooth approximation* of g.

Main properties of the inf-conv μ-smooth approximation

Theorem

Consider the setting in the Definition for g_μ^{ic}. Then,

(a) *the following "dual" formulation for g_μ^{ic} holds:*

$$g_\mu^{\mathrm{ic}}(\mathbf{x}) = (g^* + \omega_\mu^*)^*(\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{E}^*} \{\langle \mathbf{y}, \mathbf{x}\rangle - g^*(\mathbf{y}) - \mu\omega^*(\mathbf{y})\}\,.$$

(b) *g_μ^{ic} is differentiable and with gradient $\nabla g_\mu^{\mathrm{ic}}$ which is Lipschitz with constant $\frac{1}{\sigma\mu}$.*

(c) *Let $\mathbf{x} \in \mathbb{E}$. Suppose that the minimum in μ-infconv is attained at the point $\mathbf{u}_\mu(\mathbf{x})$. Then*

$$\nabla g_\mu^{\mathrm{ic}}(\mathbf{x}) = \nabla\omega\left(\frac{\mathbf{x} - \mathbf{u}_\mu(\mathbf{x})}{\mu}\right) = \nabla\omega_\mu(\mathbf{x} - \mathbf{u}_\mu(\mathbf{x})).$$

Main properties of the inf-conv μ-smooth approximation

Theorem

Consider the setting in the Definition for g_μ^{ic}. Then,

(a) *the following "dual" formulation for g_μ^{ic} holds:*

$$g_\mu^{\mathrm{ic}}(\mathbf{x}) = (g^* + \omega_\mu^*)^*(\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{E}^*} \{\langle \mathbf{y}, \mathbf{x}\rangle - g^*(\mathbf{y}) - \mu\omega^*(\mathbf{y})\}.$$

(b) *g_μ^{ic} is differentiable and with gradient $\nabla g_\mu^{\mathrm{ic}}$ which is Lipschitz with constant $\frac{1}{\sigma\mu}$.*

(c) *Let $\mathbf{x} \in \mathbb{E}$. Suppose that the minimum in μ-infconv is attained at the point $\mathbf{u}_\mu(\mathbf{x})$. Then*

$$\nabla g_\mu^{\mathrm{ic}}(\mathbf{x}) = \nabla\omega\left(\frac{\mathbf{x} - \mathbf{u}_\mu(\mathbf{x})}{\mu}\right) = \nabla\omega_\mu(\mathbf{x} - \mathbf{u}_\mu(\mathbf{x})).$$

- This shows that g_μ^{ic} satisfies property (ii) of the Smoothable function Definition.
- It remains to detect conditions under which g_μ^{ic} also satisfies property (i) of the Definition of a smoothable function.

Any Convex Function is Smoothable

Under our setting, any convex function is essentially smoothable. More precisely:

Any Convex Function is Smoothable

Under our setting, any convex function is essentially smoothable. More precisely:

Theorem

Let $X \subseteq \mathbb{E}$ be a closed convex set. Suppose that g is subdifferentiable over X. Then for any $\mu > 0$ and $\mathbf{x} \in X$ the following holds:

$$g(\mathbf{x}) - \mu\omega^*(\gamma_{\mathbf{x}}) \leq g_\mu^{\mathrm{ic}}(\mathbf{x}) \leq g(\mathbf{x}) + \mu\omega(\mathbf{0}),$$

where $\gamma_{\mathbf{x}} \in \partial g(\mathbf{x})$ is a subgradient of g at $\mathbf{x}$.

Moreover, if

$$D[g, \omega^*] = \sup_{\mathbf{x} \in X} \sup_{\mathbf{d} \in \partial g(\mathbf{x})} \omega^*(\mathbf{d}) < \infty$$

then for any $\mu > 0$, g_μ^{ic} is a μ-smooth approximation of g over X with parameters

$$\left(\frac{1}{\sigma}, D[g, \omega^*] + \omega(\mathbf{0}), 0\right).$$

Recovering Nesterov's Smoothing (05) via μ-Inf-Conv

Nesterov's Smoothing is based on a non-Euclidean extension of the *Dual* Moreau smoothing approximation.

