Nonsmooth Optimization

in 30 minutes

February 15, 2013

Napsu Karmitsa

Department of Mathematics and Statistics
University of Turku, Finland

email: *napsu@karmitsa.fi*

Contents

Preliminaries

- *Nonlinear Programming*.

The Goal

- Attendees should know the basic concepts of nonsmooth analysis and optimization. That is, *subdifferential*, *subgradient* and *optimality conditions*.

INTRODUCTION TO NONSMOOTH OPTIMIZATION

- **Nonsmooth optimization (NSO)** refers to the general problem of minimizing (or maximizing) functions that are typically **not differentiable at their minimizers** (or maximizers).

- Let us consider the NSO problem of the form

$$\begin{cases} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in G, \end{cases}$$

  where the objective function $f : G \to \mathbb{R}$ is supposed to be locally Lipschitz continuous on the feasible set $G \subseteq \mathbb{R}^n$.

- Note that **no differentiability or convexity** assumptions are made.

Introduction to Nonsmooth Optimization (Cont.)

NSO problems arise in *many fields of applications*, for example in

- image denoising,
- optimal control,
- neural network training,
- data mining,
- economics, and
- computational chemistry and physics.

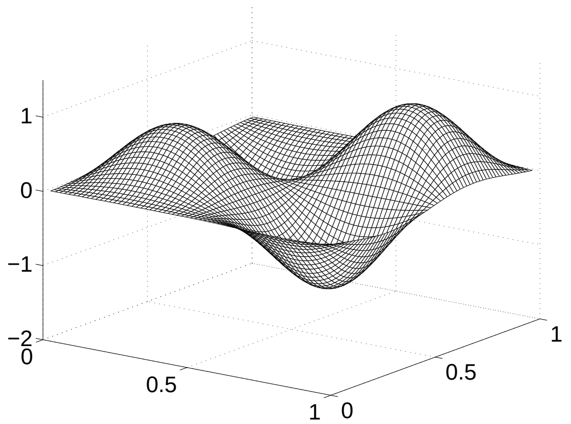
Moreover, using certain important methodologies for *solving difficult smooth problems* leads directly to the need to solve nonsmooth problems. This is the case, for instance in

- decompositions,
- dual formulations, and
- exact penalty functions.

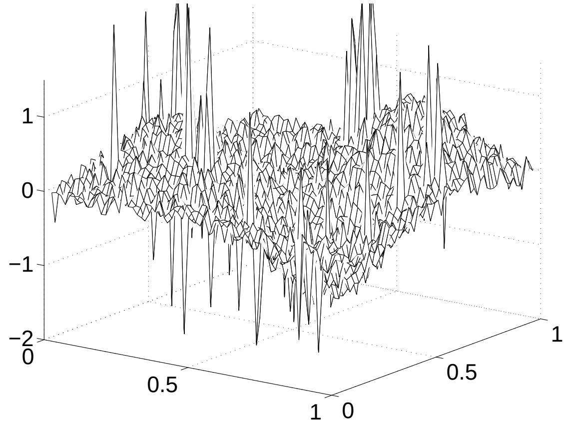
Finally, there exist so called *stiff problems* that are analytically smooth but numerically nonsmooth.

Introduction to Nonsmooth Optimization (Cont.)

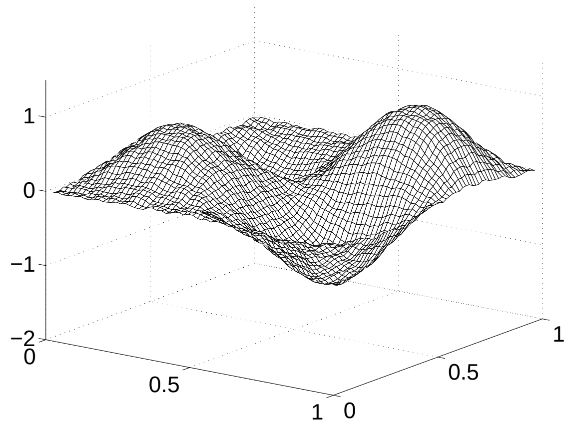
Example — Image Denoising



(a) True image



(b) Noisy image



(c) Result image

Figure 1: True and noisy images and result of NSO solver `LMBM` for formulation with L^1 fitting and smooth regularization ($n = 63 \times 63$).

Difficulties Caused by Nonsmoothness

Smooth problem:

- Descent direction is obtained at the opposite direction of the gradient $\nabla f(\boldsymbol{x})$.
- The necessary optimality condition $\nabla f(\boldsymbol{x}) = 0$.
- Difference approximation can be used to approximate the gradient.

Nonsmooth problem:

- The gradient does not exist at every point, leading to difficulties in defining the descent direction.
- Gradient usually does not exist at the optimal point.
- Difference approximation is not useful and may lead to serious failures.
- The (smooth) algorithm does not converge or it converges to a non-optimal point.

NONSMOOTH ANALYSIS: CONVEX ANALYSIS

DEFINITION. The *subdifferential of a convex function* $f : \mathbb{R}^n \to \mathbb{R}$ at $\boldsymbol{x} \in \mathbb{R}^n$ is the set $\partial_c f(\boldsymbol{x})$ of vectors $\boldsymbol{\xi} \in \mathbb{R}^n$ such that

$$\partial_c f(\boldsymbol{x}) = \{\, \boldsymbol{\xi} \in \mathbb{R}^n \mid f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{\xi}^T(\boldsymbol{y} - \boldsymbol{x}) \text{ for all } \boldsymbol{y} \in \mathbb{R}^n \,\}.$$

Each vector $\boldsymbol{\xi} \in \partial_c f(\boldsymbol{x})$ is called a *subgradient* of f at $\boldsymbol{x}$.

THEOREM. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then the classical directional derivative $f'(\boldsymbol{x}; \boldsymbol{d})$ exists in every direction $\boldsymbol{d} \in \mathbb{R}^n$ and for all $\boldsymbol{x} \in \mathbb{R}^n$

(i) $f'(\boldsymbol{x}; \boldsymbol{d}) = \max \{\, \boldsymbol{\xi}^T \boldsymbol{d} \mid \boldsymbol{\xi} \in \partial_c f(\boldsymbol{x}) \,\}$ for all $\boldsymbol{d} \in \mathbb{R}^n$, and

(ii) $\partial_c f(\boldsymbol{x}) = \{\, \boldsymbol{\xi} \in \mathbb{R}^n \mid f'(\boldsymbol{x}, \boldsymbol{d}) \geq \boldsymbol{\xi}^T \boldsymbol{d} \text{ for all } \boldsymbol{d} \in \mathbb{R}^n \,\}$.

THEOREM. If $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, then for all $\boldsymbol{y} \in \mathbb{R}^n$

$$f(\boldsymbol{y}) = \max \{\, f(\boldsymbol{x}) + \boldsymbol{\xi}^T(\boldsymbol{y} - \boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{\xi} \in \partial_c f(\boldsymbol{x}) \,\}.$$

NONSMOOTH ANALYSIS: NONCONVEX ANALYSIS

DEFINITION (Clarke). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at $\boldsymbol{x} \in \mathbb{R}^n$. The *generalized directional derivative* of f at $\boldsymbol{x}$ in the direction $\boldsymbol{d} \in \mathbb{R}^n$ is defined by

$$f^\circ(\boldsymbol{x};\boldsymbol{d}) = \limsup_{\substack{\boldsymbol{y}\to\boldsymbol{x} \\ t\downarrow 0}} \frac{f(\boldsymbol{y}+t\boldsymbol{d}) - f(\boldsymbol{y})}{t}.$$

DEFINITION (Clarke). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at a point $\boldsymbol{x} \in \mathbb{R}^n$. Then the *subdifferential* of f at $\boldsymbol{x}$ is the set $\partial f(\boldsymbol{x})$ of vectors $\boldsymbol{\xi} \in \mathbb{R}^n$ such that

$$\partial f(\boldsymbol{x}) = \{\, \boldsymbol{\xi} \in \mathbb{R}^n \mid f^\circ(\boldsymbol{x};\boldsymbol{d}) \geq \boldsymbol{\xi}^T\boldsymbol{d} \text{ for all } \boldsymbol{d} \in \mathbb{R}^n \,\}.$$

Each vector $\boldsymbol{\xi} \in \partial f(\boldsymbol{x})$ is called a *subgradient* of f at $\boldsymbol{x}$.