- The class of nonsmooth convex functions considered by (N05):

$$q(\mathbf{x}) = \max\left\{\langle \mathbf{u}, A\mathbf{x}\rangle - \phi(\mathbf{u}) : \mathbf{u} \in Q\right\}, \quad \mathbf{x} \in \mathbb{E},\ A : \mathbb{E} \to \mathbb{V} \textit{ linear map}.$$

- **Result (N05)**: The convex $q_\mu \in C^{1,1}_{L_\mu}(\mathbb{E})$; $L_\mu = \|A\|^2/\sigma\mu$:

$$\clubsuit \qquad q_\mu(\mathbf{x}) = \max\left\{\langle \mathbf{u}, A\mathbf{x}\rangle - \phi(\mathbf{u}) - \mu d(\mathbf{u}) : \mathbf{u} \in Q\right\}, \quad \mathbf{x} \in \mathbb{E},$$

where $d(\cdot)$ is a σ-strongly convex continuous function over the compact set $C \subseteq \operatorname{dom} d$ and $D := \max_{\mathbf{x}\in Q} d(\mathbf{x})$.

Recovering Nesterov's Smoothing (05) via μ-Inf-Conv

Nesterov's Smoothing is based on a non-Euclidean extension of the *Dual* Moreau smoothing approximation.

- The class of nonsmooth convex functions considered by (N05):

$$q(\mathbf{x}) = \max\{\langle \mathbf{u}, A\mathbf{x}\rangle - \phi(\mathbf{u}) : \mathbf{u} \in Q\}, \quad \mathbf{x} \in \mathbb{E},\ A : \mathbb{E} \to \mathbb{V} \textit{ linear map}.$$

- **Result (N05)**: The convex $q_\mu \in C^{1,1}_{L_\mu}(\mathbb{E})$; $L_\mu = \|A\|^2/\sigma\mu$:

$$\clubsuit \qquad q_\mu(\mathbf{x}) = \max\{\langle \mathbf{u}, A\mathbf{x}\rangle - \phi(\mathbf{u}) - \mu d(\mathbf{u}) : \mathbf{u} \in Q\}, \quad \mathbf{x} \in \mathbb{E},$$

  where $d(\cdot)$ is a σ-strongly convex continuous function over the compact set $C \subseteq \operatorname{dom} d$ and $D := \max_{\mathbf{x}\in Q} d(\mathbf{x})$.

- The function to be smoothed can be written as

$$q(\mathbf{x}) = g(A\mathbf{x})$$

  where $g := (\tilde{\phi})^*$, and $\tilde{\phi} := \phi + \delta_Q$

- With $\omega := (d + \delta_Q)^*$, invoking our results we obtain

$$q_\mu(\mathbf{x}) = g^{\mathrm{ic}}_\mu(A\mathbf{x}) \in C^{1,1}_{L_\mu}(\mathbb{E});\ L_\mu = \|A\|^2/\sigma\mu$$

  and q_μ is a μ-smooth approximation of q with parameters $\left(\frac{\|A\|^2}{\sigma}, D, 0\right)$

- Result of (N05) is recovered and **No need to restrict to *max-structure***

Two Well-Known Examples of Smooth Approximation

Example (Euclidean norm function)

Let $\mathbb{E} = \mathbb{R}^n$ endowed with the l_2 norm $\|\cdot\| = \|\cdot\|_2$. Consider the setting

$$g(\mathbf{x}) = \|\mathbf{x}\|, X = \mathbb{E}, \omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2 \;\Rightarrow\; \omega(\mathbf{0}) = 0, D[g, \omega] = \frac{1}{2}$$

Here $g_\mu^{\mathrm{ic}} \equiv g_\mu^{\mathrm{px}}$, is a μ-smooth approximation of g (over $\mathbb{R}^n$) with parameters $\left(1, \frac{1}{2}, 0\right)$, given by

$$g_\mu^{\mathrm{px}}(\mathbf{x}) = \min_{\mathbf{u}} \left\{ \|\mathbf{u}\| + \frac{1}{2\mu}\|\mathbf{u} - \mathbf{x}\|^2 \right\} = \begin{cases} \frac{\|\mathbf{x}\|^2}{2\mu} & \|\mathbf{x}\| \le \mu, \\ \|\mathbf{x}\| - \frac{\mu}{2} & \text{else}, \end{cases}$$

Two Well-Known Examples of Smooth Approximation

Example (Euclidean norm function)

Let $\mathbb{E} = \mathbb{R}^n$ endowed with the l_2 norm $\|\cdot\| = \|\cdot\|_2$. Consider the setting

$$g(\mathbf{x}) = \|\mathbf{x}\|, X = \mathbb{E}, \omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2 \;\Rightarrow\; \omega(\mathbf{0}) = 0, D[g,\omega] = \frac{1}{2}$$

Here $g_\mu^{\mathrm{ic}} \equiv g_\mu^{\mathrm{px}}$, is a μ-smooth approximation of g (over $\mathbb{R}^n$) with parameters $\left(1, \frac{1}{2}, 0\right)$, given by

$$g_\mu^{\mathrm{px}}(\mathbf{x}) = \min_{\mathbf{u}} \left\{ \|\mathbf{u}\| + \frac{1}{2\mu}\|\mathbf{u}-\mathbf{x}\|^2 \right\} = \begin{cases} \frac{\|\mathbf{x}\|^2}{2\mu} & \|\mathbf{x}\| \le \mu, \\ \|\mathbf{x}\| - \frac{\mu}{2} & \text{else}, \end{cases}$$

Example (l_1 norm -The Huber Fucntion)

With same vector space $\mathbb{E}$ and ω, let $g(\mathbf{x}) = \|\mathbf{x}\|_1$. Then $\omega(\mathbf{0}) = 0, D[g,\omega] = \frac{n}{2}$ and hence g_μ^{px}, is the sum of Huber functions on each of the components:

$$g_\mu^{\mathrm{px}}(\mathbf{x}) = \sum_{i=1}^{n} H_\mu(x_i), \quad \left(H_\mu(y) \equiv \begin{cases} \frac{y^2}{2\mu} & |y| \le \mu, \\ |y| - \frac{\mu}{2} & \text{else}, \end{cases} \right)$$

is a μ-smooth approximation of g (over $\mathbb{R}^n$) with parameters $\left(1, \frac{n}{2}, 0\right)$.

More examples, see B-T (2012) paper.

Smoothing via Asymptotic Functions

- Another general approach to smooth functions is via the concept of asymptotic (recession) functions, as introduced in BenTal-Teboulle (89).

Smoothing via Asymptotic Functions

- Another general approach to smooth functions is via the concept of asymptotic (recession) functions, as introduced in BenTal-Teboulle (89).
- Most optimization problems may be formulated as

$$(K) \qquad \inf\{r_\infty(f_1(\mathbf{x}),\ldots,f_m(\mathbf{x})) : \ \mathbf{x} \in \mathbb{E}\},$$

where r_∞ is the asymptotic function of some given function r.

Smoothing via Asymptotic Functions

- Another general approach to smooth functions is via the concept of asymptotic (recession) functions, as introduced in BenTal-Teboulle (89).
- Most optimization problems may be formulated as

$$(K) \qquad \inf\{r_\infty(f_1(\mathbf{x}),\ldots,f_m(\mathbf{x})) : \ \mathbf{x} \in \mathbb{E}\},$$

  where r_∞ is the asymptotic function of some given function r.
- **Two Recalls**: Let $r : \mathbb{E} \to (-\infty, +\infty]$ proper, closed and convex. Then,

$$\begin{aligned} r_\infty(d) &= \lim_{\mu \to 0^+} \left\{ r_\mu(\mathbf{d}) := \mu r\left(\frac{\mathbf{d}}{\mu}\right) \right\} \text{ for every } \mathbf{d} \in \operatorname{dom} r \\ r_\infty &= \sigma_{\operatorname{dom} r^*}, \ \ \sigma \text{ is the support function} \end{aligned}$$