NONSMOOTH ANALYSIS: NONCONVEX ANALYSIS (CONT.)

THEOREM. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at a point $\boldsymbol{x} \in \mathbb{R}^n$. Then

$$f^\circ(\boldsymbol{x}; \boldsymbol{d}) = \max \{ \boldsymbol{\xi}^T \boldsymbol{d} \mid \boldsymbol{\xi} \in \partial f(\boldsymbol{x}) \} \text{ for all } \boldsymbol{d} \in \mathbb{R}^n.$$

THEOREM (Rademacher). Let $S \subset \mathbb{R}^n$ be an open set. A function $f : S \to \mathbb{R}$ that is locally Lipschitz continuous on S is differentiable almost everywhere on S.

THEOREM. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at a point $\boldsymbol{x} \in \mathbb{R}^n$. Then

$$\partial f(\boldsymbol{x}) = \operatorname{conv}\{ \boldsymbol{\xi} \in \mathbb{R}^n \mid \nabla f(\boldsymbol{x}_i) \to \boldsymbol{\xi},\ \boldsymbol{x}_i \to \boldsymbol{x} \text{ and } f \text{ is differentiable at } \boldsymbol{x}_i \},$$

where $\operatorname{conv} S$ denotes the convex hull of set S.

Nonsmooth Analysis: Results and Remarks

- The subdifferential for locally Lipschitz continuous functions is a generalization of the subdifferential for convex functions: If $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, then $f'(\boldsymbol{x}; \boldsymbol{d}) = f^\circ(\boldsymbol{x}; \boldsymbol{d})$ for all $\boldsymbol{d} \in \mathbb{R}^n$, and $\partial_c f(\boldsymbol{x}) = \partial f(\boldsymbol{x})$.

- The subdifferential for locally Lipschitz continuous functions is a generalization of the classical derivative: If $f : \mathbb{R}^n \to \mathbb{R}$ is both locally Lipschitz continuous and differentiable at $\boldsymbol{x} \in \mathbb{R}^n$, then $\nabla f(\boldsymbol{x}) \in \partial f(\boldsymbol{x})$. If, in addition, $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable at $\boldsymbol{x} \in \mathbb{R}^n$, then $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$.

NONSMOOTH OPTIMIZATION

THEOREM. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at $\boldsymbol{x} \in \mathbb{R}^n$. If f attains its *local minimal value* at $\boldsymbol{x}$, then

(i) $\mathbf{0} \in \partial f(\boldsymbol{x})$ and

(ii) $f^\circ(\boldsymbol{x}; \boldsymbol{d}) \geq 0$ for all $\boldsymbol{d} \in \mathbb{R}^n$.

THEOREM. If $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, then the following conditions are equivalent:

(i) Function f attains its *global minimal value* at $\boldsymbol{x}$,

(ii) $\mathbf{0} \in \partial_c f(\boldsymbol{x})$, and

(iii) $f'(\boldsymbol{x}; \boldsymbol{d}) \geq 0$ for all $\boldsymbol{d} \in \mathbb{R}^n$.

DEFINITION. A point $\boldsymbol{x} \in \mathbb{R}^n$ satisfying $\mathbf{0} \in \partial f(\boldsymbol{x})$ is called a *critical* or a *stationary point* for f.

NONSMOOTH OPTIMIZATION:
PRACTICAL POINT OF VIEW

Usually we do not know the whole subdifferential of the function but only **one arbitrary subgradient** at each point!

$\Rightarrow$ **We need special methods to solve nonsmooth optimization problems.**

Bundle Methods

Subgradient Methods

Derivative Free Methods

Gradient Sampling Methods

Hybrid Methods

Special Methods

WHY TO USE NONSMOOTH FORMULATIONS FOR THE PROBLEMS?