Smoothing via Asymptotic Functions

- Another general approach to smooth functions is via the concept of asymptotic (recession) functions, as introduced in BenTal-Teboulle (89).
- Most optimization problems may be formulated as

$$(K) \qquad \inf\{r_\infty(f_1(\mathbf{x}),\ldots,f_m(\mathbf{x})) : \ \mathbf{x}\in\mathbb{E}\},$$

  where r_∞ is the asymptotic function of some given function r.
- **Two Recalls**: Let $r:\mathbb{E}\to(-\infty,+\infty]$ proper, closed and convex. Then,

$$\begin{aligned} r_\infty(d) &= \lim_{\mu\to 0^+}\left\{r_\mu(\mathbf{d}) := \mu r\left(\frac{\mathbf{d}}{\mu}\right)\right\} \text{ for every } \mathbf{d}\in \operatorname{dom} r \\ r_\infty &= \sigma_{\operatorname{dom} r^*}, \ \ \sigma \text{ is the support function} \end{aligned}$$

- **Idea of BT (89):** connection of support and asymptotic functions, naturally suggest approximating problem (K) whereby r_∞ is replaced by r_μ.

Smoothing via Asymptotic Functions

- Another general approach to smooth functions is via the concept of asymptotic (recession) functions, as introduced in BenTal-Teboulle (89).
- Most optimization problems may be formulated as

$$(K) \qquad \inf\{r_\infty(f_1(\mathbf{x}),\ldots,f_m(\mathbf{x})) : \ \mathbf{x}\in\mathbb{E}\},$$

  where r_∞ is the asymptotic function of some given function r.
- **Two Recalls**: Let $r:\mathbb{E}\to(-\infty,+\infty]$ proper, closed and convex. Then,

$$\begin{aligned} r_\infty(d) &= \lim_{\mu\to 0^+}\left\{r_\mu(\mathbf{d}) := \mu r\left(\frac{\mathbf{d}}{\mu}\right)\right\} \text{ for every } \mathbf{d}\in \operatorname{dom} r \\ r_\infty &= \sigma_{\operatorname{dom} r^*}, \ \ \sigma \text{ is the support function} \end{aligned}$$

- **Idea of BT (89):** connection of support and asymptotic functions, naturally suggest approximating problem (K) whereby r_∞ is replaced by r_μ.
- **A shameless commercial!** For more results see: Auslender-Teboulle Book, Chapter 3, (2003).
- Here we show that that there exists an interesting close relation between the asymptotic function-based smoothing and the inf-conv μ-smooth approximation.

A Simple Formula for inf-conv μ-smooth approximation

The inf-conv μ-smooth approximation of a convex function has a simple and special structure when the function $g := \omega_\infty$.

A Simple Formula for inf-conv μ-smooth approximation

The inf-conv μ-smooth approximation of a convex function has a simple and special structure when the function $g := \omega_\infty$.

Assumption: (satisfied for many useful cases – for all examples here)

$$\text{For any } \mu > 0, \mu\omega\left(\frac{\mathbf{x}}{\mu}\right) \geq \omega_\infty(\mathbf{x}) \text{ for all } \mathbf{x}.$$

Theorem

Let $\omega : \mathbb{E} \to \mathbb{R}$ be a $C^{1,1}$ convex function with Lipschitz gradient constant $1/\sigma$, and let g be a convex finite-valued function over $\mathbb{E}$. Suppose the Assumption holds and let $g = \omega_\infty$. Then for any $\mu > 0$,

$$g_\mu^{\mathrm{ic}}(\mathbf{x}) = \mu\omega\left(\frac{\mathbf{x}}{\mu}\right) \quad \textit{for every } \mathbf{x} \in \mathbb{E}.$$

Moreover, the function g_μ^{ic} is a μ-smooth approximation of g with parameters $(\frac{1}{\sigma}, \omega(\mathbf{0}), 0)$.

Examples for Common Nonsmooth Problems

By Theorem, simply use $\mu\omega\left(\frac{\mathbf{x}}{\mu}\right)$ to build μ-smooth approximation of g.

$\omega(\mathbf{u})$	$g(\mathbf{x}) = \omega_\infty(\mathbf{x})$	g_μ–parameters
$\sqrt{1+\|\mathbf{u}\|^2}$	$\|\mathbf{x}\|$	$(1, 1, 0)$
$\sum_i \sqrt{1+u_i^2}$	$\|\mathbf{x}\|_1$	$(1, n, 0)$
$\log(\sum_i \exp(u_i))$	$\max_{1\le i\le m} x_i$	$(1, \log n, 0)$
same ω	$\max_{1\le i\le m} \mathbf{a}_i^T\mathbf{x} + b_i$	$(\|\mathbf{A}\|^2, \log m, 0)$

Maximum of Convex Functions

This covers a broad class of problems. For a given integer $m > 1$, consider the convex function

$$g(\mathbf{x}) = \max\{f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})\},$$

- $f \equiv (f_1, \ldots, f_m)$ are m differentiable convex functions over a compact convex set $X \subseteq \mathbb{E}$ with Lipschitz gradients over X with constants $L_{f_1}, \ldots, L_{f_m}$ respectively.
- The vector space $\mathbb{E}$ has a norm denoted by $\|\cdot\|_{\mathbb{E}}$

Maximum of Convex Functions

This covers a broad class of problems. For a given integer $m > 1$, consider the convex function

$$g(\mathbf{x}) = \max\{f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})\},$$

- $f \equiv (f_1, \ldots, f_m)$ are m differentiable convex functions over a compact convex set $X \subseteq \mathbb{E}$ with Lipschitz gradients over X with constants $L_{f_1}, \ldots, L_{f_m}$ respectively.
- The vector space $\mathbb{E}$ has a norm denoted by $\|\cdot\|_{\mathbb{E}}$
- Clearly, the function g can also be rewritten as

$$g(\mathbf{x}) = \max_{\lambda \in \Delta_m} \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) = \max_{\lambda \in \Delta_m} \langle \lambda, f(\mathbf{x}) \rangle$$

Maximum of Convex Functions

This covers a broad class of problems. For a given integer $m > 1$, consider the convex function

$$g(\mathbf{x}) = \max\{f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})\},$$

- $f \equiv (f_1, \ldots, f_m)$ are m differentiable convex functions over a compact convex set $X \subseteq \mathbb{E}$ with Lipschitz gradients over X with constants $L_{f_1}, \ldots, L_{f_m}$ respectively.
- The vector space $\mathbb{E}$ has a norm denoted by $\|\cdot\|_{\mathbb{E}}$
- Clearly, the function g can also be rewritten as

$$g(\mathbf{x}) = \max_{\lambda \in \Delta_m} \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) = \max_{\lambda \in \Delta_m} \langle \lambda, f(\mathbf{x}) \rangle$$

- It has the "max" structure....But..since $f_i(\cdot)$ are **nonlinear**, the smoothing framework of (N05) ***cannot*** be applied.

Smoothing Maximum of Convex Functions

Proposition

Let $f_1, \ldots, f_m$ be m continuously differentiable convex functions over a compact convex set $X \subseteq \mathbb{E}$ whose gradients are Lipschitz over X with constants $L_{f_1}, \ldots, L_{f_m}$ respectively. Let

$$g(\mathbf{x}) = \max\{f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})\}.$$

Then for every $\mu > 0$ the function

$$g_\mu(\mathbf{x}) = \mu \log\left(\sum_{i=1}^{m} e^{f_i(\mathbf{x})/\mu}\right)$$

is a μ-smooth approximation of g with parameters

$$\left(\max_{i=1,\ldots,m} M_{f_i}^2, \log(m), \max_{i=1,\ldots,m} L_{f_i}\right),$$

where $M_{f_i} := \max\left\{\|\nabla f_i(\mathbf{x})\|_{\mathbb{E}}^* : \mathbf{x} \in X\right\}, \quad i = 1, \ldots, m.$

To Smooth or not to Smooth?

We compare *partial* smoothing versus *full* smoothing.

To Smooth or not to Smooth?

We compare *partial* smoothing versus *full* smoothing.

Consider the following $l_1 - l_1$ least fitting problem on the vector space $\mathbb{R}^n$ endowed with the norm $\|\cdot\|_1$:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ M(\mathbf{x}) \equiv \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1 + \|\mathbf{x}\|_1 \right\}, \mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m.$$

This problem does not possess any smooth component, $f \equiv 0$ in the model (G).