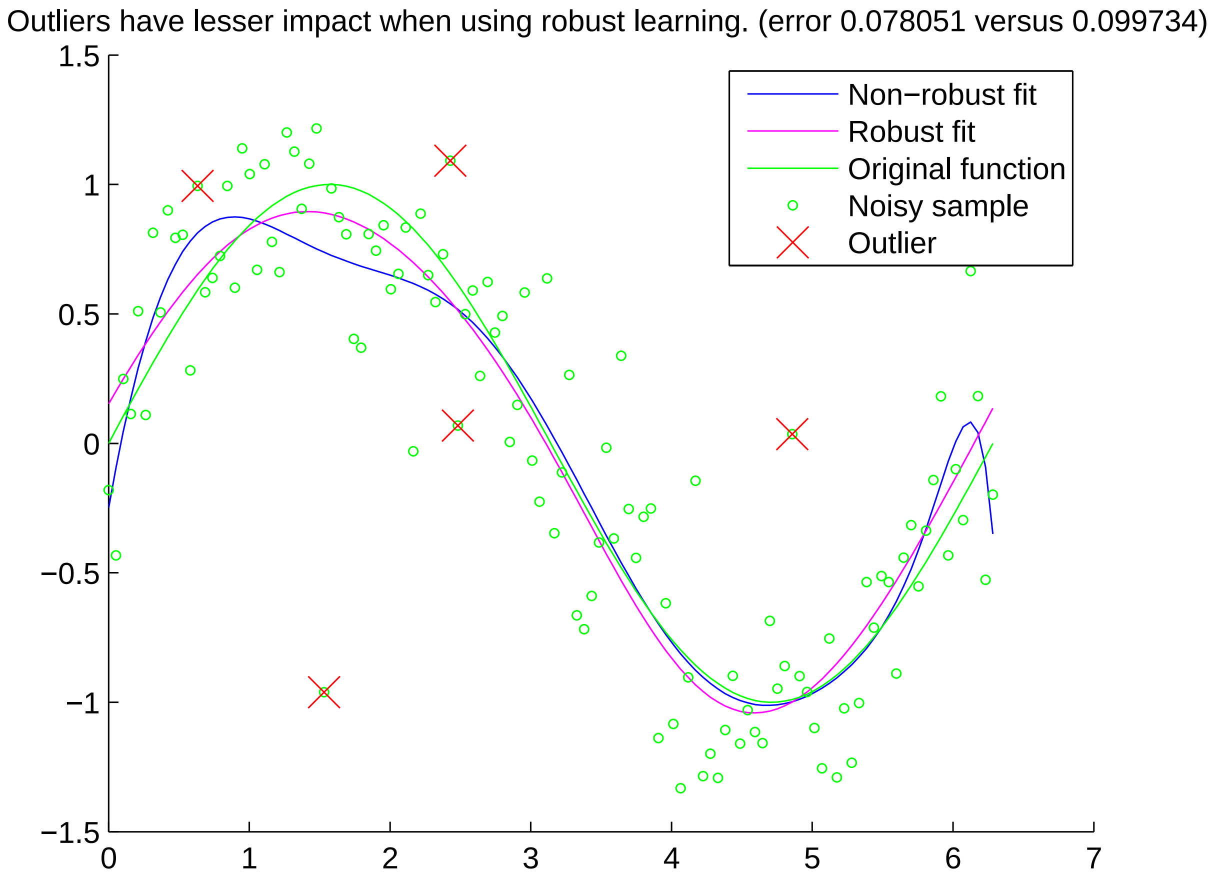


Figure 2: The robust formulations for the optimization problem arising in MLP network training: difference of the output of the traditional non-robust (smooth) data fitting and the robust (nonsmooth) data fitting when reconstructing function $f(x) = \sin(x)$.

REFERENCES

- Clarke F.H.: *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.
- Kiwiel K.C.: *Methods of Descent for Nondifferentiable Optimization*, Springer-Verlag, Berlin, 1985.
- Mäkelä M.M. ja Neittaanmäki P.: *Nonsmooth Optimization: Analysis and Algorithms with Applications to Optimal Control*, World Scientific Publishing Co. Singapore, 1992.
- Rockafellar R.T.: *Convex Analysis*, Princeton University Press, Princeton New Jersey, 1970.
- Shor N.Z.: *Minimization Methods for Non-Differentiable Functions*, Springer-Verlag, Berlin, 1985.

- Some NSO software and NSO software links can be found at

  http://napsu.karmitsa.fi/nsosoftware/