There are (at least) two possible smoothing approaches for this problem within our model (G):

- **A** **Full smoothing**. Take $g(\mathbf{x}) \equiv \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1 + \|\mathbf{x}\|_1$ and $h \equiv 0$.
- **B** **Partial Smoothing**. Take $g(\mathbf{x}) \equiv \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1, h(\mathbf{x}) = \|\mathbf{x}\|_1$.

We use Huber's function (i.e., Moreau Prox) to smooth $\|\cdot\|_1$.

Partial versus Full Smoothing

In the partial smoothing setting, the problem to be solved is

$$(\mathrm{PS}_\mu) \qquad \min_{\mathbf{x}} \left\{ \sum_{i=1}^{m} H_\mu(\mathbf{A}_i\mathbf{x} - b_i) + \|\mathbf{x}\|_1 \equiv g_\mu(\mathbf{x}) + \|\mathbf{x}\|_1 \right\}, \ (\mathbf{A}_i = i\text{-th row of } \mathbf{A})$$

Here g_μ is a μ-smooth approximation of g with parameters $\left(\|\mathbf{A}\|^2, \frac{m}{2}, 0\right)$.

In the full smoothing setting, the smooth problem to be solved is

$$(\mathrm{FS}_\mu) \qquad \min_{\mathbf{x}} \left\{ \sum_{i=1}^{m} H_\mu(\mathbf{A}_i\mathbf{x} - b_i) + \sum_{j=1}^{n} H_\mu(x_j) \equiv h_\mu(\mathbf{x}) \right\}.$$

Here h_μ is a μ-smooth approximation of h with parameters $\left(\|\mathbf{A}\|^2 + 1, \frac{m+n}{2}, 0\right)$.

Partial versus Full Smoothing

In the partial smoothing setting, the problem to be solved is

$$(\mathrm{PS}_\mu) \qquad \min_{\mathbf{x}} \left\{ \sum_{i=1}^{m} H_\mu(\mathbf{A}_i\mathbf{x} - b_i) + \|\mathbf{x}\|_1 \equiv g_\mu(\mathbf{x}) + \|\mathbf{x}\|_1 \right\}, \ (\mathbf{A}_i = i\text{-th row of } \mathbf{A})$$

Here g_μ is a μ-smooth approximation of g with parameters $\left(\|\mathbf{A}\|^2, \frac{m}{2}, 0\right)$.

In the full smoothing setting, the smooth problem to be solved is

$$(\mathrm{FS}_\mu) \qquad \min_{\mathbf{x}} \left\{ \sum_{i=1}^{m} H_\mu(\mathbf{A}_i\mathbf{x} - b_i) + \sum_{j=1}^{n} H_\mu(x_j) \equiv h_\mu(\mathbf{x}) \right\}.$$

Here h_μ is a μ-smooth approximation of h with parameters $\left(\|\mathbf{A}\|^2 + 1, \frac{m+n}{2}, 0\right)$.

Notes

1. No need to smooth the l_1 part $\|\mathbf{x}\|_1$, since in that case one can directly invoke a fast proximal gradient method, e.g., FISTA.

Partial versus Full Smoothing

In the partial smoothing setting, the problem to be solved is

$$(\mathrm{PS}_\mu) \qquad \min_{\mathbf{x}} \left\{ \sum_{i=1}^{m} H_\mu(\mathbf{A}_i\mathbf{x} - b_i) + \|\mathbf{x}\|_1 \equiv g_\mu(\mathbf{x}) + \|\mathbf{x}\|_1 \right\}, \ (\mathbf{A}_i = i\text{-th row of } \mathbf{A})$$

Here g_μ is a μ-smooth approximation of g with parameters $\left(\|\mathbf{A}\|^2, \frac{m}{2}, 0\right)$.

In the full smoothing setting, the smooth problem to be solved is

$$(\mathrm{FS}_\mu) \qquad \min_{\mathbf{x}} \left\{ \sum_{i=1}^{m} H_\mu(\mathbf{A}_i\mathbf{x} - b_i) + \sum_{j=1}^{n} H_\mu(x_j) \equiv h_\mu(\mathbf{x}) \right\}.$$

Here h_μ is a μ-smooth approximation of h with parameters $\left(\|\mathbf{A}\|^2 + 1, \frac{m+n}{2}, 0\right)$.

Notes

1. No need to smooth the l_1 part $\|\mathbf{x}\|_1$, since in that case one can directly invoke a fast proximal gradient method, e.g., FISTA.
2. Not advisable to consider the partial smoothing approach in the opposite way... Computing a proximal mapping of the l_1-fitting term seems to be as difficult as solving the original problem!

Summary of Results–Average over 100 Realizations

To compare the two approaches, we performed Monte-Carlo runs on random $(\mathbf{A}, \mathbf{b})$.

- $\mathbf{x}_{\rm FS}, \mathbf{x}_{\rm PS}$ outputs of FISTA after N iterations for each realization of $(\mathbf{A}, \mathbf{b})$.
- M^* Optimal value of original $l_1 - l_1$ problem via Sedumi
- Err-FS $= M(\mathbf{x}_{\rm FS}) - M^*$
- Err-PS $= M(\mathbf{x}_{\rm PS}) - M^*$

N	Err-FS	Err-PS	Err-FS/Err-PS
100	3.2951	1.3722	2.7152
200	1.0009	0.2740	5.0633
400	0.1741	0.0284	22.4585

Clearly, the partial smoothing approach is superior to the full smoothing approach:

1. Reaches better accuracies for a given number of iterations
2. The error in function values of the full smoothing setting is more than 22 times the error obtained by the partial smoothing setting.

Theoretical Justification

Suppose that we wish to solve a problem of the form

$$\min_{\mathbf{x}}\{g(\mathbf{x}) + q(\mathbf{x})\},$$

where *both* g and q are convex and *nonsmooth functions*

- **Full smoothing (FS):** $\min_{\mathbf{x}}\{g_\mu(\mathbf{x}) + q_\mu(\mathbf{x})\}$
- **Partial Smoothing (PS):** $\min_{\mathbf{x}}\{g_\mu(\mathbf{x}) + q(\mathbf{x})\}$
- Here g_μ, q_μ are μ-smooth approximation of g, q with their respective parameters (α_g, β_g, K_g) and (α_q, β_q, K_q).

Theoretical Justification

Suppose that we wish to solve a problem of the form

$$\min_{\mathbf{x}}\{g(\mathbf{x}) + q(\mathbf{x})\},$$

where *both* g and q are convex and *nonsmooth functions*

- **Full smoothing (FS):** $\min_{\mathbf{x}}\{g_\mu(\mathbf{x}) + q_\mu(\mathbf{x})\}$
- **Partial Smoothing (PS):** $\min_{\mathbf{x}}\{g_\mu(\mathbf{x}) + q(\mathbf{x})\}$
- Here g_μ, q_μ are μ-smooth approximation of g, q with their respective parameters (α_g, β_g, K_g) and (α_q, β_q, K_q).

Result: Apply a fast iterative method $\mathcal{M}$ on both problems with initial $\mathbf{x}_0 = 0$. Applying our results it can be shown that

$$N(FS) > N(PS)$$

- $N(FS)$:=lower bound on No. of iterations required to obtain an ε-optimal solution of problem (FS) with the fast method $\mathcal{M}$
- $N(PS)$:= lower bound on No. of iterations required to obtain an ε-optimal solution of (PS) with the same fast method.

Conclusion

Conclusion

To Smooth or Not to Smooth?

At least from last example and results....

Conclusion

To Smooth or Not to Smooth?

At least from last example and results....

Smoothing is a valuable approach to tackle nonsmooth problems, but it should be used "moderately" and only when truly necessary!

Conclusion

To Smooth or Not to Smooth?

At least from last example and results....

Smoothing is a valuable approach to tackle nonsmooth problems, but it should be used "moderately" and only when truly necessary!

Thank you for listening!

A. Beck and M. Teboulle.
Smoothing and First Order Methods: A Unified Framework.
SIAM J. Optimization – to appear